UMEÅ UNIVERSITY

Estimation of hazard ratios from observational data with applications related to stroke

Guilherme Wang de Faria Barros

Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå 2024

Doctoral Thesis
Department of Statistics
Umeå School of Business, Economics and Statistics
Umeå University
SE-901 87 Umeå

Statistical Studies No. 57
ISBN: 978-91-8070-240-9 (print)
ISBN: 978-91-8070-241-6 (pdf)
ISSN: 1100-8989
Electronic version available at http://umu.diva-portal.org/

Printed by: Cityprint i Norr AB, Umeå University
Umeå, Sweden 2024

Find out the cause of this effect,
Or rather say, the cause of this defect,
For this effect defective comes by cause.

WILLIAM SHAKESPEARE
HAMLET, ACT II, SCENE 2

Contents

List of papers

The thesis is based on the following papers:

I. Beharry, J., Yogendrakumar, V., Barros, G. W. F., Davis, S., Norrving, B., Figtree, G. A., Donnan, G., von Euler M. and Eriksson, M. (2023). Recurrent ischemic stroke and mortality in stroke patients without standard modifiable risk factors: An analysis of the Riksstroke registry. *Submitted, under review.*

II. Barros, G. W. F., Eriksson, M. and Häggström, J. (2023). Performance of modeling and balancing approach methods when using weights to estimate treatment effects in observational time-to-event settings. *PLoS ONE 18(12): e0289316.*

III. Barros, G. W. F. and Häggström, J. (2023). Impact of non-informative censoring on propensity score based estimation of marginal hazard ratios. *Submitted, under review*

IV. Barros, G. W. F. and Häggström, J. (2023). Covariate selection for the estimation of marginal hazard ratios in high-dimensional data. *Manuscript*

Abstract

The objective of this thesis is to examine some challenges that may emerge when conducting time-to-event studies based on observational data. Time-to-event (also called survival) is a setting that involves analyzing how different factors may influence the length of time until an individual experiences the event of interest. This type of analysis is commonly applied in fields such as medical research and epidemiology. In this thesis, which focuses on stroke, we are interested in the time to a recurrent stroke or the death of a patient who survived a first stroke.

Hazard ratios are one of the main parameters estimated in time-to-event studies. Hazard ratios involve comparing the risk of experiencing the event between two groups, usually a treated group and an untreated group. They can also involve other factors, such as different age groups. Hazard ratios can be estimated from the data by using the Cox regression model.

Observational data, in contrast to experimental data, involves data collected without any intervention or random assignment of treatment to the individuals. Confounders, that is, variables that distort or obscure the true relationship between treatment and outcome, are always present and need to be controlled for in observational studies.

National registers are an important source of observational data. A national registry is a centralized database or system that collects, stores, and maintains information about a specific population or group of individuals within a country. Sweden is known for its detailed and complete national registers. In this thesis, data from the Swedish Stroke Register (Riksstroke) is used to study factors related to stroke.

In time-to-event studies involving observational data, several challenges may arise for the researcher during data analysis. Some individuals may not experience the event during the observation period and thus the information about their time until the event is incomplete. These individuals are considered as censored. Some individuals may experience another event rather than the one of interest, a competing risk. Additionally, models must be properly constructed, with researchers selecting variables and determining the suitable functional form.

Four papers are included in the thesis. Paper I demonstrates how to handle competing risks in survival analysis. The study involves comparing individuals with and without standard modifiable risk factors and their

risks of a recurrent stroke or death using data from the Swedish Stroke Register.

The estimation of marginal hazard ratios is a common theme in the other three papers. All involve simulation studies in order to extend methods and explore best practices when estimating marginal hazard ratios.

Paper II explores non-parametric methods that can be used as alternatives to more traditional parametric methods when balancing datasets in order to estimate a marginal hazard ratio. A case study was also conducted using data from the Swedish Stroke Register involving the prescription of anticoagulants at hospital discharge after a stroke.

Paper III is about how censoring affects marginal hazard ratio estimation, even with perfect balancing of the dataset. We study this issue, taking into consideration varying effect sizes and censoring rates. A procedure to attenuate the problem is also studied.

Paper IV concerns covariate selection in the case of high-dimensional data. High-dimensional data involves cases in which the number of covariates in the study is comparable to the number of individuals, and therefore covariate selection methods are needed. In the paper, we explore some of these methods and suggest a best-performing procedure. As Paper II, Paper IV involves a case study of anticoagulant prescription using data from the Swedish Stroke Register.

KEYWORDS: survival analysis, causal inference, hazard ratios, marginal hazard ratio, stroke, balancing

Sammanfattning (Summary in Swedish)

Syftet med denna avhandling är att undersöka några av de utmaningar som kan uppstå när man genomför tid-till-händelse-studier baserade på observationsdata. Tid-till-händelse-analys (även känt som överlevnadsanalys) används för att undersöka hur olika faktorer påverkar tiden till att en specifik händelse inträffar för en individ. Denna analysmetod är vanlig inom medicinsk forskning och epidemiologi. I denna avhandling, som fokuserar på stroke, undersöks tiden till en återkommande stroke eller död hos patienter som överlevt en första stroke.

Viktiga parameter i tid-till-händelse-studier är hasardkvoter, som jämför risken för händelsen mellan två grupper. T.ex. en medicinskt behandlad och en obehandlad grupp eller olika åldersgrupper. Vanligtvis används en Cox regressionsmodell för att uppskatta hasardkvoter.

Observationsdata skiljer sig från experimentella data genom att de samlas in utan intervention eller slumpmässig tilldelning av behandling. Förväxlingsfaktorer, som kan förvränga eller dölja det verkliga sambandet mellan behandling och utfall, måste då tas i beaktning i observationsstudier.

Nationella register är en viktig källa till observationsdata. Ett nationellt register är en central databas eller ett system som samlar in, lagrar och underhåller information om en specifik population eller grupp av individer inom ett land. Sverige är känt för sina detaljerade och kompletta nationella register. I denna avhandling används data från det svenska strokeregistret (Riksstroke) för att studera faktorer relaterade till stroke.

I tid-till-händelse studier med observationsdata kan flera utmaningar uppstå för forskaren under dataanalysen. Vissa individer kanske inte upplever händelsen under observationsperioden och därmed är informationen om deras tid fram till händelsen ofullständig. Dessa individer betraktas som censurerade. Vissa individer kan uppleva en annan händelse än den som är av intresse, en konkurrerande risk. Dessutom är noggrann modellkonstruktion, med bra variabelval och lämplig funktionell form, avgörande.

Fyra artiklar ingår i avhandlingen. I Artikel I demonstreras hur man hanterar konkurrerande risker i överlevnadsanalys. I studien jämförs individer med och utan standardiserade modifierbara riskfaktorer och deras risk för återkommande stroke eller död med hjälp av data från det svenska strokeregistret.

Uppskattningen av marginella hasardkvoter är ett gemensamt tema i de övriga tre artiklarna. Alla innehåller simuleringsstudier för att utvidga metoderna och utforska bästa praxis vid uppskattning av marginella hasardkvoter.

I Artikel II undersöks icke-parametriska metoder som kan användas som alternativ till mer traditionella parametriska metoder när man balanserar dataset för att uppskatta en marginella hasardkvoter. En fallstudie genomfördes också med hjälp av data från det svenska strokeregistret om förskrivning av antikoagulantia vid utskrivning från sjukhus efter en stroke.

Artikel III handlar om hur censurering påverkar skattning av marginella hasardkvoter, även med perfekt balansering av datasetet. Vi studerar denna fråga och tar hänsyn till varierande effektstorlekar och censureringsfrekvenser. En procedur för att minska felskattningen vid censurering studeras också.

Artikel IV handlar om urval av kovariater vid högdimensionella data. Högdimensionella data innebär fall där antalet kovariater i studien är jämförbart med antalet individer. I uppsatsen utforskar vi några av dessa metoder och föreslår ett förfarande som ger bäst resultat. I likhet med Artikel II innehåller Artikel IV en fallstudie av förskrivning av antikoagulantia med hjälp av data från det svenska strokeregistret.

Resumo (Summary in Portuguese)

O objetivo dessa tese é examinar alguns desafios que podem surgir quando realizam-se estudos de tempo até evento usando dados observacionais. Análise de tempo até evento (também conhecida como de análise de sobrevivência) é um cenário que envolve a análise de como diferentes fatores podem influenciar o período de tempo até que um indivíduo vivencie o evento de interesse. Esse tipo de análise é comumente aplicado em áreas como pesquisa médica e epidemiologia. Nesta tese, que tem como foco acidente vascular cerebral (AVC), estamos interessados no tempo até um AVC recorrente ou até a morte de um paciente que sobreviveu a um primeiro AVC.

Razões de risco são um dos principais parâmetros estimados em estudos de tempo até o evento. Razões de risco envolvem a comparação do risco de ocorrência do evento entre dois grupos, geralmente um grupo que recebeu tratamento e um grupo não tratado. Razões de risco também podem envolver outros fatores, como diferenças de idade. Razões de risco podem ser estimadas utilizando o modelo de regressão de Cox.

Dados observacionais, em contraste com os dados experimentais, envolvem dados coletados sem qualquer intervenção ou atribuição aleatória de tratamento aos indivíduos. Fatores de confusão, ou seja, as variáveis que distorcem ou obscurecem a verdadeira relação entre o tratamento e o resultado, estão sempre presentes e precisam ser controlados em estudos observacionais.

Registros nacionais são uma importante fonte de dados observacionais. Um registro nacional é um banco de dados ou sistema centralizado que coleta, armazena e mantém informações sobre uma população específica ou um grupo de indivíduos em um país. A Suécia é conhecida por seus registros nacionais detalhados e completos. Nesta tese, os dados do Registro Sueco de AVC (Riksstroke) são usados para estudar fatores relacionados ao AVC.

Em estudos de tempo até o evento envolvendo dados observacionais, vários desafios podem surgir para o pesquisador durante a análise de dados. Alguns indivíduos podem não passar pelo evento durante o período de observação e, portanto, as informações sobre o tempo decorrido até o evento são incompletas. Esses indivíduos são considerados censurados. Indivíduos podem também passar por outro evento em vez daquele de interesse, um risco concorrente. Além disso, os modelos devem ser con-

struídos adequadamente, com os pesquisadores selecionando as variáveis e determinando a forma funcional adequada.

Quatro artigos estão incluídos na tese. O Artigo I demonstra como lidar com riscos concorrentes na análise de sobrevivência. O estudo envolve a comparação de indivíduos com e sem fatores de risco modificáveis padrão e seus riscos de um AVC recorrente ou morte usando dados do Registro Sueco de AVC.

A estimativa de razões de risco marginais é um tema comum nos outros três artigos. Todos envolvem estudos de simulação para ampliar os métodos e explorar as práticas recomendadas ao estimar razões de risco marginais.

O Artigo II explora métodos não paramétricos que podem ser usados como alternativas aos métodos paramétricos mais tradicionais ao equilibrar conjuntos de dados para estimar uma razão de risco marginal. Também foi realizado um estudo de caso usando dados do Swedish Stroke Register envolvendo a prescrição de anticoagulantes na alta hospitalar após um AVC.

O Artigo III trata de como a censura afeta a estimativa de razões de risco marginal, mesmo com o balanceamento perfeito do conjunto de dados. Estudamos essa questão, levando em consideração o tamanhos de efeito do tratamento e diferentes taxas de censura. Também é estudado um procedimento para atenuar o problema.

O Artigo IV trata da seleção de covariáveis no caso de dados de altas dimensões. Os dados de altas dimensões envolvem casos em que o número de covariáveis no estudo é comparável ao número de indivíduos e, portanto, são necessários métodos de seleção de covariáveis. Neste artigo, exploramos alguns desses métodos e sugerimos um procedimento de melhor desempenho. Como no Artigo II, o Artigo IV envolve um estudo de caso de prescrição de anticoagulantes usando dados do Registro Sueco de AVC.

Preface

"It takes a village to raise a child". The journey of a PhD student and a thesis bears a striking resemblance to that idea. The fruition of this thesis owes its existence to the support and assistance provided by numerous individuals and institutions, each playing a crucial role. Throughout the span of my PhD studies, from September 2019 to February 2024, which coincided with the challenges posed by the COVID-19 pandemic, the backing from these entities proved more indispensable than ever.

First, I extend my gratitude to my supervisor, Jenny Häggström, for affording me this opportunity and offering unwavering support and collaboration throughout these years. Our weekly meetings to discuss ideas, writing, and practical tips for navigating the academic and scientific research landscape have been invaluable. Additionally, I express deep thanks to my co-supervisor, Marie Eriksson, who adeptly guided me through the applied side of scientific research as a statistician and provided assistance during my job search towards the end of my PhD studies.

I am immensely thankful to my family for making this journey possible. The support of my mother, aunts, brothers, uncle, father, and cousins enabled me to pursue a master's degree in Sweden, laying the foundation for my subsequent PhD education.

I would like to acknowledge the pivotal role that everyone in the Department of Statistics at Umeå University played during this period. The department stands out as a special and inviting place, characterized by its friendly atmosphere and the presence of genuinely pleasant individuals. I extend this in special for Tetiana Gorbach, Johan Svensson, and Jessica Fahlén who worked alongside me in various courses, significantly contributing to my development as a university-level teacher. The camaraderie with fellow PhD students: Filip Edström, Joakim Wallmark, Mohammad Ghasempour (Amin), Kreske Ecker, Josline Otieno, Huixia Wang, and Niloofar Moosavi was indispensable, offering a shared experience during the PhD journey. Lastly, I appreciate the delightful conversations in the fika room with Anita Lindmark, Xijia Liu, and Anders Lundquist!

The last group of people to mention are all my friends who spent time with me traveling, visiting, being visited, having meals, talking on the phone and just generally being lovely, awesome people that supported me

through this entire period. The list of people is immense and spread in different cities and countries around the world: Mikaela Lillbäck, Youko Fujino, Georgios Rizothanasis, Emilia Lindblad, Natalia Fedorova, Sarah McIntyre, Jing-Jia Huang, Sanni Ranta, Henrik Karlsson, Milda Pocevičiūtė, Sara Johansson, Rafaela Stillner, Madeleine Lundgren, Natasha Guida, Frida Bylund, Atimmy Truong, Victor Attolini, Katarzyna Kozicka, Caetano Dieguez, Graziella Alves, and many others. I really hope I didn't forget anyone, but with so many it is impossible not to.

The city of Umeå itself has been an incredible experience, and living in the north was a pleasant surprise. I've learned a lot and had fun with a huge amount of activities that the city supports and enables: trying out a great variety of dancing styles, taking pictures of northern lights (like the cover of this thesis!), cross-country skiing, curling, working out at IKSU, visiting museums and art exhibitions, eating local food, going out for a run late in the evening during the summer or even to the beach in Norrmjöle or in Nydala! I appreciate the city, the university, and local groups and institutions for contributing to this enriching experience, and I look forward to continuing to enjoy my time here in Umeå in the future.

Umeå, February 2024
Guilherme Wang de Faria Barros

1 Introduction

This thesis focuses on the study of best practices and applications of hazard ratio estimation or survival analysis in a time-to-event setting, based on observational data. Observational studies involve the collection of data without imposing any treatments or interventions on the subjects, in contrast to experiments where treatments are assigned (Hernán and Robins, 2020).

In time-to-event studies, the researcher is often trying to understand the effect of a certain treatment or possible risk factors on the time from beginning of follow-up until an event occurs in individuals from a certain population. Time-to-event studies are most commonly found in epidemiology and medicine, but they are also used in a range of other scientific disciplines such as economics, the social sciences, the environmental sciences, and engineering (Kleinbaum and Klein, 2005).

Hazard ratios are some of the most commonly estimated parameters, usually relying on the Cox proportional hazards model. Hazard ratios represent the relative likelihood of an event happening in one group compared to another group over time. The two groups usually compared are one under the treatment regime and another untreated (control) group (Kleinbaum and Klein, 2005).

In scientific research the effects of treatments or interventions are traditionally studied using randomized controlled trials (RCTs), which provide a way to isolate the treatment effects on the time-to-event from other possible confounders (or confounding variables). In this case, a confounder is any variable that might distort or obscure the true relationship between treatment and outcome. However, RCTs have limitations related to the number of participants as well as the cost, generalizability, ethics, and duration of the study. For example, if we wanted to study the effects of smoking on the human body, it would be unethical to encourage study participants to smoke or to continue smoking (Hernán and Robins, 2020).

These problems can be circumvented by conducting an observational study. While such studies inevitably include confounding variables, they have become more common in recent years with advances in technology that facilitate the collection and storage of large amounts of data. There is thus growing interest in using observational data to emulate the conditions found in RCTs, for example, by removing confounding bias and establishing causal links between variables (Rosenbaum, 2017).

This thesis aims to explore a diverse range of problems that may arise when estimating hazard ratios from time-to-event data derived from observational studies. Paper I is an application of methods to estimate conditional hazard ratios of a recurrent stroke in the presence of competing risks of death using observational data from the Swedish Stroke Register (Riksstroke). Paper II examines non-parametric methods for balancing datasets and estimating the marginal hazard ratio, with an application regarding the use of anticoagulants in post-stroke recovery. Paper III and Paper IV are further extensions on marginal hazard ratio estimation considering more extreme scenarios, respectively, high censoring and high-dimensional data.

The thesis is organized as follows. Section 3 is an introduction to survival analysis. Section 4 is about hazard ratios and their role in survival analysis. Section 5 concerns the potential outcomes framework and balancing of datasets. Section 6 deals with the basis of simulating survival data with a prespecified marginal hazard ratio, a technique used in Papers II, III and IV. Section 7, summarizes the papers. Finally, Section 8 is a closing statement of the thesis, with some suggestions for further research on the topics covered in this thesis.

2 Stroke and the Swedish Stroke Register

Stroke is a major cause of death and disabilities worldwide. It is a complex disease and can be related to a wide range of risk factors, disease processes and mechanisms, and their combinations (Kuriakose and Xiao, 2020; O'Donnell et al., 2016). The two main types of stroke are ischemic stroke, caused by blockage of a blood vessel in the brain, and hemorrhagic stroke, caused by bleeding in the brain (Murphy and Werring, 2020).

The risk factors for stroke can be classified into two major groups: non-modifiable risk factor, such as age, gender, genetics, and modifiable risk factors, such as hypertension, diabetes, atrial fibrillation (AF), smoking, hyperlipidaemia, alcohol consumption and substance abuse, social economic status (SES), and obesity. Age is usually considered to be the most important risk factor, only about 10%-15% of all strokes happen in individuals below 55 years old (Murphy and Werring, 2020).

In Sweden, 72 hospitals provide acute stroke care and register stroke patients in the Swedish Stroke Register (Riksstroke). Riksstroke was

established in 1994 in order to help with monitoring and improving the quality of stroke care in Sweden. It is estimated that more than 90% of all strokes in Sweden are part of Riksstroke (Riksstroke, 2023).

Data collected by Riksstroke includes basic patient information such as age and sex, as well as the patient's cardiovascular risk factors such as AF, hypertension, diabetes and smoking, living conditions before the stroke, and dependency on others to carry out activities of daily living (ADL). Riksstroke also collects information on acute care and secondary prevention as well as follow-up data reported by the patients three months after stroke, including ADL dependency. Individual information on Riksstroke can be cross-referenced with other Swedish registers to include additional information such as date and cause of death from the Cause of Death Register (Eriksson et al., 2010; Lindmark, Glader, et al., 2014; Lindmark, Eriksson, and Darehed, 2022)

3 Survival analysis

Survival analysis, also known as time-to-event analysis or event history analysis, focuses on understanding the distribution of the time it takes for an event to happen. This type of analysis is particularly useful when we are interested in estimating the probability of an event occurring over a specified period. The main variable of interest in survival analysis is the survival time (or time-to-event). This is the time from the start of the study until the occurrence of the event or a predefined endpoint (Kleinbaum and Klein, 2005).

The event of interest could be anything that takes place over time, such as death, failure of a machine, finding a job, or recovery from a disease. Survival analysis is widely used in various fields including medicine, engineering, economics, and the social sciences.

Censoring is an important factor to be considered in survival analysis. An individual is considered to be censored when their survival time is not completely known. This can happen in two main forms: left censoring and right censoring.

In right censoring, the true survival time is not observed since the event does not occur before the end of the study. Left censoring occurs when the event of interest has already occurred for some subjects before the study began, and their event times are known only within a certain range

or interval. When studying a disease, for example, some individuals may already have developed the disease before the study and data collection began (Leung, Elashoff, and Afifi, 1997).

The relationship of the censoring mechanism to other variables and the time-to-event is another factor to be considered. Independent censoring occurs when the the survival time T and the censoring time C are independent. Non-informative censoring occurs when the probability of an individual being censored at a time t does not depend on that individuals's prognosis for failure at time t. Finally, when censoring time is dependent on the time-to-event, the censoring mechanism is referred to as informative censoring (Kleinbaum and Klein, 2005; Willems et al., 2018).

In survival analysis, one of the main goals is to estimate the survival function (or curve), denoted as $S(t)$. The survival function is the probability that an individual will survive beyond time t without experiencing the event. Given T as the time-to-event, a common estimator for the survival probability at a failure time $t_{(j)}$ is the Kaplan-Meier (KM) estimator (Kaplan and Meier, 1958):

$$\hat{S}(t_{(j)}) = \prod_{m=1}^{j} \hat{\Pr}(T > t_{(m)} | T \geq t_{(m)}). \tag{1}$$

The KM estimator is a non-parametric method that allows one to estimate the survival function from censored data and provides an estimate of the probability of surviving beyond each observed time point. It is particularly useful for visualizing survival trends and comparing different groups within a dataset. An example of Kaplan-Meier curves can can be seen in Figure 1.

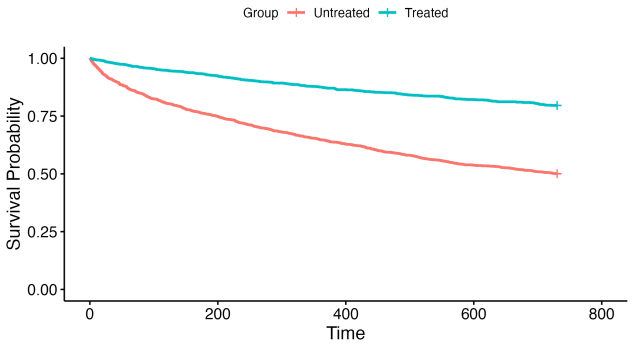


Figure 1: Example of Kaplan-Meier curves

4 Hazard ratios

Suppose for each individual $i = 1, \ldots, n$ we have p measured covariates $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$, a treatment status for each individual Z_i ($Z_i = 1$ if treated; $Z_i = 0$ if untreated), and T_i as the recorded follow-up time. T_i can be incomplete, or censored, when C_i, the censoring time, is lower than Y_i, the true time-to-event, that is, $T_i = \min\{Y_i, C_i\}$. In this case, $D_i = 1_{Y_i \leq C_i}$ is the event indicator. Throughout this thesis, we will assume that censoring is right censoring and non-informative, common assumptions in time-to-event studies. If there are no competing risks, the hazard function is defined as (Kleinbaum and Klein, 2005):

$$h(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2}$$

The hazard function is a function of time and describes the instantaneous rate of occurrence of the event of interest among the individuals still at risk. For example, if the event of interest is a stroke, the hazard function describes the instantaneous probability of one of the individuals at risk suffering a stroke.

The Cox proportional hazards model (a semi-parametric regression model) (Cox, 1972) is a common way to estimate hazard ratios and relate them to covariates and a treatment variable. It can be written as:

$$h(t|\mathbf{X}_i, Z_i) = h_0(t) e^{(\alpha_Z Z_i + \sum_{k=1}^{p} \mathbf{X}_{ik} \alpha_k)}, \tag{3}$$

where $h_0(t)$ is the baseline hazard function, e^{α_Z} is the hazard ratio be-

tween $Z_i = 1$ and $Z_i = 0$ and e^{α_k} are the hazard ratios for each of the p covariates respectively, given that all other variables are held constant. As this model depends on variables being held constant, the hazard ratios being estimated are conditional hazard ratios (CHRs).

However, in some cases we might be interested in estimating the marginal hazard ratio (MHR), that is, the effect when no other covariates are held constant. With an ideal RCT, this too can be accomplished using the Cox proportional hazards model, with the treatment as the only regressor (Austin, 2014):

$$h(t|Z_i) = h_0(t)e^{(\alpha_Z Z_i)}. \tag{4}$$

This is possible with an ideal RCT since both the treated and untreated groups would be exactly the same in all possible covariates, differing only in treatment status and outcome. Thus, there is no source of confounding between outcome and treatment, and the MHR can be directly estimated (Austin and Stuart, 2015a).

However, if a subject can experience any of a set of different events, this implies the presence of competing risks. The hazard functions of interest in that case are the cause-specific hazard function and the subdistribution hazard function. The cause-specific hazard function is defined by (Pintilie, 2006):

$$h^{cs}(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \leq T < t + \Delta t, D = k | T \geq t)}{\Delta t}, \tag{5}$$

where D denotes the type of event that occurred. In comparison, the subdistribution hazard function is defined by:

$$h^{sub}(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}\left((t \leq T < t + \Delta t, D = k | T \geq t) \text{ or } (T < t \text{ and } D \neq k)\right)}{\Delta t}. \tag{6}$$

Both cause-specific and subdistribution hazard ratios can be estimated by modified Cox regression models. In general, when competing risks are present, cause-specific models are more common for researching etiologic questions, while subdistribution models are more suited to estimating incidence or predicting prognosis. However, when publishing time-to-event studies involving competing risks, the general recommendation is to report results for both cause-specific and subdistribution models (Austin, Lee, and Fine, 2016).

5 Potential outcomes and balancing

If the data being used for analysis is observational and does not come from an RCT, estimation of the MHR will be biased due to confounding. However, it is still possible to achieve an unbiased estimation of the MHR by using balancing methods on the dataset. This procedure is based on the similar problem of estimating treatment effects in the potential outcomes framework (Austin, 2014).

In the potential outcomes framework, all individuals in the sample have a pair of potential outcomes $Y_i(0)$ and $Y_i(1)$. This pair of outcomes are, respectively, in the control regime ($Z_i = 0$) and outcome under treatment ($Z_i = 1$). For each individual, only one outcome is observed: $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$ (Hernán and Robins, 2020). In this framework, a commonly estimated parameter is the average treatment effect (ATE), defined as $\text{ATE} = \text{E}[Y_i(1) - Y_i(0)]$. An important assumption for ATE estimation in an observational study is that of unconfoundedness (or ignorability). This assumption is represented by:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp Z_i | \mathbf{X}_i. \tag{7}$$

We also define any covariate X that is part of the minimally sufficient adjustment set for this unconfoundedness definition as a *confounder* (VanderWeele and Shpitser, 2013). A common example of a confounder is a covariate that affects both the treatment and/or assignment of the treatment and the outcome itself. These relationships can be represented by graphs (Pearl, 2009) and a simple case can be seen in the example in Figure 2.

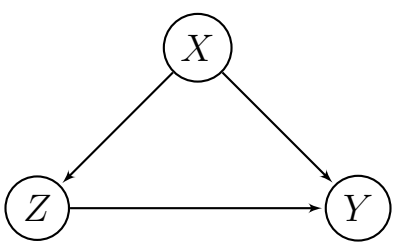


Figure 2: Example graph. X is a confounder when trying to estimate the effect of Z (treatment) on Y (outcome)

Although we only observe one of the potential outcomes for each indi-

vidual, in RCTs it is easy to estimate the ATE at a population level, since both the treated and the untreated groups are assumed to be equivalent, in which case confounding bias is not present. In practice, we can estimate how similar the treated and untreated groups are by using balancing measures such as the absolute standardized mean difference (ASMD) (Z. Zhang et al., 2019). For a covariate X, the ASMD is defined as:

$$\text{ASMD} = \frac{|\bar{X}_{\text{treated}} - \bar{X}_{\text{untreated}}|}{\sqrt{(s^2_{\text{treated}} + s^2_{\text{untreated}})/2}}, \tag{8}$$

where $\bar{X}_{\text{treated}}$ and $\bar{X}_{\text{untreated}}$ are the sample means of X in treated and untreated subjects, respectively, and s^2_{treated} and $s^2_{\text{untreated}}$ the analogous sample variances. Higher values of ASMD imply larger differences between the two groups. Usually an ASMD value of 0.25 in a covariate is considered balanced, although stricter thresholds, for example, 0.10, have also been suggested (Austin, 2009). It is common to plot ASMD values to better grasp how unbalanced a dataset is, as in Figure 3

In observational studies, this unbalancing of the covariates between groups is a source of bias due to confounding. In order to account for the bias induced by confounding, we first define the concept of propensity score in order to estimate the ATE. The propensity score is the probability of an individual receiving the treatment, given its observed covariates: $\text{Prob}(Z_i = 1|\mathbf{X}_i)$. Using the propensity score, there are two main methods that enable a researcher to control for confounders, so mimicking the conditions of a RCT and obtaining a balanced dataset: propensity score matching (PSM) and inverse probability of treatment weighting (IPTW). Concisely, PSM pairs individuals in the dataset with similar propensity scores while discarding the rest. IPTW, on the other hand, assigns weights to each individual. Hence, both methods produce a balanced dataset that is derived from the original one. The weights used for ATE estimation with IPTW can be estimated by:

$$w_i^{PS} = \frac{Z_i}{\text{Prob}(Z_i = 1|\mathbf{X}_i)} + \frac{1 - Z_i}{1 - \text{Prob}(Z_i = 1|\mathbf{X}_i)}, \tag{9}$$

where w_i^{PS} is the weight assigned to individual i.

In the time-to-event setting, both these methods can also be used to produce balanced datasets. The next step is to then use this balanced dataset in a Cox regression (equation 4). In a corresponding manner

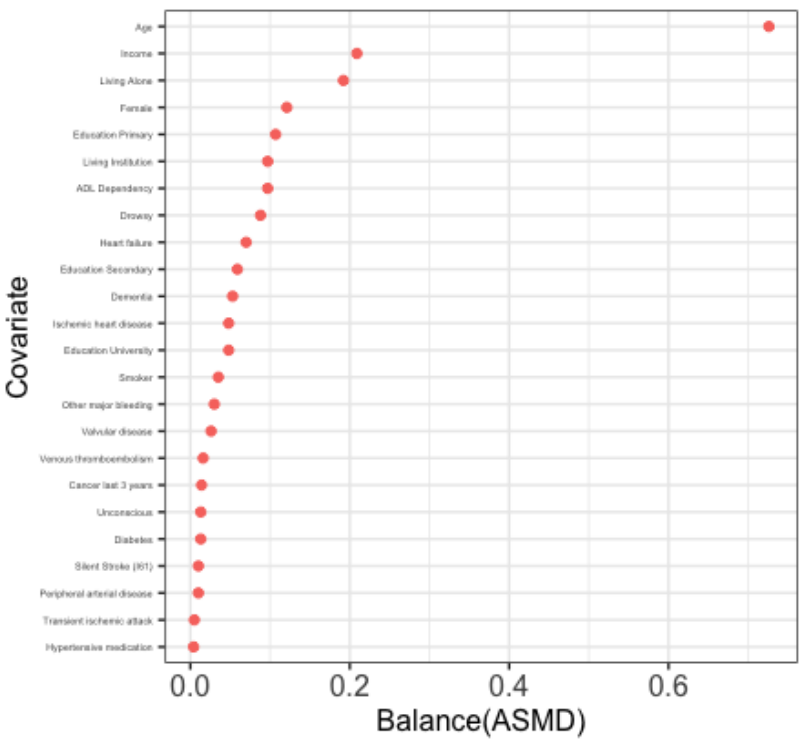


Figure 3: Example of ASMD plot

to ATE estimation and to RCTs, without any censoring involved, this produces an unbiased estimator of the MHR (Austin, 2014; Austin and Stuart, 2015b; Austin, 2013).

To estimate the propensity score, the most common method is the use of a logistic regression:

$$\Pr(Z_i = 1|\mathbf{X}_i) = \frac{e^{\beta_0+\beta_1 X_{i1}+\ldots\beta_p X_{ip}}}{1 + e^{\beta_0+\beta_1 X_{i1}+\ldots\beta_p X_{ip}}}. \tag{10}$$

In this parametric case, the selection of which covariates and their transformations should be accounted for during the logistic regression is an important factor. Model misspecification is known to cause bias during the MHR estimation process. A commonly used method for variable selection in association with logistic regression is ℓ_1 regularization, also known as lasso regression (Hastie, Tibshirani, and Friedman, 2001). This is done by constraining the parameters of the logistic regression by:

$$\sum_{k=1}^{p} |\beta_k| \leq \lambda \tag{11}$$

where $\lambda > 0$ is a tuning parameter that controls the regularization. λ can be selected in a number of ways. Two possible alternatives are cross validation (Hastie, Tibshirani, and Friedman, 2001) or, in the case of balancing, directly targeting the ASMD of the covariates (Fowler et al., 2017).

In addition to parametric methods, non-parametric methods can also be used to generate weights. Non-parametric methods such as stable balancing weights (Zubizarreta, 2015), entropy balancing (Hainmueller, 2012), and empirical balancing calibration weights (Chan, Yam, and Z. Zhang, 2016), directly target the balancing of the covariates and have been shown to be important alternatives for generating weights in causal inference.

It is important to note that, unlike the ATE, the MHR cannot be considered to be a causal parameter or causal treatment effect (Aalen, Cook, and Røysland, 2015; Hernán, 2010; Martinussen, Vansteelandt, and Andersen, 2020). Even in an idealized RCT, the MHR concerns an average between two populations, a treated population and an untreated population, across time. However, it is not possible to guarantee that the MHR will not change over time or that these two groups are comparable across the entire period.

6 Simulation settings

Three of the papers in this thesis include simulation studies. This section is a summary of the procedure to generate observational time-to-event data with a prespecified MHR. The basic simulation scenario is based on previous work by Setoguchi et al. (2008), Austin (2013) and Austin and Stuart (2015b).

The first step is to sample ten independent covariates $\mathbf{X} = X_1, \ldots, X_{10}$. In this case, X_1, X_3, X_5, X_6, X_8, and X_9 are Bernoulli($p = 0.5$) distributed, while X_2, X_4, X_7 and X_{10} are sampled according to a standard normal distribution. A true propensity score is then created for every individual i in the dataset:

$$\text{logit}(\Pr(Z_i = 1|\mathbf{X}_i)) = \boldsymbol{\zeta}^T \mathbf{X}_i, \tag{12}$$

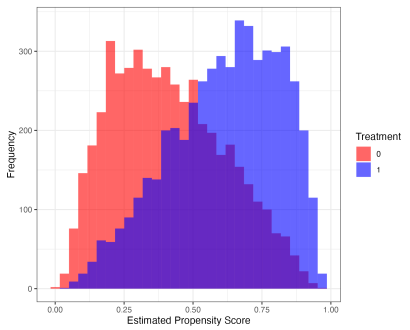
where $\boldsymbol{\zeta} = (0.8, -0.25, 0.6, -0.4, -0.8, -0.5, 0.7, 0, 0, 0)^T$. Using these true PS values, we sample a treatment status variable,Z_i, for each individual. Thus we can generate a linear predictor (LP):

$$LP_i = \alpha_z^* Z_i + \boldsymbol{\beta}^T \mathbf{X}_i, \tag{13}$$

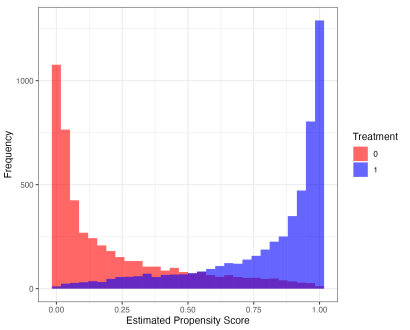
$\boldsymbol{\beta}^T$ is set as $k \times (0.3, -0.36, -0.73, -0.2, 0, 0, 0, 0.71, -0.19, 0.26)^T$ and α^* is a log conditional hazard ratio of the treatment. It should be noted that k can be used to control the degree of overlap between the treated and untreated groups. The base case of $k = 1$ represents strong overlap between the two groups. Increasing values of k result in weaker overlap. To visualize how k affects the overlap, see Figure.4: Finally, we can assign a time-to-event Y_i:

$$Y_i = \left(\frac{-\log(u_i)}{\gamma^* e^{LP_i}} \right)^{1/\eta}, \tag{14}$$

where u_i is sampled from a standard uniform distribution and η and γ^* are set tp 2 and 0.00002, respectively. In order to generate a specific MHR (α_z), we must then choose a value for α_z^* that corresponds to the desired MHR. This is done through aniterative bissection process that involves simulating both potential outcomes (for $Z_i = 1$ and $Z_i = 0$), checking the corresponding MHR associated with α_z^* and then adjusting as necessary, as suggested by Austin (2013).



(a) Overlap of a generated dataset when k = 1



(b) Overlap of a generated dataset when k = 3

Figure 4: Overlap example dataset with k = 1 and k = 3

In order to account for non-informative censoring, we can sample censoring times C_i values from other distributions. Two common distributions when considering censoring are the uniform distributions and the Weibull distribution. We consider the cases where $C_i \sim \text{uniform}(0, \theta)$ or $C_i \sim \text{Weibull}(\eta, \theta)$. To achieve a specific censoring rate, the value of θ can be chosen by solving the integral:

$$\gamma(\theta|\pi) = \int_{D'} \text{Prob}(\omega = 1|u, \theta) f_{\tau_i}(u) du - \pi, \tag{15}$$

where π is the desired censoring proportion, D' is the domain of $\tau_i = \frac{\exp(LP_i/\eta)}{\sqrt{\gamma}}$, $\omega = 1_{\{Y \geq C\}}$ is an indicator variable for censoring and $f_{\tau_i}(\cdot)$ is the density function of τ_i. If C_i is distributed according to uniform$(0, \theta)$, we have:

$$\Pr(\omega = 1|\tau_i, \eta, \theta) = \frac{\tau_i}{\eta\theta} \Gamma\left(\frac{1}{\eta}, (\theta/\tau_i)^\eta\right). \tag{16}$$

And if $C_i \sim \text{Weibull}(\eta, \theta)$:

$$\Pr(\omega = 1|\tau_i, \eta, \theta) = \frac{1}{1 + (\theta/\tau_i)^\eta}, \tag{17}$$

where $\Gamma(\cdot, \cdot)$ is the lower incomplete gamma function. It is not possible to explicitly find $f_{\tau_i}(u)$, when $\mathbf{X}$ is a mix of normal and Bernoulli distributions, as in our case. However, as recommended by Wan (2017), it can

be estimated with kernel methods, finding a θ value that results in the desired censoring proportion.

This scenario is further modified in Paper II to account for model misspecification and overlap, in Paper III to create an analogous RCT to the observational study, and in Paper IV to consider a high-dimensional case. More details can be seen in the respective parts of the thesis.

7 Summary of papers

7.1 Paper I

In Paper I, data from the Swedish Stroke Register (Risksstroke) is used to study the long-term prognosis of individuals with acute ischemic stroke in the absence of standard modifiable stroke risk factors (SMoRFs). Individuals were considered to possess a SMoRF if they had one of hypertension, diabetes, hyperlipidemia, atrial fibrillation, or an active smoking history. One in five individuals who had suffered an ischemic stroke had no SMoRFs, but little is known about the long-term prognosis of these individuals.

In total, 152,588 total individuals who suffered an ischemic stroke in Sweden between 2010 and 2020 were considered in the study. The differences between individuals with and without SMoRFs were compared using several methods, including relative risks, odds ratios, cumulative incidence curves, and hazard ratios, considering the competing risks of a recurrent stroke and death.

From this study we concluded that patients without SMoRFs have a lower risk of short and long-term mortality than patients with one or more SMoRFs. This was the largest and most comprehensive study of first-presentation ischemic stroke in patients without risk factors.

7.2 Paper II

Paper II explored the estimation of marginal hazard ratios (MHRs). Specifically, Paper II focuses on the differences between modeling and balancing approaches. Modeling approaches such as inverse propensity score weighting often rely on the correct specification of a parametric model without targeting either balance or stability. On the other hand, balancing approach methods are usually non-parametric and target co-

variate imbalances directly, allowing the researcher to explicitly set the desired balance constraints.

The finite sample properties of different modeling and balancing approach methods were evaluated by estimating the marginal hazard ratio using Monte Carlo simulations. The methods examined were logistic regression, lasso, non-parametric covariate balancing propensity score, calibration and entropy balancing, and stable balancing weights. The Monte Carlo simulations evaluated scenarios with misspecification of the model, censoring, and overlap. The methods were also illustrated by analyzing data from the Swedish Stroke Register in order to estimate the effect of anticoagulants on time to recurrent stroke or death in stroke patients with atrial fibrillation at patient discharge.

In this paper we find that in simulated scenarios the balancing approach methods with good overlap and low or no model misspecification performed similarly to the modeling approach methods. In scenarios involving bad overlap and model misspecification, the modeling approach method incorporating variable selection performed better than the other methods. This indicates that methods that target covariate balance for estimating MHRs are a valuable alternative. However, good performance is not guaranteed in situations with, for example, poor overlap, high censoring, or misspecified models/balance constraints.

7.3 Paper III

Paper III continues the study of estimating the MHR, focusing on the effects of non-informative censoring. For this goal, a Monte Carlo simulation was set up with varying degrees of censoring.

The Monte Carlo simulation was adapted to consider scenarios with varying degrees of censoring, two different censoring mechanism distributions (uniform and Weibull), and different values for the true MHR. The process of estimating the MHR was also considered using weighting and matching and involving observational or experimental data.

In this paper, we show that the estimation of the MHR is biased by non-informative censoring, even under perfect randomization. This bias increases as the rate of censoring is larger and the effect of the treatment is stronger. When there is a treatment effect, the estimation is also unbiased.

A procedure to minimize this censoring is also suggested. This proce-

dure is successful at low and medium rates of censoring, but it is not able to eliminate bias at the highest rates of censoring.

7.4 Paper IV

Paper IV is a further extension on MHR estimation. In this paper we explored the best practices in estimating the MHR using high-dimensional data. A Monte Carlo simulation was set up and a case study involving Stroke Register data was also performed to showcase the methods studied.

The methods explored are well known in causal inference literature, but have never been explored in the time-to-event case. Covariate selection was performed with lasso to find sets of covariates related to the treatment assignment and to the outcome. These groups were used in a post-lasso regression to create weights used for the MHR estimation. The union and intersection of these groups were also considered.

In addition, a multiply robust approach combining all the sets and their combinations was studied as a way to improve the MHR estimation. This inclusion of a multiply robust approach was shown to be successful in improving performance, especially in the most complex scenarios.

8 Final remarks and further research

This thesis set out to explore some of the challenges that may arise when conducting time-to-event studies based on observational data. Competing risks, censoring, misspecification of models, and covariate selection were explored in the papers that constitute this thesis.

Among these, the most important problem that remains unsolved is related to censoring when estimating the MHR. The solution explored in Paper III is only partial and was not able to solve the bias in the estimation in scenarios with high censoring. This problem needs further study as censoring is one of the most important components of time-to-event studies. A possible solution might be the use of pseudo-observations, which have been used previously to solve problems related to censoring in survival analysis (Andersen and Pohar Perme, 2010; Andersen, Syriopoulou, and Parner, 2017).

Paper IV presents recommendations for covariate selection. It should be noted that the experiments conducted in this case involved linear scenarios of data generation, and therefore lasso regression was an appropri-

ate method. In the future, it would be interesting to see the performance of more advanced machine learning methods in selecting covariates for both treatment assignment and outcome, especially in non-linear scenarios (Spooner et al., 2020).

Finally, one of the major disadvantages of the marginal hazard ratio is that it cannot be interpreted as a causal parameter (Hernán, 2010; Martinussen, Vansteelandt, and Andersen, 2020; Aalen, Cook, and Røysland, 2015). There have been advances in relation to investigating and estimating a causal HR (Axelrod and Nevo, 2023), but this is an estimator that is quite new and still needs development, especially regarding the cases involving observational data.

Ethical considerations

Statistical method development for fair comparison of stroke care and outcome was part of the EqualStroke-project, approved by the Ethical Review Board in Umeå (Dnr: 2012-321-31M, 2014-76-32M). Patients and next of kin were informed about the registration and aim of the Riksstroke register and their right to decline participation (opt-out consent).

References

Aalen, O. O., R. J. Cook, and K. Røysland. Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis* 21 (2015).

Andersen, P. K. and M. Pohar Perme. Pseudo-observations in survival analysis. *Statistical methods in medical research* 19.1 (2010), 71–99.

Andersen, P. K., E. Syriopoulou, and E. T. Parner. Causal inference in survival analysis using pseudo-observations. *Statistics in Medicine* 36.17 (2017), 2669–2681.

Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* 28.25 (2009), 3083–3107.

Austin, P. C. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine* 32 (2013).

Austin, P. C. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine* 33 (2014), 1242–1258.

Austin, P. C., D. S. Lee, and J. P. Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 133.6 (2016), 601–609.

Austin, P. C. and E. Stuart. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* 28 (2015).

Austin, P. C. and E. Stuart. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research* 26 (2015), 1654–1670.

Axelrod, R. and D. Nevo. A sensitivity analysis approach for the causal hazard ratio in randomized and observational studies. *Biometrics* 79.3 (2023), 2743–2756.

Chan, G., P. Yam, and Z. Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78.3 (2016), 673–700.

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

Eriksson, M., F. Jonsson, P. Appelros, K. H. Åsberg, B. Norrving, B. Stegmayr, A. Terént, and K. Asplund. Dissemination of thrombolysis for acute ischemic stroke across a nation: experiences from the Swedish stroke register, 2003 to 2008. *Stroke* 41.6 (2010), 1115–1122.

Fowler, P., X. de Luna, P. Johansson, P. Ornstein, S. Bill, and P. Bengtsson. Study protocol for the evaluation of a vocational rehabilitation. *Observational Studies* 3.1 (2017), 1–27.

Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20.1 (2012), 25–46.

Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

Hernán, M. A. The Hazards of Hazard Ratios. *Epidemiology* 13 (2010).

Hernán, M. A. and J. M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020. ISBN: 1420076167.

Kaplan, Edward L and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53.282 (1958), 457–481.

Kleinbaum, D. G. and M. Klein. *Survival Analysis: A Self-Learning Text*. New: Springer Science and Business Media, LLC, 2005.

Kuriakose, D. and Z. Xiao. Pathophysiology and treatment of stroke: present status and future perspectives. *International Journal of Molecular Sciences* 21.20 (2020), 7609.

Leung, K., R.M. Elashoff, and A. A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health* 18 (1997), 83–104.

Lindmark, A., M. Eriksson, and D. Darehed. Socioeconomic status and stroke severity: Understanding indirect effects via risk factors and stroke prevention using innovative statistical methods for mediation analysis. *PLOS One* 17.6 (2022), e0270533.

Lindmark, A., E. Glader, K. Asplund, B. Norrving, M. Eriksson, and Riks-StrokeCollaboration. Socioeconomic disparities in stroke case fatality–Observations from Riks-Stroke, the Swedish stroke register. *International Journal of Stroke* 9.4 (2014), 429–436.

Martinussen, T., S. Vansteelandt, and P. K. Andersen. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis* 26 (2020).

Murphy, S. J. X. and D. J. Werring. Stroke: causes and clinical features. *Medicine* 48.9 (2020), 561–566.

O'Donnell, M. J., S. L. Chin, S. Rangarajan, D. Xavier, L. Liu, H. Zhang, P. Rao-Melacini, X. Zhang, P. Pais, S. Agapay, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *The Lancet* 388.10046 (2016), 761–775.

Pearl, Judea. *Causality*. Cambridge university press, 2009.

Pintilie, M. *Competing Risks: A Practical Perspective*. John Wiley & Sons, 2006.

Riksstroke. *Riksstroke - General Information*. 2023. URL: `https://www.riksstroke.org/general-information/` (visited on 11/22/2023).

Rosenbaum, Paul R. *Observation and Experiment: An Introduction to Causal Inference*. USA: Harvard University Press, 2017. ISBN: 9780674975576.

Setoguchi, S., S. Schneeweiss, M. Brookhart, R. Glynn, and E. Cook. Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety* 17 (2008), 546–55.

Spooner, A., E. Chen, A. Sowmya, P. Sachdev, N. A. Kochan, J. Trollor, and H. Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports* 10.1 (2020), 20410.

VanderWeele, Tyler J and Ilya Shpitser. On the definition of a confounder. *Annals of Statistics* 41.1 (2013), 196.

Wan, F. Simulating survival data with predefined censoring rates for proportional hazards models. *Statistics in Medicine* 36 (2017), 838–854.

Willems, S. J. W., A. Schat, M. S. van Noorden, and M. Fiocco. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical Methods in Medical Research* 27.2 (2018), 323–335.

Zhang, Z., J. H. Kim, G. Lonjon, Y. Zhu, et al. Balance diagnostics after propensity score matching. *Annals of Translational Medicine* 7.1 (2019).

Zubizarreta, J. R. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* 110.511 (2015), 910–922.