Statistical and machine learning methods for the dynamic prediction of prognosis in haematological malignancies

Biccler, Jorne Lionel

DOI (link to publication from Publisher):
10.5278/vbn.phd.med.00127

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

STATISTICAL AND MACHINE LEARNING METHODS FOR THE DYNAMIC PREDICTION OF PROGNOSIS IN HAEMATOLOGICAL MALIGNANCIES

BY
JORNE LIONEL BICCLER

DISSERTATION SUBMITTED 2019

AALBORG UNIVERSITY
DENMARK

Statistical and machine learning methods for the dynamic prediction of prognosis in haematological malignancies

PhD Dissertation

Jorne Lionel Biccler

Dissertation submitted January 5, 2018

Dissertation submitted: January 5, 2019

PhD supervisor: Associate Professor, MD, DMSc,
Tarec C. El-Galaly
Department of Clinical Medicine
Aalborg University, Denmark

Assistant PhD supervisor: Professor, PhD, Martin Bøgsted
Department of Clinical Medicine
Aalborg University, Denmark

Professor, MD, PhD, Mats Jerkeman
Department of Oncology
Lund University, Sweden

MD, PhD, Peter de Nully Brown
Department of Haematology
Rigshospitalet, Denmark

PhD committee: Clinical Professor Soeren Paaske Johnsen (chairman)
Aalborg University

Professor Mette Noergaard
Aarhus University

Professor Paul Lambert
University of Leicester

PhD Series: Faculty of Medicine, Aalborg University

Department: Department of Clinical Medicine

ISSN (online): 2246-1302
ISBN (online): 978-87-7210-376-1

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2019

Preface

While studying for my bachelor and master degrees my interest in applied statistics started to become clear. It was, however, not until I started as a Ph.D. student at the Department of Haematology that my interest in statistical epidemiology became apparent. I would like to thank my supervisor, Tarec El-Galaly, for introducing me to the world of epidemiological research and haematological malignancies. Tarec, you always happily shared your knowledge on all the aspects of research and regularly gave me that little push in order to try something when you thought it would be beneficial. I feel privileged that you were my supervisor these past years.

I would also like to thank my co-supervisors who made invaluable contributions to the papers that make up this dissertation. First of all, I would like to thank Martin for the discussions regarding statistical issues and, occasionally, Hilbert spaces. Mats thank you for your feedback, providing Swedish register data, and some of the best salmon I have tasted. Peter thank you for sharing your knowledge on the structure of LYFO and occasionally spotting errors in my data selection procedures. In addition to my co-supervisors, I would like to thank Karin for her numerous contributions and convincing Tarec to pay my dinner that one time. Jan, Lene, and Ingrid, without you the papers on APL and HL would not have been what they are right now. I would also like to thank Tim Verdonck and Diego Villa who made my research visits at KU Leuven and British Columbia Cancer wonderful experiences.

Lasse, thank you for the interesting conversations with topics ranging from the mechanics of plumbing systems to relative survival models. Without you and our Taylor Swift jam sessions, my time at the department would have been a whole lot more boring and less (or more?) productive. Rasmus, your not so occasional cynical remarks were always appreciated and Mads thanks for reminding people about the dangers of putting shoes on top of the door.

Of course, I would like to thank my family. Freja thanks for your love, support, and patience.

Jorne Lionel Biccler
Aalborg University, January 5, 2019

Thesis outline

Thesis title:	Statistical and machine learning methods for the dynamic prediction of prognosis in haematological malignancies
PhD student:	Jorne Lionel Biccler.
Supervisors:	Clinical Professor, MD, DMSc, Tarec C. El-Galaly, Aalborg University
	Professor, PhD, Martin Bøgsted, Aalborg University
	Professor, MD, PhD, Mats Jerkeman, Lund University
	MD, PhD, Peter de Nully Brown, Rigshospitalet

This thesis is made up of the following papers

I) J.L. Biccler, S. Eloranta, P.dN. Brown, H. Frederiksen, M. Jerkeman, K.E. Smedby, M. Bøgsted, T.C. El-Galaly (2018). Simplicity at the cost of predictive accuracy in diffuse large B-cell lymphoma: a critical assessment of the R-IPI, IPI, and NCCN-IPI. Cancer Medicine. 7(1):114-122.

II) J.L. Biccler, L.S.G. Østgård, M.T. Severinsen, C.W. Marcher, P. Møller, C. Schöllkopf, L.S. Friis, M. Bøgsted, L.H. Jakobsen, T.C. El-Galaly, J.M. Nørgaard (2018). Evolution of relative survival for acute promyelocytic leukemia patients alive at landmark time-points: a population-based study. Leukemia 32, 2263-2265.

III) J.L. Biccler, S. Eloranta, P.dN. Brown, H. Frederiksen, M. Jerkeman, J. Jørgensen, L.H Jakobsen, K.E. Smedby, M. Bøgsted, T.C. El-Galaly (2018). Optimizing Outcome Prediction in Diffuse Large B-Cell Lymphoma by use of Machine Learning and Nationwide Lymphoma Registries: a Nordic Lymphoma Group Study. Clinical Cancer Informatics 2, 1-13.

IV) J.L. Biccler, T.C. El-Galaly, M. Bøgsted, J. Jørgensen, P.dN. Brown, C.B. Poulsen, J. Starklint, M.B. Juul, J.H. Christensen, P. Josefsson, A. Dessau, and L.H. Jakobsen (2018). Clinical prognostic scores are poor predictors of overall survival in various types of malignant lymphomas. Leukemia & Lymphoma (Early View).

V) J.L. Biccler, I. Glimelius, S. Eloranta, K.B. Smeland, P.dN. Brown, L.H. Jakobsen, H. Frederiksen, M. Jerkeman, A. Fosså, T.M.L. Andersson, H. Holte, M. Bøgsted, T.C. El-Galaly, K.E. Smedby. Relapse risk and loss of lifetime after modern combined modality treatment for young Hodgkin

lymphoma patients: A Nordic Lymphoma Epidemiology Group study. Journal of Clinical Oncology (Accepted).

VI) J.L. Biccler, M. Bøgsted, S. van Aelst, T. Verdonck. Outlier robust modeling of survival curves in the presence of potentially time-varying coefficients. In preparation.

A dashboard implementing the models introduced in papers I and III was developed and can be found on `https://lymphomapredictor.org`

In addition to the articles included in the thesis, the following papers not considered part of the thesis were published in the period during which the PhD took place.

1. M.A. Andersen, C.T. Eriksen, C. Brieghel, J.L. Biccler, C. da Cunha-Bang, M. Helleberg, C.U. Niemann (2018). Incidence and predictors of infection among patients prior to treatment of chronic lymphocytic leukemia: a Danish nationwide cohort study. Haematologica. 103(7): e300-e303.

2. T. Wästerlid, J.L. Biccler, P.dN. Brown, M. Bøgsted, G. Enblad, J. Mészáros Jørgensen, J.H. Christensen, B.E. Wahlin, K.E. Smedby, T.C. El-Galaly, M. Jerkeman (2018). Six cycles of R-CHOP-21 are not inferior to eight cycles for treatment of diffuse large B-cell lymphoma-A Nordic Lymphoma Group population-based study. Annals of Oncology. 29(8): 1882–1883.

3. C. K. Chin, E. Tsang, H. Mediwake, W. Khair, J.L. Biccler, G. Hapgood, P. Mollee, Z. Nizich, D. Joske, D. Radeski, G. Cull, D. Villa, T.C. El-Galaly, C. Y. Cheah (2018). Frequency of bowel perforation and impact of bowel rest in aggressive non-Hodgkin lymphoma with gastrointestinal involvement. British Journal of Haematology (Early View).

4. C. H. Nørgaard, N. B. Søgaard, J. L. Biccler, L. Pilgaard, M. H. Eskesen, T. H. Kjartansdottir, M. Bøgsted, T. C. El-Galaly (2018). Limited value of routine follow-up visits in chronic lymphocytic leukemia managed initially by watch and wait: A North Denmark population-based study. PLOS ONE (Accepted).

Abstract

The prognosis of patients with haematological malignancies in terms of survival probability and relapse risk is of interest to clinical practitioners, patients, health economists, and many more. By combining the extensive register data available in the Nordic countries with novel statistical and machine learning methods, new insights into the disease course of haematological malignancies are obtained.

Three of the papers that make up this PhD dissertation concern the use of statistical methods to predict the survival probability at different time points. Currently available prognostic indices are often based on statistically inefficient practices and can be outperformed by simple models. In Paper I, the extent of this inefficiency in the case of diffuse large B-cell lymphoma (DLBCL) is investigated. In particular, a simple new model which outperforms the standard prognostic scores is introduced and described. In Paper III, this simple model of the survival of DLBCL patients is improved upon by using a machine learning method. In Paper IV, the performance of established prognostic scores for eight common haematological malignancies is investigated. None of the established risk scores perform well and all are outperformed by a simple model which relies only on information regarding age and functional status.

For patients surviving treatment and achieving remission, information regarding the risk of relapse and death given that they survived the initial period is of significant interest. In addition to being useful for patient counselling, such knowledge facilitates the construction of rational follow-up programs. Describing such dynamic measures of survival was the main objective of Paper II and Paper V. In Paper II, the survival of acute promyelocytic leukaemia (APL) is compared to that of a similar healthy population. It is shown that the survival prospects of APL patients surviving the critical three month period post-diagnosis are nearly indistinguishable from those of a healthy population. In Paper V, it is shown that also the survival of young Hodgkin lymphoma patients in remission for over two years becomes similar to that of a similar background population. Furthermore, also the relapse risk becomes negligible and the use of follow-up strategies focusing on relapses past this two-year landmark is of limited value.

Often register data contains some errors. Potential causes of these errors include disease misclassification, measurement errors, data entry errors, etc. These outliers often have a large influence on standard statistical techniques and the resulting estimates are often of limited use when outliers are present.

The field of robust statistics aims to develop statistical methods on which outliers have only a limited influence. In paper VI a robust estimation procedure for survival models is proposed.

Resumé

Inden for hæmatologiske sygdomme er prognose et vigtigt begreb for både klinikere, patienter, pårørende, sundhedsøkonomer og andre faggrupper, og vurderes ofte ud fra døds- og recidivrisikoen. Ved at kombinere de omfattende sundhedsregistre med nye statistiske og machine learning metoder kan ny viden om hæmatologiske sygdomsforløb opnås.

Denne ph.d. afhandling indeholder seks videnskabelige artikler, hvoraf tre omhandler brugen af statistiske metoder til mere præcist at prædiktere overlevelsessandsynligheden for hæmatologiske patienter til forskellige tidspunkter. Eksisterende prognostiske modeller inden for hæmatologien er ofte baseret på suboptimale statistiske fremgangsmåder, og disse kan derfor nemt forbedres ved brugen af andre simple modeller. I artikel I undersøges hvordan små ændringer i håndteringen af kliniske variable kan føre til en langt mere præcis forudsigelse af overlevelsen for patienter med diffust storcellet B-cellet lymfom. I artikel III forbedres denne simple prognostiske model ved hjælp af machine learning teknikker. I artikel IV undersøges præcisionen af prognostiske modeller, som bliver brugt indenfor otte almindelige lymfomtyper. Denne artikel illustrerer disse modellers manglende præcision og viser at disse overgås af en prognostisk model, der baseres alene på patienternes alder og performance status målt ved diagnose.

For de hæmatologiske patienter som overlever behandlingen og opnår remission, er viden om risikoen for recidiv og død særlig værdifuld. Da denne viden er vigtig i forbindelse med patientrådgivning, som efter behandlingen gives regelmæssigt, er disse risici også interessante ved bestemte tidspunkter i opfølgningsforløbet. Denne information kan også bruges til at etablere mere rationelle opfølgningsprogrammer. Formålet med artikel II og V var netop at beskrive disse risici dynamisk. I Artikel II sammenlignes mortaliteten blandt akut promyelocytisk leukæmi (APL) patienter med mortaliteten i den danske baggrundsbefolkning. APL patienter, som overlever de første tre kritiske måneder opnår samme overlevelse som en baggrundbefolkningen.
I artikel V dokumenteres, at unge patienter med Hodgkin lymfom kan forvente en overlevelse, der modsvarer en rask baggrundsbefolkning, hvis sygdommen har været i remission i mere end to år. Desuden er risikoen for recidiv meget lille for denne patientgruppe, og derfor vil værdien af opfølgningsstrategier, som fokuserer på detektion af relaps efter denne toårs milepæl være begrænset.

Ofte indeholder registerdata fejl. Potentielle fejlkilder omfatter blandt andet misklassifikation af sygdomme, målefejl og dataindtastningsfejl. Disse

outliers har ofte stor indflydelse på forskningsresultater, når konventionelle statistiske metoder tages i brug, hvilket medfører en begrænset brugbarhed af resultaterne. Robuste statistiske værktøjer sigter mod at generere resultater, hvor outliers har en begrænset indflydelse. I Artikel VI præsenteres en robust estimationsprocedure for overlevelsesmodeller.

Contents

Part A

Background

Background

The kappa of this thesis is intended to introduce some background information regarding the data sources and techniques used in the papers. Since papers I up to V include an extensive overview of the clinical and epidemiological considerations, the focus of a large part of this kappa will be on the statistical background.

In Section 1, a short introduction to the haematological malignancies that were studied and the data-sources underlying the studies is provided. In each study in this dissertation the outcome was right-censored, an overview of the survival analysis methodology used to take the censoring into account is presented in Section 2. Papers I, III, and IV concern the use of prognostic models, a description of what prognostic models are, how to measure their performance, and how to combine different prognostic models is given in Section 3. A drawback of standard statistical measures of survival, e.g. the hazard function, is that they are not readily understood by patients or medical practitioners. In papers II and V alternative survival measures such as the standardized mortality ratio and the loss of life expectancy are used to make the results more easily interpretable. These measures are introduced in Section 4. Furthermore, Papers II and V also describe the evolution of these measures conditional on patients achieving longer periods of remission. The statistical methods used to do so are described in Section 5. Finally, paper VI considers estimation procedures for survival models in the presence of outliers, a short overview of the estimation problems occurring when outliers are present is given in Section 6.

1 Haematological malignancies

1.1 Introduction

Haematological malignancies are cancers that originate in the blood, bone marrow, or lymph system. Depending on the type of the affected cells, these malignancies are subdivided into three subgroups: leukaemias (blast cell cancers), lymphomas (lymphocyte cancers), and myelomas (plasma cell cancers). Within these three subcategories, the prognosis, treatment options, etc. vary largely and these overarching categories have been further divided into subtypes as e.g. exemplified in the WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues.[1]

The treatment of haematological malignancies plays an important role in the

history of cancer treatments. In fact, one of the first successful chemotherapy treatments was a component derived from the warfare chemical mustard gas which has been used for the treatment of a variety of haematological malignancies.[2] Currently, treatment of haematologic malignancies ranges from wait and watch strategies to therapies consisting of one or more of chemotherapeutic drugs, immunotherapy, radiotherapy, transplantation, etc.

1.2 Overview of selected haematological malignancies

Diffuse large B-cell lymphoma

Diffuse large B-cell lymphoma (DLBCL) is the most common type of lymphoma and accounts for $\approx$ 30% of all lymphomas.[3] Patients diagnosed with DLBCL have a $\approx$ 75% chance of surviving five years post-diagnosis.[4] Subclassifications of DLBCL are often based on genetic information or on the dissease location and large differences in treatment response and survival exist.[5] One of the most often used classification systems subdivides DLBCL according to its cell of origin, more particularly DLBCL cases can be classified into germinal centre B-like (GCB) DLBCL and activated B-cell (ABC) DLBCL.[6] While the GCB/ABC classification is recomended in the current WHO classification [1]. Recently, there has also been an increased interest in the treatment and epidemiology of specific gene alterations, e.g. double-hit lymphomas.[7] However, currently there is no compelling data suggesting that the presence of these alterations or the GCB/ABC classification have a clinical implication with respect to the treatment of DLBLC. The location of the involved sites, on the other hand, can have a large influence on the treatment options and survival. For example, standard therapies are often unable to reach a sufficiently high penetration rate of the blood-brain barrier and the treatment of DLBCL presenting in the central nervous system is more challenging and associated with worse outcomes than systemic DLBCL.[8]

The majority of systemic DLBLC patients receive curative intent treatment. The current standard of care consists of rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) or variations thereof.[5]

Altough there has been a surge in the interst of risk scores based on molecular data, simple clinical risk scores still form the basis of DLBCL risk stratification used in daily clinical practice. The international prognostic index (IPI), which was developed in 1993, is the most commonly used prognostic index for DLBCL.[9] The risk factors that make up the IPI are: age at diagnosis $>$ 60 years; Ann Arob stage III or IV; elevated lactate dehydrogenase; Eastern Cooperative Oncology Group performance status 2, 3, or 4; and the presence of $>$ 1 extranodal site.[9] The IPI score is then defined as the number of

present risk factors. Given the IPI score, patients are divided into four risk groups: low risk (0 or 1 risk factor); low-intermediate risk (2 risk factors); high-intermediate risk (3 risk factors); and high risk(4 or 5 risk factors).[9] Recent studies have attempted to improve upon the IPI and examples of this are the R-IPI and NCCN-IPI indices.[10, 11] Common among almost all proposed prognostic indices is that they rely heavily on dichotomized risk factors and result in the classification of cases into few, usually ≤ 4, risk categories. These practices have been commonly critizized in the statistical literature, yet remain wide-spread.[12] Papers I, III, and IV document the information loss due to the dichotomization of the risk scores for DLBCL and improve upon the currently available prognostic indices by avoiding this practice.[13–15]

Hodgkin lymphoma

In contrast to most cancer types, Hodgkin lymphoma (HL) has a bimodal incidence pattern and patients diagnosed before turning 40 years old constitute a significant portion of the HL cases. The majority of the patients receiving current standard of care treatment will be considered cured and $> 90\%$ survives at least five years post-diagnosis.[16, 17] Given these high cure and survival rates, rational survivorship care is clinically important, in particular at a time with limited ressources in available in the public health care setting.[18] Hodgkin lymphoma is usually staged according to the Ann Arbor staging system which ranges from stage I (lymphoma isolated to one location) to stage IV (disseminated involvement of one or more organ outside the lymph system).[19] Based on the Ann Arbor staging system, HL is often subdivided into limited stage disease (Ann Arbor stage I or II without B-symptoms) and advanced stage disease (Ann Arbor stage II with B-symptoms or Ann Arbor stage III/IV). Although the Ann Arbor stage is still widely used, more novel staging measures such as those based on PET/CT imaging are gaining traction.[20]

The mustargen, vincristine, prednisone and procarbazine (MOPP) chemotherapy regimen was the first treatment that succesfully cured a large proportion of advanced stage HL patients.[21] The main drawback of the MOPP regimen is its high rate of secondary malignancies, which did not became clear until more than a decade after its introduction.[22, 23] The successor of MOPP was a combination of doxorubicin, bleomycin, vinblastine, and dacarbazine (ABVD) which is highly efficacious and less toxic, and ABVD remains part of the standard of care.[24] A more recently introduced chemotherapy regimen that is used in clinical practice is the bleomycin, etoposide, doxorubicin, cyclophosphamide, vincristine, procarbazine, and prednisolone (BEACOPP) combination which has led to even higher cure rates among advanced-stage HL patients. The effect of BEACOPP on overall survival, however, remains

controversial.[25] In addition to chemotherapy, HL patients often receive radiotherapy treatment and in order to reduce toxicities, the dosage and extent of the fields used in these treatments have been substantially reduced over the last decades.[26]

Acute promyelocytic leukaemia

Acute promyelocytic leukaemia (APL) is a subtype of acute myeloid leukaemia (AML) characterized by the translocation of the retinoic acid receptor (RARA) gene.[27] The accumulation of leukemic promyelocytes causes the clinical hallmark of APL, acute and life-threatening coagulopathy. The treatment of APL is based on the ability of all-trans retinoic acid (ATRA) to force the differentiation of the leukemic promyelocytes and APL remains one of the few cancers for which molecularly targeted therapy has been successfully introduced.[28] The survival prospects of APL patients who receive an ATRA based treatment are excelent compared to other types of AML, especially if the patient survives the critical coagulopathy phase.[29, 30] Notwithstanding the good long-term survival of APL patients treated with ATRA, the early death rate remains high, likely due to a delayed administration of ATRA.[29, 30] The standard of care for APL consists of an induction treatment of ATRA combined with chemotherapy and/or arsenic trioxide.[27] After the induction treatment, most patients receive a consolidation treatment consisting of a similar regimen and with recent regimens five-year relapse risks $< 15\%$ have been reported in clinical trials.[27, 31]

1.3 Data for epidemiological studies of haematological malignancies

Two major sources of data on the survival and clinicopathologic characteristics of cancers patients are clinical trials and clinical registers.

Clinical trials are usually conducted to test the efficacy of new treatments. In order to rigorously define the study population and treatment, the enrolment is often restricted to patients who meet strict eligibility criteria and the treatments are administered in highly controlled settings.[32] For some clinical trials, patients have to undergo extensive examinations before being enrolled which can result in the exclusion of high-risk patients with acute treatment needs.[32, 33] Furthermore, due to budget and time constraints often the follow-up period and number of enrolled patients is limited which hampers the detection of late events and late toxicities.[34] Data from clinical trials tends to be meticulously gathered and is generally of high quality.

Registers often cover large unselected populations and tend to be constructed and maintained with a certain purpose in mind, e.g. quality control. Therefore, the most prominent advantages of clinical register data are that the data usually represents day to day clinical practice and describes large unselected populations. An important drawback is that the estimation of causal effects is hampered due to potential confounding. Furthermore, partly because these databases are often not constructed to accomodate all needs of scientific research, the data completeness and detail level tend to be lower than that of clinical trials. In Denmark, population-wide registers have been established for a large number of diseases. Two large clinical registers which were used in this thesis are the nationwide Danish National Lymphoma Register (LYFO) and the Danish Acute Leukemia Register.[35, 36] The coverage of these registers exceeds 95% and, hence, essentially all Danish lymphoma and leukaemia patients are included in these registers.[35, 36]

In summary, data from clinical trials and clinical registers each have their distinct advantages and disadvantages. In practice, the use of both leads to robust and generalizable results that form the foundation of clinical practice.

2 Analysis of survival data

2.1 Introduction and notation

Patient data is often collected at diagnosis and when events have occurred, such as deaths or relapses, event times are added to the database. At the time of the statistical analysis a number of patients included in the databases might not have experienced the event and the only information regarding their event date is that it has not happened by a certain date. In statistical lingo this is called censoring. The most commonly occurring type of censoring, right-censoring, occurs when for some patients only a lower bound of the event time is observed.

The analysis of censored and truncated data is the topic of survival analysis, a branch of statistics. In the papers presented in this thesis neither truncation nor left-censoring occurred and this section will focus on the concepts of survival analysis relevant for right-censored data. To ease the exposition we will, without loss of generality, assume that all time variables are reported in years.

Throughout this thesis the potentially unobserved event-time and censoring time of subject i will be denoted by T_i and C_i, respectively. In register studies, T_i often corresponds to the time between diagnosis and death or progression, and C_i corresponds to the time between diagnosis and the date at which the

event status was last updated. The information which is usually available in studies of right-censored data is the minimum of the event and censoring times, $U_i = \min(T_i, C_i)$, and the status variable, $\Delta_i = \mathbb{1}(T_i < C_i)$, which indicates whether U_i corresponds to an event or a censoring.

In time-to-event studies, the inference targets tend to be either the hazard function or the survival function. The hazard function is the event rate at time t given survival up to time t or formally

$$\alpha(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h | T \geq t)}{h}.$$

The survival function describes the probability of surviving up to a time t, i.e. the survival function is

$$S(t) = P(T > t).$$

The majority of the statistical techniques for survival analyses tend to model the hazard function, however, the survival function can easily be recovered from a known hazard function by noting that

$$S(t) = \exp\left(- \int_0^t \alpha(u) du \right).$$

A large proportion of epidemiological studies attempt to describe the survival of patients with a specific disease. In such analyses, there is often an interest in describing and comparing the survival prospects of different subgroups. Given that the necessary data on survival status and covariates, $\boldsymbol{X}$, is available, these studies can be performed by modelling the hazard, $\alpha(t|\boldsymbol{X}, \boldsymbol{\theta})$, or survival function, $S(t|\boldsymbol{X}, \boldsymbol{\theta})$, in function of some parameters $\boldsymbol{\theta}$. Depending on the dimension of the parameter space these models can be categorized as non-parametric, semi-parametric, and parametric models. Although a rigorous description of each type is beyond the scope of this thesis, for each approach one statistical model is introduced and its advantages and disadvantages are concisely described.

2.2 Survival models

Non-parametric models

The goal of most non-parametric estimators is to estimate a certain function without imposing any strong constraints on its shape. The main advantage of this is that only minimal assumptions are needed to obtain consistent estimators. Perhaps the largest drawback is that the estimation procedures can

be fairly inefficient when compared to parametric alternatives.

The most often used non-parametric estimator in survival analysis is the Kaplan-Meier (KM) estimator.[37] If we denote whether a patient is at risk of experiencing the event at time t as $Y_i(t) = \mathbb{1}(U_i \geq t)$, the number of patients at risk of experiencing the event at time t is $Y(t) = \sum_{i=1}^{n} Y_i(t)$. Defining the number of events happening at time t as $N(t) = \sum_i^n \mathbb{1}(U_i = t)\Delta_i$, the KM estimator can be written as

$$\hat{S}(t) = \prod_{s \in \{U_i \leq t | \Delta_i = 1\}} \left(1 - \frac{N(s)}{Y(s)}\right).$$

The most important assumption underlying the theoretical results for the KM estimator is the assumption of independent (or non-informative) censoring. By reporting KM estimates stratified according to some covariate the survival functions of different subgroups are often compared. Depending on the sample size within each strata this can lead to unstable estimates.

Semi-parametric models

Semi-parametric models tend to impose constraints on some components of the model, i.e. the parametric part of the model, while keeping the rest of the model very flexible, i.e. the non-parametric part of the model. The rationale behind semi-parametric models is that if the model parameters of interest are the parametrized ones they will be estimated relatively efficiently while the nuisance parameters are left unspecified to avoid model misspecification.

The Cox proportional hazard (CPH) model is without doubt the most popular semi-parametric survival model.[38] In the CPH model it is assumed that given a baseline hazard, $\alpha_0(t)$, the covariates have a proportional effect on the hazard function

$$\alpha(t|\boldsymbol{X}_i) = \alpha_0(t)\exp\left(\boldsymbol{X}^t\boldsymbol{\theta}\right).$$

The coefficients of the covariates can be estimated without specifying the baseline hazard by use of the partial likelihood estimator

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \left[\frac{\exp\left(\boldsymbol{X}_i^t\boldsymbol{\theta}\right)}{\sum_{j|Y_j(U_i)=1}\exp\left(\boldsymbol{X}_j^t\boldsymbol{\theta}\right)}\right]^{\Delta_i}.$$

This model is especially useful when only the covariate effects are of interest. The baseline hazard can also be estimated, e.g. by using the Breslow-estimator, and hence estimates of the hazard and survival functions can be obtained.[39]

Parametric models

Parametric survival models explicitly model the whole survival curve or hazard function using a finite number of parameters. The main advantage is that the survival process is fully parametrized and, if the model is correct, the usual estimators are often more efficient than their non- or semi-parametric counterparts. Their main disadvantage is that when the survival process does not fit the parametric model, the interpretation and applicability of the resulting model is limited. This disadvantage can be partly circumvented by use of models based on flexible parametric functions, e.g. splines.[40]

An example of a parametric model used in survival analysis is obtained by parametrizing the baseline hazard used in the CPH model. For example, setting $\alpha_0(t) = \lambda^p t^{p-1}$, in which λ and p are parameters, leads to a Weibull model with proportional covariate effects. The parameters in parametric survival models are usually obtained by use of a maximum likelihood estimator (MLE) [41]

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \prod_{i=1}^{n} S(U_i|\boldsymbol{X}_i, \boldsymbol{\theta})\alpha(U_i|\boldsymbol{X}_i, \boldsymbol{\theta})^{\Delta_i}.$$

Altough alternative estimation procedures exists, the MLE methodology is the most commonly used as it generally leads to estimators which are asymptotically unbiased and achieve a high efficiency.

2.3 Time-varying coefficients

Covariates available in clinical registers are often only measured at the diagnosis date. A limitation of covariates measured at diagnosis is that the magnitude of the covariate effects on the hazard function often changes over time.[42] For example, in Figure 1 the hazard ratio of stage IV versus stage I disease in DLBCL is shown. Around the time of diagnosis stage IV patients have a largely increased hazard for dying as compared to stage I patients. However, for patients surviving three or more years the stage measured at diagnosis is not as influential and its effect becomes negligible around five years post diagnosis.

One way of explicitly including this knowledge is by allowing the coefficients to change in function of the time. E.g. replacing the constant coefficients in the standard CPH model by spline effects in function of the time, $\boldsymbol{\Theta}(t, \boldsymbol{\theta})$, leads to

$$\alpha(t|\boldsymbol{X}) = \alpha_0(t)\exp\left(\boldsymbol{X}^t\boldsymbol{\Theta}(t, \boldsymbol{\theta})\right).$$

Other examples of models with time-varying effects include the parametric Royston-Parmar model and the Cox model with non-parametric time-varying

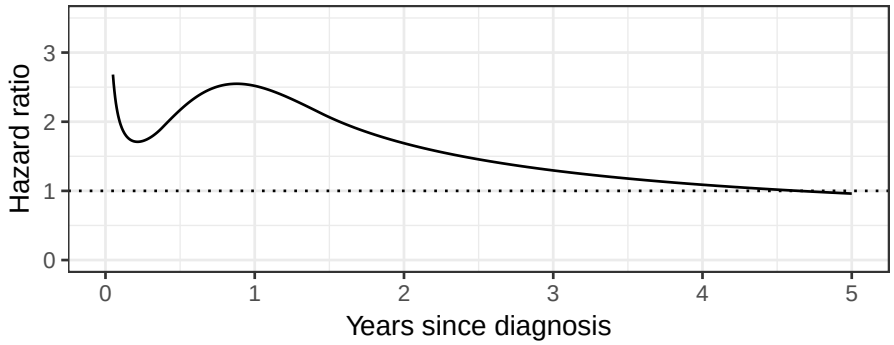

Figure 1: Hazard ratio of stage IV versus stage I disease in HL as estimated by a flexible parametric proportional hazard model.

effects.[40, 43]

2.4 Competing risk analyses

In cancer research, there are often multiple terminal endpoints, e.g. death due to cancer and death due to causes unrelated to cancer. One common approach for handling multiple endpoints is the use of a composite endpoint, e.g. death from any cause. A drawback of this approach is that the individual endpoints are not explicitly described. Competing risk analysis is a large sub-field of survival analysis which focuses on modelling multiple terminal endpoints, a visual example of the set-up of interest is given in Figure 2. If we denote the set of potential endpoints by $\mathcal{K}$, the information that is observed in competing risk studies is the time of an event or censoring, U_i, whether the observed time was a censoring or event time, Δ_i, and the type of the event, $K \in \mathcal{K}$. The two quantities that are the main focus of competing risk analyses are the cause-specific hazard and the cumulative incidence functions.

The cause-specific hazard function describes the event rate of a specific event-type at time t given survival up to time t and is defined as

$$\alpha_k(t) = \lim_{h \to 0} \frac{P(t \leq T \leq t+h, K = k | T \geq t)}{h}, \; k \in \mathcal{K}.$$

Given that the events always correspond to one death cause, it is easy to see that the hazard function of any event occurring is

$$\alpha(t) = \sum_{k \in \mathcal{K}} \alpha_k(t)$$

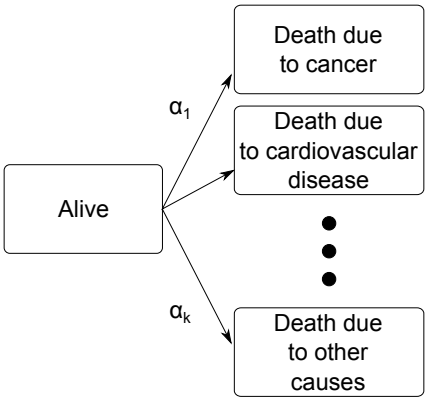

Figure 2: A diagram visualizing the usual set-up in competing-risk studies.

and the probability of survival up to time t remains $S(t) = \exp\left(-\int_0^t \alpha(u)du\right)$. The probability of event k happening by time t is known as the cumulative incidence and can be derived from the cause-specific hazard functions

$$P_k(t) = P(T \leq t, K = k) = \int_0^t S(u)\alpha_k(u)du. \tag{1}$$

The most often used non-parametric estimator of the cumulative incidence is the Aalen-Johansen estimator, which essentially substitutes the cause-specific hazard functions in Equation 1 with the increments in the cause-specific Nelson-Aalen estimators.[44]

The incorporation of covariate effects into the cause-specific hazard function can be done as in the usual survival analysis setting. Fitting cause-specific hazard models is done by treating the competing risk events times as censoring times. The modelling of the cause-specific hazard is generally the most appropriate in aetiological studies.

When the objective of a study is prognostication, the quantity of interest is usually the cumulative incidence, i.e. the probability that a specific event occurs. In this scenario, an important disadvantage of modelling the cause-specific hazard is that a model for $S(t)$ or for each cause-specific hazard function has to be developed. Additionally, due to the involvement of $S(t)$ in Equation 1, interpreting the effect of a covariate on the cause specific hazard in terms of the cumulative incidence becomes problematic. It could even be that a certain covariate is associated with increased cause-specific hazard rates but leads to a decrease in the cumulative incidence. Models that di-

rectly model the cumulative incidence function have been developed. The most popular approach is the Fine-Gray model which models the cumulative incidence as

$$P_k(t|\boldsymbol{X}, \boldsymbol{\theta}) = 1 - (1 - P_0(t))^{\exp(\boldsymbol{X}^t\boldsymbol{\theta})},$$

in which $P_0(t)$ is a baseline cumulative incidence.[45]

2.5 Pseudo-values

The objective of a survival analysis is often to determine the effect of covariates on some function of the event times, $f(T_i)$, for example the five-year survival, $f(T_i) = \mathbb{1}(T_i \geq 5)$. If $f(T_i)$ could be observed for each observation then given a link function, $g(\cdot)$, regression models of the form $E(f(T_i)|\boldsymbol{X}_i) = g^{-1}(\boldsymbol{X}_i^t\boldsymbol{\theta})$ could be estimated using the theory surrounding generalized linear models. However, due to the censoring, $f(T_i)$ is not necessarily observed. To circumvent this issue pseudo-values can be used as replacements of the potentially unobserved $f(T_i)$ values.[46]

Given that we have an estimator $\hat{\rho}$ for $\rho = E(f(T_i))$ the pseudo-value of observation i is defined as

$$\hat{\rho}_i = n\hat{\rho} - (n-1)\hat{\rho}^{-i},$$

in which $\hat{\rho}^{-i}$ is the estimate obtained using all observations except for observation i.[47] In an analysis based on uncensored data the natural $\hat{\rho}$ estimator is $\frac{1}{n}\sum_{i=1}^{n} f(T_i)$ and the pseudo-values are then $\hat{\rho}_i = f(T_i)$. When censoring is present which estimator is used depends on the $f(\cdot)$ function. E.g. when $f(T_i) = \mathbb{1}(T_i \geq 5)$ one could obtain an estimate of the five-year survival using the the Kaplan-Meier estimator. When the Kaplan-Meier estimator is used in this way, Graw et al have shown that $E(\hat{\rho}_i) \to E(f(T_i))$ and $E(\hat{\rho}_i|\boldsymbol{X}_i) \to E(f(T_i)|\boldsymbol{X}_i)$ as $n \to \infty$.[48] By relying on results from functional analysis, Overgaard et al have recently extended these results beyond the Kaplan-Meier estimator and x-year survival psuedo-values.[49]

Once the pseudo-values are calculated they can be used as outcome variables in generalized linear models of the form

$$E(f(T_i)|\boldsymbol{X}_i) = g^{-1}(\boldsymbol{X}_i^t\boldsymbol{\theta}).$$

The estimates are then obtained as the solutions to the following generalized estimating equations (GEE)[50]

$$U(\boldsymbol{\theta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} g^{-1}(\boldsymbol{X}_i^t\boldsymbol{\theta}) V_i^{-1}(\hat{\rho}_i - g^{-1}(\boldsymbol{X}_i^t\boldsymbol{\theta})) = \mathbf{0},$$

in which V_i is a working variance. To estimate the variance a sandwich estimator can be used[47]

$$\hat{\Sigma} = I^{-1}(\hat{\boldsymbol{\theta}})\hat{\text{var}}(U(\hat{\boldsymbol{\theta}}))I^{-1}(\hat{\boldsymbol{\theta}}),$$

in which

$$I(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} g^{-1}(\boldsymbol{X}_i^t\boldsymbol{\theta})^t V_i^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} g^{-1}(\boldsymbol{X}_i^t\boldsymbol{\theta})$$

and

$$\hat{\text{var}}(U_i(\boldsymbol{\theta})) = \frac{1}{n}\sum_{i=1}^{n} U_i(\theta)U_i(\theta)^t.$$

It is known that this variance estimator tends to be biased, more particularly, it tends to return conservative estimates and a number of alternative unbiased variance estimators have been proposed.[49, 51, 52] In practice, however, the sandwich estimator remains the most convenient.

3 Prognostic models

The research field of prognostic modelling is concerned with the probability that certain events happen in the future. Information on the prognosis is valueable in a large number of scenarios. First of all, the prognosis of a certain disease is obviously of interest to patients. Additionally, treatment plan decisions made by physicians are in part based on their knowledge about the prognosis. Prognostic information can also guide health-economical decision making, e.g. follow-up programs for relapse detection might be viewed as too costly if the relapse risk is limited. Other applications include the selection of patients for experimental treatment, comparing the performance in treating a disease across hospitals, etc.[53] In practice, prognostic models which include risk factors are often preferred over describing the prognosis for the whole population with the disease. In the statistical literature prognostic modelling is usually referred to as predictive modelling and a vast literature exists.[54] In the personalized medicine litterature the words prognostic and predictive have been differentiated and in that field effect modifiers are often called predictive variables.[55] The words prognostic and predictive will be used interchangeably in this dissertation.

3.1 Prediction versus explanation

One of the essential differences between explanatory and predictive analyses is which models are considered to be useful.

In explanatory analyses, the goal is to obtain a model which allows the researcher to reach a causal conclusion regarding the effects of interest, e.g. which clinical factors have a causal effect on the survival of DLBLC patients. Models which are known to incorrectly model the underlying system are usually, and with reason, not considered. An example is the exclusion of well-known confounders which would lead to biased estimates of the causal effect of interest. In conclusion, a major goal of explanatory modelling is to obtain a model in which the bias of the effect estimates is minimized.

In predictive analyses, the goal is to predict the outcome of new observations, e.g. the five-year survival probablity of a newly diagnosed DLBCL patient. A model is therefore considered to be useful when its predictions are accurate. How the model accomplishes this and whether or not the model actually resembles the underlying causal process is not necessarily of interest. One could think that using a model that perfectly models the causal structure is the way forward, however, quite often this approach will be suboptimal when considering a measure of the prediction error as optimality criterion.[56] This scenario is visualized in Figure 3 in which a biased estimator would tend to lead to estimates closer to the true value than an unbiased estimator. In fact, estimators for which it is known that they will be biased but have a lower variance than the usual unbiased estimators are commonly used in prognostic studies. Perhaps the most well-known example of this is the class of penalized estimators. For example, ridge regression techniques tend to bias the estimators towards zero in order to obtain a reduction in the variance of the estimators.

3.2 Measuring predictive error

Measures of predictive performance can be roughly categorized into two categories, calibration and discrimination measures.[57] Discrimination measures quantify the ability of a predictive model to correctly rank patients according to who is at higher risk to experience the event earlier. Calibration measures are concerned with measuring whether the modelled probability of an event is in accordance with the probability of that event being observed in the data. In this section, we introduce one measure of calibration, the Brier score, and two measures of discrimination, the C-index and time-varying area under the receiver operating curve (AUC).

Brier score

To measure the calibration of survival models at time t, the squared error between the survival status at time t, $\mathbb{1}(T \geq t)$, and the modelled probability

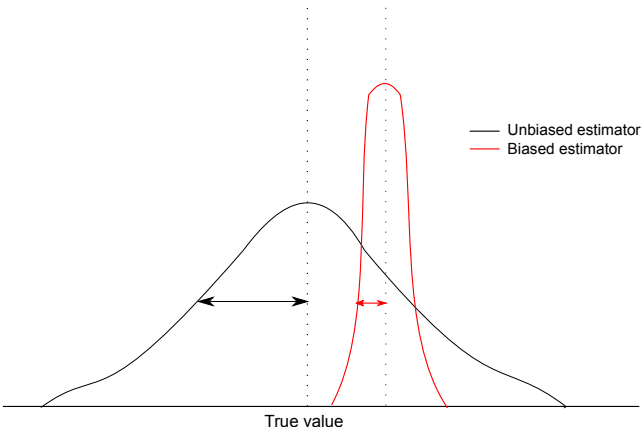

Figure 3: Example of the bias-variance tradeoff. A biased model might on average lead to more precise estimates than an unbiased model with a larger variance. The curves represent the distribution of two estimators from which the predictions are obtained.

of being alive, $S(t|\boldsymbol{X}, \boldsymbol{\theta})$, is often used.[58, 59] This squared error is called the Brier score and is formally defined as

$$B(t) = E\left((\mathbb{1}(T \geq t) - S(t|\boldsymbol{X}, \boldsymbol{\theta}))^2\right).$$

If we denote the survival function of the censoring mechanism by $G(t) = P(C \leq t)$ and an estimator thereof by $\hat{G}(t)$ the following inverse probability of censoring weighted (IPCW) estimator

$$\hat{B}(t) = \sum_{i=1}^{n} \frac{\mathbb{1}(\min(T_i, t) \leq C)}{\hat{G}(\min(T_i, t)|\boldsymbol{X}_i)} (\mathbb{1}(T_i \geq t) - S(t|\boldsymbol{X}_i, \boldsymbol{\theta}))^2,$$

can be shown to be a consistent estimator of the Brier score at time t.[59] From now on we denote the true underlying survival function, which is not necessarily of the form $S(t|\boldsymbol{X}, \boldsymbol{\theta})$, by $S_0(t|\boldsymbol{X})$. The following result is easily derived

$$\begin{aligned} &= E\left((\mathbb{1}(T \geq t) - S(t|\boldsymbol{X}, \boldsymbol{\theta}))^2\right) \\ &= E\left((\mathbb{1}(T \geq t) - S_0(t|\boldsymbol{X}))^2\right) + E\left((S_0(t|\boldsymbol{X}) - S(t|\boldsymbol{X}, \boldsymbol{\theta}))^2\right). \end{aligned} \tag{2}$$

Hence, the Brier score consists of an irreducible error term and a mean squared error arising from using an incorrect model. In practice, usually a plot of the Brier score over time is reported together with a summary measure over the study period, $[0, \tau]$, one such a measure is the integrated Brier

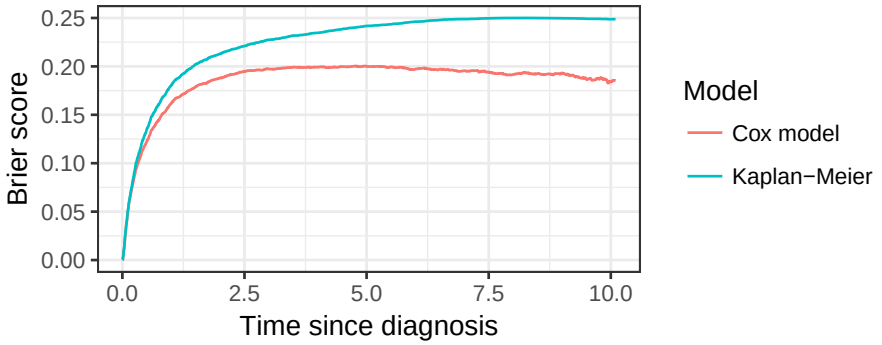

Figure 4: The Brier score of the survival of DLBCL patients as predicted by a Cox proportional hazard model which included age as covariate and the Kaplan-Meier estimator.

score[60]

$$\text{IBS} = \int_0^\tau B(u)du.$$

The Brier score curves for two models of the survival of DLBCL patients are shown in Figure 4. More particularly, the models are a CPH model with age as a covariate and a Kaplan-Meier estimator of the population survival. Since the Kaplan-Meier estimate is independent of any modelling assumptions and includes no covariate information it can be interpreted as a reference model which is hopefully be outperformed by more advanced models, as is the case in this example. The term $E\left((\mathbb{1}(T \geq t) - S_0(t|\boldsymbol{X}))^2\right)$ which shows up in Equation 2 gives some insight into the shape of the Brier score. Even if the true model was known, the Brier score would be expected to differ from zero and initially show an evolution closely related to $S_0(t|\boldsymbol{X})$.

C-index

The C-index measures the concordance between the ranking of failure times and the ranking of the risk scores. Denoting the risk score given covariates by $\mathcal{R}(\boldsymbol{X})$, the C-index of a randomly sampled pair of observations $((\mathcal{R}(\boldsymbol{X}_1), T_1), (\mathcal{R}(\boldsymbol{X}_2), T_2))$ is defined as

$$C = P(\mathcal{R}(\boldsymbol{X}_1) > \mathcal{R}(\boldsymbol{X}_2)|T_2 > T_1),$$

which corresponds to the probability that the risk score of patient one is larger than the risk score of patient two given that patient one dies before patient two. In practice this quantity is often estimated by the following

IPCW estimator[61]

$$\hat{C} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \hat{G}(U_i)^{-2}\mathbb{1}(U_i < U_j)\mathbb{1}(\mathcal{R}(\boldsymbol{X}_i) > \mathcal{R}(\boldsymbol{X}_j))}{\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \hat{G}(U_i)^{-2}\mathbb{1}(U_i < U_j)}.$$

Depending on the estimator of $G(\cdot)$, $\hat{G}(\cdot)$ can be quite unstable for large values which in turn can lead to unstable estimates of the C-index. Therefore it has been suggested to use a truncated C-index instead[61]

$$C_\tau = P(\mathcal{R}(\boldsymbol{X}_1) > \mathcal{R}(\boldsymbol{X}_2)|T_2 > T_1, T_1 < \tau),$$

which can be estimated as follows

$$\hat{C}_\tau = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \hat{G}(U_i)^{-2}\mathbb{1}(U_i < U_j, U_i < \tau)\mathbb{1}(\mathcal{R}(\boldsymbol{X}_i) > \mathcal{R}(\boldsymbol{X}_j))}{\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \hat{G}(U_i)^{-2}\mathbb{1}(U_i < U_j, U_i < \tau)}.$$

Given that the C-index is defined as a probability its value is part of the $[0, 1]$ interval. If we randomly generated risk scores from a continuous distribution independent of T_i, the C-index would be 0.5. Hence, a C-index of 0.5 corresponds with a model which has no prognostic information at all and models that are of use in practice should have a C-index > 0.5.

Time-varying AUC

A drawback of the C-index is that it does not describe the performance of the risk score at different time points of interest but instead focusses on the complete follow-up period. It is not unthinkable that some risk scores could have a good performance in the period directly following diagnosis but lose performance over time. Such evolutions of the discriminative performance over time would not be described by the C-index. A measure closely related to the C-index is the time-varying AUC. Different definitions have been proposed in the literature.[62] The most intuitive definition of the time-varying AUC at time t for a pair of observations $((\mathcal{R}(\boldsymbol{X}_1), T_1), (\mathcal{R}(\boldsymbol{X}_2), T_2))$ is[62]

$$AUC(t) = P(\mathcal{R}(\boldsymbol{X}_1) > \mathcal{R}(\boldsymbol{X}_2)|\mathbb{1}(T_1 \leq t), \mathbb{1}(T_2 > t)).$$

Hence, the time-varying AUC measures the probability that patients who have died by time t will receive a larger risk score than patients who are still alive. The IPCW estimator of this quantity is[63]

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \hat{G}(U_i)^{-1}\hat{G}(t)^{-1}\mathbb{1}(U_i \leq t, U_j > t)\mathbb{1}(\mathcal{R}(\boldsymbol{X}_i) > \mathcal{R}(\boldsymbol{X}_j))}{\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \hat{G}(U_i)^{-1}\hat{G}(t)^{-1}\mathbb{1}(U_i \leq t, U_j > t)}$$

As for the C-index, a time-varying AUC of 0.5 can be obtained at any time t by randomly assigning a continuous risk score to each observation.

3.3 Ensemble models

A large number of different survival models are available and this number keeps on growing. In general, the objective of a prognostic study is to have one prognostic model and model selection is an important issue. Ensemble modelling is a general approach which avoids the selection of just one model and instead combines a set of prognostic models, $S_1(\cdot), \ldots, S_k(\cdot)$, into one final model. The most common approach is to construct a weighted average of the considered models

$$S_{ens}(\cdot) = \sum_{i=1}^{k} \beta_i S_i(\cdot).$$

Diversity in ensemble models

Usually ensemble methods work well when combining a diverse set of models. In fact, it can be preferable to use a diverse set of weaker models instead of a set of similar models that perform well when used individually, this is exemplified with a hypothetical example.

Suppose that we have a population where HL and DLBCL are equally prevalent and that we want to make a classifier for discriminating between the DLBCL and HL cases. To accomplish this, 100 pathologists are given the same information and tell us whether the patient has DLBCL or HL. Let us assume that, conditional on the information they receive, the classifications provided by these pathologists are somehow independent of each other. Furthermore, the provided classifications are correct in 60% of the cases. The predictions are then combined and the final classification corresponds to the majority vote. After some calculations, it is seen that the ensemble of pathologists will get the answer right in around 97% of the cases. Let us now imagine a different scenario in which we have another 100 pathologists who all use the same classification algorithm which is correct in 90% of the cases. Each pathologist will therefore correctly classify a case with a probability of 90% and the predictions are perfectly correlated. If we apply the same majority voting system, the board of pathologists will deliver the correct classification in 90% of the cases. Hence, by combining independent opinions of sub-optimal pathologists we get a better classifier than by combining strongly correlated opinions of well-performing pathologists.

Selection of the weights

By imposing restrictions on the β parameters, different properties of $S_{ens}(\cdot)$ can be ensured. For example, if all included models lead to proper survival functions it is natural to assume that $\beta_i > 0$, $\forall i \in \{1, \ldots, k\}$ and $\sum_{i=1}^{k} \beta_i = 1$ which ensures that $S_{ens}(\cdot)$ is a survival function. Depending on the goal of the analysis, there are a number of different approaches for estimating the β parameters. The most simple approach is to give each model an equal weight, $\beta_i = \frac{1}{k}$, $\forall i \in \{1, \ldots, k\}$. Although this idea might seem simplistic, it can lead to powerful methods such as the random survival forest methodology in which survival trees are combined with equal weight.[64] Another popular approach is Bayesian model averaging in which the posterior probabilities of the models given the data are used as β parameters.[65]

Another popular approach is stacking (also called super-learning).[66, 67] The β parameters obtained by stacking are those that minimize a cross-validated prediction error. In the case of survival models, one can e.g. use the Brier score as measure of the prediction error.[68] Assuming that m-fold cross-validation was performed, we can denote the models estimated using the dataset in which the i'th observation was left out by $\hat{S}_1^{-i}, \ldots, \hat{S}_k^{-i}$. The β parameter estimates are then defined as

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{\mathbb{1}(\min(T_i, t) \leq C_i)}{\hat{G}(\min(T_i, t)|\boldsymbol{X}_i)} \left(\mathbb{1}(T_i \geq t) - \sum_{j=1}^{k} \beta_j \hat{S}(t)_j^{-i} \right).$$

If we add the constraint that $\beta_i > 0$, $\forall i \in \{1, \ldots, k\}$ and $\sum_{i=1}^{k} \beta_i = 1$ this is the stacking method described by Wey et al.[68]

An important advantage of this stacking approach is that by minimizing the cross-validated error, models which tend to overfit will not be given inappropriately large weights. Furthermore, both theoretical studies as well as practical applications have shown that stacking usually leads to a performance improvement over selecting one best model.[67, 69]

4 Measures of survival

The survival and hazard functions are the two most often modelled quantities in survival analyses. However, survival curves and especially hazard functions might not be easily understood by the general public. Furthermore, without a working knowledge of the research field, survival estimates might be hard to interpret due to a lack of information on the expected survival had the disease or condition not been present. As alternatives to the hazard and

survival functions, we introduce the standardized mortality ratio and the loss of life expectancy which both relate the survival of the patients to that of a similar background population.

4.1 Standardized mortality ratios

The standardized mortality ratio (SMR) is defined as the number of deaths observed, O, divided by the number of deaths that would be expected in a background population with a similar gender, age, entry-date, and potential follow-up, E. In cancer research, SMRs have often been used to compare the survival of the study population to that of the general population and recent examples can be found for DLBCL and HL.[4, 70, 71]

To calculate the SMR it is reasonable to assume that the study and background populations are independent conditional on the population characteristics, hence an estimator of the SMR can be obtained as $\frac{\hat{O}}{\hat{E}}$. The obvious estimator choice for the numerator is $\hat{O} = \sum_{i=1}^{n} \Delta_i$. For the calculation of E two methods, the person-year method and prospective method, have been proposed.[72, 73] Both rely on information about the survival in the general population which can be extracted from databases such as the Human Mortality Database.[74]

The person-year method is the most commonly used method.[75] From now on, the hazard rate of a background population with similar characteristics as observations i will be denoted by $\alpha_i^*(t)$. The person-year estimator of the expected number of deaths is then defined as

$$\hat{E}_{py} = \sum_{i=1}^{n} \int_0^{U_i} \alpha_i^*(u) du. \tag{3}$$

Under the assumption that the hazard in the study population is equal to that of the background population this estimator will lead to an unbiased estimate of the expected number of deaths. However, if one is willing to accept this assumption there is little use in estimating the SMR since, by assumption, it will be equal to one.[73, 76] If the hazard functions in the study population and background population are not equal the estimator presented in Equation 3 will be a biased estimate of the expected number of deaths.[73, 76] Intuitively this can be explained by looking at the upper limit of the integral, U_i. Since U_i depends on the hazard in the study population the estimator depends on both $\alpha_i(t)$ and $\alpha_i^*(t)$. Although the estimator $\frac{\hat{O}}{\hat{E}_{py}}$ loses its interpretation as a standardized mortality ratio it can be shown that if $\alpha_i(t) = \theta\alpha_i^*(t)$ then $\frac{\hat{O}}{\hat{E}_{py}}$ is the MLE of θ.[77]

In order to use the prospective method information on the potential follow-up for each observation i, i.e. C_i, is needed. In the register studies considered in this thesis the potential follow-up time can be assumed to be equal to the time between diagnosis and the date of the extraction of the vital status from the national register. In other scenarios, C_i is often not known. The prospective method is based on the fact that if we had a subject from the background population with the same characteristics as observation i who we also followed for C_i time then the probability of observing a death would be[73]

$$1 - S_i^*(C_i) = 1 - \exp\left(-\int_0^{C_i} \alpha_i^*(u)du\right).$$

Hence, the number of expected deaths in a background cohort with similar characteristics and follow-up as the study cohort can be estimated as

$$\hat{E}_{pr} = \sum_{i=1}^{n} \left(1 - S_i^*(C_i)\right).$$

Unlike $\hat{E}_{py}$, $\hat{E}_{pr}$ is an unbiased estimator of the expected number of deaths irrespective of whether or not the hazard functions in the study and background populations are equal. The biggest drawback of this estimator is that it requires knowledge of C_i which is unavailable in a lot of epidemiological studies.[72]

In practice, the estimator choice will partially depend on whether or not the potential follow-up times are available. In light of the bias of the person-year estimator for estimating the number of expected deaths, the practice of calling $\frac{\hat{O}}{\hat{E}_{py}}$ an estimator of the SMR seems inappropriate and this estimator should probably be referred to as a ratio of hazard rates.[78]

4.2 Loss of life expectancy

The life expectancy, or the expected time until the event happens is equal to

$$E(T) = \int_0^{+\infty} S(u)du.$$

Similarly, one can define the life expectancy of a background cohort with similar characteristics as

$$E(T^*) = \int_0^{+\infty} S^*(u)du$$

and the loss in life expectancy is defined as the difference between these two life expectancies

$$E(T^*) - E(T) = \int_0^{+\infty} (S^*(u) - S(u))\, du. \tag{4}$$

In order to estimate this quantity, one can plug-in estimates of $S_i^*(t)$ and $S_i(t)$ into Equation 4. A major problem with this approach is that for time points larger than the longest follow-up time $\hat{S}(\cdot)$ is not based on any available data and the tail of the survival function is based on extrapolation. When e.g. the Kaplan-Meier estimator is used in combination with a dataset in which the largest observed time point corresponds with a censored observation one gets $\hat{S}(t) > 0, \forall t$ and $\int_0^{+\infty} \hat{S}(u)du = \infty$. In order to avoid this behaviour, it has been recommended to use parametric estimates of $S(\cdot)$.[79, 80] Even when parametric estimators are used the life expectancy estimates will be based on extrapolation and can depend largely on the used method.[81] Especially when the study population consists of young patients for who it is expected that there is a long period over which the excess hazard changes, e.g. due to late toxicities, very long follow-up is necessary to obtain reliable estimates of the life expectancy.[80, 81]

An alternative to the unrestricted loss of life expectancy can be obtained by replacing the upper bound in Equation 4 with a value τ which falls well within the follow-up period.[82, 83] The restricted loss of life expectancy is thus formally defined as

$$E(\min(T^*, \tau)) - E(\min(T, \tau)) = \int_0^{\tau} (S^*(u) - S(u))\, du.$$

The theory surrounding pseudo-values for the restricted mean survival can be extended to define pseudo-values for the restricted loss of lifetime

$$\rho_i = \int_0^{\tau} S_i^*(u)du - n \int_0^{\tau} \hat{S}(u)du + (n-1) \int_0^{\tau} \hat{S}^{-i}(u)du.$$

Regression models can then be obtained by using the pseudo-values as outcome variables and the methodology described in Section 2.5 applies.

5 Landmark analysis

5.1 Introduction

Descriptions of the survival or risk of an event at times measured from the diagnosis date are common throughout the literature. Recently there has

been an increasing interest in the evolution of survival measures given that patients reach landmarks such as two years of event-free survival. For example, multiple studies have shown that the survival prospects of systemic DLBCL patients become similar to those of an age- and gender-matched background population given that they reach two years of event-free survival.[4, 70] This dynamic information is useful in multiple settings such as patient counselling, cost-benefit analyses of follow-up programs, and can be the motivation behind clinical trial end-points.

An example of a quantity modelled in a dynamic analysis is the probability that patients who survive up to time s will survive another x-years. Hence, the quantity of interest is $P(T > x + s|T > s)$. More generally, we will denote that the condition of interest is met at time s by $L(s) = 1$ and if it is not met we will write $L(s) = 0$. The quantity of interest is regularly a summary measure that describes the survival from the landmark time s over the next x years and we denote this quantity with $\mu(s, x)$. Usually the window x is kept fixed and the evolution of $\mu(s, x)$ in function of s is the quantity that is reported. Returning to the previous example we get that $L(s) = \mathbb{1}(T > s)$ and $\mu(s, x) = P(T > x + s|L(s) = 1) = P(T > x + s|T > s)$. In this particular scenario, an explicit expression of $\mu(s, x)$ can be obtained

$$\mu(s, x) = P(T > x + s|T > s) = \frac{S(x + s)}{S(s)}.$$

A similar expression can be constructed for the cumulative incidence given survival up to time s. However, for other measures such as standardized mortality ratios the calculation by use of conditional expectations is less straightforward.

An alternative approach is landmarking which was originally introduced as a method to avoid immortal time bias.[84] The idea of a landmark analysis is that we only focus on the observations for which it is known that the condition of interest was met. In the case where the condition is measured at a time s we will denote the corresponding dataset by $\mathcal{L}(s)$. Measures such as the standardized mortality ratio for patients for who the landmark condition was met can then be estimated by applying the usual estimators to the $\mathcal{L}(s)$ dataset.

A practical example can be found in Figure 5 in which the evolution of the five-year overall survival (OS) measured from different OS landmarks is described for DLBLC patients by applying the Kaplan-Meier estimator to landmarked datasets. In this case, it is clear that the five-year OS increases over time and seems to plateau at different times in different subgroups. A draw-

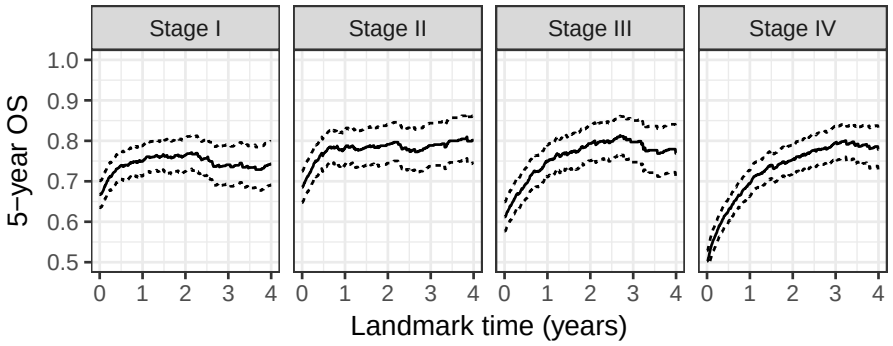

Figure 5: The evolution of the five-year overall survival (OS). The x-axis displays the landmark time, i.e. the number of years of OS reached. The y-axis shows the OS in the subsequent five years for the DLBCL patients who survived up to the landmark time (x-axis).

back of using the landmarking approach in this way is that the estimates can be quite unstable if relatively few observations meet the landmark condition.

5.2 Super models

To avoid the unstability of the landmark estimates $\hat{\mu}(s,x)$ and in order to incorporate covariate effects, Van Houwelingen et al proposed to model $\mu(s,x|\boldsymbol{X})$ by use of a spline model $\mu(s,x|\boldsymbol{X}) = g(\boldsymbol{X}^t\boldsymbol{\Theta}(s,\boldsymbol{\theta}))$ in which $\boldsymbol{\Theta}(s,\boldsymbol{\theta})$ is a spline parameterized by $\boldsymbol{\theta}$ and $g(\cdot)$ is an appropriate link function.[85] This approach of using coefficients in function of the time s has been coined super modelling and multiple suggestions on how to fit such models have been made.[85, 86] The method we focus on is the use of landmarked pseudo-values since it provides a general framework for the dynamic modelling of the restricted mean survival time and other summary measures.[86, 87]

A set of dynamic pseudo-values is defined for each of a number of uniformly spaced times, $s_1,\ldots,s_k$, in the interval $[0,\tau]$. For each s_j there is a dataset $\mathcal{L}(s_j)$ of patients who met the landmark condition and are uncensored at time s. For each observation $i \in \mathcal{L}(s_j)$ its dynamic pseudo-observation is then defined as

$$\hat{\rho}_{i,s_j} = n_{s_j}\hat{\rho}_{s_j} - (n_{s_j}-1)\hat{\rho}_{s_j}$$

in which n_{s_j} is the number of observations in the $\mathcal{L}(s_j)$ dataset and $\hat{\rho}_{s_j}$ is obtained by using an appropriate $\hat{\rho}$ estimator (see Section 2.5) on the $\mathcal{L}(s_j)$

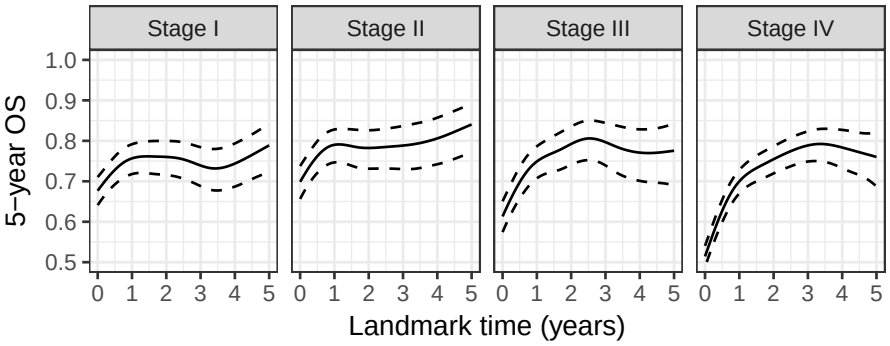

Figure 6: The evolution of the five-year overall survival (OS) modelled using a supermodel. The x-axis displays the landmark time, i.e. the number of years of OS reached. The y-axis displays the OS in the subsequent five years for the DLBCL patients who survived up to the landmark time (x-axis).

dataset.[86] The generalized estimation equations then become[86, 88]

$$U(\boldsymbol{\theta}) = \sum_{j=1}^{k} \sum_{i \in \mathcal{L}(s_j)} \frac{\partial}{\partial \boldsymbol{\theta}} g^{-1}(\boldsymbol{X}_i^t \boldsymbol{\Theta}(s_j, \boldsymbol{\theta})) V_{i,s_j}^{-1} (\hat{\rho}_{i,s_j} - g^{-1}(\boldsymbol{X}_i^t \boldsymbol{\Theta}(s_j, \boldsymbol{\theta}))) = \mathbf{0}$$

Just as in Section 2.5, an estimate of the variance can be obtained by use of a sandwich estimator.[86]

In Figure 6, the evolution of the five-year OS measured from different OS landmarks is described for DLBLC patients by using the supermodel approach. Comparing Figure 5 and Figure 6 it is seen that the supermodel estimates show less aberrant behaviour and are smoother.

6 Robust statistics

6.1 Introduction

Datasets from registers often contain some incorrect data. Potential sources of errors in the data include measurement errors, data entry errors, etc. When standard estimation methods are used the estimates can be heavily influenced by a few outliers. Estimation procedures that lead to sensible estimates even when a part of the data consists of outliers are called robust estimators. Robust statistics has been an active research field during the last decades.[89]

An example of the influence a few outliers can have on the ordinary least squares estimator (OLS) of regression coefficients is given in Figure 7. The

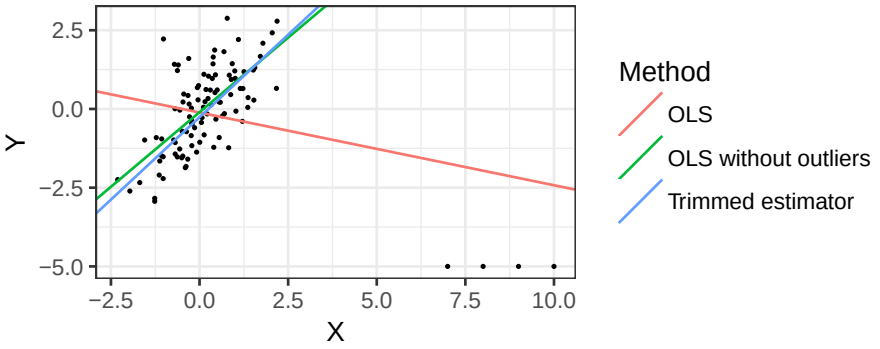

Figure 7: An example of different linear regression estimates based on a dataset with 100 observations simulated using a linear regression model and four additional outliers (the observations shown in the bottom-right part of the plot). The estimates were obtained using the ordinary least squares (OLS) estimator with/without outliers and a trimmed estimator.

dataset contained 100 observations which were simulated according to the regression model $Y_i = X_i + \epsilon_i$ in which X_i and ϵ_i were independently simulated from a standard normal distribution. Additionally, 4 observations that did not follow this relationship were added, these can be found in the bottom right of the figure. The estimated regression line obtained using OLS clearly does not represent the model that the majority of the data is based on. On the other hand, the regression line obtained using OLS on the data-set of 100 clean data-set models the majority of the data well.

The detection of observations which do not follow a model that fits the majority of the data can e.g. lead to the detection of new disease subtypes or prognostic biomarkers. An often used outlier detection approach is based on the residuals calculated using the model parameter estimates. Because standard non-robust estimates can be heavily biased in the direction of the outliers, residuals based on these estimates are often unable to recover these outliers, an effect called masking, and it might be that non-outlying observations are marked as outliers, an effect called swamping.[90] By use of more robust estimators of the model parameters, the detection of outliers using residuals becomes more feasible.

There are a number of different ways to construct robust estimators. The most intuitive robust estimators are the class of trimmed-estimators which are introduced below. Another type of estimators are M-estimators which generalize the maximum likelihood principle to other objective functions. By the appropriate selection of the objective function, M-estimators with certain robustness properties can then be constructed.[89] Other robust procedures

include MM-estimators, S-estimators, etc. For an introduction to these methods we refer to Maronna et al and Heritier et al.[89, 90]

6.2 Trimmed-estimators

Trimmed estimators attempt to find estimates which are optimal for the majority of the data by inspecting the fit on subsets of the full data-set. If we denote the set of all subsets of size k by $\mathcal{I}_k$ and the contribution of the i'th observation to the log-likelihood by $l_i(\boldsymbol{\theta})$ then the trimmed likelihood estimator is defined as:

$$\hat{\boldsymbol{\theta}} = \max_{\chi \in I_k} \max_{\boldsymbol{\theta}} \sum_{i \in \chi} l_i(\boldsymbol{\theta}).$$

If there are at most $n - k$ outliers, at least one of the subsets $\chi \in I_k$ will not contain outliers and $\max_{\boldsymbol{\theta}} \sum_{i \in \chi} l_i(\theta)$ will tend to be larger than the similar measures obtained with datasets containing some outliers. In practice often $k = \frac{n+1}{2}$ is chosen which ensures that the estimator does not break down as long as less than half the observations in the dataset are outliers.[91]

The regression line obtained by applying a trimmed log-likelihood estimator to the simulated dataset is shown in Figure 7. It is clear that this estimate is a better fit for the majority of the data than the standard OLS estimate.

An important shortcoming of this approach is that the estimate is based on only a portion of the original dataset, independent of whether or not the excluded observations are outliers. This implies that the procedure is not efficient when few or no outliers are present. This efficiency issue can be avoided by refitting the model using all observations which were not considered outlying given the trimmed log-likelihood estimates.[91] A second weakness is that the search space of subsets of size $n - k$ tends to be very large if the sample size is large. To avoid unnecessarily long computation times often an approximate solution is obtained by inspecting a limited number of subsets in $\mathcal{I}_k$, e.g. 500 randomly selected subsets.[91]

6.3 Robust estimators for censored data

Currently few robust estimators that can be used in combination with censored data exist.[90] Some exceptions include robust procedures for the estimation of the coefficients in the CPH model.[92–94] Recently, robust estimators for some parametric accelerated failure time and additive hazard models have also been introduced.[95–97] Notwithstanding these contributions, compared to the number of methods available for (generalized) linear models the selection of robust estimators that can be combined with censored data re-

mains seriously limited. In particular, to our knowledge, no outlier robust estimators exist for models with time-varying coefficients.

7 Overview of the thesis

The overarching theme of the papers that make up this thesis is the use of statistical and machine learning methods to improve our knowledge about the prognosis of haematological diseases. This is accomplished by relying on the techniques introduced in the background material.

In Paper I, the performance of the IPI, a prognostic index for DLBCL, is inspected and compared to that of a CPH model which avoids some of the major methodological issues that were present in the development of the IPI.[9] By use of the Brier score, C-index, and time-varying AUC the performance of the IPI is shown to be sub-optimal as compared to the CPH model which is implemented as a dashboard application available on `https://lymphomapredictor.org`.

In Paper II, the survival of acute promyelocytic leukaemia patients is compared to that of a similar background population. In this letter, the standardized mortality ratio and landmarking methodologies introduced in Sections 4 and 5 are applied. It is shown that the relative mortality is fairly minimal given that patients survive the critical three months period following diagnosis.

In Paper III, the predictive model introduced in Paper I is extended. A stacked model for the survival of DLBCL patients is developed by combining a number of non-, semi-, and fully parametric models. The performance of the stacked model is then compared to that of the IPI and the CPH model from paper I by use of the Brier score, C-index, and time-varying AUC. The stacked model is shown to outperform both the IPI and CPH model but the gain is relatively minimal given that additional clinicopathological variables are included in the stacked model. Also the stacked model is available on `https://lymphomapredictor.org`.

In Paper IV, eight commonly used prognostic indices for haematological malignancies are described. By use of the Brier score and time-varying AUC their performance is compared to that of a simple regression model based on age and performance status. None of the prognostic indices outperform this simple model casting doubt on their usefulness in clinical practice.

In Paper V, the survival and relapse risk for young HL patients is described.

To compare the survival of HL patients with that of a background population the restricted loss of life expectancy is combined with the landmark methodology. Furthermore, a landmark analysis based on a supermodel of pseudo-values is used to describe the relapse risk. By using pseudo-values of the restricted loss of life expectancy and relapse risk also multivariable models are described. The results of these analyses show that limited stage HL patients have excellent survival prospects. Furthermore, the five-year relapse risk is low for patients reaching two years of event-free survival which calls the use of follow-up programs focusing on detecting relapses post the two-year event-free survival landmark into question. Finally, it is shown that the relapse risk of advanced stage patients treated with 6-8 cycles of ABVD is similar to that of advanced stage patients treated with 6-8 cycles of BEACOPP. This is interesting given that the BEACOPP treated patients tended to have more adverse clinical characteristics implying that BEACOPP might have a higher efficacy with respect to preventing relapses.

Paper VI describes the use of the Brier score as a loss function in order to obtain a more robust estimation procedure for the coefficients of proportional hazard models. This new loss function is used for a penalized piecewise constant proportional hazard model which can model time-varying coefficients. The performance of this estimation procedure is then demonstrated on real and simulated data.

8 Limitations and future perspectives

8.1 Papers I, III, and IV

The prognostic models provided in Papers I and IV could be extended in multiple ways. First of all, since cancer is fundamentally a genetic disease, combining clinical data together with data describing the tumour genetics and biology could potentially lead to superior prognostic models. Secondly, in line with the landmark analyses of Papers II and V the presented prognostic models could be extended in order to provide updated prognostic information at landmark time-points such as being in remission for two years. Currently, survival curves given that the patient is alive a number of years post-diagnosis can be obtained from `http://lymphomapredictor.org`. However, also conditioning on patients not having experienced a relapse would be of interest. Two options for obtaining such dynamic prognostic models are multi state models or landmarked survival models.

It seems reasonable to assume that the developed scores work well in the

Nordic countries which tend to have similar health care systems, treatment guidelines, population characteristics, etc. However, to see whether or not the prognostic models perform well in populations covered by other health care systems, additional external validation studies will have to be performed. Furthermore, given that treatment regimens and health care systems keep evolving over time, the prognostic models developed in this thesis will have to be updated accordingly.

8.2 Papers II and V

In Paper II and Paper V the differences between the survival of patients and that of a background population are described. In the papers, quantitative measures of this difference are given together with intervals. This approach has as disadvantage that the dual hypothesis tests, i.e. a test of whether there is a difference between the two survival curves, has as null-hypothesis that there is no difference. One of the fundamental characteristics of hypothesis testing is that being unable to reject the null-hypothesis does not necessarily point towards the null-hypothesis being true. This problem is especially apparent when the sample sizes are small. To overcome this drawback, non-inferiority tests are often used in clinical trials. Similarly, when describing the differences between the survival of the patients and that of the background population non-inferiority tests could provide valuable complementary information.

A large number of young HL patients will be considered cured and are likely to be alive decades after their initial diagnosis. Given the large number of long term survivors, the importance of late onset toxicities such as cardiovascular disease is significant in HL.[98] In Paper V, the median follow-up was nine years. Since most of these late onset-toxicities happen more than a decade after treatment, our study was unable to take these into account. For example, we detected a negligible five-year restricted loss of life expectancy in limited stage HL patients. However, since the majority of late onset toxicity events would not have happened during the follow-up of this study, the unrestricted loss of life expectancy might still be increased. To which extent late onset toxicities remain problematic with contemporary treatments will have to be investigated when the follow-up is long enough and the necessary data becomes available.

Finally, in paper V the 6-8 cycles of BEACOPP and 6-8 cycles of ABVD regimens were compared. The results point towards BEACOPP leading to lower relapse risks. In future research, analyses adjusted for potential treatment confounders such as Ann Arbor stage could be of interest for the scientific community.

8.3 Paper VI

There are multiple potential extensions to the robust estimation procedure presented in Paper VI. First of all, a number of additional theoretical guarantees could be derived, e.g. results regarding the asymptotic normality of the proposed method when it is combined with proportional hazard models. Such results would allow the construction of confidence intervals and make the method more applicable in practice.

In the paper, it shown that leverage points have a limited influence on the score equations obtained when the Brier score is used as loss function. The derivation of this result currently relies on the proportional hazard assumption. Whether or not the robustness properties are similar for other classes of models remains uninvestigated. Furthermore, the current R implementation remains rather slow and more advanced optimization techniques might lead to significant calculation speed gains. Finally, by using the same methodology, robust estimation procedures for competing risk models could likely be constructed.

References

[1] S. H. Swerdlow, World Health Organization, and International Agency for Research on Cancer. *WHO classification of tumours of haematopoietic and lymphoid tissues*.

[2] L. S. Goodman and M. M. Wintrobe. Nitrogen mustard therapy; use of methyl-bis (beta-chloroethyl) amine hydrochloride and tris (beta-chloroethyl) amine hydrochloride for Hodgkin's disease, lymphosarcoma, leukemia and certain allied and miscellaneous disorders. *Journal of the American Medical Association*, 132:126–32, September 1946.

[3] J. R. Anderson, J. O. Armitage, and D. D. Weisenburger. Epidemiology of the non-Hodgkin's lymphomas: Distributions of the major subtypes differ by geographic locations. *Annals of Oncology*, 9(7):717–720, July 1998.

[4] L. H. Jakobsen, M. Bøgsted, P. d. N. Brown, B. Arboe, J. Jørgensen, et al. Minimal Loss of Lifetime for Patients With Diffuse Large B-Cell Lymphoma in Remission and Event Free 24 Months After Treatment: A Danish Population-Based Study. *Journal of clinical oncology*, 35(7):778–784, March 2017.

[5] J. O. Armitage. How I treat patients with diffuse large B-cell lymphoma. *Blood*, 110(1):29–36, July 2007.

[6] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000.

[7] J. W. Friedberg. How I treat double-hit lymphoma. *Blood*, 130(5):590–596, August 2017.

[8] A. J. M. Ferreri. How I treat primary CNS lymphoma. *Blood*, 118(3):510–22, July 2011.

[9] The International Non-Hodgkin's Lymphoma Prognostic Factors Project. A Predictive Model for Aggressive Non-Hodgkin's Lymphoma. *New England Journal of Medicine*, 329(14):987–994, September 1993.

[10] L. H. Sehn, B. Berry, M. Chhanabhai, C. Fitzgerald, K. Gill, et al. The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP. *Blood*, 109(5):1857–61, March 2007.

[11] Z. Zhou, L. H. Sehn, A. W. Rademaker, L. I. Gordon, A. S. Lacasce, et al. An enhanced International Prognostic Index (NCCN-IPI) for patients with diffuse large B-cell lymphoma treated in the rituximab era. *Blood*, 123(6):837–42, February 2014.

[12] P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, January 2006.

[13] J. Biccler, S. Eloranta, P. de Nully Brown, H. Frederiksen, M. Jerkeman, et al. Simplicity at the cost of predictive accuracy in diffuse large B-cell lymphoma: a critical assessment of the R-IPI, IPI, and NCCN-IPI. *Cancer Medicine*, December 2017.

[14] J. L. Biccler, S. Eloranta, P. de Nully Brown, H. Frederiksen, M. Jerkeman, et al. Optimizing Outcome Prediction in Diffuse Large B-Cell Lymphoma by Use of Machine Learning and Nationwide Lymphoma Registries: A Nordic Lymphoma Group Study. *JCO Clinical Cancer Informatics*, (2):1–13, October 2018.

[15] J. L. Biccler, T. C. El-Galaly, M. Bøgsted, J. Jørgensen, P. de Nully Brown, et al. Clinical prognostic scores are poor predictors of overall survival in various types of malignant lymphomas. *Leukemia & Lymphoma*, 2018.

[16] I. Glimelius, S. Ekberg, M. Jerkeman, E. T. Chang, M. Björkholm, et al. Long-term survival in young and middle-aged Hodgkin lymphoma patients in Sweden 1992-2009-trends in cure proportions by clinical characteristics. *American Journal of Hematology*, 90(12):1128–1134, December 2015.

[17] F. Merli, S. Luminari, P. G. Gobbi, N. Cascavilla, C. Mammi, et al. Long-Term Results of the HD2000 Trial Comparing ABVD Versus BEACOPP Versus COPP-EBV-CAD in Untreated Patients With Advanced Hodgkin Lymphoma: A Study by Fondazione Italiana Linfomi. *Journal of clinical oncology*, 34(11):1175–81, April 2016.

[18] A. K. Ng. Current survivorship recommendations for patients with Hodgkin lymphoma: focus on late effects. *Blood*, 124(23), 2014.

[19] P. P. Carbone, H. S. Kaplan, K. Musshoff, D. W. Smithers, and M. Tubiana. Report of the Committee on Hodgkin's Disease Staging Classification. *Cancer research*, 31(11):1860–1, November 1971.

[20] M. Hutchings, A. Loft, M. Hansen, L. M. Pedersen, A. K. Berthelsen, et al. Position emission tomography with or without computed tomography in the primary staging of Hodgkin's lymphoma. *Haematologica*, 91(4):482–9, April 2006.

[21] V. T. Devita, A. A. Serpick, and P. P. Carbone. Combination chemotherapy in the treatment of advanced Hodgkin's disease. *Annals of internal medicine*, 73(6):881–95, December 1970.

[22] C. A. Coltman and D. O. Dixon. Second malignancies complicating Hodgkin's disease: a Southwest Oncology Group 10-year followup. *Cancer treatment reports*, 66(4):1023–33, April 1982.

[23] J. M. Kaldor, N. E. Day, E. A. Clarke, F. E. Van Leeuwen, M. Henry-Amar, et al. Leukemia Following Hodgkin's Disease. *New England Journal of Medicine*, 322(1):7–13, January 1990.

[24] G. Bonadonna, R. Zucali, S. Monfardini, M. de Lena, and C. Uslenghi. Combination chemotherapy of Hodgkin's disease with adriamycin, bleomycin, vinblastine, and imidazole carboxamide versus MOPP. *Cancer*, 36(1):252–259, July 1975.

[25] V. Diehl, J. Franklin, M. Pfreundschuh, B. Lathan, U. Paulus, et al. Standard and Increased-Dose BEACOPP Chemotherapy Compared with COPP-ABVD for Advanced Hodgkin's Disease. *New England Journal of Medicine*, 348(24):2386–2395, June 2003.

[26] M. V. Maraldo, N. P. Brodin, M. C. Aznar, I. R. Vogelius, P. Munck af Rosenschöld, et al. Estimated risk of cardiovascular disease and secondary cancers with modern highly conformal radiotherapy for early-stage mediastinal Hodgkin lymphoma. *Annals of Oncology*, 24(8):2113–2118, August 2013.

[27] M. S. Tallman and J. K. Altman. How I treat acute promyelocytic leukemia. *Blood*, 114(25), 2009.

[28] M. E. Huang, Y. C. Ye, S. R. Chen, J. R. Chai, J. X. Lu, et al. Use of all-trans retinoic acid in the treatment of acute promyelocytic leukemia. *Blood*, 72(2):567–72, August 1988.

[29] J. H. Park, B. Qiao, K. S. Panageas, M. J. Schymura, J. G. Jurcic, et al. Early death rate in acute promyelocytic leukemia remains high despite all-trans retinoic acid. *Blood*, 118(5):1248–1254, August 2011.

[30] S. Lehmann, A. Ravn, L. Carlsson, P. Antunovic, S. Deneberg, et al. Continuing high early death rate in acute promyelocytic leukemia: a population-based report from the Swedish Adult Acute Leukemia Registry. *Leukemia*, 25(7):1128–1134, July 2011.

[31] H. J. Iland, M. Collins, K. Bradstock, S. G. Supple, A. Catalano, et al. Use of arsenic trioxide in remission induction and consolidation therapy for acute promyelocytic leukaemia in the Australasian Leukaemia and Lymphoma Group (ALLG) APML4 study: a non-randomised phase 2 trial. *The Lancet Haematology*, 2(9):e357–e366, September 2015.

[32] S. Jin, R. Pazdur, and R. Sridhara. Re-Evaluating Eligibility Criteria for Oncology Clinical Trials: Analysis of Investigational New Drug Applications in 2015. *Journal of clinical oncology*, 35(33):3745–3752, November 2017.

[33] M. J. Maurer, H. Ghesquières, B. K. Link, J.-P. Jais, T. M. Habermann, et al. Diagnosis-to-Treatment Interval Is an Important Clinical Factor in Newly Diagnosed Diffuse Large B-Cell Lymphoma and Has Implication for Bias in Clinical Trials. *Journal of clinical oncology*:JCO2017765198, April 2018.

[34] G. Thanarajasingam, L. M. Minasian, F. Baron, F. Cavalli, R. A. De Claro, et al. Beyond maximum grade: modernising the assessment and reporting of adverse events in haematological malignancies. *The Lancet. Haematology*, 5(11):e563–e598, November 2018.

[35] L. Ostgard, J. M. Nørgaard, K. Raaschou-Jensen, R. S. Pedersen, D. Rønnov-Jessen, et al. The Danish National Acute Leukemia Registry. *Clinical Epidemiology*, Volume 8:553–560, October 2016.

[36] B. Arboe, P. Josefsson, J. Jørgensen, J. Haaber, P. Jensen, et al. Danish National Lymphoma Registry. *Clinical epidemiology*, 8:577–581, 2016.

[37] E. L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457, June 1958.

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

[39] N. Breslow. Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society*, 34:216–217, 1972.

[40] P. Royston and M. K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, August 2002.

[41] O. O. Aalen, Ø. Borgan, and H. K. Gjessing. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer New York, New York, 2008.

[42] M. A. Hernán. The hazards of hazard ratios. *Epidemiology*, 21(1):13–5, January 2010.

[43] T. Martinussen and T. H. Scheike. *Dynamic Regression Models for Survival Data*. Statistics for Biology and Health. Springer New York, New York, 2006.

[44] O. O. Aalen and S. Johansen. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5:141–150, 1978.

[45] J. P. Fine and R. J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496, June 1999.

[46] P. K. Andersen, J. P. Klein, and S. Rosthøj. Generalised Linear Models for Correlated Pseudo-Observations, with Applications to Multi-State Models. *Biometrika*, 90:15–27.

[47] P. K. Andersen and M. Pohar Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99, February 2010.

[48] F. Graw, T. A. Gerds, and M. Schumacher. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255, June 2009.

[49] M. Overgaard, E. T. Parner, and J. Pedersen. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics*, 45(5):1988–2015, October 2017.

[50] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, April 1986.

[51] M. Jacobsen and T. Martinussen. A Note on the Large Sample Properties of Estimators Based on Generalized Linear Models for Correlated Pseudo-observations. *Scandinavian Journal of Statistics*, 43(3):845–862, September 2016.

[52] M. Overgaard, E. T. Parner, and J. Pedersen. Estimating the variance in a pseudo-observation scheme with competing risks. *Scandinavian Journal of Statistics*, May 2018.

[53] K. G. M. Moons, P. Royston, Y. Vergouwe, D. E. Grobbee, and D. G. Altman. Prognosis and prognostic research: what, why, and how? *BMJ (Clinical research ed.)*, 338:b375, February 2009.

[54] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

[55] K. V. Ballman. Biomarker: Predictive or Prognostic? *Journal of clinical oncology*, 33(33):3968–71, November 2015.

[56] G. Shmueli. To Explain or to Predict? *Statistical Science*, 25(3):289–310, August 2010.

[57] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1):128–38, January 2010.

[58] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[59] T. A. Gerds and M. Schumacher. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*, 48(6):1029–1040, December 2006.

[60] U. B. Mogensen, H. Ishwaran, and T. A. Gerds. Evaluating Random Forests for Survival Analysis using Prediction Error Curves. *Journal of statistical software*, 50(11):1–23, September 2012.

[61] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–17, May 2011.

[62] P. J. Heagerty and Y. Zheng. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1):92–105, March 2005.

[63] H. Hung and C.-T. Chiang. Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010.

[64] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008.

[65] C. T. Volinsky, A. E. Raftery, D. Madigan, and J. A. Hoeting. David Draper and E. I. George, and a rejoinder by the authors. *Statistical Science*, 14(4):382–417, November 1999.

[66] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, January 1992.

[67] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), January 2007.

[68] A. Wey, J. Connett, and K. Rudser. Combining parametric, semiparametric, and non-parametric survival models with stacked survival models. *Biostatistics*, 16(3):537–549, July 2015.

[69] S. Rose. Mortality Risk Score Prediction in an Elderly Population Using Machine Learning. *American Journal of Epidemiology*, 177(5):443–452, March 2013.

[70] M. J. Maurer, H. Ghesquières, J.-P. Jais, T. E. Witzig, C. Haioun, et al. Event-free survival at 24 months is a robust end point for disease-related outcome in diffuse large B-cell lymphoma treated with immunochemotherapy. *Journal of clinical oncology*, 32(10):1066–73, April 2014.

[71] G. Hapgood, Y. Zheng, L. H. Sehn, D. Villa, R. Klasa, et al. Evaluation of the Risk of Relapse in Classical Hodgkin Lymphoma at Event-Free Survival Time Points and Survival Comparison With the General Population in British Columbia. *Journal of Clinical Oncology*, 34(21):2493–2500, July 2016.

[72] N. Keiding and M. Væth. Calculating expected mortality. *Statistics in Medicine*, 5(4):327–334, July 1986.

[73] A. J. Hartz, E. E. Giefer, and R. G. Hoffmann. A Comparison of two methods for calculating expected mortality. *Statistics in Medicine*, 2(3):381–386, October 1983.

[74] University of California Berkeley and Max Planck Institute for Demographic Research. Human Mortality Database.

[75] R. A. Case and A. J. Lea. Mustard gas poisoning, chronic bronchitis, and lung cancer; an investigation into the possibility that poisoning by mustard gas in the 1914-18 war might be a factor in the production of neoplasia. *British journal of preventive & social medicine*, 9(2):62–72, April 1955.

[76] G. Berry. The analysis of mortality by the subject-years method. *Biometrics*, 39(1):173–84, March 1983.

[77] N. E. Breslow. Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45, April 1975.

[78] D. B. Richardson, A. P. Keil, S. R. Cole, and R. F. MacLehose. Observed and Expected Mortality in Cohort Studies. *American Journal of Epidemiology*, 185(6):479–486, March 2017.

[79] C. P. Nelson, P. C. Lambert, I. B. Squire, and D. R. Jones. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*, 26(30):5486–5498, December 2007.

[80] T. M.-L. Andersson, P. W. Dickman, S. Eloranta, M. Lambe, and P. C. Lambert. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Statistics in Medicine*, 32(30):5286–5300, December 2013.

[81] L. H. Jakobsen. *Analysis of Relative Survival Patterns in Cancer Register Data*. PhD thesis, Aalborg University, 2018.

[82] P. Royston and M. K. B. Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, 30(19):2409–2421, August 2011.

[83] P. K. Andersen, M. G. Hansen, and J. P. Klein. Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations. *Lifetime Data Analysis*, 10(4):335–350, December 2004.

[84] J. R. Anderson, K. C. Cain, and R. D. Gelber. Analysis of survival by tumor response. *Journal of clinical oncology*, 1(11):710–9, November 1983.

[85] H. C. Van Houwelingen. Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, March 2007.

[86] M. A. Nicolaie, J. C. van Houwelingen, T. M. de Witte, and H. Putter. Dynamic Pseudo-Observations: A Robust Approach to Dynamic Prediction in Competing Risks. *Biometrics*, 69(4):1043–1052, December 2013.

[87] M. K. Grand and H. Putter. Regression models for expected length of stay. *Statistics in Medicine*, 35(7):1178–1192, March 2016.

[88] B. F. Kurland and P. J. Heagerty. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6(2):241–258, April 2005.

[89] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK, March 2006.

[90] S. Heritier, E. Cantoni, S. Copt, and M.-P. Victoria-Feser. *Robust Methods in Biostatistics*. John Wiley & Sons, Ltd, Chichester, UK, 2009.

[91] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Wiley, 1987.

[92] T. Bednarski. Robust estimation in Cox's regression model. *Scandinavian Journal of Statistics*:213–225, 1993.

[93] A. Farcomeni and S. Viviani. Robust estimation for the Cox regression model based on trimming. *Biometrical Journal*, 53(6):956–973, November 2011.

[94] P. Sasieni. Some New Estimators for Cox Regression. *The Annals of Statistics*, 21(4):1721–1759, December 1993.

[95] I. Locatelli, A. Marazzi, and V. J. Yohai. Robust accelerated failure time regression. *Computational Statistics & Data Analysis*, 55(1):874–887, 2011.

[96] C. Agostinelli, I. Locatelli, A. Marazzi, and V. J. Yohai. Robust estimators of accelerated failure time regression with generalized log-gamma errors. *Computational Statistics & Data Analysis*, 107:92–106, 2017.

[97] E. E. Álvarez and J. Ferrario. Robust estimation in the additive hazards model. *Communications in Statistics - Theory and Methods*, 45(4):906–921, February 2016.

[98] F. E. van Leeuwen and A. K. Ng. Late sequelae in Hodgkin lymphoma survivors. *Hematological Oncology*, 35:60–66, June 2017.

ISSN (online): 2246-1302
ISBN (online): 978-87-7210-376-1

AALBORG UNIVERSITY PRESS