Master’s Thesis

LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT

School of Engineering Science

Industrial Engineering and Management

Business Analytics

Jenni Lunttila

PREDICTING DISEASE-SPECIFIC SURVIVAL OF COLORECTAL CANCER PATIENTS USING SERUM AND TISSUE DATA - A COMPARISON OF STATISTICAL AND MACHINE LEARNING TECHNIQUES FOR SURVIVAL ANALYSIS, IMPUTATION, AND FEATURE SELECTION

26.4.2022

Supervisors and examiners: Professor Mikael Collan, Professor Pasi Luukka

Supervisors: Docent Harri K. Mustonen (University of Helsinki), Professor of Surgery Caj Haglund (University of Helsinki)

ABSTRACT

Lappeenranta-Lahti University of Technology LUT

LUT School of Engineering Science

Industrial Engineering and Management

Jenni Lunttila

Predicting disease-specific survival of colorectal cancer patients using serum and tissue data – A comparison of statistical and machine learning techniques for survival analysis, imputation, and feature selection

Master's thesis

2022

171 pages, 43 figures, 34 tables and 23 appendices

Examiners: Professor Mikael Collan and Professor Pasi Luukka

Supervisors: Docent Harri K. Mustonen (University of Helsinki) and Professor of Surgery Caj Haglund (University of Helsinki)

Keywords: survival analysis, machine learning, imputation, Cox proportional hazards, random survival forest, colorectal cancer

Globally colorectal cancer (CRC) is the third most common cancer. The incidence rates of CRC are rising, especially in high-income countries. In Finland CRC has one of the highest mortality rates compared to other cancers. Earlier diagnosis helps to achieve better prognosis for patients. Thus, creating demand for improving diagnosis methods. Through enhanced computational capabilities the possibilities for a shift towards more patient-centred care can be developed. Advanced statistical and machine learning applications could provide as helpful in identification of biomarkers with high prognostic value and resulting in novel prospects for cancer therapy.

This thesis aims to identify variables with predictive potential and estimate survival of CRC patients. The state-of-art of survival analysis literature in the field of oncology is discussed. The issue with incomplete data is addressed using three different imputation techniques: listwise deletion, median imputation, and kNN-imputation. From those imputed datasets the important variables are identified by applying three different feature selection techniques. The sample size is artificially increased using MICE. The survival analysis is conducted utilizing Cox proportional hazards (CPH) model and random survival forests (RSF). The models are validated by holdout method and semi-stratified k-fold cross-validation (cv). The RSF models slightly outperformed CPH models. The highest performance according to c-index is obtained from kNN-imputed RSF model with log-rank splitting rule (0.751 on test data, 10-fold cv).

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT

LUT School of Engineering Science

Tuotantotalous

Jenni Lunttila

Paksu- ja peräsuolisyöpäpotilaiden tautikohtaisen elossaolon ennustaminen seerumi- ja kudosaineistosta – Vertailu tilastollisten ja koneoppimistekniikoiden välillä elinaika-analyysissä, imputoinnissa sekä muuttujavalinnassa

Diplomityö

2022

171 sivua, 43 kuvaa, 34 taulukkoa ja 23 liitettä

Tarkastajat: Professori Mikael Collan ja Professori Pasi Luukka

Ohjaajat: Dosentti Harri K. Mustonen (Helsingin yliopisto) ja Kirurgian professori Caj Haglund (Helsingin yliopisto)

Avainsanat: elinaika-analyysi, koneoppiminen, imputointi, Coxin malli, satunnaiselossaolometsä, suolistosyöpä

Paksu- ja peräsuolisyövät (suolistosyöpä) ovat kolmanneksi yleisin syöpä maailmanlaajuisesti. Näiden syöpien esiintyvyys kasvaa erityisesti teollistuneissa maissa. Yksi korkeimmista syöpäkuolleisuuksista Suomessa ilmenee suolistosyöpäpotilailla. Potilaan ennuste on sitä parempi mitä aiemmin syöpä diagnosoidaan. Tämä luo kriittisen tarpeen diagnosointimenetelmien kehittämiselle. Laskentakapasiteetin kehittymisen myötä mahdollistuu muutos kohti potilaskeskeisempää hoitoa. Edistyneet tilastolliset ja koneoppimisen sovellukset auttavat ennustavien biomarkkereiden tunnistamisessa ja täten johtavat uusien syöpäterapioiden kehittämiseen.

Tämän diplomityön tavoitteena on tunnistaa ennustamiseen soveltuvia muuttujia ja arvioida suolistosyöpäpotilaiden elinaikaa. Työssä tarkastellaan aiempaa onkologista elinaika-analyysikirjallisuutta. Epätäydellisen aineiston haasteisiin vastataan kolmella imputointimetodilla: epätäydellisten rivien poisto, mediaani-imputointi ja kNN-imputointi. Täydennetyistä aineistoista tärkeimmät muuttujat tunnistetaan kolmella piirteenvalintatekniikalla. Aineistokokoa laajennetaan keinotekoisesti MICE-tekniikalla. Elinaika-analyysiin käytetään Coxin mallia (CPH) sekä satunnaiselossaolometsiä (RSF). Mallit validoidaan holdout -menetelmällä ja ositetulla k-kertaisella ristiinvalidoinnilla (rv). RSF-mallit suoriutuvat elinajan ennustamisessa CPH-malleja hieman paremmin. Paras suorituskyky c-indeksillä mitattuna saavutetaan RSF-mallilla käyttäen log-rank jakosäännöstä ja kNN-imputoitua aineistoa (0.751 testiaineistolla, 10-kertainen rv).

ACKNOWLEDGEMENTS

I want to express my gratitude to all who supported me during this thesis process. First, I want to thank all four of my thesis supervisors Mikael Collan, Pasi Luukka, Harri Mustonen and Caj Haglund for their time, guidance, and patience. This thesis process offered the challenges I've wished for in an interesting field of science. I was given an opportunity to learn more about the different statistical and machine learning applications for medical purposes and explore oncological literature.

I'm grateful for the opportunity to continue studies right after completing a M.Sc. (Econ.) Accounting furthering my education to a second master's degree in M.Sc. (Tech.) Business Analytics, and for the people who encouraged me to pursue that goal. I'm thankful for all the people I've got a change to know during these seven years in LUT. To conclude, I owe my friends and family a great debt of gratitude for their support and encouragement during my studies.

Kuopio 26.4.2022

Jenni Lunttila

LIST OF ABBREVIATIONS

AgeAtOper	Age at operation
AIC	Akaike information criterion
AJCC	American Joint Committee of Cancer
ANN	Artificial neural network
AREG	Amphiregulin
AUC	Area under curve
BICV	Balanced incomplete CV(n_v)
BM	Benchmark
BMA	Bayesian model averaging
BMU	Best matching unit
BS	Brier score
CA-125	Carbohydrate antigen 125
CDF	Cumulative distribution function
CEA	Carcinoembryonic
CH	Cumulative hazards
CNN	Convolutional neural network
CPH	Cox proportional hazards
CRC	Colorectal cancer
CRP	C-reactive protein
cv	Cross-validation
CYR61	Cysteine-rich 61

DBN	Deep belief network
DNN	Deep neural network
DSS	Disease-specific survival
e.g.	exempli gratia
FAP	Familial adenomatous polyposis
FDR	False discovery rate
FN	False negatives
FP	False positives
GA	Genetic algorithm
GBC	Gradient boosting classifier
h(t)	Hazard function
H(t)	Cumulative hazard function
h(t)	Hazard function
HGF	Hepatocyte growth factor
HNPCC	Hereditary non-polyposis colorectal cancer
HR	Hazard ratio
i.e.	id est
IBS	Integrated Brier score
IHC	Immunohistochemistry
IL-9	Interleukin-9
IP-10	Interferon-induces protein-10
KLK13	Kallikrein 13
KM	Kaplan-Meier

kNN	k-nearest neighbour
LASSO	Least absolute shrinkage and selection operator
LOOCV	Leave-one-out cross-validation
LPC	Lassoed principal components
MAR	Missing at random
MCCV	Monte Carlo cross-validation
MCMC	Markov chain Monte Carlo
md	Minimal depth
MI	Multiple imputation
MICE	Multiple chained equations
MIP-1b	Macrophage inflammatory protein-1beta
ML	Machine learning
MLP	Multilayer perceptron
MMP	Matrix metalloproteinase
MNAR	Missing not-at random
mtry	Number of variables randomly selected as candidates for a splitting a node
MUC16	Mucin 16
NM	Naive Bayes
nodesize	terminal node size
NPV	Negative predictive value
OOB	Out-of-bag
PH	Proportional hazards
PL	Product-limit

PPV	Positive predictive value
Pr	Probability
RBM	Restricted Bolzmann machine
RF	Random forest
RNA	Ribonucleic acid
ROC	Receiver operating characteristics
RSF	Random survival forest
S(t)	Survival function
S(t)	Survival function at time t
S100A11	Calcium binding protein
SCF1	Stem cell factor 1
Se	Sensitivity
SOM	Self-Organizing map
Sp	Specificity
SPARC	Secreted protein acidic and rich in cysteine
SSVS	Stochastic search variable selection
SVM	Support vector machine
T	Event time
t	Observed time
TATI	Tumour-associated trypsin inhibitor
TN	True negatives
TNM	Tumour node metastasis
TP	True positives

TXLNA	Taxilin alpha
UICC	Union for International Cancer Control
VIMP	Variable importance
WHO	World Health Organization
WISP1	WNT1-inducible-signaling pathway protein

Table of Contents

List of appendices

APPENDIX 22. Significant markers identified by RSF models.

APPENDIX 23. DeepSurv: a preliminary artificial neural network approach for survival analysis on CRC patient data.

List of figures

List of Tables

1 Introduction

1.1 Background of the study

Cancer is the second most common cause of death globally (World Health Organization: Regional Office for Europe, 2020). In Finland every fourth death is cancer-related (Suomen virallinen tilasto (SVT), 2020). Incidence rates of colorectal cancer (hereinafter CRC) are rising, especially in countries with high human development index (World Health Organization: Regional Office for Europe, 2020). Globally CRC is the third most common cancer (World Health Organization: Regional Office for Europe, 2020). In Finland CRC has one of the highest mortality rates together with lung, prostate and breast cancer (Pitkäniemi *et al.*, 2021). Earlier diagnosis improves the prognosis of CRC patients (Colores, 2022). Thus, it is crucial to develop methodologies to help to diagnose CRC at its early stages. Currently central tool used for prediction of survival is the stage of disease scale (Terveyskirjasto, 2018).

Like many other fields of industry, medicine is also rapidly transformed by the possibilities offered by enhancements achieved in computational capabilities. There is a shift towards more patient-centred, tailored care. A possible step forward towards this goal of more personalized medicine could be identification of potential target molecules for novel, improved medical therapies. This could be achieved through advanced machine learning solutions. The aim of these current studies is to identify biomarkers with high prognostic value. This then could help to identify new therapeutic possibilities for cancer therapy. More personalized, patient-centred solutions, more immediate analyses, possibility to accurately predict future, opportunity to save lives with earlier diagnoses, save costs and resources (Bjarnadottir *et al.*, 2018). Get better insight how human bodies function, and how different illnesses affect our bodies. More accurate and earlier diagnoses result in better treatment and prognosis for patients.

Survival analysis attempts to predict the time until a certain event occurs (Englebert, Quinn and Bichindaritz, 2017). In medical studies this event typically is death or recurrence. With this type of survival data there exists observations that do not experience the event of interest within the observed time. These observations are referred as right-censored (Katzman *et al.*, 2018). One

of the challenges in survival analysis posed by the panel data used for it, is that the datasets are collected during long periods of time (Wang, Li and Reddy, 2017). Thus, the models that could accurately predict survival using only a limited time-to-event data are needed. In social sciences survival analysis is also known as event history analysis and as reliability analysis in engineering (Allignol and Latouche, 2021).

Working with real world data almost inevitably means working with incomplete data. This missing data can cause unexpected effects to the analysis. Accurate imputations are crucial since bad imputations lead to poor results and lack of generalizability. Therefore, creating a demand for techniques to handle this kind of missingness. A possible solution of removing observations with missing values might lead to ill-fitted models and incorrect results. Hence the use of methods for imputing these missing values through similarities to existing data are investigated. Imputed values are sophisticated predictions which can possess error, and this error cumulates in the analysis.

Despite many years of research and the possibilities offered by modern technology, the survival rate for patients diagnosed with colorectal cancer (CRC) is poor because the disease is often diagnosed not until at later stages (World Health Organization: Regional Office for Europe, 2020). This creates demand for studies identifying possibilities for earlier diagnosis of cancer and personalized forms of therapy. It is crucial to understand the pathogenesis of CRC to fully be able to develop computer-based analysis solutions for it. This way the correct markers could be chosen for further investigation. Specific features are connected to distinct types of cancer when those show evidence of increased levels in the body. These are referred as tumour markers. (National Cancer Institution, 2020) The objective of this thesis is to identify these features from our dataset which could potentially be used as markers for CRC.

1.2 Focus and research questions

The main objective of this thesis is to compare few selected approaches for prediction of survival of CRC patients. Also a couple methods for handling missing values and selecting the best predictive features from the data are compared. These techniques are selected based on the conducted literature review. As a secondary goal this thesis attempts to identify clinicopathologic prognostic factors which could function as markers for estimating the survival of CRC patients. Here term clinicopathologic refers to data collected from pathological laboratory examinations of different tissue, serum and blood samples collected from CRC patients. This data is then used to build a predictive model. Pathological examinations as a branch of medical science, typically include examinations of tissue samples, or cadavers in the case of performing an autopsy (McGill University, 2021).

The dataset used consists of 318 Finnish colorectal cancer (CRC) patients. From the patients, serum and tissue samples were collected and documented. More detailed description of the data is provided later in chapter 3.1. The research questions are as follows where the first is the main research question which is supported by following sub-research questions. The answers for these research questions are summarised in chapter 4.2.

How the survival of CRC patients could be modelled and predicted?

What is the state-of-art literature of survival analysis studies in oncology?

How to handle missing values whilst ensuring to preserving the characteristics of the data?

How to perform feature selection to select the features with the most predictive potential?

Which biomarkers are associated with high prognostic value for predicting survival of CRC patients?

1.3 Research methods

The research methods applied for this thesis are literature review and survival analysis techniques together with data preprocessing and validation methods. Previous literature is consulted forming the state-of-art of the research. This will form a theoretical framework for the thesis. The research for this thesis is mainly quantitative. The data used in this thesis consists of 318 Finnish CRC patients, and it is offered by University of Helsinki.

Literature review is conducted by doing several searches to multiple databases using relevant search terms. This process is documented and presented to enhance repeatability and thus reliability of the results. The focus is in presenting the previous academic literature around the survival analysis of cancer patients. Literature review provides an answer one of the research questions, and supports decisions made in following sections. The process conducted for literature review is described in more detail in chapter 2.1. The state-of-art section consists of four (4) subsections, previous studies, CRC research, data preprocessing, survival analysis techniques, and model validation. Hence a broad knowledge about current and past research around CRC is presented briefly together with some key research about statistic techniques used in those studies.

The data collected from Finnish CRC patients including values for patient's serum, plasma and tissue samples is used to build a prognostic model. To focus is the prediction of patients' survival based on these biomarker values. For analysis a few imputation and feature selection techniques are compared. After these preprocessing steps survival analysis is conducted using a couple different approaches. A considerable limitation in this thesis is the size of the data which is quite small and contains multiple missing values. Because of these characteristics two (2) research questions are formed to answer challenges regarding missing values and feature selection. Deeper analysis of medical side of this is left outside of the scope since this is an engineering thesis. The focus is on the mathematical modelling and data analysis. The analysis is discussed more in chapter 3.

1.4 Structure of the thesis

Figure 1 shows the structure of this thesis. First on chapter 1 is the introduction in which the background of the study and the research questions are presented. Here the focus of the study is set. Then in chapter 2.1 state-of-art of survival analysis studies in the field of oncology are discussed. Chapter 2.2 focuses on clarifying key concepts about cancer, especial colorectal cancer (CRC). Data preprocessing techniques for imputation and feature selection are then introduced in chapter 2.3. Following the chapter 2.4 presents survival analysis methods, and chapter 2.5 focuses on approaches for validating the model. The chapter 3 then forms the empirical part of the thesis. There the data is presented in chapter 3.1. The preprocessing steps are discussed in chapter 3.2, followed by the survival analysis in chapter 3.3. Finally the conclusions are provided in chapter 4, where the summary of the thesis is in chapter 4.1, research questions are answered in chapter 4.2, reliability is assessed in chapter 4.3, and topics for further studies are discussed in chapter 0.

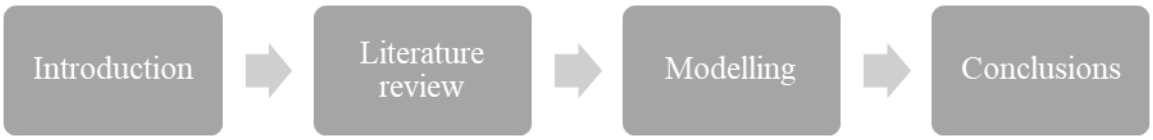


Figure 1. Structure of the study.

2 Literature review – state of the art of survival analysis studies

In this section the current state of survival analysis studies specifically around cancer is discussed. The focus will be on the studies utilizing both statistical and machine learning techniques in the field of oncology research. Also, some brief discussion about the application of survival analysis techniques outside cancer research is presented. Brief insight into colorectal cancer (CRC) is presented focusing on the classification of this cancer, reasons behind the emergence of cancer and the incidence and mortality rate both nationally in Finland and globally. Since this thesis is made as a part of technical degree, the presented medical concepts might be unfamiliar to some of the readers. For the reader it is important to comprehend these concepts in order to understand the decisions made later in modelling phase. Additionally, preprocessing techniques for imputing missing data and feature selection are presented. Then the selected methods for the analysis are introduced. To conclude this section, model validation techniques are discussed.

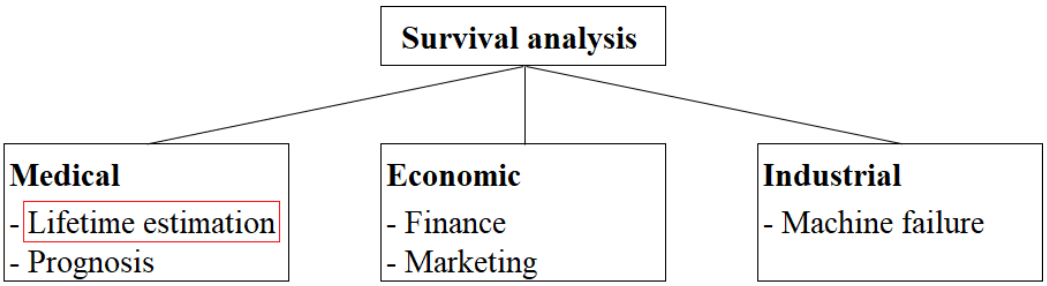


Figure 2. Example applications for survival analysis.

Survival analysis techniques have also been adopted outside medical field. The Figure 2 demonstrates some of the example fields where survival analysis could be applied to. The focus of this thesis is highlighted by a red box in the graph. In finance and marketing the customer churn prediction is one the crucial aspects of the relationship to evaluate. Lima, Mues and Baesens (2009) modelled customer churn prediction using logistic regression and decision tables. Credit scoring for evaluation of customers' credit worthiness is another important factor to consider

for businesses offering consumer finance. Pavlidis et al. (2012) conducted classification of applications for credit (CAC) using sequential and adaptive logistic regression. Baesens et al. (2005) demonstrated an another angle to credit scoring problem using neural network survival analysis. Other possible applications for survival analysis techniques discussed in the literature are bankruptcy prediction (Topaloglu and Yildirim, 2009), employee attrition and fidelity modelling (Pechacek *et al.*, 2019), and machine failure time estimation (Cai, Huang and Tian, 2009).

Besides cancer and other disease related medical research, survival analysis has been widely studied concerning predictions of transplantation success. Delen et al. (2010) compared Cox regression model, decision trees, SVMs and ANNs for prediction of thoracic transplantation success. Ershoff et al. (2020) studied post-liver transplant mortality using deep neural networks. Oztekin et al. (2009) researched decision trees, neural networks and logistic regression as classification models for the purpose of variable selection for Cox regression models to predict survival of heart-lung transplantation patients. Turgeman and May (2016) researched possibility of predicting hospital readmissions using survival analysis techniques. They developed a mixed-ensemble model combining decision trees and support vector machine obtaining accuracy rates greater than 80 %. Rahimian et al. (2018) also investigated the usage of gradient boosting classifier (GBC) and random forests (RF) compared to the traditional approach of Cox proportional hazards model as a benchmark for predicting the risk of emergency admissions.

Convolutional neural networks (hereinafter CNNs) have gained attention in recent years due to their ability to recognize patterns from medical images (Yasaka and Abe, 2018). Kather et al. (2019) used CNN for tissue sample slide images to identify specific tissue types to predict survival for CRC patients. They were able to decompose complex tissue and based on those identified prognosticators to assign a prognostic score for patients. Another study by Hosny et al. (2018) demonstrated the stratification of lung cancer patients using both 3D CNN and random forests for computer tomography (CT) images. They found the 3D CNNs to be superior to random forests in prediction performance. Furthermore, Yao et al. (2017) used CNNs to predict patient risk using pathological images. A Deep Correlational Survival model was presented for this task of analysing multi-view data. The further details of these studies are left for readers

own interest to investigate since the applications of machine learning for medical imaging is left outside of the scope of this study.

Christodoulou et al. (2019) identified an urgent need for enhancements in reporting to be able to compare different survival analysis methods for clinical risk prediction. In addition, their review demonstrated a lack of usage of calibration to assess the reliability of predictions using ML techniques. Several studies were found to have insufficient validation of results.

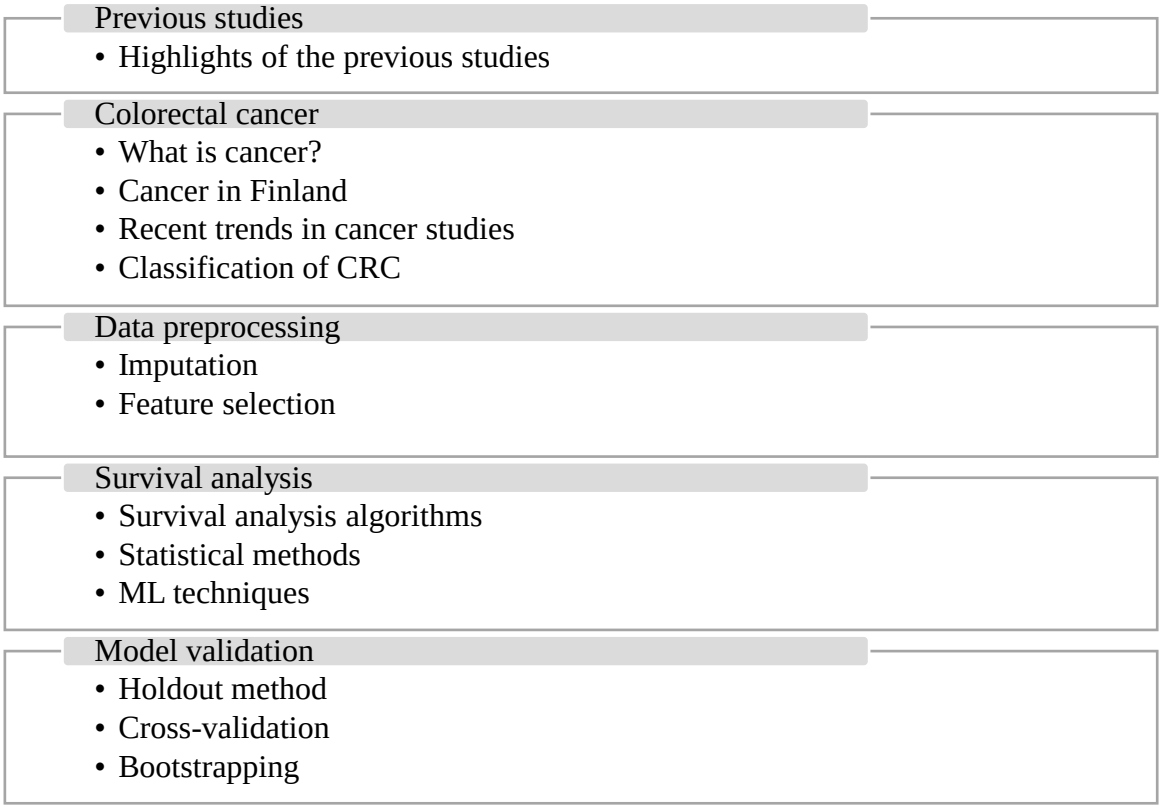


Figure 3. Structure of literature review.

2.1 Previous studies

For this literature review the focus is on studies applying ML -based survival analysis techniques on CRC patient data. However, the search is extended to the other cancers as well to provide more comprehensive review about the current state of survival analysis studies in the field of oncology. Total of six (6) databases are selected for conducting queries: ScienceDirect, Emerald Journals, EBSCO – Business Source Complete, IEEExplore, and Scopus. Some additional articles are found as referenced and suggested articles from those found using queries to aforementioned databases. These searches are performed in October 2020.

As an example, the query process in ScienceDirect is described here. In ScienceDirect all content is peer-reviewed (ScienceDirect, 2022). The search is performed in October 2020. First, a search term "survival analysis" is used which resulted in 78,568 hits. Then the search is narrowed down by selecting only research articles and adding additional search terms "machine learning" and "cancer". After these restrictions 312 hits were obtained of which 204 were in subscribed journals, i.e. the access to those is granted. Then the search is further narrowed down by only including those articles which were published in articles having JUFO classification 2 or higher, and additionally including results starting from year 2000. This resulted in 34 articles. After skimming all those 34 articles, the following 8 articles were selected for the literature review. In addition to these articles from ScienceDirect there are four (4) articles from other databases. Together this state-of-art of composed of 12 previous studies. These are displayed in Table 1, and in more detail with some additional articles in Appendix 1.

Table 1. The main articles selected for the literature review.

Author(s) and year	Name of the publication	Journal	Database
Bjarnadottir et al. (2018)	Predicting Colorectal Cancer Mortality: Models to Facilitate Patient-Physician Conversations and Inform Operational Decision Making	Production and Operations Management (2)	Wiley Online Library
Murtojärvi et al. (2020)	Cost-effective survival prediction for patients with advanced prostate cancer using clinical trial and real-world hospital registry datasets	International Journal of Medical Informatics (3)	ScienceDirect
Van Belle et al. (2011)	Support Vector Methods for Survival Analysis: A Comparison Between Ranking and Regression Approaches	Artificial Intelligence in Medicine (2)	ScienceDirect
Kleinlein & Riaño (2019)	Persistence of data-driven knowledge to predict breast cancer survival	International Journal of Medical Informatics (3)	ScienceDirect
Reijnen et al. (2020)	Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study	Plos Medicine (3)	EBSCO
Delen, Walker & Kadam (2005)	Predicting breast cancer survivability: a comparison of three data mining methods	Artificial Intelligence in Medicine (2)	ScienceDirect
Tseng et al. (2019)	Predicting breast cancer metastasis by using serum biomarkers and clinocopathological data with machine learning technologies	International Journal of Medical Informatics (3)	ScienceDirect
Zupan et al. (2000)	Machine learning for survival analysis: a case study on recurrence of prostate cancer	Artificial Intelligence in Medicine (2)	ScienceDirect
Churilov et al. (2005)	Data Mining with Combined Use of Optimization Techniques and Self-Organizing Maps for Improving Risk Grouping Rules: Application to Prostate Cancer Patients	Journal of Management Information Systems (3)	EBSCO
Jerez-Aragonés et al. (2003)	A combined neural network and decision trees model for prognosis of breast cancer relapse	Artificial Intelligence in Medicine (2)	ScienceDirect
Ruy, Chandrasekaran & Jacob (2004)	Prognosis Using an Isotonic Prediction Technique	Management Science (3)	EBSCO
Jing et al. (2019)	A deep survival analysis method based on ranking	Artificial Intelligence in Medicine (2)	ScienceDirect

Bjarnadottir et al. (2018) implemented classification trees and relaxed regularized logistic regression to predict mortality of CRC patients both in the short-term and medium-term. They based their model on the patients clinical and demographic information. As a result, their model based on the relaxed lasso evidenced the highest AUC value for all time horizons expect for the 30-day survival. Based on their research the factors affecting early mortality on CRC patients are old age, advanced stage cancer and the high amount of co-existing diseases.

Murtojärvi et al. (2020) focused on forming a cost-effective variable selection algorithm for a prognostic model for mCRPC (metastatic castration resistant prostate cancer) patients. For this they used penalized Cox regression model with LASSO regularization approach and a greedy cost-specified variable selection algorithm. Using the budget as a hard constraint, the LASSO for variable selection performed better than the greedy algorithm without compromising the prognostic value of the survival model.

Van Belle et al. (2011) compared ranking and regression based SVM techniques for survival analysis for several real-world data sets. These data sets consisted of leukaemia, lung cancer, prostate cancer, and breast cancer patients. Cox model was selected as a benchmark and compared with two SVM approaches for censored data. One way is to rephrase the task as a ranking problem (RANKSVMC) and the other is to perform regression (SVCR). They also proposed a new model similar to RANKSVMC including regression constraints. The models with regression constraints showed better performance on high dimensional data compared to SVM-based models with ranking constraints.

Comparison between multiple different machine learning based survival analysis models were tested for forming a prognosis for breast cancer patients by Kleinlein and Riaño (2019). The ML models were developed for both as stage-specific and joint models for all the stages. For the research they selected to implement naïve Bayes, logistic regression, and decision trees. Instead of resampling they decided to use cost-sensitive meta-classifier system which seems to evidence better performance for large data sets like theirs. For all breast cancer stages the joint and stage-specific logistic regression showed highest AUC values.

A recent study by Reijnen et al. (2020) proposed a Bayesian network based prognostic model to identify endometrial cancer patients' preoperative risk together with lymph node metastasis to aid medical decision making. The Bayesian networks' ability of producing easily understandable visualizations and applicability for censored data guided their choice. The data used for modelling consisted of both clinical and pathological data and the possible follow-up information. Interestingly, the predictors were selected based on previous studies, and those together with medical expert knowledge were used to construct the initial network manually. Then the data-driven approach for model optimization was performed by using hill-climbing and Tabu search algorithm, which both are score-based machine learning algorithms. With this BN approach they were able to achieve high discriminative performance and good calibration.

A bit over a decade ago Delen, Walker and Kadam (2005) presented a study comparing three different methods for breast cancer survival prediction. These techniques were artificial neural networks, decision trees and logistic regression. They treated survival as a dependent variable and adjusted the survival rate so that the non-breast cancer related mortalities are not included to the analysis. The data had both categorical and continuous variables. Measured in prediction accuracy the ANN and decision trees performed better than logistic regression. The sensitivity analysis performed for ANN's output, the most important input variables were found to be grade and stage of cancer.

Tseng et al. (Tseng *et al.,* 2019) investigates serum biomarker and clinicopathological data using ML methods in order to predict breast cancer metastasis. They modelled 30-, 60- and 90-days prognosis whether or not cancer has become metastatic. To evaluate the most important features for the model the mean decrease in Gini index and the amount of time a certain feature became a split feature was conducted. These features were found to be age, ER expression, TNM stage, and the following biomarkers, CA15-3, CEA, and sHER2. RSF achieved best predictive performance (AUC 0.746) whilst the logistic regression performed worst (AUC 0.581).

In their article Zupan et al. (2000) emphasized the machine learning applications' lack of ability for conducting survival analysis for handling with small samples with censored data. However,

this statement was made twenty years ago. Since then, the machine learning techniques have developed and are performing better measured in the prediction accuracies. They also mentioned the possibilities offered by the development of neural network applications. Cox proportional hazards model, which is commonly used in the academic medical literature, was compared with two selected machine learning techniques, naïve Bayes classifier and decision trees in making the survival prognosis for prostate cancer patient data with right-censored observations. Kaplan-Maier curves were used to estimate the probability on non-recurrence after the operation. Decision trees (concordance index 0.72) found to underperform CPH (concordance index 0.76) and NB (concordance index 0.75).

Churilov et al. (2005) implemented an approach combining multiple optimization techniques with self-organizing maps (SOM) for prostate cancer patient data to refine the rules for their risk grouping. The goal was set to avoid unnecessary treatment by reducing the number of patients categorized into the intermediate-risk group, as the cancer treatments are often quite intensive for the patient and their loved ones. First the clustering was performed with SOM using age PSA (biochemical prostate-specific antigen) at the time of the diagnosis, Gleason score and tumour stage as inputs. Gleason score is used as a staging system for the aggressiveness of prostate cancer where 1 is the lowest and 5 the highest (Prostate Conditions Education Council, 2020) After that they repeated the process this time using optimization-based clustering on patients' survival. As a result of their analysis, the usage of unsupervised learning techniques for rule refinement improved the classification accuracy of cancer patients.

Risk for breast cancer relapse after surgery was modelled using a system combining decision trees and ANNs (Jerez-Aragonés *et al.*, 2003). For right-censored observations, for which the relapse did not occur during the observation period, they classified those incidents using a dummy variable as "non-relapse" and the others as "relapse". Feature selection, from the set of previously chosen features by medical experts, before the actual modelling was performed using control of induction by sample division method, CIDIM which was originally presented by Ramos-Jiménez, Morales-Bueno and Villalba-Soria (2000). They used decision trees to select the most important prognostic factors and then use those as input for the neural network (MLP, multilayer perceptron) which gave as an output the probability of relapse. Cox regression

model was used as a benchmark model and NN seemed to perform slightly better as a prognosis model for all the chosen time intervals.

An article by Ryu, Chandrasekaran and Jacob (2004) is a bit of focus for this thesis but it proposes an interesting technique. They demonstrated an isotonic prediction technique used to a prediction problem of breast cancer patient data with multiple censored values. The final model was tested also using heart attack data. For the estimation of patients' survival, they used the isotonic separation technique which uses survival time points instead of fixed time frame survival functions. For this technique, the model is required to fulfil the monotonicity condition and isotonic consistency condition. The approach is similar to the cost closure problem as it tries to minimize the misprediction penalty functions. Backward sequential feature elimination process applied for feature selection which resulted selection of only 2 features out of possible 31 for the breast cancer dataset. The model was able to achieve the classification accuracy of around 90 %.

2.2 Colorectal cancer and its classification

More than one in three Finnish citizens get cancer during their lifetime. Of those about half (~20 % of Finnish population) die from cancer. On the global level cancer altogether is the second most common cause of death. As the infrastructure and other aspects of sufficient standards of living develop globally, the cancer-related mortality is expected to rise. (Pitkäniemi J, Malila N, Virtanen A, Degerlund H, Heikkinen S, 2020; World Health Organization: Regional Office for Europe, 2020)

Cancer cells in some part of the body start to reproduce uncontrollably. This is triggered by cell mutations caused by a factor called carcinogens. The risk for the mutations predisposing malignancy grows as the people age. The intense rate of growth is achieved through dismissing mechanisms controlling cell proliferation. This is one of the six hallmarks of cancer which are discussed briefly in following section. Malignant cells spread to other parts of the body through

a process called the invasion-metastasis cascade. Typically, the prognosis is better and there are more possible forms of therapies before the metastasis phase. That is one of the reasons why the early diagnosis is crucial to the survival of the cancer patient. (Fidler, 2003; Talmadge and Fidler, 2010; Vicente-Dueñas *et al.*, 2013; Lambert, Pattabiraman and Weinberg, 2017)

Cancerous cells appear first in one organ of the body, but later they form metastasis and spread to other parts as well. The concept of the six hallmarks of cancer was first introduced in 2000. In 2011 the revised version was presented. The hallmarks represent the nature of most cancers. These are sustaining proliferative signalling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, and resisting cell death (e.g. apoptosis). Basically, this means that the malignant tumour is capable of growing intensively without restrictions posed normally to cells and grow new blood vessels and alter existing ones to guarantee sufficient intake of nutrients and oxygen. Invasion and metastasis allow the tumours to spread to other parts of the body through the body's vascular system. Here again the connection between inflammation and cancer can be observed as the inflammation upholds many of these hallmarks. The further details of these hallmarks are not discussed here, and it is left for the readers' own interest to find out more about the biological side of this. Still today those hallmarks function as a key part of determining and developing the right therapies for cancer patients. (Hanahan and Weinberg, 2000, 2011)

2.2.1 Rising trends in cancer research

There has been a rising trend in predictive cancer studies during recent years. Querying from ScienceDirect database using search words "cancer" and "inflammation", selecting the hits from the last five (5) years (2017 – 2021) with focus on research articles. This increasing trend is demonstrated as a graph in Figure 4. The future of oncology is about digging deeper into the molecular structures of cancer cells (Murciano-Goroff *et al.*, 2020). More personalized cancer therapies could be achieved through research of predictive biomarkers (Calon *et al.*, 2012; Sreekumar *et al.*, 2018; Kilgour *et al.*, 2020). The information about the usage of these biomarkers could be used as guidance in addition to existing stage classification, e.g. AJCC

(American Joint Committee of Cancer), in decision-making for therapies. Biomarkers provide insight into the tumour which traditional AJCC classification fails to deliver (Guinney *et al.*, 2015; Chatterjee *et al.*, 2019).

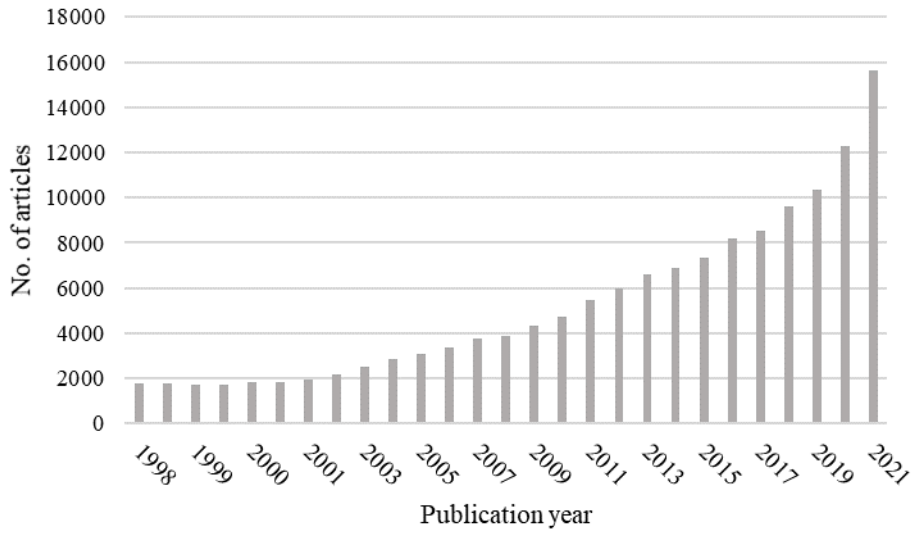


Figure 4. Number of hits from ScienceDirect for query "cancer" AND "inflammation".

A recent study (Chen *et al.*, 2020) demonstrated detection of cancer from blood sample up to four years before symptoms and the actual diagnosis. A similar study identified specific driver genes for early detection of cancer (Gerstung *et al.*, 2020), while another study (Tayel *et al.*, 2018) investigated the role of microRNAs in CRC. Also single-cell sequencing (SCS) and research concerning the role of cytokines has delivered promising results (Sino Biological, 2019; Lim, Lin and Navin, 2020). There has been proof of successful utilization of cytotoxic agents as a curative treatment for cancer (Weber, 2002). Cancer-targeted therapy is a step towards more personalized treatment of cancer patients. This treatment is based on the patient's tumour's characteristics (Mereiter *et al.*, 2019).

There will be a special focus on inflammation and cancer in future, since many of the markers determined in serum and tissues are inflammation related (Balkwill, Charles and Mantovani, 2005; Bromberg and Wang, 2009; Rumba *et al.*, 2018). The role of inflammation is estimated

to be more than 25 % of known cancer predisposing factors (Hussain and Harris, 2007). In cancer we see two different kinds of inflammation, the local and the systemic one. Local inflammation leads to increased infiltration of different inflammatory cells or changes of the expression of different inflammatory markers (Rumba *et al.*, 2018). The local changes may be part of the host' defence against the tumour or they may be initiated by the tumour to help it infiltrate and metastasize. (Kasurinen *et al.*, 2020) Part of tumours cause a systemic inflammatory response in the host. It is seen as elevated levels of several inflammatory markers in the circulation (Balkwill, Charles and Mantovani, 2005). A marked systemic inflammation in cancer patients is considered having a negative prognostic effect (Rumba *et al.*, 2018).

2.2.2 Cancer in Finland

According to the Finnish Cancer Register (2020) most common cancers are breast cancer, prostate cancer, colorectal cancer and lung cancer. The incidence rate of cancer in Finland has been increasing steadily during the last decades. However, the relative survival rate has also been on the rise because of the advanced therapies and earlier diagnoses. Table 2 displays the age-standardized incidence, mortality, and prevalence rates for different cancer types. Incidence is the new cancer cases, and prevalence demonstrates the proportion of persons alive with past cancer diagnosis (Suomen Syöpärekisteri, 2021). For the table are selected three (3) most common cancer types in Finland which are breast cancer for women (f for female), prostate cancer for men (m for male) and lung cancer. Gastric cancer is included here since it has one of the highest mortality rates (Pitkäniemi J, Malila N, Virtanen A, Degerlund H, Heikkinen S, 2020). The incidence and mortality rates for male breast cancer are also included despite the fact that those numbers are low. Colon cancer and rectal cancer are included in the table as those are in the focus of this thesis. Since Finnish Cancer Register registers colon and rectal cancers separately both of those are included to this table. In this thesis colon and rectal cancer are included in the same dataset as CRC. By using age-standardized incidence and mortality rates the effect of changes in population's age distribution and the numbers remain comparable between different decades (Tilastokeskus, 2020). The rates are calculated as a number of incidents per 100,000 people while keeping the age structure the same (here Finland 2014) during the whole observation period.

Table 2. Age-standardized incidence and mortality rates in Finland in 2018 (adapted from (Pitkäniemi J, Malila N, Virtanen A, Degerlund H, Heikkinen S, 2020)).

	Breast cancer	Prostate cancer	Colon cancer	Rectal cancer	Lung cancer	Gastric cancer
Incidence rate	165.66 (f) 1.24 (m)	190.68 (m)	34.45 (f) 42.34 (m)	16.71 (f) 29.79 (m)	32.2 (f) 64.53 (m)	7.85 (f) 13.81 (m)
Mortality rate	27.14 (f) 0.2 (m)	39.18 (m)	11.77 (f) 16.86 (m)	5.95 (f) 10.24 (m)	24.57 (f) 56.37 (m)	4.88 (f) 9.58 (m)
Prevalence	2345.7 (f) 11.1 (m)	2134.9 (m)	276.4 (f) 302.7 (m)	141.6 (f) 217.1 (m)	80.7 (f) 118.6 (m)	44.5 (f) 57.9 (m)

This thesis focuses on CRC. The following sections discuss shortly about these conditions. We will be examining the characteristics of CRC together with its classification systems. The stage of the diseases works as one of the most important predictors in survival analysis. Early diagnosis is crucial for patient's survival.

2.2.3 Classification of cancer

The stage of disease is found to partly explain the survival of patients. Cancer classification is affected by tumour biology and host-dependent factors. Classification estimated prior to the actual treatment affects the selected therapies and the prognosis of the patient. Common classification method for cancer stage is the international UICC/AJCC TNM classification which stands for Tumour, Node, and Metastasis. Other classification schemes are Dukes classification for CRC, the Lauren classification for gastric cancer and the histological tumour classification by WHO. Lauren (1965) presented his histologic classification of gastric carcinoma in 1965. This Lauren classification divides the gastric cancer tumours into two types, the intestinal, and the diffuse type. Later a third type was introduced, a mixed-type tumour. The recent WHO

(2019) classification of tumours of the digestive system defines tumour types using their molecular structures. Here the focus will be on the TNM and Dukes classifications. (Leocata *et al.*, 1998; Berlth *et al.*, 2014; Amin *et al.*, 2017)

Originally TNM staging was proposed by a French doctor between 1940s and the early 1950s. A couple of years later in 1968 Union for International Cancer Control (UICC) published their first edition of TNM classification guidelines (UICC, 2021b). In 1982 American Joint Committee of Cancer (AJCC) developed their own version of TNM staging system with separate definitions (AIJCC, 2021; UICC, 2021b). Five years later, in 1987 both of these TNM classification systems by UICC and AJCC were merged into one (UICC, 2021a). TNM is used to classify tumours based on actual tumour, lymph nodes, and the presence of metastases. The classification has five (5) stages where 0 means the presence of a small local tumour and no metastases, and in stage IV the cancer has multiple metastasis. These classes can be divided further into subclasses providing more specific information about the aspect in question. (Jiang *et al.*, 2015; Amin *et al.*, 2017)

An older classification specifically for CRC from the early 1900s is the Dukes classification, later the modified Dukes classification which is more or less similar to the TNM one. Then there were only three categories, A to C. (Dukes, 1932) In 1954 an Astler-Coller classification was proposed as a modified version of Dukes in which B and C were split into two sub-categories (Astler and Coller, 1954). Later the classification is estimated to four (4) classes A-D where A is the least invasive and D is cancer with the furthest metastasis (Cancer Research UK, 2018). Dukes classification is used in the CRC data in this thesis. It is used to classify CRC. In Table 3 the different classes for CRC and their characteristics are shortly described.

Table 3. Dukes classification (adapted from (Dukes, 1932; Cancer Research UK, 2018)).

Stage	Description
A	not invasive, no metastases
B	cancer invaded through bowel wall, no metastasis
C	regional lymphatic metastases
D	distant metastases

To clarify some of these notions for a reader not familiar with medical concepts the biological idea behind the TNM classification is briefly described here. For different types of cancer, the aspects of TNM classification vary in prognostic value. For example, in CRC the depth of tumour invasion is crucial. Depth of tumour invasion describes the depth how far into the tissue the tumour has invaded. In the case of CRC this means whether or not the actual tumour has invaded through the bowel wall. When the cancer spreads to other part of the body from its primary location, it becomes metastatic. The lymph nodes are a part of human body's immune system which helps the body to defend against infection. There are regional lymph nodes near the original tumour site and distant lymph nodes in other parts of the body. Typically, the prognosis is better when the cancer has metastasis only in the regional lymph nodes. (American Cancer Society, 2015; Amin *et al.*, 2017; Cancer.net, 2019)

New classification schemes, especially histologic-based ones, with better prognostic value are constantly researched to guide medical professional in making therapeutic decisions. A novel approach to develop better classifications is to use gene expression patterns. The Cancer Genome Atlas (TCGA) project (Amin *et al.*, 2017) presented a molecular characterization of gastric adenocarcinoma, which is the most common type of gastric cancer. This classification divides gastric adenocarcinomas into four (4) classes which are tumours positive for Epstein-Barr virus, microsatellite unstable tumours, genomically stable tumours, and tumours with chromosomal instability. The further details of this classification are left outside of the scope of this thesis and for the readers own interest to find out more. (Alizadeh *et al.*, 2001; Bass *et al.*, 2014; Amin *et al.*, 2017)

2.2.4 Colorectal cancer

CRC is formed in the gastrointestinal tract which is part of digestive system starting from mouth and ending to anus (National Institute of Diabetes and Digestive and Kidney Diseases, 2017). Colon is divided into four (4) sections, and it serves as the last part of the gastrointestinal tract right after the small intestine. These sections starting from the end of small intestine are ascending colon, transverse colon, descending colon, and sigmoid colon. (OncoLink, 2021) Both CRC together with gastric cancer are two of the most difficult cancer types to develop effective therapies for. Typically, those cancers are diagnosed at later stage when the cancer already possibly has developed metastases.

According to the American Cancer Society's recent report (2020) on CRC, the incidence rate for new CRC cases globally, especially in high-income countries, is increasing and the patients' average age is lower than before. Currently the average age for a patient diagnosed with CRC is 66 years. This declining trend in patients age seems to be accelerating in the future. Only 5 % of diagnosed CRC cases in Finland are caused by a hereditary genetic error, e.g. Lynch syndrome (hereditary non-polyposis colorectal cancer syndrome, HNPCC) (Koskenvuo, Pöyhönen and Lepistö, 2020). Thus, through healthy lifestyle habits the risk factors for CRC can be reduced. Also screening, by conducting a colonoscopy, as a preventative measure together with better treatment have decreased the mortality rate of CRC. From the UK, where CRC and gastric cancer are quite common, the results from screening as a preventative measure are very promising. (Cancer Research UK, 2017; Siegel *et al.*, 2020; World Health Organization: Regional Office for Europe, 2020)

The connection between nutrition and prevalence of cancer has been studied widely. Specifically, the correlation between sufficient vitamin D intake and prevention of CRC has been identified in several studies (Garland and Garland, 1980; Feskanich *et al.*, 2004; Wu *et al.*, 2007; Perdue *et al.*, 2014). Also, there has been evidence that high-dose vitamin D supplements benefit patients with advanced or metastatic CRC through improved overall survival (Ng *et al.*, 2019). Additionally, studies have found the patient's obesity to function as a predisposing factor

for CRC. Processed meat, alcohol and fatty foods are shown to increase the risk of getting CRC. (Garland and Garland, 1980; Garland *et al.*, 2006; Torre *et al.*, 2015; Liu *et al.*, 2019; Ng *et al.*, 2019)

The most common type of CRC are adenocarcinomas which constitute about 95 % of the CRC cases (CTCA, 2018). There exists some variation in the survival and characteristics of a cancer patients based on the location of the primary tumour. The prognosis seems to be worse for the right-sided CRC tumours (Janssens *et al.*, 2018). Almost half of the CRC patients will develop metastases. Developing earlier diagnosis and curative treatments remains a challenge for CRC studies. (Markowitz *et al.*, 2002; Calon *et al.*, 2012; Yaeger *et al.*, 2018)

2.3 Data preprocessing

In this section different imputation and feature selection methods are discussed as a part of data preprocessing. The methods are selected based on the conducted literature review. For imputation the nature of missing data patterns is briefly discussed, following by the approaches for filling those values. For feature selection methods Cox score, LASSO approach, Bayesian approach, and bootstrap resampling are presented together with mentions of some other techniques for feature selection.

2.3.1 Imputation

In real-life data applications there are typically observations with missing values. This incompleteness of data is characteristic for medical studies since patients drop-out on follow-ups and die from various causes. This results in right-censored data where the sample size decreases during time because of aforementioned reasons. Since the number of observations having missing values in our CRC data is remarkable, the possibility to impute these values to make data complete are crucial to investigate. Missing data poses multiple issues, e.g. biased estimates and inability to correctly detect associations between covariates (Carroll, Morris and Keogh,

2020), if disregarded in analysis. When imputing missing values for longitudinal data the emphasis lays on the data closer in time since those often are more crucial as predictors (van Buuren, 2014). Typically, bigger datasets are more suitable for fitting predictive models, since there are more data available the model could become more accurate and detailed. More detailed descriptions about the missingness of our CRC data are discussed later in section 3.2.1. First the imputation techniques used in the studies included in the literature revies are mentioned. After that some further details concerning the methods selected to be used in this thesis are discussed.

From the conducted literature review, the papers demonstrated multiple ways to handle missing data (see table in Appendix 1). Removal of observations with missing values (Burke *et al.*, 1997; Anand *et al.*, 1999; Delen, Walker and Kadam, 2005; Barsainya, Sairam and Patil, 2018; Bjarnadottir *et al.*, 2018; Que *et al.*, 2019; Murtojärvi *et al.*, 2020) is one of the basic approaches to handle missing data. This technique of discarding all observations with at least one missing value can is also known as complete-case analysis (Carroll, Morris and Keogh, 2020). Both Reijnen et al. (2020) and (2016) applied multiple imputation. Xu et al. (2020) used multiple interpolation. Here the missing values are determined by the data using the existing values. Van Stiphout et al. (2010) applied a simple imputation, and imputed the missing values using either average or the most common value of the variable in question.

Little and Rubin (2002) propose a framework to classify three main ways which data could be incomplete. These are the following missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). With our CRC dataset the data is MNAR. This is because some of the missingness stems from the merging of the datasets. This rules out multiple imputation (MI) and multiple imputation by chained equations (MICE) (Azur *et al.*, 2011) from the set of possible approaches to handle missing data. If we would have chosen to include only one single datasets, MCAR could be assumed. Besides determining the manner data is incomplete, both the pattern of missing data and how that affects the linkage between variables are useful to investigate (van Buuren, 2014). This is briefly discussed in section 3.2.

For this thesis three techniques for handling missing values are chosen. These are disregarding all observations with any missing values, imputation by median, and kNN-imputation. Listwise deletion approach is later referred as a benchmark. For imputation by median all the missing values are replaced using a median of that specific variable. Median imputation is one of the simple imputation approaches (Carroll, Morris and Keogh, 2020). Additionally, when artificially enlarging our CRC dataset, the multiple imputation by chained equations (MICE) is applied (Vilardell *et al.*, 2020). The usage of MICE here can be justified by the fact that the imputed data with added completely empty rows can be assumed MAR. Thus MICE can be applied. More about this process later in chapter 3.2.1.

In k-nearest neighbour imputation (kNN-imputation) kNN-algorithm is used to impute missing values where k is the number of nearest neighbours based on Euclidean distance to be used. Imputed values are mode for discreate variables, and median for continuous based on the neighbours' values (RDocumentation, 2021). The method selects the k number of neighbours with similar profiles to impute the values with (Troyanskaya *et al.*, 2001), in our case the patients with similar clinicopathological and demographic profiles. So, the missing values are filled with predictions based on similarity among observations.

Multiple imputation by chained equations (hereinafter MICE) is one of the approaches to imputing under conditionally specified models. This technique is a Markov chain Monte Carlo (MCMC) method (van Buuren, 2014). The process of imputing the missing values starts with randomly drawing observations from the data. Then the missing values are imputed using simple imputation, e.g. mean or median (Li and Razzaghi, 2019). After that these missing values are predicted based on the other available values one variable at a time through all the variables in the data. This imputation forms a single iteration. The process is then repeated m times (where m represents the number of cycles) in parallel to obtain accurate estimates for the missing values. (van Buuren and Groothuis-Oudshoorn, 2011) As the process is iterative, the convergence is crucial (van Buuren, 2014).

Although MICE is a widely used technique for handling incomplete data (Li and Razzaghi, 2019), there is a need for theoretical evidence on the appropriateness of MICE (White, Royston and Wood, 2011). Another imputation technique similar to MICE is MCMC sampling-based NORM algorithm (Schafer, 1999) which is more theoretically grounded (White, Royston and Wood, 2011). Further details about multiple imputation (MI) techniques are left outside the scope of this thesis.

2.3.2 Feature selection

Due to enhances in the field of biomedicine in many cases the number of features (e.g. covariates) exceeds the number of observations (here patients), making the data high dimensional. This concerns our data as well. However, as our sample size is insufficient to identify the underlying manifold prefer the models with in-build feature selection instead of usage of separate feature extraction methods (Pölsterl *et al.*, 2016). Thus some preselection of variables ought to be done prior actual analysis. The objective is to identify the most crucial features for predicting survival and utilize those feature in building a multivariate survival analysis model (Witten and Tibshirani, 2010) without compromising the predictive performance (Kuhn and Johnson, 2019). From the pool of all available features select only the ones relevant to the target concept which describe the phenomena most accurately.

In a case with almost equal number of observations and covariates the performance is poor with CPH using the data as is (Witten and Tibshirani, 2010) thus variable selection is needed. For example, when providing all the covariates of our data to a CPH model, there arises a convergence issue. The successful feature selection affects the overall performance of the model positively. However with random survival forests have embedded feature selection (Pölsterl *et al.*, 2016), so this does not concern that. The goal is to try to find the best performing subset of features whilst maintaining the crucial features of the data and obtaining the best possible the precision of the model. For clarity, this process is simplified in this thesis. For all the separate survival analysis models, same features are used except for the random survival forests.

Dimensionality reduction is required prior building a prognostic model with number of coefficients surpassing number of observations, i.e. when p > n. Another possible issue arises with overfitting when the model fits the noise of the data along with the survival related signal. Simplicity of the model is a preferred characteristic for better interpretability and lower costs (Witten and Tibshirani, 2010). Reducing the dimensionality of the data helps to eliminate irrelevant features and reduce noise, thus reducing time and memory required for data analysis.

Literature suggests multiple techniques for identification of significant features for predicting survival of cancer patients. In this following chapter a few of those are discussed briefly. In some studies, a false discovery rate (FDR) approach is applied to narrow down the list of potential candidates with a small number of false positives which is preferred (Witten and Tibshirani, 2010). Thus finding out whether the covariates are truly associated with the survival or not. Because of the pivotal role of this, FDR-corrected p-values are calculated with univariate Cox. More detailed comparative studies about the differences in the FDR of feature selection techniques could be carried out in the future.

For this thesis correlation analysis, univariate CPH, and random survival forests are used for feature selection. The details about implementing these techniques to our CRC data are discussed later in section 3.2.3. Further details and applications of other methods are left outside the scope of this study. Some studies (Li and Razzaghi, 2019; Reijnen *et al.*, 2020; Wang, Wang and Makond, 2020) perform feature selection based on literature review. However, this approach limits the feature space to obey previous findings and thus pre-empt the possibility of novel discoveries. Our data has more features (p) than observations (n). This case when p>n is referred as a curse of dimensionality, which can be resolved using feature selection techniques. These techniques aim to reduce the number of predictors without sacrificing the predictive performance (Kuhn and Johnson, 2019)

2.3.2.1 Cox score

A standard approach for identification of significant features is using a univariate Cox score, or Cox score test. As a tool for feature selection Cox score measures the correlation between a feature's value and patient survival, and only the features having a Cox score value surpassing a certain threshold are included into further analysis (Bair and Tibshirani, 2004). First, fit a univariate Cox for each covariate individually obtaining a score statistic, a Cox score which quantifies how well does the covariate predict survival (Beer *et al.*, 2002). High Cox scores indicate that the specific covariate is associated with the survival outcome and thus should be included. Beer et al. (2002) applied Cox scores to obtain risk indices for each patient by taking an appropriate linear combination of a subset of significant genes. Those risk indices were then used to divide the data into low- and high-risk groups. Though they decided to exclude censoring status from the analysis.

Wald scores can be used in a similar manner than Cox scores, however those perform poorly in a high-dimensional data setting (Witten and Tibshirani, 2010). Significance analysis of microarrays (SAM), a technique proposed by Tusher, Tibshirani and Chu (2001), incorporates a modified Cox score. Through an addition of a small constant the method is more stable and thus often outperforms traditional Cox score.

In its simplicity of utilizing a Cox score, this approach has its downside since it does not take into consideration possible relations between the covariates. It sometimes chooses spurious features (Bair and Tibshirani, 2004) which means that the features selected by Cox score will not be the best possible predictors for the survival. This is important factor to acknowledge since these are not independent in a way in reality but in fact affect each other in various ways. A single covariate independently can have a totally different affect to the patient's survival than when all the covariates affect simultaneously.

Bair and Tibshirani (2004) introduced an another approach to avert this aforementioned issued with selecting potentially spurious covariates. This technique is referred as a PLS-corrected

(partial least squares) Cox score which selects the features based on both survival time and underlying genetic profile of the patient. Besides the issues with using only the risk score to form predictions, they raised the question concerning the real-life situations where the patient data pool might not be sufficient to obtain the risk for an individual patient. The predictions depend so strongly on the limited pool of the other patients, and this cannot be seen as a very desired property. This leads to issues with small sample size and outliers. However, these enhancements of the basic approach of selecting the features based on the output of the univariate CPH are left outside of the scope of this thesis. Here the key is to conceive the potential issues with the chosen techniques and mention some of the solutions developed to tackle those.

2.3.2.2 LASSO approach

Tibshirani (1996) proposed a novel approach to variable selection for linear regression models. The LASSO approach (Least Absolute Shrinkage and Selection Operator) minimizes the residual sum of squares w.r.t. the sum of the absolute value of the coefficients which are less than a user-specified constant. Later, Tibshirani (1997) proposed a revisited version of lasso suited especially for CPH models. This new approach minimizes the log partial likelihood with respect to sum of the absolute values of the parameters which are bounded by a positive constant value, s (see formula ***(1)*** (Tibshirani, 1997)). This approach enables the reduction of estimation variance. For the penalization scheme to function properly the initialization by standardization of parameters is required. Lasso was tested using lung and liver cancer datasets using traditional CPH with stepwise variable selection as a benchmark model.

$$\widehat{\boldsymbol{\beta}} = argmin\, \ell(\boldsymbol{\beta}) \qquad w.r.t\ \sum|\beta_j| \leq s \tag{1}$$

Witten and Tibshirani (2008) proposed an improved version of this technique called the lassoed principal components (LPC) method. They demonstrated on a simulated dataset and a kidney cancer dataset that their LPC method outperforms a standard Cox score-based analysis. Murtojärvi et al. (2020) applied LASSO with a greedy cost-specified variable selection algo-

rithm for a prognostic model for mCRPC (metastatic castration-resistant prostate cancer) patients. Also, decision trees (Jerez-Aragonés *et al.*, 2003; Mao *et al.*, 2005; Oztekin, Delen and Kong, 2009) and regression trees (Loh, 2002) have been used for variable selection. Mitchell and Beauchamp (Mitchell and Beauchamp, 1988) applied Bayesian variable selection in linear regression for energy-conservation study. Since then the technique has also been used in oncology studies (Duan *et al.*, 2018; Zhang *et al.*, 2018; Nikooienejad, Wang and Johnson, 2020).

2.3.2.3 Bayesian approach for feature selection

Bayesian approach with a hierarchical Bayes model presented by Michell and Beauchamp (1988) proposes a solution for variable selection. In their model the probability distribution is first assigned to the dependent variable using regression model's parameters' prior distributions. The actual selection of variables from the input data is affected by a spike and slab distribution. Then the Bayesian model is multiplied by the density function of corresponding parameters and the result is integrated, the variables in the final sub-model are obtained. The mean of the loss distribution together with the goodness-of-fit plot are suggested to be used for the cross validation of this approach.

George and McCulloch (1993) presented a stochastic search variable selection (SSVS) technique for variable selection in multiple regression models using Bayesian approach with the Gibbs sampler. Gibbs sampler is a MCMC algorithm which can be applied to generating random variables indirectly from a distribution without knowledge about the density (Casella and George, 1992). Their solution utilizes the Gibbs sampler for the calculation of posterior probabilities used to identification of most relevant variables for the model.

2.3.2.4 Bootstrap resampling

A bootstrap resampling approach was introduced as an application to subset selection in the CPH by Sauerbrei and Schumacher (1992). Their solution tackles the problems with variable selection in the model building phase and presents an alternative to commonly used stepwise selection methods. For stepwise selection, there exists three basic procedures: forward selection, stepwise selection, and backward elimination. This bootstrap resampling approach is based on a paper by Efron (1979) which discussed the Quenouille-Tukey jack knife method in the context of bootstrap methods. The approach takes bootstrap replications, M, consisting of survival times, censoring indicators, and covariates drawn from the original data, and treats those as independent samples. Then the variables having enough prognostic value are identified using a selected stepwise procedure. This approach allows the research structure to be altered in a way that the level of strength of factors' prognostic value can be chosen. Sauerbrei and Schumacher tested their approach using two different cancer datasets.

2.3.2.5 Other techniques for feature selection

Some of the other possible approaches for feature selection proposed in literature are genetic algorithms (Liu *et al.*, 2013; Mansoori, Suman and Mishra, 2014; Aalaei *et al.*, 2016), Gibbs sampling (Casella and George, 1992; George and McCulloch, 1993; Herring, Ibrahim and Lipsitz, 2004), control of induction by sample division (CIDIM) (Ramos-Jiménez, G. Morales-Bueno, R. Villalba-Soria, 2000), neural networks (Jerez-Aragonés *et al.*, 2003; Oztekin, Delen and Kong, 2009; Ching, Zhu and Garmire, 2018), bootstrap resampling (Sauerbrei and Schumacher, 1992), gradient descent (Burke *et al.*, 1997), information gain measures (Anand *et al.*, 1999; Barsainya, Sairam and Patil, 2018), Cox-nnet (Ching, Zhu and Garmire, 2018), fuzzy similarity and entropy (FSAE) (Lohrmann *et al.*, 2018), clustering one less dimension (COLD) (Lohrmann and Luukka, 2019), fuzzy entropy measures with similarity classifier (Luukka, 2011) and backward sequential feature elimination process (Ryu, Chandrasekaran and Jacob, 2004). Further discussion about these feature selection techniques is left outside the scope of this thesis.

2.4 Survival analysis methods

An overview of the statistical and machine learning methods and applications in the field of medical research is provided based on the conducted literature review constitutes this following section. Cox proportional hazards model (CPH), random survival forests (RSF), and artificial neural networks (ANN). In addition discussion about support vector machines (SVM) for survival analysis is presented in Appendix 2. CPH is the most common survival analysis technique in the field of health sciences (Yavari *et al.*, 2014). ML applications enables quicker and more accurate analysis of biomedical data without being prone to human errors (Abdar and Makarenkov, 2019). Deep learning tries to model the high-level abstract structures of the data using a set of algorithms of non-linear transformations (Bengio, Courville and Vincent, 2013).

Ren et a. (2019) defined survival analysis as a process of analysing and modelling time-to-event data whilst including imputation techniques. Survival analysis is a commonly used technique in making prognosis for patients with different medical conditions, e.g. for cancer patients. Here the goal is to try to detect any maladies at their early stages, and possible prevent incurable deceases through help to identify patients at risk. Thus conducting survival analysis can help to identify biomarkers which tell information about the illness and patients' prognosis. Better understanding about these conditions could allow more efficient use of resources as the early diagnosis become more common.

Censoring is one of the most important aspects differentiating survival analysis from traditional clustering (Štajduhar, Dalbelo-Bašić and Bogunović, 2009). Survival data is typically partly right-censored (Oakes, 2000). In that case for some of the observations there is no knowledge about the survival status at end of the follow-up period (Katzman *et al.*, 2018). This can be interpreted as missing data, or the survival time is equal to the censored time. Time to event of interest, e.g. death is only known for those observations for which that event occurred during an observation period, others are considered as censored instances. The censoring in our case is unplanned which means that it is caused by change (Nelson, 1972). The focus for this thesis is on right-censoring.

The Figure 5 demonstrates the concept of right-censoring. Those observations marked using 'X' have experienced an event of interest, e.g. death. For others marked with a circle there is no observed event, and for some there are no recent information concerning their status. If the situation is investigated at time 8 (represented with a vertical red line in the figure), observation S3 experiences right-censoring. Since there is no available information about the status of S3 after time 4, it is considered as censored due to fail to follow-up or withdrawal. Observations S1, S2 and S5 are given status alive at time 8.

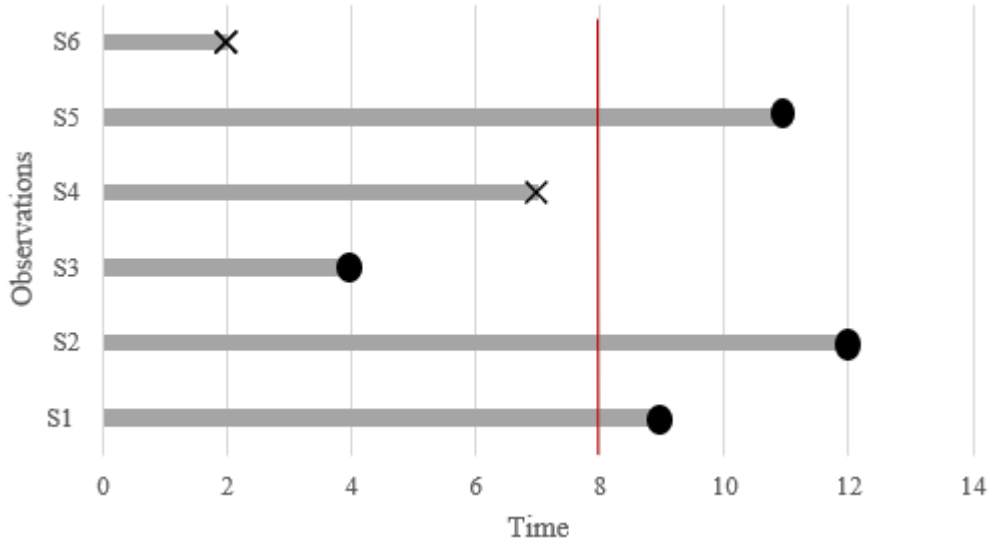


Figure 5. Demonstration of right-censoring (adapted from (Wang, Li and Reddy, 2017)).

The challenges, like non-linearities, heterogeneity of effects, and large number of predictor variables, posed to traditional regression-based survival analysis techniques could possibly be overcome using ML techniques which 'learn from the data'. These ML models attempt to find a 'balance' that minimizes a loss function by minimizing both bias and variance. This procedure is referred as tuning model parameters. (Goldstein, Navar and Carter, 2017) Survival data can be characterized as high-dimensional, low sample size (HDLSS) which might cause issues if not properly considered while conducting the analysis and selecting the appropriate techniques for that (Hao *et al.*, 2019).

This section focuses on the different steps included into the survival analysis. Both statistical and machine learning applications to survival analysis are discussed. The Figure 6 roughly presents the structure of this section. In addition, the necessary preprocessing and validation techniques are briefly presented. The arrow at the bottom of the figure demonstrates the chronological order of the phases in the process. This section begins with Cox proportional hazards analysis which is the common approach for oncologic survival analysis (Matsuo *et al.*, 2019).

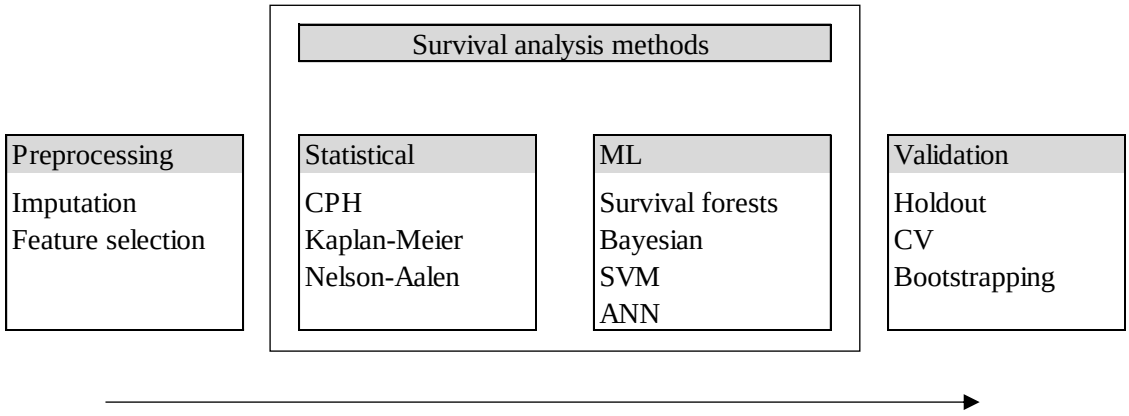


Figure 6. Example taxonomy of survival analysis methods and the overall process (adapted from (Wang, Li and Reddy, 2017)).

2.4.1 Survival analysis algorithms

Survival function ***(2)*** (Kaplan and Meier, 1958) calculates the probability of surviving past time t, e.g. the event time T is greater than the observed time t (Kaplan and Meier, 1958). Typically, this event is death. In other words, the survival function quantifies the probability that the event has not occurred at time t. Survival function possesses three (3) properties. First, the survival function only gets values on the interval $[0,1]$. Second, the cumulative distribution function of T is $1 - S(t)$. This implies the third property that the survival function of a non-increasing function of time, t. (Wang, Li and Reddy, 2017; Lifelines, 2021) Thus the probability of survival decreases over time.

$$S(t) = \Pr(T > t) \quad (2)$$

Another pivotal function is survival analysis is hazard function *(3)* (Wang, Li and Reddy, 2017). Hazard function reflects the probability of an event occurring at time t given that the event has not occurred until time t. Thus hazard function is a conditional probability. Function calculates the probability that a patient will experience an event, e.g. will not survive, for an additional extremely small amount of time δ. Higher value of hazard function represents higher risk of an event occurring. (Lee and Wang, 2003)

$$h(t) = \lim_{\delta t \to 0} \frac{\Pr(t \leq T \leq t + \delta t | T > t)}{\delta} = \frac{-S'(t)}{S(t)} \quad (3)$$

2.4.2 Kaplan-Meier estimator

An Kaplan-Meier based application on grouped interval survival data is life table analysis (Cutler and Ederer, 1958) which is utilised in actuarial context especially in life insurance (Wesley, 1998). The origins of life tables are in the mid-1600 when the method was first proposed by Graunt and Halley. Their solution applied a concept that the survivorship tables are computed using a summation of deaths per each age-group. Later de Moivre continued from the work of Halley and formulated a method for calculating annuity values which have been utilised in a commercial applications. (Greenwood, 1938)

Kaplan and Meier (Kaplan and Meier, 1958) present their nonparametric product-limit (hereinafter PL) estimate to estimate the distribution of missing values for observations. This technique is based on previous actuarial methods and the observed events are subdivided to losses, λ, and deaths, δ. The probability of surviving past time t is represented by the distribution function, $P(t)$. If the lifetime of individuals is assumed to be finite, the $P(\infty) = 0$. The product-limit estimate of $P(t)$, $\hat{P}(t)$ is right-continuous unlike the reduced-sample estimate $P^*(t)$.

2.4.2.1 Calculation of product-limit estimate

Distinguishing from the life table analysis Kaplan-Meier method assess the intervals of interest by the times of deaths instead of fixed time intervals, and there is no assumption of constant mortality during the observed intervals (Wesley, 1998). The process of calculating the product-limit (PL) starts by dividing the time into selected intervals each of which is assigned estimate, p_j, for the share of the survive individuals beyond the end of that interval. The selection of these intervals should be conducted in a manner that the

Hence, the $P(t)$ is estimated using all the previously mentioned estimates, p_j, for all the preceding intervals of t. When the PL considers both deaths and losses in the interval between u_{j-1} and u_j the formula is given by ***(4)*** (Kaplan and Meier, 1958). Thus the survival probability at time t is equal to the product of the percentage chance of surviving at time t and each time prior that.

$$\hat{P}(t) = \prod_{j=1}^{k} \left(\frac{n_j'}{n_j}\right), \quad given\ that\ u_k = t,\ \ n_j' = n_j - \delta_j \qquad (4)$$

However, if a single death event is considered enough to occupy a single interval, now the formulation of PL estimate is the following step function ***(55555)*** (Kaplan and Meier, 1958) where the deaths are arranged in the order of age. N is the number of random samples used for calculations, and r is the positive integers for which time of death is smaller than equal to the time t. Basically, this is product of all the probabilities for patient who fulfil the aforementioned condition of time of death.

$$\hat{P}(t) = \prod_r [(N-r)/(N-r+1)] \qquad (5)$$

If the observed time period contains no losses, the PL is reduced to a binomial estimate ***(6)*** (Kaplan and Meier, 1958) where $n(t)$ is the number of patients observed and survived at time

t when the deaths at that time t are subtracted. The mean life estimate can be derived from the PL estimate by taking an integral over it.

$$\hat{P}(t) = n(t)/N \qquad (6)$$

2.4.2.2 Kaplan-Meier curves

KM survival curves are widely used in medical research (Wesley, 1998). Those display the Kaplan-Meier estimate over observed time in a staircase pattern where each death is demonstrated as a drop. Figure 7 demonstrates an example Kaplan-Meier curve. This specific curve is plotted using our CRC data with listwise deletion of rows with any missing data prior enhancing the sample size. The overall survival probability decreases as the time increases. The dotted lines around the step function represent the associated confidence intervals. The line does not decrease dramatically since there are only 29 CRC related deaths in this specific data, and 48 patients are assigned status 'other' which means that the patient is either censored or cause of their death is not CRC.

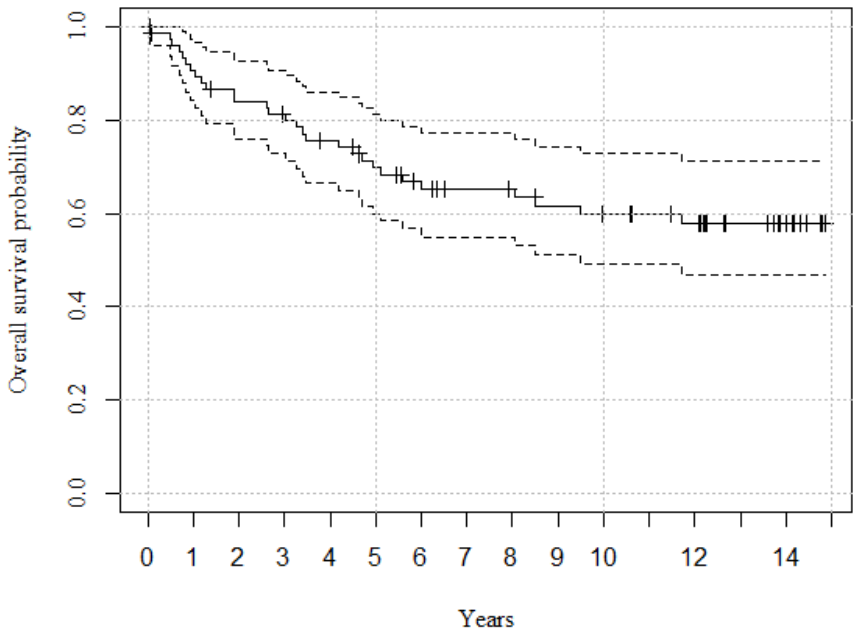


Figure 7. Example of a Kaplan-Meier curve.

2.4.2.3 Nelson-Aalen estimator

Nelson-Aalen estimator uses counting process approach to estimate the cumulative hazard function for censored data (Wang, Li and Reddy, 2017). A counting process counts the events of a point process which is a countable random collection of points (Aalen, 1978). With Nelson-Aalen estimator there are no assumptions concerning the distribution of the underlying data. Thus it is often applied to check the fit of parametric models. Nelson (Nelson, 1972) first referred this method as hazard plotting. However, this estimator is not popular since its visual result fails to be as intuitive and simple to interpret as other methods (Lewinson, 2020). Formula *(7)* (Borgan, 2005) shows the calculation of Nelson-Aalen estimator where the number of events (d_j) at time t is divided by the number of observations (e.g. patients) at risk (r_j). (Borgan, 2005)

$$\hat{A}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j} \qquad (7)$$

Nelson-Aalen estimator can be applied to estimate more complicated processes as well (Andersen *et al.*, 2012; Njamen-Njomen and Ngatchou-Wandji, 2014). An example of these situations is the multistage model where other deaths can be included as competing events and thus reducing bias in the survival predictions. Account for competing risks jointly in the analysis for more accurate predictions and thus possibly better preventative care (Lee, Yoon and Van Der Schaar, 2020). The obtained prediction models become more dynamic with competing risks incorporated. This becomes more crucial the more other deaths or events the data has, other than the event of interest e.g., here cause-specific deaths due to CRC and other events are competing risks. Cox models, which are discussed in chapter 2.4.3, can also be used to model these competing risks situations (Lunn and McNeil, 1995; Beyersmann and Scheike, 2016). Also, random survival forests (see chapter 2.4.4) can be applied to model competing risks situations (Ishwaran *et al.*, 2014). Further details about this nonparametric estimator are left outside the scope of this thesis.

2.4.3 Cox proportional hazards model

Cox's proportional hazard regression model (hereinafter CPH) is one of the most used techniques in the survival analysis research. CPH is a common choice for a benchmark model for testing the prognostic performance of different survival analysis techniques. Since Cox (1972) proposed the original version of the model in 1972, several alterations have been presented in the literature. These other models, including DeepSurv, DeepHit and many other, are discussed later in this section. In addition, some alternative techniques for variables selection for CPH are briefly discussed.

Cox (1972) continues from the previously discussed (in section 2.4.2) results of Kaplan and Meier (1958) specifically the ones relating to incorporation of regression-like arguments into life-table analysis. The Cox proportional hazards model is presented. The model examines the log-linear relationship between the independent variables and the hazard function which determines the effects of observed covariates on the risk of an event occurring. A concept of adding a stress term to the model is proposed. This allows the model to be carried out on different stress levels which allows the investigation of stress to the failure-time's distribution. A briefly presented case with bivariate life tables is left out of the scope for this thesis as well as further discussion of introducing the stress variable to the model. Another paper by Cox (1975) continues generalizing the ideas relating to the partial likelihood.

For each individual in the observed population, either death, loss or censoring is assumed. This can be formulated as (x_i, σ_i, z_i) which are observed for each individual i as possibly censored failure time, indicator for failure or censoring, and covariates. If an individual's failure time is greater than the time t, the individual survives. This probability can be modelled using a survivor function. Hazard function at time t is composed of two nonnegative functions given baseline hazard function as $\lambda_0(t)$, β as a vector of regression coefficients, and z vector of covariates, is shown in the formula *(8)* (Cox, 1972) where $\exp(\beta)$ is the hazard ratio. This ratio of two groups remains proportional over time, hence the model is proportional. (Cox, 1972, 1975)

$$\lambda(t|z) = \lambda_0(t)\exp(z\beta) \quad (8)$$

Thus there is a proportional hazards assumption associated with the CPH models. This PH assumption means that the hazard ratio of all of the covariates' is assumed constant and thus not vary with time (In and Lee, 2019). So, these hazards representing the probability of an event occurring at a certain time point t, which in our case is the probability of survival, need to remain proportional over the observation period. Neglecting to check the possible violations of this proportional hazards assumption of the CPH model undermines the final results of the study (Kuitunen *et al.*, 2021) hence the nonproportionality is assessed after the initial fit of the CPH model. Basically, this violation indicates that one or more of the model's covariates changes over time which then leads to either over or underestimated hazard ratios (Schemper, Wakounig and Heinze, 2009).

Potential solutions to tackle these violations to proportional hazards assumption are stratification (Harrell, Lee and Mark, 1996), usage of an estimator for the aggregate factor effect based on cumulative hazards estimated under a stratified CPH (Wei and Schaubel, 2008), usage of time-dependent coefficients (Harrell, Lee and Mark, 1996; Quantin *et al.*, 1999), Schemper's weighted model (Schemper, Wakounig and Heinze, 2009), fit a piecewise PH model which results in a step function of HR (Moreau, O'Quigley and Mesbah, 1985; Quantin *et al.*, 1999), switch to a different type of a model (i.e. semi-parametric proportional odds model (Bennett, 1983a; Moreau, O'Quigley and Mesbah, 1985), separate modelling for different time periods (Schemper, Wakounig and Heinze, 2009), or parametric log-logistic model (Bennett, 1983b)), and restricted mean survival time (Kuitunen *et al.*, 2021). Further investigation of dealing with nonproportionality in Cox models using the aforementioned approaches is left outside the scope of this thesis.

However, the hazard function could take other forms as well whilst the properties for that are required to be circumspect. In discrete time, the formulation of the hazard function is slightly altered as the baseline hazard function is replaced by $\frac{\lambda_0(t)}{1-\lambda_0(t)}$. For simplicity, the hazard rate, $\lambda_0(t)$, is assumed to be constant. Additionally, the β as log hazard ratio can be derived without

further insight about the hazard rate. From this follows the assumption of the exponential nature of the underlying distribution which makes the model semiparametric. This way the maximum likelihood approach can be applied to handle the plausible censoring in the data. These maximum likelihood estimates of β can be obtained by iterating the functions ***(9)*** (Cox, 1972) & ***(10)*** (Cox, 1972) derived from the function ***(12)*** (Cox, 1972). Further calculations about these formulas in a discrete time case are left out from this thesis. (Cox, 1972, 1975)

$$U_\xi(\beta) = \frac{\sigma L(\beta)}{\sigma \beta_\xi} \qquad (9)$$

$$\mathcal{J}_{\xi\eta}(\beta) = -\frac{\sigma^2 L(\beta)}{\sigma \beta_\xi \sigma \beta_\eta} \qquad (10)$$

Likelihood for inference about parameters β previously referenced (Cox, 1972) as a conditional likelihood can be calculated as shown where R_j is the risk set at time $t_j - 0$. Actually, this (formula ***(11)*** (Cox, 1975)) could be interpreted as partial likelihood. This formulation assumes the continuous time and failures to be occurring at distinct times. The need for the usage of partial likelihood stems from the complexity of calculation of the full likelihood. (Cox, 1972, 1975)

$$\prod_j \frac{\exp(\beta^T z_j)}{\left\{\sum_{k \in R_j} \exp(\beta^T z_k)\right\}} \qquad (11)$$

Now the log partial likelihood is given as formula ***(12)*** (Cox, 1972) where $\mathcal{R}(t_{(i)})$ is the risk set at a particular time $t_{(i)}$ when the failures occur, and k equals the number of observations.

$$L(\beta) = \sum_{i=1}^{k} z_i \beta - \sum_{i=1}^{k} log\left[\sum_{l \in \mathcal{R}(t_{(i)})} exp\{z_{(l)}\beta\}\right] \qquad (12)$$

Finally, after the derivation of maximum likelihood estimates of β, the estimation of failure time's distribution can be conducted. This is done by generalization of the Kaplan-Meier maximum-likelihood estimate by taking baseline hazard as zero for all the time points except the ones where failure as occurred. For those time points having failure, separate maximum likelihood estimation is then conducted. This then results in the product integral formula. (Cox, 1972, 1975)

2.4.3.1 Artificial neural networks for CPH

The idea of combining Cox regression with neural networks was first proposed by Faraggi and Simon (1995). In their solution the output of a simple feed-forward network (a single logistic hidden layer and a linear output layer) replaces the linear function typically used in CPH. It is a nonlinear extension of classical CPH. However, research has failed to demonstrate improvements for this nonlinear extension beyond classical CPH (Sargent, 2001).

DeepHit

Lee et al. (2018) presented a novel, discrete approach for survival analysis called DeepHit. Differing from the previous survival models DeepHit does not make any assumptions about the underlying stochastic processes and applies a deep NN for learning the survival times' distribution directly. This approach allows the modelled stochastic processes to depended on the covariates. DeepHit is also suitable for modelling the events in the case of multiple competing risks, e.g. multiple different diseases. The time-to-event of interest is called first hitting time hence the name DeepHit.

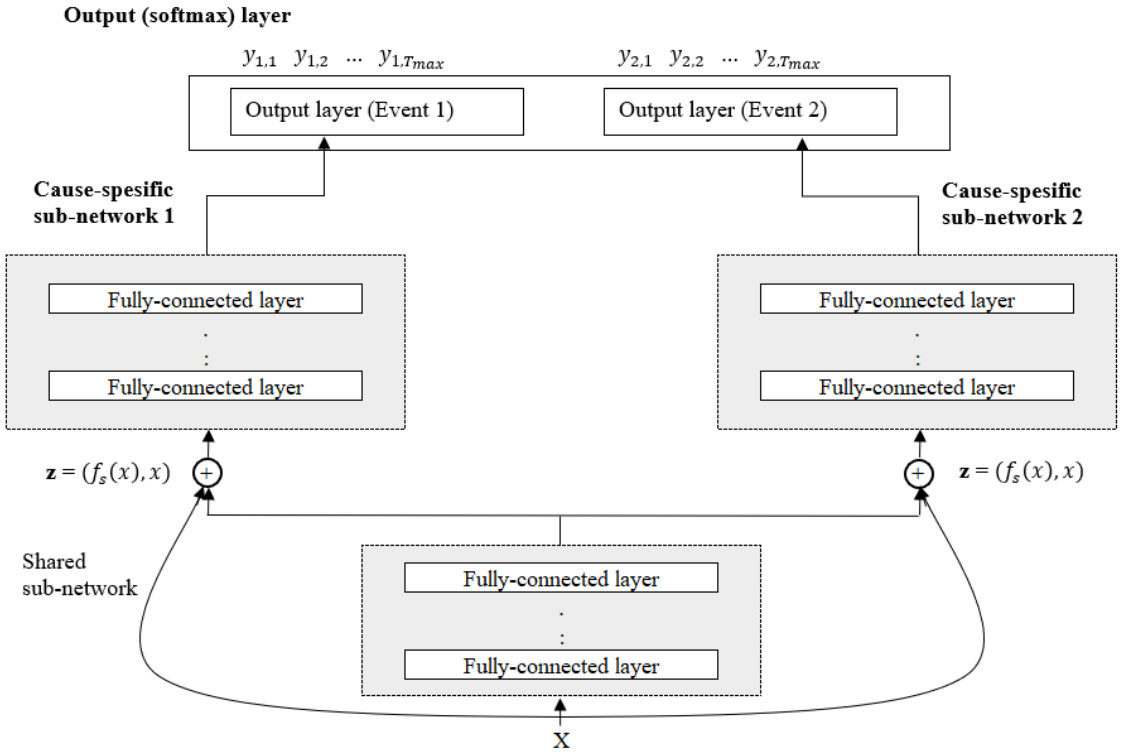


Figure 8. DeepHit structure (adapted from (Lee et al., 2018)).

Structure of the DeepHit is formed by a single shared network and multiple sub-networks (Figure 8). The softmax activation function is applied as the network's output layer. As the model is discreate the continuous time data needs to be discretised using an equidistant grid stretched between the first and last timestamp. This number of discrete time-points is one of the hyperparameters to the actual model. The output is constructed as the estimated probability mass function of the duration time which gives the estimated survival function as follows ***(13)*** (Kvamme, Borgan and Scheel, 2019) . There $y_k(x)$ represents the output of a NN with covariates x at discrete time k. (Lee *et al.*, 2018; Kvamme, Borgan and Scheel, 2019)

$$\hat{S}(\tau_j|\boldsymbol{x}) = 1 - \sum_{k=1}^{j} y_k(x) \qquad \textbf{(13)}$$

The DeepHit's loss function is a combination of two separate loss functions, $loss = loss_1 + loss_2$. The calculation of the total loss considers the censored observations typical for time-to-event data. First, the log-likelihood of the joint distribution of the first hitting time and the

corresponding event is defined as follows shown in formula ***(14)*** (Lee et al., 2018) where $\mathbb{1}(\cdot)$ is the indicator function and K the number of competing risks. The first part of $loss_1$ captures the non-censored observations and the second part the censored ones.

$$loss_1 = -\sum_{i=1}^{N}[\,\mathbb{1}(k^{(i)} \neq \emptyset) \cdot \log\left(y_{k^{(i)},s^{(i)}}^{(i)}\right) + \mathbb{1}(k^{(i)} \neq \emptyset) \cdot \log(1 - \sum_{k=1}^{K} \hat{F}_k(s^{(i)}|\boldsymbol{x}^{(i)}))] \qquad \textbf{(14)}$$

The second loss is added to enhance function's the ranking abilities. In the formula ***(15)*** (Lee et al., 2018) $A_{k,i,j}$ represents the comparison of the risk at time s between the individual who dies and the individual who survives longer than time s. $\eta(x, y)$ is a convex loss function, e.g. $exp\frac{-(x-y)}{\sigma}$. α_k are the coefficients relating to the different competing risks.

$$loss_2 = \sum_{k=1}^{K} \alpha_k \cdot \sum_{i \neq j} A_{k,i,j} \cdot \eta(\hat{F}_k(s^{(i)}|\boldsymbol{x}^{(i)}), \hat{F}_k(s^{(i)}|\boldsymbol{x}^{(j)})) \qquad \textbf{(15)}$$

The prognostic performance of DeepHit was tested against multiple cause-specific benchmark models on three real-life datasets and a synthetic-one. The 5-fold cross validation was performed. Measured using a time-dependent concordance index (Antolini, Boracchi and Biganzoli, 2005) the DeepHit was able to slightly outperform cause-specific CPH, Fine-Gray proportional sub-distribution hazards model, deep multi-task Gaussian process, RSF, threshold regression, logistic regression, and DeepSurv (Lee *et al.*, 2018). In 2020 a revisited version of the DeepHit called Dynamic-DeepHit was proposed by Lee, Yoon and Van Der Schaar (2020). Further details of this specific technique are left outside the scope of this thesis.

Later Kvamme, Borgan and Scheel (2019) presented a new loss function (function ***(16)*** (Kvamme, Borgan and Scheel, 2019)) calculated in batches to tackle potential problems associated with scaling with large datasets and the proportionality assumption. Their solution considers the censored observations, which are typical with survival data, as using a possibly right-censored event time instead of true event time. For minimization of Cox partial likelihood, they proposed stochastic gradient descent (SGD) rather than the traditional approach of applying

Newton-Raphson's method. They also suggested amendment of penalty parameter to the loss function. They used this novel loss function as part of extended Cox model with neural networks.

$$loss = \frac{1}{n}\sum_{i:D_i=1} log\big(1 + \exp\big[g\big(x_j\big) - g(x_i)\big]\big), j \in R_i \setminus \{i\} \qquad (16)$$

DeepSurv

Katzman et al. (2018) proposed a modern Cox proportional hazards fully connected, feed-forward deep neural network which is also known as DeepSurv. Purpose of this was to develop a more cost-effective solution than CPH to the problem of being able to personalize treatment by estimating individual's risk of failure, which typically is interpreted as death. DeepSurv exploits deep neural network with gradient descent to optimize the weights, θ, of the network. These weights are used to parametrize the hazard rate of the individual's covariates.

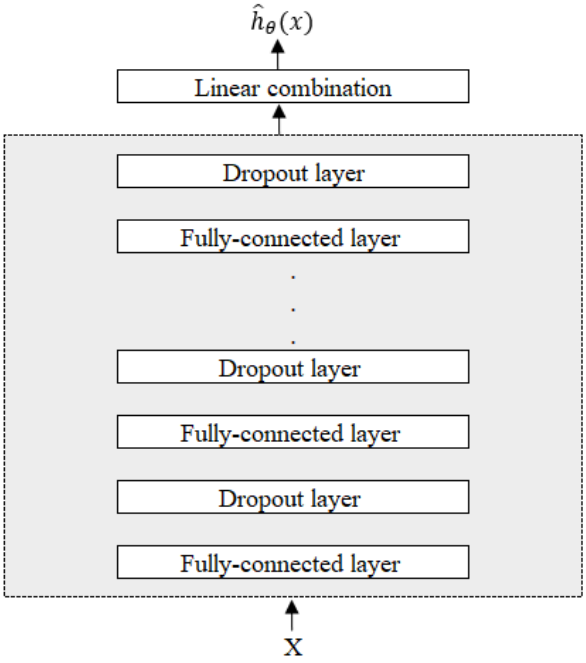


Figure 9. DeepSurv structure (adapted from (Katzman et al., 2018)).

Figure 9 describes the structure of DeepSurv. X represents the baseline data used as an input to the network. Output of the DeepSurv model, $\hat{h}_\theta(x)$, is used to estimate the Cox's log-risk function. The loss function is the average negative log partial likelihood presented in the function ***(17)*** (Katzman et al., 2018), where $N_{E=1}$ is the number of individuals with an observable event, and λ is the ℓ_2 regularization parameter. This functions also as the objective function used for the training of the network. Random search for hyper-parameter optimization (Bergstra and Bengio, 2012) is performed to identify the number of hidden layers, number of nodes in each layer, and the dropout probability. (Katzman *et al.*, 2018)

$$loss := l(\theta) = -\frac{1}{N_{E=1}}\sum_{i:E_1=1}(\hat{h}_\theta(x_i) - log\sum_{j\in R(T_i)} e^{\hat{h}_\theta(x_j)}) + \lambda \cdot \|\theta\|_2^2 \qquad \textbf{(17)}$$

For the actual treatment recommendation system Katzman et al. (2018) divided the individuals (here patients) into separate individual groups, τ, based on the treatment, i, they are given. This way the hazard function is as follows in formula ***(18)*** (Katzman et al., 2018).

$$\lambda(t; x|\tau = i) = \lambda_0(t) \cdot e^{h_i(x)} \qquad \textbf{(18)}$$

Now, the output, $\hat{h}_\theta(x)$, from the NN can be interpreted as the log-risk, $h_i(x)$, of being assigned to a specific treatment group for each individual from the group of patients. Additionally, all the patients are assumed to have the same baseline hazard function, $\lambda_0(t)$. By taking a logarithm of the hazard ratio, the recommender function ***(19)*** (Katzman et al., 2018) is obtained. This recommender demonstrates the individual's risk-ratio of being assigned to a certain treatment group, e.g. provides personalized treatment recommendations for decision support. Each patient is assigned twice, first to the treatment group i and then to the group j. From the difference of the results obtained from the NN, the treatment recommendation is achieved. (Katzman *et al.*, 2018)

$$rec_{ij}(x) = \log\left(\frac{\lambda(t; x|\tau = i)}{\lambda(t; x|\tau = j)}\right) = \log\left(\frac{\lambda_0(t)\cdot e^{h_i(x)}}{\lambda_0(t)\cdot e^{h_j(x)}}\right) = h_i(x) - h_j(x) \qquad \textbf{(19)}$$

On both simulated and real patient datasets, DeepSurv found to outperform CPH and RSF models measured in the prediction accuracy measured by c-index (Katzman *et al.*, 2018).

RankDeepSurv

Jing et al. (Jing *et al.*, 2019) presented a paper in which they proposed a new survival analysis technique, RankDeepSurv, which is an extension of DeepSurv with rank and regression constraints. They tested this method for forming prognosis for nasopharyngeal carcinoma (nasopharynx cancer) patients. The structure of RankDeepSurv is shown in the Figure 10. For the activation function in the fully connected layers of the network the Exponential Liner Units (ELUs) are used. The appropriate number of layers and nodes is determined using a random hyper-parameter search. Unlike with the DeepSurv, only one dropout layer is used to prevent overfitting of the network.

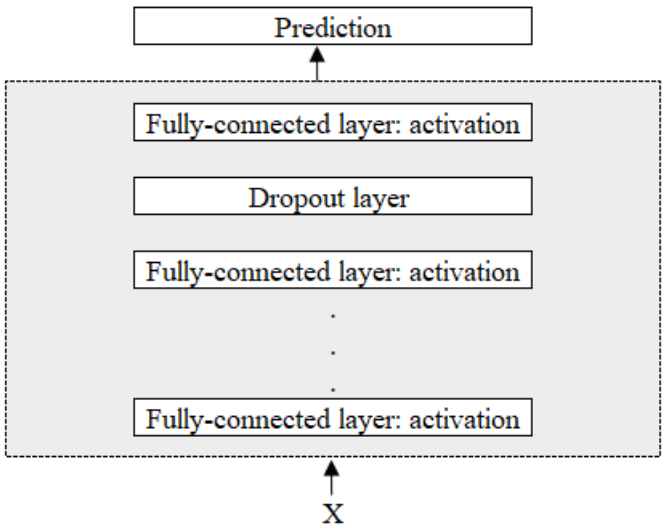


Figure 10. Structure of RankDeepSurv (adapted from (Jing et al., 2019)).

Like Kvamme, Borgan and Scheel (2019) they (Jing *et al.*, 2019) also presented a new loss function for optimization of a deep feed-forward NN. Their solution is based on the sum of an

extended mean squared error loss (L_1) and a pairwise ranking loss based survival data's ranking information (L_2). In the formula ***(22)*** (Jing et al., 2019) of the total loss function the outputs from L_1 (see formula ***(20)*** (Jing et al., 2019)) and L_2 (see formula ***(21)*** (Jing et al., 2019)) are multiplied by positive constants α and β. To that the weights, θ, determined by the network regularized with parameter λ are added resulting in the total loss. Similar to the DeepSurv, the gradient descent optimization approach is applied to find the weights, θ, for the parameters. This extended loss function converges as a convex function. (Jing *et al.*, 2019)

$$L_1 = \frac{1}{n}\sum_{j=1,I(j)=1}^{n}(yp_j - y_j)^2,\ \ I(j) = \begin{cases} 1, & if \quad \delta_j = 1 \\ & \delta_j = 0 \wedge yp_j \leq y_j \\ 0, & else \end{cases} \qquad (20)$$

$$L_2 = \frac{1}{n}\sum_{I(i,j)=1}^{n}\left[\left(y_j - y_i\right) - \left(yp_j - yp_i\right)\right]^2, I(i,j) = \begin{cases} 1, & y_j - y_i > yp_j - yp_i \\ 0, & else \end{cases} \qquad (21)$$

$$L_{total}(\theta) = \alpha * L_1(\theta) + \beta * L_1(\theta) + \lambda \cdot \|\theta\|_2^2 \qquad (22)$$

In the comparability indicator, which calculates the pairwise ranking loss as a relative value, the right-censored observations are considered since those observations cannot be compared to the ranking of survival time. This RankDeepSurv approach is used to avoid the possible issues associated with the traditional CPH and therefore outperform it. RankDeepSurv outperformed CPH, random survival forest (RSF) and DeepSurv in the case of highly right-censored data. (Jing *et al.*, 2019)

2.4.4 Random survival forests

Breiman (2001) proposed a concept of random forests (RF). These forests are an ensemble of decision trees with randomized splits. A few years later Ishwaran et al. (2008) presented an extension of random forest for regression, classification and survival analysis called random survival forests (RSF). RSF is an advanced machine learning method based on ensemble learning (Wang, Li and Reddy, 2017). This data-driven and non-parametric approach to conduct

survival analysis with embedded feature selection is able to achieve high performance in high-dimensional settings with highly correlated subsets of variables (Ishwaran *et al.*, 2010).

Decision tree classifiers disperse a classification problem into smaller simpler, usually binary, decisions (Safavian and Landgrebe, 1991). The classification process begins from the root of the tree and the decision rules are followed at each node until a terminal node (i.e. leaf) is achieved. These rules are data-driven (Churilov *et al.*, 2005). Figure 11 demonstrates the basic structure of a decision tree. At the top there is a root node (node 1) after which the tree branches based on the rule 1. Nodes 2 and 5 are internal nodes. Nodes 3, 4, 6, and 7 are terminal nodes, e.g. leaf nodes.

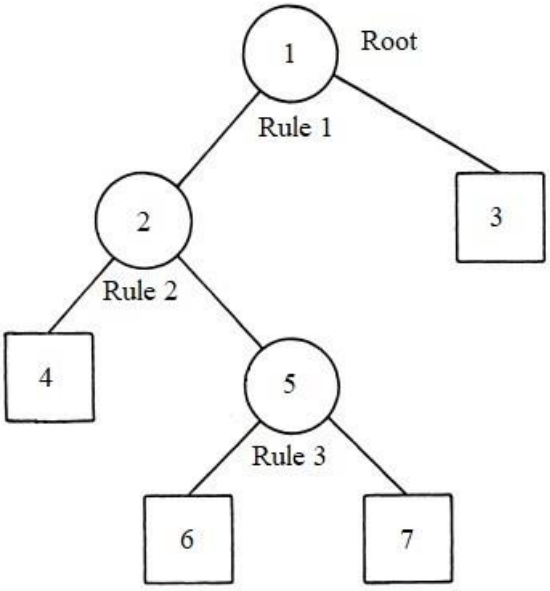


Figure 11. Example structure of a decision tree (adapted from (Segal, 1988)).

By combining multiple these independent tree-structured classifiers with identically distributed random vectors a random forest classifier is formed (Breiman, 2001). With this ensemble of classification trees a lower variance and improved accuracy could be achieved. RFs apply the Strong Law of Large Numbers and thus the overfitting will not be an issue since the model always converges. Randomization of the splits improves prediction performance compared to

AdaBoost and bagging in growing a forest. Randomization of the splits and feature selection results in a RF model robust to outliers and noise. (Breiman, 2001)

Ishwaran et al. (2008) proposed random survival forests as an extension of Breiman's (2001) random forests especially for analysis of right-censored survival data. They also presented a novel approach for missing data imputation called adaptive tree imputation. This technique imputes the missing values as the tree is grown using randomly drawn values from a set of complete in-bag data within the working node. This imputation is performed before splitting a node and thus the out-of-bag (hereinafter OOB) error estimate remains unbiased. (Ishwaran *et al.*, 2010)

The process of growing a survival forest begins by drawing a certain number of bootstrap samples from the original data. For each of these bootstrap samples a survival tree is then grown. At each node the splitting is performed based on the chosen splitting rule. The objective is to do the splitting so that the survival difference of new daughter nodes will be maximal. The splitting is continued until a given threshold for number of unique deaths at each terminal node is achieved. Finally, ensemble cumulative hazard function and OOB prediction error are determined. Another method for assessing the performance of the RSF model is the C-index (Harrell *et al.*, 1982). (Ishwaran *et al.*, 2010) Also, performance can be evaluated using integrated Brier score (Mogensen, Ishwaran and Gerds, 2012). These measures of predictive accuracy are discussed in more detail in section 2.5.4.

There are multiple available splitting rules for node splitting in survival context. Typical approaches are log-rank splitting, gradient-based Brier score splitting, and log-rank score splitting (Ishwaran and Kogalur, 2021). Log-rank can be considered as a standardized distance between the empirical hazard functions between adjacent the tree's nodes (LeBlanc and Crowley, 1993). Log-rank split preferred in noisy scenarios (Schmid, Wright and Ziegler, 2016) and performs well on censored survival data (LeBlanc and Crowley, 1993). Log-rank score splitting is based on a standardized log-rank statistic (Hothorn and Lausen, 2003). In addition to these Schmid et al. (2016) recommend the use of Harrell's C as a splitting criterion instead of log-rank splitting

when the dataset is small and censoring rate high. Other examples of possible splitting rules for RSF are random log-rank splitting and conservation-of-events splitting rule (Ishwaran *et al.*, 2008).

The variables with the most predictive importance are identified using either variable importance (VIMP), maximal subtrees, or minimal depth (Md) (Ishwaran *et al.*, 2010). VIMP quantifies the increase of prediction error if a variable is noised up (Breiman, 2001). Positive VIMP values indicate predictive variables (Ishwaran *et al.*, 2010). For maximal subtrees the importance of a certain feature is determined by its position in a tree (Ishwaran, 2007). Minimal depth (md) of a maximal subtree can be used for variable selection. Md quantifies the distance a case travelled down the tree before encountering the first split of a certain variable. More predictive variables obtain smaller values of md. (Ishwaran *et al.*, 2010) Thus meaning that these important variables are first encountered closer to the root of the tree than the terminal nodes emphasizing their importance.

However in high-dimensional ($p/n > 10$ (R Package Documentation, 2021b)) settings this md measure experiences a so called ‘ceiling effect’ which means that the tree fails to grow deep enough to identify important features (Ishwaran *et al.*, 2010). To overcome this possible issue Ishwaran et al. (2010) propose a variable hunting (RSF-VH) approach. In this approach the dimensionality of the data is reduced prior the fitting of RSF model. The process is repeated, and as a result list of significant features is obtained using a minimal depth threshold. These concepts of identifying the variables with the most predictive abilities are discussed more in section 3.2.3 in the context of our CRC data.

2.4.4.1 Tree-based survival analysis applications from the literature

Bjarnadottir et al. (2018) applied classification trees with GUIDE to predict CRC mortality. The GUIDE algorithm (Generalized, Unbiased, Interaction Detection and Estimation) was first proposed by Loh (2002). Their GUIDE model obtained higher prediction accuracy (AUC) than

CPH models only on the 30-day survival horizon. Still the model demonstrated good performance (AUC 0.88 – 0.96). Zupan (Zupan *et al.*, 2000) used decision tree induction to predict prostate cancer recurrence. The patients were divided into smaller subgroups based on their characteristics. They found out that the naïve Bayes and CPH slightly outperformed decision trees.

Jerez-Aragonés et al. (2003) combined decision trees with neural networks to predict breast cancer relapse. They applied a decision tree algorithm called CIDIM (control of induction by sample division method) to reduce the number of rules for selecting significant predictors among all variables. The resulting features are then passed on as inputs to the ANN. Delen, Walker and Kadam (2005) compared the predictive performance of ANNs, decision trees (C5) and logistic regression on breast cancer data. Differing from the other examples mentioned here, they found decision trees to perform better than the other two methods obtaining the accuracy of 93.1 %.

Katzman et al. (2018) found DeepSurv to achieve better performance RSF on providing patients with personalized treatment recommendations using both simulated and real survival data. Tseng et al. (2019) compared random forests, SVMs, logistic regression, and Bayesian classification on predicting breast cancer metastasis. In their study the RF was found to present best performance to predict BC metastasis at best three months in advance with ROC value of 0.75. Jing et al. (Jing *et al.*, 2019) examined the performance of RSFs, CPH, DeepSurv and RankDeepSurv on different medical datasets. RankDeepSurv outperformed all others in predictive abilities, and RSF achieved the worst performance. Kim et al. (2014) performed ovarian cancer survival classification using SVMs, RFs, median-based FSCOX (feature selection with Cox proportional hazard regression model), and SVM classifier using FSCOX. FSCOX with SVM performed best obtaining accuracy of 88.64 % and the RFs the worst with accuracy of 75 %.

2.4.5 Artificial neural networks

In this section the background of the artificial neural networks (hereinafter ANN) and their basic structure are presented. Also, the concepts of neuro-fuzzy systems, deep belief networks and self-organizing maps are discussed. However, these approaches are discussed only briefly as a part of this thesis, and further details are left for the readers own interest. Examples of other ANN structures for survival analysis which are left outside of the scope of this thesis are radial basis function (RBF) (Boracchi and Biganzoli, 2002), recurrent neural network (RNN) (Giunchiglia, Nemchenko and van der Schaar, 2018), and autoencoders (Macías-García *et al.*, 2017).

There have also been studies about the fuzzification of traditional ML techniques and a shift towards data-driven fuzzy systems instead of expert knowledge-driven approach (Hüllermeier, 2015). This concept of fuzzification of ANN is briefly discussed. Matsuo et al. (2019) found ANN to be superior to traditional CPH in survival prediction of cervical cancer patients. Supporting that finding, Ching et al. (2018) proposed that ANNs are better suited than other survival analysis methods for analysis of biological data because of their inherit biology-based structure, together with their ability to model complex non-linear functions (Delen, Walker and Kadam, 2005).

In 1950s soon after development of artificial intelligence, idea of neural networks started to evolve (Hof, 2013). Artificial neural networks (ANN) are inspired by the biological nervous systems (Rosenblatt, 1958), and especially by their ability to store and handle data economically, and react to several individual stimuli. The brains' cerebral cortex' structure for memory storage as a distributed system which can function economically recognizing and grouping stimuli based on their similarity is efficient. Plasticity characterises this system allowing it to adapt to changes whilst maintaining crucial basic structures. Basically, ANNs calculates response times for stimuli for a set of nodes with predefined weights. Through training this network will eventually be able to recognize specific models, e.g. speech, writing or patterns. (Hof, 2013)

Some literature refers to neural networks are connectionists systems (Kruse and Nauck, 1998) as the structure consists of different elements, nodes, connected to each other in a specific manner, and the actual information is stored in those connections between the nodes (Rosenblatt, 1958). Now, the nodes represent the neurons in the human nervous systems. The nodes gather the input stimuli given to them through the network and calculate an activation function value which is then passed forward to the network (Kruse and Nauck, 1998). To ignore weaker stimuli and disregard unnecessary noise, there is a fixed threshold value for this activation function (Rosenblatt, 1958).

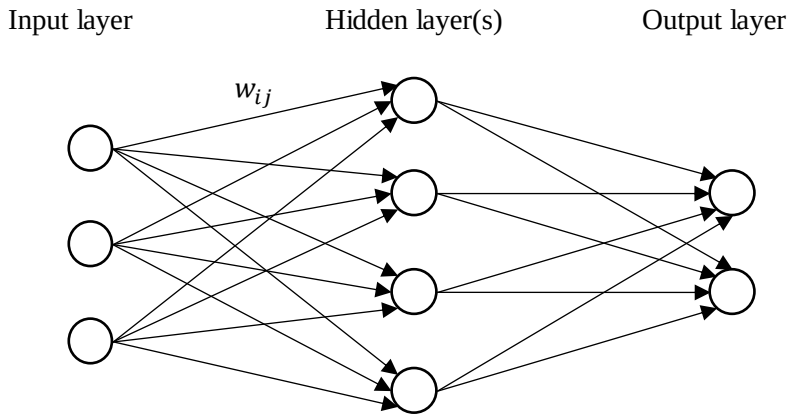


Figure 12. Structure of a simple feed-forward ANN (adapted from (Ripley, 1994; Vieira, Pinaya and Mechelli, 2017)).

Structure of a simple feed-forward ANN (Figure 12) consists of three main types of layers which are input layer, hidden layer, and output layer. The connections between the input and hidden layers (i.e. arrows in the Figure 12) are weighted (Ripley, 1994). If there are more than a single hidden layer, the network is referred as a multi-layer perceptron, MLP (Vieira, Pinaya and Mechelli, 2017). MLP is a widely used ANN structure (Delen, Walker and Kadam, 2005). Perceptron attempts to mimic this structure as a computer-based system, artificial neural network (ANN). The more hidden layers there are in the ANN, the higher the depth of the network is. In each of these layers, there are nodes as neurons which are connected to the nodes in other layers using adaptive weights. The network structure is fully-connected if and only if all the nodes in the previous layer are connected to all of the nodes in the following layer. As an output, the ANN gives out the probabilities of each input value belonging into a certain class. In the

example Figure 12, there are two possible classes for the observations to be classified into, e.g. deceased or non-deceased. The appropriate number of nodes and layers can be obtained through hyperparameter optimization before conducting the analysis with ANN (Vieira, Pinaya and Mechelli, 2017).

After the structuring of ANN is finished, the model is trained using a selected algorithm, e.g. gradient descent (Vieira, Pinaya and Mechelli, 2017). During the training progress, adaptation to the possible alterations is conducted through a back-propagation procedure, i.e. feedback from the nodes (Rosenblatt, 1958) in which the weights are readjusted "backward". The network learns through iterative training process, as it generalizes the examples fed into it (Kruse and Nauck, 1998). Now, the proposed class estimate provided by the network is compared with the actual output value, and the difference is backpropagated from the output layer back to the previous layers. Through this iterative process, the error of the network can be minimized to fall below a predetermined threshold value whilst enhancing the accuracy of the network (Abdel-Zaher and Eldeib, 2016). As an alternative to backpropagation procedure Hinton et al. (2006) proposed greedy layerwise training which consist of an unsupervised step and the following supervised step to perform the actual classification.

2.4.5.1 Neuro-fuzzy systems

Neuro-fuzzy classification (NEFCLASS) systems combine concepts of neural networks and fuzzy systems (Kruse and Nauck, 1998) as a method for supervised learning. Neural networks function here as a tool for parameter optimization which are then utilized as a part of fuzzy system (Kruse and Nauck, 1998). For these neuro-fuzzy systems the hidden layer consists of fuzzy rules (Nauck and Kruse, 1999). Fuzzy systems inherently possess the desirable feature of comprising of easily understandable linguistic rules unlike many other ML applications which may appear more of like a black box to the user of a decision-support system (Nauck and Kruse, 1999). Some users of these decision support systems might value the understandable structure of a system over higher accuracy of a more complex system. The evaluation of the system's results might seem more rational when the system itself is structured understandably.

Additionally these neuro-fuzzy systems allow the classes to have smooth, flexible boundaries instead of crisp definitions making the system less vulnerable to noise in the data (Nauck and Kruse, 1999). Basically, each observation is assigned a membership degree which describes the degree of that observation belonging to a certain cluster (Klawonn, Kruse and Winkler, 2015). Through this the system corresponds better real-life situations where the data is rarely perfect. Through fuzzification of the ML model, the uncertainty inherent to nature can be easily incorporated into the system (Hüllermeier, 2015). The rules for the fuzzy classification are based on expert knowledge about the subject. In situations where the amount of this expert knowledge to form the rules is not sufficient, the modelling needs to be supported using a data-driven approach (Kruse and Nauck, 1998). In some literature this type of an approach is referred as 'grey box modelling' where both the data and the expert knowledge are applied in modelling (Czogała and Łęski, 2000).

2.4.5.2 Deep Belief Networks

Deep belief network (hereinafter DBN) is a type of a deep neural network consisting of multiple hidden units as layers of latent variables (Hinton, 2009). These layers are connected but the units within units in each layer are not. Through unsupervised learning from the input data DBN can be applied to initialize weights for NN (Bergstra and Bengio, 2012). A drawback of this approach is its computationally intensive nature since it possesses multiple non-linear layers and many hyperparameters to optimize. As a one possible solution Hinton et al. (2006) proposed a novel fast learning algorithm for DBN utilising complementary priors in the calculation. However, DBNs are reported having high accuracy performance. Abdel-Zaher and Eldeib (2016) demonstrated application of back propagation NN with DBN for automatic breast cancer prediction problem with promising results. (Hinton, Osindero and Teh, 2006)

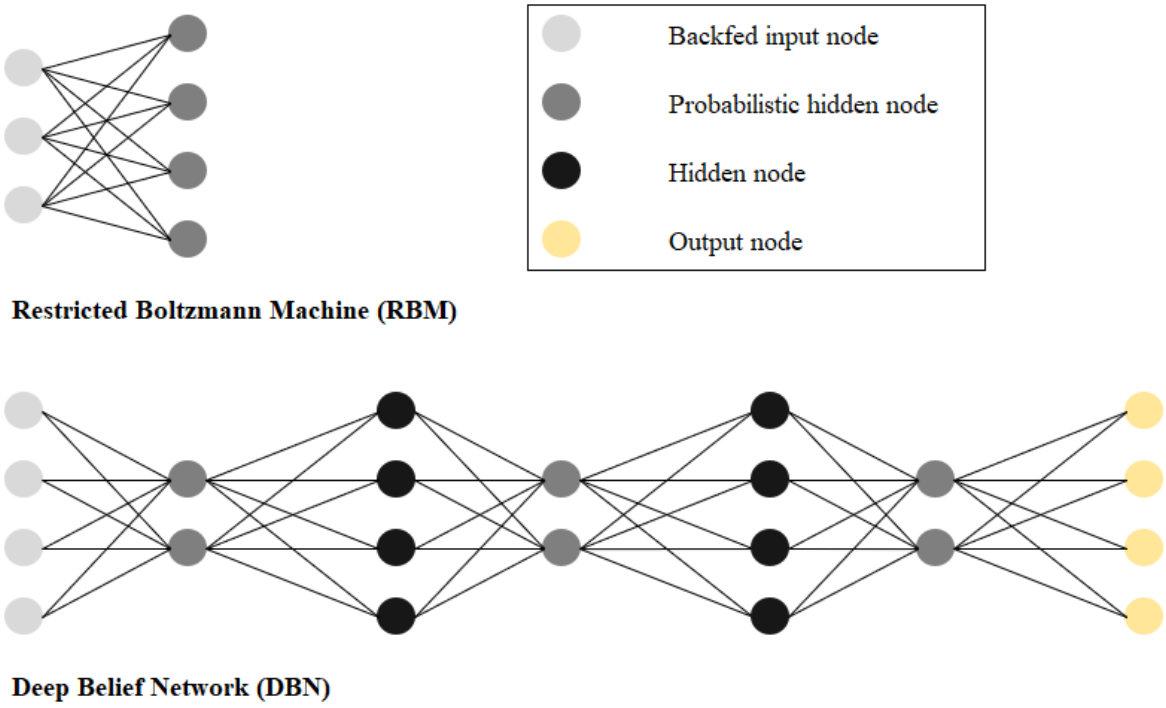


Figure 13. RBM and DBN (adapted from (van Veen and Leijnen, 2009)).

To fully understand the concept of DBN, the basic idea behind the restricted Boltzmann machine (hereinafter RBM) ought to be acknowledged. RBM is a generative stochastic ANN that uses the input data to learn their probability distribution (Hinton, 2009). Forthwith, DBN can be considered to consist of learning modules each of which is a type of RBM having a layer of visible units representing the data. In a way, the DBNs could be seen as a variant of RBM (Pacal *et al.*, 2020). The other layer of hidden units represents the features of the correlations of the data. These layers are connected using weighted connectors. These weights for DBNs can be achieved through training RBMs in a greedy manner. Figure 13 demonstrates these basic architecture of RBM and DBN. (Hinton, Osindero and Teh, 2006; Abdel-Zaher and Eldeib, 2016)

2.4.5.3 Self-Organizing Maps

The concept of self-organizing maps (hereinafter SOM) was first presented by Kohonen (1981) in 1981 and the first publication utilizing it to actual problem was three years later in a speech recognition application (Kohonen, Mäkisara and Saramäki, 1984). SOM is based on the idea of combining concepts of k means clustering together with graphical smoothing. SOM is able to project high-dimensional data into two-dimensional clusters (Churilov *et al.*, 2005). Thus performing dimensionality reduction without sacrificing the information contained in the hidden structures (Hanafizadeh and Mirzazadeh, 2011). Another beneficial characteristic of SOM is their ability to store and organize data into two and three-dimensional structures (Klement and Snášel, 2011).

SOM are able to process large datasets using their unsupervised learning algorithm (Golmah, 2014) and produce a summary of it as an output (Hanafizadeh and Mirzazadeh, 2011). The algorithm conducts clustering by dividing the input data based on their similarities and topology into clusters of approximately same size and organizes those formed clusters. In this process the intra-class (inside a cluster) similarity is maximized whilst minimizing the inter-class (between clusters) similarity. (Wei *et al.*, 2012)

Like aforementioned structure of ANN (see Figure 12), SOM also has input layer and output layer, but has hidden layers (Wei *et al.*, 2012). In input layer each node is a vector of a length equal to the number of features associated with it (Hsu *et al.*, 2009). Some literature refers output layer as map layer as the algorithm maps the input to the output as feature map (Hanafizadeh and Mirzazadeh, 2011). This structure is visualized in Figure 14 where the lines represent the weights.

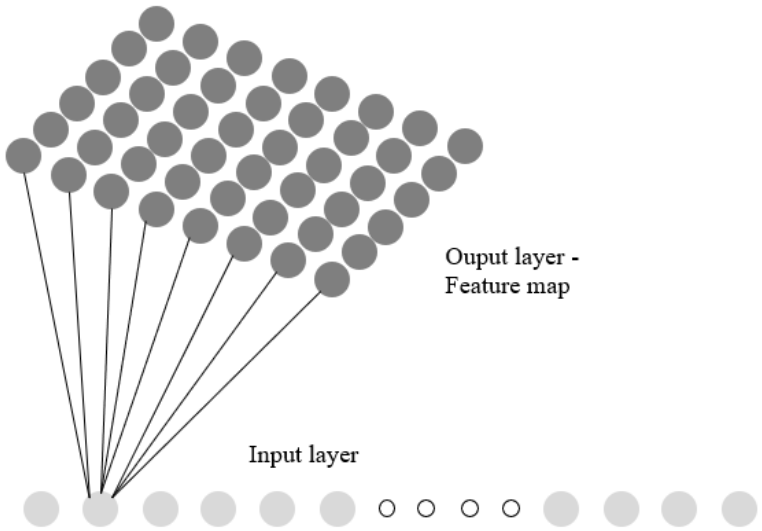


Figure 14. Structure of SOM (Hsu et al., 2009).

The SOM algorithm follows a 'winner takes it all' principle where the winning node amplifies its weights during training (Hanafizadeh and Mirzazadeh, 2011). Now the winning node referred also as best matching unit (BMU) has the weights similarly to those of input vector. Function ***(23)*** (Hsu et al., 2009) demonstrates the calculation of the amount of learning taking place in each node. From this can be observed that the key factors influencing this are neighbourhood size, $R(t)$, and learning rate, $\eta(t)$. In the function d represents the distance. As the training process continues the nodes' learning process eventually stabilizes. Now, the BMU strengthens its weights the most, and the further away a node is from this BMU the lower the learning rate is, and the weights of those nodes will not be strengthened. Nodes close to the BMU receive a bit amplified weight, so that if a stimuli with a similar pattern is introduced to the network, this process will become more efficient. (Smith and Gupta, 2002; Hsu *et al.*, 2009)

$$\eta(t) \times e^{\frac{-d}{R(t)}} \qquad (23)$$

Figure 15 demonstrates this learning occurring during the training process. X represents the input data and M_i all the different models for the network. Out of all those models M_c is the best match for the characteristics of the input vector. All of the models inside the circle in the

figure match the input vector better than the models outside of that circle. The training process focuses on those models inside the circle. The neighbourhood size affect the radius of the circle. (Kohonen, 2013)

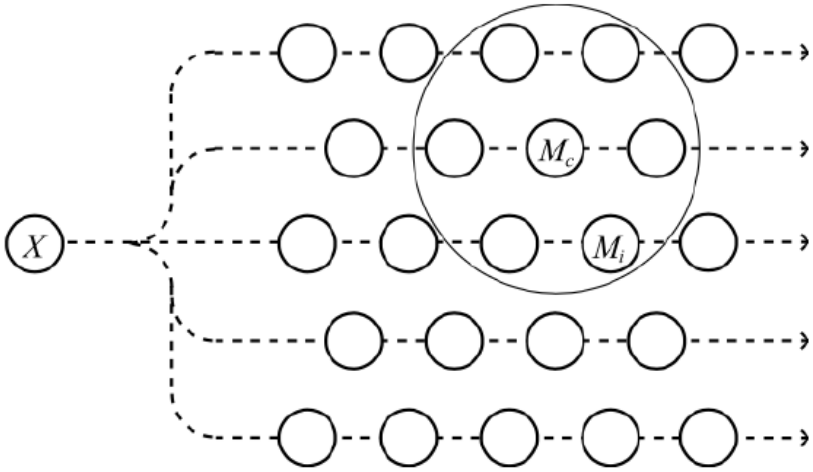


Figure 15. SOM (Kohonen, 2013).

2.5 Model validation

To characterize and measure model's predictive abilities, and validate the model, the two most common approaches are cross-validation and bootstrap technique (Picard and Cook, 1984; Kohavi, 1995). Additionally, holdout method is a widely utilized method in which the data is divided into training and testing sets (Harrell, Lee and Mark, 1996). The importance of analysts' personal experience and possible preconceptions must not be overlooked in the process of model validation alongside with validation procedures (Picard and Cook, 1984). This phenomenon is known as selection bias.

Other feasible techniques for model selection are Akaike information criterion (AIC) (Akaike, 1974; Shibata, 1981), C_p (Mallows, 1973), and the jackknife estimate of bias (Efron, 1983) which all are asymptotically equivalent to cv when the size of the validation set is 1 (Shao, 1993). The further discussion concerning these techniques is discarded from the scope of this

thesis. In this section, all these three main procedures of model's internal validation are briefly presented.

2.5.1 Holdout method

In holdout method, a dataset is divided into a training set, D_t and a test set, D_h (Figure 16). Some commonly used split ratios for training and testing are 70/30 (Que *et al.*, 2019), 75/25 (Snow *et al.*, 2001), 80/20 (Barsainya, Sairam and Patil, 2018; Bjarnadottir *et al.*, 2018), and 85/15 (Bottaci *et al.*, 1997). A test set is also referred as a holdout set, or a validation set. In some literature this method is known as the test sample estimation. A training set is then used for model building, and after that the remaining test set is applied to the model and used to validate it by calculating accuracy statistics. The estimated accuracy is calculated as the proportion of correct predictions. This is demonstrated in the formula ***(24)*** (Kohavi, 1995) where h is the size of the test set. The function calculates the average overall accuracy using a sum over all observations in the test set, D_h, divided by the number of observations in the test set. Now, the inducer function I maps the given dataset D into a classifier, and σ gets binary values (0/1) based on the accuracy of the prediction. If the predicted label for unlabelled observation v_i is the same as the possible label for instance i, y_i, the value for σ is 1, otherwise 0. (Kohavi, 1995; Harrell, Lee and Mark, 1996)

$$acc_h = \frac{1}{h}\sum_{\langle v_i, y_i\rangle \in D_h} \sigma(I(D_t, v_i), y_i) \qquad (24)$$

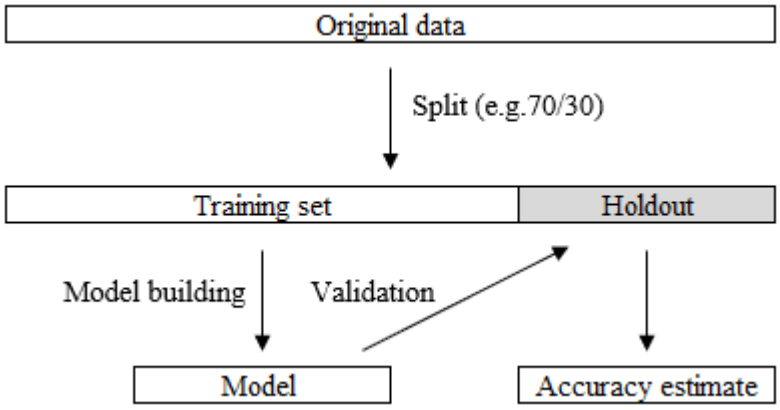


Figure 16. Holdout validation.

This method has multiple drawbacks. It is criticized of being inefficient, since a portion of the original data is left out for the validation and thus the model is not based on all data, there occurs potential loss of crucial information. This makes the holdout method a pessimistic estimator. If the test set is insufficient in size, the final accuracy estimate might experience high variability. Additionally, the results vary greatly depending on the split. The accuracy estimate can be slightly enhanced using a random subsampling in which the whole holdout process is repeated k times. Ideally, the holdout set imitates the possible nature of future observations. (Picard and Cook, 1984; Kohavi, 1995; Harrell, Lee and Mark, 1996)

2.5.2 Cross-validation

Cross-validation (hereinafter cv) is known for its ability to reduce variability in the model (Harrell, Lee and Mark, 1996). One possible drawback with cv is the high level of variability of accuracy estimates (Efron, 1983). The process is basically repeated data-splitting where the original dataset is divided into training and validation sets. The training set is referred as construction set in some studies. The model is chosen based on the predictive abilities of the competing models by minimizing the estimated prediction error. (Shao, 1993)

Cross validation can be applied for tuning and selection of model parameters (Evers and Messow, 2008; Van Belle *et al.*, 2013; Bharath *et al.*, 2018; Tseng *et al.*, 2019), and then select the best performing model (Chikha and Marzouki, 2009). Li et al. (2019) demonstrated the usage of cv for determination of regularization parameters for MSAMB (Multi-task learning based Survival Analysis for Multi-source Block-wise missing data) model. CV can be applied to both training and validation the model (Anand *et al.*, 1999; Murtojärvi *et al.*, 2020) by minimizing the possible bias (Jerez-Aragonés *et al.*, 2003) and mimicking external validation to ensure generalizability of results (Tseng *et al.*, 2019).

There are several approaches for cross-validation. The most common is the k-fold cross-validation. In addition to that also leave-one-out cross-validation (LOOCV) appear regularly in the literature. For selection of classifier 10-fold cross-validation should be preferred over the leave-one-out cross-validation (Kohavi, 1995; Schumacher, Holländer and Sauerbrei, 1997). More unusual alteration of cross-validation are generalized cross-validation (GCV) (Bates *et al.*, 1986), Monte Carlo cross-validation (MCCV) (Zhu, Li and Huang, 2019), and analytic approximate cross-validation (APCV) (Shao, 1993). These more uncommon forms of cv are discussed briefly at the end of this section. Further details considering those methods are left for the reader's own interest.

2.5.2.1 k-fold cross-validation

The objective of k-fold cross-validation procedure is to minimize the bias with random sampling and to compare different methods (Oztekin, Delen and Kong, 2009). In some studies k-fold cv is also referred as rotation validation, or rotation estimation. The process starts with dividing the original dataset, D, into k folds (subsets) of same size (Kohavi, 1995). Then the model is trained using k-1 folds, i.e. one fold is reserved for validation (Štajduhar, Dalbelo-Bašić and Bogunović, 2009). This process is repeated k-1 times, thus making each fold function as a validation set once (Kalderstam *et al.*, 2013). Figure 17 demonstrates this k-fold cv process.

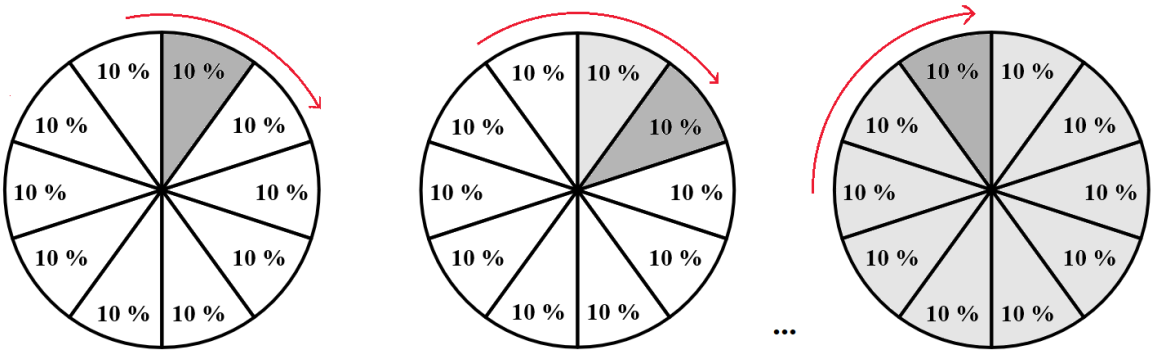


Figure 17. 10-fold cross-validation procedure (adapted from (Oztekin, Delen and Kong, 2009)).

The described k-fold cv process is then repeated using random reseeding n (e.g. 5) times, making each observation a part of the validation set n times in total (Kalderstam *et al.*, 2013; Kleinlein and Riaño, 2019). This is referred as nested cross-validation (Hosny *et al.*, 2018). Finally, the overall cross-validation accuracy estimate is obtained as the average of accuracy measures of each individual fold k (Oztekin, Delen and Kong, 2009). This overall cross-validation accuracy estimate is presented in the formula ***(25)*** (Kohavi, 1995) where D is the original dataset, $D_{(i)}$ is the test set including an instance $x_i = \langle v_i, y_i \rangle$, and it calculates the proportion of the correct classifications of all the observations, n, in the data (Kohavi, 1995). Now, both the inducer function I and σ are the same as with the holdout method discussed above.

$$acc_{cv} = \frac{1}{n} \sum_{\langle v_i, y_i \rangle \in D} \sigma(I(D \backslash D_{(i)}, v_i), y_i) \qquad (25)$$

Since the prediction accuracy estimate depends highly on the assignment of observations into the folds, by stratifying the folds the overall prediction accuracy estimate becomes less biased (Kohavi, 1995). In stratified k-fold cv, the fraction of each class in each fold is similar to the original data in size and distribution thus reducing both bias and variation of the results in comparison to the regular k-fold cv (Zupan *et al.*, 2000). Whereas with a semi-stratified k-fold cv some specific variable is applied as a basis for the splits (Štajduhar, Dalbelo-Bašić and Bogunović, 2009), e.g. stratified split w.r.t. a censored outcome variable (Kalderstam *et al.*, 2013). Stratification of the splits for cv aids to achieve the goal of obtain as unbiased accuracy estimate as possible (Oztekin, Delen and Kong, 2009).

For lower values of k (i.e. less than five), the k-fold cv is shown to give pessimistically biased results, as well as higher variance (Kohavi, 1995). This could be averted by using larger values for k (10-20), though higher values of k are shown to result in an upward bias (Efron and Tibshirani, 1997). However, low value of k could be adequate with small datasets (Tseng *et al.*, 2019) considering the number of cases in each fold is sufficient. To conclude, Kohavi (1995) and Oztekin et al. (2009) suggest the usage of stratified 10-fold cv for model building and validation.

Tseng et al. (2019) utilised nested 3-fold inner cross-validation instead of 10-fold cv since the sample size of their data was relatively small. In their approach two folds were used to train the model and the remaining fold was used to test the model. In a prostate cancer study by Murtojärvi et al. (2020) a 5-fold cv was used. 10-fold cross-validation scheme was applied in many studies (Zupan *et al.*, 2000; Jerez-Aragonés *et al.*, 2003; Delen, Walker and Kadam, 2005; Štajduhar, Dalbelo-Bašić and Bogunović, 2009; Vanya Van Belle *et al.*, 2011; Kleinlein and Riaño, 2019) identified in literature review. Kleinlein and Riaño (2019) utilised both 5-fold cv and 10-fold cv approaches for validating models predicting breast cancer survival.

2.5.2.2 Leave-one-out cross-validation

Some literature refers LOOCV also as leaving-one-out testing (Ryu, Chandrasekaran and Jacob, 2004). In LOOCV all but one sample are included to model building, and after that the finished model is testing using the sample which was left out previously. This process is repeated by using each single sample as the one being left out. Figure 18 demonstrates this LOOCV process with n observations. The total overall accuracy estimate is obtained as an average of all the individual accuracy estimates. (Mangasarian, Street and Wolberg, 1995; Chikha and Marzouki, 2009)

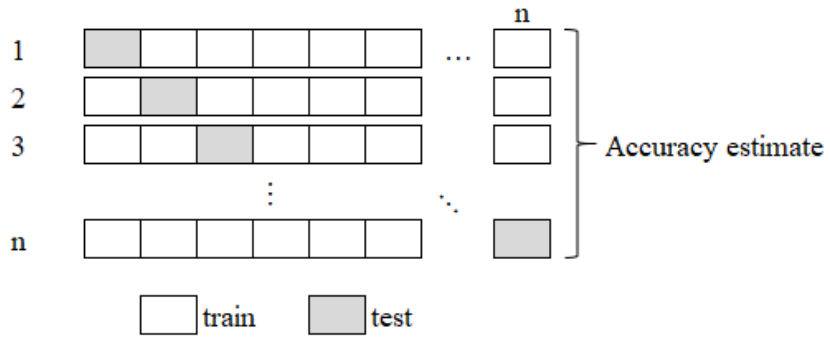


Figure 18. Example LOOCV process (adapted from (Ritari et al., 2019)).

Leave-one-out cross-validation (LOOCV) provides almost unbiased estimates which however experience unacceptable levels of variation especially with small datasets (Efron, 1983), and with unstable prediction rules (Breiman and Spector, 1992). These issues rise from the structure of this method as only a small portion of the data is used for validation at a time. In most cases, k-fold cv is more suitable of model selection compared to LOOCV (Breiman and Spector, 1992).

The complexity of calculations increases as the number of observations in the data grows, making cross-validation an inconsistent method for model selection. Thus, LOOCV is not suitable for model selection of linear models because of its inherent nature of providing inconsistent results with larger datasets. Instead, more suitable approach for ensuring the convergence would be a leave-n_v-out cv where n_v is the size of the validation set. This approach can be referred as the balanced incomplete CV(n_v) method, or abbreviated as BICV. (Shao, 1993) Further details of this method are left outside of the scope of this thesis.

2.5.2.3 Other cross-validation methods

Generalized cross validation (GCV) can be used to select a value for an estimate or aid for model selection (Bates *et al.*, 1986). In Monte Carlo cross-validation MCCV a collection $\mathcal{R}$ of b subsets are randomly drawn each having the size n_v as the size of the validation set (i.e. test

set). MCCV is also referred as repeated random subsampling validation in literature (Zhu, Li and Huang, 2019). The model is then chosen based on the minimized prediction error. Analytic approximate cross-validation (APCV) is less computationally demanding than BICV or MCCV. However, as a drawback this method is suitable only for linear models, and it fails to outperform MCCV with smaller datasets. MCCV manages to evidence more stable performance compared to the traditional cv with small samples. (Shao, 1993)

2.5.3 Bootstrapping

Bootstrapping is a data-driven procedure for calculations of statistical inferences. The origins behind the methodology are thought to be in the legendary stories about Baron Münchausen who is claimed to pull himself up by his own bootstraps to prevent from drowning into a swamp (Efron and Tibshirani, 1993). Efron (1979) presented several bootstrap estimates of prediction error to clarify the concept of the jackknife which is a nonparametric statistical method for estimation of variance and bias. This can be justified by considering the jackknife as a linear expansion for approximation of the bootstrap. In practice, bootstrap seems to outperform jackknife in estimating the variance of sample median. Here the focus is on error rate estimation using bootstrap approach. The basic bootstrap procedure and a couple of its variants, .632 bootstrap estimate and .632+ rule are briefly described. Also, OOB as a way to assess the error rate is presented.

Basically, the bootstrap procedure attempts to approximate the sampling distribution of a random variable $R(X, F)$ X is the random sample and F is the unknown distribution based on the known, observed data, x. Three alternative methods for the approximation of the bootstrap distribution are proposed to be 1st direct theoretical calculation (e.g. to estimate the variance of sample mean), 2nd Monte Carlo approximation, and 3rd expansion of the Taylor series. Now, this third approach is closely related to the concepts of the jackknife. (Efron, 1979)

The most basics bootstrapping approach begins with developing a model using all the data, in total n observations. Then, a bootstrap sample is obtained by drawing random samples B from the original data with replacing each observation back into the sample pool after it has been drawn (in-the-bag sample). This is a clear difference from the jackknife where samples are drawn without replacements (Efron, 1979). After deriving B bootstrap samples, the model is estimated in all of them resulting in B fitted models. These models are then fitted back to the original data giving B bootstrap estimates of prediction error. (Efron and Tibshirani, 1993; Zhu, Li and Huang, 2019)

This basic approach can be improved by applying the bootstrap sample to itself and measuring the prediction error between the prediction error when the reduced model is applied directly to the original data and the case when the model is applied to the bootstrap sample itself. This is referred as 'double bootstrap' estimate (Efron, 1983). Basically, this calculates the difference between the true error rate and the estimated error rate (formula ***(26)*** (Efron, 1983)). This measure is referred as optimism. Optimism describes the possible underestimation of the prediction error. (Efron and Tibshirani, 1993; Zhu, Li and Huang, 2019)

$$R\big((X,Y),(F,G)\big) = error_F - \widehat{error}_F \qquad (26)$$

Most bootstrap methods seem to outperform common cross-validation methods (Efron, 1979) which make those alluring to employ as a part of an analysis. The bootstrap approach utilizes the whole data for the building of the model which is an enormous advantage compared to other validation methods, e.g. holdout method and leave-one-out cross-validation (Harrell, Lee and Mark, 1996). It also contributes almost unbiased estimates whilst having low variance.

However, as the number of observations grows, the probability of selecting the model evidencing the highest predictive accuracy does not converge to one, making the bootstrap method

asymptotically inconsistent (Shao, 1993). It is also crucial to assess the prediction error associated with the estimated values as the bootstrap estimates are random (Efron and Tibshirani, 1993).

2.5.3.1 .632 bootstrap estimator

There exist multiple variations of the traditional bootstrap procedure. The .632 bootstrap estimator, also known as $.632(\hat{e}^{(0)} - \bar{e}rr)$, is chosen to be included as an example of bootstrap variation. The bootstrap sample is then used for training the model and remaining observations for testing to obtain an accuracy estimate. The optimism is adjusted by using only the prediction error of cases not included into the bootstrap sample. The probability that a certain given observation is included in a bootstrap sample of size n is 0.632 ($\approx 1 - (1 - 1/n)^n$). Thus, 0.368 is the probability of any given observation not being chosen after n samples. Forthwith, the bootstrap accuracy estimate is given as follows in formulas ***(27)*** (Efron, 1983; Efron and Tibshirani, 1997) and ***(28)*** (Kohavi, 1995). In ***(27)*** (Efron, 1983; Efron and Tibshirani, 1997) .632 estimator $\widehat{Err}^{(.632)}$ is the sum of observed error and the estimated error scaled using corresponding weights. In ***(28)*** (Kohavi, 1995) acc_s is the accuracy on the training set, $\epsilon 0_i$ is the accuracy estimate of bootstrap sample i (leave-one-out bootstrap), and B represents the number of bootstrap samples similarly to previous. (Efron, 1983; Kohavi, 1995; Efron and Tibshirani, 1997)

$$\widehat{Err}^{(.632)} = .368 \cdot \overline{err} + .632 \cdot \widehat{Err}^{(1)} \qquad (27)$$

$$acc_{boot} = \frac{1}{B}\sum_{i=1}^{B}(.632 \cdot \epsilon 0_i + .368 \cdot acc_s) \qquad (28)$$

Bootstrap .632 fails if the dataset is completely random, or the classifier is a perfect memorizer introducing bias to the estimates (Kohavi, 1995). With small samples .632 bootstrap seems to outperform many other variations of bootstrap methods (Efron, 1979). However, with some problems .632 experiences low variability and large bias (Bailey and Elkan, 1993). To amend possible downward biased results of .632 bootstrap estimate of error a bias-corrected version

called .632+ rule was introduced which in the presence of possible overfitting emphasizes more the leave-one-out bootstrap estimate of error (Efron and Tibshirani, 1997). Further details concerning the .632+ rule are left outside the scope of this thesis.

2.5.3.2 Out-of-bag error

Out-of-bag (OOB) error, also referred as out-of-bag estimate, can be applied to measure the possible overfitting associated with bootstrapping. OOB can be applied to measure the prediction error of random forests, boosted regression trees, and other ML models employing bootstrap aggregating for sub-sampling data. For calculation of out-of-bag (OOB) estimate the dataset is split into randomly selected samples (in-the-bag bootstrap estimates) which are used for model building, and the out-of-bag samples applied for validation. The variables are selected based on the maximum out-of-bag classification accuracy. OOB is related to the leave-one-out cv error. (Tsuji *et al.*, 2012; Suchorska *et al.*, 2019)

2.5.4 Measures of predictive accuracy

Measuring the model's predictive accuracy is crucial find a way to improve models' outcome predictions' accuracy, and thus make better choices. High accuracy of the prediction in survival analysis especially in the context of patient data is crucial for the obtaining of correct prognosis for the patient, making the treatment plans, differentiate survival analysis methods, or to investigate the effects of a certain single factor on the prognosis (Harrell, Lee and Mark, 1996). It is likewise important to identify difference in the models' discrimination abilities (Youden, 1950) to be able to select higher performing model, as well as to develop a model to be more accurate in forming predictions. Assessment of the models' calibration is the other component of predictive accuracy along with discrimination (Harrell, Lee and Mark, 1996). A few calibration metrics are presented at the end of this section.

The accuracy measures presented in this section are chosen based on the conducted literature review (Appendix 1). The metrics applied to assess the predictive accuracy of the models fitted in this thesis are discussed in this chapter. The remaining measures from the literature review are presented in Appendix 3. First confusion matrix is shortly presented since extensions of confusion matrix and R^2 are typically used to explain the accuracy of survival model's predictions (Heagerty and Zheng, 2005). Then, the receiver operating characteristics (ROC) curve and are under the curve (AUC) measures together with their time-dependent versions are discussed. After that Harrell's c-index, which is a commonly used measure in survival analysis research, is presented. To conclude this section about measures of predictive accuracy Brier score is presented. Youden index and calibration are briefly displayed in the Appendix 3.

CONFUSION MATRIX

Confusion matrix is used to assess the performance of a binary classifier. Its focus is on the number of correct classifications. It is a two-by-two matrix which differentiates the predicted and actual values of observations. Using the values in the confusion matrix, the calculation of accuracy, sensitivity, specificity, and F-measure is enabled. The basic structure of a confusion matrix is presented in Figure 19. For true positives (TP) the predicted class is true, and the actual class is also true. Whilst for false positives (FP) the predicted class is true when the actual class ought to be false. Similarly, the concepts are defined for true negatives (TN) and false negatives (FN). (Fawcett, 2006)

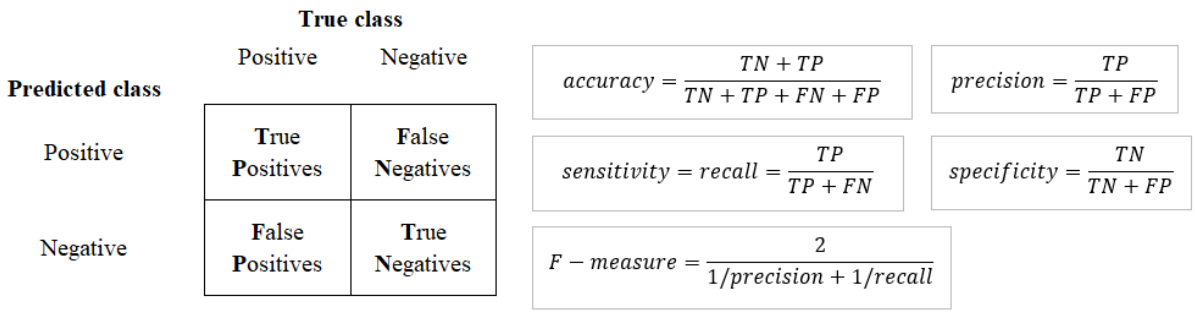


Figure 19. Confusion matrix and common performance metrics (adapted from (Youden, 1950; Fawcett, 2006)).

Classification accuracy describes the proportion of which observations were classified correctly (Zupan *et al.*, 2000). Ripley and Ripley (1998) proposed a weighted classification accuracy measure with Kaplan-Meier estimator as more suitable approach for cases experiencing censoring.

Sensitivity and specificity as correct classification rates are used for measuring the predictive accuracy of binary response models where the correct classification is conditional of the status of outcome variable, e.g. death (Heagerty and Zheng, 2005). Those both represent the predictive accuracy of the model (Antolini, Boracchi and Biganzoli, 2005). Sensitivity is the model's ability to correctly identify an individual experiencing an event, Y_i, e.g. patient with a disease, $P(\hat{p}_i > c|Y_i = 1)$, where $\hat{p}_i$ is the prediction for i and c is the criteria for classifying the prediction. Whilst specificity is the ability to correctly identify individuals who do not experience the event, e.g. a patient without a disease, $P(\hat{p}_i \leq c|Y_i = 0)$. True positive rate is the relation of true positives to total number of positives. From the definition of specificity false positive rate can be obtained as 1 subtracted with specificity. (Heagerty and Zheng, 2005) Using sensitivity and specificity values, G-mean metric ***(29)*** (Y. Wang et al., 2019) can be obtained as a geometric mean of those (Y. Wang *et al.*, 2019).

$$G - mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \qquad (29)$$

For each case individually it is important to consider which type of prediction error is more fatal. This leads to question about minimizing the misclassification costs. False positives create unnecessary burden to the limited resources which might lead to a situation in which there is not enough possibilities for treatment for those patients who are in critical need. On the other hand with false negatives patient might end up in a situation where they are given treatment until too late. (Youden, 1950)

RECEIVER OPERATING CHARACTERISTICS

Receiver operating characteristics (hereinafter ROC) curve is an extension of confusion matrix commonly used as a visual representation of the model's discrimination ability. It is commonly used especially in machine learning applications. The concept of ROC originates from signal detection theory (Meyer-Baese and Schmid, 2014). The benefit from using this approach is that is can handle cases with skewed class distributions and uneven costs of classification errors, here false positives (Fawcett, 2006). The model explains the relationship between the covariates and the outcome variable in a binary setting, e.g. 0/1. For this this model the accuracy of the predictions is of interest. ROC curve is a plot of the sensitivity, i.e. true positive rate, against false positive rate, which is 1-specificity, visualizing the accuracy of the classification rule. The curve is interpreted as the higher the curve, the more accurate the model is. (Hanley and McNeil, 1982; Antolini, Boracchi and Biganzoli, 2005; Heagerty and Zheng, 2005)

From ROC the concept of AUC, area under the curve, is derived as the probability of correct classification. AUC can be described as the model's measure of concordance between the covariates and the outcome variable. Hanley and McNeil (1982) define AUC as the probability that "a randomly chosen diseased subject is correctly rated with greater suspicion than a randomly chosen non-diseased subject". The degree or index of suspicion in medical context refers to the medical professionals' initial feeling about the probable possibility of the diagnosis prior to further examinations (The McGraw-Hill Companies, 2002). AUC is closely related to the concepts of the nonparametric Wilcoxon test statistics and that of Mann-Whitney U-statistics (Bamber, 1975).

A perfect true-positive rate is obtained with ROC curve when the graph goes through the point (0,1). Now, the AUC value is 1, meaning the model attains correct classification rate of 100 %. If the ROC curve is diagonally in 45-degree angle, the model has no discriminative abilities. The AUC in this case is 0.5 thus the model achieves only 50 % ability of discrimination which is as good as a random guess. Figure 20 visualizes these concepts. Curve A represents perfect classification, curve B adequate classification, and curve C random classification. (Hanley and McNeil, 1982; Fawcett, 2006)

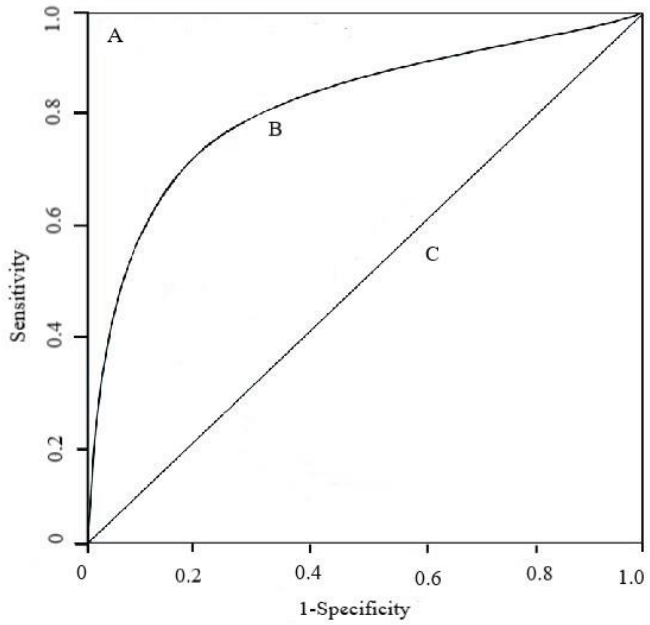


Figure 20. Example ROC curve (adapted from (Montella et al., 2020)).

TIME-DEPENDENT ROC AND AUC

Heagerty and Zheng (2005) introduced time-dependent versions of sensitivity, specificity and ROC for regression models having a censored survival time as outcome. Their approach to sensitivity and specificity as time-dependent variables is related to the concept of partial likelihood, as well as variations of Harrell's c-index as a weighted average over time of time-dependent accuracy measures. This relationship between c-index and time-dependent ROC curves conveys that an individual with a higher value of the prognostic marker, dies earlier compared to an individual with lower marker value. These sensitivity and specificity values are then used in calculations to obtain time-dependent ROC and AUC measures.

First off, incident/dynamic ROC curve ***(30)*** (Heagerty and Zheng, 2005) for incident true-positive rate given an incident false-positive rate as input at time t where p denotes the dynamic false-positive rate and is defined in the interval [0,1]. This is a combination of incident true-positive rate as sensitivity and an inverse function of dynamic false-positive rate as specificity.

$$ROC_t^{\mathbb{I}/\mathbb{D}}(p) = TP_t^{\mathbb{I}}\{[FP_t^{\mathbb{D}}]^{-1}(p)\} \qquad (30)$$

Now, the time-dependent AUC is obtained as an integral over incident/dynamic ROC curve ***(31)*** (Heagerty and Zheng, 2005).

$$AUC(t) = \int_0^1 ROC_t^{\mathbb{I}/\mathbb{D}}(p)dp \qquad \textbf{(31)}$$

Typically, this time-dependent value of AUC decreases as the time increases. Finally, a time-dependent concordance measure is given as follows ***(32)*** (Heagerty and Zheng, 2005). where τ is the upper boundary for a fixed follow-up period. Basically, this concordance measure is just a rescaled weighted average over a certain fixed time interval. The weights used can be obtained using the Kaplan-Meier estimator. This concordance is interpreted as the probability that "predictions for a random pair of subjects are concordant with their outcomes, given that the smaller event occurs in $(0, \tau)$". (Heagerty and Zheng, 2005)

$$C^\tau = \int_0^\tau AUC(t) \cdot w^\tau(t)dt \qquad \textbf{(32)}$$

HARRELL'S C DISCRIMINATION INDEX

Harrell et al. (1982) introduced concordance index based on the idea of based on the idea of Kendall's rank correlation coefficient by Brown et al. (1973), and is an extension of AUC. This goodness of fit measure is also known as c-index or Harrell's c-index. It is an index of prognostic information suitable also for cases with censored data for models which produce risk scores. It measures the model's predictive discrimination and is widely used for evaluation of survival analysis models. c-index is criticized because of its lack of ability to distinguish small differentiations in the discrimination capability between two models because of its ranking rules (Harrell, Lee and Mark, 1996).

For the calculation of this c-index a concept of permissible pairs must be familiarized. First, the data is divided into pairs and for each of those pairs a prognostic score is determined by the

underlying survival analysis model, e.g. CPH. Then, some of the pairs are left outside of the analysis. These pairs include those for which both individuals are still alive at the end of the observed time period, and if there is no knowledge of whom will outlive the other one. For the remaining pairs the prognostic scores are compared and the individual who will outlive the other one is determined. Now, the patient pair is considered to be concordant, is the individual who outlived the other one has higher prognostic score. Otherwise, the pair is disconcordant. (Harrell Jr *et al.*, 1982; Harrell, Lee and Mark, 1996)

Finally, this c-index can be calculated by dividing the number of concordant pairs by the total number of permissible pairs, which consists of both concordant pairs and disconcordant pairs (formula ***(33)*** (Harrell, Lee and Mark, 1996)). In the case of identical prognostic scores for both individuals in a patient pair, only half is added to the total number of concordant pairs instead of one. However, this kind of a pair is still considered as permissible, thus it is added as one to total number of permissible pairs. Now, c-index denotes the probability for that the individual with higher prognostic score will outlive the other from a set of randomly chosen two individuals. Essentially, the higher the c-index, the more accurate the underlying prognostic model is. This c-index could then be used to calculate the value for Somers' D rank correlation index, here denoted by γ (formula ***(34)*** (Harrell, Lee and Mark, 1996)), which measures the relationship between prognostic score and actual survival time. (Harrell Jr *et al.*, 1982; Harrell, Lee and Mark, 1996)

$$c = \frac{\#\ concordant\ pairs}{\#\ permissible\ pairs} \qquad (33)$$

$$\gamma = 2(c - .5) \qquad (34)$$

The values of c-index are in the interval [0,1] where .5 indicates that the estimated prognosis is as valid as a flip of a coin, i.e. the relationship between the prognostic score and survival time is random. Literature (Schmid, Wright and Ziegler, 2016) reports for c-index a typical variation to be in the interval of [.6,.75] in medical research applications.

Later in 2005, Antolini et al. (2005) proposed an extension of Harrell's c-index, a time-dependent discrimination index C^{td}. For this measure, the predicted individual failure times are not needed to know. Instead this approach uses for outcome prediction the whole predicted survival function, assuming the individual experiencing the event, e.g. death, has worse prognosis than the other individuals surviving longer. Therefore, the 'one-to-one' assumption between predicted times and predicted survival probabilities does not hold. C^{td} accounts for the time-dependent effect of the model covariates and the sampling variability. Thus, making it a better suited evaluation metric for cases experiencing high amounts of censoring compared to traditional c-index. Jack-knife method on correlated one-sample U-statistics is conducted to include confidence intervals. Calculation of this time-dependent discrimination index is done as a weighted average of time-dependent AUC values at time $t_{(k)}$ (formula ***(35)*** (Antolini, Boracchi and Biganzoli, 2005)) where $w(t_{(k)})$ is the probability that the concerned patient pair is concordant at time $t_{(k)}$ assuming discrete time. These AUC values represent the accuracy of the classification (of the individuals in a patient pair). C^{td} can take values in the interval of [.5, 1] where .5 is for lack of discrimination.

$$C^{td} = \frac{\sum_{k=0}^{K} AUC(t_{(k)}) \cdot w(t_{(k)})}{\sum_{k=0}^{K} w(t_{(k)})} \qquad \textbf{(35)}$$

BRIER SCORE

Expected Brier score (hereinafter BS) is a binary classification metric (Kvamme, Borgan and Scheel, 2019) originally developed for measuring the inaccuracy of weather forecasts, back when it was referred as a verification score, P (Brier, 1950). BS describes the mean squared error of prediction. Later Graf et al. (1999) proposed a generalisation of BS to account the censoring in the data (see formula ***(36)*** (Graf et al., 1999)). According to their definition for BS measures mean square error of prediction, $I(T_i > t^*)$, given estimated probabilities, $\hat{\pi}(t^* | \tilde{X}_i)$ at time t^* instead of traditional misclassification rate. In the case of individuals without observed events the event-free status is considered as an error-free prediction. Now, these event-free probabilities, $\hat{\pi}(t^*)$, are equal constants for all individuals (formula ***(37)*** (Graf et al., 1999)). $\hat{S}(t^*)$ observed rate of event-free individuals at time t^*. Misclassification rate can be

seen as a version of traditional BS. There are multiple choices for a proper loss functions to be used with this measure. The most used ones in the literature are scoring rule, logarithmic score, and empirical logarithmic score. Further details about these loss functions are left outside the scope of this thesis.

$$BS(t^*) = \frac{1}{n}\sum_{i=1}^{n}\left(I(T_i > t^*) - \hat{\pi}\left(t^*\middle|\tilde{X}_i\right)\right)^2 \qquad (36)$$

$$BS(t^*) = \left(\hat{\pi}(t^*) - \hat{S}(t^*)\right)^2 + \hat{S}(t^*)\left(1 - \hat{S}(t^*)\right) \qquad (37)$$

The function slightly alters for the case of random censorship, BS^c, see formula ***(38)*** (Graf et al., 1999). The observations for which the time of censoring happens prior to the time of interest t^*, the (survival) status at that time is unknown, so it will not be included to the BS calculations. Each individual contribution is weighted separately using Kaplan-Meier estimate of the censoring distribution, $\hat{G}(t)$. This approach of reweighting the individuals accounts for the loss of information due to censoring. Further details are left for readers' own interest to investigate. (Graf *et al.*, 1999)

$$BS^c(t^*) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(0 - \hat{\pi}\left(t^*\middle|\tilde{X}_i\right)\right)^2 I\left(\tilde{T}_i \le t^*, \delta_i = 1\right)\left(\frac{1}{\hat{G}(\tilde{T}_i)}\right) + \left(1 - \hat{\pi}\left(t^*\middle|\tilde{X}_i\right)\right)^2 I\left(\tilde{T}_i > t^*\right)\left(\frac{1}{\hat{G}(t^*)}\right)\right\} \qquad (38)$$

Integrated Brier score (IBS), see formula ***(39)*** (Graf et al., 1999), considers loss as a function of time instead of a fixed time point t^*. Essentially, IBS is obtained by integrating the empirical BS w.r.t. a weight function W(t). Time-dependent functions or alterations of marginal survival functions can used as this weight function. This IBS can be used to define R^2 residual variation measure by dividing the IBS under random censorship by Kaplan-Meier prediction, and subtracting this quotient from one. (Graf *et al.*, 1999; Štajduhar, Dalbelo-Bašić and Bogunović, 2009)

$$IBS = \int_0^{t^*} BS(t^*)\, dW(t) = \int_0^{t^*}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(I(T_i > t) - \hat{\pi}\left(t\middle|\tilde{X}_i\right)\right)^2\right\} dW(t) \qquad (39)$$

Inaccuracy can be viewed as a composition of both imprecision and inseparability. This idea is constructed into the decomposition of the expected Brier score. One of the properties of Brier score is used as a definition of a 'strictly proper scoring rule' which is a widely used forecasting scheme. This property is that Brier score takes its minimum value when the true event-free probabilities, $S(t^*|X)$, are used as estimated probabilities, $\hat{\pi}(t^*|\tilde{X})$ (Graf *et al.*, 1999). This leads to the interpretation of Brier score where a lower value indicates better accuracy of the predictions (Reijnen et al., 2020), meaning that 0 indicates a perfect prediction whilst 1 the worst (Vilardell et al., 2020).

3 Quantitative analysis

In this analysis section, we want to find out answers to the following questions shown in the Figure 21. First, our CRC data is briefly discussed, and some exploratory data analysis is performed. Then for data preprocessing the details concerning the applied imputation schemas and the validation techniques are presented following by the feature selection. For imputation three approaches are selected. First, a listwise deletion of rows with any missing values as a benchmark. In addition to this imputation by median and kNN-imputation are used. Feature selection is performed by applying three different approaches. These are correlation analysis, univariate Cox, and random survival forests (RSF). Finally the survival analysis is conducted. For this Cox proportional hazards with PH violation checks, random survival forests (RSF) and neural networks are used. Neural networks application, DeepSurv, is only briefly discussed with preliminary model fits and without further hyperparameter optimization and selection of the most predictive features. This is in Appendix 23. To validate the models, holdout method with two (2) different splits (80/20 and 85/15) and semi-stratified k-fold cross-validation (k = 5 and k = 10). To clarify the differences between RSF for feature selection and for survival analysis, the former is referred as RSF feature selection from now on in this thesis, and the latter as RSF. The evaluation of the results concludes this section. The objective is to compare the model performance with these different approaches.

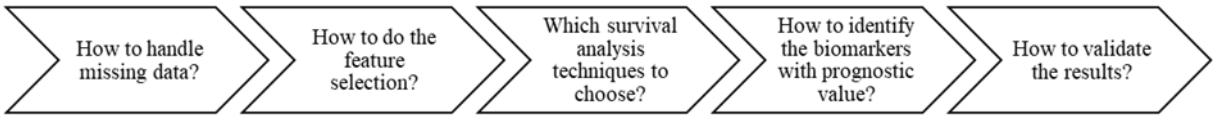


Figure 21. Questions for the analysis.

3.1 Patients and methods

In this chapter deeper insight into our CRC patient data is provided. Characteristics of the patient cohort are discussed prior preprocessing the data. For the analysis three (3) different imputation techniques are used and the removal of all observations with missing values is applied

as a benchmark. By doing this, we will be able to see which technique fits the data best. However, there is a possibility to overfit the imputation method to the data. This means that for a specific sample one method seems to work better than the others but does not if more new observations from the same distribution are available. The selected imputation techniques are imputation using median, and imputation using k-nearest neighbour (hereinafter kNN). Multiple imputation by chained equations (hereinafter MICE) is used to artificially increase the data. The usage of MICE here can be justified by the fact that the imputed data with added completely empty rows can be assumed MAR.

Statistical analyses were carried out with SPSS (IBM SPSS Statistics Version 26 Release 26.0.0.1) and R software (The R Project for Statistical Computing version 4.0.5). The main R libraries used are the following: survival (3.2-10), survminer (0.4.9), randomForestSRC (3.0.0), pec (2020.11.17), caTools (1.18.2), survcomp (1.40.0), dplyr (1.0.6), mice (3.13.0), ggplot2 (3.3.5), timeROC (0.4), stringr (1.4.0), gsubfn (0.7), and tidyverse (1.3.1). For DeepHit models survivalmodels (0.1.9) library is used. That library utilizes python through reticulate and to do that pycox and pytorch needed to be installed using Anaconda.

The CRC patient cohort consists of 318 patients and 227 variables. Out of those patients 105 (33.0 %) have survived to the end of the observation period, rest 213 (67.0 %) have deceased. The patient data consists of clinicopathologic variables describing the patients and their cancer, and tissue and serum values with added CEA, CRP, MMP-8, MMP-9 and TATI values. An oncology panel of targeted immunohistochemistry (IHC) values is included to the data as well. An overview of these 318 CRC patients is provided in the Table 4. These five (5) variables (i.e. age, sex, TNM stage, Dukes stage, and location of the tumour are chosen for a more detailed inspection out of all 227 features since those offer a broad overall picture of our CRC patient cohort. Also five (5) markers (i.e. CEA, CRP, TATI, MMP-9, and MMP-8) are briefly explained since those are included to the cohort in addition to immunopanel and Olink panel values. More features describing the treatment of CRC patients in our data are presented in Appendix 4. To capture the hazard ratio of the variables realistically, CEA and IL-6 are log transformed prior performing any analysis.

Table 4. Clinicopathologic characteristics of the 318 CRC patients (adapted from (Kasurinen, 2020)).

Clinicopathologic variable	Freq	Pct
Age < 67	155	48.7 %
Age ≥ 67	163	51.3 %
Female	152	47.8 %
Male	166	52.2 %
Tumour classification (pT)		
T1	7	2.2 %
T2	52	16.4 %
T3	154	48.4 %
T4	27	8.5 %
NA	78	24.5 %
Lymph node metastasis (pN)		
N0	131	41.2 %
N1	64	20.1 %
N2	45	14.2 %
NA	78	24.5 %
Distant metastasis (pM)		
M0	198	62.3 %
M1	42	13.2 %
NA	78	24.5 %
Dukes		
A	55	17.3 %
B	94	29.6 %
C	111	34.9 %
D	58	18.2 %
Location		
Colon dex.	87	27.4 %
Colon sin.	64	20.1 %
Rectum	167	52.5 %

CEA (carcinoembryonic antigen) is used for cancer control. Reference value for CEA for healthy non-smoking adults is below 3 µg/l. High value over 20 µg/l typically indicates cancer. (HUSLAB, 2021) CRP (C-reactive protein) is a protein produced by liver cells consequently to infections in human body, and their objective is to resist possible pathogens. Reference value

for healthy individual is below 4 mg/l (Duodecim, 2021). TATI (tumour-associated trypsin inhibitor promotes carcinogenesis (Kasurinen *et al.*, 2020) which is the process of cells transforming to cancer cells. CRP, CEA and TATI are commonly used biomarkers (Allin and Nordestgaard, 2011; Barouchos *et al.*, 2015; Tayel *et al.*, 2018) thus utilised also as biomarker for cancer. High preoperative levels of these biomarkers are associated with unfavourable survival outcome for patient (Køstner *et al.*, 2016; Tayel *et al.*, 2018; Kasurinen *et al.*, 2020). These markers and their reference values might be unknown to the reader and those are pivotal in understanding the factors affecting the survival of cancer patients, thus those are briefly presented here.

Matrix metalloproteinase (hereinafter MMP) have an essential role in metastasis, cancer invasion and thus prognosis (S. Wang *et al.*, 2019). Higher levels of MMP-8 (neutrophil collagenase) are shown to be correlated with distant metastasis, systematic inflammation and decreased survival for CRC patients (Sirniö *et al.*, 2018). However, MMP-8 possesses anti-tumorigenic and anti-metastatic functions for some cancers, e.g. breast cancer (Decock et al., 2015). MMP-9 (matrix metalloproteinase gelatinase B) activates a certain highly pleiotropic cytokine (TGF-β) that promotes tumour growth (Yu and Stamenkovic, 2000), controls apoptosis, angiogenesis and immune regulation (Prud'homme, 2007), thus increasing chances of mortality.

For both the patients' age at diagnosis and at operation the median is 67 years, and the mean is 66. For patients' age at diagnosis interquartile range (IQR) is 75.84-57.70 = 18.14. In Figure 22 the histogram and density plot of the patients' age at the time of diagnosis is displayed. The red vertical line represents the median age. In the data both sexes are represented equally, 47.8 % (152) females and 52.2 % (166) males (see Table 4).

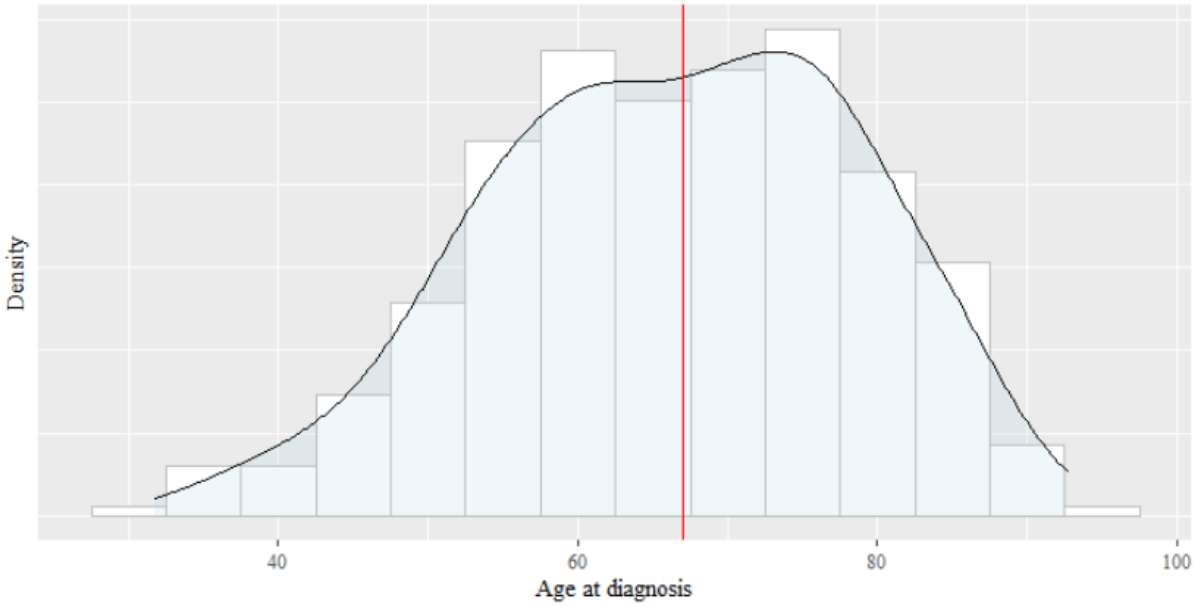


Figure 22. Histogram and density plot of patients' age at diagnosis.

The character variable describing the stage of cancer is disregarded since there is another variable describing the same using the numerical Dukes staging illustrated previously in section 2.2. The 5-year survival is defined separately for each stage of Dukes. The distribution between different stages of Dukes is shown in Table 4.

Some basic exploratory analysis of the data is conducted. Boxplots for six (6) categorical variables in the immunopanel data are displayed individually in the Appendix 5. These notions are made from just visually observing the data. These boxplots highlight the possible differences between two groups based on the censoring variable. The other group consists of patients having died from CRC and the other group of patients with other cause of death or those surviving past the observation period. In our data, patients with CRC causing their death seem to have shorter survival. On average, tumour location on the left (sin) side of the colon is associated with higher survival rate. Additionally, patients with adenocarcinomas evidence slightly higher survival rate compared to those with mucinous tumours. As could be presumed, changes of longer survival decrease as the Dukes' stage increases. Relating to this staging patient with tumours of higher grades (III - IV) seem to evidence shorter survival. Grade of the tumour refers to how differentiated it is. However, patient who did not die from CRC have remarkable higher

chances of survival. Lower grades indicate well or moderately differentiated tumours, where those are organized in a similar manner to normal, healthy tissue (University of Rochester Medical Center, 2021).

Boxplots for the continuous variables in the immunopanel data are shown in Appendix 6. From those the standardized values the trends between the two groups (those that died from CRC and the rest) of the data can be compared. Patient that died cause of CRC have higher values of interleukins 5 (IL-5), 8 (IL-8), 17 (IL-17) and 18 (IL-18), interleukin-2 receptor alpha (IL2RA), basic fibroblast growth factor (bFGF), granulocyte colony-stimulating factor (G-CSF), granulocyte-macrophage colony-stimulating factor (GM-CSF), interferon gamma (IFNγ), interferon alpha-2 (IFNα2), macrophage inflammatory protein-1 alpha (MIP-1a), migration inhibitory factor (MIF), nerve growth factor beta (b-NGF), and carcinoembryonic antigen (CEA). For the rest of the continuous variables in the Olink panel and CRP/MMP/TATI data sets the boxplots are investigated. From these illustrations can be seen the presence of multiple outliers.

3.2 Imputation and preprocessing

Some of the variables are chosen to be left out of the analysis to ease the modelling. These are all time dependent. Recurrence, post- and pre-therapies, e.g. cytostatic treatment and radiotherapy are excluded. Since all the patients having received pre-operative treatment should have survived at least until the operation for those individuals to be taken into consideration. Similarly, for patients having received post-operative treatment ought to survive from the operation at least until the start of any post-op treatment. Also, recurrence of cancer is a time-dependent variable, as it happens later post operation and other possible therapies. Thus the effect of recurrence is only accounted after it has happened. Influx and outflux values of our CRC data are briefly discussed in Appendix 7.

3.2.1 Imputation

The issue with this CRC patient cohort is the plethora of missing values. Over 33 % of the data is missing. Missing data pattern is multivariate and non-monotone. For 109 variables out of 317 and for 241 patients out of 318 there exists one or more missing values. The incompleteness of the data is remarkable. Only with 76 patients there are no missing values. The Figure 23 displays the missing value distribution of the data prior applying any imputation techniques. The number of missing values is on the x-axis, and the number of patients on the y-axis. So, e.g. there are 76 patients with complete observations, and for 73 patients 92 variable values are missing. In total 173 patients have over 90 variables with missing values. This means that over 50 % (~55 %) of the patients in our data have around 60 % of the variable values missing. For those 45 patients with only one or 2 values missing, this could be due to human error and for these observations the imputation could provide more reliable results than those with majority of values missing.

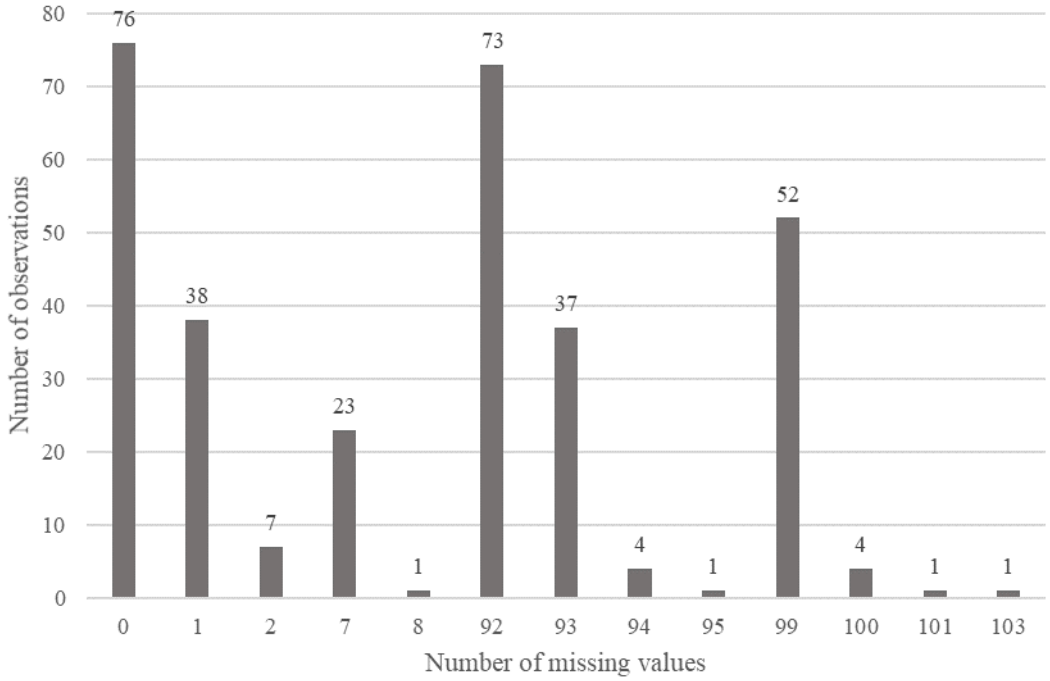


Figure 23. Missing value distribution.

Listwise deletion functions as a benchmark technique in which all the observations with any missing values are removed. Removal of rows having any missing observations could introduce bias to the final model. However, 76 observations might not be sufficient to build any survival analysis model. Even if the data is artificially augmented using imputation, the idea of using this approach ought to be considered with caution. Although this listwise deletion approach is included as a part of this thesis.

To obtain a more sufficient cohort size, imputation is used to complete the data. For the imputation three (3) techniques are chosen. Besides listwise deletion imputation by median, and kNN are utilised. MICE is not suitable for our data since it assumes MAR (Azur *et al.*, 2011), and the data has MNAR. Further details these techniques are discussed earlier in section 2.3.1. This section is supported by the conducted literature review. Three (3) different approaches for handling missing data are sufficient to fulfil the scope of this thesis.

However, after these imputation approaches the dataset still suffers from small sample size. Building a model using low sample size increases the margin of error which is not desirable. This issue is handled by enhancing data artificially using multiple imputation by chained equations (MICE). Here the use of MICE can be justified since the added empty rows in the imputed datasets can be assumed as MAR. The usage of mice at this point can be justified. New data is then validated though cross-validation. Further enhances to this data enlargement are left outside the scope of this thesis. Other approaches for this exist, e.g. fuzzy similarity based classifier utilizing Łukasiewicz structure (Luukka, 2008) which can cope with small sample size and a mega-trend-diffusion technique (Li *et al.*, 2007). Basically, these approaches predict new data based on the existing values.

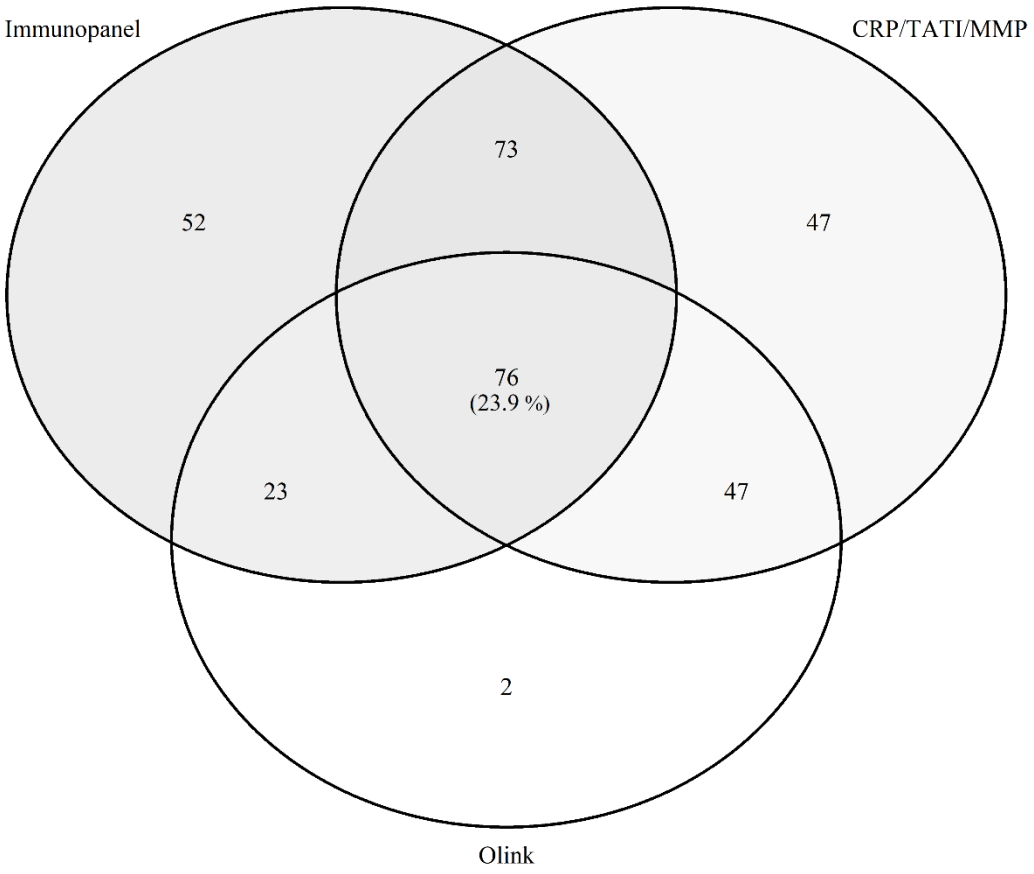


Figure 24. Venn diagram of complete patient records of all three datasets.

Prior imputation all the data is recoded as numerical data containing continuous and categorical variables. Figure 24 demonstrates the number of complete patient record values for each dataset. For the CRC immunopanel data only less than 1 % (~0.57 %) of the values are missing. 11 variables out of 59 have at least one missing value. This is almost 19 % of all the variables. Additionally, 91 patients out of 318 have at least one variable without a recorded value. Therefore nearly every third (~29 %) patient does not have a complete record. In both the dataset containing CRP, TATI and MMP values (243 observations) and the Olink panel (148 observations) all the observations are complete. After joining the immunopanel with CRP, TATI, MMP and the Olink panel values the portion of the missing values becomes more remarkable, over 33 %. This means that one third of the data is missing. Out of 156 variables 109 have at least one missing value, which means that almost 70 % of features' values are not completely recorded. For the patients 242 out of 318 have at least one variable without a value. This means that only a bit under quarter (~23.9 %, see Figure 24) of the patients have complete records,

whilst over 75 % of the patient records are incomplete. Thus, the missingness is remarkable in our data and it is important to find a way to impute these missing values. If we would conduct listwise deletion to remove missing values, the datasets would become too small for performing survival analysis. A possible approach here could be removing the features having many missing values before imputation. However, this is not done here as the focus is on imputation techniques. Another approach could be to identify variables with particularly many missing values and delete those variables. This could lead to losing a potentially important feature. All the variables with the number of valid and missing values, and the original data source are shown in Appendix 8.

The listwise deletion is applicable only if the analysis would be conducted only using just the immunopanel data without added CRP, TATI, MMP and the Olink panel values. In this case, only 0.57 % of the data would be missing, and ten (10) variables and 91 patients would have incomplete values as already mentioned above. Thus a benchmark approach of listwise deletion results a patient cohort of 227 patients which should be a sufficient size. Here, only relevant variables from the immunopanel data are included for the analysis. Essentially, this data consists of patients' serum and tissue values together with CEA, information about the location (colon dex., colon sin. or rectum), Dukes' stage (A – D) and histopathologic characteristics of the tumour (adenomatous or mucinous), disease-specific survival censor and survival time. Also, patients' age and gender are included as demographics. However, the focus in this thesis concerns all the three datasets merged. Further analysis of the immunopanel data individually is left outside of the scope of this study.

First, imputation using median is computed. In this approach all the missing values are imputed using the median value of the variable in question. Imputing using median alters the distribution of the variables slightly. This concerns especially categorical and binary variables with only a few classes. Considering our quite small cohort size, the results using this imputation technique should be interpreted with caution. However, most of the data are biomarker values which are continuous. The usage of median to complete the data generalizes the values based on the observed data. Thus, reinforcing the trend of the data.

The second approach is the imputation by kNN (k-nearest neighbours). This is done applying the R's knn.impute() function from bnstruct package (version 1.0.11). For this the indices of categorical variables need to be defined. For imputation mode is used for discrete variables, and median for continuous variables (RDocumentation, 2021). Number of neighbours used here is ten (10). All available data is used to search for neighbours. After the imputation the data has 318 complete observations.

After imputing the missing values of the data, there are 76 patients for listwise deletion, and 318 patients for both median and kNN imputed datasets. Next, the size of dataset is artificially enhanced to 500 patients using MICE. For this mice function from mice package (version 3.13.0) with method random forest and maximum iterations set to 50 is used. Prior mice imputation new observations with missing values are created at the end of each dataset. After the process each of these three (3) datasets has 500 patients with complete values for 157 features. However, the dataset with listwise deletion has increased the number of observations remarkably, almost 550 %. To avoid issues with reliability, only the actual, non-imputed data is used in the test set. Thus only training data is subject for the artificial enhancement.

To assess the reliability of the artificially created new data values some descriptive statistics are calculated. The results are summarised in Table 5. The table presents the mean, median and standard deviation values for four variables; carcinoembryonic antigen (CEA), interleukin 6 (IL-6), taxilin alpha (TXLNA), and matrix metallopeptidase 9 (MMP9) individually for each of the three datasets imputed using different methods. The selection of these variables is supported by the oncologic literature. Each of these are shown to be connected with CRC in some manner. In addition, there is at least one variable from each of the original datasets; CEA and IL-6 from the immunopanel, MMP9 from CRP/MMP/TATI data, and TXLNA from the Olink panel. The results show that the mean and standard deviation are quite similar between the original and mice imputed values for median and kNN imputed datasets. There is a slight difference in mean, median and standard deviation for CEA for listwise deletion data. However, other displayed values are relatively similar. In addition, the distributions of these chosen variables prior and post mice imputation are presented in Appendix 9999. From those can be observed that the distributions of these variables are almost the same. Also, after imputing the

missing values for the added rows. the values for categorical variables are check in order that those are within a correct range. Thus, can be presumed with caution that the imputed values can be used for conducting survival analysis.

Table 5. Comparative statistics before and after adding rows using imputation.

	Listwise deletition					
	Actual mean	Imputed mean	Actual median	Imputed median	Actual sd	Imputed sd
CEA	142.4	53.8	3.1	4.1	1139.3	633.5
IL-6	21.9	21.9	14.3	14.3	25.0	24.1
TXLNA	4.7	4.6	4.7	4.7	1.2	1.1
MMP9	232.9	228.6	220.0	220.0	121.4	109.7
	Median imputed					
	Actual mean	Imputed mean	Actual median	Imputed median	Actual sd	Imputed sd
CEA	69.0	77.5	3.3	3.4	615.0	667.3
IL-6	28.8	31.2	14.1	16.5	100.1	92.2
TXLNA	4.5	4.5	4.5	4.5	0.8	0.9
MMP9	213.3	222.1	203.2	203.2	104.4	105.8
	KNN imputed					
	Actual mean	Imputed mean	Actual median	Imputed median	Actual sd	Imputed sd
CEA	69.0	53.8	3.3	3.3	615.0	494.3
IL-6	28.8	29.2	14.1	14.9	100.1	88.4
TXLNA	4.6	4.6	4.6	4.5	0.8	0.8
MMP9	212.5	214.4	201.6	205.8	107.3	108.7

3.2.2 Validation

To validate the models, two approaches are utilized: semi-stratified k-fold cross-validation and holdout method. Considering the relatively small size of our CRC patient cohort, the data is split to five (5) and ten (10) folds. In each of these folds the percentage of censored and uncensored patients is preserved, i.e. the status variable is used as a basis of forming the folds. Cross-validation allows the model to be trained, validated, and tested properly despite the small size of the dataset (Matsuo *et al.*, 2019). The potential issue having imputed values in the validation and test sets with k-fold cross-validation is disregarded here. The concept of k-fold cross-validation is described in more detail earlier in section 2.5.2.1.

For comparison, holdout method (see section 2.5.1) for model validation is applied. All three (3) datasets are randomly divided into training and test datasets using a split ratio 80/20. This ensures that the data used to build the model (80 %) is separate from the test data (20 %) applied to obtain realistic performance measures for the model. The test set contains purely fully known observations. Any of the imputed values cannot be used for the test sets since strictly speaking those values are unknown and there is no knowledge whether the imputation is accurate or not. Thus these values cannot be used to validate the model. All the imputed observations are included to the training set and only the complete data, for which the values are known with certainty assuming there are no measurement errors, is used for testing. Furthermore, split ratio 85/15 is tested. Thus altogether two (2) validation approaches are tested.

This might create some issues considering the small size and the major incompleteness (33.2 % of the values are missing) of our CRC patient cohort. Initially, there are only 76 patients with complete observations. This restrains the maximum size of the test set to be approx. 15 %. From this arises another possible issue, that the training data consists mainly of patients with at least one imputed value. There is a slight controversy with the split to train and test sets. Now, after artificially enlarging the sample size to 500 patients, the test set size using 80/20 split would be 100 patients and there are only 76 patients with totally complete observations at the beginning (see Figure 24 in section 3.2.1). This implies that with this split ratio there has to be some observations with imputed values in the test set, which might create bias to the final results. To obtain a full 80/20 split the remaining 23 patients are selected from the set of patients which have only a single missing value. The cohort of these patients has total 38 patients. Additionally, using 85/15 split this issue is not present. Since in that case, the size of the test set would be 75 (< 76). However, this situation with validating using test data with imputed values can be avoided using cross-validation.

3.2.3 Feature selection

For this thesis correlation analysis, univariate CPH, random survival forests with both minimal depth (md) and variable importance (VIMP) are used for feature selection. As our data is quite small, and thus the sample size might be insufficient to identify the underlying manifold, it is advisable to prefer models with in-build feature selection (Pölsterl *et al.*, 2016). Further details and applications of other methods are left outside the scope of this study. Some studies (Li and Razzaghi, 2019; Reijnen *et al.*, 2020; Wang, Wang and Makond, 2020) perform feature selection based on literature review. However, this approach limits the feature space to obey previous findings and thus pre-empt the possibility of novel discoveries. Our data has more features (p) than observations (n). This case when p>n is referred as a curse of dimensionality, which can be resolved using feature selection techniques. These techniques aim to reduce the number of predictors without sacrificing the predictive performance (Kuhn and Johnson, 2019)

3.2.3.1 Correlation analysis

First approach is to remove highly correlated variables from the datasets. Sometimes it would be considerable to combine feature that have a strong correlation. However, this is not done here because of the high-dimensionality of the data. The threshold is set so that all variables having correlation higher than 0.8 are removed. Additionally, correlation bound of 0.7 is used since the higher values failed to exclude almost any features. However, with the lower value the approach still fails to perform effective feature selection. It is not reasonable to lower the correlation bound even further down. All the patients in the benchmark dataset with the immunopanel data joined with CRP, TATI, MMP, and the Olink panel have colon cancer, hence the variable describing the location of the tumour is excluded from that dataset. This might introduce some bias to the model since there are no patients with rectal cancer. This procedure is conducted from all three (3) datasets. The number of features for each of these three (3) datasets before and after removal of highly correlated variable is demonstrated in Table 6. Table indicating the included variables for each set is displayed in Appendix 10. From that table can be seen that this method excludes only a few variables and is not very effective method of

feature selection on our data, even with a lower correlation bound. Thus this approach is excluded as a feature selection technique for our CRC data and further investigation is not performed.

Table 6. Number of variables per datasets before and after removal of highly correlated variables.

Correlation bound	Imputation	Nro. of features	
		Ante	Post
0.8	Listwise deletion (BM)	153	105
	Median imputation	154	123
	kNN-impute	154	123
0.7	Listwise deletion (BM)	153	86
	Median imputation	154	105
	kNN-impute	154	106

3.2.3.2 Univariate Cox

Second approach is to perform a univariate CPH to select the features most related to predicting survival. For that all the categorical variables are converted to factors with defined levels. This concerns seven (7) features. Features are selected based on their FDR (false discovery rate) - corrected (Benjamini and Hochberg, 1995) p-value in the univariate CPH. Threshold for selecting features based on their FDR-corrected p-value is $p < 0.10$. After this exclusion the benchmark data (listwise deletion data) has 1 variable, median imputed data 12, and kNN-imputed data 49 variables.

The resulting p-values and FDR-corrected p-values for each of the datasets are shown in Appendix 11. The covariates which have FDR-corrected p-value < 0.1 are highlighted in the table. The Figure 25 summarises the features selected by univariate Cox for all the datasets. There is only one (1) variable in common for all these three approaches. This is mucin 16 (MUC16) which is also known as CA-125 (carbohydrate antigen 125) (Felder *et al.*, 2014). Since the listwise deletion data has such few observations the results from this univariate analysis can be

foreseen. For this data the model reduces to univariate. Additionally, kNN-imputation is much more sophisticated approach than imputation using median. Thus explaining the difference in chosen features for these datasets. All the features selected for median imputed data are also chosen for kNN-imputed data.

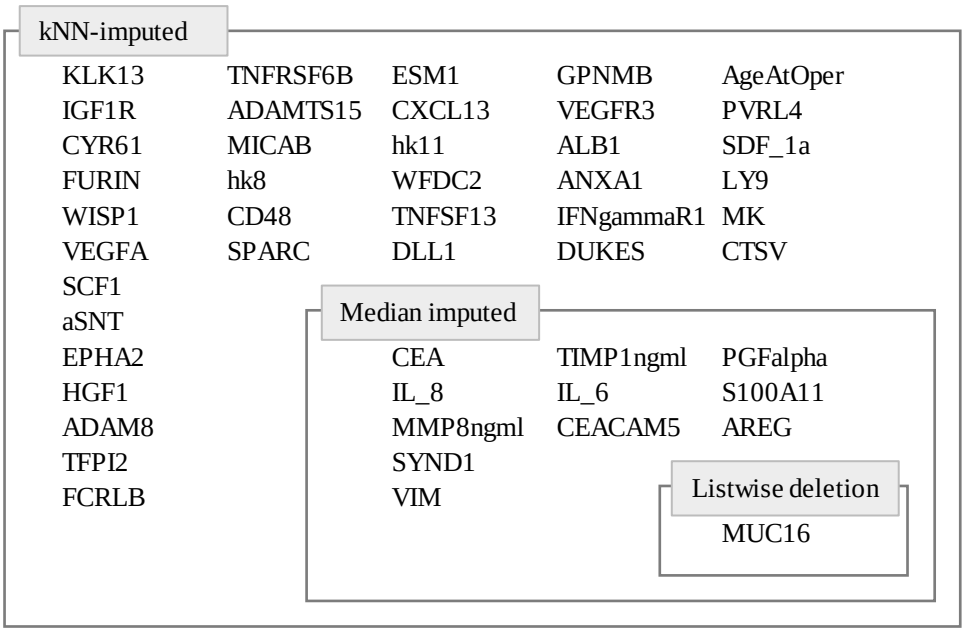


Figure 25. Venn diagram of the features selected by univariate Cox for all the datasets.

3.2.3.3 Random Survival Forest

Third approach is to utilize random survival forests for determining important variable. In this thesis random survival forests (RSF) are applied for both feature selection and survival analysis. RSF has an inbuild feature selection. Further details about the adaptation of RSF for our CRC data are discussed later in section 3.3.2. The most important variables for predicting the survival of CRC patients in our data are identified using two (2) approaches, minimal depth algorithm (md) by Ishwaran et al. (2010) and variable importance (VIMP). Since VIMP procedure contains some random elements, it would be ideal to repeat the model building a few times and average over the results to obtain the most realistic outcome. Thus the process is repeated ten (10) times. Minimal depth is an extension of maximal subtrees, and especially well-suited for

high-dimensional data cause of the lack of random elements in the calculations (Ishwaran *et al.*, 2010). Variable hunting and variable hunting with variable importance (VIMP) are not used as variable selection methods since those assume the number of features to be substantially larger than the number of observations, e.g. the ratio (p/n) is greater than ten (10) (R Package Documentation, 2021b). These ratios in our datasets are [0.308, 0.31] where the first one is for listwise deletion (benchmark) data with the location variable excluded.

The RSF models for all three (3) datasets are fitted using 1,000 trees. For splitting rule log-rank score is applied which is recommended for survival analysis (R Package Documentation, 2021a), and bootstrap protocol is to bootstrap data without replacement. Minimum size of terminal node is set to 15 (default value for survival analysis) and number of variables randomly selected as candidates for splitting a node (mtry) is a square root of the number of features. The results are presented in Appendix 12. With minimal depth variable selection the selected model sizes are 63 for listwise deletion (benchmark) data, 54 for median imputed data, and 47 for kNN-imputed data.

For VIMP block size of ten (10) is used. The VIMP procedure is repeated independently, and the variable importance values are averaged over the iterations. Thus makes the predictiveness of the covariates more interpretable. The process is repeated ten (10) times. Only variables that inhabit predictive ability (VIMP > 0) are selected. Using the mean VIMP for 10 independent iterations for listwise deletion data 44 features are selected, median imputed data for 86, and 130 for kNN-imputed data. With minimal depth approach 63 features are selected for listwise deletion data, 54 for median imputed data, and 47 for kNN-imputed data. For final selection the features chosen by both minimal depth approach and VIMP are determined for all the datasets. Thus, for listwise deletion data 29 features are chosen, for median imputed data 38, and for kNN-imputed data 47. Out of these selected features, listwise deletion data and median imputed data share 14. Listwise deletion also has 14 selected features in common with kNN-imputed data. There are 18 same features chosen for median and kNN-imputed datasets. For all these three datasets there are nine (9) in common. These variables are shown in Table 7. The rest of these results are displayed in Appendix 12.

Table 7. Features selected using md and VIMP in common for all datasets.

Listwise deletion & median	Listwise deletion & kNN	Median & kNN	All three
MUC16	MUC16	DUKES	MUC16
DUKES	DUKES	CEA	DUKES
SYND1	SYND1	MUC16	SYND1
LYPD3	LYPD3	AgeAtOper	LYPD3
CEACAM1	Frgamma	IL_6	CEACAM1
KLK13	CEACAM1	IL_8	KLK13
HGF	KLK13	LYPD3	PGFalpha
PGFalpha	IGF1R	KLK13	CPC1
IL_2Ra	FCRLB	SYND1	TIMP1ngml
MIP_1b	PGFalpha	TIMP1ngml	
MMP8ngml	MSLN	diff	
CPC1	CPC1	CEACAM1	
TIMP1ngml	CXCL13	AREG	
IL_9	TIMP1ngml	CEACAM5	
		CPC1	
		PGFalpha	
		ITGAV	
		TLR3	
Total selected features in common			
14	14	18	9

3.3 Survival analysis

The outcome of the study is to predict disease specific survival (DSS) for different time intervals. These intervals are 1-year to n-year survival where n is in {1, 3, 5}. The outcome variable describes the probability for whether the patient survived or died for a certain period of time. For clarity, the focus of this study is on analysis conducted for death due to CRC. For each of these time horizons number of patients remaining, number of CRC-related deaths and other deaths or censored observations for each of the datasets are displayed in Figure 26. Since there exists right-censoring in our CRC patient cohort shorter time intervals (e.g. 1-year compared to 5-year) have more data than in the longer time intervals.

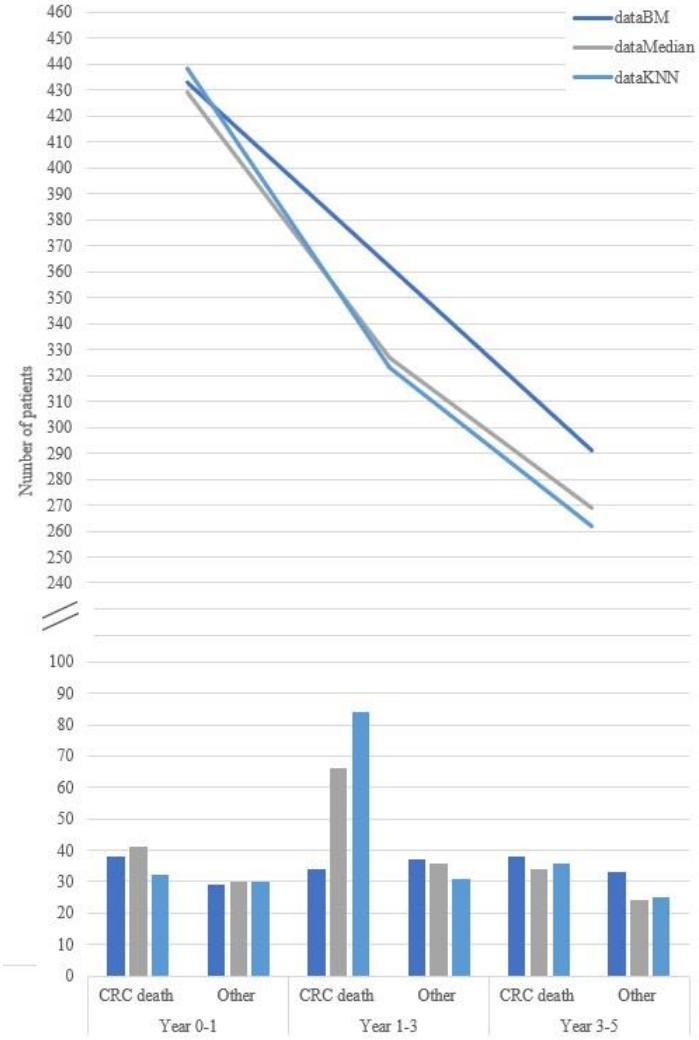


Figure 26. CRC-related deaths and censored observations for selected time intervals.

For all survival analysis methods applied, the censoring status is required to be coded in the same manner. This means that the method requires so that status 1 indicates death (event), and 0 censoring (Ishwaran and Kogalur, 2021). For this thesis all other deaths than those caused by CRC are assigned censoring status of 0. This approach is taken in all models.

First model type is Cox proportional hazards. Because of the high-dimensionality of our CRC data feature selection prior fitting the model is required to avoid convergence issues. CPH models are fitted for each of the three (3) datasets using the variables selected using univariate Cox

and RSF feature selection. Correlation analysis failed to reduce the number of features significantly. Because of that it is not applied further. The second survival model used is random survival forest (RSF). For RSF the feature selection is embedded, thus eliminating the need to perform separate feature selection.

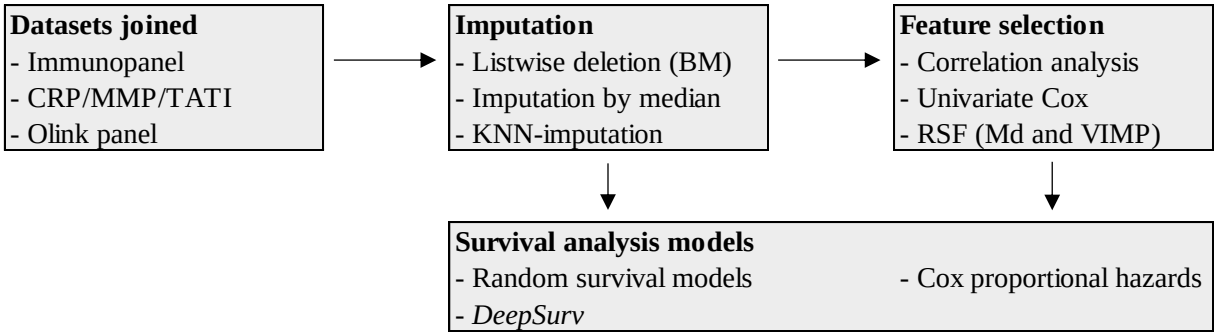


Figure 27. Data analysis scheme.

In addition neural networks are included as preliminary fits using DeepSurv without further hyperparameter optimizations. These results are provided in the Appendix 23. Deeper and more detailed analysis using these techniques could be carried out in the future. However, this more insightful insight is left outside the scope of this thesis. Figure 27 summarises the survival analysis models used with the specified imputation and feature selection techniques. Essentially, there are two (2) main models, Cox proportional hazards (CPH) and random survival forests (RSF).

Both Kaplan-Meier curves and Nelson-Aalen fitter can be used as indicators for the populations' overall survival. The theoretical framework concerning these curves is discussed in sections 2.4.2.2 and 2.4.2.3. Nelson-Aalen fitted could be applied to multistage model where other causes of death are competing risks for cancer-related death. However, this competing risks analysis is left outside the scope of this thesis. Appendix 13 demonstrates the Kaplan-Meier curves for all three (3) datasets, listwise deletion as benchmark (Figure 35), median imputed (Figure 36), and kNN-imputed (Figure 37). The curves for the median and kNN imputed datasets are almost similar. The KM graph for BM data differs from those. The difference in the

shapes of these curves could be partly explained by the fact that both in median imputed and kNN-imputed datasets there are more censored observations than CRC-related deaths.

The slope is decreasing quite rapidly until year 5 for all datasets. From there on there is only a gentle decreasing slope in the possibility of survival. The second major drop in survival probability estimates comes around year 12. The rough estimates for 5-year survival from these KM curves are 75 % for BM, 68 % for median imputed, and 71 % for kNN-imputed datasets. These estimates are roughly the same than the figures (64.15 % for males and 69.03 for females) reported by the Finnish Cancer Registry (2021). In our data, there are no significant differences in the overall survival between the genders. However, as studies have demonstrated (Janssens *et al.*, 2018) right-sided colon cancer seems to be associated with worse survival rates compared to the left-sided cancer.

Further, Nelson-Aalen estimate curves for these datasets are displayed in Appendix 14. Those demonstrate the cumulative hazard against time. The hazards cumulate exponentially over time. Like KM curves, these are used to give an average view of the population. Those correspond to the KM curves displaying the overall survival possibility. The graphs for both the median-imputed and the kNN-impute datasets are almost similar. Likewise with KM curves there can be observed a change in the slope around year 12. After that the cumulative hazard function increases much more drastically over time. The graph for listwise deletion data the cumulative hazard increases in a linear like manner until the end of the observation period when the hazard starts to increase exponentially.

3.3.1 Cox proportional hazards model

In this section the Cox proportional hazards (CPH) models are fitted for all three (3) datasets using the features identified previously in section 3.2.3. The utilised approaches for selecting the covariates that show predictiveness are univariate Cox, and RSF-based methods; md- and

VIMP- analysis. Therefore two (2) different sets of features are tested with CPH, and in total six (6) CPH models are built with different validation approaches.

Prior applying the model to the test data, the proportional hazards assumption is evaluated analysing the Schoenfeld residuals against the transformed time. If the test for trend is not statistically significant for none of the covariates and globally, the proportional hazards can be assumed (UCLA, 2021). This observation of not violating the PH assumption can be verified through visual inspection of Schoenfeld residuals plotted against transformed time for each covariate. Schoenfeld plot should demonstrate a non-zero slope with residuals plotted in a random walk manner around this zero value mean line in order the proportional hazards not be violated. (Schoenfeld, 1982). Basically, if there is no clear pattern in the residuals over time, the proportional hazards assumption holds.

Ideally, presence of the potential influential observations and outliers in the datasets ought to be checked. In the case this PH assumption is violated, potential solutions could be to add the problematic variables to the model as stratified, change the functional form of the regression variables, and adding time interaction terms. Since the focus of this thesis is about the comparison of different imputation methods, feature selection and survival analysis techniques, if there are any features experiencing nonproportional hazards it is only acknowledged, and not acted upon. Further processing of variables violating the PH assumption is left outside the scope of this thesis. These features violating the PH assumption are presented in Appendix 15.

For CPH models using features selected by univariate Cox, for both listwise deletion and median imputed data there are no PH violations. Additionally for listwise deletion data with variables from RSF feature selection there is only a single variable violating the PH assumption in both splits. For CPH with median imputed data and RSF feature selection there are four (4) features violating the PH assumption. With CPH models fitted using kNN-imputed data with both univariate Cox and RSF for feature selection there are multiple features violating the PH assumption. However, it ought to be emphasized that for kNN-imputed data more features are selected for the fitting of CPH models than for either listwise deletion data or median imputed

data. Thus could be assumed that as the model size increases the possibility of more features violating the proportional assumption increases as well.

The results for fitted CPH models are presented in Appendix 16. There are separate tables for all validation approaches. These tables contain information for each of the model about possible PH violations, concordance of the train set, and values for likelihood ratio test, Wald test, and log-rank score. In addition to this survival predictions using the test data are provided for year 1, year 3, and year 5 in Table 9. Instead of using full years for these predictions, quantiles of survival time could be used. However, this approach is left outside the scope of this thesis. Also, the statistically significant marker values identified by the CPH model are presented. kNN-imputed dataset with both splits and both feature selection techniques achieve the highest concordance on the test data, approximately 0.77. For CPH models which validated using a semi-stratified k-fold cross-validation, c-index values for train and test data with variances, and predicted survival probabilities and their dispersion are summarised in Appendix 17. The results from these k-fold cv models support the findings of holdout validated CPH models. Further analysis of those cross-validated results is left outside the scope of this thesis.

The models built using the listwise deletion (BM) data, the predicted 5-year survival probabilities seem to be a bit overoptimistic ([75.5 %, 76.2 %]). For models with median imputed data, the predicted 5-year survival probabilities are a bit lower with variables selected using univariate Cox feature selection ([72.8 %, 73.9 %]) than with those from RSF feature selection ([76.7 %, 77.6 %]). Predicted values using the kNN-imputed data experience a similar pattern to the values from median imputed data, i.e. values obtained using features from univariate Cox ([69.3 %, 72.3 %]) result in moderately lower 5-year survival prediction estimates than those from RSF feature selection ([78.2 %, 81.5 %]). Here, the higher probabilities are obtained using the variables from RSF feature selection. The presence of features violating the PH assumption with the kNN-imputed data might affect these predictions. These survival predictions are calculated on the test data, thus highly affected by the split and choice of validation set. Applying cross-validation might be a possibly solution to overcome this. These cross-validated results are summarised in Appendix 17.

The most important features with predictive potential are identified from models validated with holdout method. Figure 28 displays the most important features according to these CPH models and how many times each of these features is selected as a statistically significant in predicting CRC patients' survival. The frequency of how many times a specific feature is identified as a statistically significant marker as displayed in Appendix 18. The variables which were chosen at least three (3) times to possess predictive potential are displayed in an order of decreasing importance in Figure 28. The variables with the best predictive abilities are Dukes, followed by neutrophil collagenase (MMP8ngml), mucin 16 (MUC16), kallikrein related peptidase 13 (KLK13), amphiregulin (AREG), secreted protein acidic and rich in cysteine (SPARC), cysteine-rich 61 (CYR61), and age at operation (AgeAtOper).

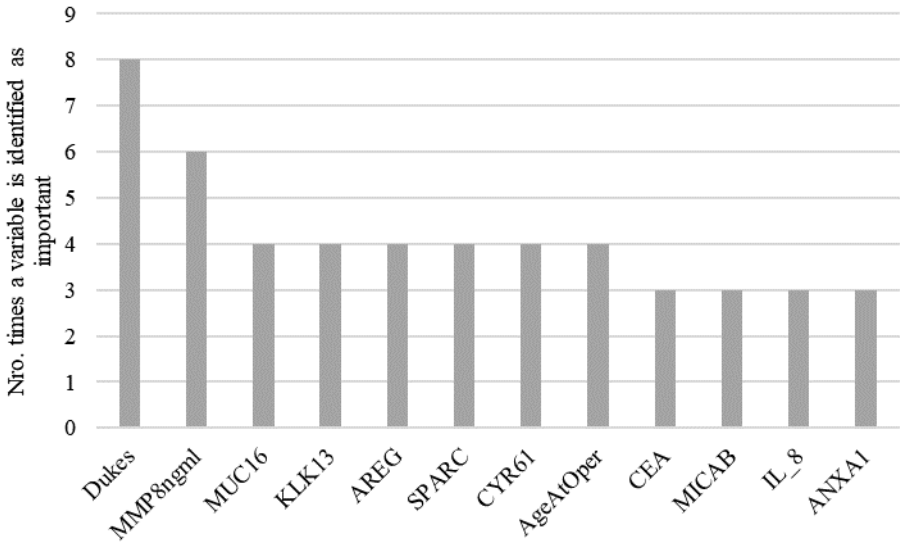


Figure 28. Most statistically significant markers by CPH models.

From fitted CPH models the highest concordance on test data are obtained with the models using kNN-imputed data. The worst concordance values are obtained from the models using listwise deletion data. The survival prediction estimates form all the models offer a decreasing series of probabilities. Thus the motivation for selecting the CPH model with kNN-imputed data with 80/20 split ratio using the variables selected by RSF feature selection for a more detailed review.

There are ten (10) features violating the PH assumption in this specific CPH model. These are Fc receptor like B (FCRLB), cysteine-rich angiogenic inducer 61 (CYR61), mucin 16 (MUC16), human kallikrein 8 (hK8), tumour necrosis factor ligand superfamily member 13 (TNFSF13), human kallikrein 11 (hK11), membrane-targeting domain (CPC1), toll-like receptor 3 (TLR3), vascular endothelial growth factor A (VEGFA), and vimentin (VIM). The output of the model is in Appendix 19. The sensitivity and specificity values are determined to form a type of a confusion matrix to assess the error type of the model.

Table 8. Survivors and censored patients at selected times, and performance metrics.

MUC16	Cases	Survivors	Censored	Se (%)	Sp (%)	PPV (%)	NPV (%)
t =1	18	350	32	78.02	51.27	7.4	97.91
t =3	85	256	59	69.74	55.04	32.53	85.41
t =5	103	216	81	71.32	58.39	41.65	83.02

Mucin 16 (MUC16) is selected as a marker for this since its higher levels are associated with poorer survival (Björkman *et al.*, 2019). Median of the marker value is used as a cutpoint. These values together with the number of survivals, and censored patients for each time interval is shown in Table 8. Cases refer to the patients with having died from CRC. Survivors are the individuals remaining in the data after a given time point. In the column 'Censored' are the number of patients for whom there are no more status information after that time, e.g. died from other causes, or left the study. The timepoints chosen for this are the same than used for survival predictions. The sensitivity and negative predictive values (NPV) of the model are quite sufficient. Specificity of the model remains almost at the same level though the predetermined time intervals. Positive predictive value (PPV) increases as the time of interest increases.

3.3.2 Random survival forests

Random survival forests (RSF) are popular technique for survival analysis for their good performance and embedded feature selection. RSFs are fitted for all three (3) datasets using R's randomForestSRC (version 2.11.0) and survival (3.2-10) packages. R's rfsrc function's inbuild option for imputing missing values is left outside the scope of this thesis. RSF models with

1,000 trees are fitted using bootstrap sampling with replacement. For splitting rule both log-rank splitting and the gradient-based Brier score splitting are applied.

Number of trees and the number of predictor variables used for each node is defined prior the RSF model. To find the optimal values for both of these variables there has been propositions in the literature (Kruppa et al., 2013; Lopes, 2015). For all datasets the optimal values of terminal node size (nodesize) and the number of variables randomly selected as candidates for a splitting a node (hereinafter mtry) are determined by applying a tuning function from R's RF-SRC package to test data. This tuning function utilizes out-of-bag (OOB) error for parameter tuning (Ishwaran and Kogalur, 2021). For comparison the default values of the R's randomForestSRC package are used. This means $q = \sqrt{155} \sim 12$ (mtry, number of variables randomly selected as candidates for splitting a node (R Package Documentation, 2021a)) covariates randomly drawn at each node, and terminal node size set to 15 which is the default value for survival analysis. All the other arguments are kept the same for all datasets to make the results comparable. RSF models with 1,000 trees are fitted using bootstrap sampling with replacement. Cumulative error rate is calculated on every tree, i.e. block size is set to be one (1). This block size affects the determination of the variable importance, and with the value one the variable importance is type Breiman-Cutler VIMP (Ishwaran *et al.*, 2008). Using smaller values with block size usually gives higher accuracy (R Package Documentation, 2021a).

Schmid, Wright and Ziegler (2016) propose the usage of Harrell's C as a splitting criterion for RSF instead of commonly used log-rank statistic. Their study shows that through Harrell's C the accuracy of prediction can be enhanced with data sets experiencing high rate of censoring. However, for this thesis this is left out of scope as well as log-rank score. Log-rank statistic and the gradient-based Brier score splitting are used instead as a split criterion. For the Brier score the 90^{th} percentile of the observed events time is used for the time horizon (R Package Documentation, 2021a). With the log-rank splitting the error rate of the models is slightly lower than with using the Brier score. The results are summarised in Appendix 20 for holdout method, and in Appendix 17 for k-fold CV results.

Performance of RSF has been reported to be depended on the censoring rate. Datasets with low censoring rate seem to achieve better performance compared to those with less deaths, i.e. higher censoring rate. (Ishwaran *et al.*, 2008) With more information about the deaths of the observed patients, the features affecting the overall survival are easier to identify, and thus predict the survival more accurately. The censoring rates are calculated as the number of patients experiencing censoring divided by the total number of patients in the dataset. In this thesis the censoring rates are before MICE imputation ~62 % for the benchmark dataset (with listwise deletion) and ~67 % for both median and kNN imputed sets. After creating more data with MICE imputation the censoring rates are ~65 % for the benchmark data, ~65 % for median imputed, and ~64 % for kNN-imputed data. Although the censoring rates slightly change after artificially enlarging the data, those remain quite high, and thus might affect unfavourably the final performance of the RSF model.

Appendix 20 summarises the RSF models' performance using train/test split ratios 85/15 and 80/20. The error rates for both test and train data using log-rank and gradient-based Brier score as a splitting criterion is presented in performance column. For all three datasets the log-rank used as a splitting criterion seems to obtain higher predictive accuracy. This observation concerns both approaches for selection of terminal node size and number of variables randomly selected as candidates for splitting a node. Using the R's tuning algorithm to determine these values (terminal node size and mtry) results in slightly better performance. Tuned values for terminal node size are quite similar to the default value of 15. However, the tuned values for mtry are significantly larger than the default value of 12. Though the effect of selecting more variables as candidates for splitting a node on the performance of the final model is not significantly better.

The performance on the benchmark data is quite poor. It seems that there is a possibility that models are overfitted with train data and thus result in considerably over concordance values on test data. This might be due to that fact that there is not much initial data left, and the MICE imputation fails to identify the pattern of the underlaying data when it was used to artificially enlarge the dataset. To recap, 84.6 % of the benchmark dataset are imputed values. Thus the successful imputation is critical for obtaining a model with high predictive ability. The highest

performance is obtained using median imputed data and log-rank splitting with RSF structure values determined using a tuning algorithm. High presence of censoring in our CRC data might be one of the reasons behind quite high error rates of the RSF models. These results from hold-out validated RSF models are supported by the findings from the cross-validated RSF models (see Appendix 17).

Survival curves are plotted for first ten (10) patients in all three datasets (see Appendix 21). For this the RSF models build with default values for terminal node size and mtry, and log-rank split rule are used. The 80/20 train/test ratio is applied. These curves in these three (3) scenarios are presented in Figure 41, Figure 42, and Figure 43. With these figures, it ought to be emphasized that these only present a fraction of the overall data. For the median imputed and the kNN-imputed data the curves differ a bit from those of benchmark data. One of the reasons for this is that in the aforementioned two datasets, there are many censored observations, i.e. there are less recorded CRC-related deaths than those alive or having died for other reasons.

The probability estimates for 1-year, 3-year, and 5-year survival are determined for all the models (see Table 9). All of the models successfully predict a decreasing series of survival probability estimates for these timepoints. There is a lot of deviation in the 5-year survival probability predictions obtained using listwise deletion (BM) data ([69.2 %, 75.8 %]). The predictions from RSF with default values for minimum size of terminal node and number of variables randomly selected as candidates for splitting a node (mtry) are as coherent than values gained from tuning function. For both median imputed and kNN-imputed datasets with both default and tuned values for terminal node size and mtry the 5-year survival prediction estimates are a bit lower with log-rank as a splitting rule (median [58.3 %, 62.4 %], kNN [53.0 %, 62.1 %]) than with gradient-based Brier score (median [68.7 %, 71.9 %], kNN [75.3 %, 80.1 %]).

The most important features for predicting the survival of CRC patients with these RSF models applying holdout method for validation are determined using a similar approach than the one taken for feature selection (see chapter 3.2.3.3). To recapitulate, both the minimal depth (md) and variable importance (VIMP) of the include features of the models are determined. Only

those features identify as important by both md and VIMP are recognized as a variable with high predictive potential. The variables which were chosen at least 12 times (50% of the fitted RSF models) to have predictive abilities are demonstrated in an order of decreasing importance in Figure 29. These variables with the best predictive capabilities are carcinoembryonic antigen (CEA), Dukes, hepatocyte growth factor (HGF), calcium binding protein (S100A11), macrophage inflammatory protein-1beta (MIP_1b), interleukin-9 (IL_9), taxilin alpha (TXLNA), secreted protein acidic and rich in cysteine (SPARC), stem cell factor 1 (SCF1), interferon-induced protein-10 (IP_10), and WNT1-inducible-signaling pathway protein 1 (WISP1). The rest of the variables with the most predictive potential that RSF models identify are presented in a table in Appendix 22.

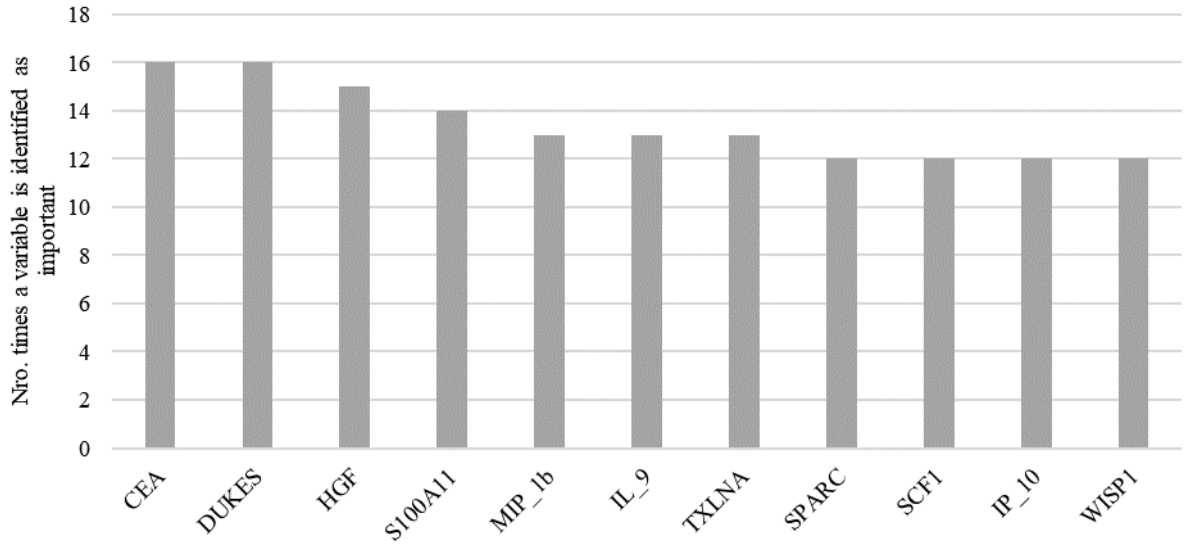


Figure 29. Most statistically significant markers by RSF models.

3.4 Evaluation of the results

In this chapter the CPH and RSF models discussed in chapter 3.3 are compared. For this the Harrell's C (concordance) is applied. The survival predictions for the selected timepoints are discussed. Additionally, the markers identified by these models to possess the most predictive potential are discussed. To conclude the findings concerning the identification of potential biomarkers for predicting the DSS of CRC patients are compared with the literature. It ought to

be emphasized that the models are built using a slightly different set of features. For CPH models features are chosen using two (2) different feature selection techniques, univariate Cox and RSF feature selection. The included variables for CPH using univariate Cox are displayed in Appendix 11, and RSF feature selection in Appendix 12. As for RSF models there is no need for prior feature selection as it has in-build feature selection. These variables are listed in Appendix 8.

The performance of the built models is assessed using Harrell's C (aka concordance index) for both train and test data. This Harrell's C as a measure of predictive accuracy is discussed in more detail chapter 2.5.4. These values are displayed in Appendix 16 for CPH and Appendix 20 for RSF with other metrics of those models. Values for c-index are summarised in Table 9 for holdout validated models and in Appendix 17 for k-fold cv. For both splits as well as 5-fold and 10-fold cv for CPH the highest concordance values are obtained using the kNN-imputed dataset. On the test data kNN with features selected using univariate Cox results in slightly higher c-index than with features by RSF feature selection method when applying holdout method. However, with k-fold cv the situation is vice versa, but the c-index values are relatively similar. This notion applies for median imputed data as well. For listwise deletion data with the univariate Cox the model reduces to univariate (*). The performance on listwise deletion data is poor. Thus could be concluded that the best performance on our CRC data with CPH according to c-index is obtained using either of the two (2) feature selection techniques on the kNN-imputed data, or the RSF feature selection technique on median imputed data. It ought to be noted that for kNN-imputed data more features are selected in the feature selection. This affects the performance of the final models.

Table 9. Summary of the models' c-index values of holdout validated models.

	Imputation	Feature selection / Splitting rule	Concordance	
			Train	Test
CPH train/test split 85/15	Listwise deletion (BM)	Univariate Cox*	0.525	0.702
		RSF	0.639	0.474
	Median	Univariate Cox	0.649	0.516
		RSF	0.780	0.746
	kNN	Univariate Cox	0.817	0.756
		RSF	0.822	0.725
CPH train/test split 80/20	Listwise deletion (BM)	Univariate Cox*	0.511	0.716
		RSF	0.649	0.424
	Median	Univariate Cox	0.642	0.555
		RSF	0.774	0.741
	kNN	Univariate Cox	0.815	0.787
		RSF	0.816	0.769
RSF train/test split 85/15 default	Listwise deletion	Log-rank	0.903	0.606
		Gradient-based Brier	0.853	0.614
	Median imputation	Log-rank	0.915	0.695
		Gradient-based Brier	0.898	0.704
	kNN-imputation	Log-rank	0.924	0.731
		Gradient-based Brier	0.889	0.684
RSF train/test split 85/15 tuned	Listwise deletion	Log-rank	0.918	0.579
		Gradient-based Brier	0.862	0.594
	Median imputation	Log-rank	0.914	0.714
		Gradient-based Brier	0.901	0.727
	kNN-imputation	Log-rank	0.931	0.719
		Gradient-based Brier	0.892	0.695
RSF train/test split 80/20 default	Listwise deletion	Log-rank	0.909	0.559
		Gradient-based Brier	0.885	0.636
	Median imputation	Log-rank	0.912	0.695
		Gradient-based Brier	0.895	0.720
	kNN-imputation	Log-rank	0.923	0.766
		Gradient-based Brier	0.887	0.722
RSF train/test split 80/20 tuned	Listwise deletion	Log-rank	0.928	0.525
		Gradient-based Brier	0.782	0.437
	Median imputation	Log-rank	0.911	0.739
		Gradient-based Brier	0.902	0.748
	kNN-imputation	Log-rank	0.932	0.745
		Gradient-based Brier	0.887	0.699

As for RSF models the highest c-index on test data is with kNN-imputed data with log-rank splitting rule for models with default values for nodeze and mtry. The models with tuned values for terminal node size and mtry the highest c-index values are with median imputed dataset using gradient-based Brier score as a splitting rule according to the holdout validated results. However, with those models the c-index for kNN-imputed data with log-rank splitting is almost the same. With k-fold cv the highest c-index values are with kNN-imputed data and log-rank

as splitting rule. Also the kNN data with log-rank splitting achieves the highest c-index on all the RSF models using both validation approaches. With listwise deletion data the performance measured using the c-index is poor, and it seems that the models might be overfitted to the train data. To conclude with the kNN-imputed data both the CPH and RSF models seem to outperform median imputed and listwise deletion datasets.

The fit can also be evaluated by comparing the survival predictions obtained from the models on the test data. If wanting to compare these figures to the one reported by Finnish Cancer Registry (2021) it ought to be emphasized that our CRC patient cohort consisted solely of patients having gone undergone surgery and thus there are no “worse” patients. This indicates that the survival prediction from our models should be higher than the reported ones. According to the Finnish Cancer Registry (2021) the survival probabilities are 82.8 % for year 1, 70.5 % for year 3, and 66.5 % for year 5. Almost all the models fitted obtain higher estimates for survival than those aforementioned figures. The timepoints of interest selected for the predictions are year 1, year 3, and year 5 counting from the time of the diagnosis. These are summarised in Table 10 for holdout validated models, and for k-fold cv in Appendix 17.

Table 10. Summary of survival probability predictions of holdout validated models.

	Imputation	Feature selection / Splitting rule	Survival probability predictions Year 1 Median	Year 1 Std	Year 3 Median	Year 3 Std	Year 5 Median	Year 5 Std
CPH train/test split 85/15	Listwise deletion (BM)	Univariate Cox*	0.944	*0.004*	0.847	*0.011*	0.759	*0.016*
		RSF	0.945	*0.001*	0.852	*0.006*	0.760	*0.013*
	Median	Univariate Cox	0.955	*0.050*	0.801	*0.139*	0.739	*0.157*
		RSF	0.969	*0.012*	0.841	*0.049*	0.776	*0.062*
	kNN	Univariate Cox	0.971	*0.183*	0.784	*0.330*	0.693	*0.345*
		RSF	0.981	*0.033*	0.851	*0.106*	0.782	*0.121*
CPH train/test split 80/20	Listwise deletion (BM)	Univariate Cox*	0.946	*0.002*	0.850	*0.005*	0.755	*0.007*
		RSF	0.951	*0.001*	0.860	*0.006*	0.762	*0.014*
	Median	Univariate Cox	0.952	*0.020*	0.792	*0.035*	0.728	*0.037*
		RSF	0.966	*0.023*	0.833	*0.062*	0.767	*0.070*
	kNN	Univariate Cox	0.974	*0.215*	0.811	*0.323*	0.732	*0.338*
		RSF	0.983	*0.042*	0.872	*0.105*	0.815	*0.119*
RSF train/test split 85/15 default	Listwise deletion	Log-rank	0.932	*0.045*	0.822	*0.086*	0.714	*0.111*
		Gradient-based Brier	0.942	*0.000*	0.848	*0.000*	0.758	*0.000*
	Median imputation	Log-rank	0.911	*0.072*	0.686	*0.127*	0.616	*0.138*
		Gradient-based Brier	0.943	*0.000*	0.758	*0.002*	0.687	*0.003*
	kNN-imputation	Log-rank	0.929	*0.060*	0.684	*0.129*	0.621	*0.133*
		Gradient-based Brier	0.963	*0.000*	0.803	*0.002*	0.753	*0.003*
RSF train/test split 85/15 tuned	Listwise deletion	Log-rank	0.934	*0.076*	0.802	*0.139*	0.694	*0.172*
		Gradient-based Brier	0.944	*0.008*	0.848	*0.019*	0.757	*0.026*
	Median imputation	Log-rank	0.916	*0.116*	0.649	*0.197*	0.583	*0.212*
		Gradient-based Brier	0.951	*0.014*	0.774	*0.054*	0.698	*0.070*
	kNN-imputation	Log-rank	0.917	*0.009*	0.632	*0.027*	0.585	*0.029*
		Gradient-based Brier	0.970	*0.011*	0.823	*0.042*	0.779	*0.055*
RSF train/test split 80/20 default	Listwise deletion	Log-rank	0.937	*0.001*	0.824	*0.006*	0.718	*0.010*
		Gradient-based Brier	0.948	*0.000*	0.849	*0.000*	0.753	*0.000*
	Median imputation	Log-rank	0.917	*0.006*	0.694	*0.017*	0.624	*0.020*
		Gradient-based Brier	0.937	*0.000*	0.763	*0.002*	0.693	*0.003*
	kNN-imputation	Log-rank	0.922	*0.005*	0.679	*0.018*	0.610	*0.019*
		Gradient-based Brier	0.959	*0.000*	0.807	*0.002*	0.759	*0.002*
RSF train/test split 80/20 tuned	Listwise deletion	Log-rank	0.938	*0.073*	0.821	*0.136*	0.692	*0.171*
		Gradient-based Brier	0.948	*0.007*	0.854	*0.014*	0.758	*0.021*
	Median imputation	Log-rank	0.922	*0.028*	0.662	*0.055*	0.597	*0.060*
		Gradient-based Brier	0.949	*0.000*	0.794	*0.004*	0.719	*0.006*
	kNN-imputation	Log-rank	0.906	*0.015*	0.613	*0.034*	0.530	*0.035*
		Gradient-based Brier	0.970	*0.011*	0.843	*0.047*	0.801	*0.064*

One of the objectives of this thesis is to identify markers with potentially predictive potential for forming the survival estimates for CRC patients. The significant features are identified for CPH and RSF models using holdout validation. The results are shown in Appendix 18 for CPH and Appendix 22 for RSF. The most important features selected by CPH models are demonstrated in Figure 28, and from RSF models in Figure 29. With RSF more potentially predictive feature were identified than with CPH. From those features both CPH and RSF identify 26 variables in common to possess predictive potential. These are displayed in Table 11. The 'Frequency' column displays the number of times each feature is selected as important by CPH

and RSF models. For CPH the maximum value for this is 12 and for RSF 24, i.e. the number of those specific models. The table is arranged in a decreasing order of importance based on the CPH frequencies. The same information is presented visually as a histogram in Figure 30.

Table 11. The features with predictive potential identified by CPH and RSF models.

	Frequency	
Variable	**CPH**	**RSF**
Dukes	8	16
MMP8ngml	6	11
MUC16	4	9
KLK13	4	10
AREG	4	9
SPARC	4	12
CYR61	4	9
AgeAtOper	4	11
CEA	3	16
MICAB	3	3
IL_8	3	8
ANXA1	3	7
diff	2	5
ADAMTS15	2	11
MMP8TIMP1molrati	2	8
SYND1	2	11
CPC1	2	2
TLR3	2	4
CD160	2	6
ALB1	2	4
VIM	2	4
IL_2Ra	2	6
PGFalpha	1	8
IL_6	1	11
MIP_1b	1	13
IGF1R	1	8

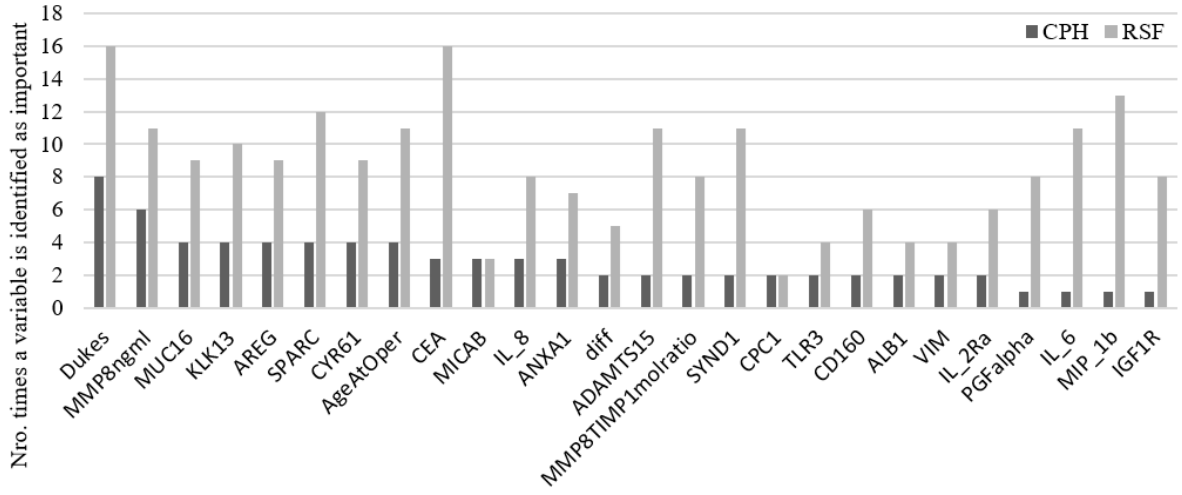


Figure 30. Histogram of frequencies of important features identified by CPH and RSF models.

Features identified as important at least by one third of both CPH, i.e four (4), and RSF, i.e. eight (8), models are discussed in more detail. There are total of eight (8) features satisfying this condition. These are Dukes staging (Dukes), neutrophil collagenase also referred as matrix metalloproteinase-8 (MMP8ngml), mucin 16 (MUC16), kallikrein 13 (KLK13), amphiregulin (AREG), secreted protein acidic and rich in cysteine (SPARC), cysteine-rich 61 (CYR61), and age at operation (AgeAtOper). Carcinoembryonic antigen (CEA) which is identified by 16 RSF models to be important, is identified only by three (3) CPH models. Similar notions can be made with ADAMTS15, syndecan 1 (SYND1), interleukin-6 (IL_6), and macrophage inflammatory protein-1beta (MIP_1b).

To assess these findings, those are compared with the previous literature. Dukes as a staging scheme for CRC (Labianca *et al.*, 2010) justifies its usage for forming predictions. Patient's age affects the decisions made whilst planning the therapeutical approaches to be taken (Labianca *et al.*, 2010). Bjarnadottir et al. (2018) identified old age and advanced stage of cancer to affect early mortality of CRC patients. Dukes staging and patients' age differ form the rest of the markers discussed here, since those are not biomarkers.

Amongst CRC patients poor survival, distant metastasis and systemic inflammation has been shown to correlate with high levels of MMP-8 (Väyrynen *et al.*, 2011; Sirniö *et al.*, 2018). Mucin 16 (MUC16, also referred as CA-125) has been identified as a prognostic marker for CRC (Gao *et al.*, 2018). High levels of mucin 16 is shown to be associated with worse prognosis for CRC patients. (Björkman *et al.*, 2019) CRC patients with low levels of kallikrein 13 are associated with poor survival (Talieri *et al.*, 2009; Björkman *et al.*, 2019). Expression levels of AREG are used as a biomarker for CRC (Lee *et al.*, 2016), and its high levels are associated with poor survival (Björkman *et al.*, 2019). Secreted protein acidic and rich in cysteine (SPARC) is shown to be related to stage of CRC and prognosis, and higher serum levels indicate better prognosis (Chew *et al.*, 2011; Carriere *et al.*, 2021). There is an evidence of presence of cysteine-rich 61 (CYR61) in CRC patients functioning as a part of carcinogenesis (Jeong *et al.*, 2014). Higher levels of CYR61 correlate with poor prognosis (Xie *et al.*, 2019). Carcinoembryonic antigen (CEA) is commonly used biomarker in CRC both for preoperative staging and postoperative screening (Labianca *et al.*, 2010). CEA has been shown to possess sensitivity as a marker (Gao *et al.*, 2018). ADAMTS15 has shown some evidence of having suppressor properties for CRC, and its levels negatively correlates with the histopathological differentiation of the tumours (Viloria *et al.*, 2009; Przemyslaw *et al.*, 2013). Syndecan 1 (SYND1) can be utilised as a prognostic marker for CRC (Wei *et al.*, 2015; Mitselou *et al.*, 2016). High levels of SYND1 are associated with poor survival (Björkman *et al.*, 2019). Pro-inflammatory cytokine interleukin-6 (IL-6) has been associated with inflammation-associated CRC. Higher levels of IL-6 are related to poorer prognosis for patients (Waldner, Foersch and Neurath, 2012). Macrophage inflammatory protein-1β (MIP-1β / MIP_1b) shown to have chemotactic abilities and thus could be a potential target for gene therapy (Luo *et al.*, 2004). Thus can be concluded that the previous literature supports the findings made in this thesis.

4 Conclusions

4.1 Summary of the study

The objective of this thesis was to identify potential biomarkers for predicting survival of CRC patients. Previous studies around techniques for survival analysis, imputation, and feature selection were consulted to form a rough understanding of the current stage of the literature around this phenomenon. Also research trends in oncology and some medical concepts, focusing on CRC, are discussed together with statistics concerning the cancer in Finland. The missing values are imputed using listwise deletion as a benchmark, imputation by median, and k-nearest neighbour imputation (kNN-imputation). In addition to this, the sample size of our CRC data is enhanced using MICE which is an imputation technique. The dimensionality of CRC data is reduced using correlation analysis, univariate Cox, and random survival forests (RSF) for feature selection. As for survival analysis Cox proportional hazards, and random survival forests are fitted. DeepSurv models are also fitted and included in the Appendix 23. However, further parameter optimization or feature selection is not performed for these artificial neural network (ANN) approaches. Thus these results are only preliminary and further research is needed to provide any conclusive remarks about the suitability of ANN solutions for predicting survival of CRC patients. The models are validated using holdout method with following split ratios; 80/20 and 85/15, additionally semi-stratified k-fold cross-validation with k = 5 and k = 10. These splits are performed in a manner which preserves the distribution of the status variable same than in the original data for each fold.

4.2 Research questions and findings

The objective of this thesis was to provide insight to the main research question:

How the survival of CRC patients could be modelled and predicted?

To support this main research question, following sub-research questions were defined:

What is the state-of-art literature of survival analysis studies in oncology?

How to handle missing values whilst ensuring to preserving the characteristics of the data?

How to perform feature selection to select the features with the most predictive potential?

Which biomarkers are associated with high prognostic value for predicting survival of CRC patients?

4.2.1 What is the state-of-art literature of survival analysis studies in oncology?

The first sub-research question was set to obtain understanding about the state-of-art of survival analysis studies in the field of oncology. This is provided an answer in chapter 2.1 and summarised also in Appendix 1. The summarised findings about the techniques used for conducting survival analysis and imputation from the studies included into the literature review are displayed in Table 12. The most applied survival analysis models were tree-based methods (e.g. classification trees, decision trees), logistic regression, Cox models, support vector machines (SVM), Bayes approaches (e.g. naïve Bayes, Bayesian network), and artificial neural networks (ANN). For many studies, there was no mention about handling possible missing values, i.e. listwise deletion can be assumed. Discarding the observations with any missing values (listwise deletion / complete-case analysis) was the most common approach taken in the literature based on the conducted literature review. Other methods were multiple imputation (MI), multiple imputation by chained equations (MICE), and simple imputation where the missing values were imputing using either mean or the most common value. Thus can be identified a need for survival analysis studies with imputation of missing values.

Table 12. Summary of methods used for survival analysis and imputation in literature.

Authors	Tree-based methods	Logistic regression	Cox	SVM	Bayes	ANN	RSF	KNN	Others	Listwise deletion	MI	MICE	Simple imputation	No mention
Bjarnadottir et al. (2018)	X	X								X				
Murtojärvi et al. (2020)			X							X				
Van Belle et al. (2011)			X	X										X
Kleinlein & Riaño (2019)	X	X			X									X
Reijnen et al. (2020)					X						X			
Delen, Walker & Kadam (2005)	X	X				X				X				
Tseng et al. (2019)		X		X	X		X							X
Zupan et al. (2000)	X		X		X									X
Jerez-Aragonés et al. (2003)	X		X			X								X
Ryu, Chandrasekaran & Jacob (2004)									Isotonic prediction					X
Anand et al. (1999)	X		X			X		X	GA	X				
Burke et al. (19979						X				X				
van Stiphout et al. (2010)		X		X									X	
Snow et al. (2001)		X				X								X
Li & Razzaghi (2019)		X				X	X		AdaBoost & imbalanced classification			X		

4.2.2 How to handle missing values whilst ensuring to preserving the characteristics of the data?

The second sub-research question was set to gain understanding about multiple ways to handle missing data using imputation techniques. From the conducted literature review can be seen that there is a demand for studies comparing different imputation techniques. In our CRC data

over 33 % of the values were missing. This concerns 109 variables and 241 observations. Thus the missingness is a remarkable characteristic and need to be addressed. In addition to this multiple imputation by chained equations (MICE) was applied to artificially increase sample size of imputed datasets. Distributions prior and post this enlargement were compared and found to be almost exact.

Different imputation techniques were considered separately. In total three (3) were applied. First approach was the listwise deletion. This means that all observations with any missing values are removed. After this only 76 observations were remained. Thus making the sample size extremely small which might not be sufficient to build a survival analysis model. When increasing the sample size using MICE for this listwise deletion dataset the fact that only a small portion of the original data was used as base. Thus these values used for MICE might generalize the final sample. Since there exist slight differences in patients in separate datasets which merged to form a bigger dataset, this merging creates missing values. Thus only a small fraction of all the observations in both the immunopanel data (~34 %) and CRP/TATI/MMP dataset (~32 %) are included after listwise deletion. For the Olink panel this percentage is slightly higher, amounting about a half (~52 %) of patients. As an additional remark this listwise deletion results in a data which contained only patients with colon cancer, and none with rectal cancer. This is an interesting notion, since the original data has more rectal cancer patients (52.5 %) than those patients with tumours located in colon (47.5 %).

The second technique was to perform imputation using median which is a type of simple imputation. Here missing values are imputed using the median value of recorded values of that specific variable. This approach of simple imputation alters the distributions of the variables by enforcing their existing trends. Finally, the third method was to use k-nearest neighbour (kNN) imputation. The neighbourhood size was set to be ten (10). This is method imputes the missing values based on those values of similar patients. This kNN-imputation is more sophisticated approach than imputation using median or simple listwise deletion.

Listwise deletion disregards many observations, and thus removes characteristics of the data making it poorer. For our small CRC patient cohort, the listwise deletion seems to be not suggestable. The whole enlarged data with 500 patients is based on just 77 patients. This makes the fitted model to be based on extremely homogenous cohort which is not the case in reality. This removal of observations with any missing values might be usable approach for larger datasets where the exclusion would only affect a small fraction of the original data.

Imputation by median generalises the patient cohort as a type of simple imputation. The value distributions become centred on the cost of more extreme values. The cohort becomes more homogenous, like with listwise deletion. The relations between different features are not acknowledged by this method. This phenomenon is emphasized when there are many features with missing values which was the case with our data. Especially with serum and tissue patient data the connections between different features are crucial. This linkage is affected negatively when imputing by median. This imputation approach might be more suitable to cases with only a small fraction of the data missing.

With kNN-imputation the data is more diverse compared to the other two (2) approaches. The missing values are imputed in a more sophisticated manner by taking into account the patients with similar clinicopathological and demographic profiles in the process of filling in the missing values. Though this dataset could be made complete, and the values of observations are realistic making the final fitted model possibly more accurate and detailed. Through experimentations with altering the neighbourhood size k, the imputation results could be optimized further. The choice of kNN-imputation is also supported by the fact that it is not computationally demanding.

When examining the performance of these imputed datasets in feature selection similar notions can be derived. With correlation analysis more features are selected for median imputed and kNN-imputed data than for listwise deletion data. With univariate Cox with FDR-corrected p-values for feature selection, the listwise deletion data reduces to univariate. For median imputed data this method chooses only a small fraction of the original set of features. A much larger set

of features is selected for kNN-imputed data. The same performance is evidenced with the RSF for feature selection. Fewest number of features is selected for listwise deletion data, then few more for median imputed data, and finally most features are chosen for kNN-imputed data. Still both the univariate Cox and RSF feature selection succeed to reduce the number of features remarkably for all datasets.

The c-index values of the fitted models are summarised in Table 9 for holdout validated models and in Appendix 17 for k-fold cv. The highest concordance values on both train and test data are obtained with kNN-imputed data. Except for RSF with tuned values the models fitted using median imputed data slightly outperform kNN-imputed ones on test data according to the hold-out validated results. The k-fold cv indicates that the kNN-imputed data with log-rank splitting rule results in a highest concordance value. However, for all models the highest c-index values on train are with kNN-imputed data. The lowest values are with listwise deletion data. The models' 1-year, 3-year and 5-year survival predictions are summarised in Table 10 for holdout validated models and in Appendix 17 for k-fold cv. All the models successfully produce a decreasing series of survival prediction estimates for the timepoints of interest. To conclude, the best fit for our CRC data overall seems to be kNN-imputation.

4.2.3 How to perform feature selection to select the features with the most predictive potential?

Similar to the second sub-research question the third is about the preprocessing of the data prior modelling. Selecting a subset of features aids to resolve potential issues associated with high-dimensionality. Despite the fact that our CRC data is quite small, and survival analysis methods with build-in feature selection are preferable, separate feature selection techniques are tested. Three (3) different approaches were used. These techniques are correlation analysis, univariate Cox proportional hazards (CPH), and random survival forests (RSF) feature selection with both minimal depth (md) and variable importance (VIMP). These methods were applied to all three (3) datasets formed after imputation.

First the correlation analysis was applied to reduce number of predictors. The threshold was set to 0.8. Correlation analysis failed to select substantially smaller datasets. For listwise deletion data the method removed 48 of the features. For both median imputed data and the kNN-imputed set 31 features were excluded. Lower correlation bound of 0.7 was also tested, and it resulted in a slightly smaller sets of features. However, that threshold is quite low and might lead to exclusion of relevant features. Thus can be concluded that the correlation analysis approach as is does not suit well for feature selection for our CRC data. The further investigation using this approach was not continued.

The second approach was to apply univariate CPH to select the features most related to predicting survival. Only the features showing statistically significant FDR-corrected p-values (< 0.10) in the univariate CPH are included in the further analysis. This showed better performance than correlation analysis. The number of features was substantially decreased. These numbers are following; for listwise deletion data one (1), median imputed data 12, and kNN-imputed data 49. However, only one (1) feature was in common in all the datasets. This is mucin 16 (MUC16) which is also known as CA-125 (carbohydrate antigen 125). Since the listwise deletion data has such few observations the results from this univariate analysis can be foreseen. For this data the model reduces to univariate. Additionally, kNN-imputation is much more sophisticated approach than imputation using median. Thus explaining the difference in chosen features for these datasets. All the features chosen for median imputed set are also selected for kNN-imputed set. However, this approach does not consider possible relationships between the features. A single covariate independently can have a totally different affect to the patient's survival than when all the covariates affect simultaneously.

The third approach was to use random survival forests with minimal depth (md) and variable importance (VIMP). For VIMP the procedure is repeated ten (10) times since the calculations contain random elements. For RSF models 1,000 trees with log-rank as a splitting rule is applied and bootstrapping without replacement as a bootstrapping protocol. Additionally parameter value for minimum terminal node size is 15, and the number of variables randomly selected as candidates for splitting a node is a square root of number of features. RSF feature selection selects more features than univariate CPH, except for kNN-imputed data, but manages to reduce

the amount of feature remarkably. For median and kNN-imputed datasets with minimal depth approach the number of chosen features is much smaller than with VIMP. For the further survival analysis, the features identified as important with both md and VIMP are chosen. Thus, the final model sizes are 29 for benchmark data, 38 for median imputed data, and 47 for KNN-imputed data. There are nine (9) features chosen by all these approaches. These are mucin 16 (MUC16), Dukes' stage, syndecan 1 (SYND1), Ly6/PLAUR domain containing 3 (LYPD3 or C4.4A), carcinoembryonic antigen cell adhesion molecule 1 (CEACAM1), kallikrein related peptidase 13 (KLK13), placenta growth factor alpha (PGFalpha), membrane-targeting domain (CPC1), and TIMP metallopeptidase inhibitor 1 (TIMP1ngml).

Here the features chosen by different imputation techniques were slightly different. Thus, can be concluded that the selection of imputation technique affects the final model, and so forth the reliability of the results. Interestingly, with univariate CPH the number of features selected is lower for both listwise deletion data and median imputed data, whilst for kNN-imputed data the situation is vice versa. However, for kNN-imputed data both of these feature selection techniques select approximately the same number of features. The features selected by univariate CPH and RSF feature selection were compared for each of the datasets individually to check for similarities. For listwise deletion data there is one (1) same feature, for median 10, and for kNN-imputed 33. These are displayed in Table 13. There can be observed that mucin 16 (MUC16) is shared features for all the datasets with these two (2) feature selection approaches. Between median and kNN-imputed datasets there are nine (9) selected features in common as can be seen from the Table 13.

Table 13. Common features selected by Univariate CPH (p < 0.05) and RSF feature selection.

	Features
Listwise deletion	MUC16
Median imputed	CEA, MUC16, IL_8, AREG, TIMP1ngml, IL_6, MMP8ngml, SYND1, CEACAM5, PGFalpha
kNN-imputed	AREG, MUC16, SYND1, CEA, KLK13, CEACAM5, VIM, IGF1R, CYR61, S100A11, FURIN, WISP1, PGFalpha, VEGFA, SCF1, IL_8, TIMP1ngml, aSNT, FCRLB, MICAB, IL_6, hk8, CD48, SPARC, ESM1, CXCL13, hk11, TNFSF13, GPNMB, ANXA1, DUKES, AgeAtOper, LY9

As these features selection techniques only affect CPH models, the performance and accuracy of survival predictions are assessed for those. The highest concordance is achieved with features from RSF feature selection for both median imputed data and kNN-imputed data. Also CPH models with the univariate Cox with kNN-imputed data are able to obtain high c-index values. Univariate Cox with median imputed data performs poorly as well as listwise deletion as a whole. Thus could be concluded that the choice of imputation affects the performance of feature selection techniques. The univariate Cox and RSF feature selection both seem to be able to choose significant features for further analysis and reduce the number of covariates quite well on kNN-imputed data. As for median imputed data the RSF feature selection method is more suitable.

4.2.4 Which biomarkers are associated with high prognostic value for predicting survival of CRC patients?

Markers with high prognostic values for predicting the survival of CRC patients were identified using CPH and RSF feature selection with md and VIMP to find potential therapeutic applications. CPH models required prior feature selection, which was performed using correlation analysis, univariate Cox, and RSF feature selection. The results from those are summarised in 4.2.3. Finally, the feature subset selected using univariate Cox and RSF feature selection were

used in the survival analysis with CPH. Thus the feature spaces for CPH and RSF were slightly different, as with RSF the feature selection is embedded.

From CPH models 27 features with potentially predictive abilities were identified. These are listed in Appendix 18. The variables which were chosen at least by three (3) models (maximum number is 12) to possess predictive potential are Dukes, followed by neutrophil collagenase (MMP8ngml), mucin 16 (MUC16), kallikrein related peptidase 13 (KLK13), amphiregulin (AREG), secreted protein acidic and rich in cysteine (SPARC), cysteine-rich 61 (CYR61), and age at operation (AgeAtOper). However, the PH violations of those CPH models was not addressed are thus might affect the choice of these significant features. Model specific PH violations are presented in Appendix 15.

From 24 fitted RSF models 144 markers with predictive potential were identified. These are presented in Appendix 22. The variables which were chosen at least by 12 models are carcinoembryonic antigen (CEA), Dukes, hepatocyte growth factor (HGF), calcium binding protein (S100A11), macrophage inflammatory protein-1beta (MIP_1b), interleukin-9 (IL_9), (TXLNA), secreted protein acidic and rich in cysteine (SPARC), stem cell factor 1 (SCF1), interferon-induced protein-10 (IP_10), and WNT1-inducible-signaling pathway protein 1 (WISP1).

Comparing the significant markers selected by CPH and RSF, 26 were found to have in common. These are displayed in Table 11 and in Figure 30. There are total of eight (8) features which were identified as important at least by one third of both CPH, i.e four (4), and RSF, i.e. eight (8), models. These are Dukes staging (Dukes), neutrophil collagenase also referred as matrix metalloproteinase-8 (MMP8ngml), mucin 16 (MUC16), kallikrein 13 (KLK13), amphiregulin (AREG), secreted protein acidic and rich in cysteine (SPARC), cysteine-rich 61 (CYR61), and age at operation (AgeAtOper). Carcinoembryonic antigen (CEA) which is identified by 16 RSF models to be important, is identified only by three (3) CPH models. Similar notions can be made with ADAMTS15, syndecan 1 (SYND1), interleukin-6 (IL_6),

and macrophage inflammatory protein-1beta (MIP_1b). The validity of these findings was assessed by comparing those with the previous studies. Literature supported these findings.

4.2.5 Concluded findings in the main research question

The aim of this study was to obtain an understanding about the factors affecting the prediction of survival of CRC patients. In addition to this different imputation and feature selection techniques were compared. For handling missing values three (3) approaches were taken. First the rows with any missing values were removed. This listwise deletion approach was popular based on the conducted literature review. However, as our data was quite small. This approach reduces the characteristics of the data population even further and increases its homogeneity. The similar issue is with median imputation. The general characteristics of the patients are enhanced. With kNN-imputation the individual features of the patients are taken into account better. Thus, the imputed data is expected to reflect the actual patient population characteristics more accurately. As the sample size is smaller more sophisticated imputation method is suggested. Here kNN-imputation.

For feature selection the number of selected features was largest with the kNN-imputed data, and the lowest with listwise deletion data. Correlation analysis failed to reduce the dimensionality of the data, and thus the further analysis with that technique was left outside the scope of this thesis. With univariate Cox with FDR-corrected p-values the listwise deletion data reduced to univariate. For median imputed data 12 features, and for kNN-imputed data 49 features were identified as statistically significant. RSF feature selection was performed by selecting those features identified as important by both md and VIMP approaches. This resulted in choosing 29 features for listwise deletion data, 38 for median imputed data, and 47 for kNN-imputed data. Thus could be concluded that univariate Cox and RSF feature selection were able to locate the important feature best from the kNN-imputed data. For median imputed data the RSF feature selection approach seemed to suit the best.

The performance of the fitted models measured using the c-index (concordance) is summarised in Table 9 for holdout validated results and in Appendix 17 for k-fold cv. The results from k-fold cv support the findings from holdout validated models. The highest c-index values for both CPH and RSF models were obtained using the kNN-imputed data (c-index on test data 0.787 CPH and 0.766 RSF with holdout validation, 0.717 CPH and 0.756 RSF using 5-fold cv, 0.714 CPH and 0.761 RSF using 10-fold cv), whilst the worst performance was with the listwise deletion data (c-index on test data 0.424 CPH and 0.437 RSF with holdout validation, 0.507 CPH and 0.545 RSF using 5-fold cv, 0.535 CPH and 0.539 RSF using 10-fold cv). The CPH models with features selected using RSF feature selection performed quite well, achieving the c-index approximately 0.75 on test data. With kNN-imputed data the performance was similar with features from both univariate Cox and RSF feature selection. For RSF models median imputed data performed better with gradient-based Brier score as a splitting rule. Log-rank splitting was more well suited for kNN-imputed data. There were no clear differences between whether the terminal node size and Number of variables randomly selected as candidates for a splitting a node (mtry) values for RSF models were selected as default or by using a tuning function.

As for the survival predictions from the models for 1-year, 3-year and 5-year survival probabilities all of the models successfully provided a decreasing series of survival estimates for these timepoints of interest. These estimates are coherent with the figures reported by Finnish Cancer Registry (2021) when emphasizing the fact that our data consists solely of CRC patients having undergone surgery, i.e. the survival prediction should be slightly higher from our models. These predictions are summarised in Table 10 for holdout validated models and in Appendix 17 for k-fold cv.

From all fitted models, markers potentially associated with predicting DSS of CRC patients were identified. Top features of predictive potential which both CPH and RSF models pointed out are Dukes staging (Dukes), neutrophil collagenase also referred as matrix metalloproteinase-8 (MMP8ngml), mucin 16 (MUC16), kallikrein 13 (KLK13), amphiregulin (AREG), secreted protein acidic and rich in cysteine (SPARC), cysteine-rich 61 (CYR61), and age at operation (AgeAtOper). Carcinoembryonic antigen (CEA), ADAMTS15, syndecan 1

(SYND1), interleukin-6 (IL_6), and macrophage inflammatory protein-1beta (MIP_1b) were identofoed as important by many RSF models but by only a few of the fitted CPH models. Some of these chosen variables are inflammation-related, thus reinforcing the notion about connection between inflammation and cancer. Examples of those are kallikrein-13 (KLK13) (Lizama *et al.*, 2015), neutrophil collagenase (MMP-8) (Sirniö et al., 2018) and interleukin-6 (IL-6) (Kumari *et al.*, 2016). Literature was found to support these findings for predicting DSS of CRC patients (see chapter 3.4).

Comparing the significant markers selected by CPH and RSF, 26 were found to have in common. These are displayed in Table 11 and in Figure 30. There are total of eight (8) features which were identified as important at least by one third of both CPH, i.e four (4), and RSF, i.e. eight (8), models. These are Dukes staging (Dukes), neutrophil collagenase also referred as matrix metalloproteinase-8 (MMP8ngml), mucin 16 (MUC16), kallikrein 13 (KLK13), amphiregulin (AREG), secreted protein acidic and rich in cysteine (SPARC), cysteine-rich 61 (CYR61), and age at operation (AgeAtOper). Carcinoembryonic antigen (CEA) which is identified by 16 RSF models to be important, is identified only by three (3) CPH models. Similar notions can be made with ADAMTS15, syndecan 1 (SYND1), interleukin-6 (IL_6), and macrophage inflammatory protein-1beta (MIP_1b). The validity of these findings was assessed by comparing those with the previous studies. Literature supported these findings.

Thus, could be concluded that based on the concordance values of both 5-fold and 10-fold cross-validation approaches, the RSF with log-rank as splitting rule using the kNN-imputed data would be suggested on our CRC data. Additionally, the survival prediction estimates obtained from that aforementioned model are giving a decreasing series of probabilities and are thus reasonable. Also, RSF models with gradient-based splitting rule fitted using kNN-imputed and median imputed datasets achieved acceptable performance. The differences between the performance measured using the c-index of the top performing models are minor. RSF models fitted on our CRC data outperform the CPH models. However, for some of these CPH models there were features violating the PH assumption. These violations were not addressed in this thesis. Potential stratification of problematic variables, alteration of the functional form of the

regression variables, or addition of time interaction terms could affect the performance of CPH models.

4.3 Reliability and validity

This chapter summarises the lessons learned from this thesis. Moreover some critique towards this thesis is presented. Also the limitations in reliability and uncertainty embedded in the results of this thesis should be reflected. Since the nature of this study is a comparative one, the methods applied are only in a general level. Further optimization and selecting the best possible approaches based on the used dataset, the performance of the model could be better. Selecting the survival analysis technique based on the underlying distribution of the time to event of interest variable of the available data is preferable. If there is an observable specific distribution of the event times parametric survival analysis methods ought to be preferred over nonparametric ones (Wang, Li and Reddy, 2017). The lack of medical knowledge affects the reliability of the results. However, the fact that this is not done as part of a medical degree rather than an engineering one.

Other survival analysis techniques could have been tested. SVMs are not included, and deeper analysis of the usability of neural networks could have been further investigated. CPH with LASSO and embedded feature selection, or Bayesian networks (Reijnen et al., 2020) and other Bayesian survival analysis techniques (Brard *et al.*, 2017) could provide interesting results. However, these are not tested as a part of this thesis. Additionally, detailed model optimization was not performed.

Another important notion concerns validation of the results. One major drawback concerning holdout validation. A portion of the data is left unused from model building. This section reserved for validation could include some important characteristics, and thus result in potential loss of crucial information. The split strongly affects the performance measures of the model on the validation set. Validating the models using k-fold cross-validation could reduce the bias

relating to this. Thus, a semi-stratified k-fold CV is tested as well. Since the CRC cohort is quite small, smaller values of k (i.e. 5) is tested as well as k = 10. The stratification of the folds is done using the status variable. For the future repeated stratified k-fold cross-validation could be applied to ensure the validity of the final model. Repetition of the k-fold cv procedure helps to reduce bias even further. Further statistical comparison of the models' performance based on the concordance values by determining their confidence intervals is left outside the scope of this thesis. This deeper examination concerning the statistical differences between the fitted models could provide interesting insight and could be performed in the future studies.

The available data was one of the major limitations in this thesis. Generalizations made for handling it was then the most fundamental delimitation. The dataset used is rather small and limited, only 318 patients. All the patients are from Finland, thus making it geologically narrow. There were a lot of missing values. One of the main reasons for this was that the final dataset was formed by merging three (3) separate datasets, and in those datasets not all the patients were represented. Thus the importance of accurate imputation is emphasized. Due to high-dimensionality of the data it is crucial to be able to select the features with the most predictive potential. This possible issue was addressed by testing multiple approaches for both imputation and feature selection. As a potential preprocessing step the dataset could have been trimmed so that the coefficients are removed if the measure of their 75^{th} percentile value is less than 100 (Beer *et al.*, 2002). By doing so, the irrelevant features with low values will be excluded from further analysis. Thus reducing the possibility of unwanted results in feature selection. To avoid possibly overfitting the imputation method to the data, it could be tested by introducing more new observations from the same distribution. This is not conducted here since the original sample size is quite small.

To keep the focus of this thesis, time-dependent variables were not included to the models. This alteration could be done in the future, and thus incorporate e.g. recurrence to the model. Also competing risk analysis could be carried out. This could provide insight into probability of competing risks in more realistic setting. An example of these competing risks could be death from some other causes, e.g. heart failure or traffic accident. To conclude, further research is needed to form definite conclusions.

4.4 Further studies

Here some suggestions for the further research in the field of oncologic survival analysis are introduced. First the other techniques for survival analysis techniques are discussed. This is followed by the remarks about imputation and feature selection. Finally, some perspectives about future oncology research and business viewpoints are briefly presented. Through dynamic prediction models with the possibility to update the model using new data (Reijnen *et al.*, 2020), and integrate medical experts' knowledge into process of building ML survival methods (Delen, Walker and Kadam, 2005) supports the development of more accurate prognosis prediction models to aid medical professionals' decision-making processes (Tseng *et al.*, 2019).

To obtain more accurate information about the survival of patients' the possibility of developing a two-stage survival analysis model could be further investigated. The first stage of this type of model predicts whether a patient will survive past a specific year x, and the objective in the second stage is to predict the accurate survival time. Also research about the combination of existing ensemble models to observe if those actions could result in improved predictive power of the model could be an interesting path to investigate. A new classification scheme for CRC could be formulated to predict therapy efficacy. Introducing competing risks analysis into survival analysis could provide more insight into predicting the survival of patients.

The possibility to apply Markov models for survival analysis could be further research. Currently those models are applied for stochastic processes, e.g. economic evaluation modelling of chronic diseases (Briggs and Sculpher, 2012). Although cancer is not considered chronic illness, Markov models could provide interesting insight into modelling prognosis of cancer patients. Other techniques worth looking into for survival analysis are neuro-evolution solutions which use evolutionary algorithms (EA) with deep neural networks (DNN) (Briggs and Sculpher, 2012), SVMs and rough sets (Delen, Walker and Kadam, 2005), and fuzzy logic systems (Seker *et al.*, 2003). Some fuzzy approaches are discussed in section 2.4.5.1.

As for imputing the missing values similarity-based classifiers could be utilised. Additionally imputation could be approached as a classification problem where all complete observations are applied to form a prediction model. Then the model is validated with cross-validation to make the prediction for missing variable values. Further the effect of different imputation techniques for the performance of the model could be studied. Additionally the techniques for artificially increasing the sample size would provide interesting results. These are discussed in section 3.2.1.

Development of more sophisticated feature selection and model optimization approaches could result in obtaining more accurate model to predict prognosis of CRC patients. An approach for selecting the most predictive variables for analysis could be done by performing a cluster analysis on the original data. Then select one feature or maximum of few features from each significant cluster and fit a model using only those features. This approach should reduce the number of features remarkably. Possibility to utilize genetic algorithms (GA) for feature selection is suggested in the literature (Ryu, Chandrasekaran and Jacob, 2004; Kalderstam *et al.*, 2013; Liu *et al.*, 2013; Mansoori, Suman and Mishra, 2014; Aalaei *et al.*, 2016). Another feature selection approach for future studies in iterative Bayesian model averaging (BMA) which uses rank-ordered list of features and applies those to traditional BMA. Englebert et al. (2017) applied CPH for breast cancer microarray data to create these rank-ordered list of genes in descending order of their log-likelihood, and let the iterative BMA process though their user-specified list of top-ranked genes. Finally, applying a backwards stepwise CPH-model built on Akaike information criterion (AIC) as a data-driven CPH based method for feature selection could be investigated more in the future.

One of the most prevalent trends in the future of medicine is the step towards more personalised healthcare, e.g. genome mapping. In determination of survival outcome, the patient's individual gene profile together with the survival time and tumour histopathology represents this shift towards tailored solutions (Beer *et al.*, 2002; Bair and Tibshirani, 2004). Early detection is crucial for better prognoses. Deeper insight into predictive biomarkers of CRC and other cancers help to develop real-time prediction models as tool for medical decision-support (Bjarnadottir *et al.*, 2018). Connection between inflammation and cancer has already studied and will be

continued even further. Possibility to utilise calprotectin as a biomarker for CRC could be investigated further since it would provide a non-invasive approach for diagnosis. Calprotectin is used to detect inflammation in the digestive system. Role of inflammation in cancer supports the use of calprotectin also in detection of cancer (Tibble *et al.*, 2001).

When approaching the problem of predicting the survival of patients with cancer or other medical condition from business viewpoint estimation of the total costs from accurate, early prognosis to all parties, e.g. patient, hospitals' resources, society though changes in employment status, versus incorrect or diagnosis is obtained too late is important. This could help to estimate the future demand of hospital resources (Bjarnadottir *et al.*, 2018), and thus allocate those more efficiently, and being able to provide the best possible care for as many patients as possible.

5 References

Aalaei, S., Shahraki, H., Rowhanimanesh, A. and Eslami, S. (2016) 'Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets', *Iranian Journal of Basic Medical Sciences*, 19(5), pp. 476–482.

Aalen, O. (1978) 'Nonparametric Inference for a Family of Counting Processes', *Annals of Statistics*, 6(4), pp. 701–726. doi: 10.1214/aos/1176344247.

Abdar, M. and Makarenkov, V. (2019) 'CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer', *Measurement: Journal of the International Measurement Confederation*. Elsevier Ltd, 146, pp. 557–570. doi: 10.1016/j.measurement.2019.05.022.

Abdel-Zaher, A. M. and Eldeib, A. M. (2016) 'Breast cancer classification using deep belief networks', *Expert Systems with Applications*. Elsevier Ltd, 46, pp. 139–144. doi: 10.1016/j.eswa.2015.10.015.

AIJCC (2021) *Cancer Staging System*. Available at: https://www.facs.org/quality-programs/cancer/ajcc/cancer-staging (Accessed: 30 December 2021).

Akaike, H. (1974) 'A New Look at the Statistical Model Identification', *IEEE Transactions on Automatic Control*, 19(6), pp. 716–723. doi: 10.1109/TAC.1974.1100705.

Alizadeh, A. A., Ross, D. T., Perou, C. M. and Van De Rijn, M. (2001) 'Towards a novel classification of human malignancies based on gene expression patterns', *Journal of Pathology*, pp. 41–52. doi: 10.1002/path.889.

Allignol, A. and Latouche, A. (2021) *CRAN Task View: Survival Analysis, CRAN R-project*. Available at: https://cran.r-project.org/web/views/Survival.html (Accessed: 17 May 2021).

Allin, K. H. and Nordestgaard, B. G. (2011) 'Elevated C-reactive protein in the diagnosis, prognosis, and cause of cancer.', *Critical reviews in clinical laboratory sciences*. England, 48(4), pp. 155–170. doi: 10.3109/10408363.2011.599831.

American Cancer Society (2015) *Lymph Nodes and Cancer*. doi: 10.1159/000423367.

Amin, M. B., Edge, S., Greene, F., Byrd, D. R., Brookland, R. K., Washington, M. K., Gershenwald, J. E., Compton, C. C., Hess, K. R., Sullivan, D. C., Jessup, J. M., Brierley, J. D.,

Gaspar, L. E., Schilsky, R. L., Balch, C. M., Winchester, D. P., Asare, E. A., Madera, M., Gress, D. M. and Meyer, L. R. (eds) (2017) *AJCC Cancel Staging Manual*. 8th edn. Springer International Publishing.

Anand, S. S., Smith, A. E., Hamilton, P. W., Anand, J. S., Hughes, J. G. and Bartels, P. H. (1999) 'An evaluation of intelligent prognostic systems for colorectal cancer', *Artificial Intelligence in Medicine*, 15(2), pp. 193–214. doi: 10.1016/S0933-3657(98)00052-9.

Andersen, P. K., Geskus, R. B., De witte, T. and Putter, H. (2012) 'Competing risks in epidemiology: possibilities and pitfalls', *International Journal of Epidemiology*. Oxford University Press, 41(3), p. 861. doi: 10.1093/IJE/DYR213.

Antolini, L., Boracchi, P. and Biganzoli, E. (2005) 'A time-dependent discrimination index for survival data', *Statistics in Medicine*, 24(24), pp. 3927–3944. doi: 10.1002/sim.2427.

Astler, V. B. and Coller, F. A. (1954) 'The prognostic significance of direct extension of carcinoma of the colon and rectum', *Annals of surgery*, 139(6), pp. 846–851. doi: 10.1097/00000658-195406000-00015.

Azur, M. J., Stuart, E. A., Frangakis, C. and Leaf, P. J. (2011) 'Multiple imputation by chained equations: what is it and how does it work?', *International Journal of Methods in Psychiatric Research*. Wiley-Blackwell, 20(1), p. 40. doi: 10.1002/MPR.329.

Baesens, B., Van Gestel, T., Stepanova, M., Van Den Poel, D. and Vanthienen, J. (2005) *Neural Network Survival Analysis for Personal Loan Data, Source: The Journal of the Operational Research Society*.

Bailey, T. L. and Elkan, C. (1993) 'Estimating the Accuracy of Learned Concepts', in *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*. AAA1 Press, pp. 895–900.

Bair, E. and Tibshirani, R. (2004) 'Semi-supervised methods to predict patient survival from gene expression data', *PLoS Biology*, 2(4), pp. 511–522. doi: 10.1371/journal.pbio.0020108.

Balkwill, F., Charles, K. A. and Mantovani, A. (2005) 'Smoldering and polarized inflammation in the initiation and promotion of malignant disease', *Cancer Cell*. Cell Press, 7(3), pp. 211–217. doi: 10.1016/j.ccr.2005.02.013.

Barouchos, Nikolaos, Papazafiropoulou, A., Iacovidou, N., Vrachnis, N., Barouchos, Nektarios, Armeniakou, E., Dionyssopoulou, V., Mathioudakis, A. G., Christopoulou, E., Koltsida, S. and Bassiakou, E. (2015) 'Comparison of tumor markers and inflammatory biomarkers in chronic obstructive pulmonary disease (COPD) exacerbations', *Scandinavian Journal of Clinical and Laboratory Investigation*. Informa Healthcare, 75(2), pp. 126–132. doi: 10.3109/00365513.2014.992944.

Barsainya, A., Sairam, A. and Patil, A. P. (2018) 'Analysis and Prediction of Survival after Colorectal Chemotherapy using Machine Learning Models', in *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*. IEEE, pp. 862–865. doi: 10.1109/ICACCI.2018.8554832.

Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., Hinoue, T., Laird, P. W., Curtis, C., Shen, H., Weisenberger, D. J., Schultz, N., Shen, R., Weinhold, N., Kelsen, D. P., Bowlby, R., Chu, A., Kasaian, K., Mungall, A. J., Robertson, A. G., Sipahimalani, P., Cherniack, A. D., Getz, G., Liu, Y., Noble, M. S., Pedamallu, C., Sougnez, C., Taylor-Weiner, A., Akbani, R., Lee, J. S., Liu, W., Mills, G. B., Yang, D., Zhang, W., Pantazi, A., Parfenov, M., Gulley, M., Piazuelo, M. B., Schneider, B. G., Kim, Jihun, Boussioutas, A., Sheth, M., Demchok, J. A., Rabkin, C. S., Willis, J. E., Ng, S., Garman, K., Beer, D. G., Pennathur, A., Raphael, B. J., Wu, H. T., Odze, R., Kim, H. K., Bowen, J., Leraas, K. M., Lichtenberg, T. M., Weaver, S., McLellan, M., Wiznerowicz, M., Sakai, R., Lawrence, M. S., Cibulskis, K., Lichtenstein, L., Fisher, S., Gabriel, S. B., Lander, E. S., Ding, L., Niu, B., Ally, A., Balasundaram, M., Birol, I., Brooks, D., Butterfield, Y. S. N., Carlsen, R., Chu, J., Chuah, E., Chun, H. J. E., Clarke, A., Dhalla, N., Guin, R., Holt, R. A., Jones, S. J. M., Lee, D., Li, H. A., Lim, E., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, K. L., Nip, K. M., Schein, J. E., Tam, A., Thiessen, N., Beroukhim, R., Carter, S. L., Cho, J., DiCara, D., Frazer, S., Gehlenborg, N., Heiman, D. I., Jung, J., Kim, Jaegil, Lin, P., Meyerson, M., Ojesina, A. I., Pedamallu, C. S., Saksena, G., Schumacher, S. E., Stojanov, P., Tabak, B., Voet, D., Rosenberg, M., Zack, T. I., Zhang, H., Zou, L., Protopopov, A., Santoso, N., Lee, S., Zhang, J., Mahadeshwar, H. S., Tang, J., Ren, X., Seth, S., Yang, Lixing, Xu, A. W., Song, X., Xi, R., Bristow, C. A., Hadjipanayis, A., Seidman, J., Chin, L., Park, P. J., Kucherlapati, R., Ling, S., Rao, A., Weinstein, J. N., Kim, S. B., Lu, Y., Bootwalla, M. S., Lai, P. H., Triche, T., Van Den Berg, D. J., Baylin, S. B., Herman, J. G., Murray, B. A., Askoy, B. A., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Lee, W., Ramirez, R., Sander, C., Senbabaoglu, Y., Sinha,

R., Sumer, S. O., Sun, Y., Iype, L., Kramer, R. W., Kreisberg, R., Rovira, H., Tasman, N., Haussler, D., Stuart, J. M., Verhaak, R. G. W., Leiserson, M. D. M., Taylor, B. S., Black, A. D., Carney, J. A., Gastier-Foster, J. M., Helsel, C., McAllister, C., Ramirez, N. C., Tabler, T. R., Wise, L., Zmuda, E., Penny, R., Crain, D., Gardner, J., Lau, K., Curely, E., Mallery, D., Morris, S., Paulauskis, J., Shelton, T., Shelton, C., Sherman, M., Benz, C., Lee, J. H., Fedosenko, K., Manikhas, G., Potapova, O., Voronina, O., Belyaev, D., Dolzhansky, O., Rathmell, W. K., Brzezinski, J., Ibbs, M., Korski, K., Kycler, W., Łaźniak, R., Leporowska, E., Mackiewicz, A., Murawa, D., Murawa, P., Spychała, A., Suchorska, W. M., Tatka, H., Teresiak, M., Abdel-Misih, R., Bennett, J., Brown, J., Iacocca, M., Rabeno, B., Kwon, S. Y., Kemkes, A., Curley, E., Alexopoulou, I., Engel, J., Bartlett, J., Albert, M., Park, D. Y., Dhir, R., Luketich, J., Landreneau, R., Janjigian, Y. Y., Cho, E., Ladanyi, M., Tang, L., McCall, S. J., Park, Y. S., Cheong, J. H., Ajani, J., Camargo, M. C., Alonso, S., Ayala, B., Jensen, M. A., Pihl, T., Raman, R., Walton, J., Wan, Y., Eley, G., Shaw, K. R. M., Tarnuzzer, R., Wang, Z., Yang, Liming, Zenklusen, J. C., Davidsen, T., Hutter, C. M., Sofia, H. J., Burton, R., Chudamani, S. and Liu, J. (2014) ‘Comprehensive molecular characterization of gastric adenocarcinoma’, *Nature*. Nature Publishing Group, 513(7517), pp. 202–209. doi: 10.1038/nature13480.

Bates, D. M., Lindstrom, M. J., Wahba, G. and Yandell, B. S. (1986) *Technical Report NO. 775 (rev.) GCVPACK - Routines for generalized cross validation*. Madison, Wisconsin.

Beer, D. G., Kardia, S. L. R., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M. G., Iannettoni, M. D., Orringer, M. B. and Hanash, S. (2002) ‘Gene-expression profiles predict survival of patients with lung adenocarcinoma’, *Nature Medicine*, 8(8), pp. 816–824. doi: 10.1038/nm733.

Van Belle, V., Neven, P., Harvey, V., Van Huffel, S., Suykens, J. A. K. and Boyd, S. (2013) ‘Risk group detection and survival function estimation for interval coded survival methods’, *Neurocomputing*, 112, pp. 200–210. doi: 10.1016/j.neucom.2012.12.049.

Van Belle, V., Pelckmans, K., van Huffel, S. and Suykens, J. A. K. (2011) ‘Improved performance on high-dimensional survival data by application of survival-SVM’, *Bioinformatics*, 27(1), pp. 87–94. doi: 10.1093/bioinformatics/btq617.

Van Belle, Vanya, Pelckmans, K., Van Huffel, S. and Suykens, J. A. K. (2011) 'Support vector methods for survival analysis: A comparison between ranking and regression approaches', *Artificial Intelligence in Medicine*, 53(2), pp. 107–118. doi: 10.1016/j.artmed.2011.06.006.

Van Belle, V., Pelckmans, K., Suykens, J. A. K. and Van Huffel, S. (2007) 'Support Vector Machine for Survival Analysis', in *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, pp. 1–8.

Bengio, Y., Courville, A. and Vincent, P. (2013) 'Representation learning: A review and new perspectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), pp. 1798–1828. doi: 10.1109/TPAMI.2013.50.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp. 289–300.

Bennett, S. (1983a) 'Analysis of survival data by the proportional odds model', *Statistics in Medicine*. John Wiley & Sons, Ltd, 2(2), pp. 273–277. doi: 10.1002/SIM.4780020223.

Bennett, S. (1983b) 'Log-Logistic Regression Models for Survival Data', *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. [Wiley, Royal Statistical Society], 32(2), pp. 165–171. doi: 10.2307/2347295.

Bergstra, J. and Bengio, Y. (2012) 'Random search for hyper-parameter optimization', *Journal of Machine Learning Research*, 13, pp. 281–305.

Berlth, F., Bollschweiler, E., Drebber, U., Hoelscher, A. H. and Moenig, S. (2014) 'Pathohistological classification systems in gastric cancer: Diagnostic relevance and prognostic value', *World Journal of Gastroenterology*, 20(19), pp. 5679–5684. doi: 10.3748/wjg.v20.i19.5679.

Beyersmann, J. and Scheike, T. H. (2016) 'Classical Regression Models for Competing Risks', in Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (eds) *Handbook of Survival Analysis*. Chapman and Hall/CRC, pp. 90–104. doi: 10.1201/b16248-11.

Bharath, K., Kurtek, S., Rao, A. and Baladandayuthapani, V. (2018) 'Radiologic image-based statistical shape analysis of brain tumours', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5), pp. 1357–1378.

Bishop, C. M. and Tipping, M. E. (2003) ‘Bayesian Regression and Classification’, in Suykens, J. A. K., Horvath, I., Basu, S., Micchelli, C., and Null, J. V. (eds) *Advances in Learning THeory: Methods, Modesl and Applications*. IOS Press (NATO Science Series, III: Computer and Systems Sciences), pp. 267–285.

Bjarnadottir, M., Anderson, D., Zia, L. and Rhoads, K. (2018) ‘Predicting Colorectal Cancer Mortality: Models to Facilitate Patient-Physician Conversations and Inform Operational Decision Making’, *Production and Operations Management*. Wiley-Blackwell, 27(12), pp. 2162–2183. doi: 10.1111/poms.12896.

Björkman, K., Mustonen, H., Kaprio, T., Haglund, C. and Böckelman, C. (2019) ‘Mucin 16 and kallikrein 13 as potential prognostic factors in colon cancer: Results of an oncological 92-multiplex immunoassay’, *Tumor Biology*. SAGE Publications Ltd, 41(7), pp. 1–10. doi: 10.1177/1010428319860728.

Black, J. E., Terry, A. L. and Lizotte, D. J. (2020) ‘Development and evaluation of an osteoarthritis risk model for integration into primary care health information technology’, *International Journal of Medical Informatics*. Elsevier Ireland Ltd, 141. doi: 10.1016/j.ijmedinf.2020.104160.

Boracchi, P. and Biganzoli, E. (2002) ‘Radial basis function neural networks for the analysis of survival data’, *Metron*, 60(1–2), pp. 191–210.

Borgan, Ø. (2005) ‘Nelson-Aalen Estimator’, in *Encyclopedia of Biostatistics*. 1st edn. John Wiley & Sons. doi: 10.1002/0470011815.

Boser, B. E., Vapnik, V. N. and Guyon, I. M. (1992) ‘Training Algorithm Margin for Optimal Classifiers’, *Perception*, pp. 144–152.

Bottaci, L., Drew, P. J., Hartley, J. E., Hadfield, M. B., Farouk, R., Lee, P. W. R., MacIntyre, I. M. C., Duthie, G. S. and Monson, J. R. T. (1997) ‘Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions’, *Lancet*. Lancet Publishing Group, 350(9076), pp. 469–472. doi: 10.1016/S0140-6736(96)11196-X.

Brard, C., Le Teuff, G., Le Deley, M. C. and Hampson, L. V. (2017) ‘Bayesian survival analysis in clinical trials: What methods are used in practice?’, *Clinical Trials*, 14(1), pp. 78–87. doi: 10.1177/1740774516673362.

Breiman, L. (2001) ‘Random Forests’, *Machine Learning*, 45(October), pp. 5–32. doi: https://doi.org/10.1023/A:1010933404324.

Breiman, L. and Spector, P. (1992) ‘Submodel Selection and Evaluation in Regression. The X-Random Case’, *International Statistical Review / Revue Internationale de Statistique*, 60(3), p. 291. doi: 10.2307/1403680.

Brier, G. W. (1950) ‘Verification of forecasts expressed in terms of probability’, *Monthly Weather Review*, 78(1), pp. 1–3.

Briggs, A. and Sculpher, M. (2012) ‘An Introduction to Markov Modelling for Economic Evaluation’, *PharmacoEconomics 1998 13:4*. Springer, 13(4), pp. 397–409. doi: 10.2165/00019053-199813040-00003.

Bromberg, J. and Wang, T. C. (2009) ‘Inflammation and Cancer: IL-6 and STAT3 Complete the Link’, *Cancer Cell*, pp. 79–80. doi: 10.1016/j.ccr.2009.01.009.

Brown, B. W., Hollander, M. and Korwar, R. M. (1973) ‘Nonparametric tests of independence for censored data, with applications to heart transplant studies’, in Proschan, F. and Serfling, R. J. (eds) *Reliability and biometry*. Philadelphia: Society for Industrial and Applied Mathematics, pp. 327–354.

Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell Jr, F. E., Marks, J. R., Winchester, D. P. and Bostwick, D. G. (1997) ‘Artificial neural networks improve the accuracy of cancer survival prediction - PubMed’, *Cancer*, 79(4), pp. 857–862. doi: 10.1002/(sici)1097-0142(19970215)79:4<857::aid-cncr24>3.0.co;2-y.

van Buuren, S. (2014) ‘Handbook of Missing Data’, in Fitzmaurice, G. M., Kenward, M. G., Molenberghs, G., Tsiatis, A. A., and Verbeke, G. (eds) *Handbook of Missing Data*.

van Buuren, S. (2018) *Flexible Imputation of Missing Data, Second Edition*. 2nd editio, *Flexible Imputation of Missing Data*. 2nd editio. Second edition. | Boca Raton, Florida : CRC Press, [2019] |: Chapman and Hall/CRC. doi: 10.1201/9780429492259.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) ‘mice: Multivariate imputation by chained equations in R’, *Journal of Statistical Software*, 45(3), pp. 1–67. doi: 10.18637/jss.v045.i03.

Cai, T., Huang, J. and Tian, L. (2009) *Regularized Estimation for the Accelerated Failure Time*

Model.

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. V. F., Iglesias, M., Céspedes, M. V., Sevillano, M., Nadal, C., Jung, P., Zhang, X. H. F., Byrom, D., Riera, A., Rossell, D., Mangues, R., Massagué, J., Sancho, E. and Batlle, E. (2012) 'Dependency of Colorectal Cancer on a TGF-β-Driven Program in Stromal Cells for Metastasis Initiation', *Cancer Cell*, 22(5), pp. 571–584. doi: 10.1016/j.ccr.2012.08.013.

Cancer.net (2019) *Stomach Cancer: Stages | Cancer.Net*, *Cancer.Net*. Available at: https://www.cancer.net/cancer-types/stomach-cancer/stages (Accessed: 2 August 2020).

Cancer Research UK (2017) *Bowel cancer statistics | Cancer Research UK*, *Bowel cancer statistics*. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero (Accessed: 31 July 2020).

Cancer Research UK (2018) *Dukes' staging system | Bowel cancer*. Available at: https://www.cancerresearchuk.org/about-cancer/bowel-cancer/stages-types-and-grades/dukes-staging (Accessed: 30 December 2021).

Carriere, P., Calvo, N., Novoa Díaz, M. B., Lopez-Moncada, F., Herrera, A., Torres, M. J., Alonso, E., Gandini, N. A., Gigola, G., Contreras, H. R. and Gentili, C. (2021) 'Role of SPARC in the epithelial-mesenchymal transition induced by PTHrP in human colon cancer cells', *Molecular and Cellular Endocrinology*, 530(October 2020). doi: 10.1016/j.mce.2021.111253.

Carroll, O. U., Morris, T. P. and Keogh, R. H. (2020) 'How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review', *BMC Medical Research Methodology*. BMC Medical Research Methodology, 20(1), pp. 1–15. doi: 10.1186/s12874-020-01018-7.

Casella, G. and George, E. I. (1992) 'Explaining the gibbs sampler', *The American Statistician*, 46(3), pp. 167–174. doi: 10.1080/00031305.1992.10475878.

Chatterjee, S. B., Hou, J., Ratnam Bandaru, V. V., Pezhouh, M. K., Syed Rifat Mannan, A. A. and Sharma, R. (2019) 'Lactosylceramide synthase β-1,4-GalT-V: A novel target for the diagnosis and therapy of human colorectal cancer', *Biochemical and Biophysical Research Communications*. Elsevier B.V., 508(2), pp. 380–386. doi: 10.1016/j.bbrc.2018.11.149.

Chen, X., Gole, J., Gore, A., He, Q., Lu, M., Min, J., Yuan, Z., Yang, X., Jiang, Y., Zhang, T.,

Suo, C., Li, X., Cheng, L., Zhang, Z., Niu, H., Li, Z., Xie, Z., Shi, H., Zhang, X., Fan, M., Wang, X., Yang, Y., Dang, J., McConnell, C., Zhang, J., Wang, J., Yu, S., Ye, W., Gao, Y., Zhang, K., Liu, R. and Jin, L. (2020) 'Non-invasive early detection of cancer four years before conventional diagnosis using a blood test', *Nature Communications*, 11(1), p. 3475. doi: 10.1038/s41467-020-17316-z.

Chew, A., Salama, P., Robbshaw, A., Klopcic, B., Zeps, N., Platell, C. and Lawrance, I. C. (2011) 'SPARC, FOXP3, CD8 and CD45 Correlation with DiseaseRecurrence and Long-Term Disease-Free Survival inColorectal Cancer', *PLoS ONE*, 6(7), pp. 1–13. doi: 10.1371/journal.pone.0022047.

Chikha, S. Ben and Marzouki, K. (2009) *Making standard SOM invariant to the initial conditions, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-642-02478-8_26.

Ching, T., Zhu, X. and Garmire, L. X. (2018) 'Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data', *PLoS Computational Biology*. Public Library of Science, 14(4). doi: 10.1371/journal.pcbi.1006076.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y. and Van Calster, B. (2019) 'A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models', *Journal of Clinical Epidemiology*. Elsevier Inc, 110, pp. 12–22. doi: 10.1016/j.jclinepi.2019.02.004.

Churilov, L., Bagirov, A., Schwartz, D. and Smith, K. (2005) *Data Mining with Combined Use of Optimization Techniques and Self-Organizing Maps for Improving Risk Grouping Rules: Application to Prostate Cancer Patients*, *Journal of Management Information Systems*.

Collins English Dictionary (2021) *Procto- definition and meaning*. Available at: https://www.collinsdictionary.com/dictionary/english/procto (Accessed: 4 April 2021).

Colores (2022) *Suolistosyövän ennuste, Tietoa suolistosyövästä*. Available at: https://www.colores.fi/suolistosyovan-ennuste/ (Accessed: 16 January 2022).

Congqin-Yi, Zhou, R. and Hu, K. (2017) 'Fuzzy Support Vector Machine for breast cancer gene classification', in *2017 2nd International Conference on Image, Vision and Computing, ICIVC 2017*. IEEE, pp. 676–679. doi: 10.1109/ICIVC.2017.7984641.

Cox, D. R. (1972) ‘Regression Models and Life-Tables’, *Journal of the Royal Statistical Society. Series B (Methodological)*. [Royal Statistical Society, Wiley], 34(2), pp. 187–220.

Cox, D. R. (1975) ‘Partial Likelihood’, *Biometrika*, 62(2), pp. 269–279.

Crohn’s & Colitis Foundation (2021) *Proctocolectomy and Colectomy*. Available at: https://www.crohnscolitisfoundation.org/what-is-crohns-disease/treatment/surgery/proctocolectomy-colectomy (Accessed: 4 April 2021).

CTCA (2018) *Types of Colorectal Cancer: Common, Rare and More Varieties*, *Cancer Treatment Centers of America*. Available at: https://www.cancercenter.com/cancer-types/colorectal-cancer/types (Accessed: 31 July 2020).

Cutler, S. J. and Ederer, F. (1958) ‘Maximum utilization of the life table method in analyzing survival’, *Journal of Chronic Diseases*. Elsevier, 8(6), pp. 699–712. doi: 10.1016/0021-9681(58)90126-7.

Czogała, E. and Łęski, J. (2000) ‘Neuro-fuzzy systems’, in *Fuzzy and Neuro-Fuzzy Intelligent Systems. Studies in Fuzziness and Soft Computing*. 47th edn. Heidelberg: Physica, pp. 141–162. doi: 10.1007/978-3-7908-1853-6_6.

Decock, J., Hendrickx, W., Thirkettle, S., Gutiérrez-Fernández, A., Robinson, S. D. and Edwards, D. R. (2015) ‘Pleiotropic functions of the tumor- and metastasis-suppressing matrix metalloproteinase-8 in mammary cancer in MMTV-PyMT transgenic mice’, *Breast Cancer Research*. BioMed Central Ltd., 17(1). doi: 10.1186/s13058-015-0545-8.

Delen, D., Oztekin, A. and Kong, Z. (2010) ‘A machine learning-based approach to prognostic analysis of thoracic transplantations’, *Artificial Intelligence in Medicine*, 49(1), pp. 33–42. doi: 10.1016/j.artmed.2010.01.002.

Delen, D., Walker, G. and Kadam, A. (2005) ‘Predicting breast cancer survivability: A comparison of three data mining methods’, *Artificial Intelligence in Medicine*, 34(2), pp. 113–127. doi: 10.1016/j.artmed.2004.07.002.

Duan, W., Zhang, R., Zhao, Y., Shen, S., Wei, Y., Chen, F. and Christiani, D. C. (2018) ‘Bayesian variable selection for parametric survival model with applications to cancer omics data’, *Human Genomics*. BioMed Central Ltd., 12(1), pp. 1–15. doi: 10.1186/s40246-018-0179-x.

Dukes, C. E. (1932) ‘The classification of cancer of the rectum’, *The Journal of Pathology and Bacteriology*. John Wiley & Sons, Ltd, 35(3), pp. 323–332. doi: 10.1002/path.1700350303.

Duodecim (2019) *Kolorektaalisyövän kansalliset hoitosuositukset*. Available at: https://www.terveysportti.fi/apps/ltk/article/hsu00007#s9 (Accessed: 5 April 2021).

Duodecim (2021) *Sairaudet ja hoito: Laboratoriotutkimusten tulkinta: CRP (P-CRP), Duodecim Terveyskirjasto*. Available at: https://www.terveyskirjasto.fi/snk03052 (Accessed: 6 April 2021).

Efron, B. (1979) ‘Bootstrap Methods: Another Look at the Jackknife’, *Source: The Annals of Statistics*, 7(1), pp. 1–26.

Efron, B. (1983) ‘Estimating the error rate of a prediction rule: Improvement on cross-validation’, *Journal of the American Statistical Association*, 78(382), pp. 316–331. doi: 10.1080/01621459.1983.10477973.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*. New York: Chapman & Hall, Inc. doi: 10.1007/978-1-4899-4541-9.

Efron, B. and Tibshirani, R. (1997) ‘Improvements on cross-validation: The .632+ bootstrap method’, *Journal of the American Statistical Association*, 92(438), pp. 548–560. doi: 10.1080/01621459.1997.10474007.

Englebert, C., Quinn, T. and Bichindaritz, I. (2017) ‘Feature Selection for Survival Analysis in Bioinformatics’, in *International Joint Conference on Artificial Intelligence (IJCAI)*.

Ershoff, B. D., Lee, C. K., Wray, C. L., Agopian, V. G., Urban, G., Baldi, P. and Cannesson, M. (2020) ‘Training and Validation of Deep Neural Networks for the Prediction of 90-Day Post-Liver Transplant Mortality Using UNOS Registry Data’, *Transplantation Proceedings*. Elsevier USA, 52(1), pp. 246–258. doi: 10.1016/j.transproceed.2019.10.019.

Evers, L. and Messow, C. M. (2008) ‘Sparse kernel methods for high-dimensional survival data’, *Bioinformatics*, 24(14), pp. 1632–1638. doi: 10.1093/bioinformatics/btn253.

Faraggi, D. and Simon, R. (1995) ‘A neural network model for survival data’, *Statistics in Medicine*. John Wiley & Sons, Ltd, 14(1), pp. 73–82. doi: 10.1002/sim.4780140108.

Faria, F. A., Perre, P., Zucchi, R. A., Jorge, L. R., Lewinsohn, T. M., Rocha, A. and Torres, R.

D. S. (2014) ‘Automatic identification of fruit flies (Diptera: Tephritidae)’, *Journal of Visual Communication and Image Representation*, 25(7), pp. 1516–1527. doi: 10.1016/j.jvcir.2014.06.014.

Fawcett, T. (2006) ‘An introduction to ROC analysis’, *Pattern Recognition Letters*, 27(8), pp. 861–874. doi: 10.1016/j.patrec.2005.10.010.

Felder, M., Kapur, A., Gonzalez-Bosquet, J., Horibata, S., Heintz, J., Albrecht, R., Fass, L., Kaur, J., Hu, K., Shojaei, H., Whelan, R. J. and Patankar, M. S. (2014) ‘MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress’, *Molecular cancer*. Mol Cancer, 13(1). doi: 10.1186/1476-4598-13-129.

Feskanich, D., Ma, J., Fuchs, C. S., Kirkner, G. J., Hankinson, S. E., Hollis, B. W. and Giovannucci, E. L. (2004) ‘Plasma Vitamin D Metabolites and Risk of Colorectal Cancer in Women’, *Cancer Epidemiology and Prevention Biomarkers*, 13(9), pp. 1502–1508.

Fidler, I. J. (2003) ‘The pathogenesis of cancer metastasis: the “seed and soil” hypothesis revisited’, *Nature Reviews Cancer*, 3(6), pp. 453–458. doi: 10.1038/nrc1098.

Freund, Y. and Schapire, R. E. (1997) ‘A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting’, *Journal of Computer and System Sciences*, 55(1), pp. 119–139. doi: 10.1006/jcss.1997.1504.

Gao, Y., Wang, J., Zhou, Y., Sheng, S., Qian, S. Y. and Huo, X. (2018) ‘Evaluation of Serum CEA, CA19-9, CA72-4, CA125 and Ferritin as Diagnostic Markers and Factors of Clinical Parameters for Colorectal Cancer OPEN’, *Scientific Reports*, 8, p. 2732. doi: 10.1038/s41598-018-21048-y.

Garland, C. F. and Garland, F. C. (1980) ‘Do Sunlight and Vitamin D Reduce the Likelihood of Colon Cancer?’, *International Journal of Epidemiology*, 9(3), pp. 227–231. doi: 10.1093/ije/9.3.227.

Garland, C. F., Garland, F. C., Gorham, E. D., Lipkin, M., Newmark, H., Mohr, S. B. and Holick, M. F. (2006) ‘The Role of Vitamin D in Cancer Prevention’, *American Journal of Public Health*, 96(2). doi: 10.2105/AJPH.

George, E. I. and McCulloch, R. E. (1993) ‘Variable selection via Gibbs sampling’, *Journal of the American Statistical Association*, 88(423), pp. 881–889. doi:

10.1080/01621459.1993.10476353.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., Mitchell, T. J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vázquez-García, I., Haase, K., Jerman, L., Sengupta, S., Macintyre, G., Malikic, S., Donmez, N., Livitz, D. G., Cmero, M., Demeulemeester, J., Schumacher, S., Fan, Y., Yao, X., Lee, J., Schlesner, M., Boutros, P. C., Bowtell, D. D., Zhu, H., Getz, G., Imielinski, M., Beroukhim, R., Sahinalp, S. C., Ji, Y., Peifer, M., Markowetz, F., Mustonen, V., Yuan, K., Wang, W., Morris, Q. D., Yu, K., Adams, D. J., Campbell, P. J., Cao, S., Christie, E. L., Cun, Y., Dawson, K. J., Drews, R. M., Eils, R., Fittall, M., Garsed, D. W., Ha, G., Lee-Six, H., Martincorena, I., Oesper, L., Peto, M., Raphael, B. J., Salcedo, A., Shi, R., Shin, S. J., Spiro, O., Stein, L. D., Vembu, S., Wheeler, D. A., Yang, T. P., Spellman, P. T., Wedge, D. C. and Van Loo, P. (2020) 'The evolutionary history of 2,658 cancers', *Nature*. Nature Research, 578(7793), pp. 122–128. doi: 10.1038/s41586-019-1907-7.

Giunchiglia, E., Nemchenko, A. and van der Schaar, M. (2018) 'RNN-SURV: A deep recurrent model for survival analysis', in *27th International Conference on Artificial Neural Networks*, pp. 23–32. doi: 10.1007/978-3-030-01424-7_3.

Goldstein, B. A., Navar, A. M. and Carter, R. E. (2017) 'Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges', *European Heart Journal*, 38(23), pp. 1805–1814. doi: 10.1093/eurheartj/ehw302.

Golmah, V. (2014) 'A case study of applying SOM in market segmentation of automobile insurance customers', *International Journal of Database Theory and Application*, 7(1), pp. 25–36. doi: 10.14257/ijdta.2014.7.1.03.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999) 'Assessment and comparison of prognostic classification schemes for survival data', *Statistics in Medicine*, 18(17–18), pp. 2529–2545. doi: https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5.

Greenwood, M. (1938) 'THE FIRST LIFE TABLE', *Notes and Records of the Royal Society of London*, 1(2), pp. 70–72.

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster,

S., Fessler, E., De Sousa .E Melo, F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G. C., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J. W., Hanahan, D., Tabernero, J., Bernards, R., Friend, S. H., Laurent-Puig, P., Medema, J. P., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L. and Tejpar, S. (2015) 'The consensus molecular subtypes of colorectal cancer', *Nature Medicine*. Nature Publishing Group, 21(11), pp. 1350–1356. doi: 10.1038/nm.3967.

Hanafizadeh, P. and Mirzazadeh, M. (2011) 'Visualizing market segmentation using self-organizing maps and Fuzzy Delphi method - ADSL market of a telecommunication company', *Expert Systems with Applications*. Elsevier Ltd, 38(1), pp. 198–205. doi: 10.1016/j.eswa.2010.06.045.

Hanahan, D. and Weinberg, R. A. (2000) *The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre, Cell*.

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: The next generation', *Cell*. Elsevier Inc., 144(5), pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

Hanley, J. A. and McNeil, B. J. (1982) 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, 143(1), pp. 29–36. doi: 10.1148/radiology.143.1.7063747.

Hao, J., Kim, Y., Mallavarapu, T., Oh, J. H. and Kang, M. (2019) 'Cox-PASNet: Pathway-based Sparse Deep Neural Network for Survival Analysis', in *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*. IEEE, pp. 381–386. doi: 10.1109/BIBM.2018.8621345.

Harrell Jr, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982) 'Evaluating the Yield of Medical Tests', *JAMA*, 247(18), pp. 2543–2546. doi: 10.1001/jama.1982.03320430047030.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982) 'Evaluating the Yield of Medical Tests', *JAMA: The Journal of the American Medical Association*, 247(18), pp. 2543–2546. doi: 10.1001/jama.1982.03320430047030.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996) 'Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing

Errors', *Statistics in Medicine*, 15, pp. 361–387. doi: 10.1002/0470023678.ch2b(i).

Heagerty, P. J. and Zheng, Y. (2005) 'Survival model predictive accuracy and ROC curves', *Biometrics*, 61(1), pp. 92–105. doi: 10.1111/j.0006-341X.2005.030814.x.

Herring, A. H., Ibrahim, J. G. and Lipsitz, S. R. (2004) *Non-ignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial*, *Appl. Statist*.

Hinton, G. (2009) *Deep belief networks*, *Scholarpedia*. Scholarpedia. doi: 10.4249/scholarpedia.5947.

Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006) 'A Fast Learning Algorithm for Deep Belief Nets', *Neural Computation*, 18, pp. 1527–1554. doi: 10.1162/neco.2006.18.7.1527.

Hof, R. D. (2013) 'Deep Learning', *MIT Technology Review*, 116(3), pp. 32–36.

Hosny, A., Parmar, C., Coroller, T. P., Grossmann, P., Zeleznik, R., Kumar, A., Bussink, J., Gillies, R. J., Mak, R. H. and Aerts, H. J. W. L. (2018) 'Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study', *PLoS Medicine*, 15(11). doi: 10.1371/journal.pmed.1002711.

Hothorn, T. and Lausen, B. (2003) 'On the exact distribution of maximally selected rank statistics', *Computational Statistics and Data Analysis*, 43(2), pp. 121–137. doi: 10.1016/S0167-9473(02)00225-6.

Hsu, S. H., Hsieh, J. P. A., Chih, T. C. and Hsu, K. C. (2009) 'A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression', *Expert Systems with Applications*. Elsevier Ltd, 36(4), pp. 7947–7951. doi: 10.1016/j.eswa.2008.10.065.

Hüllermeier, E. (2015) 'Does machine learning need fuzzy logic?', *Fuzzy Sets and Systems*. Elsevier B.V., 281, pp. 292–299. doi: 10.1016/j.fss.2015.09.001.

HUSLAB (2021) *HUSLAB - Karsinoembryonaalinen antigeeni, seerumista*. Available at: https://huslab.fi/ohjekirja/2034.html (Accessed: 6 April 2021).

Hussain, S. P. and Harris, C. C. (2007) 'Inflammation and cancer: An ancient link with novel potentials', *International Journal of Cancer*, 121(11), pp. 2373–2380. doi: 10.1002/ijc.23173.

In, J. and Lee, D. K. (2019) ‘Survival analysis: part II – applied clinical data analysis’, *Korean Journal of Anesthesiology*. Korean Society of Anesthesiologists, 72(5), p. 441. doi: 10.4097/KJA.19183.

Ishwaran, H. (2007) ‘Variable importance in binary regression trees and forests’, *Electronic Journal of Statistics*, 1, pp. 519–537. doi: 10.1214/07-EJS039.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J. and Lau, B. M. (2014) ‘Random survival forests for competing risks’, *Biostatistics*. Biostatistics, 15(4), pp. 757–773. doi: 10.1093/BIOSTATISTICS/KXU010.

Ishwaran, H. and Kogalur, U. B. (2021) *Package ‘randomForestSRC’ Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008) ‘Random survival forests’, *Annals of Applied Statistics*, 2(3), pp. 841–860. doi: 10.1214/08-AOAS169.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. and Lauer, M. S. (2010) ‘High-dimensional variable selection for survival data’, *Journal of the American Statistical Association*, 105(489), pp. 205–217. doi: 10.1198/jasa.2009.tm08622.

Janssens, K., Boeckx, N., Camp, G. Van, Beeck, K. Op De, Fransen, E., Calay, F., Damme, N. Van and Peeters, M. (2018) ‘Comparing survival in left-sided and right-sided colorectal carcinoma: A Belgian population-based study’, *Annals of Oncology*. Elsevier, 29, p. v98. doi: 10.1093/ANNONC/MDY150.017.

Jeong, D., Heo, S., Ahn, T. S., Lee, S., Park, S., Kim, H., Park, D., Bae, S. B., Lee, S. S., Lee, M. S., Kim, C.-J. and Baek, M. J. (2014) ‘Cyr61 Expression is associated with prognosis in patients with colorectal cancer’, *BMC Cancer*, 14(164). doi: 10.1186/1471-2407-14-164.

Jerez-Aragonés, J. M., Gómez-Ruiz, J. A., Ramos-Jiménez, G., Muñoz-Pérez, J. and Alba-Conejo, E. (2003) *A combined neural network and decision trees model for prognosis of breast cancer relapse, Artificial Intelligence in Medicine*.

Jiang, W. G., Sanders, A. J., Katoh, M., Ungefroren, H., Gieseler, F., Prince, M., Thompson, S. K., Zollo, M., Spano, D., Dhawan, P., Sliva, D., Subbarayan, P. R., Sarkar, M., Honoki, K., Fujii, H., Georgakilas, A. G., Amedei, A., Niccolai, E., Amin, A., Ashraf, S. S., Ye, L., Helferich, W. G., Yang, X., Boosani, C. S., Guha, G., Ciriolo, M. R., Aquilano, K., Chen, S.,

Azmi, A. S., Keith, W. N., Bilsland, A., Bhakta, D., Halicka, D., Nowsheen, S., Pantano, F. and Santini, D. (2015) 'Tissue invasion and metastasis: Molecular, biological and clinical perspectives', *Seminars in Cancer Biology*. Academic Press, pp. S244–S275. doi: 10.1016/j.semcancer.2015.03.008.

Jing, B., Zhang, T., Wang, Z., Jin, Y., Liu, K., Qiu, W., Ke, L., Sun, Y., He, C., Hou, D., Tang, L., Lv, X. and Li, C. (2019) 'A deep survival analysis method based on ranking', *Artificial Intelligence in Medicine*. Elsevier B.V., 98(July), pp. 1–9. doi: 10.1016/j.artmed.2019.06.001.

Kalderstam, J., Edén, P., Bendahl, P. O., Strand, C., Fernö, M. and Ohlsson, M. (2013) 'Training artificial neural networks directly on the concordance index for censored data using genetic algorithms', *Artificial Intelligence in Medicine*. Elsevier B.V., 58(2), pp. 125–132. doi: 10.1016/j.artmed.2013.03.001.

Kaplan, E. L. and Meier, P. (1958) *Nonparametric Estimation from Incomplete Observations*, *Journal of the American Statistical Association*.

Kasurinen, A. (2020) *Prognostic value of inflammation-related biomarkers in gastric cancer*. University of Helsinki.

Kasurinen, A., Laitinen, A., Kokkola, A., Stenman, U.-H., Böckelman, C. and Haglund, C. (2020) 'Tumor-associated trypsin inhibitor (TATI) and tumor-associated trypsin-2 (TAT-2) predict outcomes in gastric cancer', *Acta Oncologica*, 59, pp. 1–8. doi: 10.1080/0284186X.2020.1733655.

Kather, J. N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C. A., Gaiser, T., Marx, A., Valous, N. A., Ferber, D., Jansen, L., Reyes-Aldasoro, C. C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M. and Halama, N. (2019) 'Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study', *PLoS Medicine*. Public Library of Science, 16(1). doi: 10.1371/journal.pmed.1002730.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. and Kluger, Y. (2018) 'DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network', *BMC Medical Research Methodology*. BMC Medical Research Methodology, 18(1), pp. 1–12. doi: 10.1186/s12874-018-0482-1.

Kiaee, F., Sheikhzadeh, H. and Eftekhari Mahabadi, S. (2016) 'Relevance Vector Machine for

Survival Analysis', *IEEE Transactions on Neural Networks and Learning Systems*. Institute of Electrical and Electronics Engineers Inc., 27(3), pp. 648–660. doi: 10.1109/TNNLS.2015.2420611.

Kilgour, E., Rothwell, D. G., Brady, G. and Dive, C. (2020) 'Liquid Biopsy-Based Biomarkers of Treatment Response and Resistance', *Cancer Cell*. Cell Press, 37(4), pp. 485–495. doi: 10.1016/j.ccell.2020.03.012.

Kim, S., Park, T. and Kon, M. (2014) 'Cancer survival classification using integrated data sets and intermediate information', *Artificial Intelligence in Medicine*. Elsevier B.V., 62(1), pp. 23–31. doi: 10.1016/j.artmed.2014.06.003.

Klawonn, F., Kruse, R. and Winkler, R. (2015) 'Fuzzy clustering: More than just fuzzification', *Fuzzy Sets and Systems*. Elsevier B.V., 281, pp. 272–279. doi: 10.1016/j.fss.2015.06.024.

Kleinlein, R. and Riaño, D. (2019) 'Persistence of data-driven knowledge to predict breast cancer survival', *International Journal of Medical Informatics*. Elsevier Ireland Ltd, 129, pp. 303–311. doi: 10.1016/j.ijmedinf.2019.06.018.

Klement, P. and Snášel, V. (2011) 'Using SOM in the performance monitoring of the emergency call-taking system', *Simulation Modelling Practice and Theory*. Elsevier B.V., 19(1), pp. 98–109. doi: 10.1016/j.simpat.2010.07.002.

Kohavi, R. (1995) 'A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection', in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., pp. 1137–1143.

Kohonen, T. (1981) 'Automatic formation of topological maps of patterns in a self-organizing system', in Oja, E. and Simula, O. (eds) *Proceedings of 2SCIA, Scandinavian COnference of Image Analysis*. Helsinki, Finland, pp. 214–220.

Kohonen, T. (2013) 'Essentials of the self-organizing map', *Neural Networks*. Elsevier Ltd, 37, pp. 52–65. doi: 10.1016/j.neunet.2012.09.018.

Kohonen, T., Mäkisara, K. and Saramäki, T. (1984) 'Phonotopic Maps -- Insightful Representation of Phonological Features for Speech Recognition', in *Proceedings of the IEEE Seventh International Conference on Pattern Recognition*. Montreal: IEEE Computer Society,

pp. 182–185.

Koskenvuo, L., Pöyhönen, M. and Lepistö, A. (2020) ‘Familiaalinen adenomatoottinen polypoosi (FAP)’, *Duodecim*, 136(1), pp. 52–60.

Køstner, A. H., Kersten, C., Löwenmark, T., Ydsten, K. A., Peltonen, R., Isoniemi, H., Haglund, C., Gunnarsson, U. and Isaksson, B. (2016) ‘The prognostic role of systemic inflammation in patients undergoing resection of colorectal liver metastases: C-reactive protein (CRP) is a strong negative prognostic biomarker’, *Journal of Surgical Oncology*. John Wiley and Sons Inc., 114(7), pp. 895–899. doi: 10.1002/jso.24415.

Kruse, R. and Nauck, D. (1998) ‘Neuro-Fuzzy Systems’, in Kaynak, O., Zadeh, L. A., Türkşen, B., and Rudas, I. J. (eds) *Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications. NATO ASI Series (Series F: Computer and Systems Sciences), vol 162*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 230–259. doi: 10.1007/978-3-642-58930-0_12.

Kuhn, M. and Johnson, K. (2019) *Feature Engineering and Selection: A Practical Approach for Predictive Models*. 1st edn. Chapman and Hall/CRC. doi: https://doi.org/10.1201/9781315108230.

Kuitunen, I., Ponkilainen, V. T., Uimonen, M. M., Eskelinen, A. and Reito, A. (2021) ‘Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review’, *BMC Musculoskeletal Disorders 2021 22:1*. BioMed Central, 22(1), pp. 1–7. doi: 10.1186/S12891-021-04379-2.

Kumari, N., Dwarakanath, B. S., Das, A. and Bhatt, A. N. (2016) ‘Role of interleukin-6 in cancer progression and therapeutic resistance’, *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*. Tumour Biol, 37(9), pp. 11553–11572. doi: 10.1007/S13277-016-5098-7.

Kvamme, H., Borgan, O. and Scheel, I. (2019) ‘Time-to-event prediction with neural networks and cox regression’, *Journal of Machine Learning Research*, 20, pp. 1–30.

de la Chapelle, A. (2004) ‘Genetic predisposition to colorectal cancer’, *Nature Reviews Cancer*, 4(10), pp. 769–780. doi: 10.1038/nrc1453.

Labianca, R., Nordlinger, B., Beretta, G. D., Brouquet, A. and Cervantes, A. (2010) 'Primary colon cancer: ESMO Clinical Practice Guidelines for diagnosis, adjuvant treatment and follow-up', *Annals of Oncology*. Elsevier, 21(SUPPL. 5), pp. v70–v77. doi: 10.1093/ANNONC/MDQ168.

Lambert, A. W., Pattabiraman, D. R. and Weinberg, R. A. (2017) 'Emerging Biological Principles of Metastasis', *Cell*. doi: 10.1016/j.cell.2016.11.037.

LAUREN, P. (1965) 'THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. AN ATTEMPT AT A HISTO-CLINICAL CLASSIFICATION', *Acta pathologica et microbiologica Scandinavica*, 64, p. 31—49. doi: 10.1111/apm.1965.64.1.31.

LeBlanc, M. and Crowley, J. (1993) 'Survival trees by goodness of split', *Journal of the American Statistical Association*, 88(422), pp. 457–467. doi: 10.1080/01621459.1993.10476296.

Lee, C., Yoon, J. and Van Der Schaar, M. (2020) 'Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis with Competing Risks Based on Longitudinal Data', *IEEE Transactions on Biomedical Engineering*. IEEE, 67(1), pp. 122–133. doi: 10.1109/TBME.2019.2909027.

Lee, C., Zame, W. R., Yoon, J. and Van Der Schaar, M. (2018) 'DeepHit: A deep learning approach to survival analysis with competing risks', in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 2314–2321.

Lee, E. T. and Wang, J. W. (2003) *Statistical methods for survival data analysis*. 3rd edn. John Wiley & Sons.

Lee, M. S., Mcguffey, E. J., Morris, J. S., Manyam, G., Baladandayuthapani, V., Wei, W., Morris, V. K., Overman, M. J., Maru, D. M., Jiang, Z.-Q., Hamilton, S. R. and Kopetz, S. (2016) 'Association of CpG island methylator phenotype and ERE/AREG methylation and expression in colorectal cancer', *British Journal of Cancer*, (114), pp. 1352–1361. doi: 10.1038/bjc.2016.87.

Leocata, P., Ventura, L., Giunta, M., Guadagni, S., Fortunato, C., Discepoli, S. and Ventura, T. (1998) 'Gastric carcinoma: Histopathologic study of 705 cases', *Annali italiani di chirurgia*,

69, pp. 331–337.

Lewinson, E. (2020) *Introduction to Survival Analysis: the Nelson-Aalen estimator, Towards Data Science*. Available at: https://towardsdatascience.com/introduction-to-survival-analysis-the-nelson-aalen-estimator-9780c63d549d (Accessed: 7 January 2022).

Li, D. C., Hsu, H. C., Tsai, T. I., Lu, T. J. and Hu, S. C. (2007) 'A new method to help diagnose cancers for small sample size', *Expert Systems with Applications: An International Journal*. Pergamon Press, Inc. PUB1185 Elmsford, NY, USA , 33(2), pp. 420–424. doi: 10.1016/J.ESWA.2006.05.028.

Li, L., Hu, Q., Wu, X. and Yu, D. (2014) 'Exploration of classification confidence in ensemble learning', *Pattern Recognition*, 47(9), pp. 3120–3131. doi: 10.1016/j.patcog.2014.03.021.

Li, S. and Razzaghi, T. (2019) 'Personalized Colorectal Cancer Survivability Prediction with Machine Learning Methods', *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*. IEEE, pp. 2554–2558. doi: 10.1109/BigData.2018.8622121.

Li, Y., Wang, L., Zhou, J. and Ye, J. (2019) 'Multi-task learning based survival analysis for multi-source block-wise missing data', *Neurocomputing*. Elsevier B.V., 364, pp. 95–107. doi: 10.1016/j.neucom.2019.07.010.

Lifelines (2021) *Introduction to survival analysis — lifelines 0.26.4 documentation, Introduction to survival analysis*. Available at: https://lifelines.readthedocs.io/en/latest/Survival Analysis intro.html (Accessed: 7 January 2022).

Lim, B., Lin, Y. and Navin, N. (2020) 'Advancing Cancer Research and Medicine with Single-Cell Genomics', *Cancer Cell*. Cell Press, 37(4), pp. 456–470. doi: 10.1016/j.ccell.2020.03.008.

Lima, E., Mues, C. and Baesens, B. (2009) 'Domain knowledge integration in data mining using decision tables: Case studies in churn prediction', *Journal of the Operational Research Society*. Palgrave Macmillan Ltd., 60(8), pp. 1096–1106. doi: 10.1057/jors.2008.161.

Lin, C. F. and Wang, S. De (2002) 'Fuzzy support vector machines', *IEEE Transactions on Neural Networks*, 13(2), pp. 464–471. doi: 10.1109/72.991432.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical analysis with missing data*. Hoboken, N. J.:

Wiley.

Liu, P.-H., Wu, K., Ng, K., Zauber, A. G., Nguyen, L. H., Song, M., He, X., Fuchs, C. S., Ogino, S., Willett, W. C., Chan, A. T., Giovannucci, E. L. and Cao, Y. (2019) 'Association of Obesity With Risk of Early-Onset Colorectal Cancer Among Women', *JAMA Oncology*, 5(1), pp. 37–44. doi: 10.1001/jamaoncol.2018.4280.

Liu, Y., Aickelin, U., Feyereisl, J. and Durrant, L. G. (2013) 'Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data', *Knowledge-Based Systems*. Elsevier B.V., 37, pp. 502–514. doi: 10.1016/j.knosys.2012.09.011.

Lizama, A. J., Andrade, Y., Colivoro, P., Sarmiento, J., Matus, C. E., Gonzalez, C. B., Bhoola, K. D., Ehrenfeld, P. and Figueroa, C. D. (2015) 'Expression and bioregulation of the kallikrein-related peptidases family in the human neutrophil', *Innate Immunity*, 21(6), pp. 575–586. doi: 10.1177/1753425914566083.

Loh, W. Y. (2002) 'Regression trees with unbiased variable selection and interaction detection', *Statistica Sinica*, 12(2), pp. 361–386.

Lohrmann, C. and Luukka, P. (2019) 'Using Clustering for Supervised Feature Selection to Detect Relevant Features', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Cham, 11943 LNCS, pp. 272–283. doi: 10.1007/978-3-030-37599-7_23.

Lohrmann, C., Luukka, P., Jablonska-Sabuka, M. and Kauranne, T. (2018) 'A combination of fuzzy similarity measures and fuzzy entropy measures for supervised feature selection', *Expert Systems with Applications*. Elsevier Ltd, 110, pp. 216–236. doi: 10.1016/j.eswa.2018.06.002.

Lunn, M. and McNeil, D. (1995) 'Applying Cox Regression to Competing Risks', *Biometrics*. JSTOR, 51(2), p. 524. doi: 10.2307/2532940.

Luo, C., Cen, S., Ding, G. and Wu, W. (2019) 'Mucinous colorectal adenocarcinoma: Clinical pathology and treatment options', *Cancer Communications*. BioMed Central Ltd. doi: 10.1186/s40880-019-0361-0.

Luo, X., Yu, Y., Liang, A., Xie, Y., Liu, S., Guo, J., Wang, W., Qi, R., An, H., Zhang, M., Xu, H., Guo, Z. and Cao, X. (2004) 'Intratumoral expression of MIP-1beta induces antitumor responses in a pre-established tumor model through chemoattracting T cells and NK cells.',

Cellular & Molecular Immunology, 1(3), pp. 199–204. doi: 10.1182/blood.v104.11.5268.5268.

Luo, Z., Wu, X., Guo, S. and Ye, B. (2008) 'Diagnosis of breast cancer tumor based on PCA and fuzzy Support Vector Machine classifier', in *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, pp. 363–367. doi: 10.1109/ICNC.2008.932.

Luukka, P. (2008) 'Similarity classifier in diagnosis of bladder cancer', *Computer Methods and Programs in Biomedicine*, 89(1), pp. 43–49. doi: 10.1016/j.cmpb.2007.10.001.

Luukka, P. (2011) 'Feature selection using fuzzy entropy measures with similarity classifier', *Expert Systems with Applications*. Pergamon, 38(4), pp. 4600–4607. doi: 10.1016/J.ESWA.2010.09.133.

Macías-García, L., Luna-Romera, J. M., García-Gutiérrez, J., Martínez-Ballesteros, M., Riquelme-Santos, J. C. and González-Cámpora, R. (2017) 'A study of the suitability of autoencoders for preprocessing data in breast cancer experimentation', *Journal of Biomedical Informatics*, 72, pp. 33–44. doi: 10.1016/j.jbi.2017.06.020.

Mallows, C. L. (1973) 'Some Comments on C_P', *Technometrics*. [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality], 15(4), pp. 661–675. doi: 10.2307/1267380.

Mangasarian, O. L., Street, W. N. and Wolberg, W. H. (1995) 'Breast Cancer Diagnosis and Prognosis via Linear Programming', *Operations Research*, 43(4), pp. 570–577. doi: 10.1287/opre.43.4.570.

Mansoori, T. K., Suman, A. and Mishra, S. K. (2014) 'Application of Genetic Algorithm for Cancer Diagnosis by Feature Selection', *International Journal of Engineering Research & Technology*, 3(8), pp. 1295–1301.

Mao, Y., Zhou, X., Pi, D., Sun, Y. and Wong, S. T. C. (2005) 'Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection', *Journal of Biomedicine and Biotechnology*, 2005(2), pp. 160–171. doi: 10.1155/JBB.2005.160.

Markowitz, S. D., Dawson, D. M., Willis, J. and Willson, J. K. V. (2002) 'Focus on colon cancer', *Cancer Cell*. Cell Press, 1(3), pp. 233–236. doi: 10.1016/S1535-6108(02)00053-3.

Matsuo, K., Purushotham, S., Jiang, B., Mandelbaum, R. S., Takiuchi, T., Liu, Y. and Roman,

L. D. (2019) ‘Survival outcome prediction in cervical cancer: Cox models vs deep-learning model’, *American Journal of Obstetrics and Gynecology*. Elsevier Inc., 220(4), pp. 381.e1-381.e14. doi: 10.1016/j.ajog.2018.12.030.

Mayo Clinic (2021) *Colectomy*. Available at: https://www.mayoclinic.org/tests-procedures/colectomy/about/pac-20384631 (Accessed: 4 April 2021).

McBride, R., Hicks, B. M., Coleman, H. G., Loughrey, M. B., Gavin, A. T., Dunne, P. D. and Campbell, W. J. (2020) ‘Prognosis following surgical resection versus local excision of stage pT1 colorectal cancer: A population-based cohort study’, *Surgeon*. Elsevier Ltd, 18(2), pp. 65–74. doi: 10.1016/j.surge.2019.06.004.

McGill University (2021) *What is Pathology? | DEPARTMENT OF PATHOLOGY*. Available at: https://www.mcgill.ca/pathology/about/definition (Accessed: 25 January 2021).

Mereiter, S., Balmaña, M., Campos, D., Gomes, J. and Reis, C. A. (2019) ‘Glycosylation in the Era of Cancer-Targeted Therapy: Where Are We Heading?’, *Cancer Cell*. Cell Press, 36(1), pp. 6–16. doi: 10.1016/j.ccell.2019.06.006.

Meyer-Baese, A. and Schmid, V. (2014) ‘Chapter 7 - Foundations of Neural Networks’, in *Pattern Recognition and Signal Analysis in Medical Imaging*. 2nd edn. Oxford: Academic Press, pp. 197–243. doi: https://doi.org/10.1016/B978-0-12-409545-8.00007-8.

Mirinezhad, S., Moaddab, S., Bonyadi, M., Shirmohammadi, M., Eftekharsadat, A. and Somi, M. (2018) ‘Survival of familial adenomatous polyposis coexistence colorectal cancer in Iran’, *Journal of Cancer Research and Therapeutics*. Wolters Kluwer Medknow Publications, 15(1), pp. 87–91. doi: 10.4103/jcrt.JCRT_421_17.

Mitchell, T. J. and Beauchamp, J. J. (1988) ‘Bayesian Variable Selection in Linear Regression’, *Journal of the American Statistical Association*, 83(404), p. 1034. doi: 10.2307/2290131.

Mitselou, A., Galani, V., Skoufi, U., Arvanis, D. L., Lampri, E. and Ioachim, E. (2016) ‘Syndecan-1, Epithelial-Mesenchymal Transition Markers (E-cadherin/β-catenin) and Neoangiogenesis-related Proteins (PCAM-1 and Endoglin) in Colorectal Cancer’, *Anticancer Research*, 36(5), pp. 2271–2280.

Mogensen, U. B., Ishwaran, H. and Gerds, T. A. (2012) ‘Evaluating Random Forests for Survival Analysis Using Prediction Error Curves’, *Journal of Statistical Software*, 50(11), pp.

1–23. doi: 10.18637/jss.v050.i11.

Montella, A., de Oña, R., Mauriello, F., Rella Riccardi, M. and Silvestro, G. (2020) 'A data mining approach to investigate patterns of powered two-wheeler crashes in Spain', *Accident Analysis and Prevention*. Elsevier, 134(May), p. 105251. doi: 10.1016/j.aap.2019.07.027.

Moreau, T., O'Quigley, J. and Mesbah, M. (1985) 'A Global Goodness-of-Fit Statistic for the Proportional Hazards Model', *Applied Statistics*. JSTOR, 34(3), p. 212. doi: 10.2307/2347465.

Mullangi, S. and Lekkala, M. R. (2021) *Adenocarcinoma - StatPearls - NCBI Bookshelf*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK562137/ (Accessed: 12 March 2021).

Murciano-Goroff, Y. R., Taylor, B. S., Hyman, D. M. and Schram, A. M. (2020) 'Toward a More Precise Future for Oncology', *Cancer Cell*. Cell Press, 37(4), pp. 431–442. doi: 10.1016/j.ccell.2020.03.014.

Murtojärvi, M., Halkola, A. S., Airola, A., Laajala, T. D., Mirtti, T., Aittokallio, T. and Pahikkala, T. (2020) 'Cost-effective survival prediction for patients with advanced prostate cancer using clinical trial and real-world hospital registry datasets', *International Journal of Medical Informatics*. Elsevier, 133(October 2018), p. 104014. doi: 10.1016/j.ijmedinf.2019.104014.

National Cancer Institution (2020) *Pancreatic Cancer Treatment (Adult) (PDQ®)–Patient Version - National Cancer Institute*. Available at: https://www.cancer.gov/types/pancreatic/patient/pancreatic-treatment-pdq (Accessed: 10 November 2021).

National Cancer Institution (2021) *Definition: Resection*. Available at: https://www.cancer.gov/search/results?swKeyword=resection (Accessed: 4 April 2021).

National Institute of Diabetes and Digestive and Kidney Diseases (2017) *Your Digestive System & How it Works*. Available at: https://www.niddk.nih.gov/health-information/digestive-diseases/digestive-system-how-it-works (Accessed: 29 December 2021).

Nauck, D. and Kruse, R. (1999) 'Obtaining interpretable fuzzy classification rules from medical data', *Artificial Intelligence in Medicine*, 16(2), pp. 149–169. doi: https://doi.org/10.1016/S0933-3657(98)00070-0.

Nelson, W. (1972) ‘Theory and Applications of Hazard Plotting for Censored Failure Data’, *Technometrics*, 14(4), pp. 945–966.

Ng, K., Nimeiri, H. S., McCleary, N. J., Abrams, T. A., Yurgelun, M. B., Cleary, J. M., Rubinson, D. A., Schrag, D., Miksad, R., Bullock, A. J., Allen, J., Zuckerman, D., Chan, E., Chan, J. A., Wolpin, B. M., Constantine, M., Weckstein, D. J., Faggen, M. A., Thomas, C. A., Kournioti, C., Yuan, C., Ganser, C., Wilkinson, B., Mackintosh, C., Zheng, H., Hollis, B. W., Meyerhardt, J. A. and Fuchs, C. S. (2019) ‘Effect of High-Dose vs Standard-Dose Vitamin D3 Supplementation on Progression-Free Survival Among Patients With Advanced or Metastatic Colorectal Cancer: The SUNSHINE Randomized Clinical Trial’, *JAMA*, 321(14), pp. 1370–1379. doi: 10.1001/jama.2019.2402.

Nikooienejad, A., Wang, W. and Johnson, V. E. (2020) ‘Bayesian variable selection for survival data using inverse moment priors’, *Annals of Applied Statistics*. Institute of Mathematical Statistics, 14(2), pp. 809–828. doi: 10.1214/20-AOAS1325.

Njamen-Njomen, D. A. and Ngatchou-Wandji, J. (2014) ‘Nelson-Aalen and Kaplan-Meier Estimators in Competing Risks’, *Applied Mathematics*. Scientific Research Publishing, Inc, 05(04), pp. 765–776. doi: 10.4236/AM.2014.54073.

Oakes, D. (2000) ‘Survival Analysis’, *Journal of the American Statistical Association*, 95(449), pp. 282–285. doi: 10.2307/2669547.

Olén, O., Erichsen, R., Sachs, M. C., Pedersen, L., Halfvarson, J., Askling, J., Ekbom, A., Sørensen, H. T. and Ludvigsson, J. F. (2020) ‘Colorectal cancer in Crohn’s disease: a Scandinavian population-based cohort study’, *The Lancet Gastroenterology and Hepatology*, 5(5), pp. 475–484. doi: 10.1016/S2468-1253(20)30005-4.

OncoLink (2021) *All About Colon Cancer*. Available at: https://www.oncolink.org/cancers/gastrointestinal/colon-cancer/all-about-colon-cancer (Accessed: 29 December 2021).

Oztekin, A., Delen, D. and Kong, Z. (James) (2009) ‘Predicting the graft survival for heart-lung transplantation patients: An integrated data mining methodology’, *International Journal of Medical Informatics*, 78(12). doi: 10.1016/j.ijmedinf.2009.04.007.

Pacal, I., Karaboga, D., Basturk, A., Akay, B. and Nalbantoglu, U. (2020) ‘A comprehensive

review of deep learning in colon cancer', *Computers in Biology and Medicine*. Elsevier Ltd, 126(April), p. 104003. doi: 10.1016/j.compbiomed.2020.104003.

Pavlidis, N. G., Tasoulis, D. K., Adams, N. M. and Hand, D. J. (2012) 'Adaptive consumer credit classification', *Journal of the Operational Research Society*, 63(12), pp. 1645–1654. doi: 10.1057/jors.2012.15.

Pechacek, J., Gelder, A., Roberts, C., King, J., Bishop, J., Guggisberg, M. and Kirpichevsky, Y. (2019) *Institute for Defense Analyses A New Military Retention Prediction Model:: Machine Learning for High-Fidelity Forecasting*.

Perdue, D. G., Haverkamp, D., Perkins, C., Daley, C. M. and Provost, E. (2014) 'Geographic Variation in Colorectal Cancer Incidence and Mortality, Age of Onset, and Stage at Diagnosis Among American Indian and Alaska Native People, 1990–2009', *American Journal of Public Health*. American Public Health Association, 104(Suppl 3), p. S404. doi: 10.2105/AJPH.2013.301654.

Picard, R. R. and Cook, R. D. (1984) 'Cross-Validation of Regression Models', *Journal of the American Statistical Association*, 79(387), pp. 575–583. doi: 10.1080/01621459.1984.10478083.

Pitkäniemi J, Malila N, Virtanen A, Degerlund H, Heikkinen S, S. K. (2020) *Syöpä 2018. Tilastoraportti Suomen syöpätilanteesta., Suomen Syöpäyhdistyksen julkaisuja nro 93*. Helsinki: Suomen Syöpäyhdistys.

Pitkäniemi, J., Malila, N., Tanskanen, T., Degerlund, H., Heikkinen, S. and Seppä, K. (2021) *Syöpä 2019 Tilastoraportti Suomen syöpätilanteesta Syöpäjärjestöjen epidemiologinen tutkimuslaitos*.

Pölsterl, S., Conjeti, S., Navab, N. and Katouzian, A. (2016) 'Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection', *Artificial Intelligence in Medicine*. Elsevier B.V., 72, pp. 1–11. doi: 10.1016/j.artmed.2016.07.004.

Pölsterl, S., Navab, N. and Katouzian, A. (2015) 'Fast training of support vector machines for survival analysis', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 243–259. doi: 10.1007/978-3-

319-23525-7_15.

Prostate Conditions Education Council (2020) *Gleason Score - Prostate Conditions*. Available at: https://www.prostateconditions.org/about-prostate-conditions/prostate-cancer/newly-diagnosed/gleason-score (Accessed: 6 October 2020).

Prud'homme, G. J. (2007) 'Pathobiology of transforming growth factor β in cancer, fibrosis and immunologic disease, and therapeutic considerations', *Laboratory Investigation*. Nature Publishing Group, pp. 1077–1091. doi: 10.1038/labinvest.3700669.

Przemyslaw, L., Oguslaw, H. A., Elzbieta, S. and Malgorzata, S. M. (2013) 'ADAM and ADAMTS family proteins and their role in the colorectal cancer etiopathogenesis', *BMB Reports*. Korean Society for Biochemistry and Molecular Biology, 46(3), pp. 139–150. doi: 10.5483/BMBREP.2013.46.3.176.

Quantin, C., Abrahamowicz, M., Moreau, T., Bartlett, G., MacKenzie, T., Tazi, M. A., Lalonde, L. and Faivre, J. (1999) 'Variation over time of the effects of prognostic factors in a population-based study of colon cancer: Comparison of statistical models', *American Journal of Epidemiology*, pp. 1188–1200. doi: 10.1093/oxfordjournals.aje.a009945.

Que, S. J., Chen, Q. Y., Zhong, Q., Liu, Z. Y., Wang, J. Bin, Lin, J. X., Lu, J., Cao, L. L., Lin, M., Tu, R. H., Huang, Z. N., Lin, J. L., Zheng, H. L., Li, P., Zheng, C. H., Huang, C. M. and Xie, J. W. (2019) 'Application of preoperative artificial neural network based on blood biomarkers and clinicopathological parameters for predicting long-term survival of patients with gastric cancer', *World Journal of Gastroenterology*, 25(43), pp. 6451–6464. doi: 10.3748/wjg.v25.i43.6451.

R Package Documentation (2021a) *rfsrc: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. Available at: https://www.rdocumentation.org/packages/randomForestSRC/versions/2.11.0/topics/rfsrc (Accessed: 19 June 2021).

R Package Documentation (2021b) *var.select.rfsrc: Variable Selection in randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. Available at: https://rdrr.io/cran/randomForestSRC/man/var.select.rfsrc.html (Accessed: 14 June 2021).

Rahimian, F., Salimi-Khorshidi, G., Payberah, A. H., Tran, J., Ayala Solares, R., Raimondi, F.,

Nazarzadeh, M., Canoy, D. and Rahimi, K. (2018) 'Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records', *PLoS Medicine*. Public Library of Science, 15(11). doi: 10.1371/journal.pmed.1002695.

Rajpar, S. (2020) *Wide local excision*.

Ramos-Jiménez, G. Morales-Bueno, R. Villalba-Soria, A. (2000) 'CIDIM. Control of induction by sample division methods', in *Proceedings of the International Conference on Artificial Intelligence (IC-AI 2000)*. Las Vegas, pp. 1083–1087.

Ramos, M., Montaño, J., Esteva, M., Barceló, A. and Franch, P. (2016) 'Colorectal cancer survival by stage of cases diagnosed in Mallorca, Spain, between 2006 and 2011 and factors associated with survival', *Cancer Epidemiology*. Elsevier Ltd, 41, pp. 63–70. doi: 10.1016/j.canep.2016.01.001.

RDocumentation (2021) *knn.impute function - RDocumentation*. Available at: https://www.rdocumentation.org/packages/bnstruct/versions/1.0.9/topics/knn.impute (Accessed: 16 November 2021).

Reijnen, C., Gogou, E., Visser, N. C. M., Engerud, H., Ramjith, J., van der Putten, L. J. M., van de Vijver, K., Santacana, M., Bronsert, P., Bulten, J., Hirschfeld, M., Colas, E., Gil-Moreno, A., Reques, A., Mancebo, G., Krakstad, C., Trovik, J., Haldorsen, I. S., Huvila, J., Koskas, M., Weinberger, V., Bednarikova, M., Hausnerova, J., van der Wurff, A. A. M., Matias-Guiu, X., Amant, F., Massuger, L. F. A. G., Snijders, M. P. L. M., Küsters-Vandevelde, H. V. N., Lucas, P. J. F. and Pijnenborg, J. M. A. (2020) 'Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study', *PLoS medicine*. NLM (Medline), 17(5), p. e1003111. doi: 10.1371/journal.pmed.1003111.

Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L. and Yu, Y. (2019) 'Deep Recurrent Survival Analysis', in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*. AAAI Press, pp. 4798–4805. doi: 10.1609/aaai.v33i01.33014798.

Ribas, V. J., Lopez, J. C., Ruiz-Sanmartin, A., Ruiz-Rodriguez, J. C., Rello, J., Wojdel, A. and Vellido, A. (2011) 'Severe sepsis mortality prediction with relevance vector machines', in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and*

Biology Society, EMBS, pp. 100–103. doi: 10.1109/IEMBS.2011.6089906.

Ripley, B. D. (1994) 'Neural Networks and Related Methods for Classification', *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3), pp. 409–456.

Ripley, B. D. and Ripley, R. M. (1998) 'Neural networks as statistical methods in survival analysis', in Dybowski, R. and Gant, V. (eds) *Artificial Neural Networks: Prospects for Medicine*. Landes Biosciences Publishers, pp. 237–255. doi: 10.1017/CBO9780511543494.011.

Ritari, J., Hyvärinen, K., Koskela, S., Itälä-Remes, M., Niittyvuopio, R., Nihtinen, A., Salmenniemi, U., Putkonen, M., Volin, L., Kwan, T., Pastinen, T. and Partanen, J. (2019) 'Genomic prediction of relapse in recipients of allogeneic haematopoietic stem cell transplantation', *Leukemia*. Springer US, 33(1), pp. 240–248. doi: 10.1038/s41375-018-0229-3.

Rosenblatt, F. (1958) 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological Review*, 65(6), pp. 386–408. doi: 10.1037/h0042519.

Rumba, R., Cipkina, S., Cukure, F. and Vanags, A. (2018) 'Systemic and local inflammation in colorectal cancer', *Acta Medica Lituanica*. Vilnius University Press, 25(4), p. 185. doi: 10.6001/ACTAMEDICA.V25I4.3929.

Ryu, Y. U., Chandrasekaran, R. and Jacob, V. (2004) 'Prognosis using an isotonic prediction technique', *Management Science*, 50(6), pp. 777–785. doi: 10.1287/mnsc.1030.0137.

Saarto, T., Österlund, P. and Lepistö, A. (2013) 'Pahanlaatuisen leikkaukseen soveltumattoman suolitukoksen konservatiivinen hoito', *Duodecim*, 129(4), pp. 410–417.

Safavian, S. R. and Landgrebe, D. (1991) 'A Survey of Decision Tree Classifier Methodology', *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), pp. 660–674.

Sargent, D. J. (2001) 'Comparison of artificial neural networks with other statistical approaches', *Cancer*. John Wiley & Sons, Ltd, 91(S8), pp. 1636–1642. doi: 10.1002/1097-0142(20010415)91:8+<1636::AID-CNCR1176>3.0.CO;2-D.

Sauerbrei, W. and Schumacher, M. (1992) 'A bootstrap resampling procedure for model building: Application to the cox regression model', *Statistics in Medicine*, 11(16), pp. 2093–

2109. doi: 10.1002/sim.4780111607.

Schafer, J. L. (1999) ‘NORM Users Guide Multiple imputation of incomplete multivariate data under a normal model’.

Schemper, M., Wakounig, S. and Heinze, G. (2009) ‘The estimation of average hazard ratios by weighted Cox regression’, *Statistics in Medicine*. John Wiley & Sons, Ltd, 28(19), pp. 2473–2489. doi: 10.1002/SIM.3623.

Schmid, M., Wright, M. N. and Ziegler, A. (2016) ‘On the use of Harrell’s C for clinical risk prediction via random survival forests’, *Expert Systems with Applications*, 63, pp. 450–459. doi: 10.1016/j.eswa.2016.07.018.

Schoenfeld, D. (1982) ‘Partial Residuals for The Proportional Hazards Regression Model’, *Biometrika*, 69(1), p. 239. doi: 10.2307/2335876.

Schumacher, M., Holländer, N. and Sauerbrei, W. (1997) ‘Resampling and cross-validation techniques: A tool to reduce bias caused by model building?’, *Statistics in Medicine*, 16(24), pp. 2813–2827. doi: 10.1002/(SICI)1097-0258(19971230)16:24<2813::AID-SIM701>3.0.CO;2-Z.

Schwenk, H. and Bengio, Y. (2000) ‘Boosting neural networks’, *Neural Computation*, 12(8), pp. 1869–1887. doi: 10.1162/089976600300015178.

ScienceDirect (2022) *ScienceDirect.com | Science, health and medical journals, full text articles and books.* Available at: https://www.sciencedirect.com/ (Accessed: 4 January 2022).

Segal, M. R. (1988) ‘Regression Trees for Censored Data Author’, *Biometrics*, 44(1), pp. 35–47. doi: 10.2307/2531894.

Seker, H., Odetayo, M. O., Petrovic, D. and Naguib, R. N. G. (2003) ‘A fuzzy logic based-method for prognostic decision making in breast and prostate cancers’, *IEEE Transactions on Information Technology in Biomedicine*, 7(2), pp. 114–122. doi: 10.1109/TITB.2003.811876.

Shao, J. (1993) ‘Linear model selection by cross-validation’, *Journal of the American Statistical Association*, 88(422), pp. 486–494. doi: 10.1080/01621459.1993.10476299.

Shibata, R. (1981) ‘An Optimal Selection of Regression Variables’, *Biometrika*. [Oxford University Press, Biometrika Trust], 68(1), pp. 45–54. doi: 10.2307/2335804.

Shivaswamy, P. K., Chu, W. and Jansche, M. (2007) ‘A support vector approach to censored targets’, *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 655–660. doi: 10.1109/ICDM.2007.93.

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A. and Jemal, A. (2020) ‘Colorectal cancer statistics, 2020’, *CA: A Cancer Journal for Clinicians*. Wiley, 70(3), pp. 145–164. doi: 10.3322/caac.21601.

Sino Biological (2019) *Role of Cytokines in cancer*. Available at: https://www.sinobiological.com/resource/cytokines/role-of-cytokines-in-cancer (Accessed: 4 August 2020).

Sirniö, P., Tuomisto, A., Tervahartiala, T., Sorsa, T., Klintrup, K., Karhu, T., Herzig, K. H., Mäkelä, J., Karttunen, T. J., Salo, T., Mäkinen, M. J. and Väyrynen, J. P. (2018) ‘High-serum MMP-8 levels are associated with decreased survival and systemic inflammation in colorectal cancer’, *British Journal of Cancer*. Nature Publishing Group, 119(2), pp. 213–219. doi: 10.1038/s41416-018-0136-4.

Smith, K. A. and Gupta, J. N. D. (2002) ‘Neural networks in business: techniques and applications’, *Computers & Operations Research*, 27, p. 271.

Snow, P. B., Kerr, D. J., Brandt, J. M. and Rodvold, D. M. (2001) ‘Neural network and regression predictions of 5-year survival after colon carcinoma treatment’, *Cancer*, 91(8 SUPPL.), pp. 1673–1678. doi: 10.1002/1097-0142(20010415)91:8+<1673::aid-cncr1182>3.0.co;2-t.

Sreekumar, R., Harris, S., Moutasim, K., DeMateos, R., Patel, A., Emo, K., White, S., Yagci, T., Tulchinsky, E., Thomas, G., Primrose, J. N., Sayan, A. E. and Mirnezami, A. H. (2018) ‘Assessment of Nuclear ZEB2 as a Biomarker for Colorectal Cancer Outcome and TNM Risk Stratification’, *JAMA network open*. NLM (Medline), 1(6), p. e183115. doi: 10.1001/jamanetworkopen.2018.3115.

Štajduhar, I., Dalbelo-Bašić, B. and Bogunović, N. (2009) ‘Impact of censoring on learning Bayesian networks in survival modelling’, *Artificial Intelligence in Medicine*, 47(3), pp. 199–217. doi: 10.1016/j.artmed.2009.08.001.

van Stiphout, R. G. P. M., Postma, E. O., Valentini, V. and Lambin, P. (2010) ‘The contribution

of machine learning to predicting cancer outcome', in *Belgian/Netherlands Artificial Intelligence Conference*.

Suchorska, B., Schüller, U., Biczok, A., Lenski, M., Albert, N. L., Giese, A., Kreth, F. W., Ertl-Wagner, B., Tonn, J. C. and Ingrisch, M. (2019) 'Contrast enhancement is a prognostic factor in IDH1/2 mutant, but not in wild-type WHO grade II/III glioma as confirmed by machine learning', *European Journal of Cancer*. Elsevier Ltd, 107, pp. 15–27. doi: 10.1016/j.ejca.2018.10.019.

Suomen Syöpärekisteri (2021) *Syöpätilastosovellus*. Available at: https://syoparekisteri.fi/tilastot/tautitilastot/ (Accessed: 22 February 2021).

Suomen virallinen tilasto (SVT) (2020) *Kuolemansyyt*. Helsinki. doi: 1799-5051.

Suykens, J. A. K. and Vanderwalle, J. (1998) 'Least Squares Support Vector Machine Classifiers', *Neural Processing Letters*, 9, pp. 293–300. doi: 10.1023/A:1018628609742.

Talieri, M., Li, L., Zheng, Y., Alexopoulou, D. K., Soosaipillai, A., Scorilas, A., Xynopoulos, D. and Diamandis, E. P. (2009) 'The use of kallikrein-related peptidases as adjuvant prognostic markers in colorectal cancer', *British Journal of Cancer*, (100), pp. 1659–1665. doi: 10.1038/sj.bjc.6605033.

Talmadge, J. E. and Fidler, I. J. (2010) 'AACR centennial series: The biology of cancer metastasis: Historical perspective', *Cancer Research*, pp. 5649–5669. doi: 10.1158/0008-5472.CAN-10-1040.

Tayel, S. I., Fouda, E. A. M., Gohar, S. F., Elshayeb, E. I., El-sayed, E. H. and El-kousy, S. M. (2018) 'Potential role of MicroRNA 200c gene expression in assessment of colorectal cancer', *Archives of Biochemistry and Biophysics*. Academic Press Inc., 647, pp. 41–46. doi: 10.1016/j.abb.2018.04.009.

Terveyskirjasto, D. (2018) *Suolistosyöpä (ohutsuolen ja paksusuolen syövät), Lääkärikirja Duodecim*. Available at: https://www.terveyskirjasto.fi/dlk01087 (Accessed: 16 January 2022).

The McGraw-Hill Companies, I. (2002) *Index of suspicion | definition of index of suspicion by Medical dictionary, McGraw-Hill Concise Dictionary of Modern Medicine*. Available at: https://medical-dictionary.thefreedictionary.com/index+of+suspicion (Accessed: 19 November 2020).

Theodoridis, S. and Koutroumbas, K. (2009) ‘Linear Classifiers’, in Theodoridis, S. and Koutroumbas, K. B. T.-P. R. (Fourth E. (eds) *Pattern Recognition*. Boston: Academic Press, pp. 91–150. doi: https://doi.org/10.1016/B978-1-59749-272-0.50005-0.

Tibble, J., Sigthorsson, G., Foster, R., Sherwood, R., Fagerhol, M. and Bjarnason, I. (2001) ‘Faecal calprotectin and faecal occult blood tests in the diagnosis of colorectal carcinoma and adenoma’, *Gut*, 49, pp. 402–408. doi: 10.1136/gut.49.3.402.

Tibshirani, R. (1996) ‘Regression Shrinkage and Selection Via the Lasso’, *Journal of the Royal Statistical Society: Series B (Methodological)*. Wiley, 58(1), pp. 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R. (1997) ‘The lasso method for variable selection in the cox model’, *Statistics in Medicine*, 16(4), pp. 385–395. doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

Tilastokeskus (2020) *Ikävakioitu kuolleisuusluku | Käsitteet | Tilastokeskus*. Available at: https://www.stat.fi/meta/kas/ikavak_kuoll.html (Accessed: 28 July 2020).

Tipping, M. E. (2000) ‘The relevance vector machine’, *Advances in Neural Information Processing Systems*, (x), pp. 653–658.

Tipping, M. E. (2001) ‘Sparse Bayesian Learning and the Relevance Vector Machine’, *Journal of Machine Learning Research*, 1, pp. 211–244. doi: 10.1162/15324430152748236.

Tong, E. N. C., Mues, C. and Thomas, L. C. (2012) ‘Mixture cure models in credit scoring: If and when borrowers default’, *European Journal of Operational Research*, 218(1), pp. 132–139. doi: 10.1016/j.ejor.2011.10.007.

Topaloglu, Z. and Yildirim, Y. (2009) *Bankruptcy Prediction*. doi: 10.2139/ssrn.1362077.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J. and Jemal, A. (2015) ‘Global cancer statistics, 2012’, *CA: A Cancer Journal for Clinicians*, 65(2), pp. 87–108. doi: 10.3322/caac.21262.

Tóth, N. and Pataki, B. (2008) ‘Classification confidence weighted majority voting using decision tree classifiers’, *International Journal of Intelligent Computing and Cybernetics*, 1(2), pp. 169–192. doi: 10.1108/17563780810874708.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001) *Missing value estimation methods for DNA microarrays*, *Bioinformatics*. doi: 10.1093/bioinformatics/17.6.520.

Tseng, Y. J., Huang, C. E., Wen, C. N., Lai, P. Y., Wu, M. H., Sun, Y. C., Wang, H. Y. and Lu, J. J. (2019) 'Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies', *International Journal of Medical Informatics*. Elsevier Ireland Ltd, 128, pp. 79–86. doi: 10.1016/j.ijmedinf.2019.05.003.

Tsuji, S., Midorikawa, Y., Takahashi, T., Yagi, K., Takayama, T., Yoshida, K., Sugiyama, Y. and Aburatani, H. (2012) 'Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis', *British Journal of Cancer*, 106, pp. 126–132. doi: 10.1038/bjc.2011.505.

Turgeman, L. and May, J. H. (2016) 'A mixed-ensemble model for hospital readmission', *Artificial Intelligence in Medicine*. Elsevier B.V., 72, pp. 72–82. doi: 10.1016/j.artmed.2016.08.005.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) 'Significance analysis of microarrays applied to the ionizing radiation response', *Proceedings of the National Academy of Sciences of the United States of America*, 98(9), pp. 5116–5121. doi: 10.1073/pnas.091062498.

UCLA (2021) *Testing the proportional hazard assumption in Cox models*. Available at: https://stats.idre.ucla.edu/other/examples/asa2/testing-the-proportional-hazard-assumption-in-cox-models/ (Accessed: 30 July 2021).

UICC (2021a) *TNM History , Evolution and Milestones*. Available at: https://www.uicc.org/sites/main/files/atoms/files/TNM-History-2021.pdf (Accessed: 31 December 2021).

UICC (2021b) *UICC and the TNM Classification of Malignant Tumours, About UICC*. Available at: https://www.uicc.org/who-we-are/about-uicc/uicc-and-tnm-classification-malignant-tumours (Accessed: 30 December 2021).

University of Rochester Medical Center (2021) *Grading and Staging of Cancer - Health Encyclopedia - University of Rochester Medical Center*. Available at: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=85&contentid=p00

554 (Accessed: 14 July 2021).

Vapnik, V. N. and Lerner, A. Y. (1963) ‘Pattern Recognition Using Generalized Portraits’, *Avtomatika i Telemekhanika*, 24(6), pp. 774–780.

Väyrynen, J. P., Vornanen, J., Tervahartiala, T., Sorsa, T., Bloigu, R., Salo, T., Tuomisto, A. and Mäkinen, M. J. (2011) ‘Serum MMP-8 levels increase in colorectal cancer and correlate with disease course and inflammatory properties of primary tumors’, *International Journal of Cancer*, 131(4), pp. e463–e474. doi: 10.1002/ijc.26435.

van Veen, F. and Leijnen, S. (2009) *The Neural Network Zoo - The Asimov Institute*. Available at: https://www.asimovinstitute.org/neural-network-zoo/ (Accessed: 4 February 2021).

Vicente-Dueñas, C., Romero-Camarero, I., Cobaleda, C. and Sánchez-García, I. (2013) ‘Function of oncogenes in cancer development: A changing paradigm’, *EMBO Journal*, 32(11), pp. 1502–1513. doi: 10.1038/emboj.2013.97.

Vieira, S., Pinaya, W. H. L. and Mechelli, A. (2017) ‘Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications’, *Neuroscience and Biobehavioral Reviews*. Elsevier Ltd, 74(January), pp. 58–75. doi: 10.1016/j.neubiorev.2017.01.002.

Vilardell, M., Buxó, M., Clèries, R., Martínez, J. M., Garcia, G., Ameijide, A., Font, R. and Civit, S. (2020) ‘Missing data imputation and synthetic data simulation through modeling graphical probabilistic dependencies between variables (ModGraProDep): An application to breast cancer survival’, *Artificial Intelligence in Medicine*. Elsevier B.V., 107. doi: 10.1016/j.artmed.2020.101875.

Viloria, C. G., Obaya, A. J., Moncada-Pazos, A., Llamazares, M., Astudillo, A., Capellá, G., Cal, S. and López-Otín, C. (2009) ‘Genetic inactivation of ADAMTS15 metalloprotease in human colorectal cancer’, *Cancer research*. Cancer Res, 69(11), pp. 4926–4934. doi: 10.1158/0008-5472.CAN-08-4155.

Waldner, M. J., Foersch, S. and Neurath, M. F. (2012) ‘Interleukin-6--a key regulator of colorectal cancer development’, *International journal of biological sciences*. Int J Biol Sci, 8(9), pp. 1248–1253. doi: 10.7150/IJBS.4614.

Wang, J., Yang, D., Duan, X. H., Ji, J. Z. and Peng, B. (2013) ‘Application of relevance vector

machine in the engine oil wear particle fault diagnosis', in *Proceedings - 2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation, IMSNA 2013*. IEEE, pp. 982–985. doi: 10.1109/IMSNA.2013.6743445.

Wang, K. M., Wang, K. J. and Makond, B. (2020) 'Survivability modelling using Bayesian network for patients with first and secondary primary cancers', *Computer Methods and Programs in Biomedicine*. Elsevier B.V., 196, p. 105686. doi: 10.1016/j.cmpb.2020.105686.

Wang, P., Li, Y. and Reddy, C. K. (2017) 'Machine Learning for Survival Analysis: A Survey', *ACM Computing Surveys*, 1(1), pp. 1–38.

Wang, S., Jia, J., Liu, D., Wang, M., Wang, Z., Li, X., Wang, H., Rui, Y., Liu, Z., Guo, W., Nie, J. and Dai, H. (2019) 'Matrix Metalloproteinase Expressions Play Important role in Prediction of Ovarian Cancer Outcome', *Scientific Reports*. Nature Publishing Group, 9(1), pp. 1–11. doi: 10.1038/s41598-019-47871-5.

Wang, Y., Wang, D., Ye, X., Wang, Yanzhang, Yin, Y. and Jin, Y. (2019) 'A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction', *Information Sciences*. Elsevier Inc., 474, pp. 106–124. doi: 10.1016/j.ins.2018.09.046.

Weber, B. L. (2002) 'Cancer genomics', *Cancer Cell*, 1(1), pp. 37–47. doi: https://doi.org/10.1016/S1535-6108(02)00026-0.

Wei, G. and Schaubel, D. E. (2008) 'Estimating Cumulative Treatment Effects in the Presence of Nonproportional Hazards', *Biometrics*. John Wiley & Sons, Ltd, 64(3), pp. 724–732. doi: 10.1111/J.1541-0420.2007.00947.X.

Wei, H. tang, Guo, E. na, Dong, B. guo and Chen, L. sheng (2015) 'Prognostic and clinical significance of syndecan-1 in colorectal cancer: a meta-analysis', *BMC Gastroenterology*. BioMed Central, 15(1). doi: 10.1186/S12876-015-0383-2.

Wei, J. T., Lin, S. Y., Weng, C. C. and Wu, H. H. (2012) 'A case study of applying LRFM model in market segmentation of a children's dental clinic', *Expert Systems with Applications*. Elsevier Ltd, 39(5), pp. 5529–5533. doi: 10.1016/j.eswa.2011.11.066.

Wesley, D. (1998) 'Life Table Analysis', *Journal of Insurance Medicine*, 30(4), pp. 247–254.

White, I. R., Royston, P. and Wood, A. M. (2011) 'Multiple imputation using chained

equations: Issues and guidance for practice', *Statistics in Medicine*, 30(4), pp. 377–399. doi: 10.1002/sim.4067.

WHO (2019) *WHO Classification of Tumours: Digestive System Tumours*. 5th, Volum edn. Edited by WHO Classification of Tumours Editorial Board.

Witten, D. M. and Tibshirani, R. (2008) 'Testing significance of features by lassoed principal components', *Annals of Applied Statistics*, 2(3), pp. 986–1012. doi: 10.1214/08-AOAS182.

Witten, D. M. and Tibshirani, R. (2010) 'Survival analysis with high-dimensional covariates', *Statistical Methods in Medical Research*. SAGE Publications Ltd, pp. 29–51. doi: 10.1177/0962280209105024.

World Health Organization: Regional Office for Europe (2020) *WORLD CANCER REPORT : cancer research for cancer development.* Lyon: IARC.

Wu, K., Feskanich, D., Fuchs, C. S., Willett, W. C., Hollis, B. W. and Giovannucci, E. L. (2007) 'A Nested Case-Control Study of Plasma 25-Hydroxyvitamin D Concentrations and Risk of Colorectal Cancer', *Journal of the National Cancer Institute*, 99(14), pp. 1120–1129. doi: 10.1093/jnci/djm038.

Xie, L., Song, X., Lin, H., Chen, Z., Li, Q., Guo, T., Xu, T., Su, T., Xu, M., Chang, X., Wang, L. K., Liang, B. and Huang, D. (2019) 'Aberrant activation of CYR61 enhancers in colorectal cancer development', *Journal of Experimental and Clinical Cancer Research*. Journal of Experimental & Clinical Cancer Research, 38(1), pp. 1–16. doi: 10.1186/s13046-019-1217-9.

Xu, Y., Ju, L., Tong, J., Zhou, C. M. and Yang, J. J. (2020) 'Machine Learning Algorithms for Predicting the Recurrence of Stage IV Colorectal Cancer After Tumor Resection', *Scientific Reports*. Springer US, 10(1), pp. 1–9. doi: 10.1038/s41598-020-59115-y.

Yaeger, R., Chatila, W. K., Lipsyc, M. D., Hechtman, J. F., Cercek, A., Sanchez-Vega, F., Jayakumaran, G., Middha, S., Zehir, A., Donoghue, M. T. A., You, D., Viale, A., Kemeny, N., Segal, N. H., Stadler, Z. K., Varghese, A. M., Kundra, R., Gao, J., Syed, A., Hyman, D. M., Vakiani, E., Rosen, N., Taylor, B. S., Ladanyi, M., Berger, M. F., Solit, D. B., Shia, J., Saltz, L. and Schultz, N. (2018) 'Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer', *Cancer Cell*. Cell Press, 33(1), pp. 125-136.e3. doi: 10.1016/j.ccell.2017.12.004.

Yao, J., Zhu, X., Zhu, F. and Huang, J. (2017) ‘Deep Correlational Learning for Survival Prediction from Multi-modality Data’, in Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D., and Duchesne, S. (eds) *Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science, vol 10434*. Springer, Cham, pp. 406–414. doi: 10.1007/978-3-319-66185-8_46.

Yasaka, K. and Abe, O. (2018) ‘Deep learning and artificial intelligence in radiology: Current applications and future directions’, *PLoS Medicine*. Public Library of Science. doi: 10.1371/journal.pmed.1002707.

Yavari, P., Abadi, A., Dehghani-Arani, M., Alavi-Majd, H., Ghasemi, E., Amanpour, F. and Bajdik, C. (2014) ‘Cox Models Survival Analysis Based on Breast Cancer Treatments’, *Iran Journal of Cancer Prevention*, 3(Summer), pp. 124–129.

Youden, W. J. (1950) ‘Index for rating diagnostic tests’, *Cancer*, 3(1), pp. 32–35. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Yu, Q. and Stamenkovic, I. (2000) ‘Cell surface-localized matrix metalloproteinase-9 proteolytically activates TGF-β and promotes tumor invasion and angiogenesis’, *Genes and Development*. Cold Spring Harbor Laboratory Press, 14(2), pp. 163–176. doi: 10.1101/gad.14.2.163.

Zhang, Z., Sinha, S., Maiti, T. and Shipp, E. (2018) ‘Bayesian variable selection in the AFT model with an application to the SEER breast cancer data’, *Statistical Methods in Medical Research*, 27(4), pp. 971–990. doi: 10.1177/0962280215626947.

Zhu, Y., Li, L. and Huang, X. (2019) *Landmark linear transformation model for dynamic prediction with application to a longitudinal cohort study of chronic disease, Journal of the Royal Statistical Society Series C: Applied Statistics*.

Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R. and Bratko, I. (2000) ‘Machine learning for survival analysis: a case study on recurrence of prostate cancer’, *Artificial Intelligence in Medicine*, 20(1), pp. 59–75.

Appendices

APPENDIX 1. Literature review summary.

Authors	Year	Article	Publication	Data	Cancer type	Output	Time horizon	Imputation	Variables	Feature selection	Method (ML)	Split ratio (train / test)	Performance metrics	Cross validati	Results
Wang, Wang, Ye, Wang, Yin, Jin	2019	A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction	Information Sciences	advanced-stage CRC patients	CRC					for the 2nd stage of prediction of survival time, a priori knowledge as	1st tree-based imbalanced ensemble classification method, 2nd selective ensemble regression				
Barsasma, Sairam & Patil	2018	Analysis and prediction of survival after colorectal chemotherapy using machine learning models	2018 International Conference on Advantages in Computing, Communications and Informatics	CRC patients after diagnosis and chemo	CRC			removal of NAs	event type (prediction, recurrence, death), gender, age, number of lymph nodes attacked, differentiation of	informatio n gain & Gini index (with DT)	Decision tree: CART with RPART, KNN, LDA	80/20	accuracy, Kappa values	10-fold	CART best with feature selection of large datasets, KNN and LDA worse
Nakatsu, Zhou et al.	2018	Alterations in enteric virome are associated with colorectal cancer and survival	Gastroenterology	74 CRC patients & 92 controls	CRC					parallel random forests model-based backward	random survival forest (RSF), multivariate CPH regression		AUC, median out-of-bag concordance index	10-fold	
Tsuji, Midorikawa, Takahashi, Takayama, Yoshida, Sugiyama & Aburatani	2012	Potential responders to FOLFOX therapy for colorectal cancer by random forests	British Journal of Cancer	83 CRC patients, gene expression data	CRC					Random forests	Random forests	65/35	sensitivity, specificity, accuracy		After removal of outliers accuracy over 90% achieved. Goal is to identify classifier genes
Bjarnadottir, Anderson, Zia & Rhoads	2018	Predicting colorectal cancer mortality: Models to facilitate patient-physician	Production and Operations Management (2)	CRC patients, clinical and demographic info	CRC		30-day and n-year survival/ mortality (n = [0.5:0.5:5	observations with missing values discarded	top 5: if tumor is stage IV, if cancer has spread to liver, if patient requested resuscitation, if patient disposition to a non-standard	forward selection (percentage of improvement, adjusted	classification trees with GUIDE and relaxed regularized logistic regression	80/20 randomized	ROC, AUC		Relaxed lasso had the highest AUC (0.88 - 0.96) with 162 variables. Old age, advaced stage cancer

Authors	Year	Article	Publication	Data	Cancer type	Output	Time horizon	Imputation	Variables	Feature selection	Method (ML)	Split ratio (train / test)	Performance metrics	Cross validati	Results
Anand, Smith, Hamilton, Anand, Hughes, Bartels	1999	An evaluation of intelligent prognostic systems for colorectal cancer	Artificial Intelligence in Medicine	216 CRC patients with censored observations	CRC	lenght of survival after diagnosis		observations with missing values discarded	15 clinico-pathological features, top 3 with Cox's: Dukes stage, age, fibrosis with configuration marginal	information measure (regression trees), backpropagation (NN), GAs for attribute	KNN with GAs, NN (multiplayered backpropagation), regression trees, Cox's regression		mean absolute error	10-fold	ANNs accurate but poor perspicuity, regression trees high perspicuity but poor accuracy. Usage of GAs with KNN and
Burke, Goodman, Rosen, Henson, Weinstein, Harrell Jr, Marks, Winchester & Bostwick	1997	Artificial neural networks improve the accuracy of cancer survival prediction	Cancer	CRC patients	CRC & BC		5-year, 10-year	observations with missing values discarded	TNM prognostic factors + additional commonly collected demographic and anatomic variables, for CRC: age, race, gender, signs & symptoms, diagnostic and extent-of-disease tests, primary site of	backpropagation, maximum likelihood, gradient descent	ANN, TNM staging system	62/38	ROC, AUC		ANNs more accurate than TNM staging system
Grumett, Snow & Kerr	2003	Neural networks in the prediction of survival in patients with	Clinical Colorectal Cancer	pathological data of 403 CRC patients	CRC	disease recurrence	5-year		top4 for NN: vascular invasion, infiltrative growth pattern, tumor site, sex	Kaplan-Meier and log-rank significance test for	logistic regression, NN (MLP)		ROC, AUC, sensitivity, specificity		NN (ROC 78%) outperforms logistic regression
Bottaci, Drew, Hartley, Hadfiled, Farouk, Lee, Macintyre, Duthie & Monson	1997	Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions	The Lancet	334 CRC patients	CRC		9, 12, 15, 18, 21 & 24 months		42 clinocopathological variables	backpropagation	ANN (multilayer feedforward)	85/15	sensitivity, specificity, overall accuracy, positive predictive value, negative predictive value		ANN achieved higher overall accuracy (90%) than colorectal surgeons (79%, 75%)

Authors	Year	Article	Publication	Data	Cancer type	Output	Time horizon	Imputation	Variables	Feature selection	Method (ML)	Split ratio (train / test)	Performance metrics	Cross validati	Results
van Stiphout, Postma, Valentini & Lambin	2010	The contribution of machine learning to predicting cancer		1,552 rectal cancer patients	rectal	local recurrenc e, metastase s, survival	5-year	substitution using average or most common value	age, gender, distance, surgery type, surgery group, residual, tumor stage, nodal stage, overall pathological	exhaustiv e feature search, univariate analysis	logistic regresssion, proximal support vector machine		AUC, ROC	10-fold	Both models perform equally well, accuracy aroung 75% on test data.
Snow, Kerr, Brandt & Rodvold	2001	Neural network and regression predictions of 5-year survival colon carcinoma treatment	Cancer (Conference on Prognostic Factors and Staging in Cancer Management: Contributions of Artificial Neural Networks and Other	approx. 375,000 colon carcinom a patients	colon	5-year survival after primary treatment	5-year		sex, age, no of poditive regional lymph nodes, no of regional nodes examined, TNM, AJCC, residual tumor, surgery	sensitivity analysis for ANN	ANN (MLP), logistic regression	75/25 randomized	ROC, sensitivity, specificity		ANN (ROC = 87.6%) outperformed logistic regression (ROC = 82%)
Li & Razzaghi	2019	Personalized colorectal cancer survivability prediction with machine learning methods	2018 IEEE International Conference on Big Data	CRC patients	CRC		2-year	multivariate imputation by chained equations (MICE)	marital status, sex, primary site, histology, behavior, grade, diagnostic confirmation, extension, lymph nodes, metastasis, tumor size, node evaluation, metastasis	literature review	logistic regression, NN, random forest, AdaBoost, imbalanced classification		ROC, AUC, G-mean, sensitivity, specificity	5-fold	All methods except logistic regression achieved higher AUC with single-ethnicity population than with mixed (approx. 0.85)
Al-Bahrani, Agrawal & Choudhary	2017	Survivability prediction of colon cancer patients using neural networks	Health Informatics Journal	188,336 colon patients from SEER dataset	colon	survival and conditiona l survival	1-, 2- & 5-year		marital status at diagnoses, race, year of birth, birth place, grade, diagnostic confirmation, EOD 10-extension, EOD 10-lymph nodes examined, RX	scikit-learn	Deep NN, random forest, logistic regression	50/30/20 (train/validation/te st)	ROC, positive predictive value, negative predictive value, sensitivity, specificity		DNN able to achieve high AUC (0.8616), high PPV (95.09%) but low NPV (248.29%). For conditional survival
Xu, Ju, Tong, Zhou & Yang	2020	Machine learning algorithms for predicting the recurrence of stage IV	Scientific Reports	999 stage IV CRC patients	CRC	risk of cancer recurrenc e		multiple interpolation	top 8: CT, age, anesthesia time, LogCEA, CEA, ASA, RT, AJCC	scikit-learn	logistic regression, decision tree, GradientBoos ting, lightGBM	80/20	F1 score, AUC, ROC, sensitivity, specificity, accuracy, precision		GB and gbm outperformed other two models. Top 5 for GB for increasing the

Authors	Year	Article	Publication	Data	Cancer type	Output	Time horizon	Imputation	Variables	Feature selection	Method (ML)	Split ratio (train / test)	Performance metrics	Cross validati	Results
Ramos, Montaño, Esteva, Barceló & Franch	2016	Colorectal cancer survival by stage of cases diagnosed in	Cancer Epidemiology	2889 CRC patients	CRC	lenght of survival after diagnosis		multiple imputation (MI)	sex, age, diagnostic method, site, histology, TNM, stage, date of diagnosis, date of	extented Cox	actuarial, Kaplan-Meier		log-rank test		Worse survival for patients with advanced age, mucinous histology and
Murtojärvi, Halkola, Airola, Laajala, Mirtti, Aittokallio & Pahikkala,	2020	Cost-effective survival prediction for patients with advanced prostate cancer using clinical trial and real-	International Journal of Medical Informatics		mCRPC			observations with missing values discarded		greedy cost-specified variable selection algorithm, LASSO	penalized Cox regression with LASSO		concordance index		LASSO performed bettter for the variable selection than greedy algorithm
Van Belle, Pelckmans, Van Huffel & Suykens,	2011	Support Vector Methods for Survival Analysis: A Comparison Between	Artificial Intelligence in Medicine		leukaemia, lung cancer, prostate cancer, BC						Cox, 2 SVM methods: RANKSVM C & SVCR		concordance index, log rank test statistics, normalized hazard ratio	10-fold	SVM models with regression contraints showed better performance on high
Kleinlein & Riaño	2019	Persistence of data-driven knowledge to predict breast	Internationa Journal of Cancer		BC						naïve Bayes, logistic regression, decision trees			10-fold & 5-fold	Joint and stage-specific logistic regression showed the
Reijnen, Gogou, Visser et al.	2020	Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development	Plos Medicine	1,1999 patients	endometrial	LNM & 5-year DSS	5-year	multiple imputation (MI)	mylmetrial invasion, estrogen receptor, progesterone receptor, L1CAM, p53, atypical endometrial in cervical cytology, postoperative grade, lymph node	hill-climbing and Tabu search algorithm & literature review	Bayesian network		Brier score, decision analysis curves, calibration, AUC		AUC 0.82 ja Brier score 0.09 for LNM, and AUC 0.82 and Brier score 0.12 for 5-year DSS
Delen, Walker & Kadam	2005	Predicting breast cancer survivability; a comparison of	Artificial Intelligence in Medicine		BC			observations with missing values discarded	From the sensitivity analysis for ANN, the most important input variables		ANN, decision trees, logistic regression			10-fold	Measured in prediction accuracy the ANN and

Authors	Year	Article	Publication	Data	Cancer type	Output	Time horizon	Imputation	Variables	Feature selection	Method (ML)	Split ratio (train / test)	Performance metrics	Cross validati	Results
Ryu, Chandrasekaran & Jacob	2004	Prognosis using an isotonic prediction	Management Science		BC					backward sequential	isotonic prediction				Classification accuracy of around 90%
Zupan, Demšar, Kattan, Beck & Bratko	2000	Machine learning for survival analysis: a case study on recurrence of	Artificial Intelligence in Medicine		prostate						CPH, naïve Bayes, decision trees		Concordance index, sensitivity, specificity	10-fold	Decision trees found to underperform CPH and NB
Tseng , Huang, Wen, Lai, Wu, Sun, Wang & Lu	2019	Predicting breast cancer metastasis by using serum biomarkers and clinocopathological data with	International Journal of Medical Informatics		BC		30-, 60- and 90-day prognosis			mean decrease in Gini index and the amount of time a	random survival forest (RSF), logistic regression, naïve Bayes, SVM		ROC, AUC, Youden index	3-fold inner	RSF achieved best predictive performance whilst the logistic regression performed
Jerez-Aragonés, Gómez-Ruiz, Ramos-Jiménez & Alba-Conejo	2003	A combined neural network and decision trees model for prognosis of breast cancer	Artificial Intelligence in Medicine		BC	risk for relapse after surgery				CIDIM	decision trees, ANN (MLP), Cox regression		ROC, AUC, sensitivity, specificity	10-fold	NN seemed to slightly outperform Cox
Que, Chen et al.	2019	Application of preoperative artificial neural network based on blood biomarkers and	World Journal of Gastroenterology	clinicopathological data of 1,608 GC patients	GC	3-year survival and the probability of that estimate	3-year survival	observations with missing values discarded	age, sex, BMI, ASA score, tumor location, tumor size, PNI, AGR, NLR, platelet-lymphocyte ratio, TNM stage,	logistic univariate analysis	ANN (MLP)	70/30	concordance index, ROC, AUC, likelihood ratio chi-square, AIC	5-fold	ANN outperformed both clinical TNM and pathological TNM staging
Wang, Wang & Makond	2020	Survivability modelling using Bayesian network for patients with first and secondary	Computer Methods and Programs in Biomedicine	7,845 patients	several	survival of patients with 2 cancers	5-year		age, gender, FPC, primary FPC treatment, SPC, primary SPC treatment	literature review	Bayesian network, BPNN, logistic regression, SVM, Naïve Bayes		sensitivity, accuracy, specificity	5-fold	BN outperforms NN, logistic regression, SVM and NB

APPENDIX 2. Support vector machines.

This section presents the concept of support vector machines (hereinafter SVM) and few versions, i.e. SVM that formulates survival as a ranking problem (RANKSVMC) (Van Belle *et al.*, 2007) SVM for censored data support vector classification regression (hereinafter SVCR) (Shivaswamy, Chu and Jansche, 2007), ranking-based linear survival SVM (hereinafter SSVM) (Pölsterl, Navab and Katouzian, 2015), and a combination of two SVMs together with boosting ANNs Confidence-Weighted Voting Boosting Artificial Neural Network Support Vector Machine (hereinafter CWV-BANN-SVM) (Abdar and Makarenkov, 2019), developed specifically for survival analysis. The short insight into relevance vector machines (Tipping, 2001) and fuzzy-SVMs (Lin and Wang, 2002) concludes this section.

Origins of the idea of support vector machines date back to 1963 when Vapnik and Lerner (1963) proposed a novel classification method based on pattern recognition called generalised portrait method. Later Boser, Vapnik and Guyon (1992) proposed a technique suitable for SVMs to utilize kernel tricks to maximum margin hyperplanes to invert the non-linear classification task to linear classification problem in high-dimensional feature space. Their solution maximizes the margin from the decision boundary, i.e. hyperplane, whilst minimizing the classification error. SVMs can be roughly divided into two categories, the ones with classification function and the ones with regression function. First ones categorize data, and the latter predict the numerical value of the output (Delen, Oztekin and Kong, 2010). In the context of survival analysis with Kernel-based models the outcome of the model is the prognostic index (Vanya Van Belle *et al.*, 2011).

Figure 31 demonstrates SVM with classification into two separate classes. Now, the support vectors are those observations in each class closest to the hyperplane. In Figure 31 these are represented as circled shapes. The algorithm finds a decision function, represented as a hyperplane, determining into which class each observation belongs. In the case of overlapping classes, additional cost term using slack variables allowing misclassification only when absolutely necessary can be introduced to the model (Theodoridis and Koutroumbas, 2009).

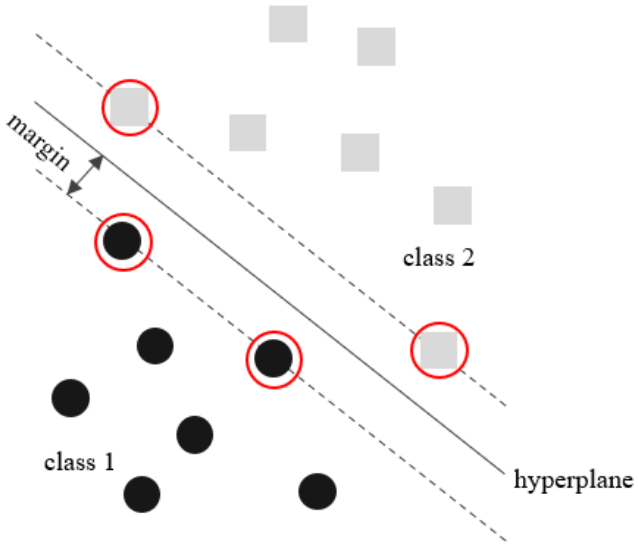


Figure 31. SVM classifier (adapted from (Boser, Vapnik and Guyon, 1992; Faria et al., 2014)).

Now, the weights of the features are computed using an optimization-based approach. The weights are minimized to maximize the margin (Theodoridis and Koutroumbas, 2009). If the optimization is constrained, Lagrange multipliers can be applied for this. The choice of the kernel function is considered to be the biggest downfall of the SVM approach (Abdel-Zaher and Eldeib, 2016). Furthermore, the procedure demands high computational abilities given the complex algorithms requiring substantial memory capacity (Bishop and Tipping, 2003).

SVM for survival analysis

Typically with survival data arises an issue to handle censored and missing data, e.g. time to event is missing when the individual is non-diseased at the end of the observation period. This creates demand for more tailored methods for conducting the actual analysis. Literature presents two SVM-based approaches for censored data (V. Van Belle *et al.*, 2011). These approaches are regression approach (Shivaswamy, Chu and Jansche, 2007), and rephrasing the problem

using the concordance index to a ranking problem (Van Belle *et al.*, 2007). Van Belle et al. (2011) proposed a novel SVM model which combines these two strategies.

Shivasmy et al. (2007) proposed a novel SVM techniques specifically designed for censored data called SVCR. Their approach utilises regression and is developed based on the baseline concept of support vector regression (hereinafter SVR). SVR is unable to process censored data which generated demand for a more sophisticated solution that could also be able to handle censored observations. Fundamentally, SVCR combines the methodology of SVR and support vector classification, SVC. Algorithm formulation of SVCR leads to a standard quadratic programming problem making this approach less computationally intensive. Compared to other ranking-based approaches SVCR is shown to evidence high generalization capabilities.

During the same year Van Belle et al. (2007) submitted a paper suggesting a novel ranking-based SVM approach for censored data. The model utilises so called health index and concordance index in determination of each individual's risk by ranking. With this method each comparable pair with the order in prognostic score differing from the observed order is exposed to penalization. Now, the concept of comparable pairs differs from the permissible pairs which will be discussed later in section 2.5.4 concerning Harrell's C discrimination index as a measure of model's predictive accuracy. Data pair is considered to be comparable if the order of the event times is known, e.g. in the case of right-censored observation, the other observation of the pairs experiences an event prior to the censoring time of the other one. Thus, it is possible to compare the observations forming the pair. Formally, the concept of comparable pairs is displayed as follows where the pair consists of observations i and j for which the event times are represented using δ_i and δ_j respectively (see formula ***(40)*** (Vanya Van Belle et al., 2011)). RANKSVMC suffers from being computationally intensive. A possible solution for this could be alter to formulations to be more consistent with least squares SVM which allows the problem to be formulated as a linear system instead of a quadratic programming problem (Suykens and Vanderwalle, 1998). Further details about this modified version of SVM are left outside the scope of this thesis. (Van Belle *et al.*, 2007; Vanya Van Belle *et al.*, 2011)

$$comp(i,j) = \begin{cases} 1 & if\ \delta_i = 1\ and\ \delta_i = 1 \\ & \delta_i = 1\ and\ \delta_j = 0\ and\ y_i \leq y_j \\ 0 & otherwise \end{cases} \quad (40)$$

Van Belle et al. (2011) compared ranking and regression based SVM techniques for survival analysis for several real-world data sets. These data sets consisted of leukaemia, lung cancer, prostate cancer, and breast cancer patients. Cox model was selected as a benchmark and compared with two SVM approaches for censored data. As mentioned above one possible way is to rephrase the task as a ranking problem (RANKSVMC) and the other is to perform regression (SVCR). They also proposed a new model similar to RANKSVMC including regression constraints. Their solution applies Lipshitz constant for regularisation instead of maximal margin. The models with regression constraints showed better performance on high dimensional data compared to SVM-based models with ranking constraints. As a possible explanation for this advantageous performance Van Belle et al. (Vanya Van Belle *et al.*, 2011) suggest the similarities to CPH. Evers and Messow (2008) also emphasize the similarities between CPH and ranking-based SVM methods, especially with their loss functions.

Another ranking-based SVM was introduced by Pölsterl et al. (2015). Their ranking-based linear SVM was designed specifically for survival analysis, hence the name Survival SVM (hereinafter SSVM). The method ranks the observations based on their predicted survival time. To improve this ranking, order statistic trees can be applied. The target is to minimize the objective function whilst accounting for the possible presence of right-censoring. Further details about this objective function and the computations of the support vectors are left outside of the scope of this thesis and can be read from the paper by Pölsterl et al. (2015). As aforementioned with RANKSVMC, not all the data is included into training, only comparable pairs. As a advantage to this approach is its ability to handle multicollinearity in data (Pölsterl *et al.*, 2016).

A year later Pölsterl et al. (2016) conducted survival analysis on breast cancer, coronary artery disease and Framingham offspring for coronary vessel disease datasets using different methods. They compared the performance of CPH with ridge penalty (ℓ_2) as a benchmark to CPH with

LASSO (ℓ_1), SSVM, random survival forests, and gradient boosted Cox with randomized regression trees and component-wise least squares. Depending on the dataset, non-linear survival models with embedded feature selection, e.g. random survival forests, seemed to achieve highest performance. As a conclusion, the selected method for survival analysis and feature selection ought to be done based on the characteristics of the given data. (Pölsterl, Navab and Katouzian, 2015; Pölsterl *et al.*, 2016)

Abdar and Makarenkov (2019) applied a combination of two SVMs together with boosting ANNs for an ensemble learning classifier. This novel method is referred as a Confidence-Weighted Voting Boosting Artificial Neural Network Support Vector Machine, CWV-BANN-SVM. Confidence-Weighted Voting (hereinafter CWV) in an ensemble method based on the ranking of class membership probabilities (Tóth and Pataki, 2008; Li *et al.*, 2014). Here CWN was applied to combine two SVMs for the model. Boosting is a ML technique for increasing the accuracy and performance of learning algorithms, e.g. ANNs (Schwenk and Bengio, 2000). A well-known boosting algorithm for ANNs is the AdaBoost introduced by Freund and Schapire (1997).

The ANNs selected for the model are MLP and RBF (radial basis function). The selected SVMs are both polynomial SVM with parameter values selected through training. The structure of the model is presented in Figure 32. They applied their method for diagnosis prediction for breast cancer patient data. Using measures of predictive accuracy, the model was able to outperform other methods including traditional ML algorithms, e.g. MLP, SVM and boosting radial basis function (RBF).

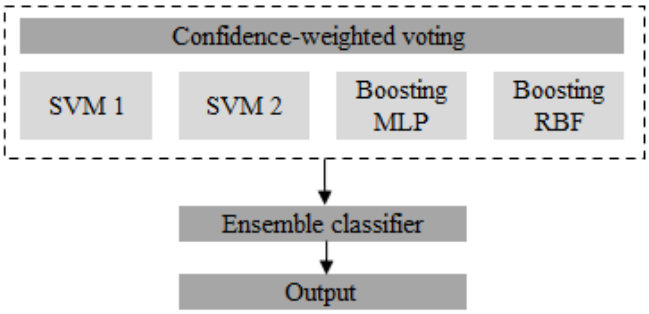


Figure 32. CWV-BANN-SVM (adapted from (Abdar and Makarenkov, 2019)).

Relevance Vector Machine

To overcome issued associated with traditional SVM relevance vector machine (hereinafter RVM) (Wang *et al.*, 2013) was developed by Tipping (2000). RVM is a type of a sparse Bayesian kernel (SBK) model (Kiaee, Sheikhzadeh and Eftekhari Mahabadi, 2016). RVM is a specialisation of a probabilistic Bayesian learning framework for obtaining sparse solutions to regression and classification task. Compared to SVM RVM is able to form prediction using fewer kernel functions (Tipping, 2001). In addition to providing an accurate estimate of the outcome, RVM offers an automated ranking of relevance of the predictors affecting the outcome variable (Ribas *et al.*, 2011). With survival data the predicted outcome variable seldom possesses a feature of being normally distributed, rather it follows an exponential or Weibull distribution, and thus violating the assumptions of standard RVM (Kiaee, Sheikhzadeh and Eftekhari Mahabadi, 2016).

To specifically suit the requirements set by the characteristics of survival data, a version of RVM is proposed by Kiaee et al. (2016) called relevance vector machine survival model (hereinafter RVMS. This RVMS model is based on the concept of Weibull AFT model. AFT, Accelerated Failure Time, model typically utilised for survival analysis or failure time computations due to its capabilities handling censored data. RVMS overtakes the mandatory log-linearity relation assumption of AFT model. Training of RVM model could be accelerated either using an efficient smooth prior or fast marginal likelihood maximization procedure. Former

approach is referred as smooth RVMS and the latter as fast RVMS. These approaches offer a slightly degraded performance than the standard RVMS.

Fuzzy SVM

Fuzzification of SVM might not be as beneficial as with other ML techniques (Hüllermeier, 2015). Lin and Wang (2002) fuzzified the membership calculation of each data point to form a novel approach to traditional SVM. This way the contribution of each point to the formulation of the decision plane could vary independently of each other. Thus making it possible for other data points to be more valuable to the model than the rest. Basically, this allows the possible misclassification of less valuable observations whilst focusing on the accurate classification of more meaningful data points. In practice there are no crisp class designations for each observation rather fuzzy membership degree. This degree, its value varying in the interval [0,1], announces to which degree a specific observation belongs to which class. For example, with binary classification task, an observation could belong to a class A with membership degree 0.80 and to the class B with the membership degree 0.20.

The support vectors are then obtained after solving the optimal hyperplane problem using Lagrangian equations. Differing from the traditional SVMs the type of a support vector, whether on the margin or one being misclassified, depends on the factor, s_i which represents the fuzzy membership degree for observation i. Through free parameter C the tradeoff between the misclassifications and the width of the margin could be altered. (Lin and Wang, 2002) Overall, this approach allows to reduce the effect of outliers and noise in the data (Congqin-Yi, Zhou and Hu, 2017).

Lin and Wang (2002) demonstrated the usage of this fuzzy SVM (hereinafter FSVM) method with different types of datasets. Additionally, there have been applications of this FSVM method in the field of oncology. Congqin-Yi et al. (2017) applied FSVM for identification of breast cancer gene achieving high accuracy of 98.9 %. However, the training for FSVM is

more computationally intensive than the reference method standard SVM. Diagnosis of breast cancer tumour using PCA with FSVM was studied by Luo et al. (2008). Their results evidenced superior predictive performance and accuracy for FSVM compared to standard SVM. The model was able to identify unconventional tumour structures for further examination by the medical experts. Mao et al. (2005) utilised FSVM and binary decision tree for multiclass cancer classification using gene expression data. The datasets applied consisted of patients having breast cancer, acute leukaemia, or small round blue-cell tumours. Their study showed that the FSVM was able to identify the most crucial genes with the highest accuracy compared to the other tested methods.

APPENDIX 3. Measures of predictive accuracy.

YOUDEN INDEX

Youden index named after Youden who proposed this novel index for rating performance of diagnostic tests in 1950. It measures the discriminative success of the test. This index is based on confusion matrix. The function for the calculation of the index is presented in the formula ***(41)*** (Youden, 1950). In the numerator from the product of true positives (TP) and true negatives (TN) is subtracted the product of false positives (FP) and false negatives (FN). The product of the number of diseased patients (a + b) and number of controls (c + d) is the in denominator. The Youden index receives values in the interval [0,1] where value 0 is the worst and value of 1 the best predictive accuracy. The value 0 is obtained when the share of false positives and false negatives is the same predicted by the test which makes the test worthless. The value 1 is obtained if and only if the test predicts perfectly, so there are neither false positives nor false negatives. To compare two separate tests using the Youden index, the means of the t-test of the standard errors of the indexes are taken. (Youden, 1950)

$$J = \frac{ad-bc}{(a+b)(c+d)} = \frac{TP*TN-FN*FP}{(TP+FN)(FP+TN)} \qquad \textbf{(41)}$$

CALIBRATION

Calibration is the other component of assessment of model's predictive accuracy together with discrimination (Harrell, Lee and Mark, 1996; Tong, Mues and Thomas, 2012) which was discussed above. It is used to measure the presence of bias in the predictions (Harrell, Lee and Mark, 1996), and thus describing how well does the predictions align with the true observed events. Model's calibrations compares the predicted number of observations against the observed number (Reijnen *et al.*, 2020). It can be assessed visually through calibration plots (Black, Terry and Lizotte, 2020) in which the predicted outcome is plotted against the observed outcome. Typically this plot is a scatter plot (Harrell, Lee and Mark, 1996).

APPENDIX 4. More detailed descriptions about the treatment of CRC patients in our cohort.

The length of the patients' hospital stay varies between 0 and 45 days (approx. one and a half (1.5) months) with median 11 days and mean of 12.59 days making the distribution of the length of hospital stay slightly right skewed (Figure 33). Patients having possibly incorrect dates for their hospital stay are removed as those could be considered as outliers. The value indicating the length of the hospital stay includes the admission day. The vertical red line demonstrates the median whilst the blue line indicates the mean of the length of patients' hospital stay.

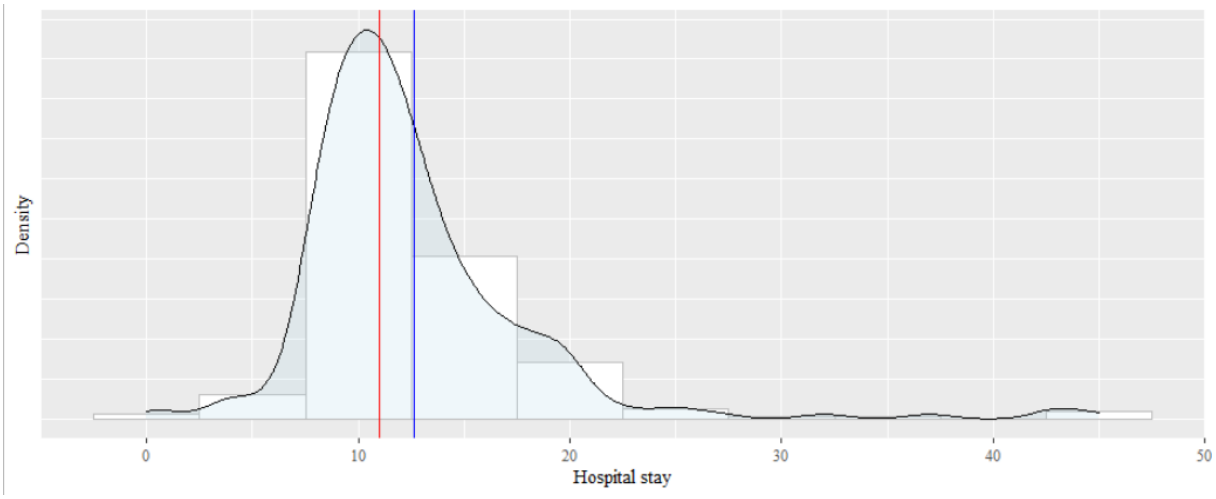


Figure 33. Length of patients' treatment period.

The information about pre-existing diseases is recoded as a binary variable. For 30.5 % (97) of the patients in this dataset there is no pre-existing diseases, and for the remaining 45.0 % (143) patients they have at least one pre-existing disease. For 78 patients (24.5 %) this information is missing. The common diseases within this dataset are hypertension, morbus cordis coronarius (MCC), diabetes mellitus, heart failure, asthma, and Lynch syndrome (HNPCC, hereditary non-polyposis colorectal cancer).

Lynch syndrome, which is referred also as hereditary non-polyposis colorectal cancer (hereinafter HNPCC) and familial adenomatous polyposis (hereinafter FAP) both are hereditary conditions known to predispose patients to CRC typically through highly penetrant mutations (de la Chapelle, 2004). In addition, patients diagnosed with Crohn's disease suffer from an increased risk of CRC diagnosis (Olén *et al.*, 2020). Out of these three (3) conditions FAP (Mirinezhad *et al.*, 2018) and Crohn's (Olén *et al.*, 2020) are shown to increase also the risk of mortality in the case of CRC. Prior any imputations there are six (6) patients with FAP, two (2) with HNPCC and four (4) with Crohn's disease in our data.

Variable describing the location of the tumour has three (3) levels. These levels are in the intestinal system starting from the end of the small intestine as follows colon dextex/the right colon, colon sinister/the left colon and the rectum. Table 4 displays the frequencies of the different tumour locations in our CRC data. The number of rectal tumours (52.5 %) slightly outnumbers tumours located in colon (47.5 %). For the analysis colon locations are combined, so two (2) levels are applied.

The Table 14 demonstrates tumour's histopathologic diagnosis. Basically histopathologic diagnosis is a microscopical study of a tissue sample, and the results are used to determine appropriate diagnosis and required form of treatment. From the table can be observed that the majority of the CRC patients in our dataset have some type of adenocarcinoma, and the rest mucinous carcinoma. There also exists some missing values. Carcinoma adenomatosum or colonic adenocarcinoma is a malignant neoplasm typically arising from epithelial cells (Mullangi and Lekkala, 2021). Carcinoma mucinosum or mucinous carcinoma is another type of adenocarcinoma characterised by the large volume of mucus it produces (Luo *et al.*, 2019). Our data supports the statistics stating adenocarcinoma to be the most prominent type of CRC globally (CTCA, 2018).

Table 14. Histopathological diagnoses of the CRC patient data.

Histopathological diagnosis	Freq	Pct
ca adenomatosum	218	68.6 %
NA	78	24.5 %
ca adenomatosum partim mucinosum	10	3.1 %
ca mucinosum	7	2.2 %
ca adenomatosum mucinosum	5	1.6 %

Operations took place in the interval of 2000 to 2003. Performed operations are generalized to form five (5) categories. These are colectomy (*colectomia*), hemicolectomy (*hemicolectomia*), proctocolectomy (*proctocolectomia*), excision (*excision*), and resection (*resectio*). Table 15 displays for each operations the number and fraction of patients receiving that. The most common type of operations are resection (36.99 %) and hemicolectomy (21.94 %). Here Hartmann's operation is included as a resection type operation.

Table 15. Operations.

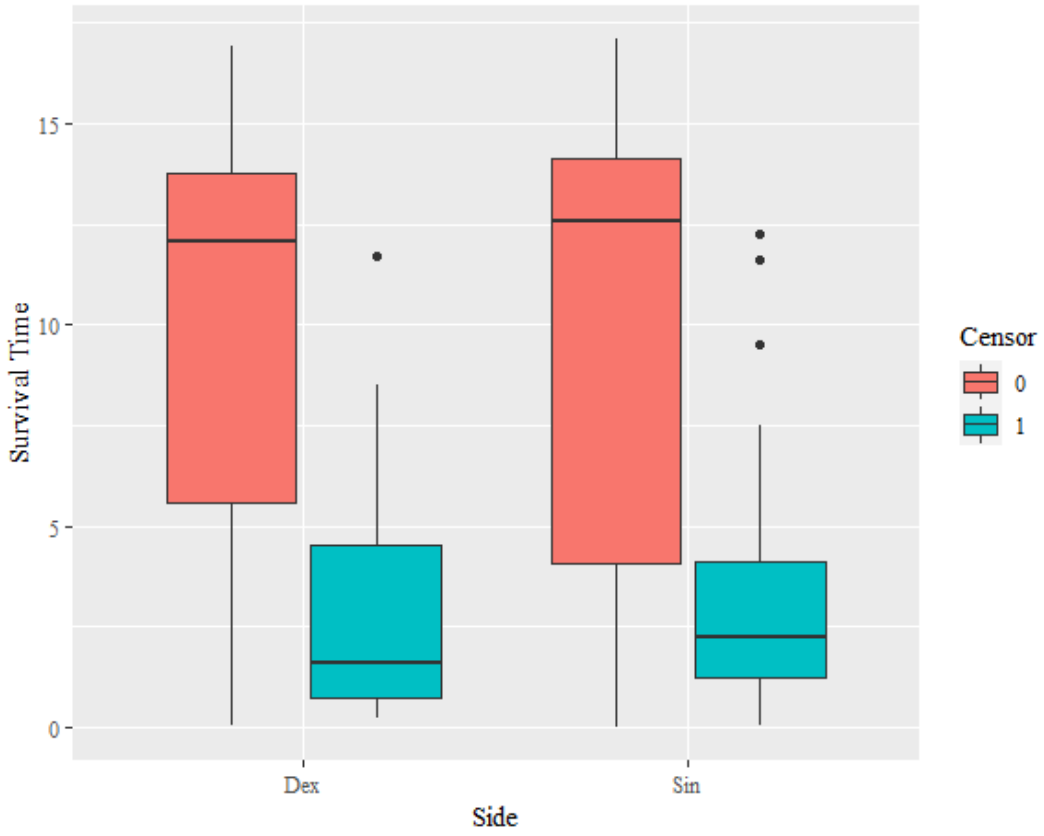
Operation	Freq	Pct
resection	117	36.79 %
NA	78	24.53 %
hemicolectomy	70	22.01 %
excision	26	8.18 %
colectomy	21	6.60 %
proctocolectomy	6	1.89 %

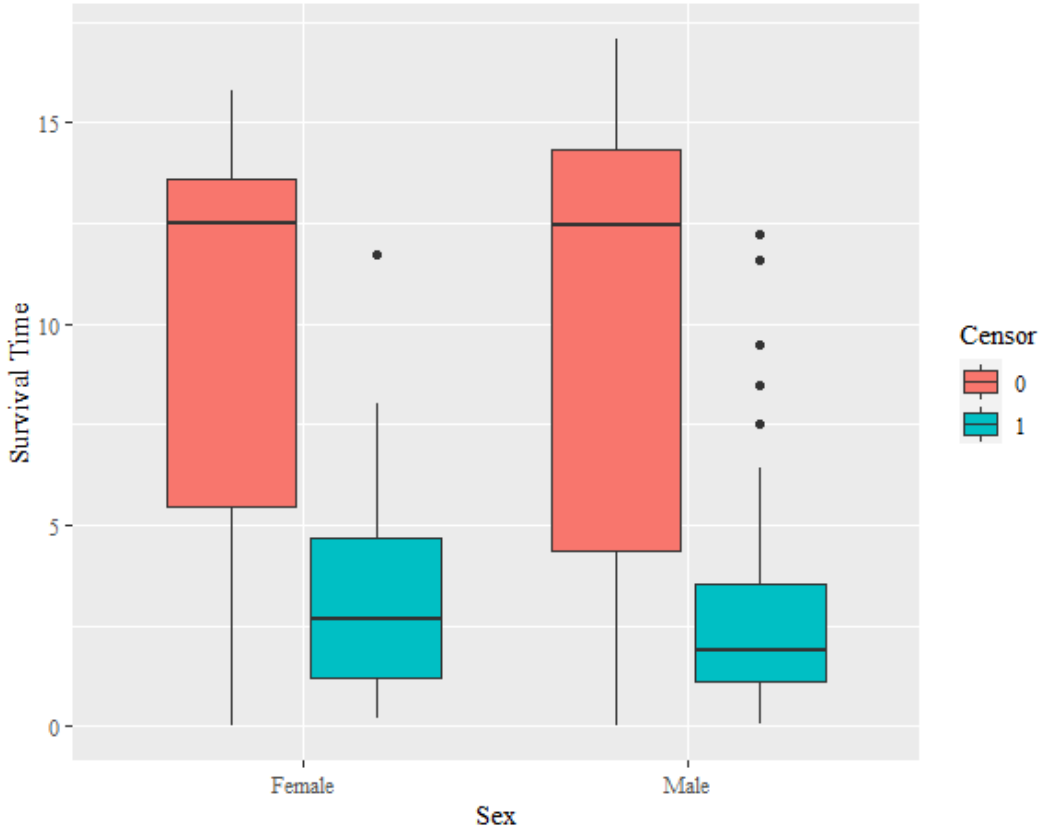
In colectomy all parts of the colon are surgically removed (Mayo Clinic, 2021), whereas in hemicolectomy only part of the colon is removed (Mayo Clinic, 2021), as the prefix hemi suggests. In proctocolectomy colon and rectum are both removed surgically (Crohn's & Colitis Foundation, 2021). Here prefix procto- refers to anus and rectum (Collins English Dictionary, 2021). In excision both the malignant area and some of the surrounding healthy tissue are removed surgically (Rajpar, 2020). Finally, resection is a surgery where malignant tissue is removed (National Cancer Institution, 2021). Differing from excision, resection is more invasive (McBride *et al.*, 2020).

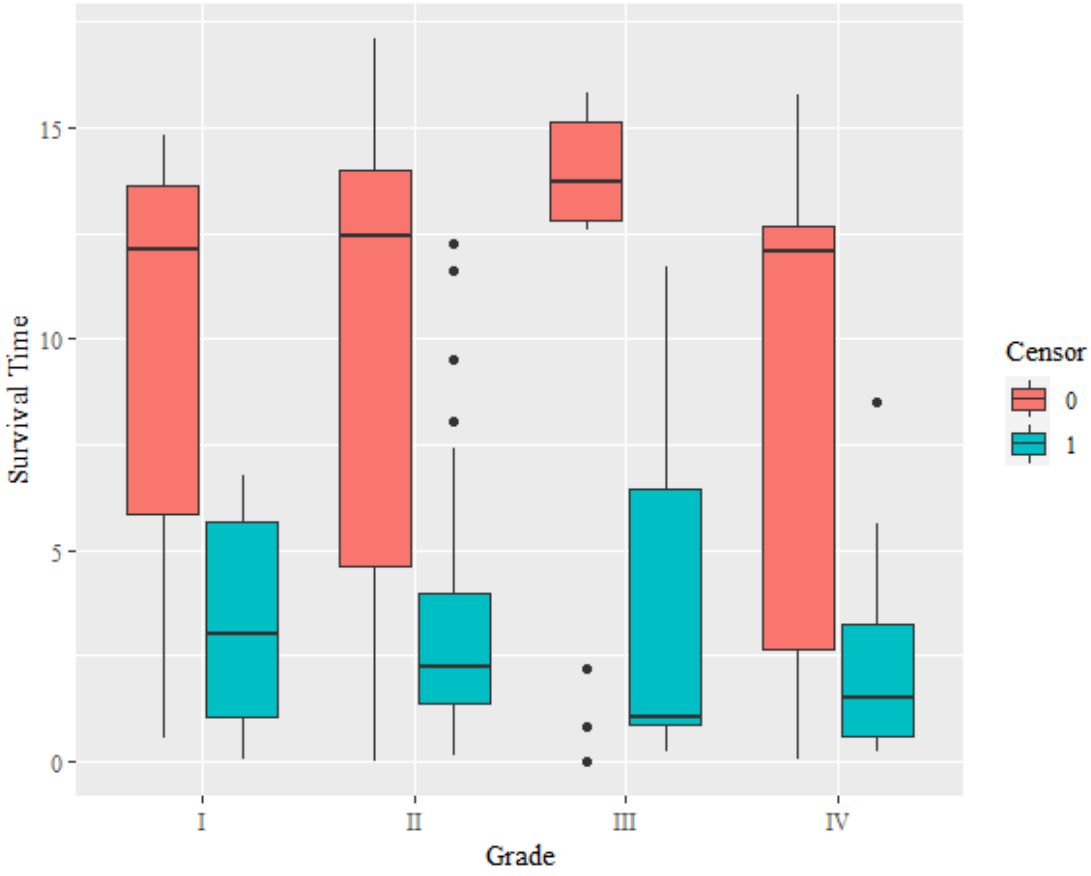
In our CRC data majority (198, 62.3 %) did not experience any complication, only 13.2 % (42) experienced with one or more post-operative complications, and for the rest (78, 24.5 %) this information is missing. The most common type of complication is adhesion occlusion which is an intestinal obstruction caused by adhesions in the intestinal tract. These are typical complications for patients with advanced peritoneal cancer (Saarto, Österlund and Lepistö, 2013). There are a few cases experiencing abscesses, peritonitis, suture leakage, infection of the wound, cardiovascular events, and pneumonia. Rest of the complication were sporadic, and thus further examination of those is left outside the scope of this thesis. Similarly to the pre-existing diseases, the variable for complications is recoded as binary where one reflects occurrence of one or more complications, whilst zero indicates that no complications occurred.

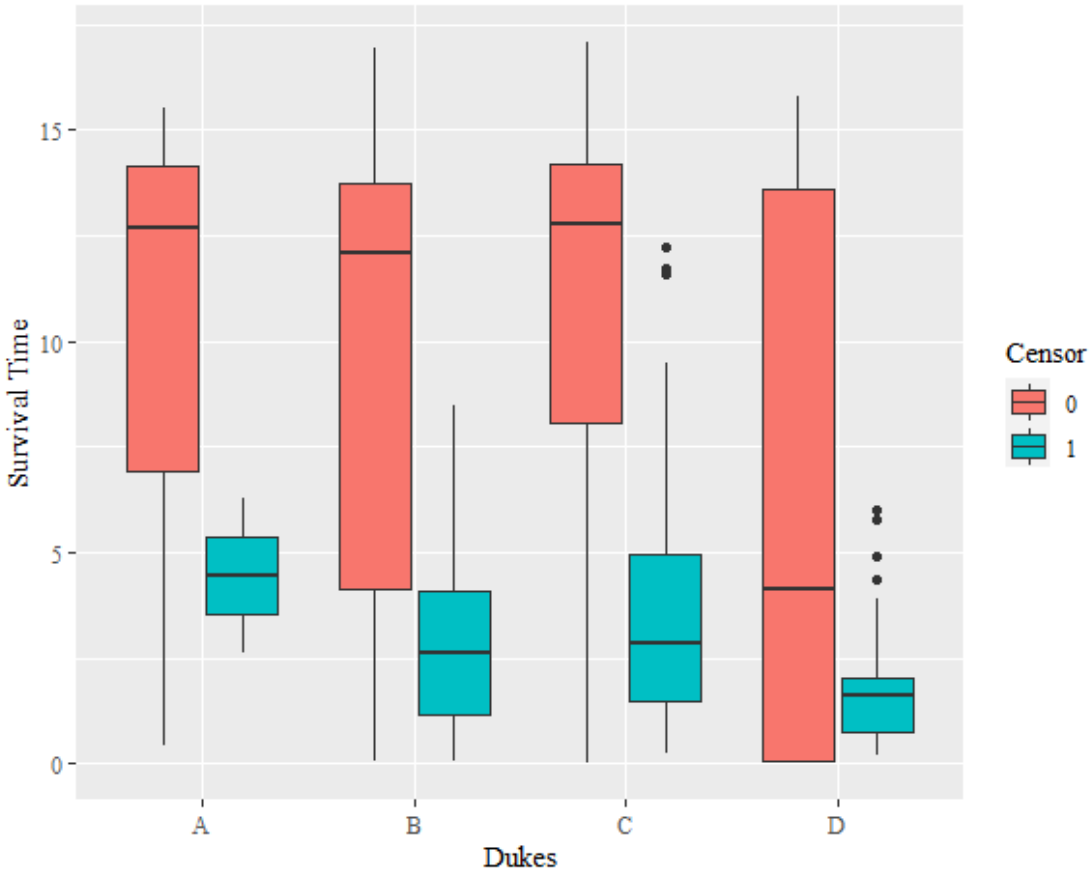
Recurrence of cancer and the more detailed description of the type and stage are excluded from the analysis. Cancer recurred with 57 (17.9 %) patients, 183 (57.5 %) of the patients did not experience recurrence of cancer during the observation period. Similarly, information about the applied chemotherapy and radiotherapy are recoded as binary variables. In total 92 (28.9 %) patients received cytostatic treatment of which mainly post-operative, 143 (45.0 %) did not and for 83 (26.1 %) patients this information is missing. Radiotherapy used typically as a preoperative treatment for 59 (18.6 %) patients, 181 (56.9 %) did not receive radiotherapy, and for 78 (24.5 %) this information is missing. Besides surgery, chemotherapy and radiotherapy are suggested methods of treatment for CRC. Chemotherapy is performed as postoperative treatment for both colon and rectal cancer. Radiotherapy can given as a preoperative treatment for patients with rectal tumours to shrink the tumour and decrease the local recurrence. This is not done for colon cancer patients. (Duodecim, 2019)

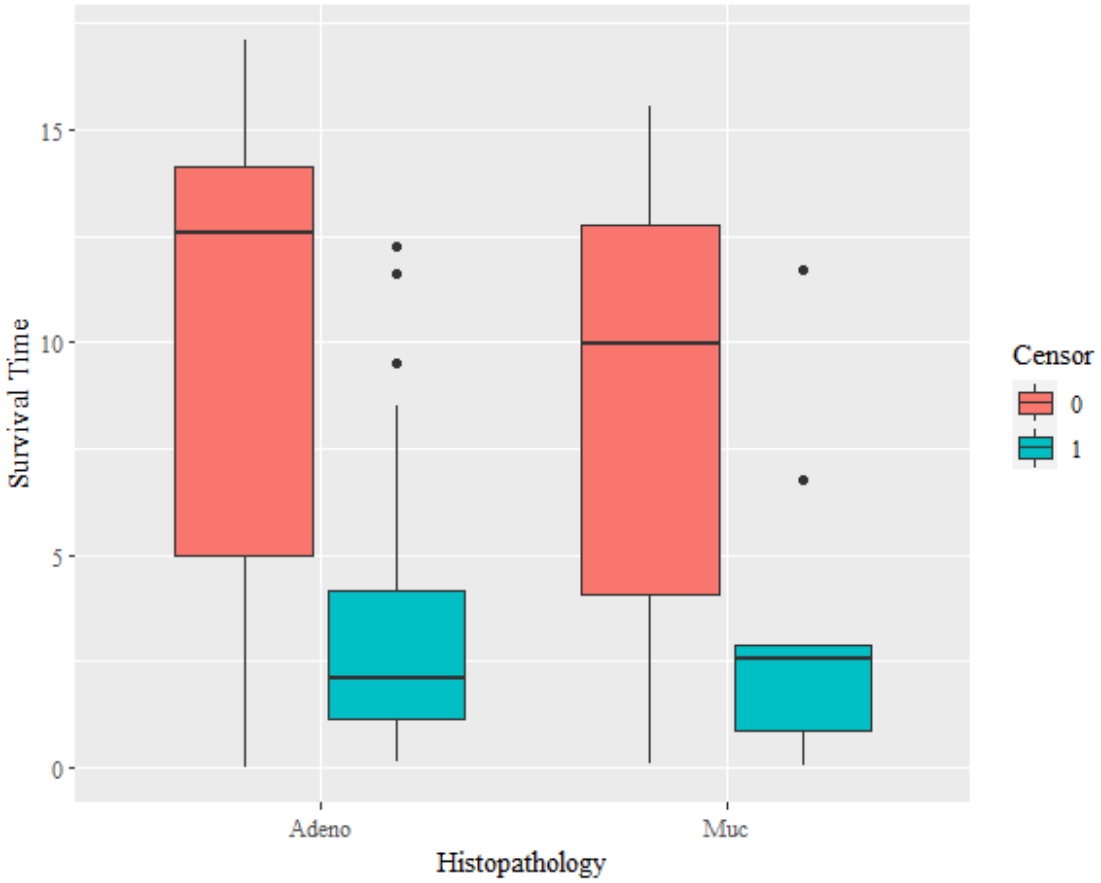
APPENDIX 5. Boxplots of the categorical features.

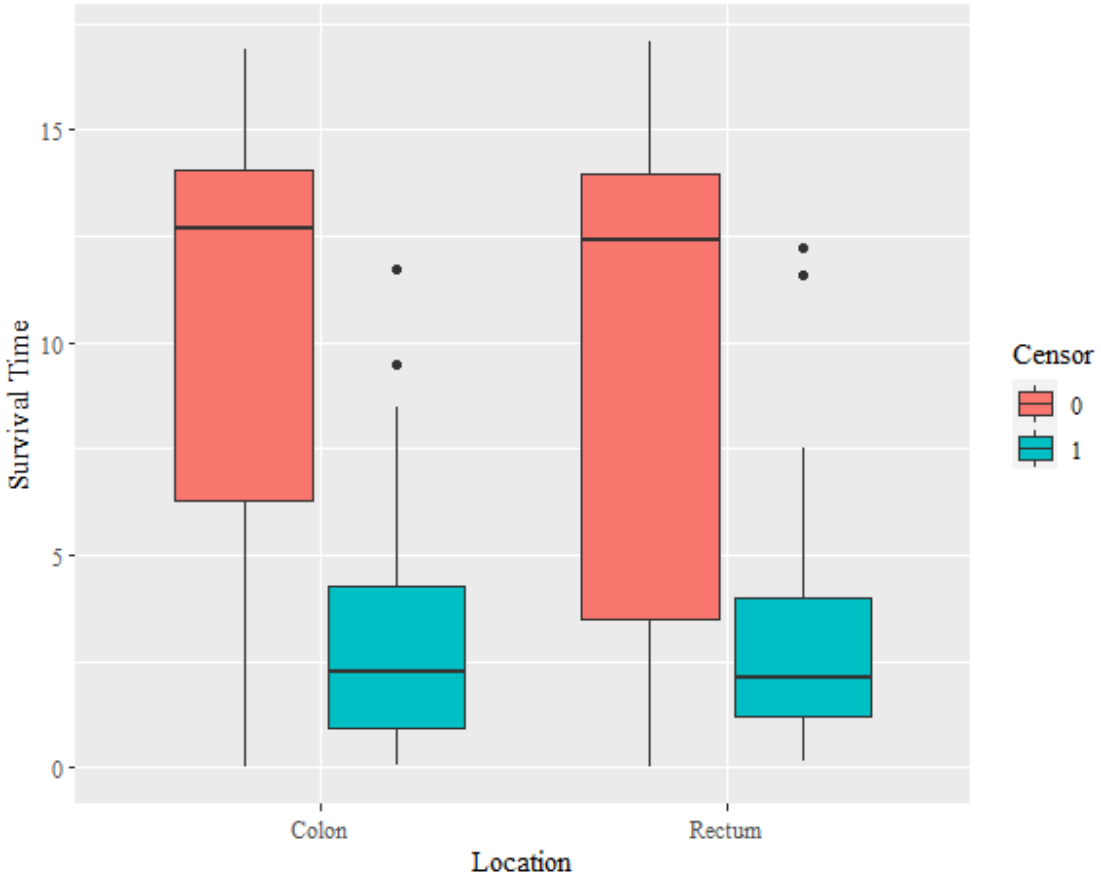




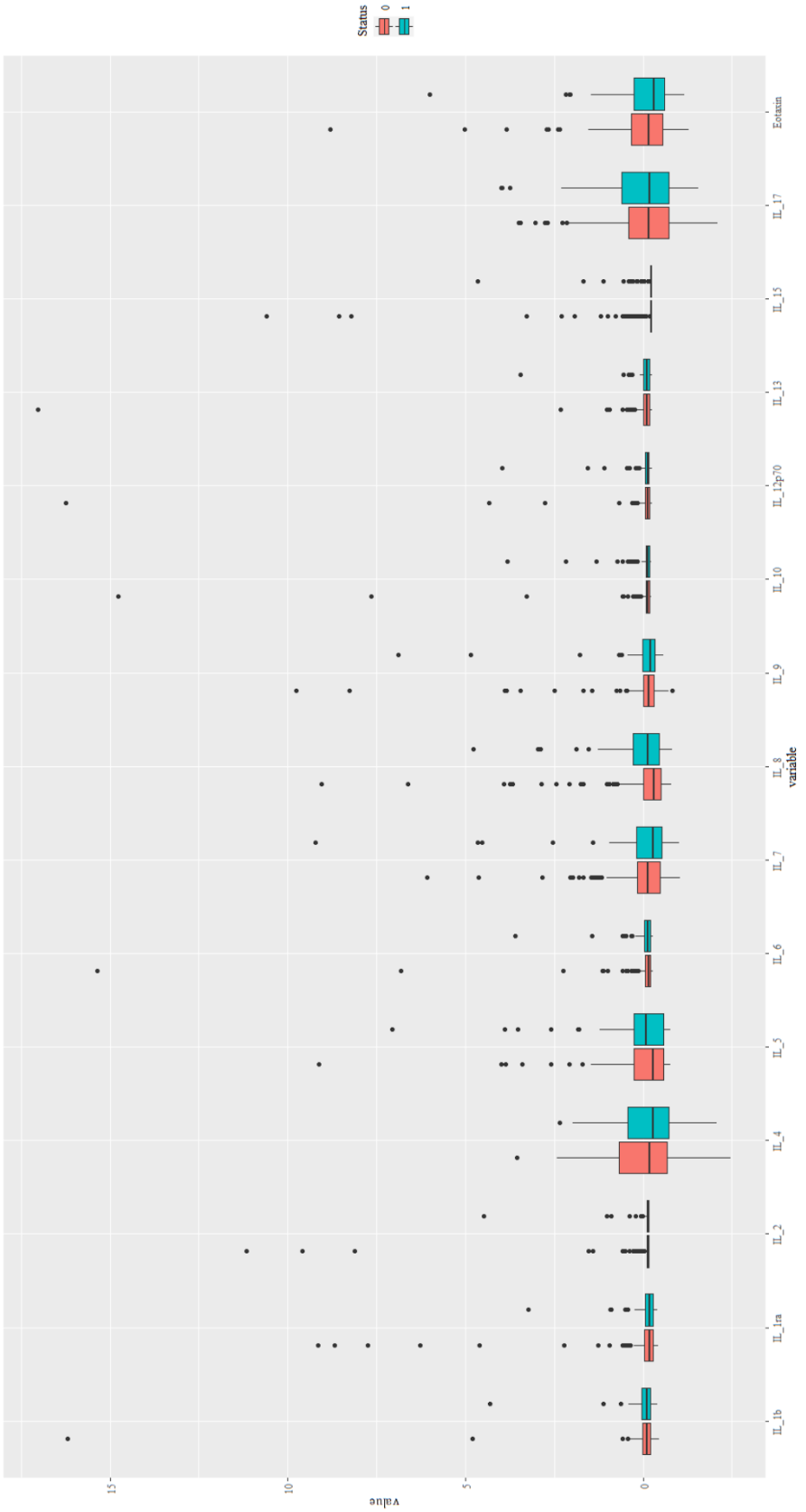

Survival Time
15
10
5
0
I
II
III
IV
Grade
Censor
0
1

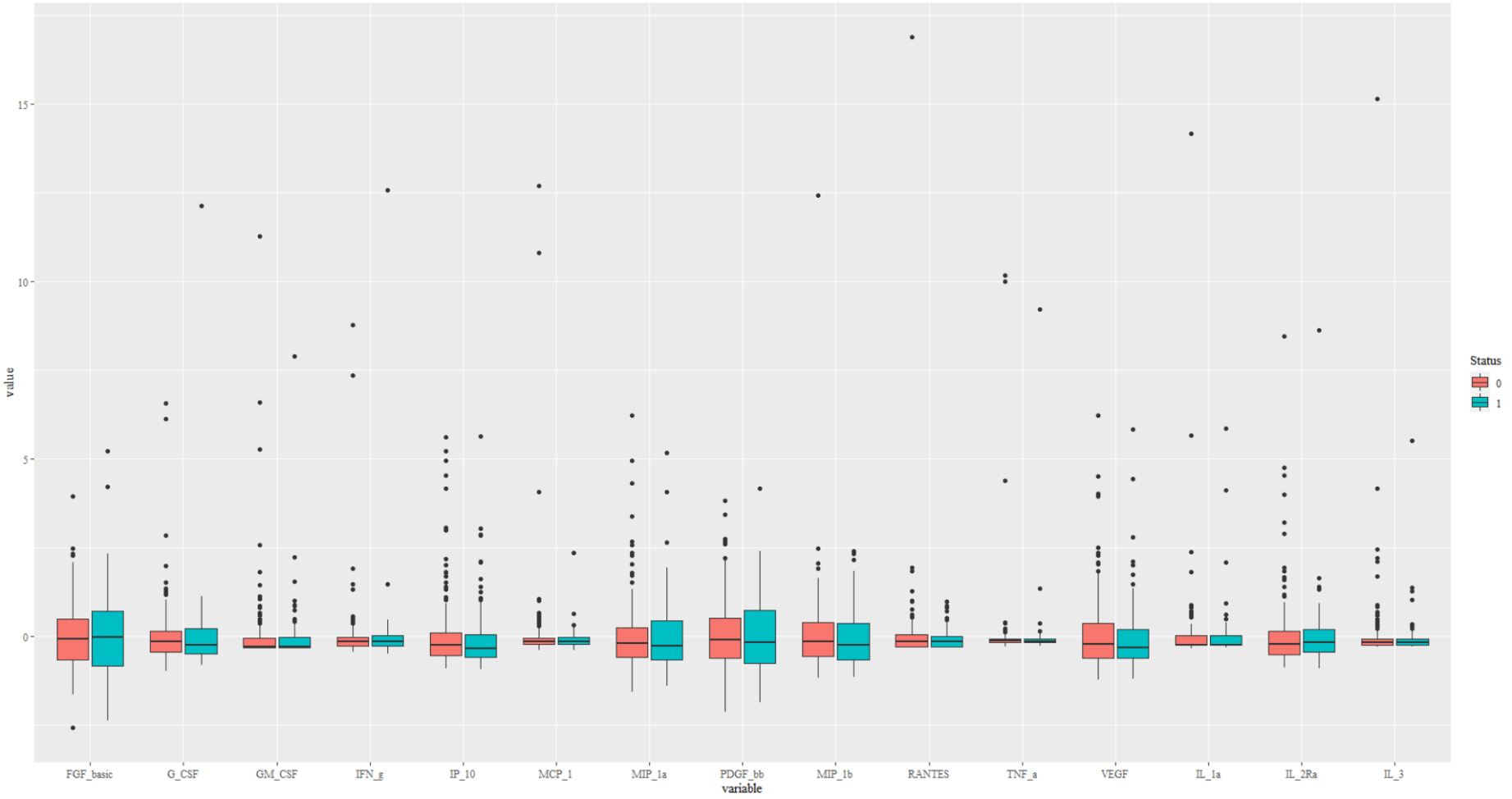

Survival Time
15
10
5
0
A
B
C
D
Dukes
Censor
0
1

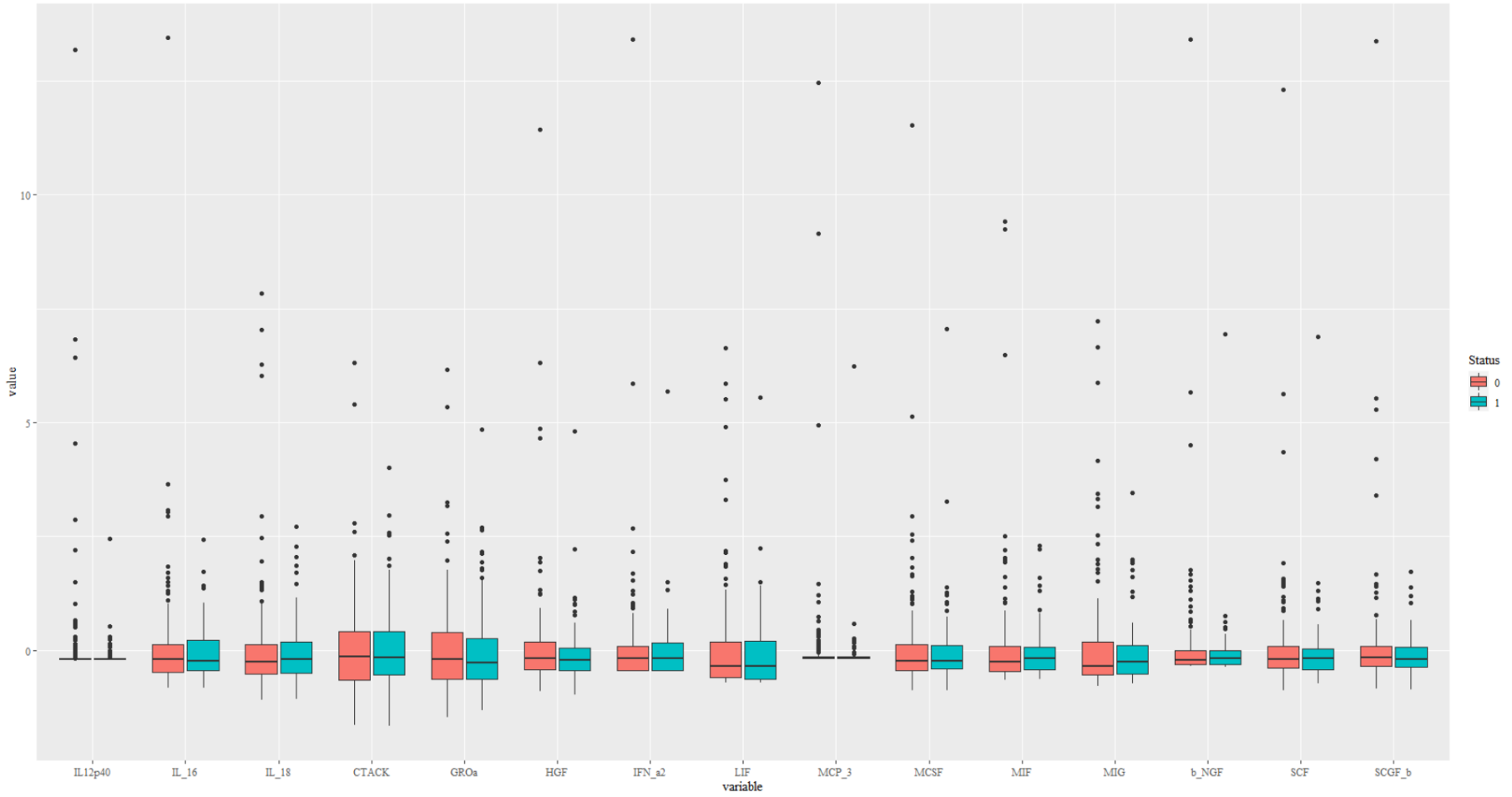

Survival Time
15
10
5
0
Adeno
Muc
Histopathology
Censor
0
1

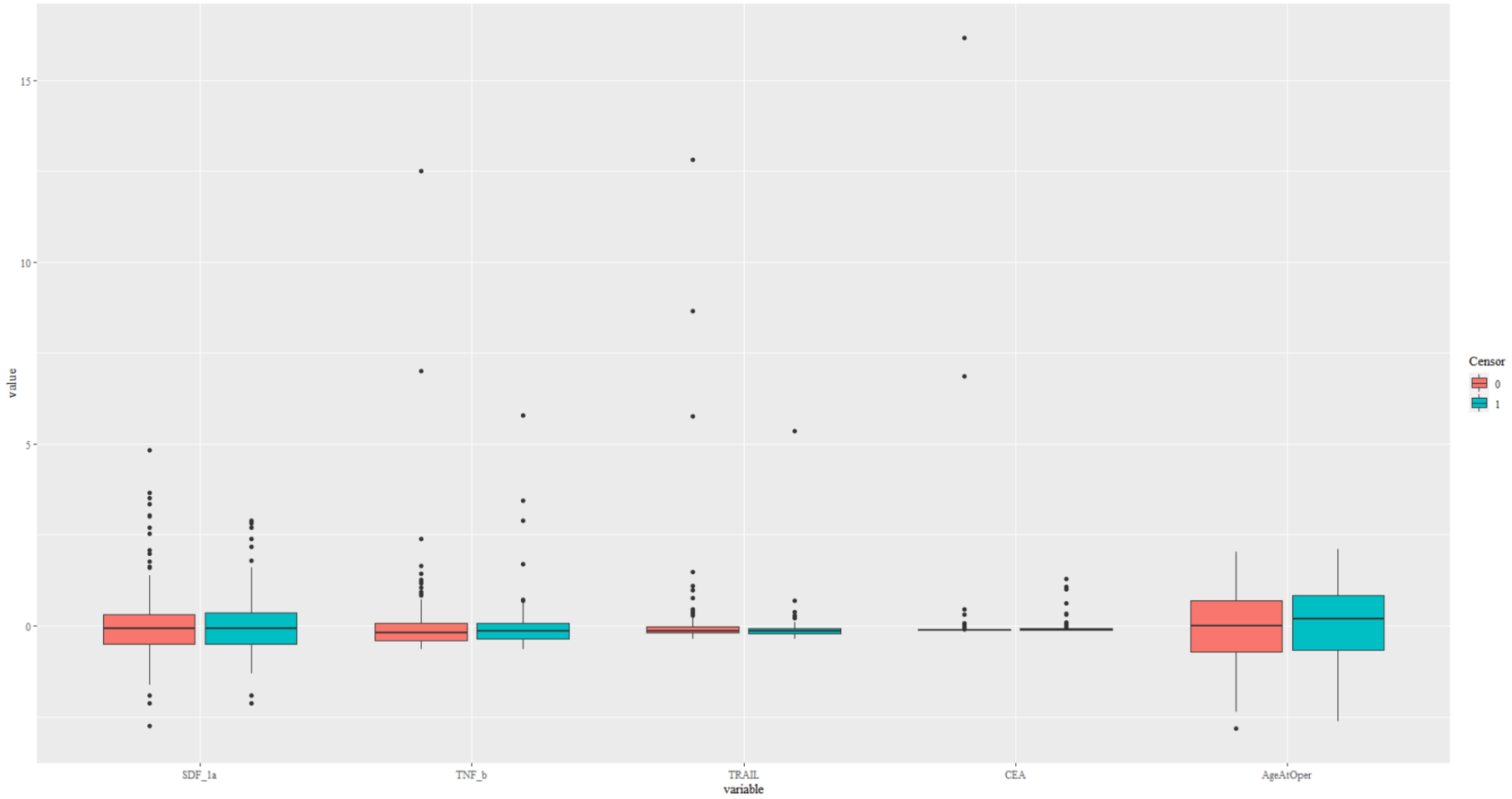

Survival Time
15
10
5
0
Colon
Rectum
Location
Censor
0
1

APPENDIX 6. Boxplots of continuous features.




15
10
5
0
value
FGF_basic
G_CSF
GM_CSF
IFN_g
IP_10
MCP_1
MIP_1a
PDGF_bb
MIP_1b
RANTES
TNF_a
VEGF
IL_1a
IL_2Ra
IL_3
variable
Status
0
1


value
10
5
0
IL12p40
IL_16
IL_18
CTACK
GROa
HGF
IFN_a2
LIF
MCP_3
MCSF
MIF
MIG
b_NGF
SCF
SCGF_b
variable
Status
0
1


15
10
5
0
value
SDF_1a
TNF_b
TRAIL
CEA
AgeAtOper
variable
Censor
0
1

APPENDIX 7. Influx and outflux.

From the calculated influx and outflux values can be observed how each variable is overall connected to the others. Influx is defined as the proportion of usable cases for imputing variable Y_j from variable Y_k. Essentially this means that a number of variable pairs (Y_j, Y_k) with Y_j missing and Y_k observed is divided by the total number of observed data cells. Data is $n \times p$ matrix Y. Influx gets values in the interval [0,1] where one (1) indicates fully missing variable, and zero (0) fully observed variable. In the case of two (2) variables having same proportion of missing data, the one having a higher value of influx is better connected to the rest of the data. Thus making it better choice for imputation. The formula ***(42)*** (van Buuren, 2018) displays the calculation of influx, I_j where j and k are indices. (van Buuren, 2018)

$$I_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1-r_{ij}) r_{ik}}{\sum_k^p \sum_i^n r_{ik}} \qquad \textbf{(42)}$$

Similarly, outflux measures how well observed values in Y_j connect to missing data in other values. Differing from the calculation of influx, the number of variable pairs is divided by the total number of incomplete data cells, O_j (see formula (***43***) (van Buuren, 2018)). Outflux also gets values in the interval [0,1] where one (1) indicates fully observed variable, and zero (0) fully missing variable. In the case of two (2) variables having same proportion of missing data, the one having a higher value of influx is better connected to the missing of the data.

$$O_j = \frac{\sum_j^p \sum_k^p \sum_i^n r_{ik} (1-r_{ij})}{\sum_k^p \sum_i^n 1-r_{ij}} \qquad \textbf{(43)}$$

These influx and outflux values are calculated for our CRC patient data to support the imputation. The proportion of data observed for different variables varies in the interval of [0.714, 1]. This means that at least 71.4 % of the values are observed for each variable in this cohort. For variables having one or more variable values missing influx varies in the interval [0, 0.268] and

outflux gets values in the interval [0.137, 1]. Figure 34 demonstrates in what manner influx and outflux values for different variables are connected. Blue dashed line represents the influx values whereas red solid line the outflux values.

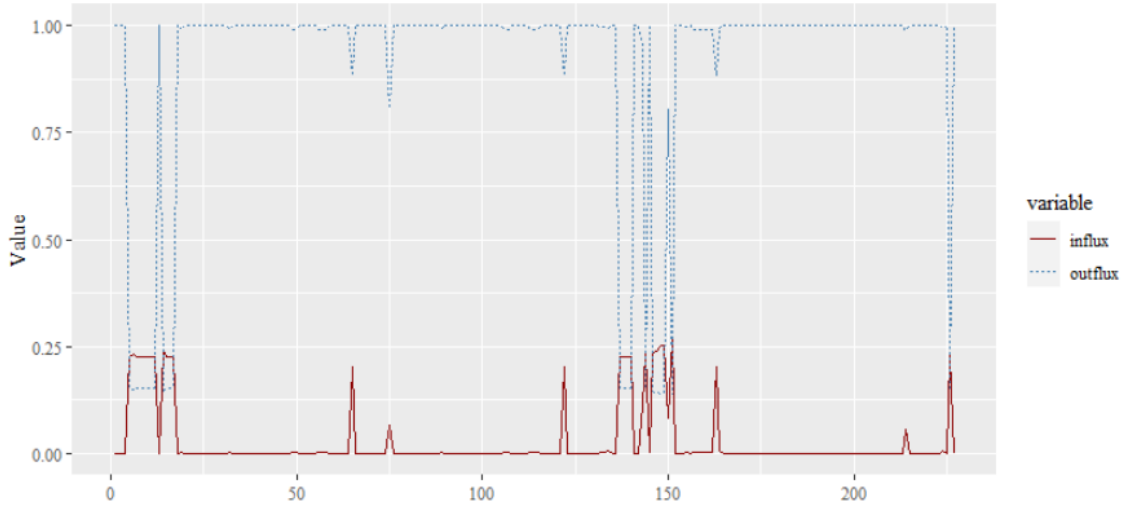


Figure 34. Influx and outflux values of the variables.

APPENDIX 8. Included variables.

The following table displays first the variable name, then the number of valid, recorded values, followed by the number of missing values, and in the last column the original data source is named. Rows having any missing values are highlighted.

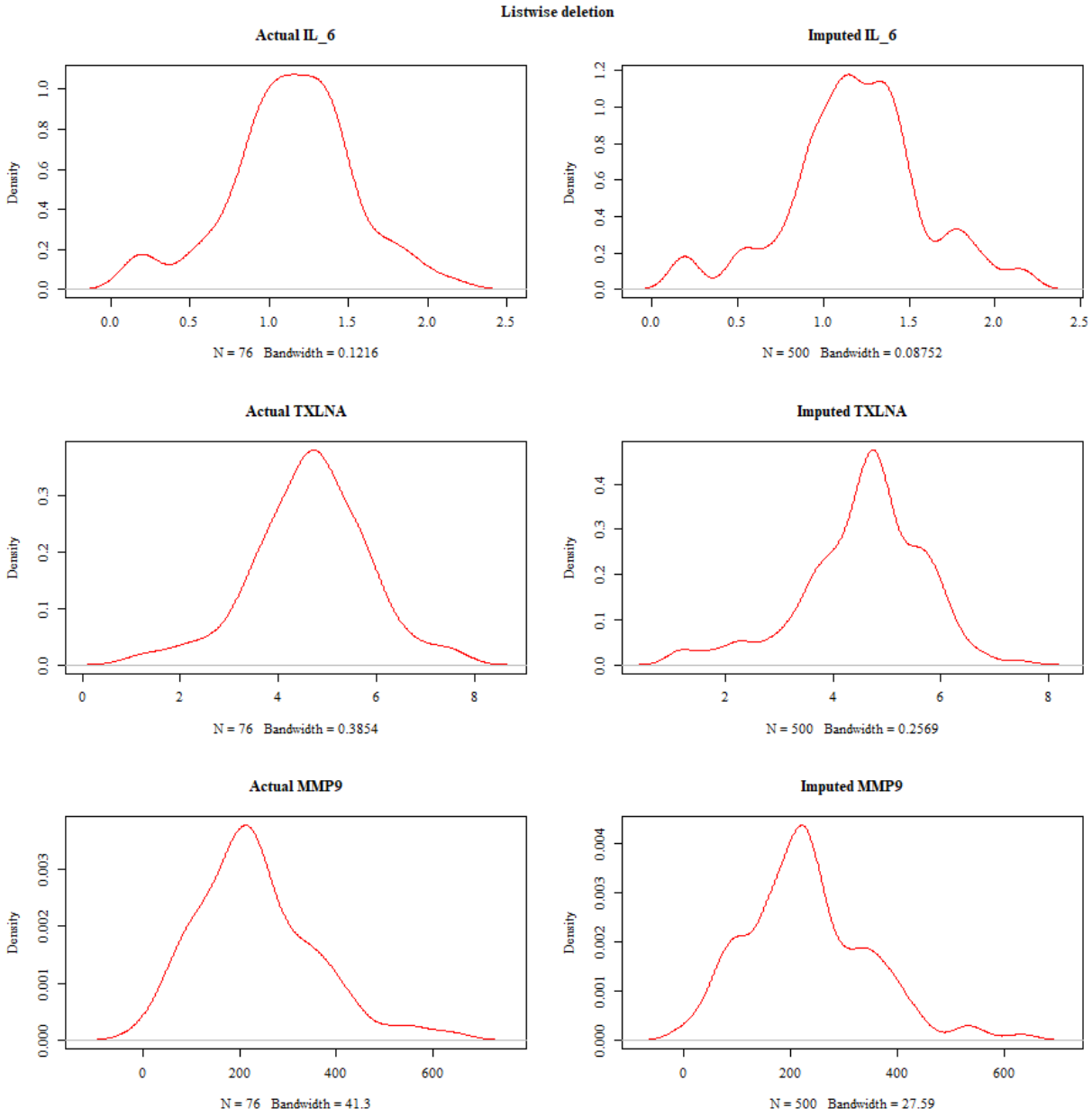
	Number of recorded values	**Number of missing values**	**Data source**
IL_1b	318	0	Immunopanel
IL_1ra	318	0	Immunopanel
IL_2	318	0	Immunopanel
IL_4	318	0	Immunopanel
IL_5	318	0	Immunopanel
IL_6	316	**2**	Immunopanel
IL_7	318	0	Immunopanel
IL_8	318	0	Immunopanel
IL_9	318	0	Immunopanel
IL_10	318	0	Immunopanel
IL_12p70	318	0	Immunopanel
IL_13	318	0	Immunopanel
IL_15	317	**1**	Immunopanel
IL_17	318	0	Immunopanel
Eotaxin	318	0	Immunopanel
FGF_basic	318	0	Immunopanel
G_CSF	318	0	Immunopanel
GM_CSF	318	0	Immunopanel
IFN_g	318	0	Immunopanel
IP_10	318	0	Immunopanel
MCP_1	318	0	Immunopanel
MIP_1a	318	0	Immunopanel
PDGF_bb	318	0	Immunopanel
MIP_1b	318	0	Immunopanel
RANTES	318	0	Immunopanel
TNF_a	318	0	Immunopanel
VEGF	318	0	Immunopanel
IL_1a	318	0	Immunopanel
IL_2Ra	318	0	Immunopanel
IL_3	317	**1**	Immunopanel
IL12p40	317	**1**	Immunopanel
IL_16	318	0	Immunopanel
IL_18	318	0	Immunopanel
CTACK	318	0	Immunopanel
GROa	318	0	Immunopanel
HGF	318	0	Immunopanel
IFN_a2	317	**1**	Immunopanel

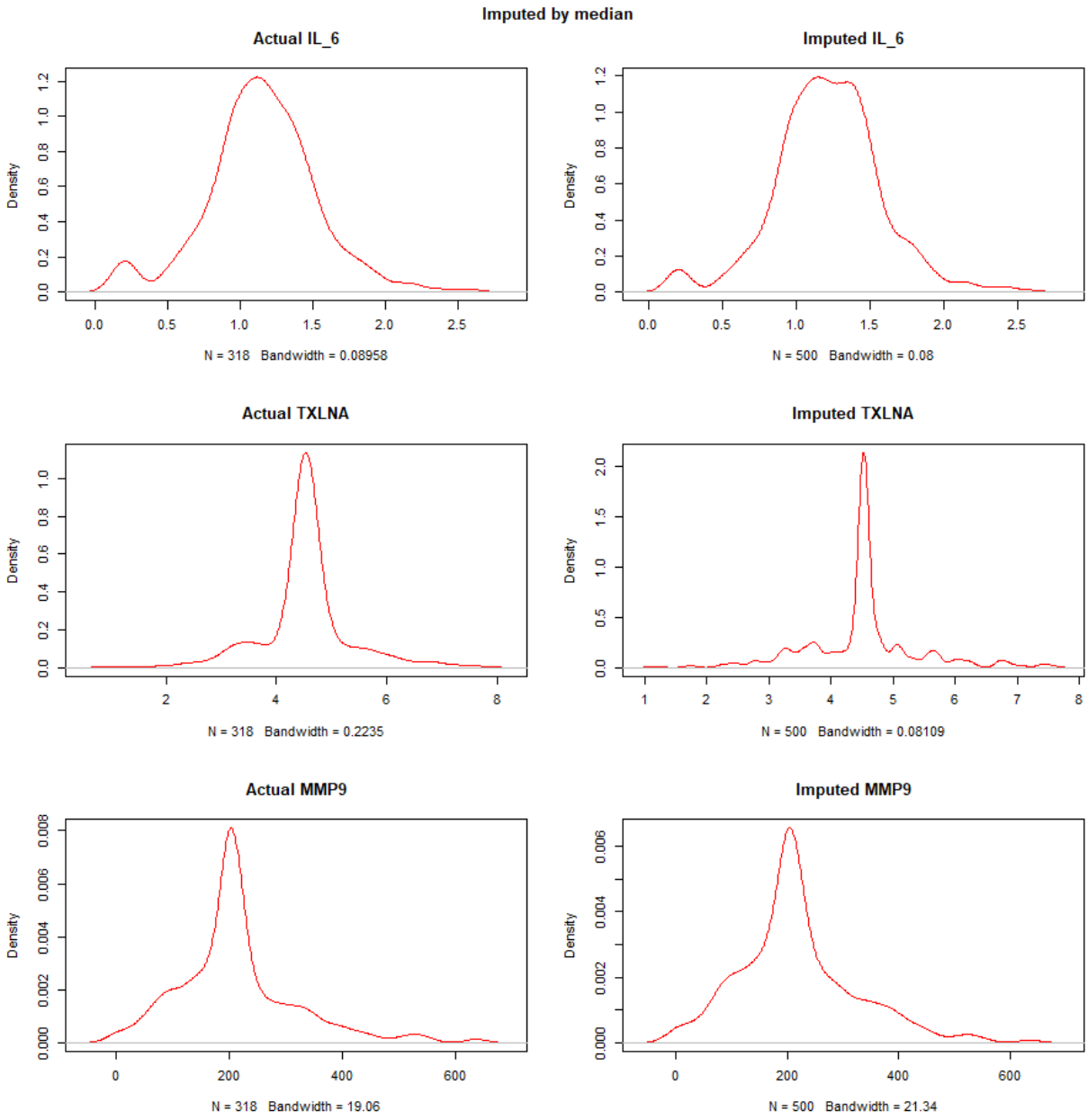
LIF	317	**1**	Immunopanel
MCP_3	317	**1**	Immunopanel
MCSF	318	0	Immunopanel
MIF	318	0	Immunopanel
MIG	318	0	Immunopanel
b_NGF	318	0	Immunopanel
SCF	318	0	Immunopanel
SCGF_b	318	0	Immunopanel
SDF_1a	254	**64**	Immunopanel
TNF_b	318	0	Immunopanel
TRAIL	318	0	Immunopanel
CEA	315	**3**	Immunopanel
Side	318	0	Immunopanel
diff	283	**35**	Immunopanel
muc_adeno_old	317	**1**	Immunopanel
DSS_censor	318	0	Immunopanel
DSS_SurvivalTime	318	0	Immunopanel
AgeAtOper	318	0	Immunopanel
SEX	318	0	Immunopanel
DUKES	318	0	Immunopanel
location	318	0	Immunopanel
CRPmgl	236	**82**	CRP/TATI/MMP
TATIngml	236	**82**	CRP/TATI/MMP
MMP8ngml	236	**82**	CRP/TATI/MMP
MMP9ngml	236	**82**	CRP/TATI/MMP
TIMP1ngml	236	**82**	CRP/TATI/MMP
MMP8TIMP1molratio	236	**82**	CRP/TATI/MMP
MMP9TIMP1molratio	236	**82**	CRP/TATI/MMP
TXLNA	145	**173**	Olink
VEGFA	145	**173**	Olink
CPE	145	**173**	Olink
KLK13	145	**173**	Olink
CEACAM1	145	**173**	Olink
MSLN	145	**173**	Olink
TNFSF13	145	**173**	Olink
EGF	145	**173**	Olink
TNFRSF6B	145	**173**	Olink
SYND1	145	**173**	Olink
TGFR2	145	**173**	Olink
CD48	145	**173**	Olink
SCAMP3	145	**173**	Olink
LY9	145	**173**	Olink
IFNgammaR1	145	**173**	Olink
ITGAV	145	**173**	Olink
TRAIL1	145	**173**	Olink
hk11	145	**173**	Olink

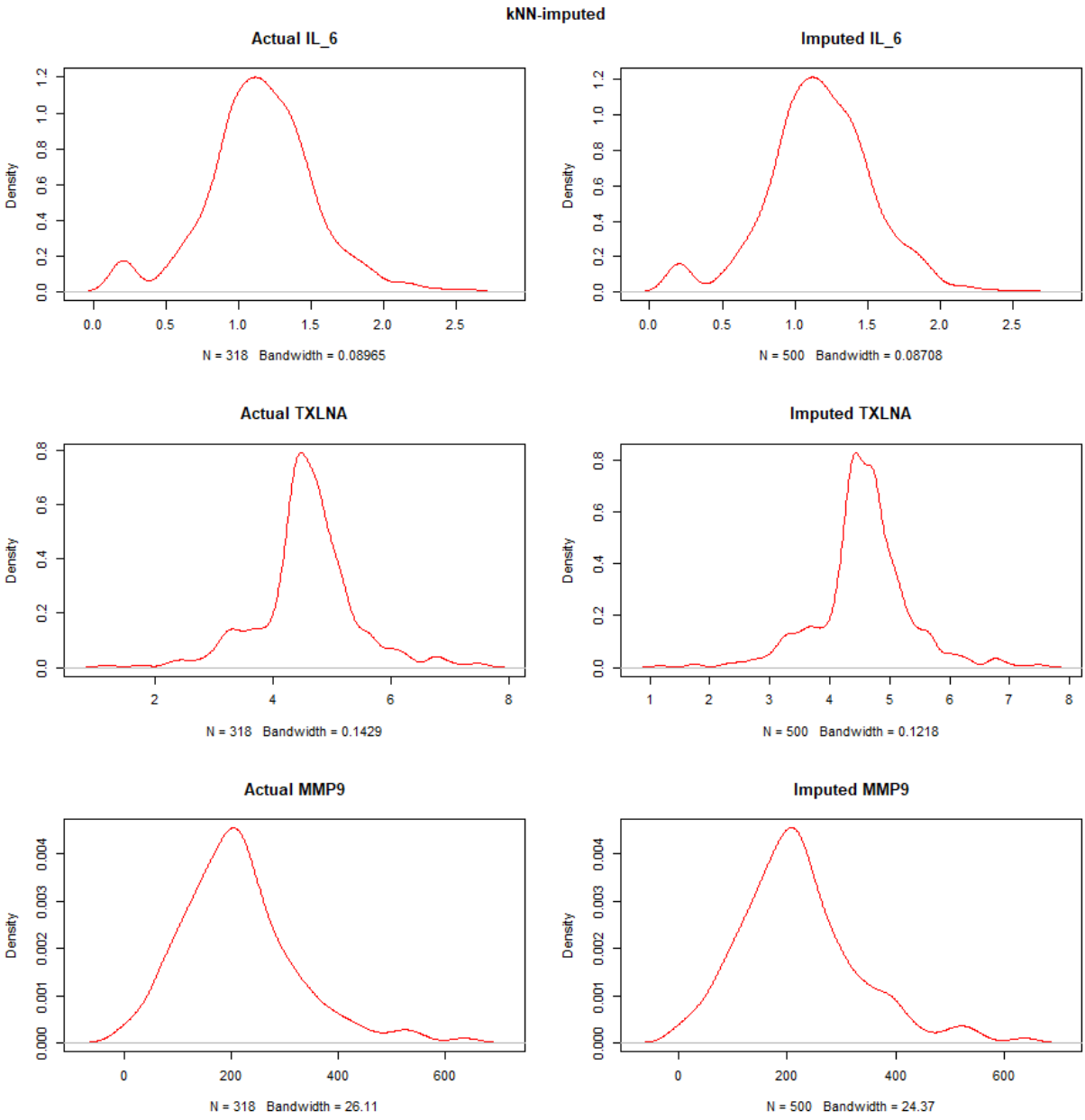
CPC1	145	**173**	Olink
TFPI2	145	**173**	Olink
hk8	145	**173**	Olink
VEGFR2	145	**173**	Olink
LYPD3	145	**173**	Olink
PODXL	145	**173**	Olink
S100A4	145	**173**	Olink
IGF1R	145	**173**	Olink
ERBB2	145	**173**	Olink
ERBB3	145	**173**	Olink
SCF1	145	**173**	Olink
SPARC	145	**173**	Olink
GZMH	145	**173**	Olink
PGFalpha	145	**173**	Olink
FURIN	145	**173**	Olink
CYR61	145	**173**	Olink
hk14	145	**173**	Olink
FADD	145	**173**	Olink
MetAP2	145	**173**	Olink
PVRL4	145	**173**	Olink
FASLG	145	**173**	Olink
EPHA2	145	**173**	Olink
ITGB5	145	**173**	Olink
Gal1	145	**173**	Olink
SEZ6L	145	**173**	Olink
GPNMB	145	**173**	Olink
CAIX	145	**173**	Olink
MIA	145	**173**	Olink
CTSV	145	**173**	Olink
CD27	145	**173**	Olink
XPNPEP2	145	**173**	Olink
ERBB4	145	**173**	Olink
HGF1	145	**173**	Olink
ADAM8	145	**173**	Olink
aSNT	145	**173**	Olink
DKN1A	145	**173**	Olink
DLL1	145	**173**	Olink
MK	145	**173**	Olink
ALB1	145	**173**	Olink
FGFBP1	145	**173**	Olink
TLR3	145	**173**	Olink
LYN	145	**173**	Olink
RET	145	**173**	Olink
VIM	145	**173**	Olink
TNFRSF19	145	**173**	Olink
CRNN	145	**173**	Olink

TCL1A	145	**173**	Olink
CD160	145	**173**	Olink
TNFRSF4	145	**173**	Olink
MICAB	145	**173**	Olink
WISP1	145	**173**	Olink
CXL17	145	**173**	Olink
PPY	145	**173**	Olink
S100A11	145	**173**	Olink
AREG	145	**173**	Olink
ESM1	145	**173**	Olink
CD207	145	**173**	Olink
ICOSLG	145	**173**	Olink
WFDC2	145	**173**	Olink
CXCL13	145	**173**	Olink
MADhomolog5	145	**173**	Olink
ADAMTS15	145	**173**	Olink
CD70	145	**173**	Olink
RSPO3	145	**173**	Olink
Frgamma	145	**173**	Olink
CEACAM5	145	**173**	Olink
VEGFR3	145	**173**	Olink
MUC16	145	**173**	Olink
WIF1	145	**173**	Olink
GZMB	145	**173**	Olink
FCRLB	145	**173**	Olink
ANXA1	145	**173**	Olink
Fralpha	145	**173**	Olink

APPENDIX 9. Comparison of the imputed and artificially increased data sets' distributions.




Imputed by median
Actual IL_6
Density
N = 318 Bandwidth = 0.08958
Imputed IL_6
Density
N = 500 Bandwidth = 0.08
Actual TXLNA
Density
N = 318 Bandwidth = 0.2235
Imputed TXLNA
Density
N = 500 Bandwidth = 0.08109
Actual MMP9
Density
N = 318 Bandwidth = 19.06
Imputed MMP9
Density
N = 500 Bandwidth = 21.34


kNN-imputed
Actual IL_6
Density
N = 318 Bandwidth = 0.08965
Imputed IL_6
Density
N = 500 Bandwidth = 0.08708
Actual TXLNA
Density
N = 318 Bandwidth = 0.1429
Imputed TXLNA
Density
N = 500 Bandwidth = 0.1218
Actual MMP9
Density
N = 318 Bandwidth = 26.11
Imputed MMP9
Density
N = 500 Bandwidth = 24.37

APPENDIX 10. Selected features using correlation analysis.

The following table shows the features selected using correlation analysis for all the datasets. The black horizontal line at the end of the table separates the variables selected when the correlation bound is set to 0.8. Before that are the variables chosen using the correlation bound of 0.7. Finally, after the table is another table summarising the number of features chosen using either one of this correlation bounds.

Listwise deletion	Median imputed	kNN-imputed
IL_5	IL_1ra	IL_1ra
IL_6	IL_4	IL_4
IL_8	IL_5	IL_5
FGF_basic	IL_6	IL_6
IP_10	IL_7	IL_7
MIP_1a	IL_8	IL_8
PDGF_bb	IL_9	IL_9
MIP_1b	Eotaxin	Eotaxin
RANTES	FGF_basic	FGF_basic
VEGF	IP_10	IP_10
CTACK	MIP_1a	MIP_1a
GROa	PDGF_bb	PDGF_bb
MIF	MIP_1b	MIP_1b
SCGF_b	RANTES	RANTES
SDF_1a	TNF_a	TNF_a
Side	VEGF	VEGF
diff	IL12p40	IL12p40
muc_adeno_old	IL_18	IL_18
AgeAtOper	CTACK	CTACK
SEX	GROa	GROa
DUKES	MIF	MIF
TATIngml	MIG	MIG
MMP8ngml	b_NGF	b_NGF
MMP9ngml	SCGF_b	SCGF_b
MMP8TIMP1molra-tio	SDF_1a	SDF_1a
MMP9TIMP1molra-tio	TNF_b	TNF_b
TXLNA	TRAIL	TRAIL
VEGFA	CEA	CEA
CPE	Side	Side
TNFSF13	diff	diff
TGFR2	muc_adeno_old	muc_adeno_old
SCAMP3	AgeAtOper	AgeAtOper
LY9	SEX	SEX
ITGAV	DUKES	DUKES
TRAIL1	location	location
hk11	MMP8ngml	MMP8ngml
CPC1	MMP9ngml	MMP9ngml

TFPI2	TIMP1ngml	TIMP1ngml
hk8	TXLNA	TXLNA
VEGFR2	VEGFA	VEGFA
PODXL	CPE	CPE
S100A4	KLK13	KLK13
IGF1R	CD48	CD48
ERBB2	LY9	LY9
ERBB3	IFNgammaR1	IFNgammaR1
SPARC	ITGAV	ITGAV
GZMH	TRAIL1	TRAIL1
PGFalpha	CPC1	CPC1
FADD	TFPI2	TFPI2
PVRL4	hk8	hk8
FASLG	VEGFR2	VEGFR2
EPHA2	LYPD3	LYPD3
ITGB5	S100A4	S100A4
Gal1	IGF1R	IGF1R
SEZ6L	ERBB2	ERBB2
GPNMB	ERBB3	ERBB3
MIA	SCF1	SCF1
CTSV	GZMH	GZMH
CD27	PGFalpha	PGFalpha
ERBB4	FURIN	FURIN
DLL1	MetAP2	MetAP2
MK	FASLG	FASLG
ALB1	EPHA2	EPHA2
FGFBP1	ITGB5	ITGB5
RET	Gal1	Gal1
VIM	SEZ6L	SEZ6L
TCL1A	GPNMB	GPNMB
TNFRSF4	CAIX	CAIX
MICAB	CTSV	CTSV
CXL17	CD27	CD27
PPY	XPNPEP2	XPNPEP2
S100A11	ERBB4	ERBB4
AREG	HGF1	HGF1
CD207	DKN1A	DKN1A
ICOSLG	MK	MK
WFDC2	ALB1	ALB1
CXCL13	FGFBP1	FGFBP1
ADAMTS15	TLR3	TLR3
CD70	VIM	VIM
RSPO3	TNFRSF19	TNFRSF19
Frgamma	CRNN	CRNN
CEACAM5	CD160	CD160
VEGFR3	TNFRSF4	TNFRSF4
MUC16	MICAB	MICAB
WIF1	WISP1	WISP1
GZMB	PPY	CXL17
	S100A11	PPY

		AREG	S100A11
		ESM1	AREG
		CD207	ESM1
		ICOSLG	CD207
		WFDC2	ICOSLG
		CXCL13	WFDC2
		MADhomolog5	CXCL13
		ADAMTS15	MADhomolog5
		CD70	ADAMTS15
		RSPO3	CD70
		Frgamma	RSPO3
		CEACAM5	Frgamma
		VEGFR3	CEACAM5
		MUC16	VEGFR3
		WIF1	MUC16
		GZMB	WIF1
		FCRLB	GZMB
		Fralpha	FCRLB
			Fralpha
0.8	IL_4	G_CSF	G_CSF
	IL_7	IFN_g	IFN_g
	IL_9	LIF	LIF
	IL_17	SCF	SCF
	Eotaxin	CRPmgl	CRPmgl
	IL_16	MMP9TIMP1molra-tio	MMP9TIMP1molra-tio
	MSLN	TNFSF13	TNFSF13
	TNFRSF6B	EGF	EGF
	IFNgammaR1	SYND1	SYND1
	LYPD3	hk11	hk11
	SCF1	PODXL	PODXL
	XPNPEP2	hk14	hk14
	aSNT	aSNT	aSNT
	DKN1A	DLL1	DLL1
	TLR3	LYN	LYN
	TNFRSF19	RET	RET
	WISP1	TCL1A	TCL1A
	ESM1	CXL17	
	MADhomolog5		

Total features with correlation bound			
0.7	86	105	106
0.8	105	123	123

APPENDIX 11. Selected features using univariate CPH.

The following tables show the features for all the datasets with p-values and false discovery rate (FDR) corrected p-values obtained from the univariate CPH analysis. For the feature selection the criterion of statistical significance with threshold $p < 0.10$ is chosen. The features satisfying this criterion are highlighted in the tables. At the end of each table there is a summary informing the number of chosen features.

Listwise deletion FDR-corrected p-values highlighted for **p-value < 0.1**

	p-value	FDR-corrected p-value
IL_1b	0.99	1
IL_1ra	0.55	0.923057851
IL_2	0.78	0.938863636
IL_4	0.97	1
IL_5	0.45	0.923057851
IL_6	0.21	0.714
IL_7	0.62	0.923057851
IL_8	**0.074**	0.597073171
IL_9	0.19	0.660681818
IL_10	0.88	0.954893617
IL_12p70	0.85	0.942391304
IL_13	0.73	0.923057851
IL_15	0.7	0.923057851
IL_17	0.66	0.923057851
Eotaxin	0.46	0.923057851
FGF_basic	0.71	0.923057851
G_CSF	0.73	0.923057851
GM_CSF	0.7	0.923057851
IFN_g	0.69	0.923057851
IP_10	0.37	0.867761194
MCP_1	0.58	0.923057851
MIP_1a	0.69	0.923057851
PDGF_bb	0.96	1
MIP_1b	0.66	0.923057851
RANTES	0.11	0.597073171
TNF_a	0.81	0.938863636
VEGF	0.65	0.923057851
IL_1a	0.79	0.938863636
IL_2Ra	0.8	0.938863636
IL_3	0.8	0.938863636
IL12p40	0.65	0.923057851
IL_16	0.84	0.942391304
IL_18	0.59	0.923057851
CTACK	0.34	0.867761194

GROa	0.5	0.923057851
HGF	0.45	0.923057851
IFN_a2	0.71	0.923057851
LIF	1	1
MCP_3	0.73	0.923057851
MCSF	0.71	0.923057851
MIF	0.73	0.923057851
MIG	0.61	0.923057851
b_NGF	0.99	1
SCF	0.84	0.942391304
SCGF_b	0.66	0.923057851
SDF_1a	0.36	0.867761194
TNF_b	0.75	0.938863636
TRAIL	0.71	0.923057851
CEA	0.73	0.923057851
AgeAtOper	0.38	0.867761194
CRPmgl	0.5	0.923057851
TATIngml	0.11	0.597073171
MMP8ngml	0.14	0.597073171
MMP9ngml	0.72	0.923057851
TIMP1ngml	**0.081**	0.597073171
MMP8TIMP1molra-tio	0.51	0.923057851
MMP9TIMP1molra-tio	0.87	0.950785714
TXLNA	0.71	0.923057851
VEGFA	0.19	0.660681818
CPE	**0.097**	0.597073171
KLK13	0.11	0.597073171
CEACAM1	0.37	0.867761194
MSLN	0.73	0.923057851
TNFSF13	0.35	0.867761194
EGF	0.57	0.923057851
TNFRSF6B	0.23	0.748723404
SYND1	**0.014**	0.597073171
TGFR2	0.31	0.867761194
CD48	0.16	0.597073171
SCAMP3	0.9	0.962937063
LY9	0.14	0.597073171
IFNgammaR1	**0.083**	0.597073171
ITGAV	0.96	1
TRAIL1	0.86	0.946618705
hk11	**0.085**	0.597073171
CPC1	0.26	0.811836735
TFPI2	0.14	0.597073171
hk8	0.79	0.938863636
VEGFR2	0.67	0.923057851
LYPD3	0.12	0.597073171
PODXL	0.67	0.923057851

S100A4	0.14	0.597073171
IGF1R	**0.018**	0.597073171
ERBB2	**0.086**	0.597073171
ERBB3	**0.092**	0.597073171
SCF1	0.98	1
SPARC	1	1
GZMH	0.64	0.923057851
PGFalpha	**0.038**	0.597073171
FURIN	**0.096**	0.597073171
CYR61	**0.072**	0.597073171
hk14	0.99	1
FADD	0.54	0.923057851
MetAP2	0.23	0.748723404
PVRL4	0.11	0.597073171
FASLG	0.81	0.938863636
EPHA2	**0.028**	0.597073171
ITGB5	0.47	0.923057851
Gal1	0.16	0.597073171
SEZ6L	0.38	0.867761194
GPNMB	**0.026**	0.597073171
CAIX	0.59	0.923057851
MIA	0.63	0.923057851
CTSV	0.8	0.938863636
CD27	0.34	0.867761194
XPNPEP2	0.48	0.923057851
ERBB4	**0.053**	0.597073171
HGF1	0.12	0.597073171
ADAM8	**0.039**	0.597073171
aSNT	0.11	0.597073171
DKN1A	0.64	0.923057851
DLL1	0.16	0.597073171
MK	0.38	0.867761194
ALB1	0.38	0.867761194
FGFBP1	0.93	0.988125
TLR3	0.7	0.923057851
LYN	0.53	0.923057851
RET	0.89	0.958943662
VIM	0.15	0.597073171
TNFRSF19	0.73	0.923057851
CRNN	0.48	0.923057851
TCL1A	0.52	0.923057851
CD160	0.79	0.938863636
TNFRSF4	0.32	0.867761194
MICAB	0.77	0.938863636
WISP1	0.15	0.597073171
CXL17	0.85	0.942391304
PPY	0.58	0.923057851
S100A11	0.14	0.597073171
AREG	**0.041**	0.597073171

ESM1	0.11	0.597073171
CD207	0.67	0.923057851
ICOSLG	0.33	0.867761194
WFDC2	0.28	0.84
CXCL13	0.36	0.867761194
MADhomolog5	**0.089**	0.597073171
ADAMTS15	**0.08**	0.597073171
CD70	0.82	0.942391304
RSPO3	0.39	0.874285714
Frgamma	0.46	0.923057851
CEACAM5	0.4	0.874285714
VEGFR3	0.18	0.655714286
MUC16	**0.000095**	0.014535
WIF1	0.4	0.874285714
GZMB	0.32	0.867761194
FCRLB	0.11	0.597073171
ANXA1	0.47	0.923057851
Fralpha	0.28	0.84
Side	0.122021943	0.597073171
diff	0.84659695	0.942391304
muc_adeno_old	0.252840596	0.805929399
SEX	0.154673686	0.597073171
DUKES	0.293915514	0.864789876

Total selected 1

Imputed by median

FDR-corrected p-values highlighted for **p-value < 0.1**

	p-value	**FDR-corrected p-value**
IL_1b	0.75	0.9384375
IL_1ra	0.61	0.877943925
IL_2	0.52	0.870434783
IL_4	0.17	0.565714286
IL_5	0.42	0.849655172
IL_6	**1.20E-03**	0.0308
IL_7	0.48	0.849655172
IL_8	**0.000024**	0.001232
IL_9	0.71	0.926610169
IL_10	0.94	0.978108108
IL_12p70	0.92	0.963809524
IL_13	0.71	0.926610169
IL_15	0.47	0.849655172
IL_17	0.29	0.708888889
Eotaxin	0.34	0.781492537
FGF_basic	0.52	0.870434783
G_CSF	0.68	0.926610169

GM_CSF	0.9	0.9625
IFN_g	0.74	0.9384375
IP_10	0.78	0.9384375
MCP_1	0.44	0.849655172
MIP_1a	0.8	0.939852941
PDGF_bb	0.52	0.870434783
MIP_1b	0.34	0.781492537
RANTES	0.39	0.843835616
TNF_a	0.91	0.963809524
VEGF	0.47	0.849655172
IL_1a	0.7	0.926610169
IL_2Ra	0.18	0.565714286
IL_3	0.77	0.9384375
IL12p40	0.27	0.681639344
IL_16	0.96	0.979072848
IL_18	0.22	0.627407407
CTACK	0.1	0.452941176
GROa	0.25	0.675438596
HGF	0.76	0.9384375
IFN_a2	0.99	0.99
LIF	0.6	0.877943925
MCP_3	0.61	0.877943925
MCSF	0.23	0.644
MIF	0.22	0.627407407
MIG	0.58	0.877943925
b_NGF	0.81	0.939852941
SCF	0.86	0.952805755
SCGF_b	0.55	0.877943925
SDF_1a	**0.035**	0.241043478
TNF_b	0.47	0.849655172
TRAIL	0.4	0.843835616
CEA	**8.90E-11**	1.37E-08
AgeAtOper	**0.019**	0.162555556
CRPmgl	0.92	0.963809524
TATIngml	0.3	0.721875
MMP8ngml	**0.0024**	0.0462
MMP9ngml	0.78	0.9384375
TIMP1ngml	**0.00076**	0.023408
MMP8TIMP1molra-tio	0.12	0.499459459
MMP9TIMP1molra-tio	0.36	0.815294118
TXLNA	0.7	0.926610169
VEGFA	**0.036**	0.241043478
CPE	0.59	0.877943925
KLK13	**0.01**	0.11
CEACAM1	0.49	0.8575
MSLN	0.9	0.9625
TNFSF13	0.17	0.565714286

EGF	0.48	0.849655172
TNFRSF6B	**0.023**	0.1848
SYND1	**0.0022**	0.0462
TGFR2	0.24	0.66
CD48	0.22	0.627407407
SCAMP3	0.79	0.939852941
LY9	0.27	0.681639344
IFNgammaR1	**0.079**	0.4158
ITGAV	0.22	0.627407407
TRAIL1	0.21	0.627407407
hk11	0.16	0.565714286
CPC1	0.87	0.957
TFPI2	**0.067**	0.382148148
hk8	**0.086**	0.427225806
VEGFR2	0.4	0.843835616
LYPD3	0.74	0.9384375
PODXL	0.61	0.877943925
S100A4	0.38	0.843835616
IGF1R	**0.033**	0.241043478
ERBB2	0.77	0.9384375
ERBB3	0.97	0.982763158
SCF1	**0.066**	0.382148148
SPARC	0.54	0.877943925
GZMH	0.95	0.979072848
PGFalpha	**0.007**	0.089833333
FURIN	**0.014**	0.126823529
CYR61	**0.081**	0.4158
hk14	0.66	0.926610169
FADD	0.55	0.877943925
MetAP2	0.29	0.708888889
PVRL4	**0.099**	0.452941176
FASLG	0.81	0.939852941
EPHA2	**0.011**	0.112933333
ITGB5	0.5	0.865168539
Gal1	0.17	0.565714286
SEZ6L	0.84	0.944233577
GPNMB	0.46	0.849655172
CAIX	0.59	0.877943925
MIA	0.83	0.939852941
CTSV	**0.061**	0.38192
CD27	0.27	0.681639344
XPNPEP2	0.57	0.877943925
ERBB4	0.68	0.926610169
HGF1	**0.024**	0.1848
ADAM8	**0.074**	0.407
aSNT	0.18	0.565714286
DKN1A	0.98	0.986405229
DLL1	**0.091**	0.4379375
MK	0.18	0.565714286

ALB1	0.34	0.781492537
FGFBP1	0.55	0.877943925
TLR3	0.39	0.843835616
LYN	0.68	0.926610169
RET	0.67	0.926610169
VIM	**0.0034**	0.058177778
TNFRSF19	0.89	0.9625
CRNN	0.77	0.9384375
TCL1A	0.66	0.926610169
CD160	0.82	0.939852941
TNFRSF4	0.45	0.849655172
MICAB	0.41	0.849655172
WISP1	0.13	0.526842105
CXL17	0.86	0.952805755
PPY	0.48	0.849655172
S100A11	**0.0068**	0.089833333
AREG	**0.00011**	0.004235
ESM1	0.16	0.565714286
CD207	0.82	0.939852941
ICOSLG	0.44	0.849655172
WFDC2	0.16	0.565714286
CXCL13	0.16	0.565714286
MADhomolog5	0.83	0.939852941
ADAMTS15	**0.012**	0.1155
CD70	0.69	0.926610169
RSPO3	0.26	0.681639344
Frgamma	0.6	0.877943925
CEACAM5	**0.0042**	0.06468
VEGFR3	**0.062**	0.38192
MUC16	**0.00001**	0.00077
WIF1	0.76	0.9384375
GZMB	0.96	0.979072848
FCRLB	0.11	0.484
ANXA1	0.15	0.565714286
Fralpha	0.57	0.877943925
Side	0.115633312	0.494653613
diff	0.459264924	0.849655172
muc_adeno_old	0.881345108	0.9625
SEX	0.578654307	0.877943925
DUKES	**0.009622212**	0.11
location	0.428047558	0.849655172

Total selected 12

kNN-imputed	FDR-corrected p-values highlighted for **p-value < 0.1**	
	p-value	**FDR-corrected p-value**
IL_1b	0.75	0.855555556
IL_1ra	0.61	0.776363636
IL_2	0.52	0.708672566
IL_4	0.17	0.385
IL_5	0.42	0.653333333
IL_6	**0.0014**	0.0077
IL_7	0.48	0.684444444
IL_8	**2.40E-05**	2.17E-04
IL_9	0.71	0.822105263
IL_10	0.94	0.965066667
IL_12p70	0.92	0.963809524
IL_13	0.71	0.822105263
IL_15	0.47	0.676448598
IL_17	0.29	0.513333333
Eotaxin	0.34	0.575384615
FGF_basic	0.52	0.708672566
G_CSF	0.68	0.818125
GM_CSF	0.9	0.955862069
IFN_g	0.74	0.850447761
IP_10	0.78	0.870434783
MCP_1	0.44	0.664313725
MIP_1a	0.8	0.88
PDGF_bb	0.52	0.708672566
MIP_1b	0.34	0.575384615
RANTES	0.39	0.63893617
TNF_a	0.91	0.959863014
VEGF	0.47	0.676448598
IL_1a	0.7	0.822105263
IL_2Ra	0.18	0.390422535
IL_3	0.77	0.865547445
IL12p40	0.27	0.489176471
IL_16	0.96	0.976339869
IL_18	0.22	0.44
CTACK	0.1	0.252459016
GROa	0.25	0.458333333
HGF	0.76	0.860588235
IFN_a2	1	1
LIF	0.61	0.776363636
MCP_3	0.61	0.776363636
MCSF	0.23	0.44275
MIF	0.22	0.44
MIG	0.58	0.763418803
b_NGF	0.81	0.884680851
SCF	0.86	0.932676056
SCGF_b	0.55	0.736521739

SDF_1a	**0.021**	0.070304348
TNF_b	0.47	0.676448598
TRAIL	0.4	0.641666667
CEA	**9.10E-11**	3.50E-09
AgeAtOper	**0.019**	0.065022222
CRPmgl	0.97	0.976339869
TATIngml	0.3	0.525
MMP8ngml	**2.10E-04**	1.61E-03
MMP9ngml	0.4	0.641666667
TIMP1ngml	**2.40E-05**	0.000217412
MMP8TIMP1molra-tio	**0.042**	0.124384615
MMP9TIMP1molra-tio	0.1	0.252459016
TXLNA	0.11	0.268888889
VEGFA	**1.20E-05**	1.32E-04
CPE	0.65	0.807258065
KLK13	**3.00E-10**	9.24E-09
CEACAM1	0.93	0.965066667
MSLN	0.43	0.655643564
TNFSF13	**0.0052**	0.022244444
EGF	0.37	0.612688172
TNFRSF6B	**6.80E-04**	4.19E-03
SYND1	**4.70E-11**	2.41E-09
TGFR2	**0.042**	0.124384615
CD48	**1.80E-03**	9.24E-03
SCAMP3	**0.035**	0.1078
LY9	**0.023**	0.075361702
IFNgammaR1	**0.0092**	0.033733333
ITGAV	**8.70E-02**	0.235052632
TRAIL1	**0.044**	0.127849057
hk11	**0.0038**	0.017211765
CPC1	0.13	0.303333333
TFPI2	**0.00029**	0.001941739
hk8	**1.50E-03**	7.97E-03
VEGFR2	0.35	0.585869565
LYPD3	0.64	0.801300813
PODXL	0.24	0.445301205
S100A4	0.31	0.536404494
IGF1R	**0.000000093**	1.79025E-06
ERBB2	7.10E-01	8.22E-01
ERBB3	0.22	0.44
SCF1	**0.000017**	0.000174533
SPARC	**2.50E-03**	0.01203125
GZMH	0.52	0.708672566
PGFalpha	**0.000011**	0.000130308
FURIN	**9.00E-06**	1.26E-04
CYR61	**1.00E-06**	1.71E-05
hk14	5.40E-01	7.29E-01

FADD	0.52	0.708672566
MetAP2	0.13	0.303333333
PVRL4	**0.019**	0.065022222
FASLG	0.21	0.437027027
EPHA2	**0.000087**	0.000705158
ITGB5	1.80E-01	3.90E-01
Gal1	0.23	0.44275
SEZ6L	0.46	0.676448598
GPNMB	**0.0062**	0.025126316
CAIX	0.24	0.445301205
MIA	**0.09**	0.238965517
CTSV	**0.027**	0.084857143
CD27	0.28	0.501395349
XPNPEP2	0.69	0.822105263
ERBB4	0.59	0.77
HGF1	**0.00022**	0.00161
ADAM8	**2.30E-04**	1.61E-03
aSNT	**5.30E-05**	0.000453444
DKN1A	4.10E-01	0.644285714
DLL1	**0.0059**	0.024556757
MK	**0.026**	0.083416667
ALB1	**0.008**	0.0308
FGFBP1	0.24	0.445301205
TLR3	0.41	0.644285714
LYN	**0.046**	0.131185185
RET	0.79	0.875251799
VIM	**4.3E-09**	9.46E-08
TNFRSF19	8.70E-01	9.37E-01
CRNN	0.2	0.421917808
TCL1A	0.62	0.782622951
CD160	**0.059**	0.1652
TNFRSF4	**0.097**	0.252459016
MICAB	**0.0013**	0.007414815
WISP1	**0.00001**	0.000128333
CXL17	6.70E-01	8.18E-01
PPY	0.23	0.44275
S100A11	**0.0000028**	0.00004312
AREG	**1.80E-15**	2.77E-13
ESM1	**2.50E-03**	1.20E-02
CD207	0.97	0.976339869
ICOSLG	0.68	0.818125
WFDC2	**0.0049**	0.02156
CXCL13	**0.0026**	0.012133333
MADhomolog5	9.40E-01	9.65E-01
ADAMTS15	**0.0012**	0.007107692
CD70	0.68	0.818125
RSPO3	**0.062**	0.1705
Frgamma	0.19	0.406388889
CEACAM5	**3.6E-09**	9.24E-08

VEGFR3	**6.40E-03**	2.53E-02
MUC16	**2E-12**	1.54E-10
WIF1	1.50E-01	3.45E-01
GZMB	0.18	0.390422535
FCRLB	**0.00047**	0.003015833
ANXA1	**0.0086**	0.032302439
Fralpha	0.11	0.268888889
Side	0.115633312	0.278242658
diff	0.459264924	0.676448598
muc_adeno_old	0.881345108	0.94254963
SEX	0.578654307	0.763418803
DUKES	**0.009622212**	0.034460946
location	0.428047558	0.655643564

Total selected	49

APPENDIX 12. Selected features using RSF feature selection.

The Table 16 shows the features selected by variable importance (VIMP) technique for each of the three (3) datasets individually in a decreasing order of VIMP. VIMP is run for ten individual times and the variable importance are calculated as an average of the values of all iterations. Then the Table 17 demonstrates the features chosen for these three (3) datasets selected using the minimal depth (md) approach. Finally, the common features selected by both of these approaches are identified for each of the datasets. The results are presented in Table 18. Duplicate values between the datasets are shown in Table 7 in chapter 3.2.3.3. For CPH modelling only the variables selected by both md and VIMP are chosen.

Table 16. Features selected using RSF feature selection with VIMP.

Listwise deletion		**Median imputation**		**kNN-imputation**	
variable	VIMP	variable	VIMP	variable	VIMP
MUC16	0.02113	DUKES	0.06662	DUKES	0.01565
DUKES	0.01155	CEA	0.01667	S100A11	0.01536
SYND1	0.00698	MUC16	0.00407	MUC16	0.01210
HGF1	0.00608	AgeAtOper	0.00312	AREG	0.01165
LYPD3	0.00532	IL_6	0.00311	SYND1	0.01152
Frgamma	0.00405	MMP8ngml	0.00241	Frgamma	0.01042
CEACAM1	0.00234	IL_8	0.00201	KLK13	0.01017
KLK13	0.00205	IL_10	0.00191	aSNT	0.00827
IGF1R	0.00144	IP_10	0.00189	IGF1R	0.00688
ADAM8	0.00126	LYPD3	0.00178	WISP1	0.00684
FCRLB	0.00123	KLK13	0.00156	FURIN	0.00571
HGF	0.00088	SYND1	0.00149	VIM	0.00498
WISP1	0.00086	S100A11	0.00134	SPARC	0.00450
PGFalpha	0.00085	PDGF_bb	0.00130	SCF1	0.00372
EPHA2	0.00083	TIMP1ngml	0.00129	MIA	0.00289
IL_2Ra	0.00083	CTACK	0.00123	CD160	0.00288
hk11	0.00080	MIG	0.00100	CYR61	0.00286
GROa	0.00080	HGF	0.00099	ANXA1	0.00263
SCGF_b	0.00060	diff	0.00098	CEACAM5	0.00249
MIP_1b	0.00057	IL_1b	0.00095	MICAB	0.00243
G_CSF	0.00057	HGF1	0.00088	CEA	0.00236
IL_12p70	0.00049	CEACAM1	0.00086	CXCL13	0.00231
IL_18	0.00048	IL_2Ra	0.00086	GPNMB	0.00196
S100A11	0.00047	MMP8TIMP1molratio	0.00086	VEGFA	0.00195
MSLN	0.00046	AREG	0.00084	EPHA2	0.00190
MMP8ngml	0.00036	VIM	0.00079	CPC1	0.00187
PVRL4	0.00034	MIP_1b	0.00067	ESM1	0.00185

CPC1	0.00034	CEACAM5	0.00064	PODXL	0.00173
RANTES	0.00033	IFNgammaR1	0.00062	TFPI2	0.00160
CXCL13	0.00032	CPC1	0.00057	hk11	0.00139
TIMP1ngml	0.00024	IL_18	0.00056	CD48	0.00127
IL_8	0.00023	EPHA2	0.00052	FCRLB	0.00123
IFNgam-maR1	0.00023	CTSV	0.00047	TNFSF13	0.00111
IL_7	0.00023	ADAMTS15	0.00043	TIMP1ngml	0.00107
PPY	0.00023	FURIN	0.00041	MMP8ngml	0.00102
IL_9	0.00021	IL_9	0.00039	VEGFR3	0.00101
ERBB2	0.00016	ADAM8	0.00035	CRNN	0.00094
CD27	0.00013	DLL1	0.00035	IL_6	0.00093
CPE	0.00011	VEGFR3	0.00033	IFNgammaR1	0.00089
DLL1	0.00008	PGFalpha	0.00033	LYN	0.00082
Side	0.00005	SCF1	0.00031	GZMH	0.00082
IL_6	0.00004	Frgamma	0.00029	LY9	0.00082
FURIN	0.00003	Eotaxin	0.00029	CD27	0.00079
DKN1A	0.00001	FGF_basic	0.00027	FASLG	0.00077
		PODXL	0.00026	PGFalpha	0.00075
		aSNT	0.00025	S100A4	0.00074
		MCSF	0.00024	location	0.00071
		IGF1R	0.00023	ERBB3	0.00069
		b_NGF	0.00023	ALB1	0.00068
		hk11	0.00023	CAIX	0.00067
		PVRL4	0.00022	ITGAV	0.00066
		ITGAV	0.00021	Side	0.00062
		WISP1	0.00020	ADAM8	0.00058
		TNFRSF6B	0.00016	DLL1	0.00058
		FCRLB	0.00016	SCAMP3	0.00057
		ERBB2	0.00015	LYPD3	0.00056
		Gal1	0.00014	HGF1	0.00053
		LY9	0.00013	ADAMTS15	0.00051
		CD27	0.00011	IL_8	0.00049
		TNF_a	0.00011	hk8	0.00048
		GZMH	0.00009	MSLN	0.00047
		TLR3	0.00008	RET	0.00045
		TNFSF13	0.00008	TNFRSF6B	0.00045
		ERBB3	0.00007	TNFRSF4	0.00043
		ESM1	0.00007	WFDC2	0.00038
		SCAMP3	0.00006	CXL17	0.00036
		CXCL13	0.00006	MK	0.00036
		hk8	0.00005	SEZ6L	0.00035
		Fralpha	0.00005	CTSV	0.00034
		ALB1	0.00005	TNFRSF19	0.00034
		TRAIL	0.00005	SEX	0.00033
		CD48	0.00005	CD207	0.00032
		IL_5	0.00005	FGFBP1	0.00032
		IL_4	0.00005	TCL1A	0.00031
		FGFBP1	0.00005	CEACAM1	0.00030
		MIA	0.00004	diff	0.00030

	MSLN	0.00003	PVRL4	0.00029
	CAIX	0.00003	TXLNA	0.00029
	SEX	0.00003	AgeAtOper	0.00029
	EGF	0.00002	MMP8TIMP1molra-tio	0.00028
	IL_12p70	0.00002	Gal1	0.00028
	ITGB5	0.00002	RSPO3	0.00025
	VEGFA	0.00002	MIP_1b	0.00025
	TXLNA	0.00002	WIF1	0.00024
	Side	0.00001	TLR3	0.00024
	TGFR2	0.00001	GZMB	0.00022
			IL_18	0.00022
			IL_2Ra	0.00021
			SDF_1a	0.00021
			ERBB2	0.00021
			FADD	0.00020
			TRAIL1	0.00019
			ERBB4	0.00018
			ICOSLG	0.00018
			PPY	0.00017
			ITGB5	0.00016
			TGFR2	0.00016
			IFN_a2	0.00014
			XPNPEP2	0.00014
			VEGFR2	0.00013
			MetAP2	0.00013
			MIG	0.00012
			IL_10	0.00012
			MADhomolog5	0.00012
			IL_7	0.00011
			IL_1b	0.00010
			CPE	0.00010
			HGF	0.00010
			EGF	0.00010
			PDGF_bb	0.00010
			IL_17	0.00009
			CD70	0.00009
			CTACK	0.00008
			IL_9	0.00008
			IL_4	0.00008
			GROa	0.00008
			SCF	0.00007
			TRAIL	0.00007
			DKN1A	0.00006
			MCSF	0.00006
			hk14	0.00005
			RANTES	0.00004
			FGF_basic	0.00004
			IL_15	0.00004
			Fralpha	0.00003

		GM_CSF	0.00003
		IL_3	0.00002
		b_NGF	0.00002
		G_CSF	0.00001
		IL_5	0.00001

Total selected features using VIMP (>0)		
44	**86**	**130**

Table 17. Features selected using RSF feature selection with md.

Listwise deletion	**Median imputation**	**kNN-imputation**
MUC16	DUKES	DUKES
TLR3	CEA	SYND1
LYPD3	IL_6	Frgamma
SYND1	IL_8	AREG
HGF1	AgeAtOper	KLK13
CEACAM1	MMP8ngml	MUC16
CEACAM5	IL_10	S100A11
DUKES	TIMP1ngml	SPARC
KLK13	IL_1b	WISP1
CD207	PDGF_bb	MIA
GROa	LYPD3	aSNT
HGF	MUC16	CEA
TIMP1ngml	diff	CEACAM5
ERBB2	CEACAM5	IGF1R
MIP_1b	IL_2Ra	SCF1
SCGF_b	IP_10	CD160
SPARC	CRPmgl	ANXA1
IL_9	HGF	MICAB
CPE	CTACK	VIM
AREG	MMP8TIMP1molra-tio	FURIN
CTSV	TNF_a	CYR61
FCRLB	FGF_basic	LYPD3
ICOSLG	IL_5	PODXL
MADhomolog5	KLK13	CPC1
MMP8ngml	IL_1a	TLR3
G_CSF	MIP_1b	PGFalpha
MetAP2	Eotaxin	IL_8
FASLG	TLR3	FCRLB
IL_10	b_NGF	CXCL13
MIF	SDF_1a	SCAMP3
IGF1R	IL_18	diff
TATIngml	CEACAM1	hk11
ERBB4	SCGF_b	IL_6
IL_7	AREG	FASLG
TXLNA	SYND1	GPNMB
IL_12p70	VEGF	ESM1

CD160	IL_9	TIMP1ngml
TCL1A	MCSF	MSLN
IL_2Ra	GROa	VEGFA
MICAB	IL_17	ITGAV
CPC1	SCF	AgeAtOper
GZMH	TRAIL	hk8
ITGAV	IL_7	CD48
IL_1ra	MMP9TIMP1molra-tio	TNFSF13
TRAIL	IFN_g	CD27
MSLN	CPC1	CEACAM1
aSNT	IL_15	LY9
ADAM8	TNF_b	
VEGFR3	ADAMTS15	
PGFalpha	MMP9ngml	
PODXL	PGFalpha	
TNF_a	ITGAV	
MMP8TIMP1molra-tio	MIP_1a	
CEA	GM_CSF	
LY9		
PPY		
TRAIL1		
S100A4		
CD48		
hk8		
CTACK		
Frgamma		
CXCL13		

Total selected features using md		
63	54	47

Table 18. Features selected using RSF feature selection with both VIMP and md.

Listwise deletion	**Median imputation**	**kNN-imputation**
MUC16	DUKES	DUKES
DUKES	CEA	S100A11
SYND1	MUC16	MUC16
HGF1	AgeAtOper	AREG
LYPD3	IL_6	SYND1
Frgamma	MMP8ngml	Frgamma
CEACAM1	IL_8	KLK13
KLK13	IL_10	aSNT
IGF1R	IP_10	IGF1R
ADAM8	LYPD3	WISP1
FCRLB	KLK13	FURIN
HGF	SYND1	VIM

PGFalpha	PDGF_bb	SPARC
IL_2Ra	TIMP1ngml	SCF1
GROa	CTACK	MIA
SCGF_b	HGF	CD160
MIP_1b	diff	CYR61
G_CSF	IL_1b	ANXA1
IL_12p70	CEACAM1	CEACAM5
MSLN	IL_2Ra	MICAB
MMP8ngml	MMP8TIMP1molra-tio	CEA
CPC1	AREG	CXCL13
CXCL13	MIP_1b	GPNMB
TIMP1ngml	CEACAM5	VEGFA
IL_7	CPC1	CPC1
PPY	IL_18	ESM1
IL_9	ADAMTS15	PODXL
ERBB2	IL_9	hk11
CPE	PGFalpha	CD48
	Eotaxin	FCRLB
	FGF_basic	TNFSF13
	MCSF	TIMP1ngml
	b_NGF	IL_6
	ITGAV	LY9
	TNF_a	CD27
	TLR3	FASLG
	TRAIL	PGFalpha
	IL_5	ITGAV
		SCAMP3
		LYPD3
		IL_8
		hk8
		MSLN
		CEACAM1
		diff
		AgeAtOper
		TLR3

Total selected features using VIMP and md		
29	38	47

APPENDIX 13. Kaplan-Meier curves.

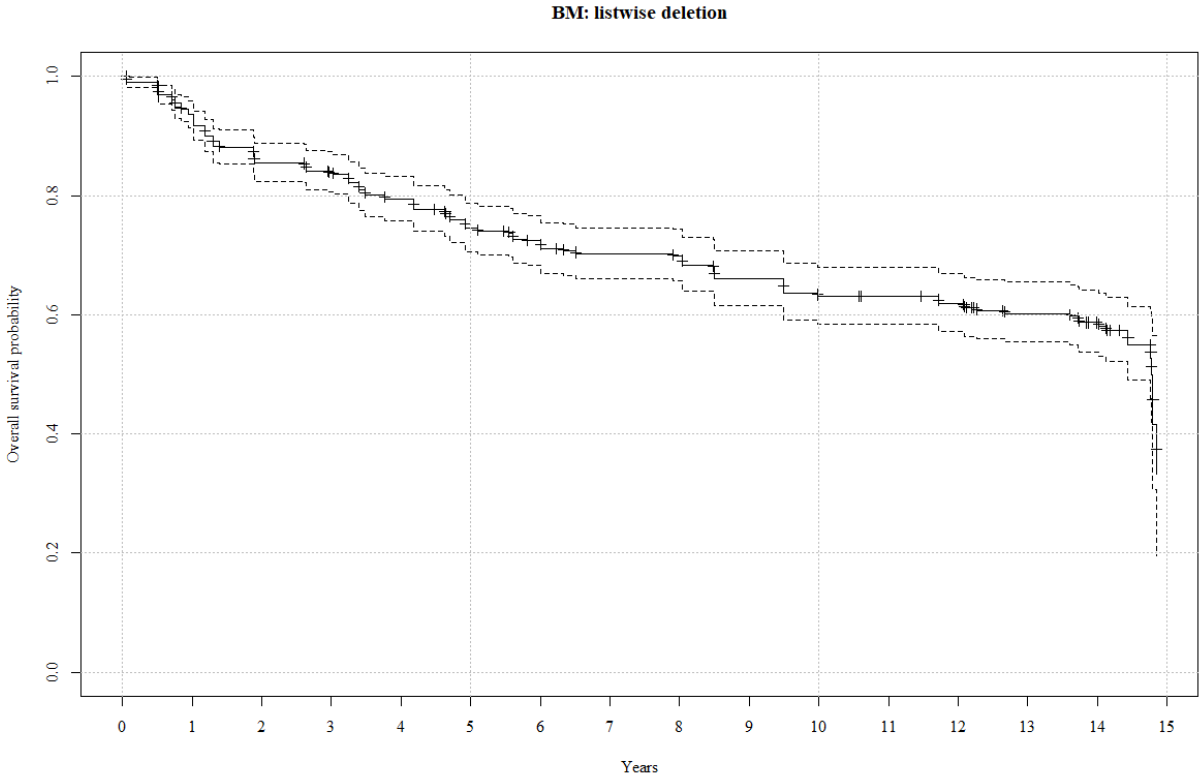


Figure 35. Kaplan-Meier curves for listwise deletion (BM) data.

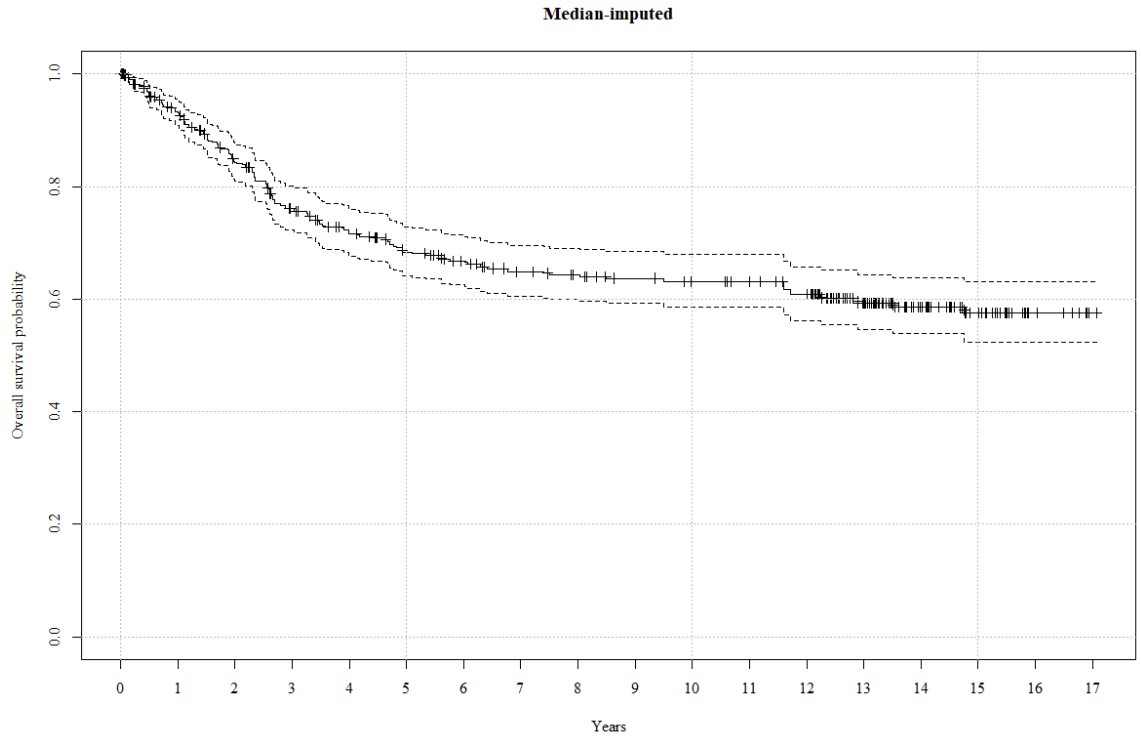


Figure 36. Kaplan-Meier curves for median imputed data.

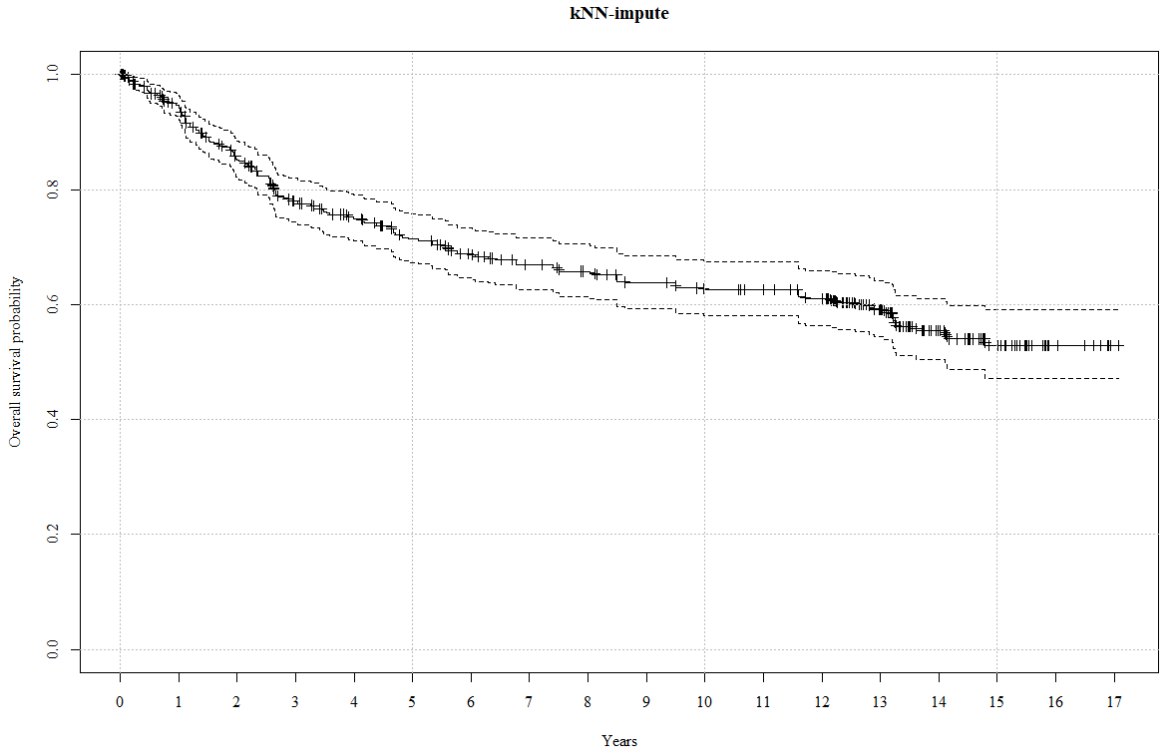


Figure 37. Kaplan-Meier curves for kNN-imputed data.

APPENDIX 14. Nelson-Aalen estimate curves.

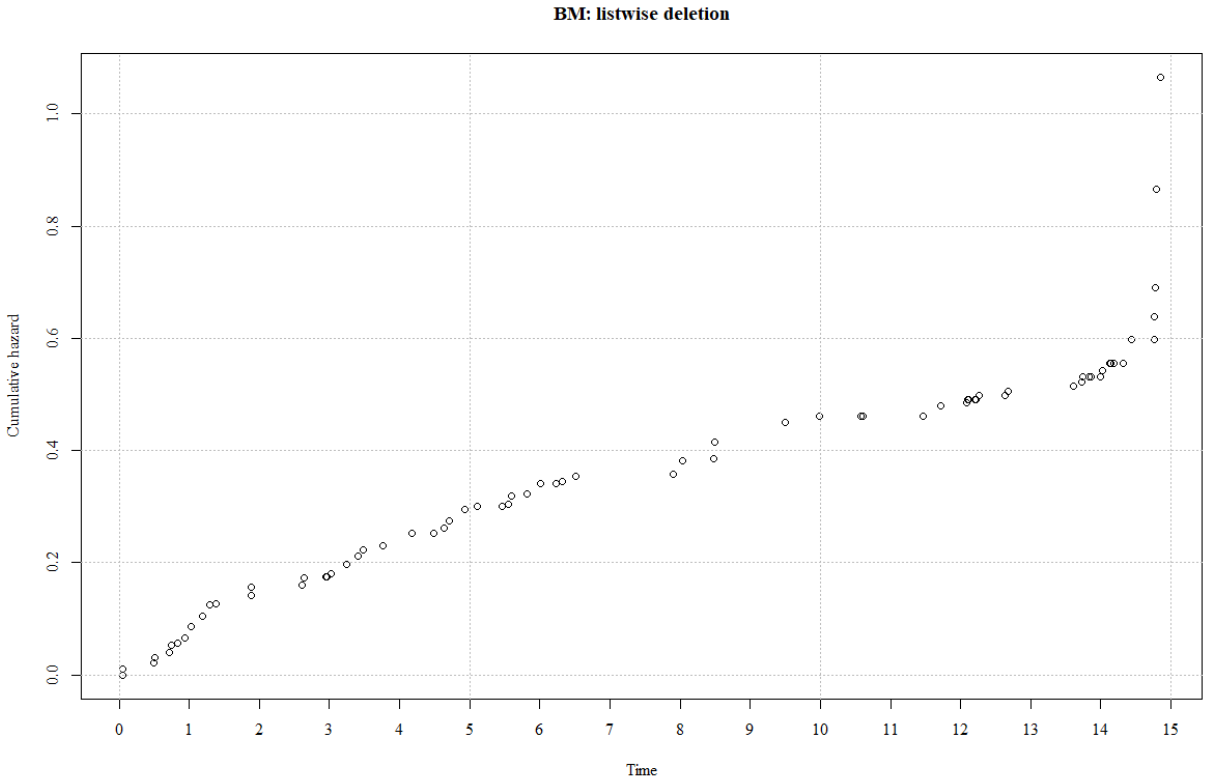


Figure 38. Nelson-Aalen estimate curve for listwise deletion (BM) data.

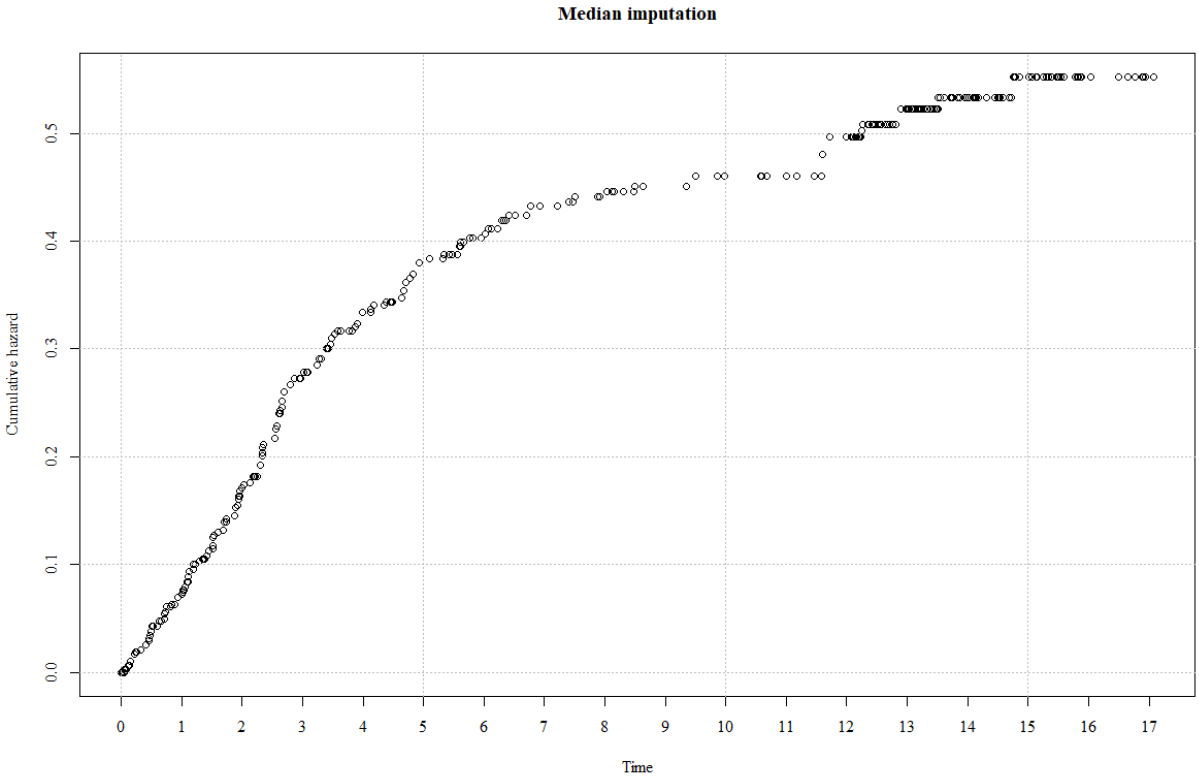


Figure 39. Nelson-Aalen estimate curve for median imputed data.

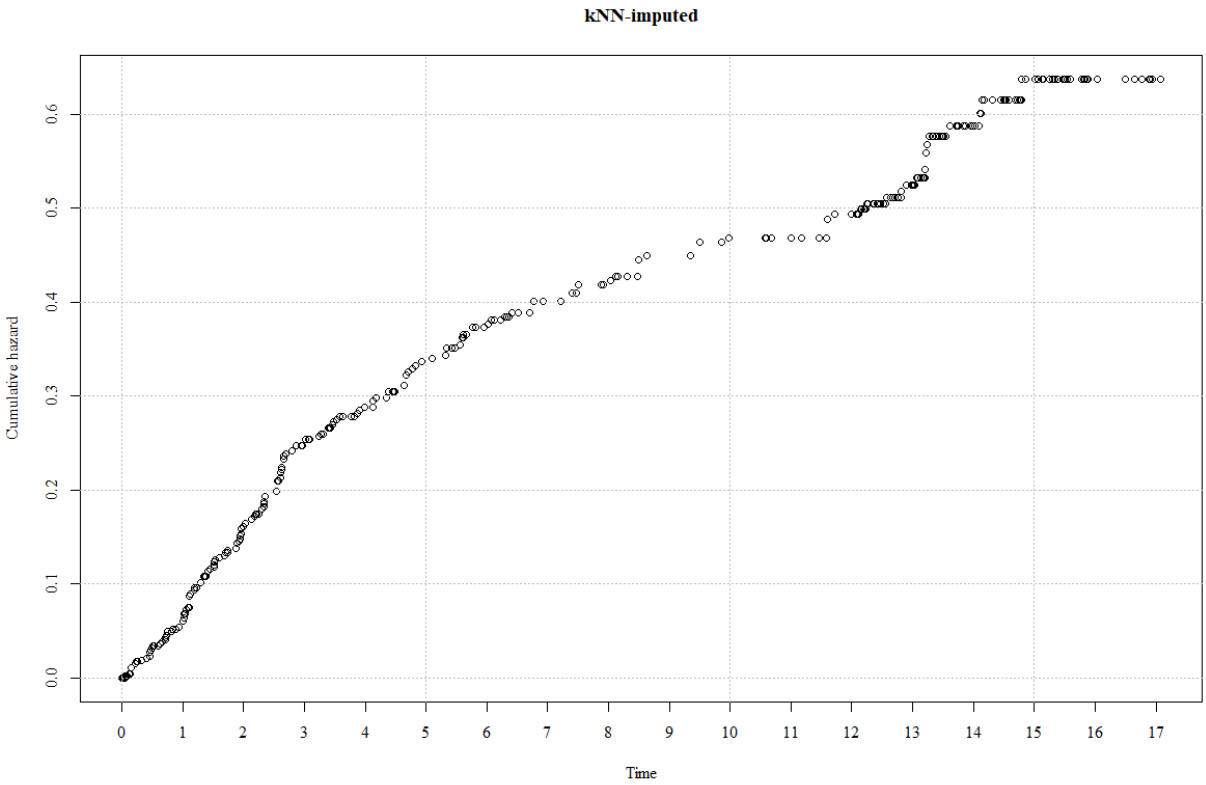


Figure 40. Nelson-Aalen estimate curve for kNN-imputed data.

APPENDIX 15. Features violating the proportional hazards assumption in CPH models.

Tables display the features violating the PH assumption in CPH models. There are separate tables for both splits and features selected using both univariate cox and random survival forests for each of the three datasets.

Univariate feature selection

85/15 split	listwise deletion	median	kNN
	-	-	hk8 MUC16 CYR61 TNFSF13 FCRLB hk11 DLL1 VEGFA SCF1 VIM

80/20 split	listwise deletion	median	kNN
	-	-	hk8 CYR61 FCRLB TNFSF13 MUC16 DLL1 hk11 VEGFA SCF1

Random survival forest for feature selection

85/15 split	listwise deletion	median	kNN
	IL_9	TLR3 ITGAV diff IL_5	MUC16 FCRLB CYR61 hk8 TNFSF13 hk11 CPC1 S100A11 TLR3 VIM

80/20 split	listwise deletion	median	kNN
	CEACAM1	TLR3 IL_5 diff ITGAV	FCRLB CYR61 MUC16 hk8 TNFSF13 hk11 CPC1 TLR3 VEGFA VIM

APPENDIX 16. Summarised results of holdout validated CPH models.

* These violating features are presented in a separate table in Appendix 11

** With listwise deletion data using the features selected by univariate Cox the CPH model reduced to a univariate model since only a single feature is included.

Table 19. Summarised results of CPH with split 85/15.

Train/Test split 85/15				
Imputation	Feature selection	PH violations*	Concordance	
			Train	Test
Listwise deletion (BM)	Univariate Cox**	No	0.525	0.702
	RSF	Yes (1)	0.639	0.474
Median	Univariate Cox	No	0.649	0.516
	RSF	Yes (4)	0.780	0.746
kNN	Univariate Cox	Yes (10)	0.817	0.756
	RSF	Yes (10)	0.822	0.725

Table 20. Significant markers identified by CPH with 85/15 split.

Imputa-tion	Feature selection	Significant markers (p-value < 0.1)
Listwise deletion (BM)	Univariate Cox**	-
	RSF	MIP_1b, Dukes, SYND1, CPC1, IGF1R
Median	Univariate Cox	CEA, MMP8ngml
	RSF	IL_2Ra, CEA, AgeAtOper, Dukes, MMP8ngml, MMP8TIMP1molratio, TLR3
kNN	Univariate Cox	IL_8, Dukes, MMP8ngml, KLK13, SPARC, CYR61, ALB1, MICAB, AREG, ADAMTS15, MUC16, ANXA1
	RSF	diff, AgeAtOper, Dukes, KLK13, SPARC, CYR61, CD27, VIM, CD160, MICAB, AREG, MUC16, ANXA1

Table 21. Summarised results of CPH with split 80/20.

Train/Test split 80/20				
Imputation	Feature selection	PH violations*	Concordance	
			Train	Test
Listwise deletion (BM)	Univariate Cox**	No	0.511	0.716
	RSF	Yes (1)	0.649	0.424
Median	Univariate Cox	No	0.642	0.555
	RSF	Yes (4)	0.774	0.741
kNN	Univariate Cox	Yes (10)	0.815	0.787
	RSF	Yes (10)	0.816	0.769

Table 22. Significant markers identified by CPH with 80/20 split.

Imputa-tion	Feature selection	Significant markers (p-value < 0.1)
Listwise deletion (BM)	Univariate Cox**	-
	RSF	MIP_1b, Dukes, SYND1, CPC1, PGFalpha
Median	Univariate Cox	CEA, MMP8ngml
	RSF	IL_2Ra, CEA, AgeAtOper, Dukes, MMP8ngml, MMP8TIMP1molratio, TLR3
kNN	Univariate Cox	IL_6, IL_8, Dukes, MMP8ngml, KLK13, SPARC, CYR61, ALB1, MICAB, AREG, ADAMTS15, MUC16
	RSF	IL_8, diff, AgeAtOper, Dukes, KLK13, SPARC, CYR61, VIM, CD160, MICAB, AREG, MUC16, ANXA1

APPENDIX 17. Summarised results from semi-stratified k-fold cross-validated CPH and RSF models.

In Table 23, for each validation approach the highest obtained concordance values for both train and test data are highlighted.

Table 23. Concordance values for semi-stratified k-fold cross-validated CPH and RSF models.

	Imputation	Feature selection / Splitting rule	Concordance			
			Train		Test	
			Average	Var	Average	Var
CPH 5-fold cv	Listwise deletion (BM)	Univariate Cox*	0.556	5.26E-04	0.549	0.008
		RSF	0.662	1.01E-04	0.507	0.002
	Median	Univariate Cox	0.642	6.41E-04	0.602	0.010
		RSF	0.786	6.78E-05	0.704	0.002
	kNN	Univariate Cox	0.819	4.17E-04	0.709	0.013
		RSF	0.819	4.59E-04	0.717	0.015
CPH 10-fold cv	Listwise deletion (BM)	Univariate Cox*	0.557	1.26E-04	0.553	0.011
		RSF	0.650	1.55E-04	0.535	0.004
	Median	Univariate Cox	0.640	1.94E-04	0.609	0.014
		RSF	0.782	4.98E-05	0.710	0.004
	kNN	Univariate Cox	0.813	1.18E-04	0.712	0.018
		RSF	0.813	1.20E-04	0.714	0.020
RSF 5-fold cv default	Listwise deletion	Log-rank	0.916	1.35E-05	0.553	0.001
		Gradient-based Brier	0.871	1.51E-04	0.572	0.002
	Median imputation	Log-rank	0.917	7.64E-06	0.698	0.001
		Gradient-based Brier	0.894	1.36E-05	0.698	0.002
	kNN-imputation	Log-rank	0.923	2.86E-06	0.756	0.008
		Gradient-based Brier	0.881	6.78E-05	0.734	0.007
RSF 5-fold cv tuned	Listwise deletion	Log-rank	0.912	3.55E-05	0.545	0.005
		Gradient-based Brier	0.882	3.35E-04	0.556	0.001
	Median imputation	Log-rank	0.911	1.59E-05	0.701	0.002
		Gradient-based Brier	0.903	5.27E-06	0.730	0.001
	kNN-imputation	Log-rank	0.912	9.36E-06	0.756	0.008
		Gradient-based Brier	0.891	4.08E-05	0.741	0.007
RSF 10-fold cv default	Listwise deletion	Log-rank	0.908	2.33E-05	0.567	0.008
		Gradient-based Brier	0.870	4.64E-05	0.555	0.010
	Median imputation	Log-rank	0.917	6.54E-06	0.699	0.003
		Gradient-based Brier	0.894	1.05E-05	0.696	0.004
	kNN-imputation	Log-rank	0.923	3.28E-06	0.761	0.014
		Gradient-based Brier	0.882	1.90E-05	0.745	0.012
RSF 10-fold cv tuned	Listwise deletion	Log-rank	0.908	3.67E-05	0.539	0.007
		Gradient-based Brier	0.888	8.97E-03	0.549	0.082
	Median imputation	Log-rank	0.909	3.48E-05	0.711	0.003
		Gradient-based Brier	0.904	1.13E-05	0.726	0.004
	kNN-imputation	Log-rank	0.912	6.15E-05	0.751	0.013
		Gradient-based Brier	0.896	1.05E-05	0.749	0.011

	Imputation	Feature selection / Splitting rule	Survival probability predictions					
			Year 1		Year 3		Year 5	
			Median	Std	Median	Std	Median	Std
CPH 5-fold cv	Listwise deletion (BM)	Univariate Cox*	0.938	*0.014*	0.843	*0.028*	0.749	*0.041*
		RSF	0.947	*0.043*	0.860	*0.098*	0.771	*0.140*
	Median	Univariate Cox	0.946	*0.074*	0.795	*0.141*	0.721	*0.157*
		RSF	0.968	*0.128*	0.849	*0.245*	0.776	*0.279*
	kNN	Univariate Cox	0.982	*0.137*	0.890	*0.272*	0.830	*0.303*
		RSF	0.984	*0.134*	0.897	*0.275*	0.842	*0.307*
CPH 10-fold cv	Listwise deletion (BM)	Univariate Cox*	0.937	*0.012*	0.841	*0.026*	0.749	*0.038*
		RSF	0.945	*0.037*	0.857	*0.087*	0.769	*0.126*
	Median	Univariate Cox	0.947	*0.074*	0.796	*0.138*	0.724	*0.156*
		RSF	0.964	*0.126*	0.838	*0.235*	0.759	*0.269*
	kNN	Univariate Cox	0.980	*0.137*	0.879	*0.261*	0.816	*0.291*
		RSF	0.984	*0.131*	0.898	*0.271*	0.842	*0.304*
RSF 5-fold cv default	Listwise deletion	Log-rank	0.925	*0.025*	0.817	*0.050*	0.712	*0.065*
		Gradient-based Brier	0.938	*0.011*	0.840	*0.017*	0.744	*0.020*
	Median imputation	Log-rank	0.935	*0.056*	0.748	*0.129*	0.665	*0.146*
		Gradient-based Brier	0.946	*0.025*	0.784	*0.086*	0.710	*0.104*
	kNN-imputation	Log-rank	0.943	*0.055*	0.759	*0.133*	0.676	*0.150*
		Gradient-based Brier	0.952	*0.024*	0.792	*0.092*	0.721	*0.110*
RSF 5-fold cv tuned	Listwise deletion	Log-rank	0.921	*0.059*	0.811	*0.101*	0.715	*0.124*
		Gradient-based Brier	0.941	*0.014*	0.849	*0.023*	0.762	*0.030*
	Median imputation	Log-rank	0.938	*0.109*	0.744	*0.205*	0.662	*0.229*
		Gradient-based Brier	0.934	*0.052*	0.761	*0.120*	0.678	*0.136*
	kNN-imputation	Log-rank	0.939	*0.074*	0.750	*0.164*	0.674	*0.178*
		Gradient-based Brier	0.959	*0.027*	0.811	*0.109*	0.743	*0.127*
RSF 10-fold cv default	Listwise deletion	Log-rank	0.925	*0.028*	0.812	*0.051*	0.709	*0.066*
		Gradient-based Brier	0.937	*0.010*	0.841	*0.016*	0.745	*0.019*
	Median imputation	Log-rank	0.925	*0.061*	0.727	*0.124*	0.649	*0.137*
		Gradient-based Brier	0.932	*0.023*	0.757	*0.069*	0.676	*0.082*
	kNN-imputation	Log-rank	0.947	*0.055*	0.765	*0.142*	0.681	*0.160*
		Gradient-based Brier	0.951	*0.025*	0.792	*0.103*	0.720	*0.123*
RSF 10-fold cv tuned	Listwise deletion	Log-rank	0.916	*0.056*	0.805	*0.089*	0.708	*0.109*
		Gradient-based Brier	0.939	*0.014*	0.847	*0.023*	0.758	*0.029*
	Median imputation	Log-rank	0.942	*0.109*	0.742	*0.203*	0.672	*0.223*
		Gradient-based Brier	0.939	*0.045*	0.777	*0.115*	0.698	*0.131*
	kNN-imputation	Log-rank	0.943	*0.076*	0.750	*0.169*	0.666	*0.186*
		Gradient-based Brier	0.955	*0.032*	0.803	*0.122*	0.733	*0.142*

APPENDIX 18. Significant markers identified by CPH.

Table 24 displays the frequency of features selected as significant marker in CPH. Total of twelve (12) models were built using the holdout method for validation, thus the maximal value of frequency is twelve. The table is arranged in an order of decreasing frequency.

Table 24. Significant markers identified by CPH.

Variable	Frequency
Dukes	8
MMP8ngml	6
MUC16	4
KLK13	4
AREG	4
SPARC	4
CYR61	4
AgeAtOper	4
CEA	3
MICAB	3
IL_8	3
ANXA1	3
diff	2
ADAMTS15	2
MMP8TIMP1molratio	2
SYND1	2
CPC1	2
TLR3	2
CD160	2
ALB1	2
VIM	2
IL_2Ra	2
CD27	1
PGFalpha	1
IL_6	1
MIP_1b	1
IGF1R	1

APPENDIX 19. Summary output for the CPH model with features chosen using RSF feature selection and fitted with kNN-imputed data 80/20 split ratio.

Call:
coxph(formula = Surv(DSS_SurvivalTime, DSS_censor) ~., data = trn,
x = TRUE, y = TRUE)

n = 400, number of events = 147

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
IL_6	0.294987	1.343109	0.284504	1.037	0.299806	
IL_8	-0.00447	0.995541	0.002696	-1.658	0.097337	.
CEA	-0.09748	0.90712	0.165252	-0.59	0.555265	
diff2	0.846133	2.330618	0.493012	1.716	0.086116	.
diff3	0.951098	2.58855	0.567002	1.677	0.093461	.
diff4	1.907369	6.735347	0.6119	3.117	0.001826	**
AgeAtOper	0.019444	1.019634	0.009566	2.033	0.042103	*
DUKES2	0.953397	2.594507	0.400834	2.379	0.017382	*
DUKES3	1.375678	3.957759	0.376797	3.651	0.000261	***
DUKES4	1.652552	5.220286	0.433235	3.814	0.000136	***
TIMP1ngml	0.002205	1.002207	0.001384	1.593	0.111144	
VEGFA	0.020249	1.020455	0.374468	0.054	0.956876	
KLK13	-0.70578	0.493724	0.225561	-3.129	0.001754	**
CEACAM1	-0.18334	0.832484	1.056244	-0.174	0.862196	
MSLN	-0.27736	0.757781	0.227316	-1.22	0.222405	
TNFSF13	-0.44959	0.637888	0.482605	-0.932	0.351545	
SYND1	0.271437	1.311848	0.284734	0.953	0.340438	
CD48	0.317228	1.373315	0.497843	0.637	0.523991	
SCAMP3	0.213946	1.238556	0.18145	1.179	0.238363	
LY9	0.519135	1.680573	0.441917	1.175	0.240101	
ITGAV	-0.14672	0.863535	0.521776	-0.281	0.77856	
hk11	0.324664	1.383566	0.359273	0.904	0.36617	
CPC1	-0.80225	0.44832	0.495286	-1.62	0.105282	
hk8	-0.0829	0.920441	0.301143	-0.275	0.783092	
LYPD3	-0.13562	0.873171	0.331578	-0.409	0.682522	
PODXL	-1.18051	0.307123	0.737627	-1.6	0.109507	
IGF1R	0.701244	2.016259	0.539997	1.299	0.194078	
SCF1	0.276241	1.318165	0.334185	0.827	0.408458	
SPARC	-2.77859	0.062126	0.929694	-2.989	0.002802	**
PGFalpha	0.061145	1.063053	0.548386	0.112	0.911219	
FURIN	0.468972	1.59835	0.529821	0.885	0.376075	
CYR61	1.324871	3.761699	0.34307	3.862	0.000113	***
FASLG	-0.06075	0.941061	0.351672	-0.173	0.862858	
GPNMB	0.439807	1.552407	0.883561	0.498	0.618649	
MIA	-0.64815	0.523011	0.494947	-1.31	0.190351	
CD27	-0.84687	0.428757	0.523709	-1.617	0.105867	
aSNT	0.132933	1.142174	0.186938	0.711	0.477017	
TLR3	-0.29107	0.747464	0.195989	-1.485	0.13751	

VIM	-0.48886	0.613326	0.259004	-1.887	0.059099	.
CD160	0.417897	1.518764	0.25357	1.648	0.099342	.
MICAB	0.243515	1.275725	0.081577	2.985	0.002835	**
WISP1	0.430631	1.538227	0.326319	1.32	0.186948	
S100A11	0.205896	1.228625	0.413946	0.497	0.618908	
AREG	0.679331	1.972558	0.198736	3.418	0.00063	***
ESM1	-0.52964	0.588817	0.341807	-1.55	0.121255	
CXCL13	-0.00245	0.997557	0.197121	-0.012	0.990099	
Frgamma	-0.00974	0.990309	0.081579	-0.119	0.904979	
CEACAM5	0.041794	1.042679	0.112306	0.372	0.709787	
MUC16	0.445622	1.561461	0.196338	2.27	0.023228	*
FCRLB	0.035075	1.035697	0.184201	0.19	0.848984	
ANXA1	0.646676	1.909185	0.364421	1.775	0.075975	.

Signif. codes 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

	exp(coef)	exp(-coef)	lower 0.95	upper 0.95
IL_6	1.34311	0.7445	0.76903	2.3457
IL_8	0.99554	1.0045	0.9903	1.0008
CEA	0.90712	1.1024	0.65615	1.2541
diff2	2.33062	0.4291	0.88679	6.1252
diff3	2.58855	0.3863	0.85197	7.8648
diff4	6.73535	0.1485	2.03006	22.3465
AgeAtOper	1.01963	0.9807	1.00069	1.0389
DUKES2	2.59451	0.3854	1.18267	5.6917
DUKES3	3.95776	0.2527	1.89112	8.2828
DUKES4	5.22029	0.1916	2.23318	12.203
TIMP1ngml	1.00221	0.9978	0.99949	1.0049
VEGFA	1.02046	0.98	0.48983	2.1259
KLK13	0.49372	2.0254	0.31731	0.7682
CEACAM1	0.83248	1.2012	0.10503	6.5986
MSLN	0.75778	1.3196	0.48535	1.1831
TNFSF13	0.63789	1.5677	0.24771	1.6426
SYND1	1.31185	0.7623	0.75079	2.2922
CD48	1.37332	0.7282	0.51761	3.6436
SCAMP3	1.23856	0.8074	0.86789	1.7675
LY9	1.68057	0.595	0.7068	3.9959
ITGAV	0.86353	1.158	0.31056	2.4011
hk11	1.38357	0.7228	0.68421	2.7978
CPC1	0.44832	2.2305	0.16982	1.1835
hk8	0.92044	1.0864	0.51011	1.6608
LYPD3	0.87317	1.1453	0.45589	1.6724
PODXL	0.30712	3.256	0.07235	1.3037
IGF1R	2.01626	0.496	0.69968	5.8102
SCF1	1.31817	0.7586	0.68472	2.5376
SPARC	0.06213	16.0963	0.01004	0.3843
PGFalpha	1.06305	0.9407	0.36288	3.1142
FURIN	1.59835	0.6256	0.56583	4.515
CYR61	3.7617	0.2658	1.92027	7.369
FASLG	0.94106	1.0626	0.47236	1.8748

GPNMB	1.55241	0.6442	0.27474	8.7719
MIA	0.52301	1.912	0.19825	1.3798
CD27	0.42876	2.3323	0.15361	1.1967
aSNT	1.14217	0.8755	0.79179	1.6476
TLR3	0.74746	1.3379	0.50905	1.0975
VIM	0.61333	1.6305	0.36917	1.019
CD160	1.51876	0.6584	0.92396	2.4965
MICAB	1.27573	0.7839	1.08722	1.4969
WISP1	1.53823	0.6501	0.81144	2.916
S100A11	1.22863	0.8139	0.54584	2.7655
AREG	1.97256	0.507	1.33618	2.912
ESM1	0.58882	1.6983	0.30132	1.1506
CXCL13	0.99756	1.0024	0.67787	1.468
Frgamma	0.99031	1.0098	0.84398	1.162
CEACAM5	1.04268	0.9591	0.83667	1.2994
MUC16	1.56146	0.6404	1.06269	2.2943
FCRLB	1.0357	0.9655	0.72184	1.486
ANXA1	1.90918	0.5238	0.93466	3.8998

Concord-ance	= 0.816	(se = 0.019)	
Likelihood ratio test	= 216 on 51 df,		p=<2e-16
Wald test	= 180.2 on 51 df,		p=3e-16
Score (logrank) test	= 239.5 on 51 df,		p=<2e-16

APPENDIX 20. Summarised results of holdout validated RSF models.

The tables present the used terminal node size (nodesize) and number of variables randomly selected as candidates for a splitting a node (mtry) values for fitting RSF models. The resulting concordance (C-index) values on train and tests data are displayed. Also, the survival predictions calculated on the test data are demonstrated. Separate tables are for different splits (80/20 and 85/15). Additionally, for values for nodesize and mtry both default values and values obtained using a tuning function are tested. Below each table describing the models built using specific splits a table presenting the identified significant markers for those models.

* Features selected using both the minimal depth (md) approach and the variable importance (VIMP) over ten (10) individual iterations.

Table 25. Summarised results of RSF with split 80/20 and default values for nodesize and mtry.

	Train/Test split 80/20					
	Imputation	Splitting rule	Nodesize	mtry	Concordance	
					Train	Test
Default: Terminal node size set to 15 and mtry as a squareroot of number of covariates.	Listwise deletion	Log-rank	15	12	0.909	0.559
		Gradient-based Brier	15	12	0.885	0.636
	Median imputation	Log-rank	15	12	0.912	0.695
		Gradient-based Brier	15	12	0.895	0.720
	kNN-imputation	Log-rank	15	12	0.923	0.766
		Gradient-based Brier	15	12	0.887	0.722

Table 26. Significant markers identified by RSF with 80/20 split and default values for nodesize and mtry.

Imputation	Splitting rule	Significant markers*	Nro selected
Listwise deletion (BM)	Log-rank	LIF, RET, ANXA1, MCP_3, MIP_1b, HGF, EPHA2, ESM1, LY9, HGF1, SEZ6L, LYPD3, SYND1, ICOSLG, DLL1, MIF, FADD, IL_9, hk14, GZMB, RANTES, b_NGF	22
	Gradient-based Brier	HGF, TNFRSF4, CD48, IL_3, FADD, AREG, CEACAM1, CD207, IL_5, FGF_basic, MetAP2, TXLNA, MADhomolog5, IL_17, FASLG, TNF_b, MMP9TIMP1molratio, ADAMTS15, CEA, SDF_1a, IL_2Ra, SYND1, WISP1, IL_9, Fralpha, IFN_g	26
Median	Log-rank	DUKES, CEA, IP_10, AgeAtOper, TIMP1ngml, MMP8ngml, IL_9, SCGF_b, b_NGF, SDF_1a, MIG, TGFR2, IL_8, PDGF_bb, IL_10, IL_6, VEGFR3, CTACK, ITGAV, ADAM8, TRAIL, WFDC2, IL_5, CRPmgl, MMP9ngml, TXLNA, TNF_b, MIP_1b, MCSF, MMP8TIMP1molratio, IL_17, IL_1b, IL_1a, CRNN, FASLG, GROa, MADhomolog5, FGF_basic, IFN_a2, HGF, IL_2Ra, PODXL, IL_3, ERBB3, GZMB, IFN_g, MMP9TIMP1molratio, hk14, VEGF, IFNgammaR1, IL_7, SCF1, ADAMTS15	53
	Gradient-based Brier	DUKES, CEA, AgeAtOper, MIG, IL_8, MIP_1b, TRAIL, TNF_b, IL_1b, IP_10, IL_12p70, IL_13, MCSF, CRPmgl, TIMP1ngml, TATIngml, HGF, MMP8ngml, IL_16, MCP_3, IL_9, MIF, SDF_1a, MMP9ngml, IL_4, IL_2Ra, IL_10, RANTES, PDGF_bb, IL_1a, b_NGF, MIP_1a, IL_7, TNF_a, G_CSF, SCGF_b, S100A11, TNFSF13, aSNT, IL_6, IFN_a2, VEGF, MMP8TIMP1molratio, MMP9TIMP1molratio, LIF, IL_1ra, IL_3, MCP_1, IL_15, MSLN, IL_18, IL_5, FGF_basic, Gal1	54
kNN	Log-rank	Frgamma, AREG, S100A11, aSNT, KLK13, DUKES, MUC16, SPARC, SYND1, MIA, IGF1R, SCF1, VEGFA, CYR61, VEGFR3, GPNMB	16
	Gradient-based Brier	S100A11, Frgamma, AREG, TNFRSF6B, IGF1R, aSNT, GPNMB, DLL1, WISP1, CYR61, MUC16, EPHA2, WFDC2, KLK13, hk11, VEGFA, ADAM8, VIM, SCAMP3, TRAIL1, S100A4, VEGFR2, PGFalpha, CEACAM5, MK, WIF1, ESM1, FCRLB, Fralpha, SPARC, ADAMTS15, SCF1, MIA, IFNgammaR1, CD160, ALB1, TLR3, LY9, CXL17, CAIX, GZMH, FASLG, DUKES, CRNN, MIP_1b, TNFSF13, PVRL4, ANXA1, CEACAM1, HGF1, CEA, CPE, ITGAV, FGFBP1, FURIN, TXLNA	56

Table 27. Summarised results of RSF with split 80/20 and tuned values for nodesize and mtry.

	Train/Test split 80/20					
	Imputation	Splitting rule	Nodesize	mtry	Concordance	
					Train	Test
Terminal node size and mtry defined by tuning algorithm from rfsrc package.	Listwise deletion	Log-rank	20	97	0.928	0.525
		Gradient-based Brier	25	153	0.782	0.437
	Median imputation	Log-rank	20	151	0.911	0.739
		Gradient-based Brier	8	154	0.902	0.748
	kNN-imputation	Log-rank	20	151	0.932	0.745
		Gradient-based Brier	20	97	0.887	0.699

Table 28. Significant markers identified by RSF with 80/20 split and tuned values for nodesize and mtry.

Imputation	Splitting rule	Significant markers*	Nro selected
Listwise deletion (BM)	Log-rank	ESM1, EPHA2, ANXA1, LIF, SEZ6L, MIP_1b, DLL1, IL_1a, RET, RANTES, hk11, MCP_3, ICOSLG, GZMB, S100A11, hk14, MMP9TIMP1molratio, MIF, LYPD3, FADD, HGF1, ITGB5, IL_7, SYND1, IL_8, TXLNA, VEGFR3, LY9, b_NGF, HGF, CD160, IL_13, ADAM8, ALB1, CPC1, FURIN, IL_2, PDGF_bb, TATIngml, WISP1, IP_10, IL_9, TNF_a, SCGF_b, ITGAV, GM_CSF, PVRL4, SPARC	48
	Gradient-based Brier	IL_1b, TXLNA, MetAP2, FCRLB, IL_3	5
Median	Log-rank	DUKES, IP_10, CRPmgl, TNF_a, TXLNA, LYN, MADhomolog5, CTACK, FASLG, IL_1b, IL_10, ITGAV, MMP8ngml, b_NGF, ERBB3, CRNN, SCF1, FADD, TCL1A, IL_6, IL_9, ADAMTS15, SDF_1a, GZMB, VEGF, FCRLB, AgeAtOper, IL_17, IFN_g, hk14, TGFR2, SCGF_b, Frgamma, PGFalpha, ADAM8, WISP1, IL_5, IL_3, KLK13, TNF_b, HGF, CYR61, TIMP1ngml, SYND1	44
	Gradient-based Brier	DUKES, CEA, IL_1b, AgeAtOper, IP_10, IL_13, IL_1ra, IL_8, TATIngml, IL_10, IL_4, IL_2, IL_5, IL_6, TNF_a, SDF_1a, Eotaxin, MIP_1b, IL_12p70, MIG, MCP_3, TRAIL, IL_1a, IL_9, TNF_b, IL_16, IFN_g, IL_7, RANTES, PDGF_bb, CPE, IL_3, CRPmgl, IL_15, IL_18, CTACK, MSLN, MCSF, MMP8ngml, MMP8TIMP1molratio, IL_2Ra, FGF_basic, HGF, MIF	44
kNN	Log-rank	Frgamma, S100A11, aSNT, DUKES, VEGFR3, SYND1, AREG, KLK13, SPARC, CRPmgl, hk14, diff, AgeAtOper, VEGF, IGF1R, ITGAV, TCL1A, Fralpha, TNF_a, MADhomolog5, LYN, MMP8TIMP1molratio, GROa, SCGF_b, CTACK, SCF1, ADAM8, MIA, PDGF_bb, MUC16, IL_1b, IP_10, FADD, TIMP1ngml, SCAMP3, PODXL, FASLG, IL_6, DLL1, CEA, CEACAM5, TXLNA, b_NGF, CD70, PGFalpha, ADAMTS15, CYR61, MICAB, TNFRSF19, IL_10, IL_3	51
	Gradient-based Brier	S100A11, TNFRSF6B, Frgamma, WISP1, AREG, DLL1, IFNgammaR1, CEA, aSNT, IGF1R, S100A4, VEGFR2, WFDC2, CYR61, IL12p40, EPHA2, GPNMB, KLK13, ADAMTS15, VEGFA, hk11, CEACAM5, MCP_1, TRAIL1, CPE, TLR3, Fralpha, ERBB4, TNF_b, SPARC, MUC16, CXL17, MetAP2, FASLG, PGFalpha, WIF1, IL_1a, CRNN, ITGAV, TFPI2, CTACK, ESM1, MSLN, SCF1, IL_6, CEACAM1, DUKES, SEZ6L	48

Table 29. Summarised results of RSF with split 85/15 and default values for nodesize and mtry.

	Train/Test split 85/15					
	Imputation	Splitting rule	Nodesize	mtry	Concordance	
					Train	Test
Default: Terminal node size set to 15 and mtry as a squareroot of number of covariates.	Listwise deletion	Log-rank	15	12	0.903	0.606
		Gradient-based Brier	15	12	0.853	0.614
	Median imputation	Log-rank	15	12	0.915	0.695
		Gradient-based Brier	15	12	0.898	0.704
	kNN-imputation	Log-rank	15	12	0.924	0.731
		Gradient-based Brier	15	12	0.889	0.684

Table 30. Significant markers identified by RSF with 85/15 split and default values for nodesize and mtry.

Imputation	Splitting rule	Significant markers*	Nro selected
Listwise deletion (BM)	Log-rank	LIF, MCP_3, RET, ANXA1, EPHA2, HGF, LYPD3, IL_1a, WISP1, ESM1, MIP_1b, hk14, HGF1, IL_2, DLL1, SEZ6L, hk11, ICOSLG, RANTES, IL_15, LY9, MIF, S100A11, IP_10, SPARC, IL_9, SYND1, IL_18, SCGF_b, PDGF_bb, b_NGF, CPC1, FADD, TGFR2, LYN, EGF, IL_2Ra, MUC16, IL_16, Eotaxin, MMP9TIMP1molratio, VEGFR3, IL_10, G_CSF, CD160, GM_CSF, Frgamma, CXCL13	48
	Gradient-based Brier	IFNgammaR1, TNFRSF4, MetAP2, MMP8ngml, HGF, SYND1, CPE, Fralpha, FCRLB, MIP_1a, TNF_b, VEGFR3, WISP1, RET, IL_5, ANXA1, TXLNA, CTSV, IL_13, CEA, CRPmgl	21
Median	Log-rank	DUKES, CEA, AgeAtOper, IP_10, SDF_1a, MMP8ngml, IL_9, TIMP1ngml, IL_8, SCGF_b, MIG, b_NGF, IL_6, PDGF_bb, IL_5, CRPmgl, ITGAV, IL_1b, IL_10, MCSF, ADAM8, GROa, MMP9ngml, CTACK, TXLNA, IL_17, TGFR2, MADhomolog5, MMP8TIMP1molratio, FASLG, IL_2Ra, IL_1a, FGF_basic, hk14, SCF1, MIP_1b, ERBB3, WFDC2, MMP9TIMP1molratio, TRAIL, PODXL, VEGFR3, IFN_a2, diff, TNF_a, TNF_b, IL_18, IFN_g, S100A11, MIF, CRNN, ADAMTS15, HGF, IFNgammaR1, CEACAM1, IL_7, IL_1ra	57
	Gradient-based Brier	DUKES, CEA, AgeAtOper, MIG, IL_8, MIP_1b, MCP_3, IL_13, IP_10, MCSF, IL_1b, MIF, TRAIL, SDF_1a, TIMP1ngml, IL_12p70, IL_10, IL_9, MMP8ngml, TNF_b, PDGF_bb, HGF	22
kNN	Log-rank	DUKES, Frgamma, AREG, S100A11, aSNT, KLK13, SYND1, MUC16, SPARC, IGF1R, MIA, VEGFA, SCAMP3, SCF1, PGFalpha, WISP1, VEGFR3, CYR61, TNFRSF19, CD160, ALB1, VIM, ANXA1, CEA, DLL1, diff, AgeAtOper, GPNMB, MMP8ngml, LYN, FURIN, CEACAM5, ADAM8, MICAB, Fralpha, TCL1A, FCRLB, CRNN, MADhomolog5, hk14, IL_8, ADAMTS15, MMP8TIMP1molratio, LY9, IL_6, TNFSF13, TXLNA, hk11, EPHA2, TIMP1ngml, GZMH, ITGAV, GROa	53
	Gradient-based Brier	S100A11, AREG, Frgamma, aSNT, IGF1R, TNFRSF6B, DLL1, VEGFA, CYR61, EPHA2, GPNMB, WISP1, VIM, hk11, WFDC2, MUC16, ADAM8, KLK13, SCAMP3, S100A4, SPARC, Fralpha, CEA, MK, ALB1, CD160, SCF1, PVRL4, TRAIL1, ESM1, MIP_1b, PGFalpha, WIF1, CEACAM5, TLR3, FASLG, DUKES, IFNgammaR1, MIA, RSPO3, VEGFR2, TNFSF13, TFPI2, LY9, CAIX, CRNN, MICAB, CEACAM1, TNFRSF4, ADAMTS15	50

Table 31. Summarised results of RSF with split 85/15 and tuned values for nodesize and mtry.

	Train/Test split 85/15					
	Imputation	Splitting rule	Nodesize	mtry	Concordance	
					Train	Test
Terminal node size and mtry defined by tuning algorithm from rfsrc package.	Listwise deletion	Log-rank	10	51	0.918	0.579
		Gradient-based Brier	10	51	0.862	0.594
	Median imputation	Log-rank	15	51	0.914	0.714
		Gradient-based Brier	15	51	0.901	0.727
	kNN-imputation	Log-rank	15	51	0.931	0.719
		Gradient-based Brier	15	51	0.892	0.695

Table 32. Significant markers identified by RSF with 85/15 split and tuned values for nodesize and mtry.

Imputation	Splitting rule	Significant markers*	Nro selected
Listwise deletion (BM)	Log-rank	EPHA2, LIF, ANXA1, ESM1, RET, RANTES, MIP_1b, hk11, SEZ6L, hk14, MCP_3, WISP1, DLL1, HGF, IL_15, IL_18, ICOSLG, LYPD3, S100A11, IL_2, LY9, MMP9TIMP1molratio, IL_9, MIF, IP_10, SYND1, VEGFR3, IL_7, SCGF_b, SPARC, MIA, TXLNA, ADAM8, ITGB5, CD160, IL_16, IL_6	37
	Gradient-based Brier	IFNgammaR1, TNFRSF4, MMP8ngml, FCRLB, SDF_1a, TXLNA, WISP1, CRPmgl, MetAP2, TNFRSF6B, FASLG, FGF_basic, Frgamma, SPARC, HGF	15
Median	Log-rank	DUKES, CEA, IP_10, CRPmgl, IL_10, AgeAtOper, MMP8ngml, TXLNA, SDF_1a, TNF_a, ITGAV, b_NGF, IL_1b, SCGF_b, IL_9, CTACK, ERBB3, IL_6, TIMP1ngml, MADhomolog5, ADAMTS15, ADAM8, hk14, TGFR2, IL_3, IL_17, MIG, CRNN, SCF1, IL_5, TCL1A, IFN_g, KLK13, HGF, FADD, TNF_b, CYR61, PODXL, LYN, FASLG	40
	Gradient-based Brier	DUKES, CEA, AgeAtOper, MIG, IL_8, IL_13, IP_10, IL_1b, MCP_3, TATIngml, TRAIL, MIP_1b, MCSF, IL_12p70, SDF_1a, IL_10, TNF_b, IL_16, TNF_a, IL_1a, IL_9, RANTES, IFN_g, MMP8ngml, S100A11, CRPmgl, MIF, MMP9TIMP1molratio, TNFSF13, TIMP1ngml, Eotaxin, IL_5, b_NGF, PDGF_bb, IL_3, IL_4, IL_7, SCGF_b, FGF_basic, IFN_a2, IL12p40, diff, IL_6, HGF, MIP_1a, CTACK, MSLN, GM_CSF, MMP9ngml, MMP8TIMP1molratio, IL_15, IL_1ra, IL_2	53
kNN	Log-rank	Frgamma, AREG, S100A11, DUKES, KLK13, aSNT, SPARC, SYND1, VEGFR3, diff, MIA, MUC16, SCF1, MMP8TIMP1molratio, AgeAtOper, IGF1R, ITGAV, hk14, LYN, TCL1A, PGFalpha, CEA, GROa, TIMP1ngml, SCGF_b, VEGF, TNF_a	27
	Gradient-based Brier	S100A11, Frgamma, AREG, TNFRSF6B, WISP1, aSNT, DLL1, IGF1R, CYR61, VEGFA, CEA, IFNgammaR1, S100A4, GPNMB, WFDC2, hk11, EPHA2, MUC16, TLR3, KLK13, SPARC, CEACAM5, VIM, IL12p40, WIF1, FASLG, ADAMTS15, VEGFR2, ESM1, GM_CSF, PGFalpha, CRNN, SCF1, TRAIL1, PPY, IL_1a, LY9, DUKES, TNFSF13, Fralpha, CPE, TFPI2, CAIX, CXL17, ITGAV, MIP_1b	46

APPENDIX 21. Survival curves from the RSF models.

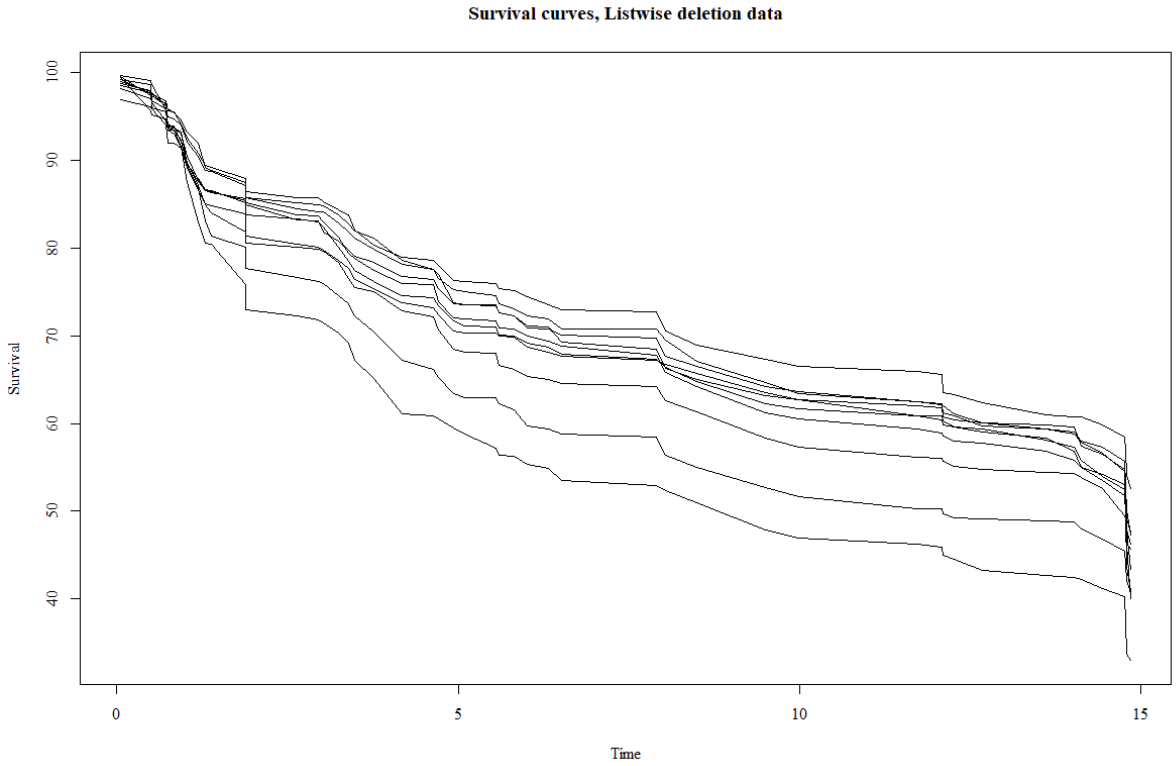


Figure 41. Survival curves for the first 10 patients of BM data.

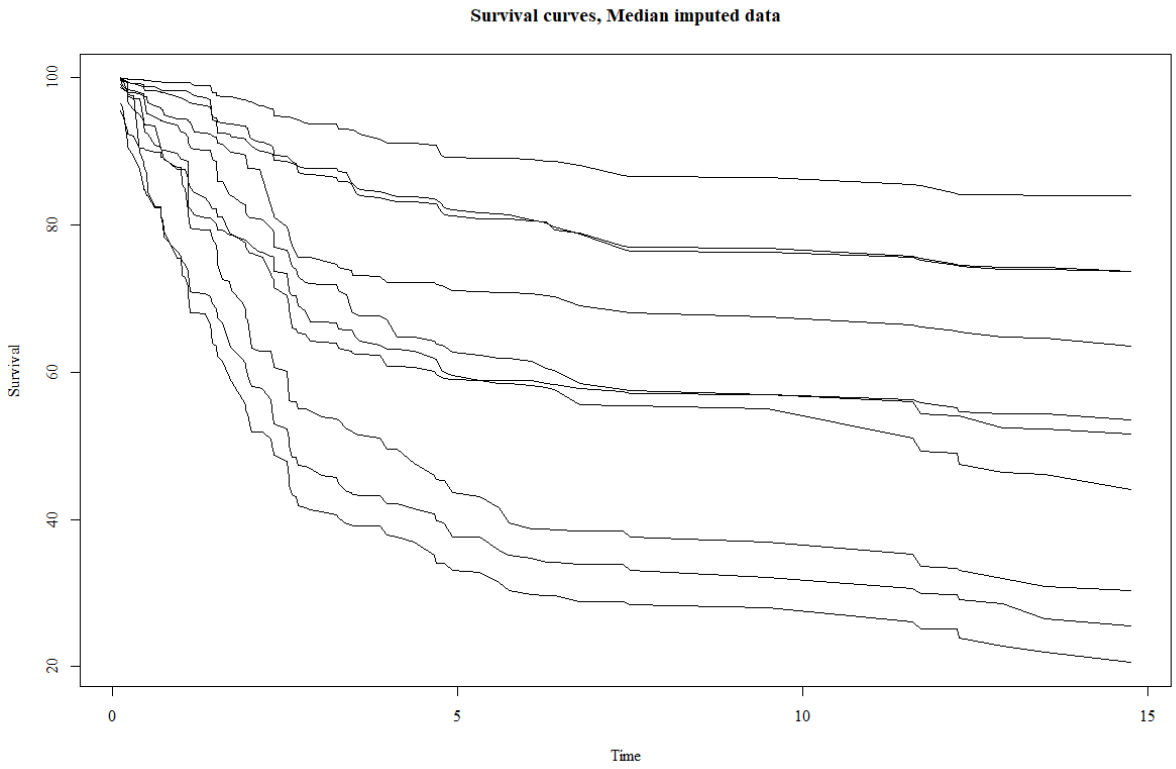


Figure 42. Survival curves for the first 10 patients of median imputed data.

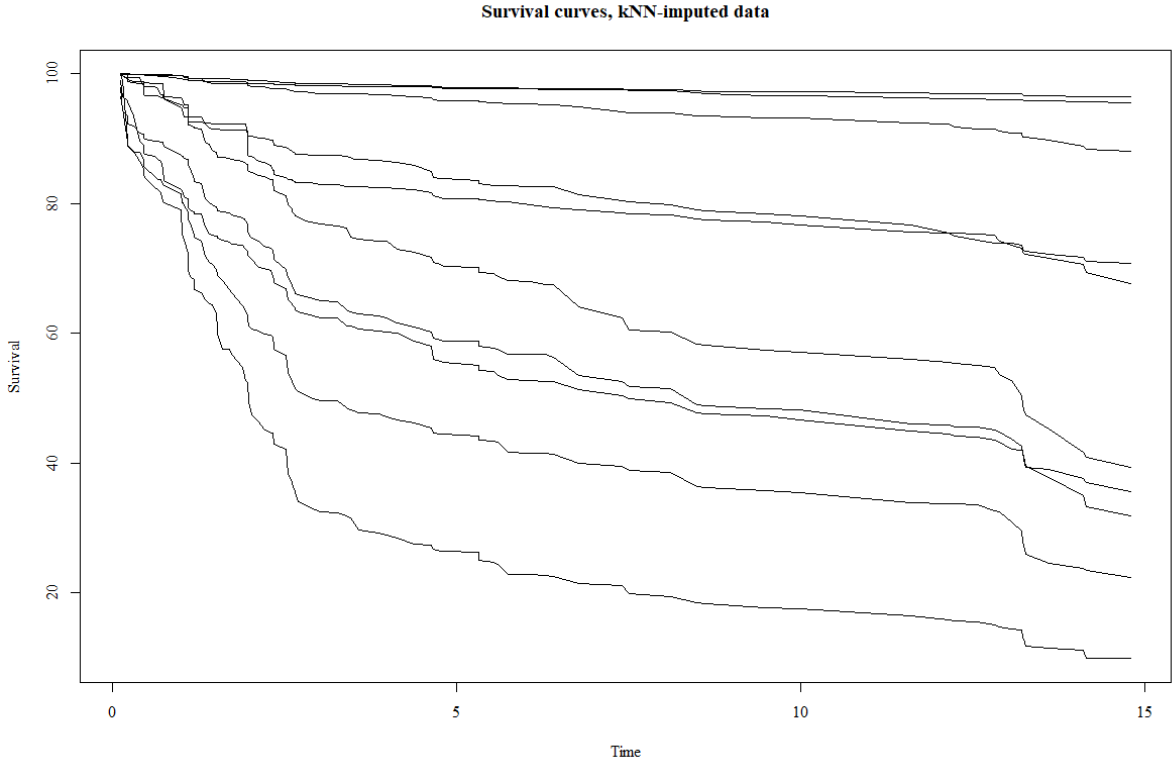


Figure 43. Survival curves for the first 10 patient of kNN-imputed data.

APPENDIX 22. Significant markers identified by RSF models.

Table 33 displays the frequency of features selected as significant marker in RSF. Total of 24 models were built, thus the setting the maximal value of frequency to 24. The table is arranged in an order of decreasing frequency.

Table 33. Significant markers identified by RSF.

Variable	Frequency
CEA	16
DUKES	16
HGF	15
S100A11	14
MIP_1b	13
IL_9	13
TXLNA	13
SPARC	12
SCF1	12
IP_10	12
WISP1	12
SCGF_b	11
hk14	11
SYND1	11
Frgamma	11
AgeAtOper	11
ITGAV	11
TNF_b	11
ADAMTS15	11
FASLG	11
MMP8ngml	11
IL_6	11
TIMP1ngml	10
IL_1b	10
DLL1	10
b_NGF	10
KLK13	10
CRPmgl	10
VEGFR3	10
IL_10	10
SDF_1a	10
ADAM8	10
PDGF_bb	9
TNF_a	9

IL_1a	9
CRNN	9
MUC16	9
aSNT	9
CYR61	9
MIF	9
AREG	9
IL_3	9
EPHA2	9
IL_5	9
CTACK	8
IFNgammaR1	8
IL_8	8
LY9	8
IGF1R	8
MCP_3	8
Fralpha	8
PGFalpha	8
MMP9TIMP1molratio	8
ESM1	8
MMP8TIMP1molratio	8
hk11	8
IFN_g	7
RANTES	7
FGF_basic	7
MIG	7
MIA	7
ANXA1	7
FADD	7
MADhomolog5	7
IL_7	7
IL_2Ra	6
TRAIL	6
GPNMB	6
LYN	6
WFDC2	6
MCSF	6
CEACAM5	6
CD160	6
TNFSF13	6
IL_13	6
FCRLB	6
VEGFA	6
TGFR2	5
CPE	5
TNFRSF6B	5
MetAP2	5
VEGF	5
IL_15	5

IL_17	5
SEZ6L	5
IL_18	5
diff	5
GROa	5
IL_16	5
TCL1A	5
CEACAM1	5
IL_2	5
RET	5
MSLN	4
GM_CSF	4
PODXL	4
ERBB3	4
TNFRSF4	4
S100A4	4
TRAIL1	4
SCAMP3	4
GZMB	4
IL_1ra	4
MMP9ngml	4
ALB1	4
LYPD3	4
TATIngml	4
ICOSLG	4
IL_12p70	4
IFN_a2	4
HGF1	4
VEGFR2	4
WIF1	4
VIM	4
TLR3	4
LIF	4
PVRL4	3
MIP_1a	3
IL12p40	3
CAIX	3
CXL17	3
Eotaxin	3
MICAB	3
TFPI2	3
FURIN	3
IL_4	3
CPC1	2
GZMH	2
G_CSF	2
ITGB5	2
MCP_1	2
TNFRSF19	2

MK	2
ERBB4	1
FGFBP1	1
CD48	1
CD70	1
CD207	1
RSPO3	1
CXCL13	1
PPY	1
CTSV	1
Gal1	1
EGF	1

APPENDIX 23. DeepSurv: a preliminary artificial neural network approach for survival analysis on CRC patient data.

DeepSurv as an ANN approach for survival analysis is applied for all three (3) datasets. 10-fold cross-validation is used as a resampling strategy. For tuning each individual learner three (3) custom autotuner functions are utilized with C-index optimization and 60 iteration random search. For tuning 85/15 and 80/20 split holdouts, as well as 10-fold cross-validation are performed. For DeepSurv, the maximum number of epochs is set to be ten (10), and the Adamax algorithm is used as an optimizer. Same number of nodes for each layer is assumed for clarity. These models are fitted using all available variables and no prior feature selection is used. For evaluation of performance the Harrell's C and Integrated Brier (aka integrated Graf) are calculated. Thus the model fits could possibly be suboptimal and are included here only to mention a possibility to apply ANNs for our CRC data.

The results are displayed in a Table 34. From there can be observed that this preliminary approach fails to perform with listwise deletion data, which is expected. The best performance of these models is obtained using the kNN-imputed data. This observation is in line with the notions made in section 4. However, the metrics show the predictive performance to be quite poor. With prior selection of used features and hyperparameter optimization, a higher predictive performance could be achieved. For now, with the initial setup the DeepSurv approach seems to suit our kNN-imputed CRC data best. This outcome ought to be considered with caution. Further analysis using these ANN-based techniques are left outside of the scope of this thesis. Further investigation about the possibilities of utilising these approaches to conduct survival analysis in the field of oncology could provide interesting results.

Table 34. DeepSurv initial results on all three (3) imputed and MICE enhanced datasets.

Autotuner	Imputation	Harrell's C	Integrated Brier
85/15 holdout, C-index optimization, 60 iteration random search	Listwise deletion	0.491	0.202
	Median imputed	0.598	0.199
	kNN-imputed	0.655	0.181
80/20 holdout, C-index optimization, 60 iteration random search	Listwise deletion	0.491	0.202
	Median imputed	0.598	0.199
	kNN-imputed	0.655	0.181
10-fold cv, C-index optimization, 60 iteration random search	Listwise deletion	0.517	0.201
	Median imputed	0.557	0.207
	kNN-imputed	0.650	0.185