Association of Cigarette Smoking with the Relapse of Anti-neutrophil Cytoplasmic Antibody Associated Vasculitis: a survival analysis.

Bo Peng

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Intelligent Systems)

Supervisor: James Ng

August 2022

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Bo Peng

August 19, 2022

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Bo Peng

August 19, 2022

Association of Cigarette Smoking with the Relapse of Anti-neutrophil Cytoplasmic Antibody Associated Vasculitis: a survival analysis.

Bo Peng, Master of Science in Computer Science

University of Dublin, Trinity College, 2022

Supervisor: James Ng

Anti-neutrophil cytoplasmic antibodies (ANCA)-associated vasculitis (AAVs) is a group of diseases involving severe, systemic, small-vessel vasculitis. The disease itself is extremely dangerous and is associated with a high rate of recurrence. This paper therefore examines the available research on ANCA-AVV recurrence and identifies smoking as a potential influence on ANCA-AAV recurrence. This paper examines whether smoking affects relapse through survival analysis including Kaplan-Meier curve, cox regression, accelerated failure time model. It was concluded that smoking was a significant factor in the relapse of ANCA-AVV patients.

Acknowledgments

First of all, I would like to thank my parents for their support in giving me the valuable opportunity to study at TCD and to experience the wider world first-hand.

I would then like to thank my supervisor, Prof. James Ng. It would have been difficult to complete the research for this thesis without his patient and meticulous teaching. I would like to thank him for his guidance and recognition during my research process. Also, thanks to Dr. Susan Connolly for her guidance and advice.

In addition, I would like to thank the data providers for this thesis, without whom the experiments would not have been supported by the data.

Finally, I would like to thank my friend Jianzheng Li, whose encouragement kept me going when I was down and out.

Bo Peng

University of Dublin, Trinity College
August 2022

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

The anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitides (AAVs) are a group of disorders involving severe, systemic, small-vessel vasculitis(AR et al. (2020)) which includes granulomatosis with polyangiitis (GPA, Wegener's granulomatosis), microscopic polyangiitis (MPA), eosinophilic granulomatosis with polyangiitis (EGPA, Churg-Strauss syndrome) and renal-limited vasculitis (RLV). Each has unique clinical characteristics including the lungs, kidneys, ear, nose, and throat (ENT), skin, heart, and central and peripheral neurological systems.(Mercuzot and Simon Letertre (2021)).Not just due to the disease or infection, but also due to comorbidities, especially cardiovascular disease, patients with AAV continue to have a mortality rate that is 2.7 times that of the general population. Comorbidities are additional disabling diseases, especially cardiovascular disease, that co-occur or co-exist with AAV. Comorbidities can influence morbidity and mortality, decrease quality of life (QoL), and raise the complexity and cost of AAV care. Having comorbidities at the time of diagnosis increases mortality, and comorbidities might impact a patient's quality of life even when remission is established. This has become an important finding in AAV patient research(Holle and Gross (2013)).

Anti-neutrophil cytoplasmic antibody (ANCA) -associated vasculitis (AAV) is a relatively rare autoimmune disease that affects about 20 people per million, and the prevalence is now about 200 per million as current improvements in treatment can extend the life expectancy of those affected(Karangizi and Harper (2021)). The disease can occur at any age, and even young children can be at risk, but it is most common in older people (peaking at 55-70 years of age) and is equally prevalent in men and women (Karangizi and Harper (2021)). They are uncommon but severe diseases with a variable clinical and serological presentation across geographic regions. (Paramalingam et al. (2019)). Prompt identification and treatment are crucial to limit the incidence of permanent scars and fatalities caused by vasculitis. Due to the disease's various, non-specific manifestations, there may be considerable diagnostic delays(Karangizi and Harper (2021)).

The treatment of AAV is based on the concept of a combination remission induction and maintenance strategy and is stage- and activity-specific (Holle and Gross (2013)). The current treatment plan for AAV aims to maximise the rate of induction and maintenance of remission while reducing the accumulation of irreversible damage and side effects(Jones and Hiemstra (2019)). Standard treatment includes induction of remission with high-dose glucocorticoids and high-dose oral or intravenous cyclophosphamide for 3 to 6 months, followed by maintenance of remission with azathioprine or methotrexate while the glucocorticoids are gradually lowered and removed. The optimum duration of maintenance treatment is not known and practice differs widely between centres. The optimum duration of maintenance treatment is not known and practice differs widely between centres(Berden and Göçeroğlu (2012). The figure 1.1 shows the ANCA-associated vasculitis and treatment procedure.

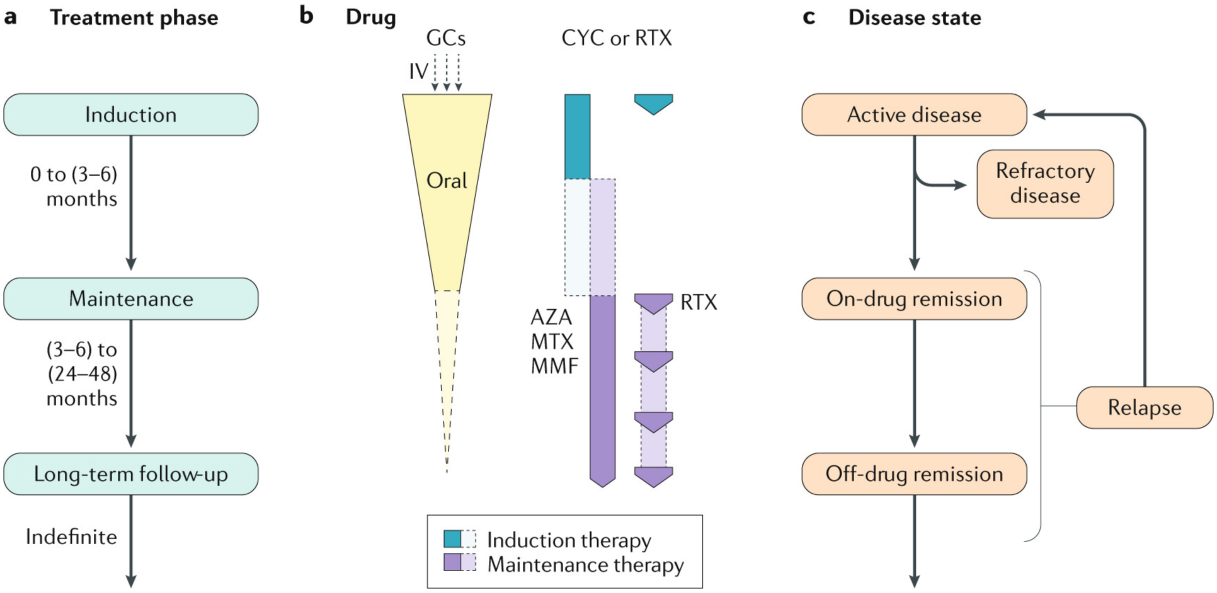


Figure 1.1: ANCA-associated vasculitis and treatment procedure(A and Anders (2020))

Modern treatments have turned ANCA-related vasculitis from a life-threatening disorder to a chronic condition prone to relapse throughout a patient's lifespan(Berden and Göçeroğlu (2012)). Using immunosuppression has increased the 5-year survival rate to 80%. After treatment, these diseases tend to follow a chronic course of relapse-remission, with 50% of patients relapsing within five years of diagnosis(Karangizi and Harper (2021)). A 5-year follow-up of 107 patients in a major observational research revealed that approximately 50 percent of those treated would have experienced one or more relapses(Berden and Göçeroğlu (2012)). Clinical relapses are frequent in anti-neutrophil cytoplasm antibody (ANCA)–associated vasculitis, demanding repeated immunosuppressive treatments and elevating the risk of serious adverse effects(Salama (2019)). Certain susceptibility factors for relapse, such as the clinical features of protease-3-ANCA and granulomatous disease with polyangiitis, have been identified. However, little progress has been made in elucidating the pathophysiological factors for relapse and why they vary significantly between patients with different ANCA subtypes. Because clinically and immunologically clear disease relapses are evident, modest immune disease activity is frequently ignored(Salama (2019)). Modern induction regimens are generally quite effective at inducing illness remission. However, maintaining remission depends on the drug employed and the maintenance regimen to which patients are shifted. This suggests that specific medications may modulate certain components of the immune response or that they may do so to varying degrees. Various cohort studies and long-term follow-up of international trials have indicated that relapse rates at 5 years range from 21% to 89%, depending on the induction and maintenance regimens used(McDermott and Xiaoqing Fu (2020)).

A multivariate Cox proportional hazards model indicated an association between current smoking and flare. A correlation was also observed between relapse risk and cumulative pack-years of smoking. Cigarette smoking is a major, dose-related risk factor for ANCA-AAV flare activities. Cessation of smoking should be acknowledged as a viable therapy option for ANCA-AAV patients(Yamaguchi and Ando (2018)). In a large case-control study of patients with AAV and matched controls from the same health care system, smoking history was strongly associated with an elevated risk of getting AAV. In addition, a dose-response relationship was discovered, in which the cumulative amount of smoking increased the likelihood of developing AAV. This connection was exceptionally robust in MPO-ANCA-positive individuals and remained in subgroups of patients with renal, head/neck, and lung disease. Multiple sensitivity analyses confirm the findings to be reliable(McDermott and Xiaoqing Fu (2020)).

1.2 Motivation of this research

Although modern treatments have transformed ANCA-associated vasculitis from a life-threatening disease into a chronic condition prone to recurrence throughout a patient's lifetime (Berden and Göçeroğlu (2012)), this has allowed patients to survive longer. Immunosuppressive drugs have increased the 5-year survival rate to 80%. However, after treatment, these diseases tend to follow a chronic course of relapse-remission, with 50% of patients relapsing within five years of diagnosis (Karangizi and Harper (2021)). This recurrence has a significant negative impact on the long-term survival of patients with

ANCA-AAV, and if research could be done to identify factors that influence the recurrence of ANCA-associated vasculitis so that earlier interventions could be made, this could lead to a significant improvement in survival, a reduction in the likelihood of subsequent suffering and an improvement in the quality of life of patients with the disease.

From the above background, a summary can be found that the Anti-neutrophil cytoplasm antibody (ANCA) associated vasculits (AAV) disease itself is extremely organ and life threatening and has a high recurrence rate. The recurrence of the disease is not only influenced by the treatment, but also by the patient's own state and the surrounding environment. Therefore, the risk and suffering of anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis relapse can be minimised if it is prevented in advance.

After collating and summarising the existing research in this area, there is currently very little information in the field about whether smoking induces relapse in Anti-neutrophil cytoplasm antibody (ANCA) associated vasculits patients. It has been shown that there is an association between smoking and the development of ANCA-AAV, and it has also been suggested that smoking may be associated with the recurrence of ANCA-AAV. Because of the high risk of ANCA-AAV and the high recurrence rate, it would be useful to analyse the association between smoking and ANCA-AAV recurrence through survival analysis and, if there is an association, to advise patients to stop smoking or not to smoke, thereby reducing the recurrence rate and increasing the quality of life of the patient.

As a widely used statistical method in disease research, survival analysis has produced many good results in disease research. However, the number of studies using survival analysis as a research method in the area of ANCA-AAV is currently low. Therefore this is a direction and topic worth exploring. This study therefore seeks to fill a gap in that research by exploring whether smoking was an influential factor in the relapse of ANCA-AAV patients through survival analysis, and adds examples of survival analysis in this field of study, aiming to identify the association between smoking and ANCA-AAV recurrence and to give patients preventive advice to reduce the recurrence rate and enhance their quality of life.

1.3 Dissertation Overview

The paper is structured as follows.

- Chapter 2 reviews the current literature and summarises current research in the field related to factors associated vasculitis recurrence and survival analysis models in relation to disease recurrence as well as disease related aspects, including some comparisons of Cox regression models and Accelerated Failure Time models.

- Chapter 3 presents the datasets used in this study, including the source of this data and a privacy and ethics statement for the data in this research. As well as a description of the data and missing statistics. An introduction to the unique type data associated with survival analysis is also introduced.

- Chapter 4 presents the survival analysis methods, the various models and the associated methodology for the Proportional hazard Assumption test used in the study.

- Chapter 5 describes the implementation of the various models for survival analysis including the Kaplan-Meier, cox regression model and the Accelerated Failure Time model and the analysis of the model results. A comparative analysis of the model results was also carried out

- Chapter 6 reflects on the findings of the study and discusses future research directions.

Chapter 2

Literature Review

This chapter reviews the research findings related to the research questions of this thesis. The first part presents the relevant research on ANCA-Associated Vasculitis regarding recurrence factors. The second part presents research on survival analysis in relation to the disease, particularly in relation to factors influencing disease recurrence.

2.1 Related Work on Relapse of ANCA-Associated Vasculitis

Although there is no shortage of research on the study of AAV recurrence, there are often inconsistencies in the results that people come up with. In Catherine King's study, the authors note that despite the existence of a number of randomised controlled trials (RCTs) and observational studies examining clinical, histological, and biochemical predictors of AAV recurrence, the risk variables revealed by these studies vary.(King and Druce (2021))

No quantitative meta-analysis of these risk factors has been undertaken to far. In addition, this data has not been practically translated into individualised therapy regimens based on an individual's risk of recurrence. Using a systematic review and meta-analysis, the authors identified risk variables for AAV recurrence. And a model was constructed to estimate the risk of recurrence. The authors specified a literature search strategy to restrict studies to identify and quantify independent predictors of AAV recurrence through further review and full text review of eligible studies using multivariate analysis or randomised controlled trials with recurrence or sustained remission as the endpoint(King and Druce (2021)).

Non-compliant studies were excluded following data feature extraction for those that met the requirements, sensitivity analysis of the data, assessment of quality using the QUPS tool and discussion by all authors. A meta-analysis of the screened studies was performed to identify available risk factor data, and Cox regression analysis was applied to test the possible score combinations of combined risk factors.The final conclusion was that Anti-PR3 positive, a lower serum creatinine level, and CVS involvement at the time of diagnosis are all associated with an increased A combination of these risk variables can be utilised to predict an individualised risk of relapse(King and Druce (2021)).

In Neil Basu's paper, Neil Basu points out that the income of MPA and GPA patients has shifted in recent decades, although premature death is still evident compared to the general population. Basu explains that the change in mortality from vasculitis is largely attributable to cyclophosphamide: an important drug but also a serious complication, such as infections and cancer. As a result, a number of multicentre randomised controlled trials have evaluated more judicious use of cyclophosphamide, including the pivotal CYCAZERAM study, in which long-term cyclophosphamide treatment was replaced by azathioprine.The French WEGENT trial showed that methotrexate and azathioprine were similarly effective in maintaining relapse-free survival in AAV, while the EUVAS IMPROVE trial showed that azathioprine was superior to mycophenolic acid in preventing relapse(Basu (2021)).This means that the choice of treatment does make a difference to the recurrence of AAV, as some treatments may have a suppressive effect, while others may instead induce recurrence.

In addition to treatment modalities and the characteristics of the disease itself, studies have shown that environmental exposures such as smoking also contribute to a higher recurrence rate of AAV disease. In Makoto Yamaguchi's study, the authors summarise the current research and so far environmental factors, such as silica exposure, infections (especially Staphylococcus aureus) and drug exposure, are considered to play a potentially important role in the development of AAV.

In several cohort studies, smoking has also been considered an environmental component that influences the onset of autoimmune disorders. However, there are currently few research addressing the impacts of smoking on AAV. In addition, previous research has not examined the impact of environmental factors, like as smoking, on the recurrence of AAV. The authors examined microscopic polyangiitis (MPA) as a study to see if a smoking history is an independent risk factor for the recurrence of MPA and if this risk is dose-dependent. The Wilcoxon rank sum test or Fisher's exact test was utilised to analyse differences in clinical features between never smokers and current/former smokers(Yamaguchi and Ando (2018)).

To identify relevant factors independently linked with each outcome, log-rank tests and/or univariate and multivariate Cox proportional risk models were utilised. The findings suggest that differences in the intensity of immunosuppressive therapy between smokers and nonsmokers did not influence the occurrence of relapse, and that smoking increased the risk of relapse in Japanese MPA patients in a dose-dependent manner, even after adjusting for clinically relevant factors. In addition to smoking status at baseline, the authors explain that cumulative smoking history is an important predictor of MPA relapse. This emphasises the necessity of smoking cessation in the treatment of MPA patients(Yamaguchi and Ando (2018)).

Greg McDermott, MD, points out that the studies that have been done on the relationship between smoking and AAV risk have produced conflicting results. Some have found no association, others have documented a non-statistically significant trend suggesting an association, and still others have reported a potential protective effect of smoking.However, these studies have been limited by small sample sizes, inconsistent use of reference groups, and a focus on patients with polyangiitis granulomatosa (GPA) who are usually PR3-ANCA positive(McDermott and Xiaoqing Fu (2020)).

The authors show that to date, from their findings, there are no large case-control studies investigating the relationship between smoking and AAV. Or to examine differences in these relationships depending on the type of ANCA.The authors obtained enough cases from the Partners inception AAV cohort, Controls were obtained from the Partners HealthCare Biobank, and grouped them according to patient smoking status into Never smoking, former smoking and For those who were current or past smokers, cumulative smoking exposure in pack-years was determined by calculating the product of the number of years of smoking and the average number of packs smoked per day. The authors found through a statistical experimental study that a higher proportion of former smokers (current or former) were found among AAV patients compared to controls, and that smoking was associated with increased odds of AAV. When stratified by smoking status, both former and current smokers were more likely to have AAV than never smokers. And the authors also observed a strong dose-response relationship, whereby the odds of AAV increased with increasing pack years of exposure. Patients with the greatest cumulative number of years of smoking had the greatest odds of developing AAV compared to never-smokers. And the results of this study remained robust even in multiple sensitivity analyses, an association that was particularly strong for patients with MPO-ANCA(McDermott and Xiaoqing Fu (2020)).

2.2 Related Work of Application of Survival Analysis in Diseases Research

Current studies using survival analysis to examine factors influencing AAV disease recurrence are very limited. However, the use of survival analysis in relapse studies and in disease-related studies is relatively widespread and applicable.This study is therefore also providing a viable option for disease research in AAV, and may provide some informative suggestions and assistance in determining the appropriate treatment options for patients.

In J . C . Carter's study for anorexia nervosa (AN), the authors suggest that knowledge of factors predicting relapse in anorexia nervosa (AN) is needed to develop effective relapse prevention treatments and may also advance the understanding of the psychopathology of AN. The authors used Kaplan-Meier survival analysis and Cox regression on data from 51 patients to examine the rate, timing and prediction of relapse, and Cox proportional hazards regression models were used to examine the effect of continuous predictor variables on the hazard function.Through the study it was found that among AN patients who remained well in the first year after discharge, there was still a significant risk of relapse. A number of variables were shown to be associated with an increased risk of relapse. These findings have implications for the initial treatment of AN and the development of relapse prevention strategies(CARTER and BLACKMORE (2004)).

In the study of conditional survival of patients after frontline therapy for diffuse large B-cell lymphoma (DLBCL), the authors used the Kaplan-Meier method to generate survival curves for different EFS time points (12, 24, 36, 48 and 60 months) for patient survival curves. The study used competing risk models to calculate the 5-year cumulative incidence of lymphoma death(Assouline and Shen Li (2020)).

In the study, authors NingNing, Fan noted that epidemiologically based studies have shown that there are ethnic differences in the clinical presentation and ANCA specificity of AAV, with PR3-ANCA and GPA being the most common in Western populations, while MPO-ANCA and MPA predominate in East Asian countries. The search for biomarkers associated with disease activity and prognosis can help in the clinical management and treatment of AAV. Therefore, in order to identify biomarkers associated with disease activity and prognosis, the authors collected data on clinical indicators at the initial visit and at follow-up through an electronic medical record system. A composite endpoint event was defined as death or end-stage renal disease, and patients were divided into endpoint and non-endpoint groups based on whether they reached the endpoint. The data were then analysed using Kaplan-Meier curves for survival analysis and Cox risk proportional regression models for risk factors affecting renal prognosis in AAV in the study(Ningning (2021)).

In Paul A. Monach's study, he proposed that anti-neutrophil cytoplasmic antibody-associated vasculitis (AAV) requires improved biomarkers of current disease activity and prediction of relapse. To be clinically meaningful, biomarkers must have good longitudinal performance in treated patients and in patients with non-severe relapses. In differentiating active vasculitis from remission, the predictive power of the marker combinations was modelled in order to investigate, adjusting for the effect of treatment on marker levels, the inclusion of patients in sustained remission and those in exacerbation. The authors used Cox proportional hazards models to complement the strengths as well as the weaknesses with other statistical models. In contrast, when predicting AAV relapse, the authors used the Cox proportional hazards model to determine whether marker concentrations in remission prior to an episode differed from those in remission not immediately preceding the episodeMonach and Warner (2022).

After reviewing the above-mentioned studies on the factors influencing ANCA-AAV relapse and survival analysis in the field of the disease. This study hypothesizes that smoking may have a potential impact on the recurrence of ANCA-AAV.

Chapter 3

Data Introduction

All the data used in this research is from Rare Kidney Disease Registry and Biobank, authorised by PARADISE Group HQ. The data contains real clinical data about Irish vasculitis patients.

3.1 Data privacy protection

Identifiable patient/control data are pseudonymised after recruitment by assigning a study ID; their consent form with medical record number will be stored in a secure facility at the local hospital. The pseudonymised data will be uploaded to the eCRF database, which will be mapped to a dedicated password-protected computer using an IP address. The database will be protected behind a mainframe and institutional firewall and only dedicated personnel will have access to it. Coded biological samples are processed and stored centrally at the RKD Biorepository and archived by the Biorepository Technician using the industry standard Freezerworks software; only the Biorepository Technician and Principal Research PI have access to this software(Little (2019)).

The database operates on multiple access levels, with only the Principal Research PI and Research Nurse having full access to the database and gaining modification rights, while other members will have limited access, with only upload and read rights; new access requests will be approved by the Principal Research PI. The registry database will be shared with the College's IT service provider as its security and access rights are managed by them. The database will be backed up regularly on an external third party server located in Dublin(Little (2019)). This diagram 3.1 shows the flow chart for data security and access.

In order to protect the privacy of patients contributing to the study, all researchers, including myself, signed a strict ethics statement which can be found in the appendix.

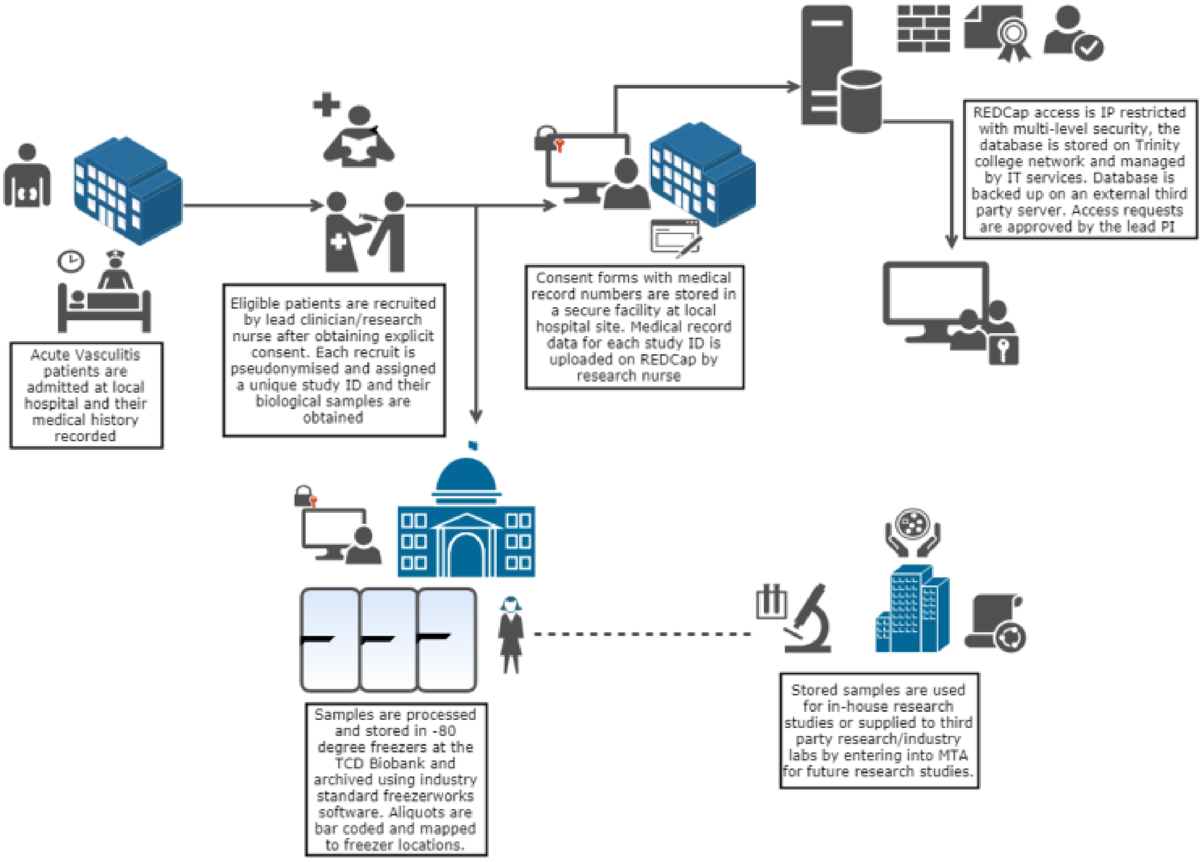

Acute Vasculitis patients are admitted at local hospital and their medical history recorded
Eligible patients are recruited by lead clinician/research nurse after obtaining explicit consent. Each recruit is pseudonymised and assigned a unique study ID and their biological samples are obtained
Consent forms with medical record numbers are stored in a secure facility at local hospital site. Medical record data for each study ID is uploaded on REDCap by research nurse
REDCap access is IP restricted with multi-level security, the database is stored on Trinity college network and managed by IT services. Database is backed up on an external third party server. Access requests are approved by the lead PI
Samples are processed and stored in -80 degree freezers at the TCD Biobank and archived using industry standard freezerworks software. Aliquots are bar coded and mapped to freezer locations.
Stored samples are used for in-house research studies or supplied to third party research/industry labs by entering into MTA for future research studies.

Figure 3.1: Flow diagram of data security and access (Little (2019))

3.2 Data Description

The primary study data was extracted from a comma-delimited CSV file with 4921*496 data entries. The data contained 642 patients' personal information including gender, age, smoking status, etc. and medical information including date of AAV diagnosis, treatment modality, date of consultation, date of relapse, etc.

The volume of data varies considerably from patient to patient, as a record is generated for each patient visit, and some patients may be cured and relapse-free after a few visits for review, while others have recurrent flares.

3.3 Overview of research data

This section describes the data fields used in this study, including data interpretation as well as data loss cases. Table 3.1 shows a general overview of the data used in the study.

The data on the various treatment modalities of induction were missing, and in order

Field Name	**Description**	**Missingness(%)**
RDK ID	RKD ID	0
Gender	gender of patient	0
Age_at_diagnosis	Age at diagnosis of AAV	0
Date Of Visit	Date Of Visit	/
Date of diagnosis	Date of diagnosis	/
Flare_Edit_3	Date of AAV relapse	/
smoking	Smoking status (Current/Previous/Never)	16

Table 3.1: Data used overview

not to affect other data studies, induction was taken out as a separate subject in the subsequent study and not added to the multivariate analysis.The maintenance treatment data field was not included in the study because there was only one data field for whether or not the treatment was maintenance, no more detailed data on which drug was received, and the distribution of true and false data for maintenance itself was severely imbalanced.

3.4 Data processing

3.4.1 Survival Time

Survival data (survival times) represent the simplest form of event history data. A survival time is the amount of time it takes for an event to occur, calculated from a well-defined starting event. Thus, Event History and Survival Data must describe three fundamental elements: a time origin, a scale for measuring time, and an event.(Broström (2021)) Survival times are data that quantify the amount of time that has passed since a specific starting point until the occurrence of a given event during follow-up. This could be the amount of time that has passed since the beginning of remission or since the diagnosis of a disease until the individual has passed away. Rarely is the distribution underlying the data normal, and the data is frequently 'censored', thus standard statistical methods cannot be employedBewick and Cheek (2004).

In a statistical examination of such data, the response is the exact time elapsed from the time of origin to the time of occurrence.Then in practice, it is very difficult to get a very accurate survival time. For example, in this experiment, the SURVIVAL TIME would be the date when the patient entered the experiment for observation from the time of diagnosis until the patient experienced the first relapse, which is the date of the event. However, since there is usually an indeterminate length of time between the onset of relapse and the patient's voluntary attendance at the hospital, this depends on the patient's disease condition and other environmental factors. This factor causes the measured event date to be later than the actual event date.

The AAV patient's diagnosis date is used in this study as the starting date of the patient's entry into the study. If there is a relapse of the patient's condition at a subsequent visit, the date of the relapse is recorded as the date of the event. If the patient did not relapse, the time of the patient's last visit was used as the date of the patient's 'censoring'/end time in the study. Figure 3.2 shows a flow chart of how the study handles survival time.

After removing the data with recorded issues, of which the survival time is less than 0, the distribution of the data is plotted. The figure 3.3 shows the distribution of survival time.

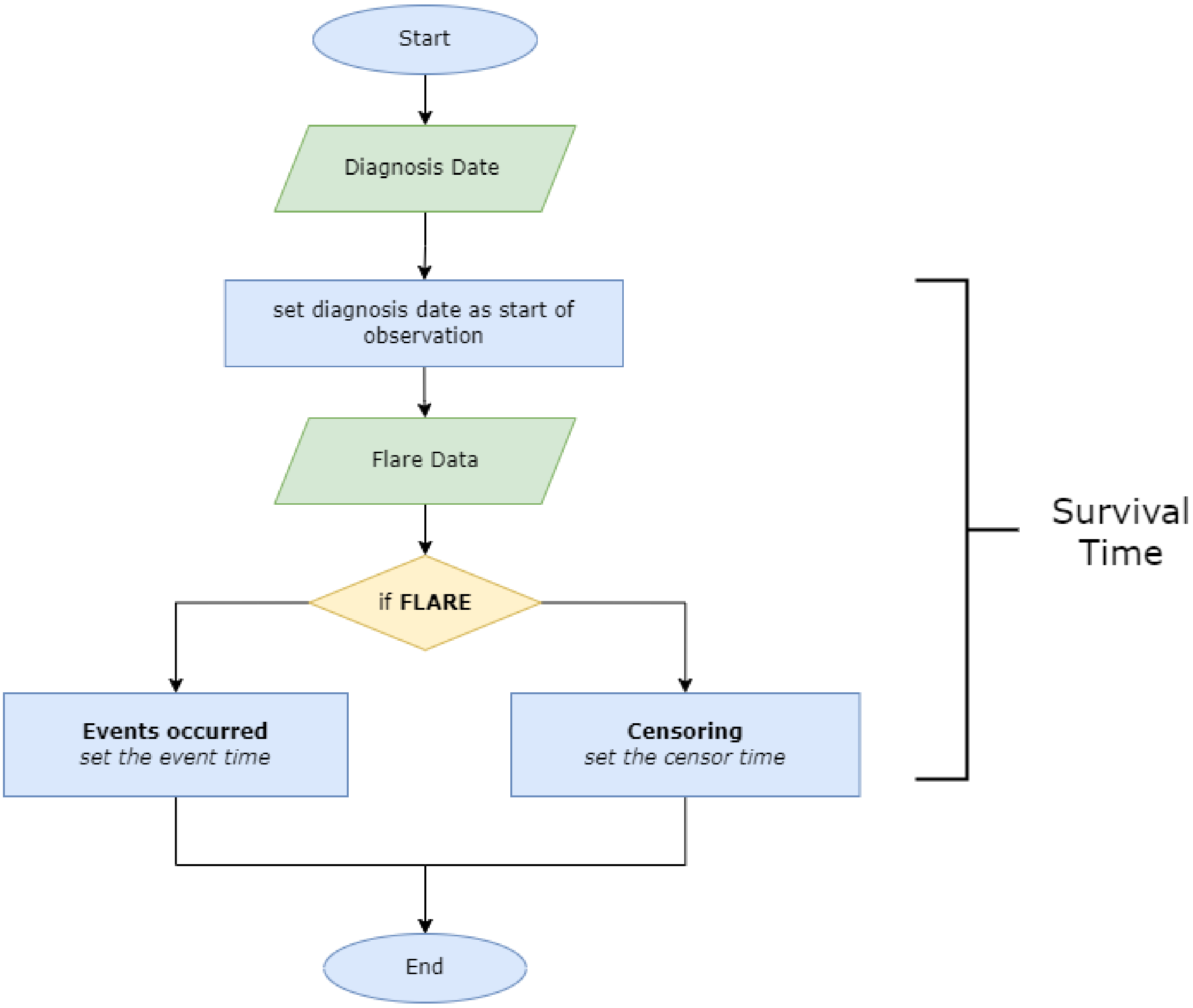


Figure 3.2: Flow chart of survival time

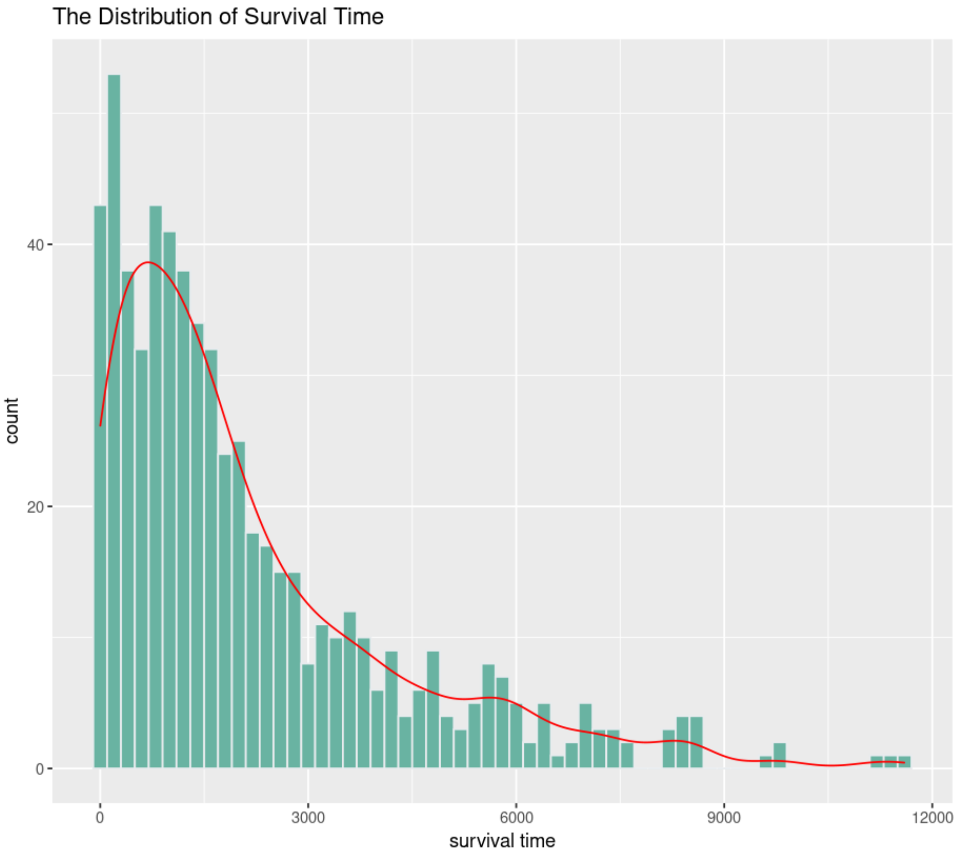


Figure 3.3: Distribution of survival time

3.4.2 Censoring

Censoring is a type of missing data unique to survival analysis. There exists left censoring and right censoring. It occurs when the sample or subject is followed to the end of the investigation, but the event never occurs. It may also be the result of the sample/subject dropping out of the research for reasons other than death, or another type of loss to follow-up. When the event has not occurred, it is right-censored. Left censoring is a circumstance in which the only information known about a event time is that it is less than a specified value.(Broström (2021)) For instance, a patient's duration in remission is censored if he or she is still in remission at the conclusion of the trial. If a patient withdraws from the study before to the conclusion of the study period, then that patient's follow-up time is also considered censored(Bewick and Cheek (2004)). Figure 3.4 shows a sample of right censoring.

In this study, all data censoring was right-censoring.This means that although a proportion of the AAV patients entered the study, the patients did not experience a relapse throughout the study recording period until the last recorded visit or they dropped out of the study halfway through and may have moved away from Ireland. These patients are lost to follow up. But we are not sure if they had a flare at the following time, it is possible that patients may experience a relapse at a subsequent time or remain non-relapsing.

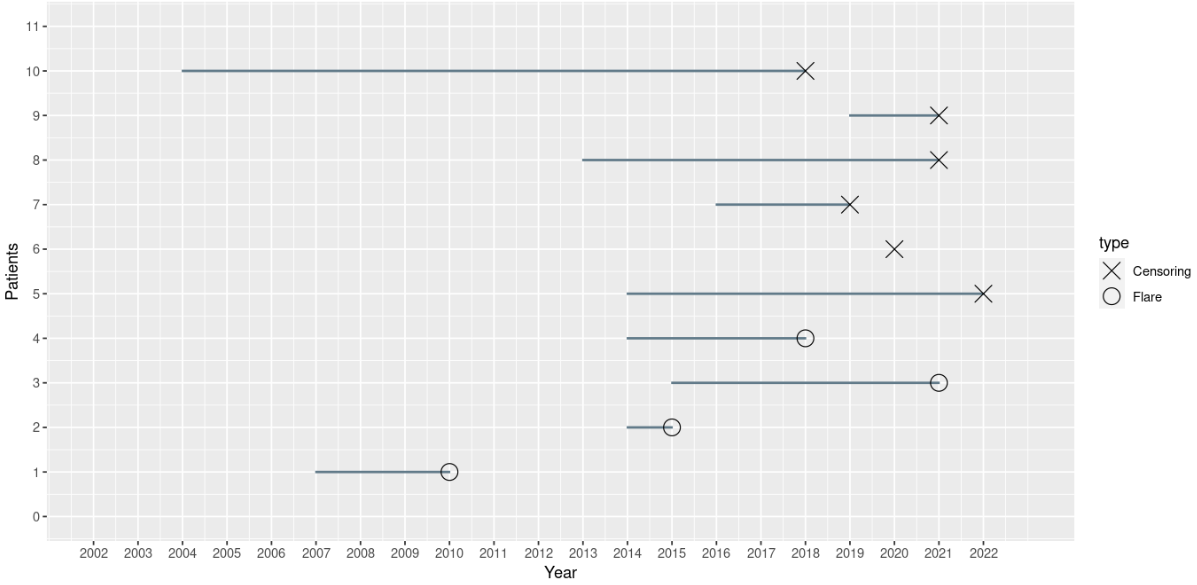


Figure 3.4: Diagram of right- censoring

Chapter 4

Model Introduction

This chapter presents the methods used in the study and the corresponding model presentations.

4.1 Survival Analysis

Survival analysis is used to analyze the rates of occurrence of events over time, without assuming the rates are constant(Broström (2021)). The practice of survival analysis uses rationality to characterise, measure and analyse events to make predictions not only about survival but also about the 'time-event process', for example, the length of time until a state change or event occurs, e.g. from singleness to engagement, from disease onset to disease remission, from addiction to withdrawal.

Given that life expectancy in genetic, biological, or mechanical terms may be decreased by sickness, conflict, the environment, or other causes, much of the study in survival analysis is comparing groups or categories or investigating variables that influence the survival process.

Clinical trials are frequently used in medical research to evaluate the efficacy of novel medications or disease treatment approaches. In these types of studies, survival analysis is frequently used to assess the risk of death or illness recurrence across patients receiving various medications or treatments. The outcomes of such analysis can also yield relevant and significant data. Similarly, survival analysis has been utilised in biological studies. Mathematical biologists have long been interested in evolutionary perspectives on ageing in human and other animal populations. Using survival analysis as a fundamental research technique, the life history of a species is demarcated, and their survival processes are associated with a variety of physical and behavioural characteristics in order to examine their reaction to their environment(Liu (2012)).

4.2 Basic lifetime functions

When analysing survival data, two time-dependent functions are crucial: the survival function and the hazard function. Density and cumulative distribution functions are typically used to describe statistical model characteristics. These routines are not appropriate for usage with censored or shortened data. The survival and hazard functions are more appropriate. The definition of the survival function $S(t)$ is the likelihood of living until at least time t. The hazard function $h(t)$ is the conditional likelihood of living until death occurs at time t. There is also a cumulative hazard function $H(t)$ based on the hazard function $h(t)$.Liu (2012)

Describing time as a continuous process. Let $f(t)$ be the probability density function (p.d.f.) of event time T. Then, according to probability theory, the cumulative distribution function (c.d.f.) over the time interval $(0, t)$, denoted by $F(t)$, represents the probability the random variable T takes from time 0 to time $t(t = 0, 1, ..., \infty)$, given by

$$F(t) = Pr(T \leq t) = \int_0^t f(u)du.$$

4.2.1 Survival Function

Survival function $S(t)$ is defined as the probability of surviving past time t, and at time t, denoted by $S(t)$, is simply the complement of the c.d.f.:,

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t) = 1 - F(t)$$

where T is the (random) life length under study.

4.2.2 Hazard Function

The hazard function is crucial to comprehending survival analysis. It measures the risk of dying in a short interval $(t, t+s)$ immediately after t, hazard function at time t is defined as the instantaneous rate of failure at time t, denoted by $h(t)$ and mathematically defined by,

$$h(t) = -\frac{1}{S(t)}\frac{dS(t)}{d(t)} = \frac{f(t)}{S(t)}$$

In survival analysis, given standardisation and its unique sensitivity to the change in the survival function, the hazard function is the preferred indicator for expressing the relative risk of experiencing a certain event.

4.2.3 Cumulative Hazard Function

The cumulative hazard function is defined as the integral of the hazard function,

$$H(t) = \int_0^t h(s)ds, t \geq 0$$

4.3 Life table method

As one of the descriptive approaches of survival analysis, life tables could be usually found in the contexts of epidemiology and demography. It records the proportion of surviving (event doesn't occur) to the end of each time interval. Comparing statistics such as "n-year recurrence rate", the life table method is a better representation which reflects sample size loss over time. Compared to other model-based analysis methods, life tables are easier to interpret and provide an intuitive reference of risks and expectations.

Based on how the data is observed and summarized, life tables can be categorized into two types: (1) cohort life table and (2) current life table. Cohort life table needed to be built on information of groups of persons who joined the observation (e.g., born, diagnosed, . . .) at the same time and the cohort is followed up until the event occurs (e.g., death, relapse, . . .) for the last member of the group. In contrast, the current life tables are grouping members by how long they have been added to the samples (e.g., by age or length of time since diagnosis).

The cohort life table could reflect the implicit difference of survival experience from people under different backgrounds of the times. For instance, in a decade, there may be significant changes in economic and medical standards. However, this requirement is making the generation of cohort life tables challenging in practice among human population as the information of a long period is rarely available (Lahiri (2018)). And, datasets with limited number of observations are also likely to encounter this problem when adopting cohort life tables.

Current life tables, on the other hand, assume that an individual's mortality is independent of the time point they joined the observations but is related to the mortality observed from people at that age in the given population.With the assumption, the current life tables usually contain data during a specified calendar year of a short period of time and is not giving the exact pattern reflecting the mortality experience. Instead, it provides a cross-sectional (Lahiri (2018)) "effective means of summarizing mortality and survival experiences of a population and under study for that calendar year" (WHO (1977)).

Current life table in survival analysis usually consists of the following columns: (i)

Interval, (ii) Number failed, (iii) Number censored, (iv) Effective sample size, (v) Survival function, and (vi) Standard error of survival.

In this study, we adopt the current life table, which means the survival function in it is a function of the number of days since diagnosis rather than of time.

4.4 Kaplan–Meier Curve

The graph of S(t) against t is known as the survival curve(Bewick and Cheek (2004)). The Kaplan-Meier curve is a non-parametric estimator frequently used to estimate the survival distribution, i.e., to compute the proportion of participants who survive a particular period following an intervention or treatment. Even if people drop out or are observed for varying amounts of time, it measures survival across time. When individuals are lost, survival rates frequently decline. At each loss, the curve will demonstrate a fall, but between losses, it will be flat(Kalra (2016)).

In any clinical or community-based study, the intervention's effectiveness is determined by the number of participants alive or successfully avoiding adverse outcomes, including death. However, the reality is that not all participants can continue for the duration of the research; a proportion typically drops out in the middle of the study, maybe because they moved out of the study area or were lost to follow-up. In this instance, the Kaplan-Meier curve estimate is a straightforward and trustworthy method that may be utilised to make more accurate conclusions on the survivability of participants(Kalra (2016)). Let d_i be the number of events at time t_i ($d_i = 1$ if there are no tied cases at t_i); then the Kaplan–Meier estimator for the probability of survival at time t is,

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

The length of the horizontal line along the x-axis of the time series shows the interval's survival time. The vertical axis depicts the predicted survival probability. The number of observations influences the precision of the estimate. At the conclusion of the trial, when a small number of individuals are at hazard of an event, estimations of survival may be incorrect. Notably, variables are not controlled for in this test, which requires categorical predictors. Additionally, it cannot accept variables whose values change over time(Sedgwick and Joekes (2013)).

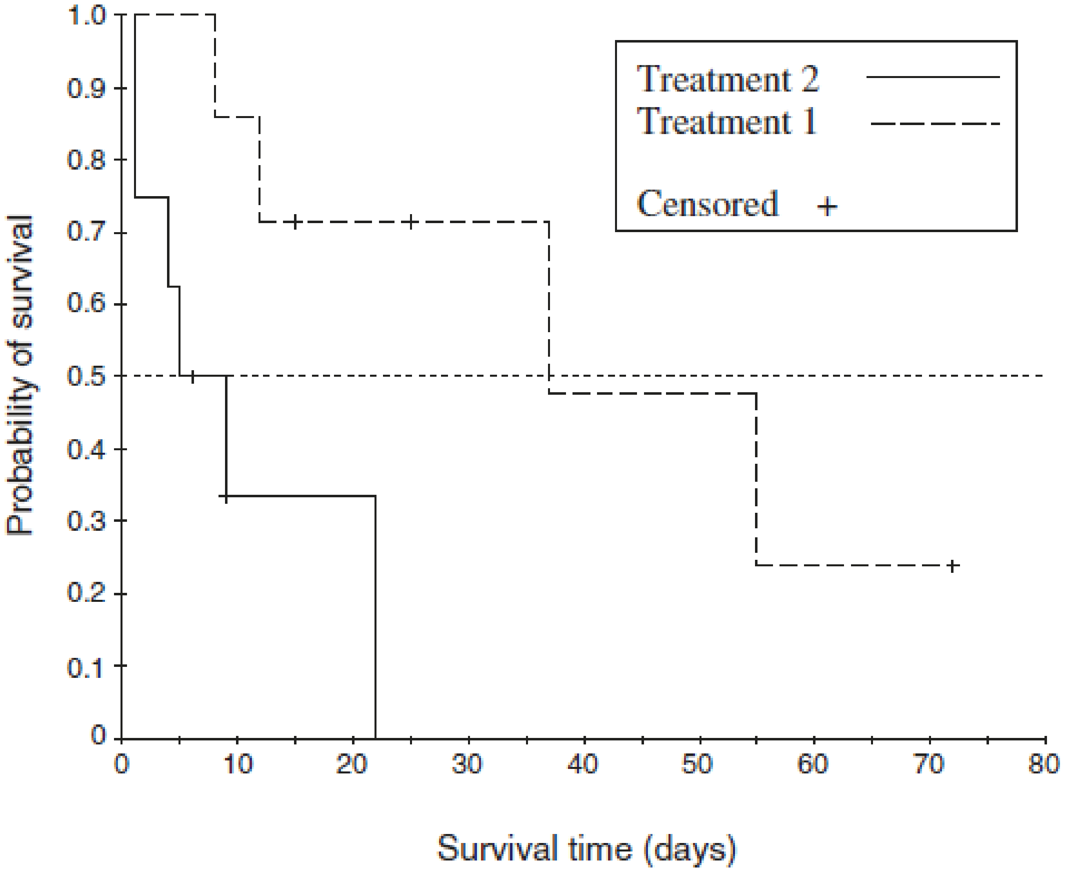


Figure 4.1: Example of a Kaplan Meier curve study (Bewick and Cheek (2004))

4.4.1 Log-Rank test

Often in research, we want to be able to compare the survival probability of two or more groups of individuals, for example by examining the differences in the effect of different therapeutic agents on patient recovery. Although the Kaplan-Meier curve can help us to see differences between groups, it is difficult to draw direct conclusions about which group is more effective in a study.

Although it is possible to calculate Kaplan-Meier curves for different groups and also to compare the proportion of survival at any given time. However, there are some problems with this method in that it does not provide a comparison of the overall survival probability of the different groups.The Log-rank test can be used to solve this problem.The logrank test was employed to test the null hypothesis that the probability of an event (in this case, an AAV recurrence) happening at any given time did not vary between patient groups. The analysis was based on the occurrence's timing. For each time point, we calculate the number of recurrences observed in each group and the number of recurrences expected if there had been little difference between the groups(Bland and Altman (2004)).The test statistic is calculated as follows,

$$\chi^2(logrank) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2},$$

where the O_1 and O_2 are the total numbers of observed events in groups 1 and 2, respectively, and E_1 and E_2 the total numbers of expected events.The overall expected number of events for a group is equal to the sum of the expected number of occurrences at each event. The estimated number of events at the time of an event can be computed by multiplying the risk of death at that moment by the number of individuals still alive(Bewick and Cheek (2004)).

When an event occurs, the same computations are run each time. If a survival time is filtered, a person is regarded to be at risk of dying during the week of the censoring but not later weeks. This method of dealing with censored observations is identical to the Kaplan-Meier survival curve(Bland and Altman (1998)). It is a powerful test against proportional hazards alternatives, but may be very weak otherwise.

The logrank test is most likely to detect differences between groups when the risk of events in one group is constantly greater than in the other. Trying to detect differences is more difficult for the logrank test when the survival curves of different groups cross, as may be the case when comparing medical and surgical interventions. Survival curves are therefore usually always plotted when analysing survival data. And the logrank test is a pure test of significance, it does not provide an estimate or confidence interval for the size of the difference between groups(Bland and Altman (2004)).

4.5 Cox (Proportional Hazard) Regression model

The log-rank test is used to test for variations in survival times between groups, however it does not permit the consideration of additional explanatory variables. Similar to a multiple regression model, Cox's proportional hazards model can test for changes in survival times for specific groups of patients when several covariates are considered.

4.5.1 Proportional Hazard Assumption

Proportional hazards is a property of survival models that is fundamental for the development of non-parametric regression modelsCox (1972).

The proportional hazards property is crucial to Cox regression. If $h_1(t)$ and $h_0(t)$ are hazard functions from two separate distributions, the definition of proportional hazard assumption is presented as,

$$h_1(t) = \psi h_0(t), for\ all\ t \geq 0$$

where ψ is positive and constant.

Provided that the above expression holds, based on this expression, cumulative risk functions $H_1(t)$ and $H_0(t)$ also have the same properties,

$$H_1(t) = \psi H_0(t), for\ all\ t \geq 0$$

where ψ is positive and constant and same with the ψ in hazard function. For all $t \geq 0$ means the constant ψ is independent with time t.

4.5.2 Proportional Hazard Assumption test

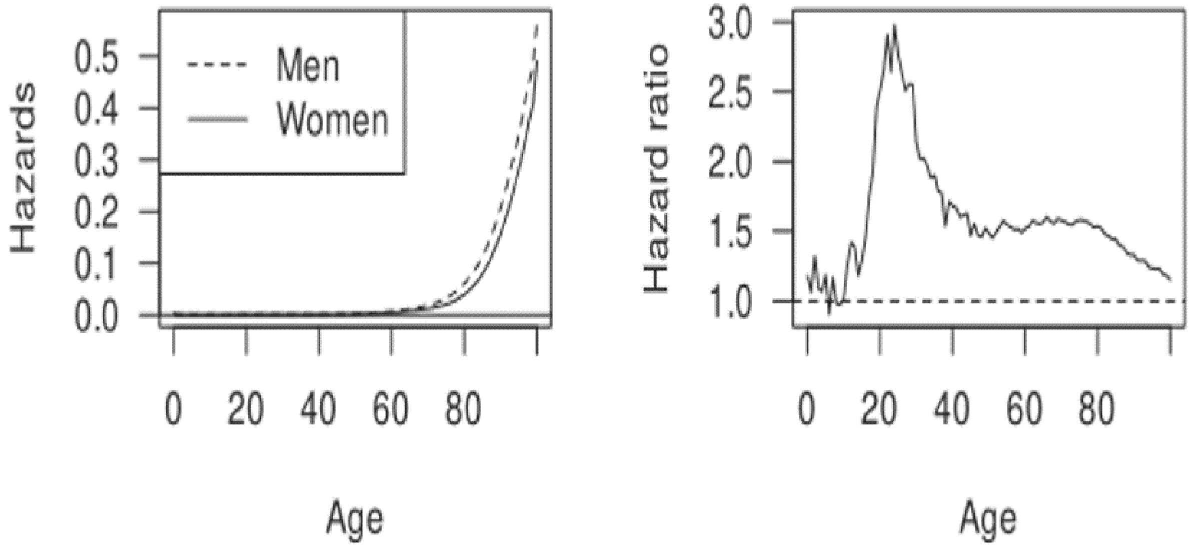


Figure 4.2: Example of two survival curves(Broström (2021))

In Figure 4.2, and in the graph on the left, the usual first reaction is that the two curves are consistent with the proportional hazard assumption. However, if the difference in survival probability between the two groups at each point is calculated and the curves are plotted, the result on the right shows that the difference in survival probability between the two curves at different times is very large and that the two hazard functions are not proportional.Therefore a better way to present the survival curve is needed.

Log scale could help to see smaller numbers, as well as in log-scales. The proportional hazard should be presented as a constant vertical distance between the curves.In order to test the proportional hazards hypothesis, the most common graphical technique for analysing the PH assumption is comparing estimated log-minus-log survival curves across various (combinations of) categories of examined variables(Sestelo (2017)).

A log–log survival curve is just a modification of an estimated survival curve produced by taking the natural log of an estimated survival probability twice. If the hazards are proportional, the stratum-specific log-minus-log plots should demonstrate constant differences, or be roughly parallel. These visual solutions are straightforward to apply but have certain restrictions. Kaplan-Meier plots are ineffective for distinguishing non-proportionality when the covariate has more than two levels because the graphs become too cluttered . Similarly, although the PH assumption may not be violated, log-minus-log curves are rarely completely parallel in actuality and tend to grow sparser and hence less accurate over longer time intervals. It is impossible to estimate how close to parallel is close enough and, hence, how proportional the risks are. Often, the decision to adopt the PH hypothesis rests on whether or not these curves intersect(Bellera et al. (2010)).

In figure 4.3, there is a example of two survival curves are a perfect match for the proportional hazard and can be seen to be perfectly parallel in the log-log scale. Such as a variation of a test originally proposed by Schoenfeld (1982). This is a test of correlation between the Schoenfeld residuals and survival time. A correlation of zero indicates that the model met the proportional hazards assumption (the null hypothesis).

In addition to this, there are also more convenient and accurate functions in some packages to calculate whether the data meets the PH assumption. Such as a variation of a test originally proposed by Schoenfeld. This is a test of correlation between the Schoenfeld residuals and survival time. The Schoenfeld residuals computation is best represented by fitting the Cox Proportional Hazards model to a data sample. A strategy for determining residuals in various directions was proposed by Schoenfeld. These residuals, properly known as Schoenfeld residuals, are defined and calculated for the cox proportional hazards model(Schoenfeld (1982)). In the same way as the specification of the partial probability does not depend on the passage of time, the derivation of the Schoenfeld residual is determined by the rank order of survival times(Liu (2012)). The Schoenfeld residuals are the discrepancies that exist between the covariate vector that represents the person at the event time t_i and the expectation of the covariate vector across the risk set $R(t_i)$. A correlation of zero indicates that the model met the proportional hazards assumption (the null hypothesis)(Sestelo (2017)).

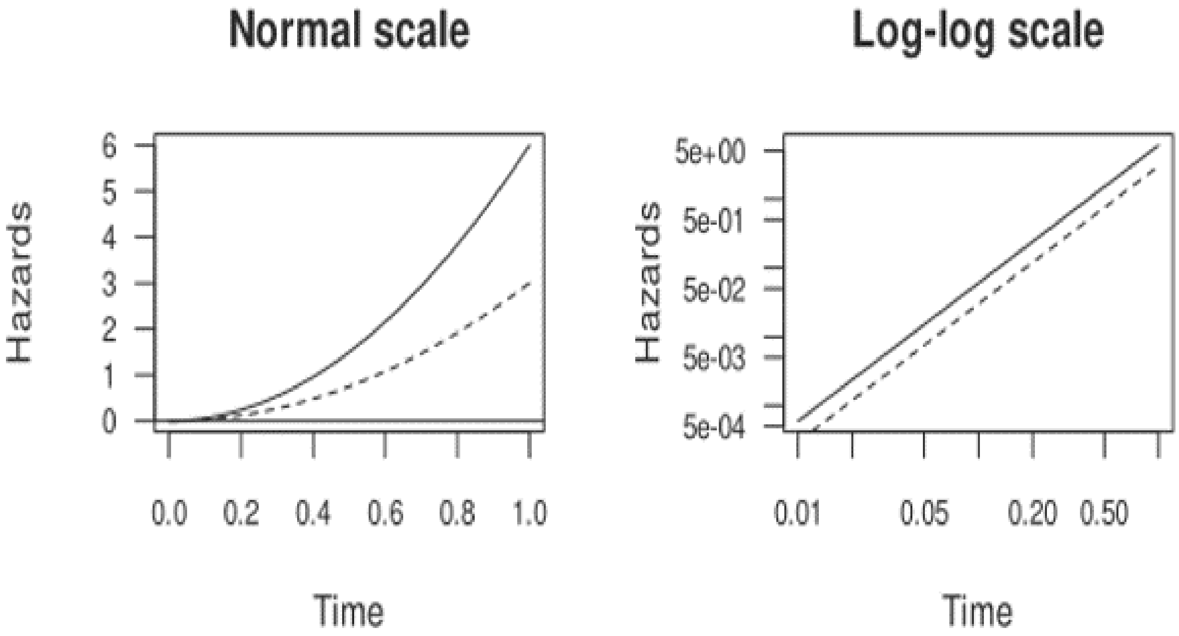


Figure 4.3: Two hazard functions are constant(Broström (2021))

4.5.3 Cox (Proportional Hazard) Regression model

Cox's proportional hazards model is comparable to a multiple regression model and can test for changes in survival times for certain patient groups when additional variables are considered. The response (causal) variable in this model is 'hazard'. The hazard is the chance of dying (or experiencing the event in question) or the danger of dying at that moment, assuming the patient survives to that point(Bewick and Cheek (2004)).

In Cox's model, no assumptions are made about the probability distribution of hazards. However there is an assumption of proportional hazards for the probability of survival for both groups. The expressions of the model are as follows,

$$ln\frac{h(t)}{h_0(t)} = b_1x_1 + ... + b_px_p$$

where $h(t)$ is the hazard at time t; $x_1, x_2 \ldots x_p$ are the explanatory variables; and $h_0(t)$ is the baseline hazard when all the explanatory variables are 0. The coefficients $b_1, b_2 \ldots b_p$ are estimated from the data using a statistical package(Bewick and Cheek (2004)).

4.6 Accelerated Failure Time Model

The proportional hazards (PH) model is the most popular for analysing the impact of multivariate factors on survival time. Nonetheless, the data do not always support the assumption of continuous hazards in the PH model. The violation of the PH assumption results in the misinterpretation of estimation results and a reduction in the statistical power of the associated tests. As opposed to the PH model, accelerated failure time (AFT) models do not assume constant hazards in survival data. The AFT models can be utilised as an alternative to the PH model if the premise of constant hazards is violated(Faruk (2018)).

Similar to a multiple linear regression model, the AFT models can be rewritten to indicate a direct relationship between the logarithm of the survival time and the explanatory variables(Orbe et al. (2002)). Suppose that $Y = log\ T$ is linearly associated with the covariate vector x . Then,

$$Y = logT = \mu^* + x'\beta^* + \tilde{\sigma\varepsilon},$$

where β^* is a vector of regression coefficients on Y or $log\ T$, μ^* is referred to as the intercept parameter, and the parameter $\tilde{\sigma}$ is an unknown scale parameter. The term ε represents random errors that follow a particular parametric distribution of T with survival function S , cumulative distribution function F , and probability density function f .Formulating AFT with respect to the random variable T, rather than to log T, we have,

$$T = exp(\mu^* + x'\beta^*)exp(\tilde{\sigma\varepsilon})$$

$$= exp(\mu^* + x'\beta^*)\tilde{E}$$

where $\tilde{E} = exp(\tilde{\sigma\varepsilon}) > 0$ has the hazard function $h_0(\tilde{e})$ and is independent of β^*.

The distribution of event time T often follows some distinct patterns. The commonly parametric distributions, which correspond to the AFT model, are Weibull, exponential, log-normal, gamma, and log-logistic. For example, if $exp(\tilde{\sigma\varepsilon}) \sim Exp(...)$, then this AFT model would a Exponential AFT model, while if $exp(\tilde{\sigma\varepsilon}) \sim Weibull(...)$, the model would be a Weibull AFT model.Figure 4.4,Figure 4.5,Figure 4.6 show the probability density functions of Weibull distribution, Exponential distribution and Gamma distribution under different parameters respectively. We can choose the one that best matches our a prior beliefs about the hazard function or we can compare different parametric models and choose among them using a criterion like Akaike Information Criterion(AIC).

AIC is an estimate of the prediction error and thus the relative quality of the statistical model for a given data setStoica and Selen (2004). The Akaike Information Criterion (AIC) is a mathematical approach that can be used to evaluate how well a model fits the data that it generates. In the field of statistics, AIC is a tool that is utilised to evaluate and contrast the various alternative models in order to choose the one that provides the best fit for the data. The AIC is computed by employing the highest likelihood estimate of the model in conjunction with the number of independent variables that were utilised in the construction of the model. According to the AIC, the model that explains the most variation with the fewest independent variables is the one that is the most suitable representation of the data.

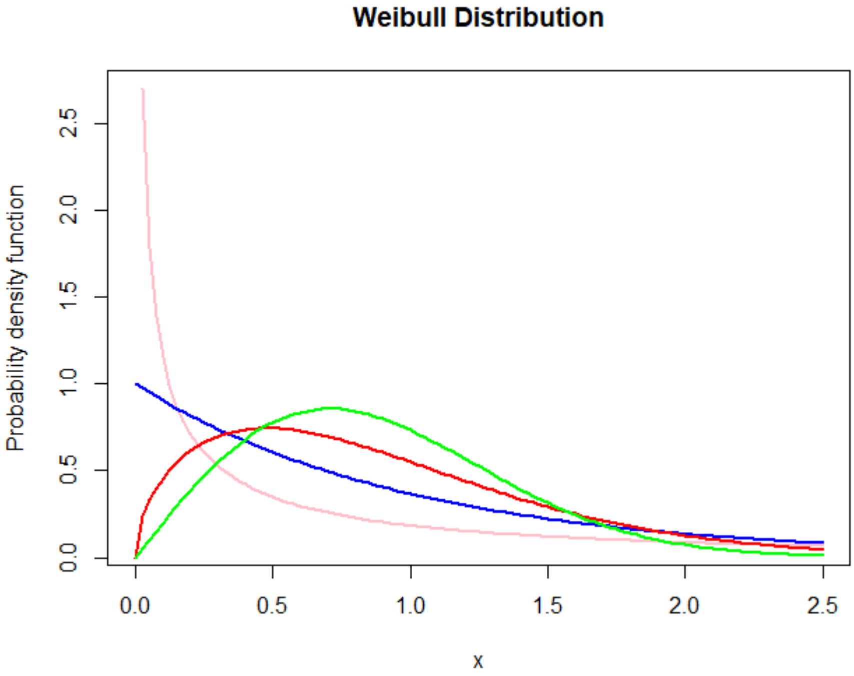


Figure 4.4: P.D.F of Weibull distribution in different parameters

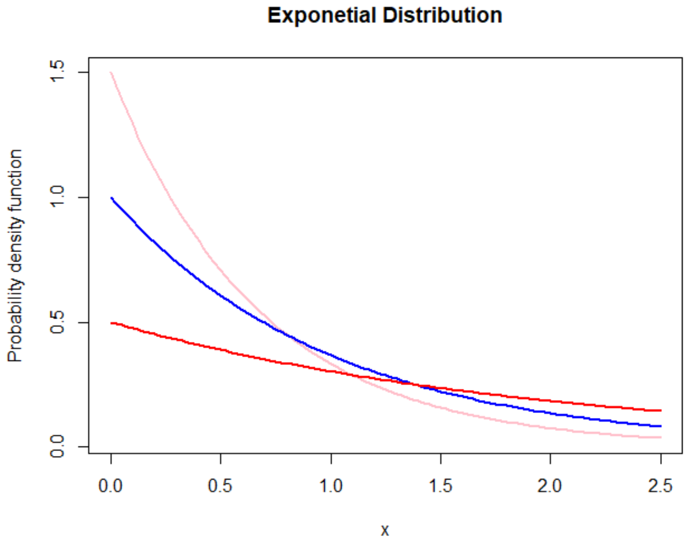


Figure 4.5: P.D.F of Exponential distribution in different parameters

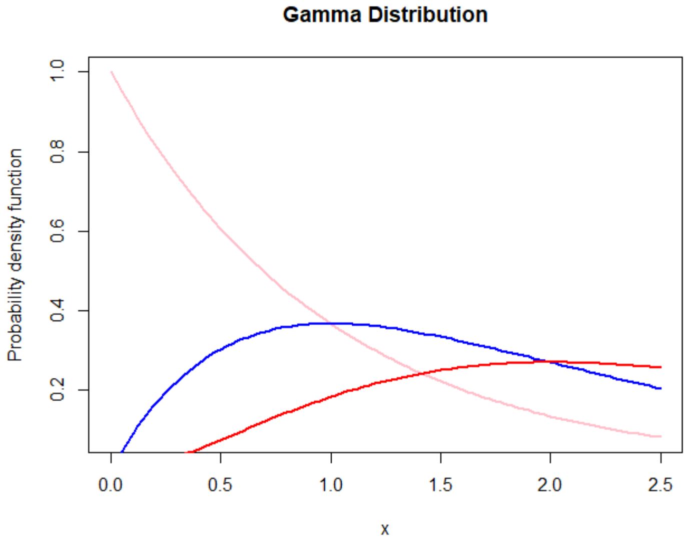


Figure 4.6: P.D.F of Gamma distribution in different parameters

Chapter 5

Results and Evaluation

To find out if smoking is an important factor of the relapse of ANCA-AAV. This chapter uses flare as the event and replaces the experimental data with the life table, Kaplan-Meier curve, Cox Proportional Hazard regression model, and Accelerated Failure Time model in the survival analysis presented in Chapter 4. The experimental results were obtained and analysed by testing the model hypotheses in the Failure Time model. The results of the different models are also compared and combined to reach a final conclusion.

5.1 Life Table Method

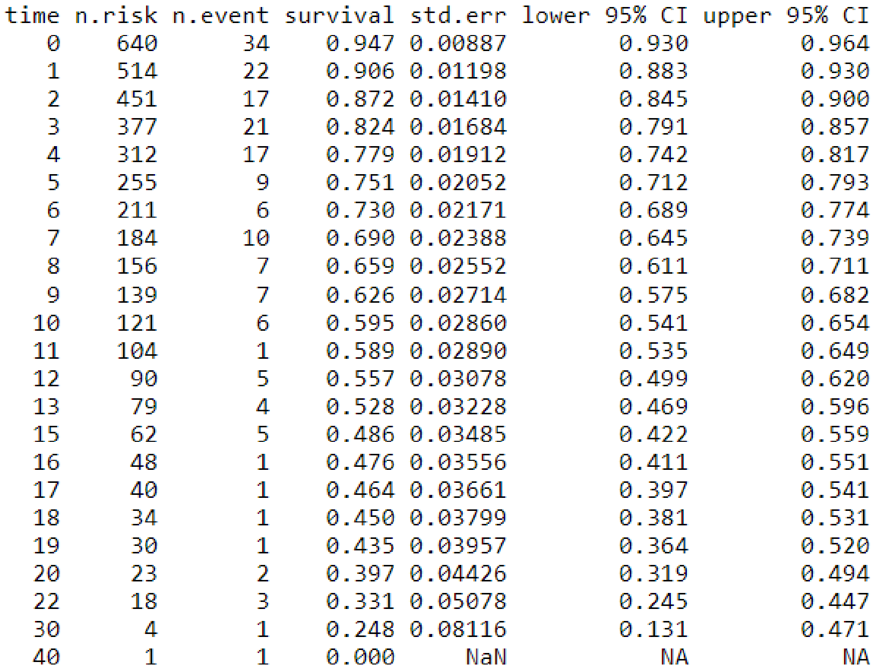
We constructed the following life table for our sample population:

From the first row, we can see that within one year after diagnosis (time=0), 34 patients out of a sample of 640 experienced a relapse. About 50% of the patients will experience a relapse within 15 years. The number of patients in risk in the second year drops from 640 to 514 because of the event (n.event=34 in time=0) and censoring.

In order to explore the effect of smoking on relapse, we can first categorise patients according to their smoking history and then create life tables for each group.

Fig 5.2, 5.3, and 5.4 illustrates the change in the odds of relapse over time for patients with different smoking histories. Here samples with unknown smoking histories (smoking=4) are ignored. From the maximum time to relapse, first year relapse rate (survival probability in time = 0) we can see that current smokers and previous smokers have a higher risk of recurrence. Among current smokers, 50% of the patients experienced relapse within 11 years. For previous smokers, 50% of the patients experienced within 13 to 14 years. For those who have never smoked, the number is 19 years.

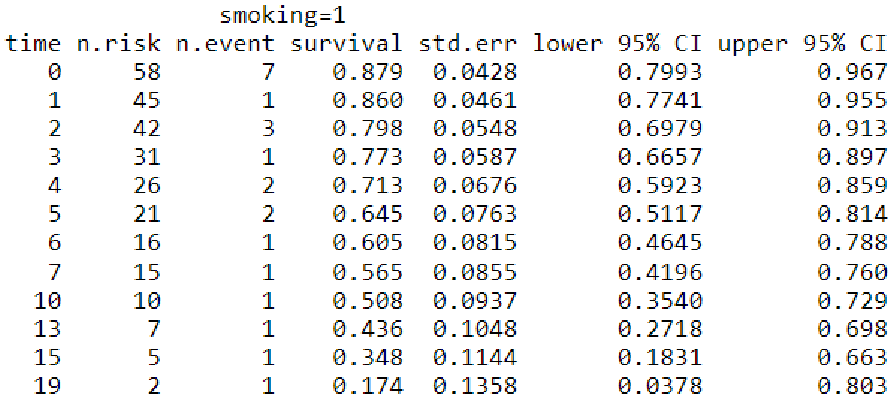
We can easily see the difference in these three tables. However, it still not safe to conclude the negative effects of smoking on relapse from the differences in life tables alone. However, as the volume of data rises in the future and based on our subsequent



```
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0    640      34    0.947 0.00887        0.930        0.964
   1    514      22    0.906 0.01198        0.883        0.930
   2    451      17    0.872 0.01410        0.845        0.900
   3    377      21    0.824 0.01684        0.791        0.857
   4    312      17    0.779 0.01912        0.742        0.817
   5    255       9    0.751 0.02052        0.712        0.793
   6    211       6    0.730 0.02171        0.689        0.774
   7    184      10    0.690 0.02388        0.645        0.739
   8    156       7    0.659 0.02552        0.611        0.711
   9    139       7    0.626 0.02714        0.575        0.682
  10    121       6    0.595 0.02860        0.541        0.654
  11    104       1    0.589 0.02890        0.535        0.649
  12     90       5    0.557 0.03078        0.499        0.620
  13     79       4    0.528 0.03228        0.469        0.596
  15     62       5    0.486 0.03485        0.422        0.559
  16     48       1    0.476 0.03556        0.411        0.551
  17     40       1    0.464 0.03661        0.397        0.541
  18     34       1    0.450 0.03799        0.381        0.531
  19     30       1    0.435 0.03957        0.364        0.520
  20     23       2    0.397 0.04426        0.319        0.494
  22     18       3    0.331 0.05078        0.245        0.447
  30      4       1    0.248 0.08116        0.131        0.471
  40      1       1    0.000     NaN           NA           NA
```

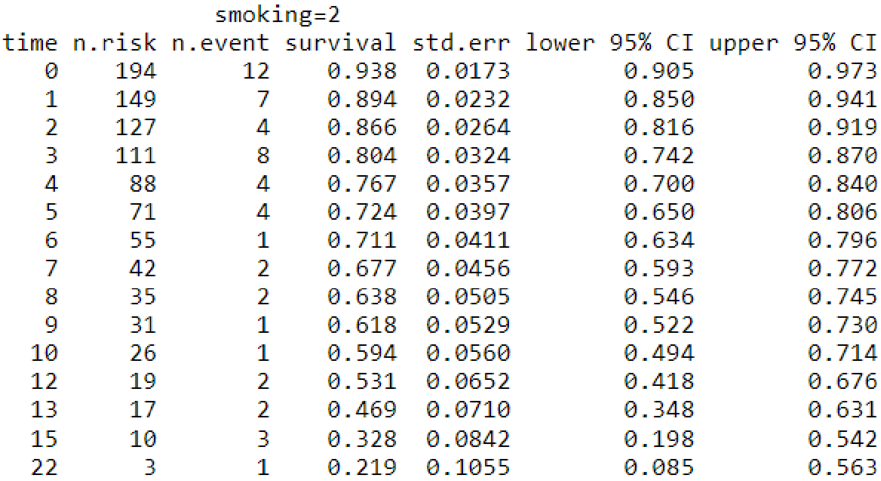
Figure 5.1: Life table of relapse, event=Flare, unit of time is year

consistent findings using other methods, we believe life tables can still be used as a simple quick reference or a material for patient education because of its intuitive nature.



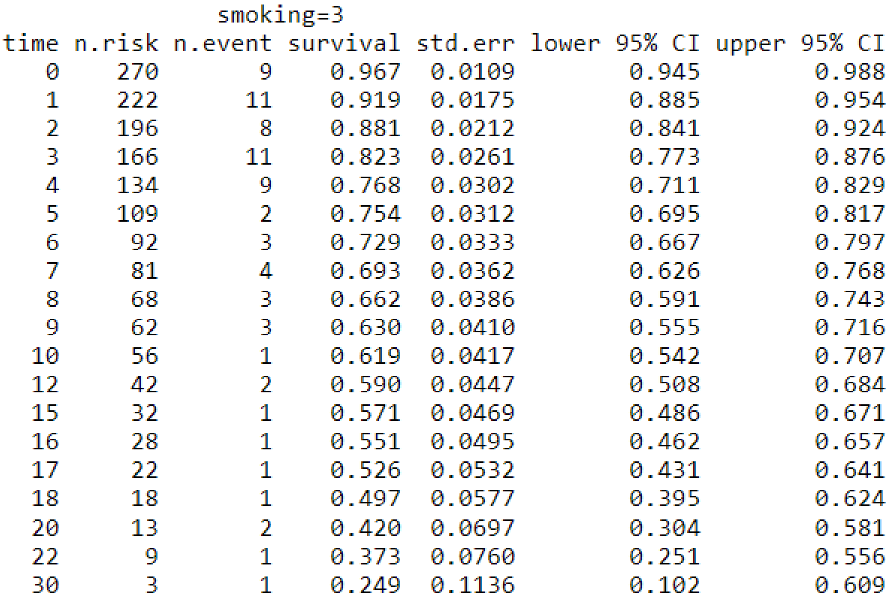
```
                smoking=1
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0     58       7    0.879  0.0428       0.7993        0.967
   1     45       1    0.860  0.0461       0.7741        0.955
   2     42       3    0.798  0.0548       0.6979        0.913
   3     31       1    0.773  0.0587       0.6657        0.897
   4     26       2    0.713  0.0676       0.5923        0.859
   5     21       2    0.645  0.0763       0.5117        0.814
   6     16       1    0.605  0.0815       0.4645        0.788
   7     15       1    0.565  0.0855       0.4196        0.760
  10     10       1    0.508  0.0937       0.3540        0.729
  13      7       1    0.436  0.1048       0.2718        0.698
  15      5       1    0.348  0.1144       0.1831        0.663
  19      2       1    0.174  0.1358       0.0378        0.803
```

Figure 5.2: Life table of current smokers (smoking=1)



```
                smoking=2
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0    194      12    0.938  0.0173        0.905        0.973
   1    149       7    0.894  0.0232        0.850        0.941
   2    127       4    0.866  0.0264        0.816        0.919
   3    111       8    0.804  0.0324        0.742        0.870
   4     88       4    0.767  0.0357        0.700        0.840
   5     71       4    0.724  0.0397        0.650        0.806
   6     55       1    0.711  0.0411        0.634        0.796
   7     42       2    0.677  0.0456        0.593        0.772
   8     35       2    0.638  0.0505        0.546        0.745
   9     31       1    0.618  0.0529        0.522        0.730
  10     26       1    0.594  0.0560        0.494        0.714
  12     19       2    0.531  0.0652        0.418        0.676
  13     17       2    0.469  0.0710        0.348        0.631
  15     10       3    0.328  0.0842        0.198        0.542
  22      3       1    0.219  0.1055        0.085        0.563
```

Figure 5.3: Life table of previous smokers (smoking=2)



```
                smoking=3
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0    270       9    0.967  0.0109        0.945        0.988
   1    222      11    0.919  0.0175        0.885        0.954
   2    196       8    0.881  0.0212        0.841        0.924
   3    166      11    0.823  0.0261        0.773        0.876
   4    134       9    0.768  0.0302        0.711        0.829
   5    109       2    0.754  0.0312        0.695        0.817
   6     92       3    0.729  0.0333        0.667        0.797
   7     81       4    0.693  0.0362        0.626        0.768
   8     68       3    0.662  0.0386        0.591        0.743
   9     62       3    0.630  0.0410        0.555        0.716
  10     56       1    0.619  0.0417        0.542        0.707
  12     42       2    0.590  0.0447        0.508        0.684
  15     32       1    0.571  0.0469        0.486        0.671
  16     28       1    0.551  0.0495        0.462        0.657
  17     22       1    0.526  0.0532        0.431        0.641
  18     18       1    0.497  0.0577        0.395        0.624
  20     13       2    0.420  0.0697        0.304        0.581
  22      9       1    0.373  0.0760        0.251        0.556
  30      3       1    0.249  0.1136        0.102        0.609
```

Figure 5.4: Life table of non-smokers (smoking=3)

5.2 Kaplan-Meier Curve results

The processed data were divided into three groups: current smokers, previous smokers, and patients have never smoked, depending on the patient's smoking status. The data were then substituted into the Kaplan-Meier curve and the results are shown in figure 5.5. From this figure, we can see that there is a difference in the survivability probability between the three groups and observe the results of the hypothesis testing experiment to test whether there is a difference between the three groups. p-value is 0.08, which is higher than the common standard 95% confidence interval of 0.05, but less than 0.1. The results are still significant.

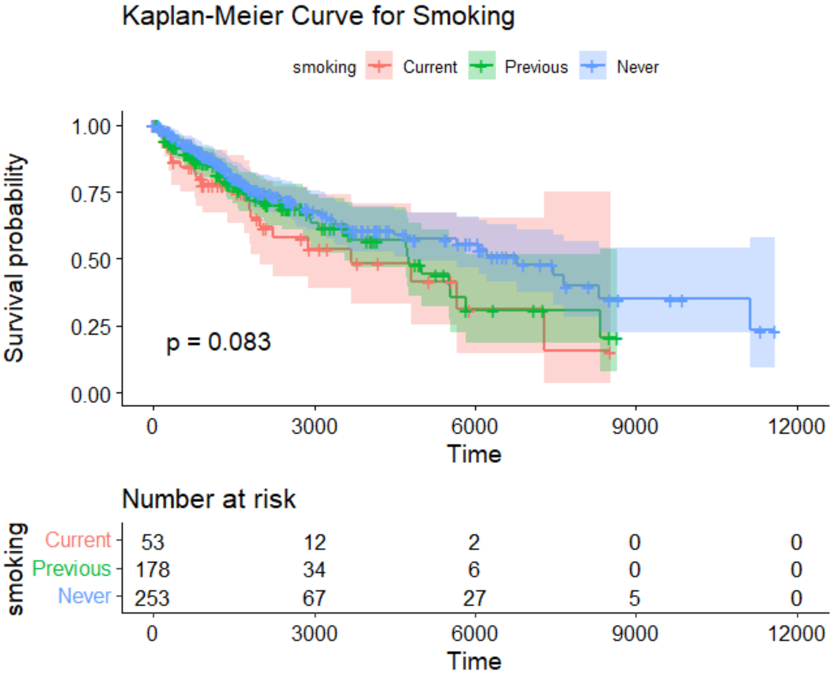


Figure 5.5: Kaplan-Meier curve of smoking

The result of plotting the survival curve on a log scale is fig 5.6,it can be seen that the three curves are approximately parallel.

In order to see if the difference between groups are statistically significant, a log-rank test was performed on the survival probability of the three groups. The results are shown in figure 5.7.

The p-value in the log-rank test is 0.1, and with a alpha level of 0.1, there is evidence against the null hypothesis that the survival curves are the same. The hazard and survival of this three groups are moderately significant different.

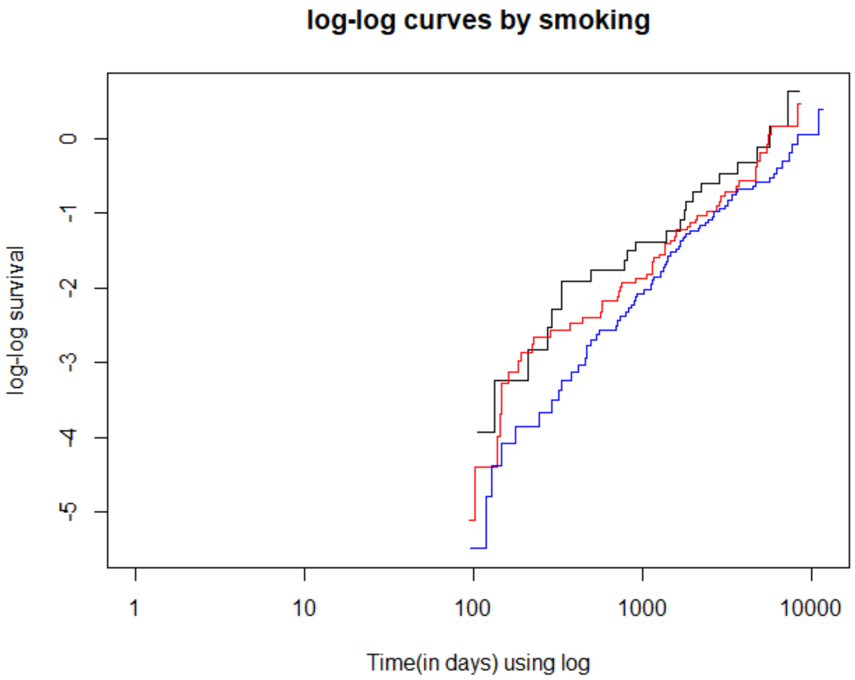
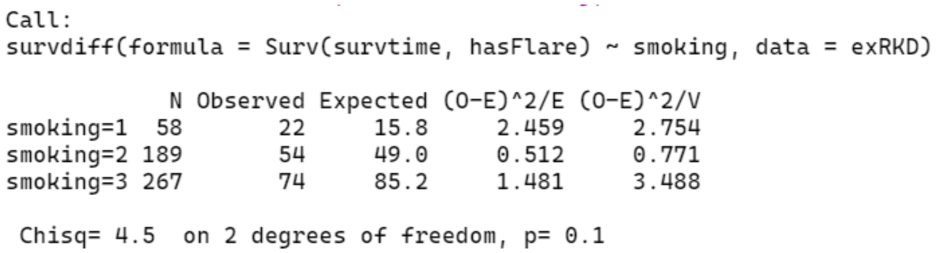


Figure 5.6: Log-Log curve of smoking



```
Call:
survdiff(formula = Surv(survtime, hasFlare) ~ smoking, data = exRKD)

             N Observed Expected (O-E)^2/E (O-E)^2/V
smoking=1  58       22     15.8     2.459     2.754
smoking=2 189       54     49.0     0.512     0.771
smoking=3 267       74     85.2     1.481     3.488

 Chisq= 4.5  on 2 degrees of freedom, p= 0.1
```

Figure 5.7: Log rank test of smoking

5.3 Cox Regression Model results

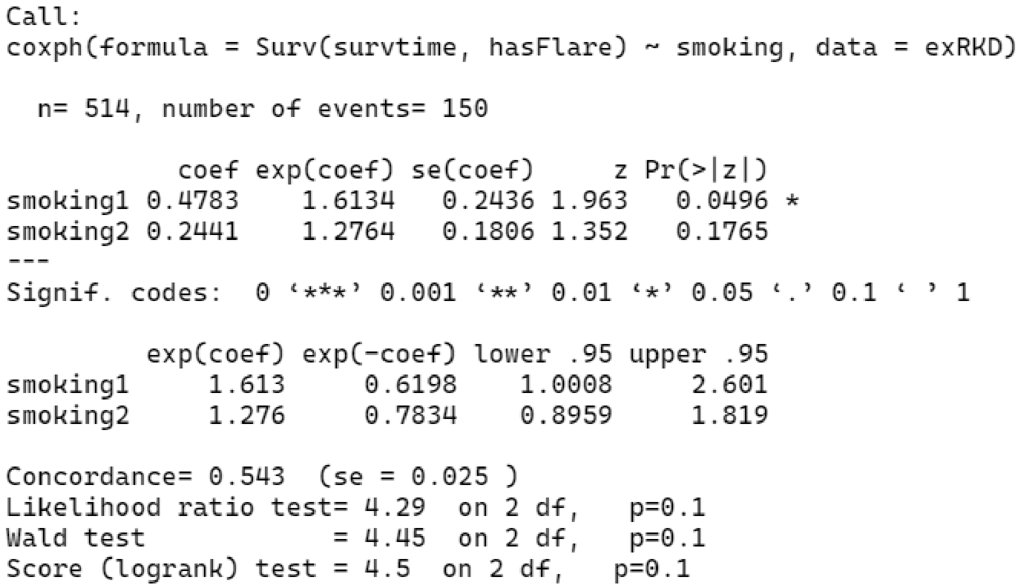
A cox regression was performed on smoking and a proportional hazard assumption test was performed on it. the model results are shown in figure 5.8. The hazard ratio (HR) is the ratio of the hazard rates corresponding to the conditions described by the two levels of the explanatory variable.The result $exp(coef)$ represents the hazard ratio. For the risk ratio,

- $HR = 1$, No effect.
- $HR > 1$, the variable will increase in hazard.
- $HR < 1$, the variable will decrease in hazard.

Hazard means the flare of AAV in this research.

As we can see from the results in fig 5.8, both smoking1 and smoking2 have a hazard ratio of larger than 1, and based on the p-value, we can conclude that the probability of flare for patients who are current smokers is 161%(95% CI,1.0008-2.601) of the probability of flare for patients who never smoke. We have 95% confidence that the true value of this hazard ratio is between 1.0008 and 2.601 (or 100.08% to 260.1%), but our best guess is that it's 161%.

Although the results for previous smokers represent a 127.6% probability of AAV relapse compared to patients never smoke. And the hazard ratio is between 0.8959 and 1.819(or 89.59% to 181.9%).Although the p-value results are not significant enough and its hazard ratio is not consistent at the 95% confidence interval to allow us to conclude from this that the probability of AAV relapse is higher in ex-smokers than in patients who never smoke, this result still has implications for our study, as patients who quit smoking may have a higher risk of relapse compared to never smokers and a relatively lower risk of relapse compared to current smokers.

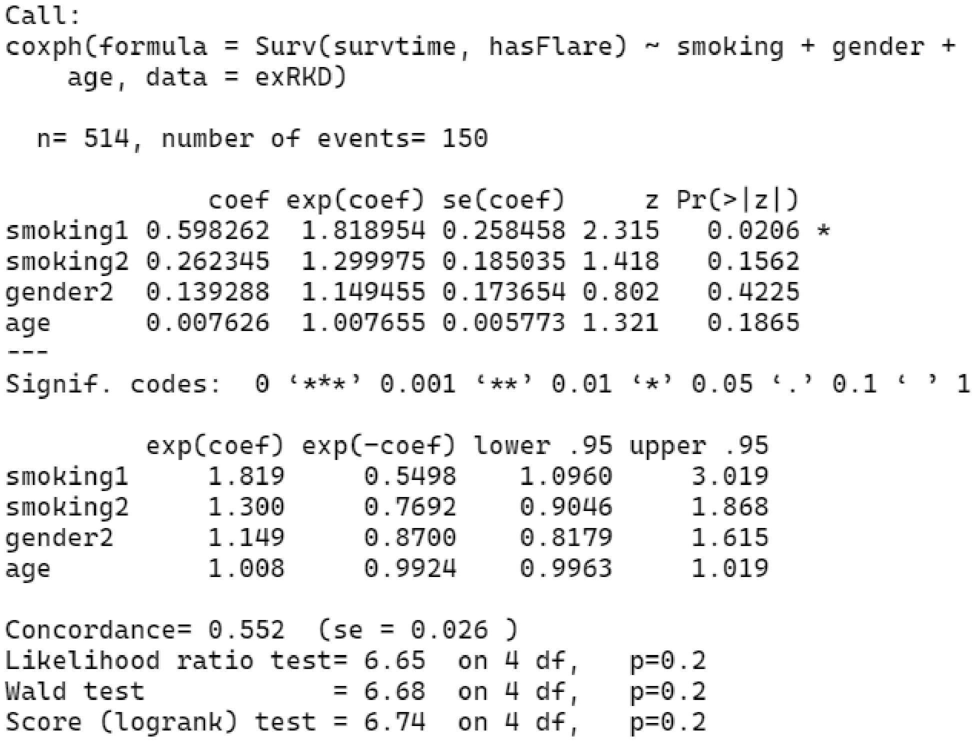


```
Call:
coxph(formula = Surv(survtime, hasFlare) ~ smoking, data = exRKD)

  n= 514, number of events= 150

           coef exp(coef) se(coef)     z Pr(>|z|)
smoking1 0.4783    1.6134   0.2436 1.963   0.0496 *
smoking2 0.2441    1.2764   0.1806 1.352   0.1765
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
smoking1     1.613     0.6198    1.0008     2.601
smoking2     1.276     0.7834    0.8959     1.819

Concordance= 0.543  (se = 0.025 )
Likelihood ratio test= 4.29  on 2 df,   p=0.1
Wald test            = 4.45  on 2 df,   p=0.1
Score (logrank) test = 4.5  on 2 df,   p=0.1
```

Figure 5.8: Cox regression model of smoking

In order to verify the stability of the results, we add some other covariates, age, gender to the cox regression. And then we get the results in figure 5.9.From the results, we can see that the model results are consistent with those in figure 5.8, despite the addition of the new covariates, which again demonstrates that probability of flares for AAV patients who are current smokers is 184.7%(95% CI,1.112-3.068) of patients who never smoke.At a 95% confidence interval, the probability of relapse for current smokers patients is 111% to 307% of patients never smoke.

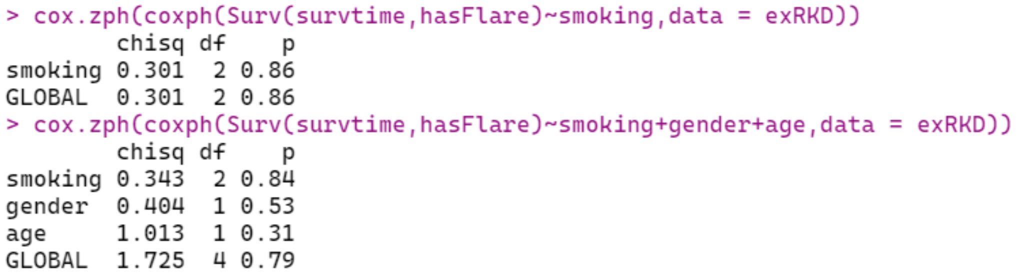
Although the results for previous smokers represent a 131.4% probability of AAV relapse compared to patients never smoke. And the hazard ratio is between 0.914 and 1.898(or 91.4% to 189.8%).Although the p-value results are not small enough and its hazard ratio is not consistent at the 95% confidence interval to allow us to conclude from this that the probability of AAV relapse is higher in ex-smokers than in patients who never smoke, this result still has implications for our study, as patients who quit smoking may have a higher risk of relapse compared to never smokers and a relatively lower risk of relapse compared to current smokers.



```
Call:
coxph(formula = Surv(survtime, hasFlare) ~ smoking + gender +
    age, data = exRKD)

  n= 514, number of events= 150

             coef exp(coef) se(coef)     z Pr(>|z|)
smoking1 0.598262  1.818954 0.258458 2.315   0.0206 *
smoking2 0.262345  1.299975 0.185035 1.418   0.1562
gender2  0.139288  1.149455 0.173654 0.802   0.4225
age      0.007626  1.007655 0.005773 1.321   0.1865
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
smoking1     1.819     0.5498    1.0960     3.019
smoking2     1.300     0.7692    0.9046     1.868
gender2      1.149     0.8700    0.8179     1.615
age          1.008     0.9924    0.9963     1.019

Concordance= 0.552  (se = 0.026 )
Likelihood ratio test= 6.65  on 4 df,   p=0.2
Wald test            = 6.68  on 4 df,   p=0.2
Score (logrank) test = 6.74  on 4 df,   p=0.2
```

Figure 5.9: Cox regression model of smoking with covariates

Although the previous log-log curve showed that the survival curves of different groups of smokers were approximately parallel, the results were not rigorous enough. Therefore, the cox regression model was tested again for proportional hazard assumption with Schoenfeld residuals test. If global value and every covariate p-value are all larger than 0.05, then the proportional hazard assumption holds true. The test result is in figure 5.10,from the result, we can see neither the cox regression model with only smoking as a variable nor the cox regression model with other covariates added was statistically significant for each covariate tested, nor was the global test statistically significant. Therefore, we can assume proportional hazards hold.



```
> cox.zph(coxph(Surv(survtime,hasFlare)~smoking,data = exRKD))
        chisq df    p
smoking 0.301  2 0.86
GLOBAL  0.301  2 0.86
> cox.zph(coxph(Surv(survtime,hasFlare)~smoking+gender+age,data = exRKD))
        chisq df    p
smoking 0.343  2 0.84
gender  0.404  1 0.53
age     1.013  1 0.31
GLOBAL  1.725  4 0.79
```

Figure 5.10: Proportional Hazard Test for cox regression models

5.4 Accelerated Failure Time model results

Although cox regression is the most widely used survival analysis model and through the above experiments we have verified the PH assumption and obtained the results of cox regression, we can still validate our previous model with the accelerated failure time model.

Based on a comparison of AIC of different parameter AFT models, the difference between the AIC values of the different models is very small, based on this result, we chose the most commonly used Weibull Accelerated Failure Time model in this experiment. After performing the same operation on the data and running it on the Weibull AFT model, and we get the results in figure 5.11. The exp(coef) represents the Time Ratio(TR) in AFT model.For Time Ratio,

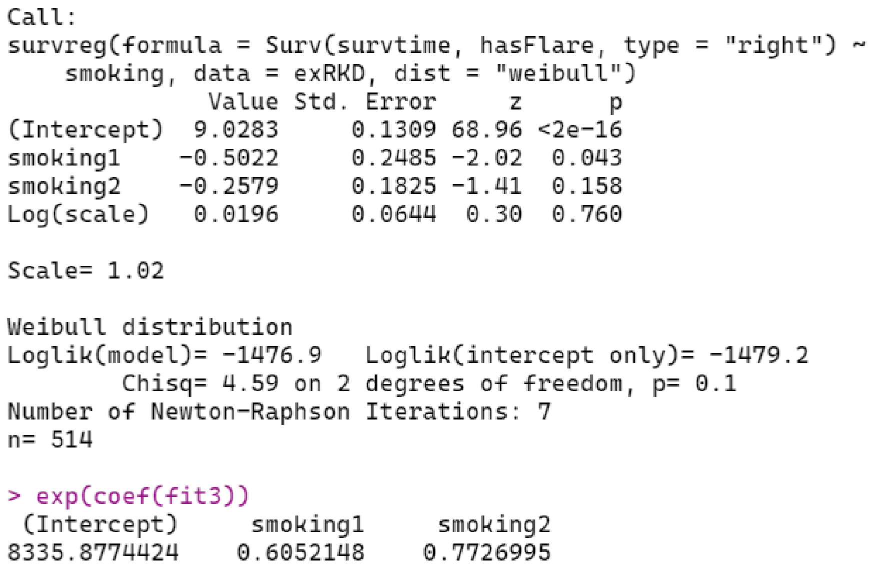
- $TR > 1$, the variable would make the survival time increase.
- $TR < 1$, the variable would make the survival time decrease.

From the results, we can see that both the time ratio of smoking1 and smoking2 are less than 1. This means patients who are current smokers and have quit smoking have shorter survival time(this means the time to flare in AAV) than patients who never smoked. We can conclude that patients who are current smokers have a 39% higher probability(95% IC) of flare than patients never smoked.

The result that patients who have quit smoking have a 23% probability of flare than patients never smoked is not strongly significant. The result is consistent with cox regression model without covariates.

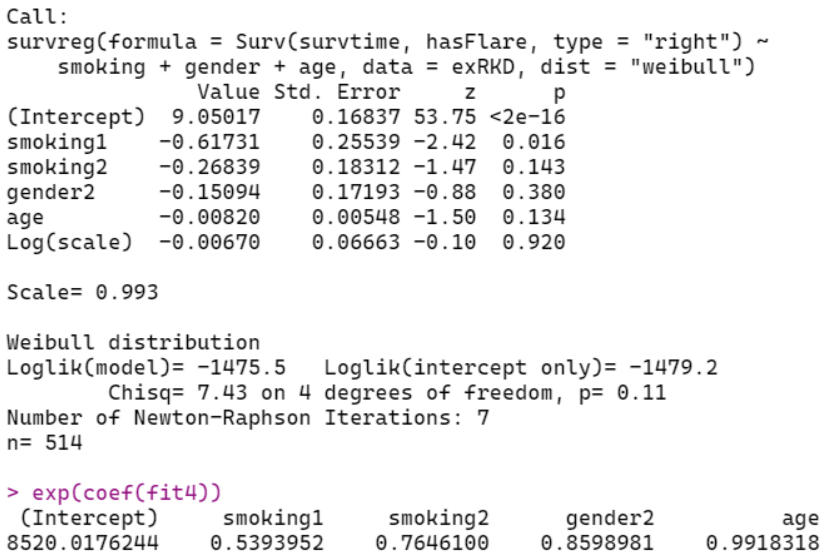
Add covariates age, gender, and maintenance treatment to the AFT model, we get results in figure 5.12. From the results, we can see that both the time ratio of smoking1 and smoking2 are less than 1. This means patients who are current smokers and have quit smoking have shorter survival time(this means the time to flare in AAV) than patients who never smoked. We can conclude that patients who are current smokers have a 46.1% higher probability(95% IC) of flare than patients never smoked.

The result that patients who have quit smoking have a 23.6% probability of flare than patients never smoked is not strongly significant. The result is consistent with AFT model without covariates. And this result is also consistent with cox regression model with covariates.



```
Call:
survreg(formula = Surv(survtime, hasFlare, type = "right") ~
    smoking, data = exRKD, dist = "weibull")
              Value Std. Error     z      p
(Intercept)  9.0283     0.1309 68.96 <2e-16
smoking1    -0.5022     0.2485 -2.02  0.043
smoking2    -0.2579     0.1825 -1.41  0.158
Log(scale)   0.0196     0.0644  0.30  0.760

Scale= 1.02

Weibull distribution
Loglik(model)= -1476.9   Loglik(intercept only)= -1479.2
        Chisq= 4.59 on 2 degrees of freedom, p= 0.1
Number of Newton-Raphson Iterations: 7
n= 514

> exp(coef(fit3))
 (Intercept)     smoking1     smoking2
8335.8774424    0.6052148    0.7726995
```

Figure 5.11: Weibull AFT model of smoking

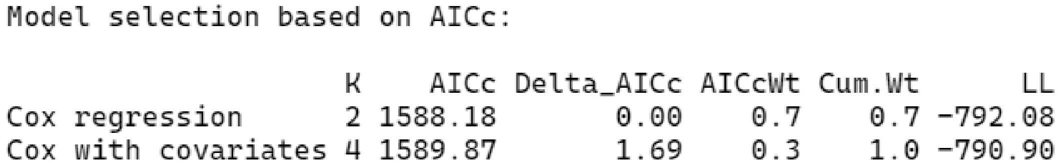


```
Call:
survreg(formula = Surv(survtime, hasFlare, type = "right") ~
    smoking + gender + age, data = exRKD, dist = "weibull")
               Value Std. Error     z      p
(Intercept)  9.05017    0.16837 53.75 <2e-16
smoking1    -0.61731    0.25539 -2.42  0.016
smoking2    -0.26839    0.18312 -1.47  0.143
gender2     -0.15094    0.17193 -0.88  0.380
age         -0.00820    0.00548 -1.50  0.134
Log(scale)  -0.00670    0.06663 -0.10  0.920

Scale= 0.993

Weibull distribution
Loglik(model)= -1475.5   Loglik(intercept only)= -1479.2
        Chisq= 7.43 on 4 degrees of freedom, p= 0.11
Number of Newton-Raphson Iterations: 7
n= 514

> exp(coef(fit4))
 (Intercept)     smoking1     smoking2      gender2          age
8520.0176244    0.5393952    0.7646100    0.8598981    0.9918318
```

Figure 5.12: Weibull AFT model of smoking with covariates

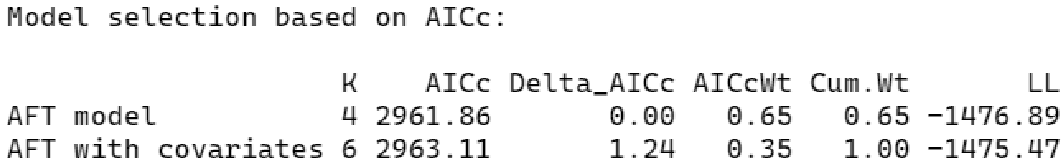
5.5 Comparison of Cox regression and Accelerated Failure Time model

Consistent experimental results were obtained for the cox regression model and the Accelerated Failure Time model, but it was difficult to directly determine which model was better. We therefore compare the two models here using the Akaike Information Criterion (AIC),a measure of the goodness of fit of an estimated statistical model, used to select the optimal model with the knowledge that a smaller AIC suggests a higher likelihood(Habibi et al. (2018)).



```
Model selection based on AICc:

                      K    AICc Delta_AICc AICcWt Cum.Wt      LL
Cox regression        2 1588.18       0.00    0.7    0.7 -792.08
Cox with covariates   4 1589.87       1.69    0.3    1.0 -790.90
```

Figure 5.13: Weibull AFT model of smoking with covariates



```
Model selection based on AICc:

                      K    AICc Delta_AICc AICcWt Cum.Wt       LL
AFT model             4 2961.86       0.00   0.65   0.65 -1476.89
AFT with covariates   6 2963.11       1.24   0.35   1.00 -1475.47
```

Figure 5.14: Weibull AFT model of smoking with covariates

The results show that the cox regression model outperforms the AFT model, while the univariate cox regression model has better results than the model with covariates, but the difference in AIC between the two models is very small and not significant. Because the AIC statistic assigns a lower penalty to more complicated models, it may place a greater amount of weight on how well a model performs on the training dataset, and as a result, choose more complicated models. So although the cox regression model performs better in the AIC assessment, it may also be the more complex model.

5.6 Conclusions

McDermott and Xiaoqing Fu (2020) has found both former and current smokers were more likely to have AAV than never smokers. And Yamaguchi and Ando (2018) found that smoking increased the risk of relapse in Japanese MPA patients in a dose-dependent manner, even after adjusting for clinically relevant factors. From the findings and experimental results of this paper, we can conclude that smoking does cause a higher probability of recurrence of Ireland's ANCA-AAV patients. In order to prevent disease relapse in patients with ANCA-AAV, patients who have been diagnosed with ANCA-AAV should be advised to stop smoking in order to prevent ANCA-AAV recurrence.

Chapter 6

Reflections & Future Work

6.1 Reflections

- The data in this paper are somewhat limited. The conclusion that smoking triggers relapse in patients with ANCA-AAV might be valid only for the population the data refers to (i.e. patients in the region of Ireland), and data from more areas are needed to derive more general conclusions.
- Although the best guess results of the model suggest that ex-smokers have a higher risk of relapse compared to never-smokers and a lower risk of relapse compared to patients who are still smoking, the results are only somewhat speculative as the model results are not significant enough.
- The discussion of smoking as an environmental factor in this paper is relatively homogeneous and does not take into account the interaction of other medical factors including treatment modalities, preventive treatment modalities and other factors on patient relapse with smoking.
- This article only discusses the fact that smoking leads to an increased probability of recurrence, but not the causes that contribute to this outcome.

6.2 Future Work

- Additional data could be attempted to explore whether patients who quit smoking have a higher probability of relapse compared to those who have never smoked, and whether they have a lower probability of relapse compared to those who are still smoking.

- When it comes to the construction of predictive models from data, machine learning techniques are quite popular and in the spotlight. In creating non-linear correlations, they offer a significant advantage over typical statistical methods. However, the application of machine learning to survival analysis is currently uncommon due to the censored data.

  In Zupan et al. (2000) 's work, the authors present a model that enables classification methods - including machine learning classifiers - to be used for survival analysis and validate the utility of the model using a case study on pre- and post-operative recurrence prediction in prostate cancer, discovering that by incorporating such weighting techniques, machine learning tools sit alongside modern statistical methods and may provide further insight into relationships within the model data by inducing symmetrized regression. A number of machine learning algorithms have been tailored to deal with the complex challenges that arise in survival data and other real-world dataWang et al. (2019).

  Therefore, attempts could be made to add machine learning to survival analysis studies to improve experimental results, and attempts could be made to construct predictive models.

- It is possible to add treatment modalities to existing models, to study their interactions and to obtain more convincing models.

Bibliography

A, R. K. and Anders, Hans-Joachim, e. a. (2020). ANCA-associated vasculitis. *Nature Reviews Disease Primers*, 6(1):71.

AR, K., HJ, A., and Basu N, e. a. (2020). Anca-associated vasculitis. *Nature reviews. Disease primers.*

Assouline, S. and Shen Li, e. a. (2020). The conditional survival analysis of relapsed dlbcl after autologous transplant: a subgroup analysis of ly.12 and coral. *Blood Advances*, 4:2011–2017.

Basu, N. (2021). Mind the gap: Balancing remission and risk of relapse in anca-associated vasculitis. *RHEUMATOLOGY.*

Bellera, C. A., MacGrogan, G., Debled, M., De Lara, C. T., Brouste, V., and Mathoulin-Pélissier, S. (2010). Variables with time-varying effects and the cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC medical research methodology*, 10(1):1–12.

Berden, A. and Göçeroğlu, Arda, e. a. (2012). Diagnosis and management of anca associated vasculitis. *BMJ*, 344.

Bewick, V. and Cheek, Liz, e. a. (2004). Statistics review 12: Survival analysis. *Critical Care*, 8(4).

Bland, J. M. and Altman, D. G. (1998). Survival probabilities (the kaplan-meier method). *Bmj*, 317(7172):1572–1580.

Bland, J. M. and Altman, D. G. (2004). The logrank test. *Bmj*, 328(7447):1073.

Broström, G. (2021). *Event history analysis with R.* Chapman and Hall/CRC.

CARTER, J. C. and BLACKMORE, E, e. a. (2004). Relapse in anorexia nervosa: a survival analysis. *Psychological Medicine*, 34(4):671–679.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Faruk, A. (2018). The comparison of proportional hazards and accelerated failure time models in analyzing the first birth interval survival data. In *Journal of Physics: Conference Series*, volume 974, page 012008. IOP Publishing.

Habibi, D., Rafiei, M., Chehrei, A., Shayan, Z., and Tafaqodi, S. (2018). Comparison of survival models for analyzing prognostic factors in gastric cancer patients. *Asian Pacific journal of cancer prevention: APJCP*, 19(3):749.

Holle, J. U. and Gross, W. L. (2013). Treatment of anca-associated vasculitides (aav). *Autoimmunity Reviews*, 12(4):483–486.

Jones, R. B. and Hiemstra, Thomas F, e. a. (2019). Mycophenolate mofetil versus cyclophosphamide for remission induction in anca-associated vasculitis: a randomised, non-inferiority trial. *Annals of the Rheumatic Diseases*, 78(3):399–405.

Kalra, Aakshi, e. a. (2016). The basics of kaplan–meier estimate. *Journal of the Practice of Cardiovascular Sciences*, 2(3):187.

Karangizi, A. H. and Harper, L. (2021). Small vessel vasculitides. *Medicine*, 50.

King, C. and Druce, Katie L, e. a. (2021). Predicting relapse in anti-neutrophil cytoplasmic antibody-associated vasculitis: a systematic review and meta-analysis. *Rheumatology Advances in Practice*, 5.

Lahiri, S. (2018). Chapter 12 - survival probabilities from 5-year cumulative life table survival ratios (tx+5/tx): Some innovative methodological investigations. In Srinivasa Rao, A. S. and Rao, C., editors, *Integrated Population Biology and Modeling, Part A*, volume 39 of *Handbook of Statistics*, pages 481–542. Elsevier.

Little, M. (2019). Rare kidney disease registry and biobank protocol version 8.

Liu, X. (2012). *Survival analysis: models and applications.* John Wiley & Sons.

McDermott, G. and Xiaoqing Fu, e. a. (2020). Association of cigarette smoking with antineutrophil cytoplasmic antibody-associated vasculitis. *JAMA internal medicine.*

Mercuzot, C. and Simon Letertre, e. a. (2021). Comorbidities and health-related quality of life in patients with antineutrophil cytoplasmic antibody (anca) - associated vasculitis. *Autoimmunity Reviews*, 20(1):102708.

Monach, P. A. and Warner, R. L. e. a. (2022). Serum biomarkers of disease activity in longitudinal assessment of patients with anca-associated vasculitis. *ACR Open Rheumatology*, 4(2):168–176.

Ningning, F. (2021). Clinical analysis of anca associated vasculitis and refractory nephrotic syndrome in single center.

Orbe, J., Ferreira, E., and Núñez-Antón, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in medicine*, 21(22):3493–3510.

Paramalingam, S., Raymond, W., and Chanakya Sharma, e. a. (2019). Disease flares, damage accrual and survival in anca-associated vasculitis in western australia. *International Journal of Clinical*, 14.

Salama, A. D. (2019). Relapse in anti-neutrophil cytoplasm antibody (anca)-associated vasculitis. *Kidney international reports*.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.

Sedgwick, P. and Joekes, K. (2013). Kaplan-meier survival curves: interpretation and communication of risk. *Bmj*, 347.

Sestelo, M. (2017). A short course on survival analysis applied to the financial industry.

Stoica, P. and Selen, Y. (2004). Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

WHO (1977). *Manual of mortality analysis: a manual on methods of analysis of national mortality statistics for public health purposes*. World Health Organization.

Yamaguchi, M. and Ando, Masahiko, e. a. (2018). Smoking is a risk factor for relapse of antimyeloperoxidase antibodies-associated vasculitis. *Journal of clinical rheumatology : practical reports on rheumatic amp; musculoskeletal diseases*, 24(7):361—367.

Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R., and Bratko, I. (2000). Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, 20(1):59–75.

Appendix

Ethics Declaration

’RKD data is saved, if this is necessary, on a password encrypted device. RKD data is not emailed to yourself or anyone or stored on cloud services without being encrypted. RKD data is not shared with anyone else or discussed with anyone else. Demonstrations, reports and publications about the project will not display actual individual level patient data