Stockholms universitet

Probabilities instead of hazards in survival analysis

Pär Villner

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:10
Matematisk statistik
Juni 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2020:10**
http://www.math.su.se

Probabilities instead of hazards in survival analysis

Pär Villner*

June 2020

Abstract

Survival analysis is a set of tools used to analyze time between events. It is often used in biostatistics, economics and actuarial mathematics. One of the most fundamental concepts of survival analysis is the hazard function, which tells us the rate at which events are happening at a given moment. The hazard function is widely used, for example, to describe the difference in effectiveness of different types of medical treatment. In this thesis, it is pointed out that the hazard function is difficult to interpret: there are many examples of researchers mistaking the hazard function for a probability, even though it is only a rate. There is, however, an alternative to the hazard function: the instantaneous geometric rate. This measure describes the actual probability of the event happening, and it is therefore easy to interpret. A regression method for the instantaneous geometric rate is presented and then applied to real biostatistical data.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: par.villner@gmail.com. Supervisor: Ola G Hössjer, Taras Bodnar, Matteo Bottai.

Acknowledgements

This thesis is based on the research of Matteo Bottai at Karolinska institutet. As external supervisor, he helped me structure the thesis, understand the methods and, not least, use the software package Stata. Therefore, it is no exaggeration to say that this thesis would not have been written if it were not for Matteo Bottai. Giola Santoni at Karolinska Institutet provided me with the dataset I have used, and she kindly answered questions regarding the data. Ola G Hössjer and Taras Bodnar were my supervisors at Stockholm University. During our weekly meeting, Taras and Ola helped me with everything from understanding basic concepts in survival analysis to getting the mathematical notation right. Their comments on two drafts were much more detailed than I could have hoped for. Moreover, their warm and encouraging attitude made this thesis a pleasure to write.

Contents

1 Introduction

Survival analysis is a set of statistical tools used for measuring the time from a start-point to the occurence of an event. It is commonly practiced in biostatistics, but also in economics, actuarial mathematics and many other fields. [1]

A central part of survival analysis is the ability to calculate the probability that an event will happen over a certain time period, or the instantaneous rate of it happening. Usually the probability of an event over a time period is measured in terms of an incidence rate, and the instantaneous rate is measured in terms of the so-called hazard function. This thesis will present an alternative approach, where the geometric rate is used instead of the incidence rate and the instantaneous geometric rate is used instead of the hazard rate.

It will be argued that at least for some types of survival data, the alternative approach is superior. The thesis is largely theoretical, but the presented methods are also applied to real biostatistical data.

2 Survival analysis

2.1 Introduction

Survival analysis is a set of tools used to analyze the time up to and in between events. In medical research, survival analysis can be used to determine how long it takes for cancer patients to die after surgery. In a factory setting, it can be used to determine how long it takes until a new machine starts malfunctioning.

One of the most influential tools from this field is the so-called "Cox semiparametric regression model", which is used to model the rate at which events are happening. (Cox 1972) In the Cox model, a baseline rate, which is usually unknown and which varies with time, is multiplied by an exponential function of covariates. Thereby, the difference in event rate between subjects with different characteristics can be estimated. Some prefer to have a model where the baseline rate is modelled with a parametric function. For example, Royston and Parmar (2002) proposed to model the baseline rate with restricted cubic splines, thereby making it possible to see not only the difference in rate between subjects with different characteristiscs, but also to see what the actual rate is.

Another example of a method used in survival analysis is the "Accelerated survival time model", where the proportion of studied subjects estimated to have experienced the event of interest at a particular time is modelled with a regression model of logarithmic time. (Klein and Moeschberger 2003, Chapter 2)

The occurence of events over time is commonly studied with ordinary linear regression models. With survival data, that method tends to be implausible. The main reason is that in survival data, we usually cannot observe the event of interest for

[1] A good introductory textbook in survival analysis that explains most of the methods and topics mentioned in this thesis in an intuitive way is Klein and Moeschberger (2003). A more advanced book is Aalen, et al. (2008).

all subjects, because observations are censored or truncated – or both. These terms will now be explained.

2.2 Censoring and trunctation

The following is based on Klein and Moeschberger (2003, Chapter 3).

Censoring means that the event of interest is unobserved for some subjects. There is left and right censoring. Left censoring means that it is known that the event happened *prior* to some time-point t, but not when. For example, in a study of at what age children have the flu for the first time, there may be a child who says she has had the flu, but she cannot remember when. In the case of right censoring, it is unknown when the event occurred *after* a certain time-point. This usually happens because when a study ends, not all subjects have experienced the event of interest. It is then known that it did *not* happen until time t, but not when it happened after that. Another common reason for right censoring is that subjects die from other reasons than the event of interest.[2]

Survival data may suffer from only left censoring, only right censoring, or both left and right censoring.

Truncation means that only individuals who experience events during a certain timespan are observed, meaning that truncated subjects give no information. There is left and right truncation. Left truncation means that only individuals who have not yet experienced the event of interest can enter the study. For example, in a study of how long patients survive after they have been diagnosed with cancer, it may be that patients become part of the study one week after the diagnosis has been given. Patients who die within one week will not be part of the study, meaning that the study will not have information about short-lived patients. This is different from left censoring, because in that situation, there is *some* information about the survival time.

Right truncation is when only people who experience the event of interest before a given time point can enter the study. This is common in retrospective studies. For example, if we want to know long it took for Swedes who went skiing in the Alps during the first week of February to get ill with Corona virus, we may collect the relevant data for all persons who had been diagnosed with the virus before March 1. We will not have information about patients for whom it took longer to get ill.

In the data analysis part of this thesis, data that is right-censored, not left-censored, will be studied. Also, the analysed data is not truncated.

2.2.1 The problem with censored and truncated data

What makes censoring and truncation problematic with regards to ordinary linear regression, is that the mean value and variance for the survival time and death rate

[2] In many cases, there are several events that may happen to a person, but if one of them happens, the others cannot happen. For example, you may die from cancer or from a traffic-accident, but if you have died from one of them, you cannot die from the other. If we are interested in several such events, we say that they are "competing events". There are methods within survival analysis to study the occurrence of several competing events, but this thesis is focused on the single-event case.

Table 1: Example of survival data (not based on real data). The ID column contains ID numbers to differentiate subjects. The Censored column contains 0 for subjects that were not censored, and 1 for subjects that were censored. The Entered and Left columns contain the time unit in which subjects entered and left the study, and the difference is displayed in the Time in study column.

ID	Entered	Left	Time in study	Censored
100	1	4	3	0
101	0	5	5	1
102	3	10	7	0
103	4	7	3	0

cannot be estimated, since there is no information about the time to event for some subjects. This is problematic for ordinary linear regression with regression variable Y and explanatory variables x, since it assumes that $E(Y|x)$ and $Var(Y|x)$ can be estimated. Yet, while being imperfect, truncated and censored data contain valuable information, and the tools of survival analysis have been developed to exploit that information.

2.2.2 The non-information assumption

The most commonly used methods in survival analysis assume that censoring is non-informative, meaning that the occurrence of censoring is not related to the probability of the event of interest. This means that in a cancer study, it must not be the case that the patients who drop out of a study (and hence are censored) tend to be the sickest patients, who will soon die.

2.2.3 Example of survival data

Table 1 shows an example of what right-censored survival data might look like. There is information on when subjects entered and left the study, if they were censored, and how long in total they spent in the study.

2.3 Important measures in survival analysis

The following section is based on Klein and Moeschberger (2003, Chapter 2).

Three important concepts in survival analysis are the survival function, the hazard function, and the cumulative hazard function. Let t denote a specific (non-negative) point in time. T is then a stochastic variable, which denotes the point in time when an event occurs: the distribution of T is $T \sim F(.)$ where $F(t) = P(T \leq t)$. A subject "dies" when the event of interest happens to the subject. That a subject "survives until t" then means that the event has not happened to that subject at time t. If T is a continuous random variable, the probability density function of T is defined as:

$$f(t) = \lim_{h \to 0} \frac{P(t \leq T \leq t+h)}{h}$$

2.3.1 Survival function

The survival function $S(t)$ is the probability that a randomly chosen individual survives until t or longer. On a continuous time-scale:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du.$$

This implies that

$$f(t) = -\frac{dS(t)}{dt}.$$

Moreover, $S(t)$ is monotonously decreasing with $S(0) = 1$, and $S(t) \to 0$ as $t \to \infty$.

A common way of estimating the survival function is with the Kaplan-Meier method, which is non-parametric. Let t_i denote a point in time where one event happened and let n_i denote the number of individuals known to have survived until just before t_i. This way, right-censoring is taken into account because censored subjects are removed from the risk group in the denominator. Then the survival function for time point t is estimated by the Kaplan-Meier estimator:

$$\hat{S}(t) = \Pi_{i:t_i \leq t} \left(1 - \frac{1}{n_i}\right).$$

The Kaplan-Meier estimator is often illustrated with a diagram with time on the x axis and proportion of survivors on the y axis. The diagram has the form of a stair going downwards. It is easy to see what proportion of subjects are estimated to survive by time t, and at what time a certain proportion of subjects survive. Since the Kaplan-Meier estimator is non-parametric, it can only differ between categorical differences between groups of subjects, and it is difficult to have many categories in the same figure. There is also no easy way of summarizing information, as can be done with a slope coefficient in regression analysis.

2.3.2 The hazard function

The hazard function or hazard rate $h(t)$ [3] is the rate at which an individual dies immediately after t, given that the individual survived until t:

$$h(t) = \lim_{\triangle t \to 0} \frac{P(t < T \leq t + \triangle t | T > t)}{\triangle t}$$

Since we are dividing a probability by the width of a time interval, the result is a rate of event occurrence per (each infinitely small) time unit. This means that the hazard

[3] Another common notation for the hazard function/rate is $\lambda(t)$

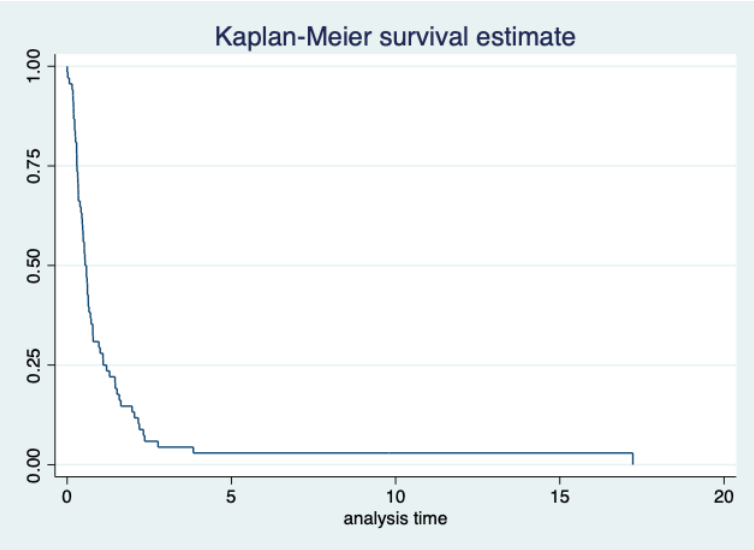

Figure 1: Plot of a Kaplan-Meier estimator for 68 patients with esophageal cancer. The data is a subset of the data analysed later in the thesis, to make the staircase shape more visible than it is with all observations. The data is right-censored.

function is *not* a probability; however, the larger the hazard function the greater the risk. It is always positive, but it has no upper bound, unlike a probability. (Sutradhar & Austin, 2018)

However, for a small $\Delta t > 0$, we have that $h(t)\Delta t$ approximates the probability of death in the interval $[t, t+\Delta t]$ given that the subject has survived until t.

The hazard function can be expressed in terms of the probability density function and the survival function (for proof, see the appendix):

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d\log(S(t))}{dt}. \tag{1}$$

Figure 2 shows a plot of a hazard function for real data. It is notable that the hazard function is above 1 at the beginning of the studied period, indicating that we are not dealing with a probability.

2.3.3 Cumulative hazard

The cumulative hazard function $H(t)$ describes the accumulate hazard of dying until time t, i.e. $H(t) = \int_0^t h(u)du$. Since $h(t)$ is not a probability, $H(t)$ does *not* describe the probability of death by time t; however, the larger the $H(t)$ the bigger the risk of death by time t. Since $h(t) = -d\log(S(t))/dt$, it is true that

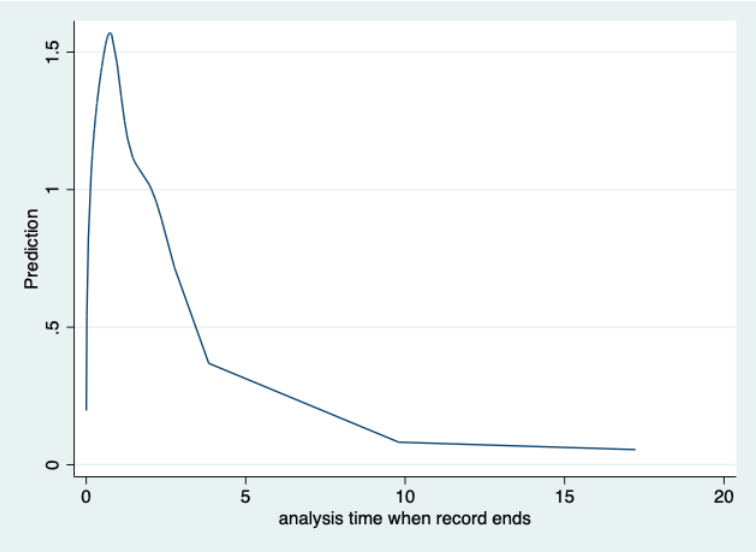

Figure 2: Plot of a hazard function. The data is the same as the one used in Figure 1.

$$H(t) = \int_0^t h(u)\,du = -\log(S(t))$$

which implies that

$$S(t) = \exp(-H(t)).$$

The Cumulative hazard can be estimated with the non-parametric Nelson-Aalen estimator. As with the Kaplan-Meier estimator of the Survival function, let n_i denote the number of people at risk just prior to time t_i where t_i is a timepoint where one event happened. Then the Nelson-Allen estimator is:

$$\hat{H}(t) = \sum_{i:t_j<t} \frac{1}{n_i}.$$

2.4 Regression models in survival analysis

The following section is based on Klein and Moeschberger (2003, Chapter 2.6).

It is often of interest to know whether the survival time of different subjects depends on certain features of the subjects. For example, cancer patients may have different survival times depending on age, gender, blood pressure or cancer treatment. If one is studying only a few characteristics, non-parametric methods can be used. For example, the difference between patients who take medicine A and B can be measured by creating Kaplan-Meier estimators of the survival function, and

compare the graphs. However, the difference is then difficult to express in a compact manner, and if there are additional features of the patient that are of interest, non-parametric measures require additional stratification.

A more promising strategy is to create parametric regression models. Most common is to estimate the hazard function with regression methods. Two common ways of doing this is with the "Cox semi-parametric model" and with fully parametric models. In biostatical journal papers, regression models of the hazard function is used *frequently* to determine the effect of various treatments, and the Cox model is used very often. [4]

2.4.1 Cox semiparametric model

This section is based on Klein and Moeschberger (2003, Chapter 8).

In the Cox semiparametric model, the hazard function based on the covariates $x = (x_1, \ldots, x_p)'$ has the form:

$$h(t|x) = h_0(t)\exp(\beta' x).$$

Here, $h_0(t)$ is a "baseline hazard" function that only depends on time, and it is assumed to be the same for all subjects and $\beta = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients. If all covariates have value 0, the exponential expression has value 1. Therefore, the baseline hazard can be seen as the hazard function for the control-group in a study, where the control-group has covariate vector $x = (0, \ldots, 0)$.

If we to the covariate vector x add a vector $a = (1, 0, 0, .., 0)'$ with 1 in the first place and 0 in all other places , the hazard function changes with a factor of $\exp(\beta_1)$.

The relative effect of adding vector a to x, can be measured with the hazard ratio. Using the same a as before:

$$\frac{h(t|x+a)}{h(t|x)} = \frac{h_0(t)\exp(\beta'(x+a))}{h_0(t)\exp(\beta' x)} = \frac{\exp(\beta'(x+a))}{\exp(\beta' x)} = \exp(\beta^t(x+a-x)) = \exp(\beta_1)$$

According to this model, the hazard ratio is constant over time, since t is not part of the exponential function. This is called "the proportional hazards assumption". It implies that, for example, the difference in death rate between patients who received different types of cancer surgeries is the same right after surgery as it is five years after surgery. Whether this assumption is true or not can be tested with using so-called Schoenfeld residuals. If the propoportional hazard assumption turns out not to be true, one may extend the model to include interactions between the β parameters and time.

The baseline hazard, on the other hand, is only a function of time and not of any parameters, making it difficult to predict at a given time. Researchers often focus on the hazard ratio and consider the baseline hazard as a nuisance parameter of little interest. This approach, however, makes it difficult to know how important the

[4] Some recent examples of journal papers where the hazard function plays a central part in evaluating treatment effects: Cao, B. et al. (2020); Zeiser, R. et al (2020); Kausik, R. et al. (2020); and Petrie, M. et al. (2020).

difference in hazard ratio is: the higher the baseline hazard, the bigger the benefits or problems of a hazard ratio different from 1. One way of solving this problem is to use a parametric baseline hazard function, such as a Weibull distribution, or to model the baseline hazard with restricted cubic splines, which will be described later in the thesis. (Royston and Lambert 2011, chapters 4 & 7)

2.4.2 Fully parametric regression models

It is possible to use common parametric distribution functions to model the hazard function. The advantage of using such functions is that parameters are easy to estimate, and the model is easy to interpret. For example, if T follows an exponential distribution, i.e. $T \sim \text{Exp}(\lambda)$, then $h(t) = \lambda$. The downside of using fully parametric models is that they rely on assumptions which are unreasonable in many cases. For example, if $T \sim \text{Exp}(\lambda)$, it is assumed that T has the same distribution for all observations, and that the expected future survival time is the same, no matter the age of the subject. This is an unreasonable assumption when people are studied. Another common model for survival data is the Weibull distribution, $T \sim \text{Weib}(a, b)$ which has the hazard function $h(t) = abt^{a-1}$. However, this model is based on the assumption that the baseline hazard is monotonically increasing or decreasing, which is not always reasonable to assume.[5] On the contrary, it is often unreasonable to assume that the baseline hazard has this property.

3 The difficulty of interpreting the hazard function

This section is based on Sutradhar and Austin (2018).

For a patient who suffers from a fatal disease, it is surely of interest to know how much longer she can expect to survive, given that she has survived until present. As we saw previously, the hazard function tells us something about this. More precisely, $h(t)\Delta t$ where $\Delta t > 0$ is small, approximates the risk of death in $[t, t + \Delta t]$ given that the subject survived until t. Therefore, it is not strange that the hazard function is used so often.

Yet, there seems to be confusion regarding the hazard function. As we also saw above, the hazard function denotes a rate, not a probability or risk. In spite of this, it is often mistaken for a risk or probability. There are several examples of journal papers in biostatistics making this mistake. [6]

The hazard ratio tells us *something* about the risk though, namely the direction of the risk. So if a hazard ratio is 3, the risk is higher in one group than in the other,

[5] Whether it is increasing or decreasing depends on if $a > 1$ or $a < 1$. If $a = 1$, the Weibull distribution collapses into an exponential distribution.

[6] For example, Berge et. al. (2016) write "the fully adjusted model showed 73% increased risk of incident IHD among cases with health anxiety", referring to table where the Hazard ratio is 1.73. Similarly, Chan et al (2007) write: "Regular aspirin use was associated with a multivariate relative risk of colorectal cancer of 0.73" even though it is a Cox hazard ratio that has been used. Emdin et al (2015) write: "Overall, a 20 mm Hg higher than usual systolic blood pressure was associated with a 63% higher risk of peripheral arterial disease (hazard ratio 1.63, 95% confidence interval 1.59 to 1.66)...". Jordanova et al (2008) write: "Univariate and multivariate Cox proportional models were used to determine the hazard ratio (HR) that represents the relative risk of death among patients in the different groups."

although the risk is not three times higher. As Strudhadar & Austin (2018) shows, to translate the hazard ratio into a relative risk, we need to know both the hazard function and the baseline hazard. Since it is common in survival analysis in general and biostatistics in particular to only focus on the hazard ratio and not caring about the baseline hazard, it is often difficult to translate the hazard ratio into a risk ratio.

This is not only a theoretical problem. It can have large impact. If we compare cancer treatment A and B, and see that the hazard ratio is 1.73 and draw the conclusion that the risk of death is 73% for cancer patients having a treatment A rather then treatment B, we may see this as a strong reason to change cancer treatment for patients at large. This could be costly and complicated. If it turned out that the real risk ratio was a lot smaller, it might have been better to spend the time and resources on something else.

4 Estimating the probability of events

In the previous section, we saw that the hazard function is often mistaken for a probability when it is in fact a rate. This misunderstanding could be overcome by education. But the fact that researchers so often seem to be interested in the instantaneous probability of an event can also be seen as a reason to develop a measure of the instantaneous probability. That is what we will try to do now.

In the next two sections, we will distinguish between two ways of measuring the rate at which events happen. The incidence rate is good for measuring the rate at which events happen, when we are interested in events that can happen several times to a single person. As we will see, the hazard rate is in fact the instantaneous incidence rate.

If we are interested in events that can only happen once to a person, we should use a different measure: the geometric rate. This rate can in fact be seen as a probability of an event happening to a person over a given time period. We will then introduce the instantaneous geometric rate, which is the instantaneous probability of an event happening.

Then, having established the geometric rate as a good measure for survival data, we will have a look at a regression model for the geometric rate.

4.1 Incidence rate

According to Bottai (2015) the common way in biostatistics to measure the occurrence of events over a time interval is the incidence rate, which is the average number of events per person-time. For example, in order to find out at what rate people have headaches, we may study 5 persons during 2 days. Assume that there are 4 headaches during the 2 days. The incidence rate is then the total number of headaches divided by the number of days multiplied with the number of persons:

$$\frac{\text{Headaches}}{\text{Days} \cdot \text{Persons}} = \frac{4}{2 \cdot 5} = \frac{2}{5} = 0.4$$

According to Bottai (2015), the incidence rate is a natural way of measuring occurrence of events that can happen several times to the same person (such as headaches). However, biostatisticians often study events that one person can only experience once, such as death. For that purpose, it may be better to use the geometric rate.[7]

4.2 Geometric rate

The geometric rate is the average probability of event per unit of time over a given time interval, for a subject. The geometric rate is denoted $g(t_1, t_2)$ where t_1 is the starting point of the interval and t_2 is the end point of the interval. We calculate the geometric rate by studying a number of people over a period of $t_2 - t_1$ time units. We count how many people *did not* experience the event of interest, and calculate

$$1 - \left[\frac{\text{Persons with no event}}{\text{Total number of persons}}\right]^{1/(t_2-t_1)}.$$

The ratio is the proportion of persons who did not experience the event out of the total amount of people in the study. The $(t_2 - t_1)$'th square root of it, is the average proportion of people who did not experience the event per time unit. Subtracting this number from 1 gives us the geometric rate: the average probability of experiencing the event in one time unit. The geometric rate can be expressed in terms of the survival function $S(t)$. If the start point is 0 and the end point is t, we have

$$g(0, t) = 1 - S(t)^{1/t}$$

More generally, the geometric rate between any two time points is

$$g(t_1, t_2) = 1 - \left[\frac{S(t_2)}{S(t_1)}\right]^{1/(t_2-t_1)}.$$

To contrast, the incidence rate over time interval $[0, t]$ can be expressed in terms of the survival function in a different way[8]:

$$\frac{1 - S(t)}{\int_0^t S(u)du}$$

This makes it clear that the incidence rate and the geometric rate are different. The following example illustrates the difference a bit more, and shows that the geometric rate is a better measure of the probability of certain events than the incidence rate.

[7] Of course, headaches can also be studied with the geometric rate, described in the next section. That would be reasonable if we are interested in analysing the probability that an individual will have at least one headache. Then we would only count the first headache that an individual has, and not bother counting the other headaches the individual has. That way, we could use the geometric rate to study the probability of having at least one headache.

[8] As we saw in the previous section, the incidence rate makes sense for events that can happen several times to a person, but here we are using it to study events that can happen only once to a person. The formula below makes no sense for the headache example in the previous section, for example, as we have no clue if each headache was experienced by different individuals or if it was the same person who had all the headaches

4.3 Geometric rate vs Incidence rate: an example

Bottai (2015, page 2700) gives the following example: "100 subjects are followed up for 2 days. At day 1 the number of survivors is 20, and at day 2 it is 16. The geometric daily mortality rates in day 1 and day 2 are $1-20/100 = 0.80$ and $1-16/20 = 0.20$, respectively. The average daily geometric mortality rate over the two days is $1-(16/100)^{1/2} = 0.60$. The latter represents the average probability, or risk, of dying in a day's time: if the 100 subjects died at a constant daily rate of 0.60, then $100 \cdot (1-0.60)^2 = 16$ would be expected to survive day 2, which is equal to the observed number of survivors. Conversely, the incidence rate, $84/(80 \cdot 1 + 20 \cdot 2) = 0.70$ deaths per person-day, does not represent the risk for a person to die in a day: if the 100 subjects died at a constant daily rate of 0.70, then $100 \cdot (1-0.70)^2 = 9$ would be expected to survive on day 2, which differs from the observed number of survivors."

4.4 Instantaneous geometric rate

The geometric rate is the average event probability over a time unit. One may also be interested in instantaneous probability of event per unit of time. This is measured by the instantaneous geometric rate at time t, which is defined as the geometric rate over an infinitely small time interval:

$$g(t) = \lim_{h \to 0} 1 - \left[\frac{S(t+h)}{S(t)} \right]^{1/h}$$

Since $g(t)$ is a probability, it is always between 0 and 1 – unlike the hazard function, that could take any positive number. By expanding this formula, using formula (1), a connection to the hazard function is seen :

$$\begin{aligned} \lim_{h \to 0} 1 - \left[\frac{S(t+h)}{S(t)} \right]^{1/h} &= \lim_{h \to 0} 1 - \exp\left[\frac{\log S(t+h) - \log S(t)}{h} \right] \\ &= 1 - \exp\left(\frac{d \log S(t)}{dt} \right) \\ &= 1 - \exp\left(\frac{-f(t)}{S(t)} \right) \\ &= 1 - \exp[-h(t)] \end{aligned}$$

As a contrast, the instantaneous incidence rate is in fact the hazard function, if we again use formula (1):

$$\lim_{h \to 0} \frac{S(t) - S(t+h)}{\int_t^{t+h} S(u)\,du} = \frac{f(t)}{S(t)} = h(t).$$

From the above, the following conclusions seem reasonable:

1. When studying events that can only happen once to a person, the geometric rate seems to be preferable to the incidence rate.

2. When studying the instantaneous probability of an event that can only happen once to a person, it therefore seems inappropriate to use the hazard function. The instantaneous geometric rate seems better, and while it is related to the hazard function, it is somewhat different.

3. The instantanous geometric rate is a probability, which makes it easy to interpret. The hazard function is harder to interpret. This is also a reason for preferring the instantaneous geometric rate when we are studying events that can only happen once to a person.

4.5 Regression models for the geometric rate

Despite the fact that the geometric rate has these nice properties, and despite its frequent use in fields such as demographics, it is seldomly used in biostatistics. That is probably because there have been no natural regression models for the geometric rate. (Bottai 2015, page 2700-1) In the next section follows the description of a regression model based on the geometric rate, which has recently been suggested by Discacciati and Bottai (2017).

4.5.1 Proportional instantaneous geometric rate and odds

The regression model for the instantaneous proportional geometric rate is

$$g_i(t|x_i) = g_0(t)exp(\beta' x_i),$$

where x_i is the covariate of individual i. The baseline geometric rate $g_0(t)$ depends on time. If we assume that x is a single covariate and β is a single parameter, then if 1 is added to x, the proportional geometric rate is changed by a factor of $exp(\beta)$. An alternative and equivalent interpretation is that the covariate parameter β is a geometric rate ratio, since if we again assume that x is a single covariate and β a single parameter

$$\frac{g_i(t|x_i+1)}{g_i(t|x_i)} = \frac{g_0(t)\exp(\beta'(x_i+1))}{g_0(t)\exp(\beta' x_i)} = \exp()$$

A related concept is the Proportional instantaneous geometric odds, defined as

$$\frac{g_i(t|x_i)}{1-g_i(t|x_i)} = \frac{g_0(t)}{1-g_0(t)}\exp(\beta' x_i).$$

Here $g_0(t)/(1-g_0(t))$ is a baseline geometric odds ratio, and $\exp(\beta)$ is the multiplicative instantaneous geometric odds ratio increase in x_i, since

$$\frac{\frac{g(t|x_i+1)}{1-g(t|x_i+1)}}{\frac{g(t|x_i)}{1-g(t|x_i)}} = \frac{\frac{g_0(t)}{1-g_0(t)}\exp(\beta'(x_i+1))}{\frac{g_0(t)}{1-g_0(t)}\exp(\beta' x_i)} = \exp(\beta).$$

4.5.2 Strategy for estimating parameters in geometric rate

We can estimate the parameters in the geometric rate and odds models by using the relation that exists between the instantaneous geometric proportional rate and the hazard function:

$$g(t|x_i) = 1-\exp(-h(t|x_i)).$$

Based on this relation, since it is true that

$$\log(g(t|x_i)) = \log(g_o(t)) + \beta' x_i$$

then it must also be true that

$$\log(1 - \exp(-h(t|x_i)) = s(t|\gamma) + \beta' x_i$$

where $s(t|\gamma)$ is assumed to be a parametric function.

An idea then is to find $s(t|\gamma)$ and β by expressing the instantaneous geometric rate and odds in terms of the hazard function, and use it in statistical software packages for finding regression models of the hazard function.

This is done by a user-written Stata command called Stpreg.[9] It is based on the command Stgenreg, which is used to estimate parametric hazard models. In order to understand how Stpreg works, we will proceed in the following steps:

1. Describe the general method for estimating hazard regression models.
2. Describe how Stgenreg estimates parametric hazard models.
3. Describe how Stpreg modifies Stgenreg to generate the instantaneous geometric rate and odds.

4.6 Estimating the instantaneous geometric rate

4.6.1 Estimating the hazard function

Estimation of regression parameters of the hazard function is based on maximizing the likelihood function. However, what the likelihood function looks like and how it is maximized, depends on what hazard model is being used and how the data is trunctated and censored. In this thesis, only right-censored data is analyzed, so the estimation methods described below holds for right-censored data.

For a subject i who dies at time t, her contribution to the likelihood is the probability density function

$$L_i = f(t_i) = S(t_i)h(t_i).$$

while for a subject who is censored at time t, we only know that she was alive until at least time t, meaning that her contribution to the likelihood is

$$L_i = S(t_i).$$

In order to turn this into a nice formula, let each observation i have three values: t_i, δ_i and x_i. Here, t_i is the amount of time from the start point until the event, δ_i is an indicator variable for whether the event occurred before censoring (i.e. $\delta_i = 1$ if the event was observed and 0 if the observation was censored). x_i is a covariate vector. Based on the reasoning above, the likelihood function should be such that

[9] This command was written by Matteo Bottai, Andrea Discaciatti and Giola Santoni. Stgenreg was written by Michael J Crowther and Paul C Lambert.

censored observations contribute with their observed survival time, while uncensored observations contribute with their time to the event. The following likelihood for n individuals serves this purpose:

$$L(\beta|t) = \Pi_{i=1}^{n} \left[f(t_i|\beta, x_i)^{\delta_i} S(t_i|\beta, x_i)^{1-\delta_i} \right], \text{ for} \{t_1, \ldots, t_n\}$$

For a fully parametric hazard model, the parameters are estimated by finding the maximum likelihood. For simple models, such as a model that follows the exponential distribution, this can be done analytically. In the case of more complicated models, numerical methods are used.

The instantaneous geometric rate is a fully parametric model, and hence the above likelihood equation will be used. However, it is worth mentioning that for the Cox semiparametric model, a slightly different method is used: one where we maximize the so-called "partial likelihood". (Aalen et al. 2008, page 135)

4.6.2 Estimating the hazard function with the Stata command Stgenreg

This section is based on Crowther and Lambert (2013).

Stgenreg maximizes parameters of the fully-parametric hazard function based on the likelihood function for survival data described above. Since $h(t_i|\beta, x_i) = f(t_i|\beta, x_i)/S(t_i|\beta, x_i)$, it is true that

$$L(\beta|t) = \Pi_{i=1}^{n} \left[h(t_i|\beta, x_i)^{\delta_i} S(t_i|\beta, x_i) \right].$$

The log likelihood is then

$$l(\beta|t) = \Sigma_{i=1}^{n} \left[\delta_i \log h(t_i|\beta, x_i) + \log S(t_i|\beta, x_i) \right].$$

Since $H(t_i|\beta, x_i) = -\log(S(t_i|\beta, x_i))$, the log likelihood can be written

$$l(\beta|t) = \Sigma_{i=1}^{n} \left[\delta_i \log h(t_i|\beta, x_i) - H(t_i|\beta, x_i) \right].$$

The likelihood function is now formulated only in terms of the hazard function. Stgenreg estimates parameters in a hazard regression model by maximizing the likelihood function through the Newton-Raphton method. To allow for complex parametric hazard models, the baseline hazard may be modelled with functions that cannot be integrated analytically. Therefore, Gaussian quadrature is used to evaluate the integral $H(t_i|\beta, x_i) = \int_0^{t_i} h(s|\beta, x_i) ds$.

The baseline hazard can be modelled in many ways. One way is with restricted cubic splines. It is this method that Stpreg uses to model the baseline geometric rate/odds, and therefore we will only focus on this method.

In the following section, the Newton-Raphson method, Gaussian quadrature and restricted cubic splines will be described.

4.6.3 Gaussian quadrature

This section is based on Crowther and Lambert (2013, section 2.2) and Givens and Hoeting (2013, Chapter 5.3).

Gaussian quadrature is a numerical technique for calculating the integral of a function $g(.)$. The main idea is that for integration limits $[-1,1]$, it is possible to approximate any function with an polynomial; the larger the degree, the closer the approximation. A polynomial of degree $2N-1$ can then be integrated exactly by evaluating the polynomial in N points $z_i, i = 1,2,..,N$ and multiply evaluations with weights $w_i, i = 1,2,...,N$. Since the polynomial was an approximation of the function $g(.)$, we have that:

$$\int_{-1}^{1} g(z)dz \approx \sum_{i=1}^{N} w_i g(z_i)$$

There are several possible weight functions w_i that can be combined with ways of choosing Z_i values. Stgenreg chooses the z_i values based on so-called Legendre polynomials, which is a set of orthogonal polynomials that were originally developed for another purpose. The z_i values that Stgenreg chooses are the roots of Legendre polynomials of the same order as the polynomial that we wish to integrate. For example, let's say we wish to integrate a linear function. Then it can be approximated by a polynomial of order 1. Since $1 = 2 \cdot 1 - 1$, we use the Legendre polynomial of order order 1. It is z_i, and the root of this polynomial is 0. If we instead think that our function is better approximated by a polynomial of order 3 then we use the Legendre polynomial of order 2 (because $3 = 2 \cdot 2 - 1$). This polynomial is $(3z^2 - 1)/2$, with roots $\pm 1/\sqrt{3}$.

The weights w_i are given by the formula $2/[(1 - z_i^2)(P'_N(z_i))^2]$, where $P_N(.)$ is the Legendre polynomial of order N. From this formula, we get that for polynomials of degree 1, have have only one weight: $w_1 = 2$. For quadratic polynomials, there are two weights with values $w_1 = 1$ and $w_2 = 1$. The default number of nodes in Stgenreg is 15 nodes.

When the integration limits are not $[-1,1]$ but $[a,b]$, a transformation is needed. Let $z = mt + c$. Assume that $a = m(-1) + c$ and $b = m(1) + c$. Solving for m and c gives

$$m = \frac{b-a}{2} \quad c = \frac{b+a}{2}$$

meaning that $z = (b-a)t/2 + (b+a)/2$ and $dz = [(b-a)/2]dt$.

This results in the following formula for getting an approximate integration of a function with integration limits $[a,b]$:

$$\int_a^b g(z)dz = \int_{-1}^{1} g\left(\frac{b-a}{2}z + \frac{a+b}{2}\right)\frac{b-a}{2}dz \approx \frac{b-a}{2}\sum_{i=1}^{N} w_i g\left(\frac{a-b}{2}z_i + \frac{a+b}{2}\right).$$

4.6.4 Newton-Raphson method

This section is based on Held and Bové (2014, Appendix C.1.3).

Newton-Raphson's method is a numerical method for finding (among other things) the maximum likelihood estimate of a likelihood function. To begin with, assume that parameter we want to estimate is the single parameter θ. Let $U(\theta)$ be the Score

function of θ, that is $U(\theta) = \frac{d \log L(\theta,x)}{d\theta}$. Also, let $I(\theta)$ be the Fischer information of θ, that is $-\frac{d^2 \log L(\theta,x)}{d\theta^2}$. The $(k+1)$'th approximation of θ is

$$\theta^{k+1} = \theta^k - \frac{U(\theta^k)}{I(\theta^k)}.$$

In case of a multivariate $\theta = (\theta_1, \ldots, \theta_p)$, we have that for

$$U(\theta) = \left(\frac{\partial \log L(\theta)}{\partial \theta_1}, ..., \frac{\partial \log L(\theta)}{\partial \theta_k} \right)$$

and for

$$I(\theta) = \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\theta) \right),\ 1 \leq i, j \leq k$$

the $(k+1)'th$ approximation of θ becomes

$$\theta^{k+1} = \theta^k - I(\theta^k)^{-1} U(\theta^k).$$

For a given start value $\theta^{(0)}$, recursion stops either after a pre-determined number of n steps, or when convergence occurs, i.e. when $|\theta^{(k+1)} - \theta^{(k)}| < \alpha$ for some pre-determined number $\alpha > 0$.

4.6.5 Restricted cubic splines

Based on Royston and Lambert (2011, Chapter 4); and Heinzl and Kaider (1996).

The baseline geometric rate/odds may have many different shapes. Since the risk of event may go up and down over time, so may the baseline geometric rate/odds. Yet it is desirable to model the baseline function with a function that is smooth without using unnecessarily many degrees of freedom. Restricted cubic splines (RCS) is a way of achieving this.

RCS is performed by distributing m knots $t_1 < t_2 < ... < t_m$ over the x axis. The two knots of each pair t_i, t_{i+1} are connected by a cubic function. Call this cubic function $R_i(.)$ for interval i. A function $R_{i+1}(.)$ shares its first knot with $R_i(.)$ and its second knot with $R_{i+2}(.)$.

The idea is that all the $R_i(.)$ functions will together form a smooth curve. To achieve this, some restrictions are imposed:

1. If t_i is the knot that $R_i(.)$ and $R_{i+1}(.)$ share, then $R_i(t_i) = R_{i+1}(t_i)$. This ensures that curve is continuous, without gaps or jumps.

2. The first and second derivative of the two functions that share a knot, must be identical at the knots. $R'_i(t_i) = R'_{i+1}(t_i)$ and $R''_i(t_i) = R''_{i+1}(t_i)$. This is to ensure that the curve that is formed by the functions is smooth.

3. The functions in the intervals before the first knot and after the last knot, are linear. $R''_0(t_1) = R''_n(t_m) = 0$

If we use the knots $t_1 < t_2 < ... < t_k$ and let t denote event time, the formula of RCS is:

$$R(t) = \beta_o + \beta_1 t + \sum_{j=1}^{k-2} \theta_j R_j(t)$$

where $R_1(t), .., R_{k-2}(t)$ are cubic terms of the form

$$R_j(t) = (t - t_j)_+^3 - \frac{(t - t_{k-1})_+^3 (t_k - t_j)}{t_k - t_{k-1}} + \frac{(t - t_k)_+^3 (t_{k-1} - t_j)}{t_k - t_{k-1}}$$

for $j = 1, ..., (k-2)$, if we let $x_+ = \max(0, x)$. According to this formula, $R(t)$ is linear when $t < t_1$ or $t > t_k$, meaning that $R(t)$ is "linear in the tails". From the formula, it is also seen that at least three knots are needed in order to have a cubic term (two knots would be mean that only $\beta_0 + \beta_1 t$ is ever evaluated). It is common to use 3-5 knots, which are placed (roughly) uniformly at certain percentiles of the distribution of t values. For example, 3 knots are usually placed at the $5, 50, 95$ percentiles, and 4 knots at the $5, 25, 75, 95$ percentiles.

If k is the number of knots, an RCS has $k-1$ degrees of freedom, in addition to the intercept. If RCS is used in a model that already has an intercept, this means that the RCS adds $k-1$ degrees of freedom.

When formulating a restricted cubic spline function in Stgenreg (and hence also in Stpreg), we specify the degrees of freedom. For k degrees of freedom, $k-1$ internal knots are placed uniformly on percentiles over the event times, and 2 boundary knots are positioned at percentile 0 and 100. (The fact that no observation-time can be smaller than the first knot or bigger than the last knot, should mean that the whole restricted cubic spline model is cubic, and not linear in the tails.) For example, if we specify 5 degrees of freedom, 2 boundary knots are defined at percentiles 0 and 100, and 4 internal knots are placed at the percentiles $20, 40, 60, 80$. (Crowther and Lambert 2013, Page 9.)

4.6.6 How Stpreg works

Previously, it was shown that the log likelihood function for the survival data can be expressed in terms of the hazard function:

$$l(\beta|t) = \Sigma_{i=1}^{n} \delta_i \log h(t_i|\beta, x_i) - H(t_i|\beta, x_i).$$

It has also been shown that the instantaneous geometric rate can be expressed in terms of the hazard function.

$$g(t|\beta, x_i) = 1 - \exp[-h(t|\beta, x_i)].$$

By inverting this relationship, the hazard function can be expressed as a function of the instantaneous geometric rate:

$$h(t|\beta, x_i) = -\log[1 - g(t|\beta, x_i)].$$

Similarly, if we are after the instantaneous geometric odds, the hazard can be expressed as

$$h(t|\beta, x_i) = \log\left[1 + \frac{g(t|\beta, x_i)}{1 - g(t|\beta, x_i)}\right].$$

This way, Stpreg gives maximum likelihood estimates the parameters. The variance of these estimates come from the observed Fisher information matrix.

5 Data analysis

Now that we have established that the instantaneous geometric rate is a good idea conceptually, it is time to use it in practice. In the remainder of the thesis, we will study data from patients who underwent surgery for esophageal cancer, which is cancer in the esophagus: a long tube that connects the throat and stomach. Esophageal cancer is the eighth most common form of cancer in the world, annually affecting 482,300 persons. (Markar et al. 2016, page 1528).

The data consists of 800 observations, that were randomly selected from all patients that underwent surgery for for esophageal cancer in Sweden during the period from $1987-01-02$ to $2010-12-14$. The follow-up for survival lasted until $2016-01-01$.

This means that we consider surgery to be the start event and death (from cancer) to be the event of interest. The data is right-censored, since some patients had not died at the end of the study. The data is not left-censored: that would mean that we had patients in our data that were dead before some date, but we did not know exactly what date. Neither is the event left-truncated: that would mean that patients became part of the study some time after the surgery, and patients who had died by then did not enter the study at all. Lastly, the data is not right-truncated: that would be the case if we for example based the study on records of patients who died before a certain date, and hence did not get information on patients who performed surgery during the period of interest but who had not died before the specified date.

For each observation we have information about when the patient joined the investigation and when they left. This information is in dates. We also have information for each patient whether he or she died from the cancer or was censored. 677 of the 800 patients died from cancer. The mean survival time after surgery was 1313.77 days, which is around 3.6 years. The longest survival time was 8740 days, roughly 24.5 years. The shortest survival time was for a patient who died on the day of operation.

Each patient also has values in categories that we will use as covariates in our analysis later on:

Age: A float number that indicates the age in years. The age of the patient is their age on the day of surgery. The mean age of the patients is 65.4 years. The oldest one was 88.6 and the youngest 19.5 years.

Sex: A binary variable, taking the number 1 for females and 0 for males. There are

Table 2: Correlation matrix between the four covariates of the data set.

	Stage	Resection	Sex	Age
Stage	1	-	-	-
Resection	0.31	1	-	-
Sex	-0.05	-0.05	1	-
Age	-0.07	-0.05	0.04	1

187 female and 613 male patients in the data set.

Stage: An integer from 1 to 4. It is a classification of the tumor according to the TNM classification for tumors. A larger value means that the tumor was bigger and more spread in the body, so that a value of 1 means that the tumor was relatively small, whereas a value of 4 means that that tumor was large and that tumors had spread in the body.[10]

The distribution of patients in the different stage categories are the following:

1: 189 patients. (23.6 %)
2 : 274 patients. (34.3 %)
3: 269 patients. (33.6 %)
4: 68 patients. (8.5 %)

Resection: When a tumor is resected (removed), some cancer cells may remain at the margin of the resection area, so that cancer cells remain in the body. The Resection variable is binary and it gives information about this. It takes value 0 if no cancer cells were left after surgery, and 1 if cancer cells remained. There are 132 patients with resection value 1, and 668 patients with resection value 0 in the data set.

5.1 Characterising the data

Before fitting a model to data, some preparatory analysis was done to detect possible collinearity between covariates that may affect the analysis. There were no obvious associations between any of the variables except for Stage and Resection, which were positively correlated. The higher the value of Stage, the higher the value of Resection, meaning that patients with a more developed cancer tended to have cancer cells left after surgery. Table 2 shows a correlation matrix for all variables

The pattern is illustrated in Figure 3, where the proportion of patients with Resection=1 is shown for the four different values of Stage. The impact of this for the present investigation is that the effect of Resection may be difficult to distinguish from the effect of Stage.

It was also investigated whether the values of the variables have changed over time. Both the number of surgeries per year and the survival time of patients has increased slightly over the years. (For an illustration, see Figure 1 of the appendix)

[10] TNM is the mostly widely used classification system for cancer tumors. It consist of five stages: 0, I, II, III and IV. In our dataset, 1 stands for 0-I, 2 for II, 3 for III and 4 for IV. (National Cancer Institute 2015).

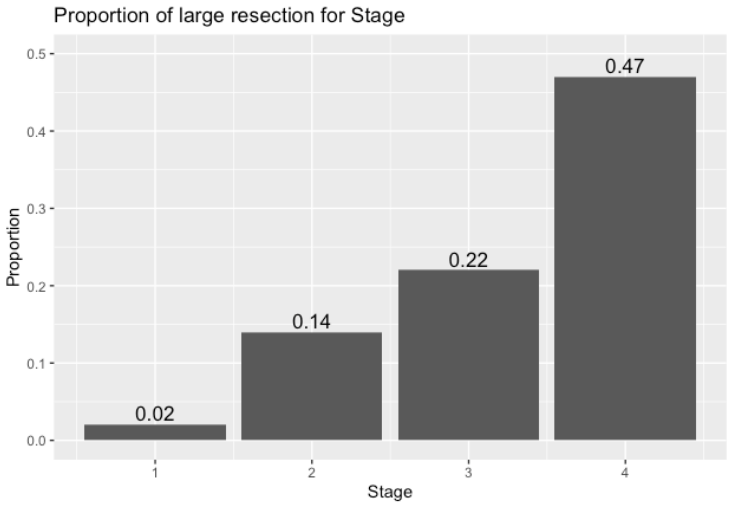

Figure 3: Proportion of surgeries where not all cancer cells were resected, for each level of Stage.

The proportion of patients with Stage=1 has increased while the proportion of patients with Stage=3 has decreased. This means that patients have tended to have less developed tumors in recent years. (For an illustration, see Figure 2 of the appendix)

For Resection, Resection=0 has increased whereas Resection=1 has decreased. This means that in recent years, it has been more common to remove all cancer cells during surgery. (For an illustration, see Figure 3 of the appendix)

The proportion of patients at various ages and sexes has been quite constant over the years. (For an illustation, see Figure 4 and 5 of the appendix)

What these changes over time means for the investigation, is that to some extent, the effect of the values of Stage and Resection on survival time may be difficult to distinguish from the general development of cancer treatment over time.

Finally, Figure 4 shows a Kaplan-Meier curve, estimating the survival function for the patients. We see that most patients did not survive long after surgery: around 50% were dead after 2 years. After about 4 years, the curve starts to plan out. Yet, only around 25% remain alive after 5 years.

5.2 Data preparation

The statistical software package Stata was used to perform the survival analysis. The data set is declared as survival data. The dates for entering and exiting the study are changed to the number of days after a Stata's "start date" 1960-01-01 . That way, the difference between when patients entered and left the investegation can be calculated. This is done next, generating the number of days a patient survived after

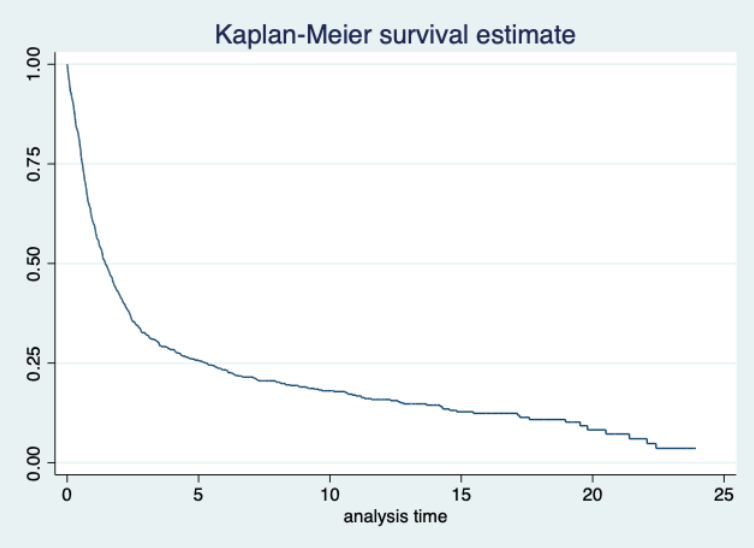

Figure 4: Kaplan-Meier estimator of the survival function for the data set, with time given in years.

surgery. One patient died on the day of operation, the lifetime of that patient was set to 0.5 since it does not make sense that a patients dies at time 0 and we lack more detailed information about exactly when the patient died in relation to they surgery.

Sex and resection are both binary variables, so they are treated as factors with 0 as the baseline level.

Stage is a little more complicated. It has four levels. We may treat these as factors. That seems reasonable if the effect of the different stages are very different from each other. One would then use a baseline level and have the other three levels as factors, meaning the model has three Stage parameters. Alternatively, Stage may have a single parameter value, which is then multiplied by the level of Stage. Then the value of the parameter would describe the instantaneous geometric odds ratio (or the geometric rate ratio, depending on what model we use) between a higher and a lower Stage level. This strategy makes sense if the effect of Stage is additive in terms of the odds ratio (rate ratio). We will investigate both methods.

5.3 Outline of the analysis

The purpose of this investigation is to find a good model for the instantaneous geometric odds of patients dying within a year when time t has passed since surgery. The investigation will proceed in several steps, where models will be selected according to the AIC criterion in bootstrap simulations.

AIC (Akaike's Information Criterion) is a way of measuring model fit according to the formula $-2l(\beta|t)+2p$, where p is the number of parameters. The lower the AIC, the better. This means that the AIC judges models by the value of the likelihood

function and punishes models for having many parameters.

Bootstrap simulation means, in this case, that a random sample with replacement is taken from the dataset, and models are fitted to the sample. In this thesis, 1000 simulations will be performed at every stage, and in every simulation 800 observations will be drawn with replacement. [11]

The investigation will proceed in the following steps:

First, we will determine how many RCS parameters to use in the baseline geometric odds. This is done by creating several baseline odds functions, differing in how many RCS parameters they have. Using bootstrap simulation, these models will be compared, and the model that is selected most often in terms of AIC will be chosen as the baseline geometric odds in all future models created.

At the second step, we will determine what covariates to use and how many interactions with time they shall have. First, univariate models based on the original data will be created. It will be ascertained that the covariate parameters are significant, and information from prior research will be taken into account. Covariates that are deemed unnecessary at this stage will be ignored in the future. Then, interaction with time will be added to the remaining univariate models. Bootstrap simulation will be used to see what is the optimal number of time-interactions. However, when later fitting multivariate models, calculations in Stpreg are greatly simplified if the same number of time interactions are used for all covariates. Therefore, the goal will be to find a number of time interactions that can be used for all covariates.

At the third step, bootstrap simulation is performed in order to compare all models that are possible to form with the remaining covariates. Normally when performing model selection and one has to deal with many parameters, a forward-, backward, or forward-backward selection algorithm is used because testing all possible models is too time-consuming. In this case, there are not that many covariates to deal with and hence all models can be compared at once. [12]

5.4 Data analysis

5.4.1 Step 1: Unconditional proportional odds

A model with no covariates is an unconditional odds model, on the form

$$\frac{g_0(t)}{1-g_0(t)}.$$

Such a model gives a picture of the instantaneous risk of dying for the population as a whole. 10 models were created, that differed in how many restricted cubic splines are used, from 1 to 10. 1000 bootstrap simulations were made, and in each simulation, the best model was picked out, using the AIC criterion for a model fit with

[11] In other words, this is non-parametric bootstrap, as opposed to parametric bootstrap, where new data is generated through simulations from a parametric model of interest in order to estimate variance and other properties of parameter estimates.

[12] For more information on different formard and backward selection schemes, see Agresti 2013, sections 6.1.3-6.1.5. My supervisor Taras Bodnar convinced me that it was better to compare all possible models instead of a forward selection scheme, as I had first planned.

no covariates. The distribution is presented in the graph below. As can be seen, the model with 7 RCS parameters is chosen the most times. Therefore, it is selected.

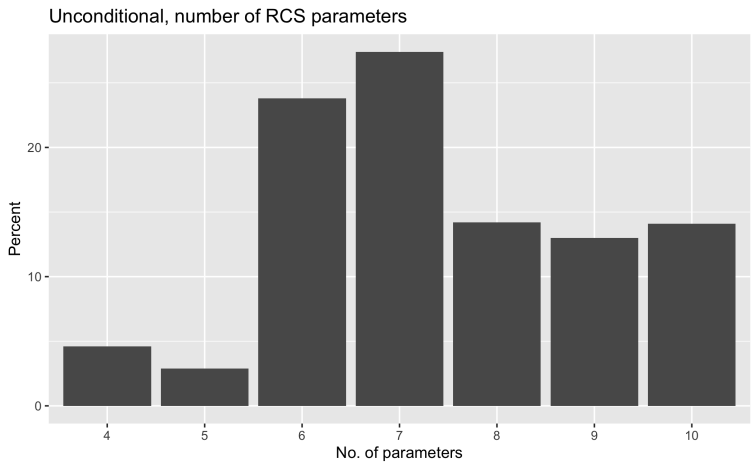

Figure 5: Distribution of selected number of RCS parameters from Bootstrap simulations for the unconditional model.

A table with parameter estimates based on the original data, along with confidence intervals, and AIC all models can be found in the appendix.

Figure 6 shows a graph of the instantaneous geometric rate for the model with 7 RCS parameters, over the studied time period. At time 0, we see that the risk of dying is around 0.7 and then it moves down and eventually it almost follows a straight line. Shortly after time 0, the risk of death increases until the risk reaches a local maximum after about a year. Then it starts decreasing again, until it settles at a risk of around 0.1 after 5 years. The increased risk of death during the first year could be because patients die from problems that occurred in relation to surgery. That the risk of death starts to decrease after one year could be because patients who survived that long in general had successful surgery, and the chance of them having beat cancer increases.

We can also relate the graph to the the Kaplan-Meier curve, in Figure 4. There, we saw that there was a lot of death in the first year, and then the death rate started to decrease a little until it stabilized after around 4 years. The plot of the unconditional geometric rate corresponds pretty well to this picture.

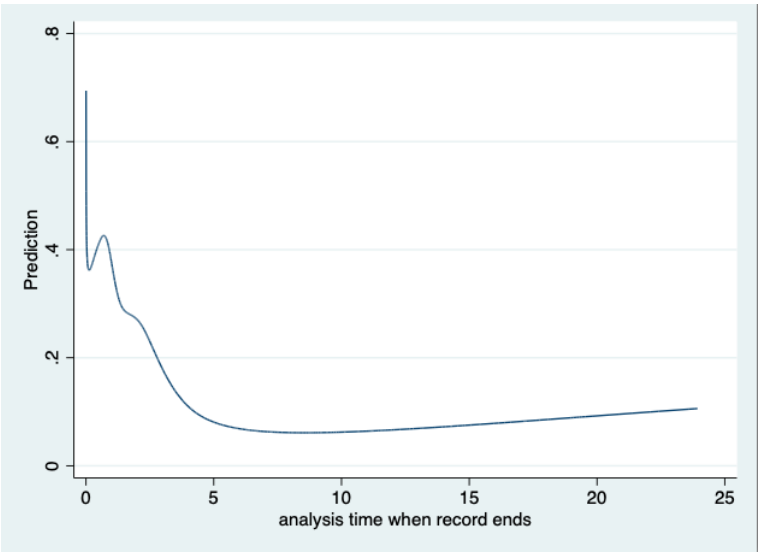

Figure 6: Predicted instantaneous geometric rate of dying within a year according to the unconditional model with 7 RCS parameters.

5.4.2 Step 2: One covariate

At this stage, new models are created by having a baseline function with 7 RCS parameters and adding one covariate. This gives an instantaneous geometric odds of the form

$$\frac{g_0(t)}{1-g_0(t)}\exp(\beta' x_i).$$

for individual i with covariate x_i. The possible covariates to include are Sex, Age, Resection and Stage. For Stage there are two different models, one where Stage is a single parameter and one where it is a factorial with four levels. Parameter estimates and AIC values for all models can be found in Table 4 in the appendix.

Sex is non-significant at the 0.05 level, meaning that we cannot reject the hypothesis that Sex does not impact the odds. This is in line with previous medical research, which has shown that when patients have a disease that is as serious as the one in our data, sex has very little impact on survival.[13]

Resection is clearly significant. Stage is also clearly significant both when treated as a single parameter and as a factorial. Which way of treating Stage is preferable? When Stage is treated as a factorial, the difference in odds ratio between the factor levels is roughly linear: for every increase in Stage level, the odds roughly doubles. For that reason, it is preferable to treat Stage as a single parameter: it is easier to interpret and it gives very similar predictions. The degrees of freedom are also fewer.

[13] My supervisor Matteo Bottai told me this during a meeting.

From the AIC, this trade-off is confirmed, since the AIC is (slightly) lower for the model where Stage is a single parameter. [14]

5.4.3 Step 3: One covariate with time interactions

Now interactions with time are added for the covariates in the univariate models. This gives a model of the form

$$\frac{g_i(t|x_i)}{1-g_i(t|x_i)} = \frac{g_o(t)}{1-g_o(t)} \exp[\beta' x_i + C(t)x_i]$$

where $C(t)$ is a spline function of time.

The interpretation of a model with covariate-time interactions, is that the odds ratios change over time: the difference in the odds of dying within a year for persons with different levels of a covariate may be larger and smaller at different times.

For Stage, Resection and Age, models were constructed with 7 RCS parameters and a single covariate. Then 1000 bootstrap simulations were made to determine the optimal number of time-interactions for each covariate. In each simulation, models with between 1 and 7 time-interactions were compared. This means that in addition to the restricted cubic spline function that makes up the baseline odds, a new set of restricted cubic splines (made with 1 to 7 degrees of freedom) are created, and the covariates interact with these new splines. In each simulation, the model with the lowest AIC was considered the best. The result of the simulations are presented in Figures 7-9.

[14] When doing bootstrap simulations, it turned out that the factorial model of Stage tended to be selected more often than the single parameter case. However, even though the factorial model was better in terms of AIC, the difference was very small in most cases. Since it is easier to interpret the single parameter version, it was deemed better.

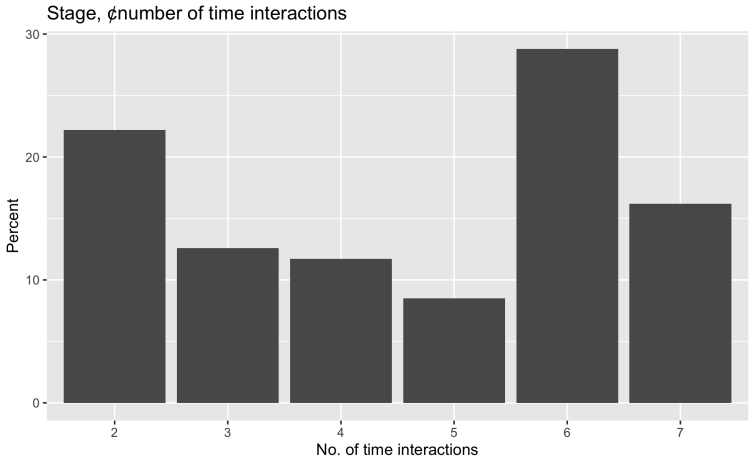

Figure 7: Distribution of selected number of time-interaction RCS parameters from Bootstrap simulations for a univariate model with Stage as the only covariate, and with a time-varying instantaneous geometric odds ratio

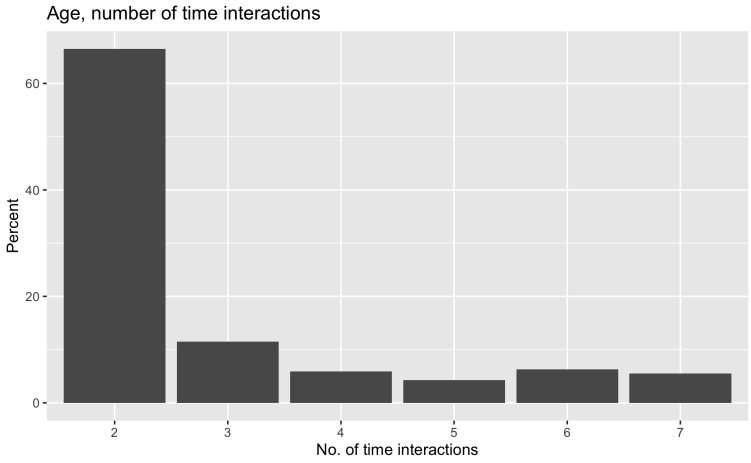

Figure 8: Distribution of selected number of time-interaction RCS parameters from Bootstrap simulations for a univariate model with Age as the only covariate, and with a time-varying instantaneous geometric odds ratio.

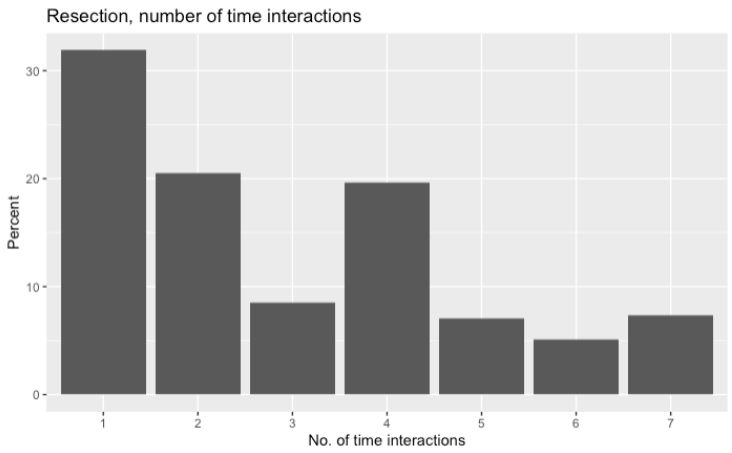

Figure 9: Distribution of selected number of time-interaction RCS parameters from Bootstrap simulations for a univariate model with Resection as the only covariate, and with a time-varying instantaneous geometric odds ratio.

For Stage, we get the lowest AIC most of the time by including 6 interactions with time, while 2 interactions are second best. For Resection, we get the best result most of the time by including 1 interaction with time. Again, 2 interactions is second best. For Age, 2 interactions is best.

Since it was preferable to have the same number of interactions with time for all covariates later on in the model fit, 2 interactions seems to be the best choice.

5.4.4 Step 4: Find the best model

Having fixed the number of RCS parameters for the baseline proportional odds ratio at 7 degrees of freedom and the number of time interactions for the covariates at 2 degrees of freedom, it is now possible to compare all possible models. Let S=Stage, R=Resection and A=Age. Two covariates joined with a hyphen denotes an interaction, so for example, "S-A" is an interaction between Stage and Age. In all models, the baseline odds is a restricted cubic spline with 7 degrees of freedom and each covariate has two interactions with time, but interactions between covariates do not have time-interactions. The models that have been compared are:

(1) No covariates
(2) S
(3) A
(4) R
(5)S, A
(6) S, R
(7) A, R
(8) S, A, S-A
(9) S, R, S-R
(10) A, R, A-R
(11) S, A, R
(12) S, A, R, S-A
(13) S, A, R, S-R
(14) S, A, R, A-R
(15) S, A, R, S-A, S-R
(16) S, A, R, S-A, A-R
(17) S, A, R, S-R, A-R
(18) S, A, R, S-R, S-A, A-R
(19) S, A, R, S-R, S-A, A-R, S-A-R

The result is presented in Figure 10. Models that were never chosen are removed.

Model 12 is chosen most of the times. This is the model with Stage, Age, Resection, Stage-Age. However, model 19, which is the saturated model, is chosen almost as many times. Model 11, which contains the covariates Stage, Age and Resection but no interactions is also performing well.

To get a better picture of the model fit, we may look at other likelihood-based methods for evaluating model fit. One of them is BIC (the Bayesian Information Criterion) which quantifies model fit with $-2l(\beta) + \log(n)p$, where n is the sample size

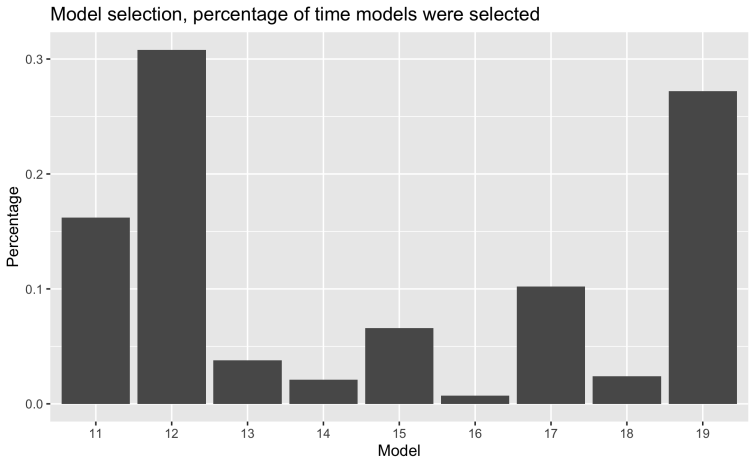

Figure 10: Proportions of the 1000 simulations when the models with different covariates and covariate interactions were chosen, with AIC as the selection criterion.

and p is the number of parameters. This means that BIC gives a greater punishment for additional parameters compared to the AIC.

Interestingly, if the BIC is used for model seletion instead of AIC, the picture changes. In Figure 11, we see how often the different models were chosen according to the BIC criterion.

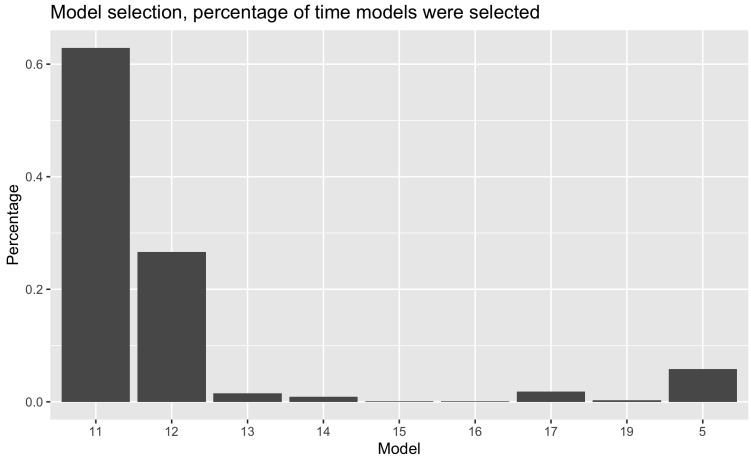

Figure 11: Proportions of the 1000 simulations when the models with different covariates and covariate interactions were chosen, with BIC as the selection criterion.

Now Model 11 if by far the best model, with Model 12 the second best, while all other models are performing very poorly and are hardly ever chosen. In particular,

the saturated model (model 19) is only chosen in 3 out of 1000 simulations.

A third way of comparing models is with the Likelihood Ratio test. This method has a more limited use than the AIC and BIC criteria, because the Likelihood Ratio test can only be used to compare nested models. However, some of our models are nested, so it works fine in our case and the Likelihood Ratio test can be used as an anlternative to AIC and BIC.

The Likelihood Ratio test is used to compare two nested models (meaning that one model is a special case of the other), with the null hypothesis that the simpler model is the true model. This means that we compare models with parameter vectors β_0 and β_a where β_0 is true under the null hypothesis while β_a is a larger parameter vector, which is used under the alternative hypothesis. Under the null hypothesis, the ratio $-2\log\left[\frac{l(\beta_0}{l(\beta_a)}\right]$ follows a χ^2 distribution with k degrees of freedom, where k is the difference in number of parameters between β_0 and β_a. If the value of the ratio is higher than a critical value (depending on the k value and the significance level), the null hypothesis is rejected, otherwise not.

Likelihood ratio tests were conducted in 1000 simulations in order to compare the three nested models 11, 12 and 19. In each simulation round, if one of the models was superior to both the other models in (pairwise) Likelihood Ratio tests, that model was considered the winner of that simulation round. It turned out that Model 12 was the winner in 369 rounds, Model 11 was the winner in 356 rounds and Model 19 was the winner in 261 rounds, when 5% was used as significance level. In 15 rounds there was no clear winner.

This means that Model 12 performed slightly better than Model 11 in terms of the likelihood ratio test, with model 19 in third place.

The conclusion we can draw from this is that model selection is very sensitive to what particular measure we are using, even though AIC, BIC and the LR test are all likelihood based. However, this was a side-note that could be better developed in a thesis dedicated to model fit and likelihood theory. AIC was our main criterion, so we consider Model 12 to be our chosen model.

5.5 Multicollinearity

As we can see from Table 5 in the appendix, the instantaneous geometric odds ratio for Resection in a univaraite model with 2 RCS interactions is 0.911 on the log scale, meaning that the odds ratio is $\exp(0.911) = 2.487$. In Model 12, where Stage, Age and Stage-Age are added, the odds ratio estimate for Resection is 1.839. The change is thus -0.648 from the univariate model to Model 12. This is a non-trivial change in odds, indicating that there is some degree of multicollinearity with Stage.[15]

This is not surprising, as it was seen previously that Resection and Stage were correlated. Multicollinearity means that several covariates to some extent contain the same information. If they contain exactly the same information, it is easy to see

[15] Solely based on this observation, it is not obvious that multicollinearity is with Stage specifically. It could also be with Age, or with both Age and Stage. However, only adding Age to the univariate model with Resection does not change the parameter estimate much for Resection, whereas adding Stage makes the difference

that at least one covariate may be omitted. If there is overlap only to some degree, it is a question of judgment whether one should omit a covariate or not.

To investigate this further, new regression models with Resection as the only coviariate were fitted to subsets of the data where Stage was fixed (to levels 1, 2, 3 and 4). The results are presented in Table 9 of the appendix. There it can be seen that the parameter estimate of the odds ratio for Resection is significant on the 0.05 level when Stage is at levels 3 and 4, but not for levels 1 and 2. This means that for long-developed cancer tumors it does have an effect whether all cancer cells were removed during surgery. But for less developed cancer, it it cannot be ruled out that it does not have an effect whether all cancer cells were resected or not.

Another way of deciding whether to include Resection or not is to conduct prior research. While I, the author of this thesis, have not had the time to read all relevant literature, it seems like prior research on Resection is non-conclusive as to whether the variable is important for survival. (see Karstens (2018) and Depypere (2018) for differing results).

Since the model with Resection added to Stage and Age clearly outperforms the model with only Stage and Age in terms of AIC and BIC, and since parameter estimates for fixed levels of Stage indicates that Resection is important at least for some levels, the best strategy seems to include Resection, in spite of the collinearity.

5.6 Interpretation of the model

The final model has 7 RCS parameters, one Resection parameter, one Stage parameter, one Age parameter, two parameters for interactions with time for each of Stage, Age and Resection, and an interaction parameter between Stage and Age. Parameter estimates for this model can be found in Table 8 of the appendix. Here is an interpretation of the model:

The instantaneous geometric odds ratio for Stage is 3.8. This means that at time 0, holding the Resection and Age covariates fixed, the odds of dying within a year is 3.8 times higher for patients who have a Stage value of x, compared to patients with an Stage value of $x-1$.

The instantaneous geometric odds ratio for Resection is 1.8. This means that at time 0, holding Age and Stage fixed, the odds of dying within a year is 1.8 times higher for patients who have a Resection value of 1 compared with patients who have a Resection value of 0.

The instantaneous geometric odds ratio for Age is 1.06. This means that at time 0, holding Stage and Resection fixed, the odds of dying within a year is 1.06 higher for a patient who is x years old compared to a patient who is $x-1$ years old.

The contribution to the instantaneous geometric odds ratio for the interactions Stage-Age is 0.99. This means that, holding Resection fixed, the odd ratio of Stage for a patient who is x years old is 0.99 times the odds ratio of Stage for a person who is $x-1$ years old. Alternatively and equivalently, it can be interpreted as, holding Resection fixed, for a Stage=1 patient, the odds ratio of Age is 0.99 times the odds ratio of Age for a Stage=0 patient. What this means is that the risk of dying within a year goes down slightly the more serious the cancer is. In itself, this may sound strange, as it seems intuitive that old age and serious cancer should increase the risk

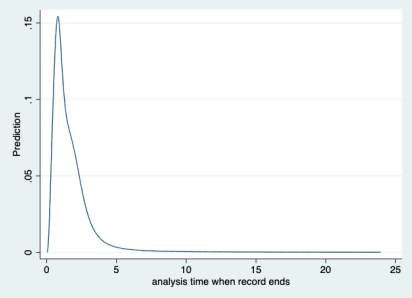

Figure 12: The probability of death within a year for patients with the covariate values Resection=0, Stage=1, Age=70

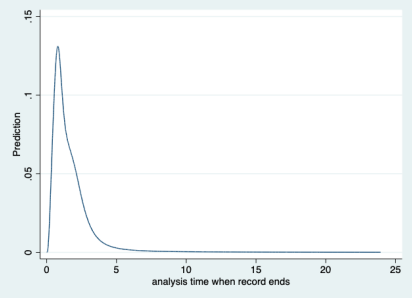

Figure 13: The probability of death within a year for patients with the covariate values Resection=0, Stage=1, Age=50

of death. But as we saw above, the Stage and Age parameters are taking care of that aspect.

For all three covariates, there is interaction with time. This means that the impact of each covariate on the risk of dying within a year changes over time. Interpretation of the parameter estimates of the time interactions is very difficult (at least I am unable to understand them).

Figures 9 and 10 show graphs with the instantaneous geometric rate for one group of patients that have Resection=0, Stage=1 and Age=50, and another group with the same Resection and Stage values but Age=70. Both graphs have a high spike right after surgery and then the risk diminishes. This is reasonable given that in the real data most of the deaths happened within five years after surgery. We also see that the risk of death overall is fairly low, which is not surprising given that the patients had mild forms of cancer and all cancer cells were removed in surgery. It is also clear that the older patients have a slightly higher risk of dying.

Figures 11 and 12 show the instantaneous geometric rate for patients with Resection=1, Stage=4 and Age=50, and patients with the same Stage and Resection values, but with Age=70. Again, we see a sharp spike of the geometric rate right after surgery which then decreases. This time, the maximum becomes high in the spike (around 0.45 and 0.6 respectively) which is reasonable given that the patients had a serious form of cancer and not all cancer cells were removed in surgery. What is surprising is that the older patients have a lower risk of dying than the younger ones. It is difficult to interpret the interactions with time, but what seems to happen is that with time, the odds ratio for Age decreases slightly. The Stage-Age interaction is constant, on the other hand, and is constant at 0.99, so the effect of this parameter may overtake the effect of Age.

The most important conclusion that can be drawn from the graphs is that if all of the cancer tissue is removed during surgery and if surgery is performed when the tumor is not well developed, the risk of death is a lot lower than if not all cancer cells are removed and surgery is performed when the tumor is well developed.

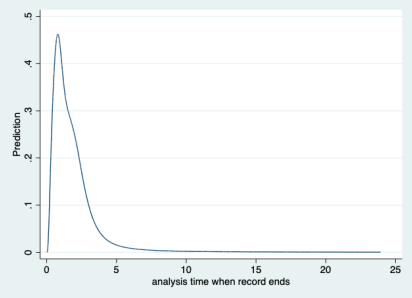

Figure 14: The probability of death within a year for patients with the covariate values Resection=1, Stage=4, Age=70

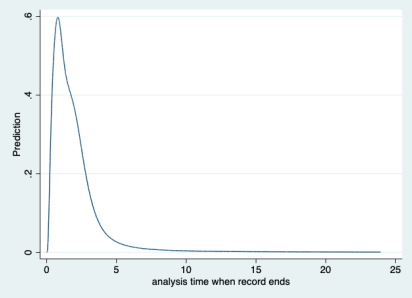

Figure 15: The probability of death within a year for patients with the covariate values Resection=1, Stage=4, Age=50

6 Discussion

In this thesis, we have proposed to use the geometric rate instead of the incidence rate when studying events that can only happen once to a person, and to use the instantaneous geometric rate instead of the hazard function to describe the instantaneous risk of an event happening to a person. We also used the instantaneous geometric rate and odds on a real data-set.

The main benefit of using the instantaneous geometric rate is that it describes the risk of an event happening. This concept is easy to understand. The hazard function, on the other hand, is a little opaque: it is related to the risk of an event happening, but it is not the same thing as the risk. This makes it it difficult to understand what the hazard function is, perhaps in particular for persons who are not statisticians. Since the hazard function is often used by non-statisticians (such as physicians and economists) it seems important to use concepts that are intuitive.

An interesting topic that has not been discussed in this thesis is how large the benefit is of using the geometric rate instead of the incidence rate. For example, it would be good to know how much the hazard ratio tends to deviate from the instantaneous geometric rate. If the gap is large, we would have a stronger argument in favor of using the geometric rate instead of the hazard function.

7 References

Aalen, O.A. et al. (2008). *Survival and Event History Analysis: A Process Point of View*. New York: Springer.

Agresti, A. (2014). *Categorical Data Analysis*, 3rd ed. New Jersey ,USA. John Wiley Sons, Inc.

Bottai, M. (2017). A regression method for modelling geometric rates. *Statistical Methods in Medical Research*, Vol. 26(6): 2700-2707.

Cao, B. et al. (2020). A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19. *New England Journal of Medicine*, Vol 382(19), p. 1787-1799.

Cox, D.R. (1972). Regression models and life tables. *Journal of the royal statistical society. Series B (Methodological)*. Volume 34(2). 187-220.

Depypere L. et al. (2018). Prognostic value of the circumferential resection margin and its definitions in esophageal cancer patients after neoadjuvant chemoradiotherapy. *Diseases of the Esophagus*, Volume 31(2).

Derogar M, et al. (2013). Hospital and Surgeon Volume in Relation to Survival After Esophageal Cancer Surgery in a Population-Based Study. *Journal of Clinical Oncology*, Vol 21(5), p 551-7.

Discacciati, A. and Bottai, M. (2017). Instantaneous geometric rates via generalized linear models. *The Stata Journal*, Vol. 17(2): 358:371.

Givens, G.H. and Hoeting, J.A (2013). *Computational Statistics*, 2nd ed. Hoboken: John Wiley Sons.

Guan, W. and Gutierrez,R. G. (2002). Programmable GLM: Two user-defined links. *Stata Journal*, Vol. 2(4): 378–390.

Heinzl, H. and Kaider, A. (1997). Gaining more exibility in COX proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine*, Vol. 54(3): 201-208.

Held, L. and Bové, D. (2014). *Applied Statistical Inference: Likelihood and Bayes*. New York: Springer.

Karstens K. et al.(2018). Does the Margin Matter in Esophageal Cancer. *Digestive Surgery*, Vol 35, p. 196-203.

Kausik, R. et al. (2020). Effect of Apabetalone Added to Standard Therapy on Major Adverse Cardiovascular Events in Patients With Recent Acute Coronary Syndrome and Type 2 Diabetes. *JAMA*, Vol 323(16), p. 1565-1573.

Klein, J.P. and Moeschberger, M.L. (2003). *Survival analysis: techniques for censored and truncated data* , 2nd ed. New York: Springer.

Markar S.R. et al. (2016). Surgical Proficiency Gain and Survival After Esophagectomy for Cancer. sl Journal of Clinical Oncology, Vol. 34(13).

National Cancer Institute (2015). *Cancer Staging*. From the website: https://www.cancer.gov/about-cancer/diagnosis-staging/staging. Accessed 7 May 2020.

Petrie, M. et al. (2020). Effect of Dapagliflozin on Worsening Heart Failure and Cardiovascular Death in Patients With Heart Failure With and Without Diabetes. *JAMA*, Vol 323(14), p. 1353-1368.

Royston, P. andLambert, P.C. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with applications to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*. Volume 21(15), 2157-2197.

Royston, P. andLambert, P.C. (2011). *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, TX: Stata Press.

Sutradhar, R. and Austin, P.C. (2018). Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. *Annals of Epidemiology* Vol. 28, 54-57.

Zeiser, R. et al (2020). Ruxolitinib for Glucocorticoid-Refractory Acute Graft-versus-Host Disease. *New England Journal of Medicine*, Vol 382(19), p. 1800-1810.

Appendix

Theory and proofs

Proof of formula (1)

The proof can be found in Bottai (2015).

$$\begin{aligned} h(t) &= \lim_{\triangle t \to 0} \frac{P(t<T\leq t+\triangle t | T>t)}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P(t<T\leq t+\triangle t, T>t)}{P(T>t)\cdot \triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P(t<T\leq t+\triangle t)}{P(T>t)\cdot \triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P(t<T\leq t+\triangle t)}{S(t)\cdot \triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P(T\leq t+\triangle t)-P(T\leq t)}{S(t)\cdot \triangle t} \\ &= \frac{dF(t)}{dt}\frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

In the last step, we exploited the fact that the derivative of a probability distribution in a certain point, is the probability density function evaluated in the same point.

Since $f(t) = \frac{-dS(t)}{dt}$ we can use the chain rule to conclude that

$$h(t) = f(t)/S(t) = \frac{-dS(t)}{dt\, S(t)} = -\frac{d\log(S(t))}{dt}.$$

Graphs

Mean survival over time

In Figure 16, each dot represents the mean survival time for patients who had surgery a given year. There appears to be a trend towards longer survival times. It is broken in the last three years. The explanation is due to right censoring, i.e. that people who had surgery those years had little time to survive before resection.

Stage over time

In Figure 17, each dot represents the proportion of patients who had cancer of the possible stages (1,2,3,4) for a given year. The blue lines are regressions lines, inserted to help the reader see trends. We can se that the proportion of Stage=1 increases while Stage=3 decreases.

Resection over time

Figure 18 shows the proportion of patients that have had Resection=0 and Resection=1 each year. As we can see, there is a clear trend towards more patients having Resection=0, meaning that for an increasing proportion of surgeries, all cancer cells are removed.

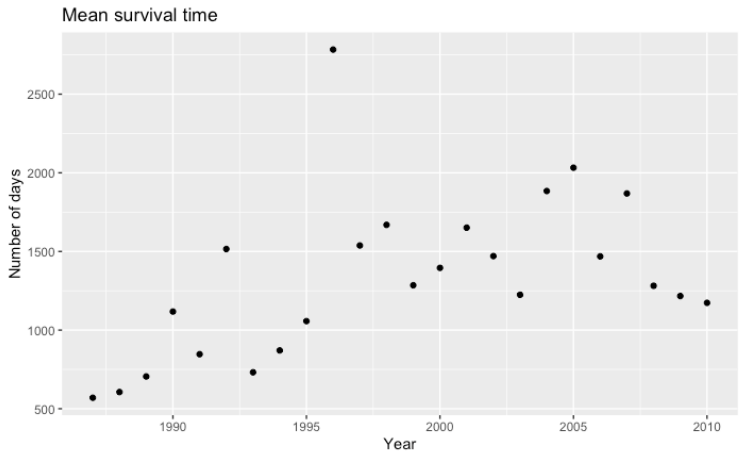

Figure 16: Mean survival time of patients who had surgery, per year. Each dot represents the average survival time (in days) for patients who underwent surgery a particular year.

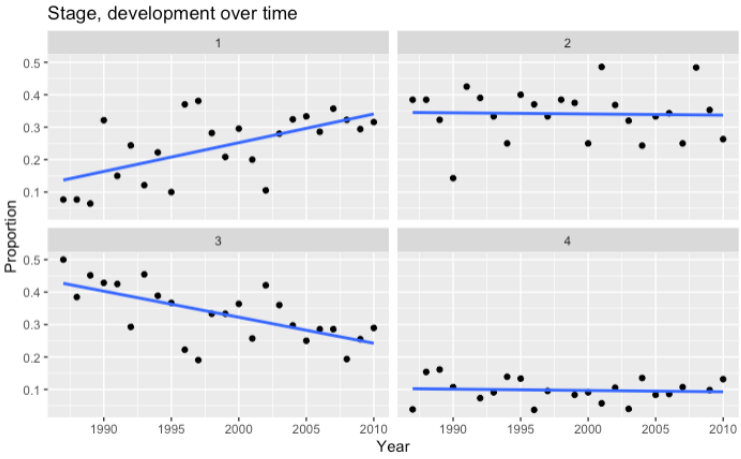

Figure 17: A plot showing how the proportion of patients with various stages changes over time. The subplots illustrate the proportions for Stage 1 (upper left), Stage 2 (upper right), Stage 3 (lower left) and Stage 4 (lower right). The lines are simple linear regression fit lines.

Age over time

Figure 19 shows the distribution of ages for each year. All patients are represented with a transparent dot, so a black dot means that several paitents had the same age that year. The blue regression line shows that there is no apparent trend in mean

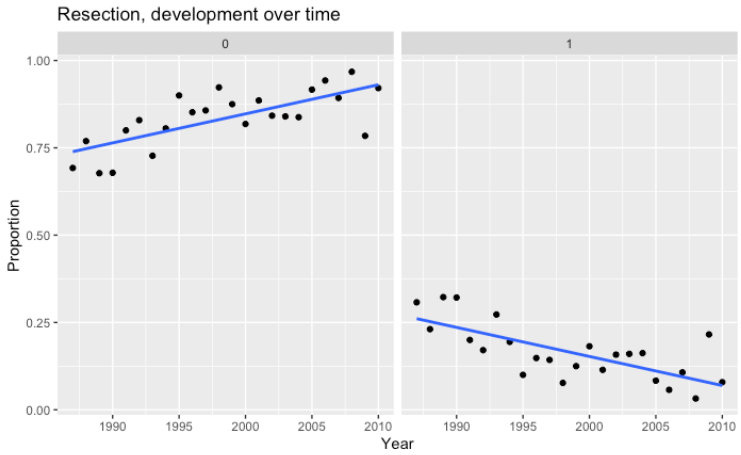

Figure 18: The plot illustrates how the proportion of patients with Resection=0 (left) and Resection=1 (right) varies over time. The lines are simple linear regression fit lines.

age.

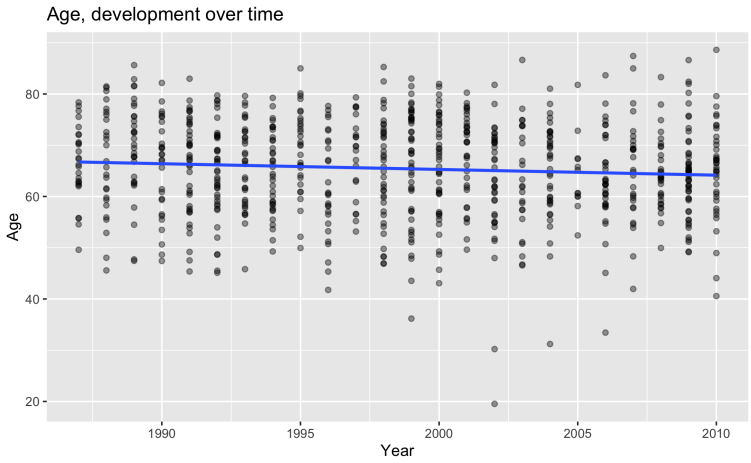

Figure 19: The plot illustrates how the age ditribution (plots) of patients varies over time. Each dot represents the age of a patient when that patient had surgery. The line is a simple linear regression fit.

Sex over time

Figure 20 shows the proportion of female each year. While the proportion has changed between years, there is no clear trend of change.

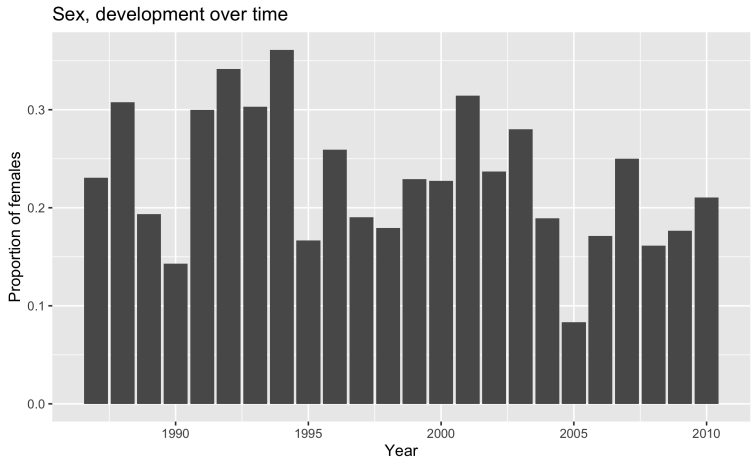

Figure 20: The plot illustrates how the proportion of patients with Resection=0 (left) and Resection=1 (right) varies over time.

Tables

In all of the tables, the parameter estimates are log instantaneous geometric odds ratios.

Table 3: Parameters estimates, 95% confidence intervals of parameter estimates and AIC values for unconditional models with no covariates and 1, 2, …, 10 RCS parameters for the baseline instantaneous geometric odds ratio. Significant parameter estimates are marked with stars.

	Number of RCS parameters									
	1	2	3	4	5	6	7	8	9	10
Spline1	-0.762***	-0.818***	-0.834***	-0.787***	-0.780***	-0.774***	-0.774***	-0.775***	-0.775***	-0.774***
	[-0.854,-0.669]	[-0.918,-0.718]	[-0.942,-0.725]	[-0.891,-0.683]	[-0.883,-0.676]	[-0.877,-0.672]	[-0.876,-0.672]	[-0.877,-0.673]	[-0.877,-0.673]	[-0.876,-0.672]
Spline2		0.436***	0.438***	0.378***	0.369***	0.356***	0.353***	0.354***	0.354***	0.352***
		[0.328,0.543]	[0.331,0.546]	[0.276,0.480]	[0.267,0.470]	[0.255,0.457]	[0.253,0.453]	[0.254,0.454]	[0.254,0.454]	[0.252,0.452]
Spline3			0.0576	0.0701	0.0946	0.113*	0.119*	0.125**	0.133**	0.143**
			[-0.0465,0.162]	[-0.0272,0.167]	[-0.00192,0.191]	[0.0168,0.209]	[0.0237,0.214]	[0.0300,0.220]	[0.0381,0.228]	[0.0481,0.238]
Spline4				-0.234***	-0.216***	-0.201***	-0.180***	-0.155**	-0.129**	-0.0978*
				[-0.328,-0.139]	[-0.310,-0.121]	[-0.295,-0.106]	[-0.273,-0.0863]	[-0.248,-0.0622]	[-0.222,-0.0360]	[-0.191,-0.00458]
Spline5					-0.121*	-0.129**	-0.154***	-0.170***	-0.183***	-0.185***
					[-0.215,-0.0267]	[-0.221,-0.0374]	[-0.245,-0.0630]	[-0.261,-0.0791]	[-0.275,-0.0906]	[-0.278,-0.0913]
Spline6						-0.161***	-0.124**	-0.0806	-0.0654	-0.0743
						[-0.256,-0.0665]	[-0.216,-0.0326]	[-0.173,0.0119]	[-0.158,0.0276]	[-0.167,0.0184]
Spline7							-0.163***	-0.175***	-0.141**	-0.102*
							[-0.257,-0.0687]	[-0.267,-0.0825]	[-0.232,-0.0501]	[-0.194,-0.00996]
Spline8								-0.0963*	-0.148**	-0.138**
								[-0.189,-0.00403]	[-0.240,-0.0562]	[-0.230,-0.0462]
Spline9									-0.0544	-0.132**
									[-0.145,0.0363]	[-0.222,-0.0413]
Spline10										-0.0154
										[-0.106,0.0756]
_cons	-1.219***	-1.260***	-1.269***	-1.223***	-1.219***	-1.216***	-1.218***	-1.217***	-1.216***	-1.214***
	[-1.307,-1.130]	[-1.351,-1.169]	[-1.363,-1.175]	[-1.316,-1.129]	[-1.312,-1.125]	[-1.309,-1.123]	[-1.311,-1.125]	[-1.310,-1.124]	[-1.309,-1.122]	[-1.308,-1.121]
n	800	800	800	800	800	800	800	800	800	800
AIC	3013.3	2945.3	2947.2	2923.5	2922.4	2918.3	2917.7	2918.7	2919.9	2921.3

95% confidence intervals in brackets

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

Table 4: Parameters estimates, 95% confidence intervals of parameter estimates and AIC values for univariate models with one covariate and 7 RCS parameters for the baseline instantaneous geometric odds ratio. Significant parameter estimates are marked with stars.

	Resection	Age	Stage	Sex	Stage (factor)
Spline1	-0.717***	-0.743***	-0.616***	-0.768***	-0.616***
	[-0.820,-0.613]	[-0.846,-0.641]	[-0.723,-0.510]	[-0.870,-0.666]	[-0.722,-0.509]
Spline2	0.352***	0.341***	0.309***	0.350***	0.307***
	[0.250,0.454]	[0.240,0.442]	[0.205,0.412]	[0.249,0.450]	[0.204,0.410]
Spline3	0.132**	0.105*	0.137**	0.115*	0.135**
	[0.0353,0.229]	[0.00966,0.201]	[0.0389,0.235]	[0.0201,0.210]	[0.0364,0.233]
Spline4	-0.171***	-0.192***	-0.129**	-0.182***	-0.131**
	[-0.266,-0.0765]	[-0.286,-0.0980]	[-0.226,-0.0322]	[-0.276,-0.0887]	[-0.228,-0.0343]
Spline5	-0.155***	-0.159***	-0.145**	-0.155***	-0.144**
	[-0.247,-0.0627]	[-0.250,-0.0674]	[-0.238,-0.0515]	[-0.246,-0.0637]	[-0.237,-0.0501]
Spline6	-0.119*	-0.129**	-0.106*	-0.126**	-0.106*
	[-0.212,-0.0264]	[-0.221,-0.0364]	[-0.200,-0.0122]	[-0.218,-0.0341]	[-0.200,-0.0120]
Spline7	-0.163***	-0.167***	-0.160**	-0.165***	-0.159**
	[-0.258,-0.0680]	[-0.261,-0.0719]	[-0.256,-0.0639]	[-0.259,-0.0707]	[-0.255,-0.0630]
Resection	1.115***				
	[0.842,1.388]				
Age		0.0275***			
		[0.0179,0.0370]			
Stage			0.682***		
			[0.575,0.790]		
Sex				-0.200	
				[-0.411,0.0105]	
Stage1					-2.041***
					[-2.459,-1.622]
Stage2					-1.218***
					[-1.616,-0.819]
Stage3					-0.631**
					[-1.032,-0.229]
_cons	-1.346***	-2.998***	-2.636***	-1.167***	0.0196
	[-1.446,-1.246]	[-3.631,-2.365]	[-2.890,-2.383]	[-1.273,-1.060]	[-0.350,0.389]
n	800	800	800	800	800
AIC	2857.9	2886.7	2758.4	2916.2	2760.7

95% confidence intervals in brackets

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

Table 5: Parameters estimates, 95% confidence intervals of parameter estimates and AIC values for univariate models with Resection as covariate, 7 RCS parameters for the baseline instantaneous geometric odds ratio, and 1,...,7 interaction parameters between Resection and time. Significant parameter estimates are marked with stars.

	Number of interactions between Resection and time							
	0	1	2	3	4	5	6	7
Resection	1.115*** [0.842,1.388]	0.996*** [0.691,1.302]	0.911*** [0.567,1.256]	0.894*** [0.519,1.268]	0.827*** [0.400,1.254]	0.793*** [0.323,1.264]	0.776** [0.275,1.276]	0.726* [0.149,1.304]
Spline1	-0.717*** [-0.820,-0.613]	-0.675*** [-0.786,-0.563]	-0.676*** [-0.787,-0.565]	-0.675*** [-0.786,-0.564]	-0.674*** [-0.785,-0.564]	-0.673*** [-0.784,-0.563]	-0.673*** [-0.784,-0.563]	-0.673*** [-0.784,-0.563]
Spline2	0.352*** [0.250,0.454]	0.369*** [0.263,0.475]	0.338*** [0.225,0.451]	0.339*** [0.226,0.453]	0.335*** [0.221,0.449]	0.335*** [0.222,0.449]	0.335*** [0.222,0.449]	0.336*** [0.222,0.449]
Spline3	0.132** [0.0353,0.229]	0.135** [0.0359,0.235]	0.147** [0.0475,0.246]	0.142** [0.0348,0.249]	0.154** [0.0450,0.264]	0.151** [0.0407,0.261]	0.150** [0.0394,0.260]	0.148** [0.0376,0.258]
Spline4	-0.171*** [-0.266,-0.0765]	-0.172*** [-0.268,-0.0760]	-0.172*** [-0.267,-0.0767]	-0.172*** [-0.268,-0.0767]	-0.190*** [-0.290,-0.0907]	-0.184*** [-0.287,-0.0814]	-0.183*** [-0.287,-0.0787]	-0.181*** [-0.286,-0.0770]
Spline5	-0.155*** [-0.247,-0.0627]	-0.157*** [-0.249,-0.0648]	-0.155*** [-0.248,-0.0628]	-0.155** [-0.248,-0.0627]	-0.163*** [-0.256,-0.0700]	-0.168*** [-0.262,-0.0732]	-0.169*** [-0.266,-0.0716]	-0.176*** [-0.274,-0.0773]
Spline6	-0.119* [-0.212,-0.0264]	-0.120* [-0.212,-0.0270]	-0.120* [-0.212,-0.0271]	-0.119* [-0.212,-0.0263]	-0.120* [-0.213,-0.0269]	-0.127** [-0.221,-0.0322]	-0.128** [-0.224,-0.0324]	-0.122* [-0.219,-0.0247]
Spline7	-0.163*** [-0.258,-0.0680]	-0.164*** [-0.259,-0.0694]	-0.165*** [-0.259,-0.0699]	-0.164*** [-0.259,-0.0697]	-0.160*** [-0.256,-0.0653]	-0.160*** [-0.256,-0.0653]	-0.163*** [-0.258,-0.0665]	-0.172*** [-0.271,-0.0736]
Resection-Spline1		-0.321* [-0.630,-0.0115]	-0.442* [-0.814,-0.0705]	-0.471* [-0.911,-0.0316]	-0.561* [-1.089,-0.0323]	-0.621* [-1.230,-0.0128]	-0.650 [-1.313,0.0117]	-0.725 [-1.517,0.0681]
Resection-Spline2			0.223 [-0.0954,0.542]	0.249 [-0.128,0.625]	0.327 [-0.117,0.772]	0.383 [-0.129,0.896]	0.408 [-0.145,0.960]	0.476 [-0.186,1.139]
Resection-Spline3				0.0530 [-0.281,0.387]	0.144 [-0.255,0.543]	0.175 [-0.264,0.615]	0.190 [-0.280,0.660]	0.246 [-0.308,0.801]
Resection-Spline4					0.226 [-0.126,0.579]	0.257 [-0.138,0.653]	0.242 [-0.178,0.663]	0.280 [-0.215,0.776]
Resection-Spline5						0.203 [-0.175,0.581]	0.263 [-0.137,0.663]	0.340 [-0.107,0.786]
Resection-Spline6							0.139 [-0.250,0.528]	0.201 [-0.246,0.648]
Resection-Spline7								0.217 [-0.215,0.650]
_cons	-1.346*** [-1.446,-1.246]	-1.347*** [-1.447,-1.247]	-1.344*** [-1.444,-1.245]	-1.344*** [-1.443,-1.244]	-1.342*** [-1.442,-1.242]	-1.342*** [-1.442,-1.242]	-1.342*** [-1.442,-1.243]	-1.343*** [-1.442,-1.243]
n	800	800	800	800	800	800	800	800
AIC	2857.9	2855.5	2855.6	2857.5	2857.9	2859.4	2861.3	2862.4

95% confidence intervals in brackets

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

Table 6: Parameters estimates, 95% confidence intervals of parameter estimates and AIC values for univariate models with Age as covariate, 7 RCS parameters for the baseline instantaneous geometric odds ratio, and 1,. . . ,7 interaction parameters between Age and time. Significant parameter estimates are marked with stars.

	Number of interactions between Age and time							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
age	0.0275*** [0.0179,0.0370]	0.0280*** [0.0184,0.0377]	0.0330*** [0.0230,0.0431]	0.0331*** [0.0230,0.0433]	0.0328*** [0.0226,0.0430]	0.0328*** [0.0226,0.0431]	0.0326*** [0.0224,0.0429]	0.0326*** [0.0224,0.0428]
Spline1	-0.743*** [-0.846,-0.641]	-1.322*** [-2.009,-0.634]	-1.418*** [-2.190,-0.647]	-1.441*** [-2.269,-0.613]	-1.446*** [-2.268,-0.624]	-1.462*** [-2.290,-0.634]	-1.469*** [-2.295,-0.643]	-1.498*** [-2.320,-0.676]
Spline2	0.341*** [0.240,0.442]	0.331*** [0.230,0.432]	2.276*** [1.419,3.133]	2.263*** [1.383,3.143]	2.184*** [1.276,3.092]	2.199*** [1.287,3.111]	2.162*** [1.255,3.070]	2.118*** [1.225,3.011]
Spline3	0.105* [0.00966,0.201]	0.101* [0.00538,0.196]	0.291*** [0.160,0.422]	0.337 [-0.342,1.015]	0.419 [-0.322,1.159]	0.437 [-0.355,1.228]	0.499 [-0.310,1.308]	0.576 [-0.233,1.385]
Spline4	-0.192*** [-0.286,-0.0980]	-0.196*** [-0.290,-0.102]	-0.169*** [-0.266,-0.0717]	-0.139 [-0.474,0.196]	-0.251 [-0.779,0.277]	-0.260 [-0.874,0.353]	-0.303 [-0.969,0.362]	-0.445 [-1.159,0.269]
Spline5	-0.159*** [-0.250,-0.0674]	-0.164*** [-0.256,-0.0722]	-0.167*** [-0.259,-0.0742]	-0.153* [-0.287,-0.0181]	-0.270 [-0.729,0.189]	-0.185 [-0.665,0.295]	-0.240 [-0.791,0.310]	-0.142 [-0.803,0.519]
Spline6	-0.129** [-0.221,-0.0364]	-0.133** [-0.225,-0.0401]	-0.144** [-0.238,-0.0512]	-0.141** [-0.236,-0.0463]	-0.187 [-0.388,0.0149]	-0.0961 [-0.606,0.414]	-0.0519 [-0.551,0.447]	-0.0406 [-0.716,0.635]
Spline7	-0.167*** [-0.261,-0.0719]	-0.171*** [-0.266,-0.0756]	-0.183*** [-0.279,-0.0869]	-0.183*** [-0.279,-0.0874]	-0.185*** [-0.282,-0.0889]	-0.173* [-0.323,-0.0229]	-0.0982 [-0.384,0.187]	-0.197 [-0.892,0.498]
Age-Spline1		0.00884 [-0.00153,0.0192]	0.0107 [-0.000724,0.0220]	0.0110 [-0.00121,0.0232]	0.0110 [-0.00107,0.0231]	0.0113 [-0.000919,0.0235]	0.0114 [-0.000784,0.0236]	0.0118 [-0.000306,0.0240]
Age-Spline2			-0.0294*** [-0.0421,-0.0167]	-0.0292*** [-0.0421,-0.0164]	-0.0281*** [-0.0414,-0.0148]	-0.0283*** [-0.0416,-0.0149]	-0.0276*** [-0.0410,-0.0143]	-0.0269*** [-0.0401,-0.0137]
Age-Spline3				-0.00206 [-0.0135,0.00935]	-0.00331 [-0.0147,0.00810]	-0.00419 [-0.0160,0.00761]	-0.00585 [-0.0179,0.00616]	-0.00737 [-0.0194,0.00469]
Age-Spline4					0.00247 [-0.00803,0.0130]	0.00151 [-0.00880,0.0118]	0.00210 [-0.00823,0.0124]	0.00367 [-0.00705,0.0144]
Age-Spline5						-0.000904 [-0.0104,0.00858]	0.000304 [-0.00912,0.00972]	-0.000401 [-0.0104,0.00955]
Age-Spline6							-0.00196 [-0.00854,0.00463]	-0.00159 [-0.0117,0.00856]
Age-Spline7								0.000224 [-0.0102,0.0107]
_cons	-2.998*** [-3.631,-2.365]	-3.031*** [-3.669,-2.392]	-3.368*** [-4.043,-2.693]	-3.376*** [-4.057,-2.694]	-3.354*** [-4.039,-2.669]	-3.357*** [-4.042,-2.671]	-3.341*** [-4.026,-2.655]	-3.340*** [-4.025,-2.654]
n	800	800	800	800	800	800	800	800
AIC	2886.7	2885.9	2864.3	2866.3	2868.0	2870.1	2871.4	2873.2

95% confidence intervals in brackets

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

Table 7: Parameters estimates, 95% confidence intervals of parameter estimates and AIC values for univariate models with Stage as covariate, 7 RCS parameters for the baseline instantaneous geometric odds ratio, and 1,...,7 interaction parameters between Stage and time. Significant parameter estimates are marked with stars.

	Number of interactions between Stage and time							
	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Stage	0.682*** [0.575,0.790]	0.672*** [0.563,0.781]	0.624*** [0.508,0.739]	0.603*** [0.483,0.722]	0.621*** [0.500,0.741]	0.624*** [0.503,0.745]	0.623*** [0.503,0.744]	0.620*** [0.499,0.740]
Spline1	-0.616*** [-0.723,-0.510]	-0.357* [-0.643,-0.0717]	-0.318* [-0.588,-0.0478]	-0.282* [-0.553,-0.0110]	-0.294* [-0.561,-0.0260]	-0.285* [-0.554,-0.0162]	-0.279* [-0.548,-0.00956]	-0.284* [-0.553,-0.0148]
Spline2	0.309*** [0.205,0.412]	0.329*** [0.221,0.438]	-0.425** [-0.715,-0.136]	-0.405** [-0.689,-0.121]	-0.387** [-0.668,-0.107]	-0.389** [-0.669,-0.108]	-0.373** [-0.655,-0.0922]	-0.375** [-0.656,-0.0938]
Spline3	0.137** [0.0389,0.235]	0.143** [0.0423,0.244]	0.151** [0.0491,0.254]	-0.0455 [-0.300,0.209]	-0.131 [-0.399,0.136]	-0.158 [-0.428,0.113]	-0.149 [-0.420,0.121]	-0.142 [-0.413,0.130]
Spline4	-0.129** [-0.226,-0.0322]	-0.125* [-0.223,-0.0275]	-0.118* [-0.217,-0.0180]	-0.192** [-0.321,-0.0630]	-0.0440 [-0.253,0.165]	0.0270 [-0.228,0.282]	0.100 [-0.173,0.373]	0.104 [-0.174,0.382]
Spline5	-0.145** [-0.238,-0.0515]	-0.150** [-0.244,-0.0565]	-0.128** [-0.224,-0.0321]	-0.142** [-0.238,-0.0450]	-0.0153 [-0.193,0.162]	-0.0420 [-0.233,0.149]	-0.167 [-0.406,0.0717]	-0.188 [-0.460,0.0839]
Spline6	-0.106* [-0.200,-0.0122]	-0.107* [-0.201,-0.0136]	-0.104* [-0.199,-0.00915]	-0.0957 [-0.192,0.000695]	-0.0629 [-0.168,0.0426]	-0.107 [-0.316,0.102]	-0.0196 [-0.229,0.190]	0.0248 [-0.239,0.288]
Spline7	-0.160** [-0.256,-0.0639]	-0.162*** [-0.258,-0.0666]	-0.162*** [-0.259,-0.0657]	-0.155** [-0.253,-0.0579]	-0.167*** [-0.266,-0.0694]	-0.164** [-0.264,-0.0647]	-0.0738 [-0.230,0.0820]	-0.0572 [-0.307,0.193]
Stage-Spline1		-0.117 [-0.238,0.00285]	-0.179** [-0.303,-0.0546]	-0.208** [-0.340,-0.0766]	-0.196** [-0.325,-0.0667]	-0.199** [-0.329,-0.0688]	-0.199** [-0.328,-0.0691]	-0.196** [-0.325,-0.0673]
Stage-Spline2			0.338*** [0.213,0.462]	0.353*** [0.227,0.478]	0.340*** [0.218,0.463]	0.341*** [0.218,0.464]	0.332*** [0.209,0.454]	0.329*** [0.208,0.451]
Stage-Spline3				0.115 [-0.00517,0.235]	0.121* [0.00594,0.237]	0.139* [0.0250,0.253]	0.145* [0.0318,0.258]	0.149** [0.0364,0.261]
Stage-Spline4					-0.0803 [-0.198,0.0377]	-0.0805 [-0.197,0.0361]	-0.0826 [-0.198,0.0329]	-0.0808 [-0.197,0.0356]
Stage-Spline5						-0.000487 [-0.118,0.117]	0.0135 [-0.102,0.129]	0.0171 [-0.0995,0.134]
Stage-Spline6							-0.0690 [-0.172,0.0344]	-0.0637 [-0.182,0.0549]
Stage-Spline7								-0.0590 [-0.177,0.0590]
_cons	-2.636*** [-2.890,-2.383]	-2.638*** [-2.892,-2.383]	-2.584*** [-2.842,-2.326]	-2.551*** [-2.812,-2.290]	-2.587*** [-2.851,-2.322]	-2.595*** [-2.861,-2.330]	-2.595*** [-2.860,-2.329]	-2.590*** [-2.856,-2.325]
n	800	800	800	800	800	800	800	800
AIC	2758.4	2756.6	2727.7	2727.1	2726.1	2726.9	2726.0	2728.1

95% confidence intervals in brackets

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

Table 8: Parameters estimates, 95% confidence intervals of parameter estimates and AIC values for multivariate models with covaraites S=Stage, R=Resection, A=Age, 7 RCS parameters for the baseline instantaneous geometric odds ratio, and 2 interaction parameters between each covariate and time. Significant parameter estimates are marked with stars.

	Models		
	(S,R,A)	(S,R,A,S-A)	(Full model)
Stage	0.521***	1.338**	1.563***
	[0.403,0.640]	[0.526,2.151]	[0.679,2.447]
Age	0.0313***	0.0575***	0.0637***
	[0.0207,0.0418]	[0.0294,0.0856]	[0.0341,0.0933]
Resection	0.622***	0.601***	5.224
	[0.276,0.968]	[0.253,0.948]	[-1.933,12.38]
Spline1	-1.059*	-0.804	-0.807
	[-1.887,-0.230]	[-1.674,0.0654]	[-1.682,0.0672]
Spline2	2.528***	2.408***	2.385***
	[1.602,3.454]	[1.469,3.347]	[1.444,3.325]
Spline3	0.447***	0.432***	0.429***
	[0.302,0.592]	[0.286,0.578]	[0.284,0.575]
Spline4	-0.0937	-0.0984	-0.0991
	[-0.198,0.0102]	[-0.202,0.00540]	[-0.203,0.00469]
Spline5	-0.143**	-0.144**	-0.145**
	[-0.241,-0.0444]	[-0.242,-0.0458]	[-0.243,-0.0463]
Spline6	-0.127**	-0.129**	-0.129**
	[-0.224,-0.0305]	[-0.226,-0.0319]	[-0.226,-0.0318]
Spline7	-0.189***	-0.188***	-0.187***
	[-0.288,-0.0906]	[-0.287,-0.0898]	[-0.286,-0.0888]
Stage-Spline1	-0.00251*	-0.00240*	-0.00242*
	[-0.00443,-0.000594]	[-0.00433,-0.000469]	[-0.00438,-0.000451]
Stage-Spline2	0.00605***	0.00588***	0.00586***
	[0.00406,0.00804]	[0.00388,0.00788]	[0.00386,0.00786]
Age-Spline1	0.0127	0.00874	0.00886
	[-0.000378,0.0258]	[-0.00486,0.0223]	[-0.00477,0.0225]
Age-Spline2	-0.0466***	-0.0444***	-0.0441***
	[-0.0614,-0.0318]	[-0.0593,-0.0295]	[-0.0589,-0.0292]
Resection-Spline1	-0.428*	-0.430*	-0.432*
	[-0.795,-0.0615]	[-0.797,-0.0622]	[-0.817,-0.0471]
Resetion-Spline2	-0.0602	-0.0624	-0.0549
	[-0.400,0.279]	[-0.401,0.277]	[-0.401,0.291]
Age-Stage		-0.0123*	-0.0155*
		[-0.0244,-0.000217]	[-0.0286,-0.00245]
Age-Resection			-0.0683
			[-0.178,0.0417]
Stage-Resection			-1.824
			[-4.405,0.757]
Age-Stage-Resection			0.0270
			[-0.0129,0.0670]
_cons	-4.502***	-6.234***	-6.664***
	[-5.232,-3.772]	[-8.110,-4.358]	[-8.651,-4.678]
n	800	800	800
AIC	2653.4	2651.4	2655.4

95% confidence intervals in brackets

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

Table 9: Parameters estimates, 95% confidence intervals of parameter estimates and AIC values for univariate model with Resection as the covariate, 7 RCS parameters for the baseline instantaneous geometric odds ratio, and 2 interaction parameters between each Resection and time. THe model has been fitted to subset of the original data, with only patients that have levels 1, 2, 4, and 4 of Stage..

	Levels of Stage that patients belong to			
	Stage=1	Stage=2	Stage=3	Stage=4
Resection	1.789	0.327	0.744**	1.443*
	[-3.665,7.242]	[-0.183,0.837]	[0.228,1.260]	[0.235,2.651]
Spline1	-0.444***	-0.441***	-0.617***	1.176
	[-0.642,-0.246]	[-0.624,-0.259]	[-0.857,-0.378]	[-1.832,4.183]
Spline2	-0.113	0.288**	0.709***	2.062
	[-0.310,0.0836]	[0.107,0.470]	[0.461,0.956]	[-1.109,5.233]
Spline3	-0.157	0.172*	0.458***	0.287
	[-0.351,0.0372]	[0.0163,0.327]	[0.254,0.663]	[-0.587,1.162]
Spline4	-0.173	-0.00947	0.0261	-0.579*
	[-0.361,0.0150]	[-0.165,0.146]	[-0.170,0.222]	[-1.084,-0.0743]
Spline5	-0.0869	-0.139	-0.120	0.0288
	[-0.278,0.105]	[-0.289,0.0118]	[-0.309,0.0693]	[-0.453,0.511]
Spline6	-0.0379	-0.132	0.0622	-0.152
	[-0.229,0.153]	[-0.290,0.0245]	[-0.135,0.259]	[-0.607,0.304]
Spline7	-0.0840	-0.0858	-0.0945	0.163
	[-0.315,0.147]	[-0.238,0.0663]	[-0.288,0.0987]	[-0.225,0.550]
Resection-spline1	1.624	-0.729**	-0.0940	-1.536
	[-1.082,4.329]	[-1.249,-0.208]	[-0.680,0.492]	[-3.985,0.912]
Resection-spline2	1.055	-0.184	0.333	-2.356
	[-4.236,6.347]	[-0.729,0.361]	[-0.257,0.923]	[-6.135,1.423]
_cons	-2.371***	-1.289***	-0.435***	0.0492
	[-2.581,-2.161]	[-1.450,-1.129]	[-0.639,-0.231]	[-0.849,0.948]
n	189	274	269	68
AIC	629.6	991.8	883.3	214.6

95% confidence intervals in brackets

* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001