Universitetet i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme/specialisation: Master of Mthematics and Physics Statistics	Spring / ~~Autumn semeste~~r, 2019.. Open/C~~onfidentia~~l
Author: Suhb-eldin Abdulaziz Nusr	..... (signature of author)
Programme coordinator: Supervisor(s): Jan Terje Kvaløy	
Title of master's thesis: COMPETING RISKS ANALYSIS OF NORWEGIAN MORTALITY	
Credits: 60	
Keywords: Survival Analysis, Kaplan-Meier, Cumulative incidence Cox model, Fine-Gray model, Competing risks	Number of pages: 59..... + supplemental material/other: ..... Stavanger,.....14.06.2019..... date/year

COMPETING RISKS ANALYSIS OF NORWEGIAN MORTALITY

Thesis submitted at the University of Stavanger
in partial fulfillment of the requirements for
the degree of Master of Physics and Mathematics
(Statistics)

By:
Suhb-Eldin Abdulaziz Nusr

Department of Mathematics and Natural Science
University of Stavanger
Submission Date: June 2019
Supervisor: Jan Terje Kvaløy

Contents

List of Figures

List of Tables

Preface

I would like to express my deepest appreciation and it gives me great pleasure in acknowledging the support and help of my supervisor Jan Terje Kvaløy. He continually and convincingly guided me throughout the year. I would like to thank all staffs in the Department of Mathematics and Natural Science - University of Stavanger for their efforts to make it easier for the all students.

Abstract

Survival data analysis is a set of statistical methodologies that is used to model time until a certain event occurs. Competing risks data arise frequently in survival data from medical research in situations when individuals under study are exposed to more than one type of event such as death from different causes, and occurrence of one of these events prevent the occurrence of the event of interest.

This thesis introduces the conventional methods of survival analysis such as Kaplan-Meier and Cox proportional model, and methods which are used in presence of competing risks such as cumulative incidence and Fine-Gray model. The Norwegian mortality data in three countries where individuals were at risk to death from four death causes was used in this thesis to make comparisons between estimates of Kaplan-Meier and cumulative incidence and between the hazard rates estimated by Cox model and Fine-Gray model.

The low rate of overall death in the data of Norwegian mortality in three counties resulted in very small differences between the estimates of survival probabilities of Kaplan-Meier and cumulative incidence, and between hazard rates estimated by Cox model and Fine-Gray model, but there are some differences between the two models in estimating the impact of some covariates.

Chapter 1

1 Introduction to survival data

Survival data (or survival times) is the simplest form of event history data. Survival data analysis is one of the statistical methodologies that is used to model time until a certain event occurs (time to event). It is been used for a long time in many different fields of study and research, for instance, economics, demography, and is widely used in medical statistics. With this methodology, we usually use collected information about an event under consideration itself (an event of the interest), as well, information about how much time it took, or the time elapsed from a well defined start time for each of individuals to the event of interest occurs (a survival time). Thus, as it is illustrated in figure 1.1, there are three essential elements that should be distinctly defined: a time origin (start point), a scale for measuring time(day, month, year,..etc) and an event to occur, or end point, in other words (e.g. death of a patient, failure of a machine, ..etc) [5].

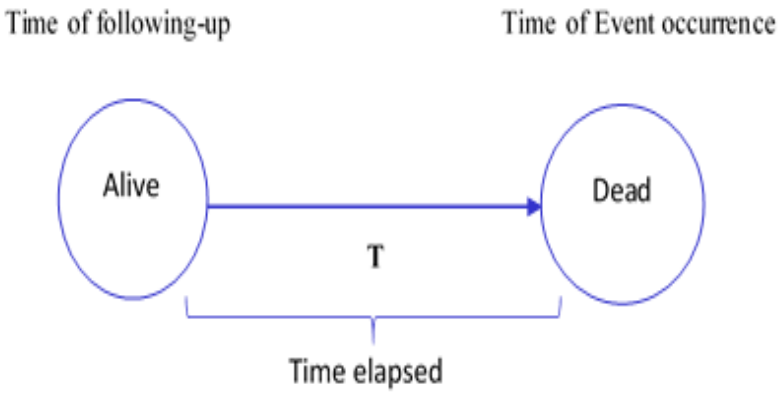

Figure 1.1: Survival data

1.1 Censoring

In the statistical analysis of survival, the response (survival time) is the exact elapsed time from the start point (the time origin), which often is the time since each of the individuals entered the study, to the end point, or time of event occurrence (death, failure, ..etc). Nevertheless, something may happen and hinder following-up some of individuals along the study period, such as the individual may suddenly disappear for some reasons. Hence, we might lose it from the sample. Therefor with long time studies we will, most probably, work on data set that contains such individuals. These individuals constitute so called (censored) observations [7]. Discarding censored observations from the data set will affect on the consistency and lead to bias estimate [5]. Hence, we need methods which takes the censoring into account.

1.1.1 Right sencoring

An individual is defined as right censored when it is lost to follow-up from some time point and onwards. This means we just know that the event of interest occurs some time after a certain time. Figure 1.2 illustrates the concept of right censoring assuming two individuals, $unit1$ and $unit2$. $Unit1$ who would have had the event of interest at time T_{unit1}, which is the true time for the event of interest (e.g. death), but due to right censoring we only know that T_{unit1} is greater than a certain number C_{unit1}. This implies that the only information we have about the survival time for this individual, $unit1$ here, is it does not experienced the event of interest until time C_{unit1}. The number C_{unit1} here is the time when $unit1$ was last seen before losing following-up [5]. This information is important and has to be used in subsequent analysis. An individual is right censored, as well, if it does not experienced the event of interest along to the end of the study period. $Unit2$ in figure 1.2 is a second example for right censoring, but this time due to that the individual, $Unit2$ here, does not experienced the event of interest (e.g. was still working, did not die, ..., etc) until the study period has finished.

1.1.2 Informative and non-informative censoring

In survival analysis data, as it has been described in 1.1.1, a subject is censored when it is lost to follow-up due to one of some reasons that are unrelated to the study (drop out of the study, end of the study,..etc). This usual type of censoring is known as *non-informative censoring*. However, in some special situations cen-

soring of a subject occurs due to a reason related to the study, and this special type of censoring is called *informative censoring* [27].

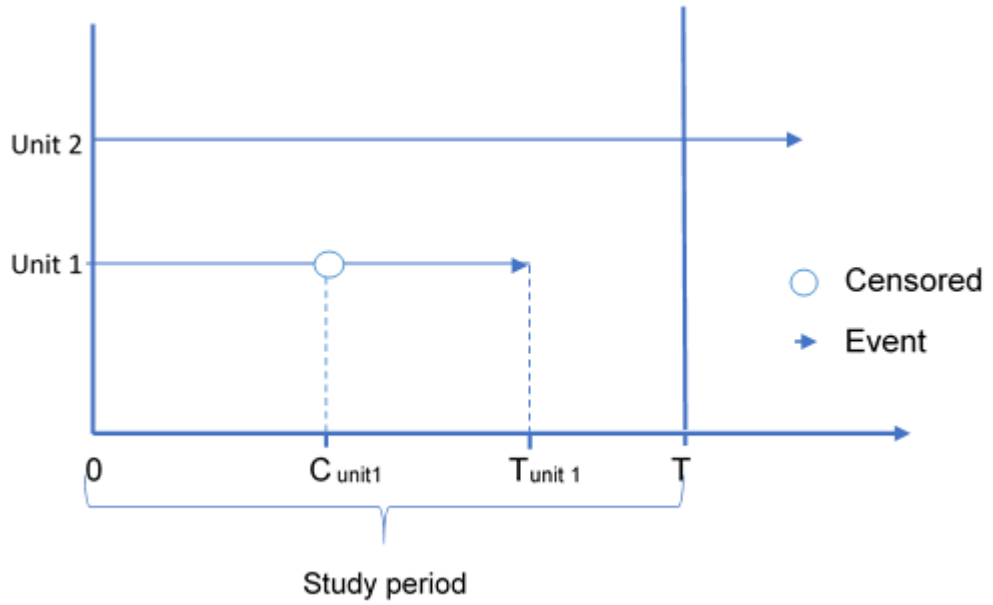

Figure 1.2: Right censoring

1.2 Left truncation

Left truncation, or delayed entry, is a well known concept in survival studies. To explain the concept of the left truncation, let us assume that we are about to make a study about the mortality of a cohort who is living in somewhere, for instance, a county. Let's say, the follow-up study started on the first of January 1960 and included all those persons alive and aged 60 years old or above on that date, and it was decided in advance that the study (following up) will take 20 years. This means the study should end on 31 December 1979. So, the start event is becoming 60 aged and the final event is death. The individuals who entered late, say 65 years old, on the first of January 1960, would not have been included if he had died at age, let's say, 63. Hence, in the analysis, we must condition on that this individual was alive at 65, or in other words, we say this individual is left truncated.

1.3 Survival function and hazard function

The survival and hazard function are a key concept in survival analysis. The survival curve is defined as a statistical graph of the survival studies of a group of patients, machine ...etc showing the survival percentage along a study time.

In studies of time to an event, the function that evaluates the probability that a patient, a machine or any other subject of interest will survive beyond a certain time $t > 0$ is a well known as the survival function [9]:

$$S(t) = \mathrm{P(T > t)} = \int_t^{\infty} f(u)\,du = 1 - F(t) \tag{1.1}$$

Where S(t) is the survival function and T is the time to event (is a random variable with density function $f(u)$ and cumulative distribution $F(t)$) . The survival function is known, as well, as the reliability function [20]. Figure 1.3 below shows an example for survival function of a study with data of 10 years period. In this figure the x-axis represents time in years, and the y-axis is represents the probability of subjects surviving. From the figure the proportion surviving, or the probability of that a subject will survive more than one year is, obviously, equal to 0.84 .

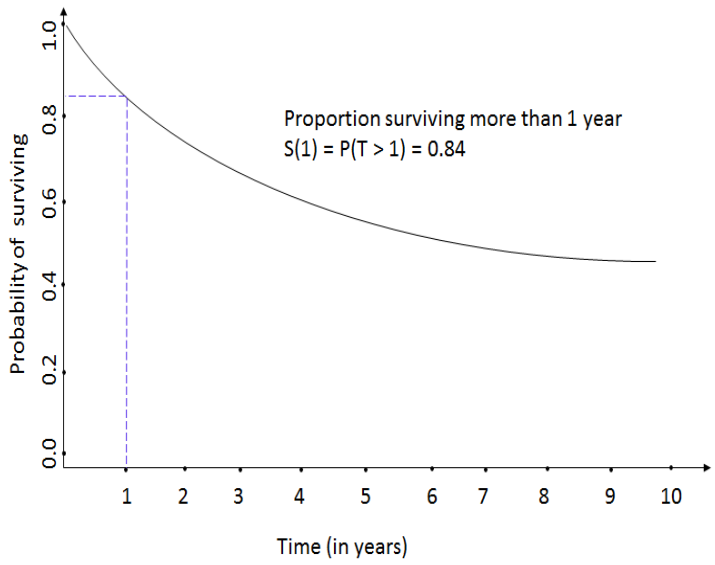

Figure 1.3: Survival function $S(t)$ for a study with 10 years period.

When we deal with survival time data , it is usually an aspect of interest to estimate periods of time which have the lowest and the highest probabilities of death (or generally of experience the event of interest) among individuals (subjects) who are still alive (did not experience the event of interest), therefore theses individuals are exposed to risk of experiencing the event of interest. The convenient method to estimate this risk is is the *hazard function h(t)*. The hazard function $h(t)$ is defined as the probability that an individual who still alive at time t dies or experiences the event of interest in a very small interval, assuming that the individual has already survived until the beginning of that interval divided by the length of the interval. This function has many different names such as the force of mortality, the conditional failure rate and the intensity function [21]. It could be expressed in term of *limit* as:

$$h(t) = lim_{\triangle t \to 0} \frac{P(t \leq T < (t + \triangle t)/T \geq t)}{\triangle t} \tag{1.2}$$

1.4 Kaplan-Meier

Kaplan-Meier , or product limit (PL) estimate, is one of the statistical methods that is used in survival analysis and aims to estimate a population survival curve from a sample where some of the data are right censored. It is one of the best methods that could be used to estimate survival function from censored data [1]. Kaplan-Meier (1958) were the first ones who brought a solution to estimate the survival curve in a very simple way taking into account the right censored data [2]. The Kaplan-Meier survival curve can simply be defined as the prospect of surviving along a certain interval (length of time), considering time in many small intervals [6]. The Kaplan-Meier analysis is based on three assumptions. The first assumption is that subjects which are censored, at any time, have the same survival probabilities as those others subjects which are still to be followed at that time.The second assumption is that the survival probabilities for subjects enlisted early and late in the enrollment period of the study are the same. The third assumption is that the event happens at the time specified. This method is performed by calculating the survival probability for each interval as the number of individuals who are surviving divided by the total number of individuals who are at risk. Individuals who have died or been censored are not regarded as individuals at risk. Individuals who censored will not be counted in the denominator after the censoring time. Then, we compute the probability of success which is equal to surviving beyond

a certain time point, then we multiply overall event times seen in the data. Thus the Kaplan-Meier estimator of the survival function $S(t)$ is [3, 4]:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{R(t_i)}\right) \tag{1.3}$$

Where $\hat{S}(t)$ is the estimated survival function, t_i, $0 < t_1 < t_2 < \ldots < t_n$ are the time points at which events occurred seen in the data, d_i is number of failure (number of occurrence of the event of interest) at point i, and $R(t_i)$ is risk size at time t_i (gives the number of individuals at risk before time t_i).

1.4.1 Example

The following data shows the survival time for some patients who entered a clinical study: 9, 12, 21, 27, 32, 39, 43, 43, 46*, 89, 115*, 139*, 181*, 211*, 217*, 261, 263, 270, 295*, 311, 335*, 346*, 365* (* refers to patients who were right censored on the corresponding day number).

Table 1.1 shows how to calculate the Kaplan-Meier estimates [1].

Time of event (t)	No of pnts died (d)	Live at the start of the day (n)	Proportion at risk surviving $(1 - \frac{d}{n})$	Probability of surviving beyond time t
9	1	23	0.9596	0.9596
12	1	22	0.9545	$0.9596 \cdot 0.9545 = 0.9130$
21	1	21	0.9524	$0.9130 \cdot 0.9524 = 0.87$
27	1	20	0.9500	0.8260
32	1	19	0.9474	0.7826
39	1	18	0.9444	0.7391
43	2	17	0.8824	0.6522
89	1	14	0.9286	0.6056
261	1	8	0.875	0.5299
263	1	7	0.8571	0.4542
270	1	6	0.8333	0.3785
311	1	4	0.75	0.2839

Table 1.1: Kaplan-Meier estimates for example 1.4.1 [1].

Figure 1.4 Kaplan Meier curve, with 95% confidence intervals (dashed lines), for example 1.4.1.

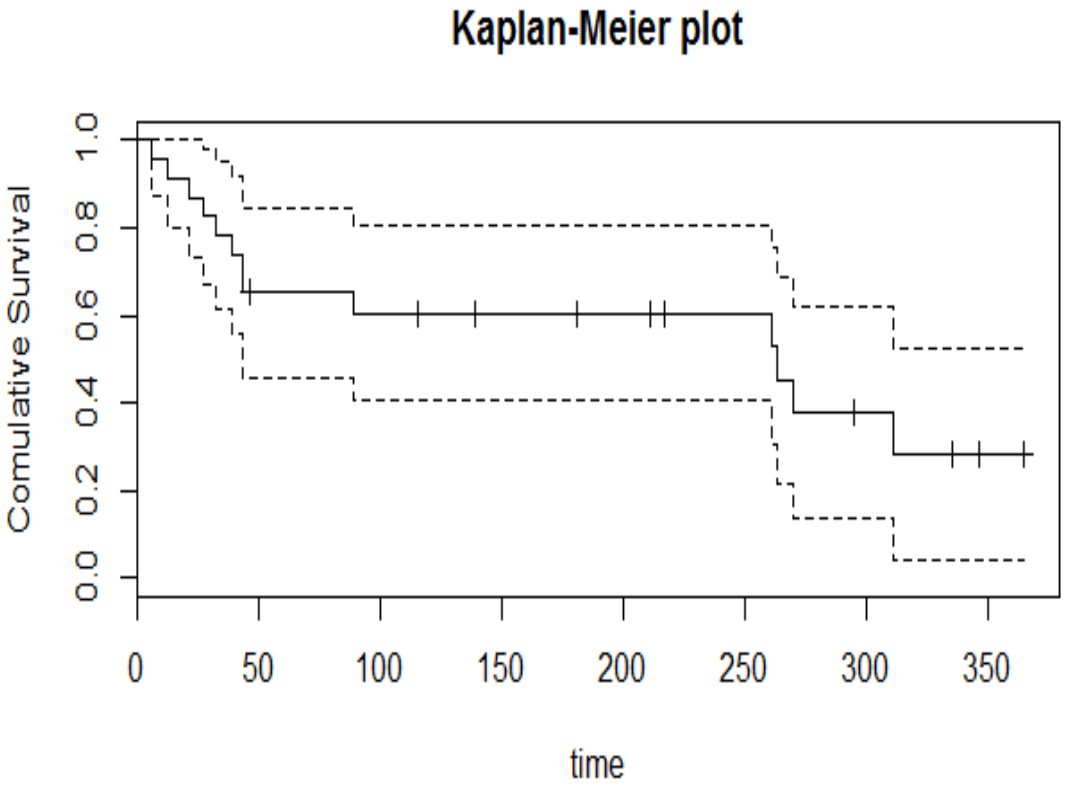

Figure 1.4: Kaplan-Meier curve for example 1.4.1, with 95% confidence intervals (dashed lines) [1].

1.5 The Log-rank Test

The log-rank test is a method in survival analysis used to test the equality of survival functions for *k-groups*. This test is performed under the null hypothesis that there is no difference in survival between the *k-groups* versus the alternative hypothesis that at least one of the curves differs [5]. In this test, we compute the expected number of events in each group at each time point. Suppose we have

two groups, $Group_1$and $Group_2$, for instance, then we can calculate $E_1$and E_2 which are the expected number of events summarized over all events time points in $Group_1$ and $Group_2$ respectively. Now, let $O_1$and O_2 refer to the total number of observed events in $Group_1$and $Group_2$ respectively. The test statistic for log-rank test can be calculated as [1]:

$$T = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{1.4}$$

Under the null hypothesis (H_0) of no difference between the groups, T follows a $\chi^2 - distribuation$. Therefore to test whether the null hypothesis (H_0) is significant or not, we compare the calculated test statistic T with the critical value from χ^2 table with a degree of freedom equal to $k-1$ where k is the number of groups.

1.5.1 Example

Suppose that the following data shows the survival times (in day) for some patients who entered a clinical study: 9, 13, 27, 38, 45*, 49, 49, *, 93, 118*, 118*, 126, 159*, 211*, 218, 229*, 263*, 298*, 301, 333, 346*, 353*, 362*(* refers to patients who were right censored on the corresponding day number). We are interested to test whether there is any difference in the survival times between these patients, and the patients in example 1.4.1 using log-rank test. Table 1.2 illustrates the log-rank test calculations for data of example 1.4.1 and 1.5.1.

Time of event	Total No of patients died in both groups (D)	No of patients died in $Group_2$ (O_2)	Alive at the start of the day	Alive at the start of the day in $Group_2$	Probability of death at the end of time	Expected Probability of death in $Group_2$ (E_2)	Expected Probability of death in $Group_1$ (E_1)
6	1	0	46	23	0.021134	0.5	
9	1	1	45	23	0.02222	0.52111	
12	1	0	44	22	0.02273	0.5	
13	1	1	43	22	0.02326	0.511628	
21	1	0	42	21	0.02381	0.5	
27	2	1	40	21	0.05	1.05	
32	1	0	39	20	0.02564	0.512821	
38	1	1	38	20	0.02633	0.526316	
39	1	0	37	19	0.02702	0.513514	
43	2	0	36	19	0.05556	1.055556	
49	2	2	32	18	0.0626	1.125	
89	1	0	31	16	0.03226	0.516129	
93	1	1	29	15	0.03448	0.517241	
126	1	1	25	12	0.04	0.48	
218	1	1	19	9	0.05263	0.473684	
261	1	0	17	8	0.05882	0.470588	
263	1	0	15	7	0.06666	0.466667	
270	1	0	14	7	0.07243	0,5	
301	1	1	12	6	0.09091	0.545455	
311	1	0	10	5	0.1	0.5	
313	1	1	9	4	0.11111	0.44444	
	24	11				12.22015	11.77985

Table 1.2: Log-rank test statistic for the data in examples 1.4.1 and 1.5.1.

Then we calculate the log-rank statistic using the above mentioned expression, expression (1.4), as:

$$T = \left(\frac{13-11.78}{11.78}\right)^2 + \left(\frac{11-12.22}{12.22}\right)^2 = 0.2481$$

The calculated test statistic T is equal 0.2481 . This value is smaller compared to the value of χ^2 table with degree of freedom 1, which is equal to the number of samples (2 here) minus 1. Hence, we accept H_0 that there is no significant difference in survival times between the two samples.

1.6 Cox Regression

Cox regression, or proportional hazards regression, is a method to model the effect of covariates on the time to event of interest. Cox regression is a semi-parametric model, and the major assumption in this model is that the effects of given covariates upon survival do not change over time. Once the assumptions of Cox regression are met, the Cox regression method can provide survival estimates that are better than the estimates that could be provided by the Kaplan-Meier function [8]. Kaplan-Meier is suitable when we have one categorical covariate, and the log-rank test, as well, is used when we have two or more groups and we are about to test whether there is any difference in the survival times between these groups or not. Using one of these two methods would not make us able to estimate the effect of other covariates upon survival times. One of the benefits of Cox regression model is it provides us with a way to estimate the effect of one or more than one covariates on survival times, and it can be used with discrete , continuous and dichotomous covariates [1].

1.6.1 Proportional Hazards

Cox regression can be formulated as follows:

$$h(t) = h_0(t)e^{(\beta_1 x_1 + ... + \beta_k x_k)} = h_0(t)e^{\beta' X} \tag{1.5}$$

Where $h_0(t)$ is unspecified baseline hazard, $\beta' = [\beta_1, \beta_2, ..., \beta_k]$ is $1 \times k$ vector of unknown parameters to estimate, and $X = [x_1, x_2, ..., x_k]$ is $k \times 1$ covariates vector. Estimating of these parameters, the parameters vector β', will give us information about the effect of covariates on the hazard rate.

The concept of proportional hazards is essential in Cox regression. If $h_1(t)$and $h_2(t)$ are two different hazards functions for two different individuals, then these two hazards functions are proportional if:

$$h_1(t) = \psi h_2(t) \Longrightarrow \psi = \frac{h_1(t)}{h_2(t)} \tag{1.6}$$

where $t \geq 0$, $\psi > 0$ (*positive constant*) is the proportionality constant. Equation (1.6) holds, as well, for the corresponding cumulative hazard function $H_i(t)$ with the same proportionality constant ψ [5].

Assume that we have two cases, *case*1 and *case*2 . To explain the concept of proportional hazards for the Cox model assume that each of these two cases has a hazard function:

*case*1= $h_1(t) = h_0(t)\, e^{\beta' X_1}$

*case*2= $h_2(t) = h_0(t)\, e^{\beta' X_2}$

using equation (1.6) gives: $\psi = \frac{h_1(t)}{h_2(t)} = \frac{h_0(t)e^{\beta' X_1}}{h_0(t)e^{\beta' X_2}} = \frac{e^{\beta' X_1}}{e^{\beta' X_2}} = e^{\beta'(X_1 - X_2)}$.

The proportionality constant $\psi = e^{\beta'(X_1 - X_2)}$is independent of time. This shows that the Cox-model implies a proportional hazard assumption.

If we consider the covariate j and assume that other covariates equal, then:

$\psi = e^{\beta_j(X_{1j} - X_{2j})}$

If $X_{1j} - X_{2j} = 1 \Longrightarrow \psi = e^{\beta_j} = HR_j$ which is the hazard ratio when the values of the covariates of the two cases equal, and there is just one unit difference between the two cases in the j^{th} covariate . That means the hazard ratio can simply be defined as change in hazard when a value of a covariate changes one unit. This gives us the interpretation of the hazard ratio HR, which can simply be interpreted as rate of the increase of the hazard when a covariate increases one unit.

1.6.2 Proportional hazard (PH) assumption

As it is mentioned above, Cox regression model is based on the proportionality assumption. This means that the proportionality constant, or the hazard ratio (ψ), should be constant along time. Which, obviously, means that the hazard ratio (ψ) is independent of time. From graphically aspect it means the hazard curves for various individuals should not cut across each other or, more precisely, they should be parallel on a log scale. If the hazard curves of each two different individuals do intersect each other, this is an indicator that the proportional hazard assumption is not met, then Cox proportional hazard model is unsuitable [14].

The above mentioned graphical methods for checking the violence of proportionality assumption of the Cox- model is based on the *scaled Schoenfeld residuals.* The scaled Schoenfeld residuals are basically independent of time. This

means the proportionality assumption holds only when a plot shows a smooth pattern against time [23].

1.6.3 The partial likelihood function PL

The likelihood function of Cox proportional hazard model considers only probabilities for events (not censoring), that is why it has been called a "partial" likelihood function PL [10]. The partial likelihood for β proposed by Cox [15, 16] without involving the baseline hazard $h_0(t)$, and it works similarly as the full likelihood. The baseline hazard in each term, will cancel out each other of the likelihood, therefor it will not be necessary to estimate it in the estimation of coefficients [10].

Let n be the number of individuals under study, δ_i be an indicator for failure or censoring (1= fail, 0= censored) for the event at time t_i, $i = 1, 2, \ldots, n$, and let $R(t_i)$ refer to the set of individuals who are surviving at time t_i (risk set at t_i). If $h_j(t)$ is the hazard function for the j^{th} individual at time t, then if an event occurred at time t_i (i.e. failure / death time), the probability of that the i^{th} individual may experience that event is [11]:

$$P([i] \diagdown t_i) = \frac{h_i(t_i)}{\sum_{j \in R(t_i)} h_j(t_i)}$$

Looking back on equation (1.5) this probability can be rewritten as:

$$P([i] \diagdown t_i) = \frac{h_0(t_i)\, e^{\beta' X_i}}{\sum_{j \in R(t_i)} h_0(t_i)\, e^{\beta' X_j}} =$$

$$\Longrightarrow P([i] \diagdown t_i) = \frac{e^{\beta' X_i}}{\sum_{j \in R(t_i)} e^{\beta' X_j}}$$

The above formula is known as risk probability of individual i at time t_i.

Assume that just one individual observes the event independently at each event occurrence time, then the partial likelihood for the coefficient β can be given by:

$$PL(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta' X_i}}{\sum_{j \in R(t_i)} e^{\beta' X_j}} \right]^{\delta_i} \tag{1.7}$$

It means we only multiply over the event times. All individuals contribute to the likelihood. The censored individuals contribute by being a part of $R(t)$until

their censoring time. Finally, we can estimate $\hat{\beta}$ by maximizing the log partial likelihood, $log\,[PL\,(\beta)]$.

In the following subsection we would explain some steps that are used in order to estimate the covariates parameters vector $\hat{\beta}$ based on the partial likelihood proposed by Cox (equation (1.7)).

1.6.3.1 Estimating the covariates parameters using PL

For the estimating of the parameters $\hat{\beta}$, Cox [16] recommended to treat the partial likelihood exactly as the full likelihood(*regular likelihood*). In this subsection we would show how to estimate parameters by making inference on the partial likelihood given by Cox.

First, we start with calculating the logarithm for equation (1.7). This will give us the log likelihood $(log\ partial-likelihood)$.

$$\begin{aligned} l\,(\beta) = log\,(PL(\beta)) &= log\left(\prod_{i=1}^{n}\left[\frac{e^{\beta' X_i}}{\sum_{j\in R(t_i)} e^{\beta' X_j}}\right]^{\delta_i}\right) \\ &= \sum_{i=1}^{n}\delta_i\left[\beta' X_i - log\left(\sum_{j\in R(t_i)} e^{\beta' X_j}\right)\right] \\ &= \sum_{i=1}^{n} l_i\,(\beta) \end{aligned} \tag{1.8}$$

Here, l_i is the contribution of the i^{th}individual in the log partial-likelihood. $i = 1,2,\ldots,n$.

Then we obtain the partial likelihood score function $U\,(\beta)$, by taking the first partial derivative of the log partial-likelihood (equation 7) with respect to the parameter (β).

$$U\,(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n}\delta_i\left[X_i - \left(\frac{\sum_{j\in R(t_i)} X_j e^{\beta' X_j}}{\sum_{j\in R(t_i)} e^{\beta' X_j}}\right)\right] \tag{1.9}$$

Finally, we can obtain the maximum partial-likelihood estimator, which estimate the parameter $\hat{\beta}$, by setting the score function $U\,(\beta)$ equal to zero, and then solve it.

The variance of the regression parameters can be estimated by calculating the negative of the second partial derivatives of the log partial-likelihood (equation

7), then the inverse of this matrix. Thus, we first obtain the so called observed information matrix $I\left(\hat{\beta}\right)$, and the inverse of this estimates the covariance matrix of the estimated regression parameters i.e:

$$I\left(\hat{\beta}\right) = -\left[\frac{\partial^2 l(\beta)}{\partial^2(\beta)}\right] \tag{1.10}$$

Now $I\left(\hat{\beta}\right)^{-1}$ estimates the covariance matrix, and in particular the diagonal elements of the inverse information matrix $I\left(\hat{\beta}\right)^{-1}$ provides the estimated variances of the corresponding coefficients.

1.6.4 Partial likelihood inference

Inferences about the regression parameters can be treated by hypothesis tests or confidence intervals. The two most common tests for testing the significance of one or more of the regression coefficients are Wald test and likelihood ratio test *LR* [24]. Many simulation studies revealed that the likelihood ratio test gives better inference, but more calculations, relative to the Wald test [25].

1.6.4.1 Statistical tests

As it is mentioned above, there are two tests used to test the statistical significance of one or more coefficients, the likelihood ratio test *LR* , and the Wald test. These two tests are essentially used to compare the difference between two models. One large model, and the second one with imposing some restrictions on some of the parameters of the first model, generally by assuming these parameters equal to zero(restrictive model), and it can be accomplished by omitting variables who are associated with these parameters. The likelihood ratio statistics is [12]:

$$LR = 2\left[l\left(\hat{\beta}\right) - l\left(\hat{b}\right)\right] \tag{1.11}$$

Where $l\left(\hat{\beta}\right)$ is the log likelihood of the large model, and $l\left(\hat{b}\right)$ is the log likelihood of the restricted model.

As an example, assume we were about to estimate the effect of some given variables on the mortality rate. Assume we first fitted the model considering some

variables such as , for instance, age, weight, civilian status, county and smoking habits, then we omitted some of these variables and fitted the model again considering, for instance, the variables age, weight and smoking habits. Here, the first fitted model is the large model, and the second one is the restricted model.

The likelihood ratio statistics is chi-square distributed with degree of freedom equal to the difference in the number of the parameters between the two models. Then the $p-value$ is calculated from $\chi^2(m)$, where m is the degree of freedom which is equal the difference in the number of the parameters between the two models. We judge the full model (the one with more variables) to significantly fit the data better than the restrictive model if the $p-value < 0.05$. The $p-value$ is calculated as $p-value = P(Y > LR)$, where Y is $\chi^2(m)$.

The Wald test is commonly used in multiple regression for testing the significance of the coefficient. In Cox regression, Wald test is used, as well, for testing the significance of a particular regression coefficient. Wald statistic has the following expression under the null hypothesis $\beta_j = 0$ [13]:

$$Z_j = \frac{\hat{\beta}_j}{S_j} \tag{1.12}$$

Where, $\hat{\beta}_j$ is the estimated coefficient, S_j is the estimated standard error of $\hat{\beta}_j$. S_j is provided by the square root of the corresponding j diagonal element of the inverse information matrix given by equation (1.10). Z_j here is approximately standard normal distributed.

1.6.4.2 Confidence interval

The calculation of the confidence interval for the Cox regression model coefficients is based on Wald statistic. The upper and lower limits of an approximate $(1-\alpha)100\%$ confidence interval can be calculated using the following formula [13]

$$\hat{\beta}_j \pm Z_{\frac{\alpha}{2}} S_j \tag{1.13}$$

Where $Z_{\frac{\alpha}{2}}$ is the critical value of the standard normal distribution.

Chapter 2

2 Data presentation

2.1 Causes of death and mortality in three Norwegian counties

Background

During the years 1974–78 all Norwegian (men and women) aged 35–49 years, who were living in three different Norwegian counties Oppland, Sogn og Fjordane, and Finmark were invited to a cardiovascular health screening test. A great per cent of the inhabitants participated in the screening and they gave, in addition, a self-report on their smoking habits. To the end of the year 2000, mortality of about 50 000 individuals was followed-up by record linkage with the cause of death registry at Statistics Norway. Here, the survival times are left-truncated at 40 year, and that is because of the risk of death for the individuals aged below 40 years old is low. In addition, all individuals are right- censored when they reach 70 years (unless they already died or censored before they turn 70 years old).

Table 2.1, below, shows the header and the first four rows of the above mentioned data of the causes of death and mortality in three Norwegian counties. In this work, we will use a subset of of 4000 individuals (2086 males and 1914 females) of a total set of 50 000 individuals, which is described above[17, 18]. These 4000 individuals were randomly selected from this cohort to study the mortality from the four causes of death:

- Death from cardiovascular disease (including sudden death).
- Death from cancer.
- Death from other medical causes.
- Death from alcohol abuse, chronic liver disease, accidents and violence

agesta	agesto	dead	dead1	dead2	dead3	dead4	Sex	Con	sbp	bmi	smk strt	smkgr
40.00	60.80	0	0	0	0	0	2	14	110	21.8	NA	1
44.43	57.65	1	0	0	1	0	2	14	120	30.4	NA	1
40.00	60.38	0	0	0	0	0	2	5	156	28.1	NA	1
41.11	66.29	0	0	0	0	0	2	14	130	24.9	26	2

Table 2.1: Data for causes of death and mortality in three Norwegian counties [18].

Coding[18]

- agesta: age of the individual when the health examination was tested (or 40 years if screened before that age).
- age sto: age of the individual in years at death or censoring.
- dead: refers to death from the all four causes (0 = censored, 1 = dead).
- dead1: refers to death from cancer (0 = censored or dead by other cause than cancer, 1 = dead from cancer).
- dead2: refers to death from cardiovascular disease, including sudden death (0 = censored or dead from other cause than cardiovascular disease, 1 = dead from cardiovascular disease)
- dead3: refers to death from other medical causes (0 = censored or dead from the other three death causes cancer, cardiovascular, and alcohol abuse, 1 = dead from other medical causes)
- dead4: refers to death from alcohol abuse, violence and accidents, and liver disease (0 = censored or dead from other causes than alcohol abuse, violence and accidents, and liver disease. 1 = dead from alcohol abuse, liver disease, and violence and accidents)
- sex: refers to individual sex (1 = male, 2 = female)
- Con: refers to three counties in Norway (5 = Oppland, 14 = Sogn og Fjordane, 20 = Finmark)

- sbp: refers to systolic blood pressure at health screening exam
- bmi: refers to body mass for the individual when the health screening exam was taken
- smk strt: refers to age when the individual started smoking
- smk gr: refers to four different smoking categories(1 = never smoked, 2 = former smoker, 3 = 1-9 cigarettes per day, 4 =10-19 cigarettes per day, 5 =20+ cigarettes per day, 6 = pipe or cigar)

2.2 Kaplan-Meier

In this section we will apply the procedures of Kaplan-Meier survival curve estimator (which is mentioned in 1.4), on the data of causes of death and mortality in three Norwegian counties to plot and discuss some survival curves.

Figure 2.1 shows Kaplan-Meier survival curves for the overall causes of death (death from cancer, cardiovascular diseases, alcohol abuse and other medical causes). Using this figure we could determine whether there was any difference in the mortality rate caused by the four causes of death among the gender or not. It is clear that the mortality from all causes among males and females who aged between 40 to 70 was grater among males (the blue curve) than females (the red curve) along the study period. By reaching 70 years old, about 91% of females had not experienced death, whereas about 76% of males had not experienced death.

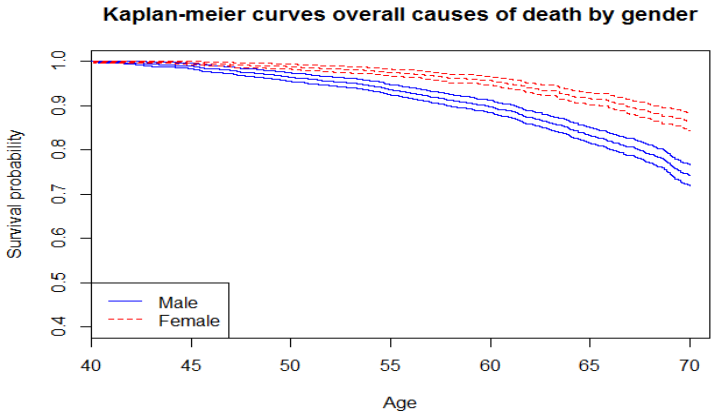

Figure 2.1: Kaplan-Meier curves overall causes of death by gender

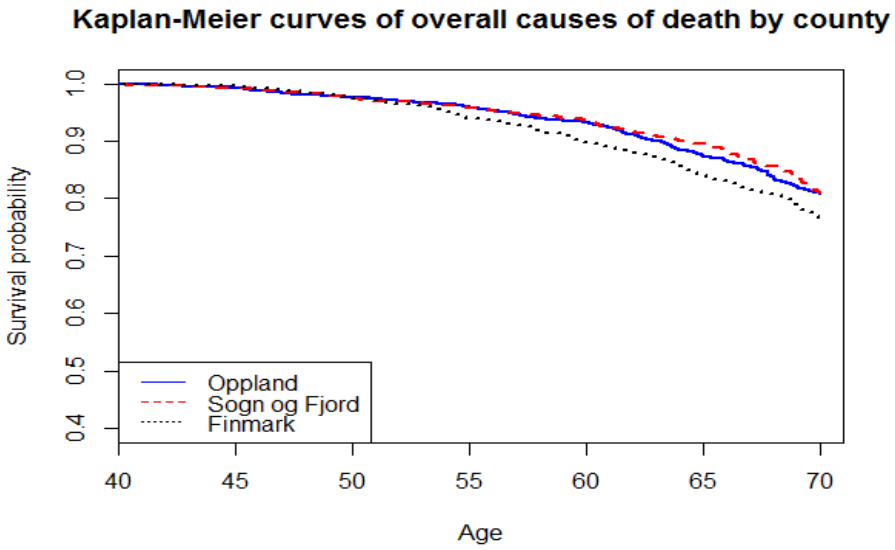

Figure 2.2: Kaplan-meier curves overall causes of death by county

The Kaplan-Meier survival curve by county is shown by figure 2.2. The survival probabilities in the three counties are approximately equal until turning about 54 years old when the survival probability for Finmark trend to be lower than the survival probability in Oppland and Sogn og Fjordane. The survival probability for Oppland became lower than the survival probability in Sogn og Fjordane after turning 61 years old. After getting 70 years old, the survival probabilities are approximately 0.78, 0.81 and 0.80 for Finmark, Oppland and Sogn og Fjordane, respectively.

Figure 2.3 shows Kaplan-Meier curves of death form each cause of death individually adjusted by gender (who were aged 40-70 years old). The mortality rates from cancer, cardiovascular, alcohol abuse and other medical causes among males were higher compared to rates of mortality from the same causes among females. The highest rate of death among females was death from cancer, whereas the highest rate of death among males was death from cardiovascular diseases. The probability of death from cancer after turning 70 years old for females was about 93%, whereas it was about 89% for males. The biggest difference in the mortality rate between males and females was in the death from cardiovascular diseases. There were about 96% females who had turned 70 years old and had not experienced death from cardiovascular, whereas there were bout 87% males had

survived. The rate of death from other medical causes and from alcohol abused is, obviously, not much. The difference between males and females in the mortality rate from these two causes was, as well, not too much, specially for the death from other medicals causes. Death from alcohol abuse among the males was higher compared to death from other medical causes, while the opposite holds for the death from these two causes among females. Probability of death from alcohol abuse and other medical causes after turned 70 year old among the males was about 97% and 98%, respectively, whereas it was 99% and 97% among the females.

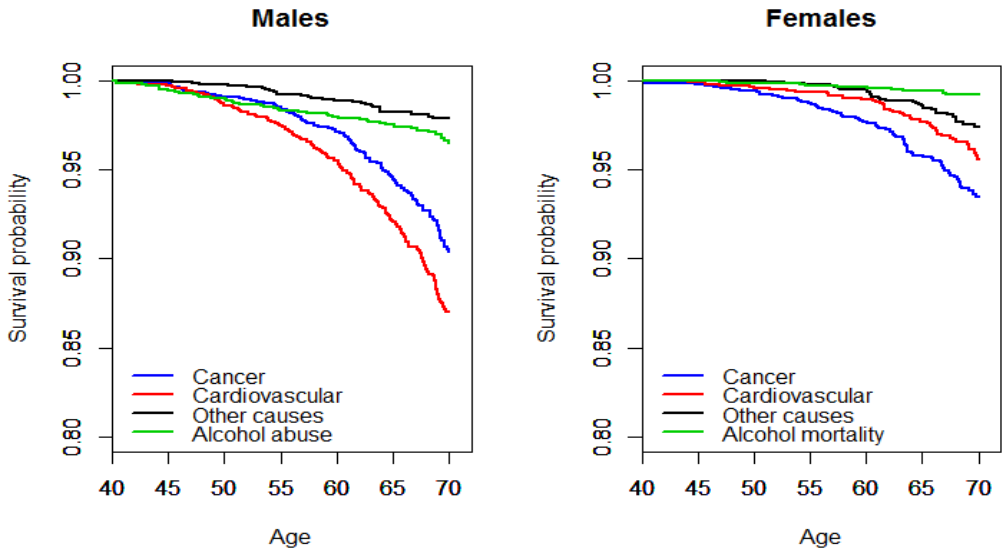

Figure 2.3: Kaplan-Meier curves for the four death causes among males *(to the left)* and females *(to the right)*.

Figure 2.4 illustrates Kaplan-Meier survival curves for rates of the mortality from the four death causes adjusted by the smoking habits. For the both causes, the smoking habits were worse prognosis in the survival probabilities of cancer and cardiovascular diseases, more than in the survival probabilities of alcohol abuse and other medical causes. The survival probabilities related to smoking habits was worse for cardiovascular diseases than for cancer among the all six levels of the smoking habits except for individual who were smoking 20+ cigarettes per day and never smoker. the biggest difference in the survival probabilities of

cancer and cardiovascular diseases related to the smoking habits were among pipe smoker which showed the worst survival probabilities for cardiovascular diseases. The probability of surviving of pipe smokers after turned 70 years old was about 78% from cardiovascular diseases , while it was 87% from cancer. Mortality rates from alcohol and other medical causes related to smoking habits was not large, and there were not even big differences in the survival probabilities between these two death causes except among pipe smoker where we could observe large differences *(compared to the other smoking habits)* specifically after turning 55 years old. More over, the survival probability of other medical causes was constantly greater, or almost equal, along ages 40-70 years old than survival probabilities of alcohol abuse, but it was not among individuals who were pipe smoker. The survival probability of cancer and cardiovascular diseases was clearly directly proportional with the smoking habits, specially with cardiovascular diseases.

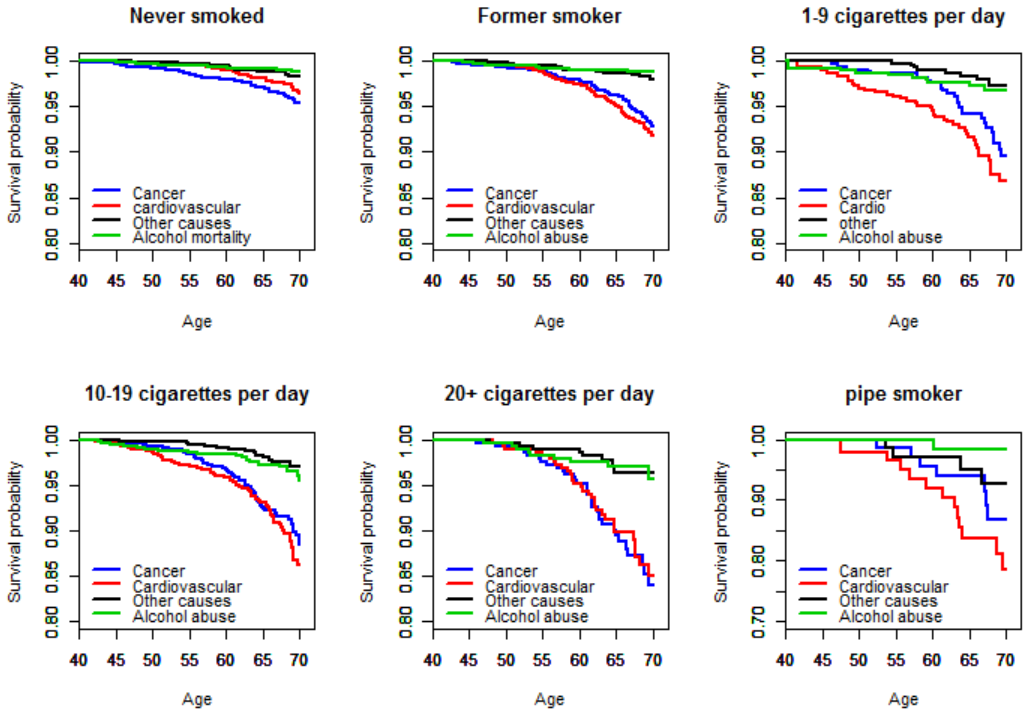

Figure 2.4: Kaplan-Meier curves for four causes of death adjusted by smoking habits.

Figure 2.5 represents Kaplan-Meier curves adjusted by the three counties Op-

pland, Sogn og Fjord og Finmark. The survival probability of cancer and cardiovascular diseases in the three counties was less compared to the survival probability of other medical causes and alcohol abuse. Cancer probability of surviving of individuals who turned 70 years old and were living in Oppland was 92.28%, 91.88% for cardiovascular disease, 96.99 % for other medical causes and 98.21% for alcohol abuse. There was not big difference between the two counties Oppland and Sogn og Fjord in the survival probabilities, but a bit notable observation was that survival probability for other medical causes was a little bit larger than for alcohol in the all three counties, but it was less survival probability for other medical cause than for alcohol for individuals who were living in Oppdal after they turned 64 years old. The worst survival prognosis was of cancer and cardiovascular diseases for individuals who were living in Finmark. There was even a notable difference between cancer and cardiovascular mortality in this county. Cardiovascular mortality was worse compared cancer mortality. Probability of survival after 70 years old for cardiovascular was about 88%, whilst for cancer it was about 91%. The difference in mortality rate between this Finmark country compared to its counterparts Oppland and Sogn og Fjord could only be interpreted due to lifestyle

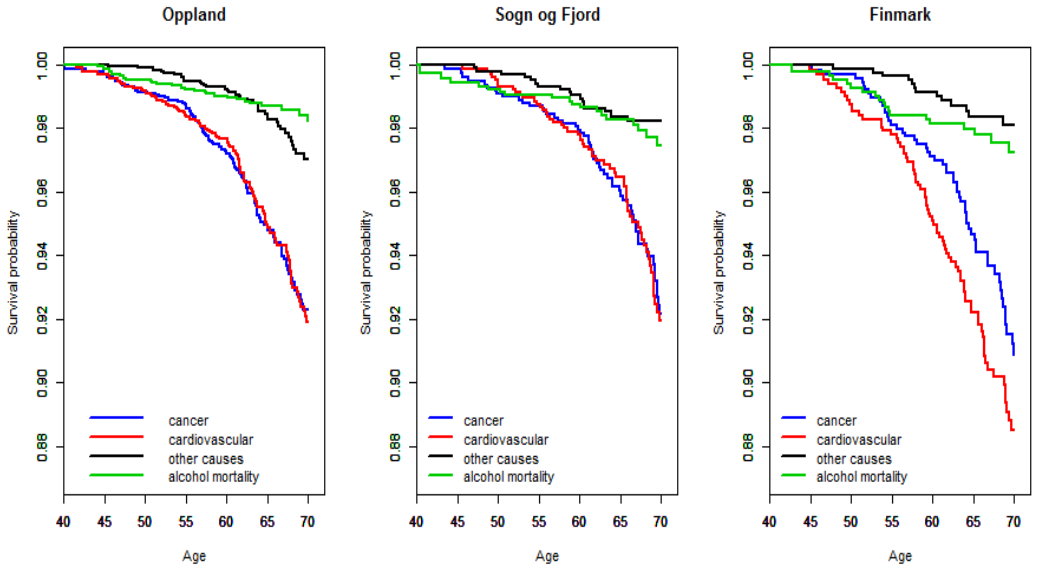

Figure 2.5: Kaplan-Meier curves for the four death causes adjusted by the three counties.

In Kaplan-Meier survival analysis, the assumption of independence of censoring is essential, and this method deals with only one type of failure, death for example, regardless the cause of this particular failure. That means Kaplan-Meier survival method will not provide good estimations if this assumption of censoring independency does not hold. Looking on the data of Norwegian mortality in three counties which we have just worked on by conducting the Kaplan-Meier survival method in this chapter. There are four different death causes that compete each other and might lead to death which is the event of interest. If we just look at the data of the death, for example, from cancer, we could find many individuals have been registered as censored while, in fact, they died from one of the three other causes. Then if the assumption of independence of censoring does not hold the Kaplan-Meier method will fail to estimate the survival times of such data. Therefore we need to find some alternative methods to estimate the survival time for these types of data when there are more than one cause for the event.

Analysis of such survival data, when the subjects are exposed to experience more than one type of event of interest, or to experience the same event of interest from multiple causes called competing risk analysis.

2.3 Cox regression model estimates for data of Norwegian mortality

In this section we conducted an analysis by performing Cox proportional hazards model on the above mentioned data of the Norwegian mortality in three different counties. We fitted the model using the six covariates, which we have described in 2.1, to estimate their effects on the survival time or, in other words, to investigate differences in mortality between these covariates. For every cause of death, we performed the uni-variable analysis by fitting the model using every covariate individually, and then performed the multi-variables analysis by including the all covariates simultaneously. After fitting multi-variable analysis we removed the non-significant covariates. In addtion, we did not include the covariate *started smoking (smk strt in the tables header)* when we fitted the multi-variable analysis. The reason for this is simply because this covariate has many missing values (many individuals were not smoker), and then we end up analysing a subset of data only containing smokers. Firstly, we fitted the model with respect to the overall causes of death, then with respect to every cause of death individually *(The cause specific Cox proportional model). P-values* of each reference level of the categorical covariates (smoking habits and county) are found by likelihood

ratio tests, while the others are found by Wald tests.

2.3.1 Cox proportional hazard model for overall causes mortality

Table 2.2 shows the results of uni-variable and multi-variable analysis of Cox proportional hazard model for overall causes of death. Staring with the uni-variable model we can, obviously, see that the all six covariates are statistically significant. For sex, the hazard ratio is 0.49 indicates that being female decreases the probability of experiencing death from overall causes (risk of death) by a factor of 0.49, or by 51% compared to male, holding the other covariates constant. In the multi-variables analysis, the sex is significant as well, and the hazard ratio is equal to 0.62 which means being female reduces the risk of death by 38% compared to being male. The systolic blood pressure is statically significant for both, uni-variable and milt-variables analysis. In the uni-variate analysis, its hazard ratio is equal to 1.02 indicating a positive association between systolic blood pressure and the overall causes mortality. This means the excepted hazard of death is 1.02 times higher in an individual who is one unit systolic blood pressure higher than another. The hazard ratio of systolic blood pressure is the same for the multi-variables analysis as for the uni-variable. The categorical covariate smoking habits has six factors with the factor never smoked as a reference level. All five factors are statically significant compared to the reference level, never smoked. Considering the smoking habits and holding the other covariates constant, the hazard ratio compared to persons who never smoked is 1.67 times higher for former smoker persons, 2.64 times higher for 1-9 cigarettes a day smokers, 2.82 times higher for 10-19 cigarettes a day smokers, 3.8 times higher for 20+ cigarettes a day smokers and the highest hazard ratio is for pipe smokers whose hazard ratio is 4.3 times higher compared to persons never smoked. In the multi-variables analysis, however, the all five factors compared to the reference level remain statically significant. The hazard ratio compared to persons who never smoked (reference level) is 1.38 times higher for former smoker persons, 2.43 times higher for 1-9 cigarettes a day smokers, 2.52 times higher for 10-19 cigarettes a day smokers, 3.08 times higher for 20+ cigarettes a day smokers and the highest hazard ratio is again for pipe smokers whose hazard ratio is 3.19 times higher compared to persons who never smoked. The covariate categorical county has three factors Sogn og Fjordane = couny 14, Finmark = county 20 , and the reference level Oppland = county 5. The overall causes mortality for Sogn og Fjordane county compared to oppland county is not statically significant neither in the uni-variable analysis nor the multi-variables analysis. However, the overall causes mortality for Finmark

county is statistically significant with a hazard ratio equal to 1.30 which means living in this county increases the hazard ratio by 30 % than living in Oppland county. Not in a uni-variable model. The body mass has a positive association with the overall causes mortality in the uni-variable analysis, whereas it is not significant in the multi-variables analysis. The body mass hazard ratio is equal 1.02. This means holding the all other covariates constant, a one unit increase in the body mass is associated with 2% increase in the expected hazard. The covariate smoking start, which referring to the age when an individual started smoking, is highly significant. Time since an individual started smoking is negatively associated with the overall causes mortality. The hazard ratio of 0.95 indicates that the expected hazard ratio decreases by 95% for any persons who started smoking at age of one year older than other person who started smoking one year younger.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
	Uni-variable			Multi-variables		
sex	0.49	(0.41 , 0.58)	$1.06{\cdot}10^{-15}$	0.62	(0.52 , 0.75)	$4.49\cdot10^{-7}$
sbp	1.02	(1.01 , 1.02)	$2.00\cdot10^{-16}$	1.02	(1.01 , 1.02)	$2.00\cdot10^{-16}$
never smoked	1	ref	$2{\cdot}10^{-16}$	1	ref	$2.2\cdot10^{-16}$
former smk	1.67	(1.29 , 2.15)	$8.4\cdot10^{-5}$	1.38	(1.06 , 1.79)	0.02
1-9 cigar	2.64	(1.98 , 3.51)	$2.58\cdot10^{-11}$	2.43	(1.82 , 3.23)	$1.28\cdot10^{-9}$
10-19 cigar	2.82	(2.21 , 3.61)	$2.00\cdot10^{-16}$	2.52	(1.96 , 3.23)	$4.09\cdot10^{-13}$
20+ cigar	3.8	(2.84 , 5.08)	$2.00\cdot10^{-16}$	3.08	(2.27 , 4.16)	$3.00\cdot10^{-13}$
pipe-cigar	4.31	(2.77 , 6.71)	$9.59{\cdot}10^{-11}$	3.19	(2.03 , 5.02)	$5.61\cdot10^{-7}$
county OPP-L	1	ref	0.01			
county S&F	0.94	(0.77 , 1.15)	0.56			
county F	1.30	(1.07 , 1.57)	0.01			
bmi	1.02	(1 , 1.05)	0.05			
smk strt	0.95	(0.94, 0.97)	$7.19{\cdot}10^{-12}$			

Table 2.2: Uni-variable and multi-variables Cox regression estimates overall four death causes

2.3.2 Cox proportional hazard model for cancer mortality

Table 2.3 shows the uni-variable and multi-variable analysis to relate the six covariates to time to death from cancer by conducting a Cox proportional hazard regression model on the data of the Norwegian mortality in three counties. The covariate sex is significant only in the uni-variable model (the same as it was in

the overall death cause). The blood pressure, in contrast to result of overall death model, is not significant neither in uni-variable model nor multi-variable model. county seems to not having any effect on death from cancer neither in uni-variable model nor multi-variable model, and the same was body mass. Age when start-smoking was statistically significant and negatively associated with cancer mortality, but not in multi-variate model. The smoking grade is highly significant and associated with cancer mortality in both, uni-variable and multi-variables analysis. It is the same compared to overall cause of death model with an exception for the level *former smoker* which is significant in overall cause of death model, but not significant for cancer mortality. The smoking habits is worse related to prognosis of cancer mortality.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
		Uni-variable			Multi-variables	
sex	0.71	(0.54 , 0.93)	0.01			
sbp	1.00	(1.00 , 1.01)	0.36			
never smoked	1	ref	$3{\cdot}10^{-8}$	1	ref	$3{\cdot}10^{-8}$
former smk	1.48	(0.99 , 2.23)	0.06	1.48	(0.99 , 2.23)	0.06
1-9 cigar	2.16	(1.34 , 3.46)	0.00	2.16	(1.34 , 3.46)	0.00
10-19 cigar	2.54	(1.72 , 3.76)	$3.09{\cdot}10^{-6}$	2.54	(1.72 , 3.76)	$3.09{\cdot}10^{-6}$
20+ cigar	3.82	(2.42 , 6.02)	$8.46{\cdot}10^{-9}$	3.81	(2.42 , 6.02)	$8.46{\cdot}10^{-9}$
pipe-cigar	3.06	(1.38 , 6.80)	0.01	3.06	(1.38 , 6.80)	0.01
county OPP-L	1	ref	0.55			
county S&F	0.91	(0.66 , 1.26)	0.58			
county F	1.12	(0.81 , 1.57)	0.48			
bmi	1.00	(0.96 , 1.04)	1.00			
smk strt	0.95	(0.93 , 0.98)	$4.42{\cdot}10^{-5}$			

Table 2.3: Uni-variable and multi-variables Cox regression estimates for death from cancer.

2.3.3 Cox proportional hazard model for death from cardiovascular disease

Table 2.4 shows estimates of the Cox proportional hazard model for death from cardiovascular. All covariates are statically significant, in the uni-variate model, except the level of county *county S&F* (Sogn og Fjordane). Tthe smoking habits, as it in death of cancer, is significant in both, uni-variable and multi-variable model. In addition, smoking habits level *former smoker* is significant in contrast

it in death from cancer model. Sex and blood pressure are highly significant in uni-variable and multi-variable model. The covariate county F (Finnmark county) is significant in the uni-variable and multi-variable model (county OPP-L is Oppland wich is the reference level of the covariate county). Body mass and smoking-start age have effect on the death from cardiovascular in the nin-variate model, but they are not significant in the multi-variable model.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
	Uni-variable			Multi-variables		
sex	0.30	(0.22 , 0.41)	$9.38{\cdot}10^{-15}$	0.38	(0.27 , 0.52)	$1.92{\cdot}10^{-9}$
sbp	1.03	(1.02 , 1.03)	$2{\cdot}10^{-16}$	1.03	(1.02 , 1.04)	$2{\cdot}10^{-16}$
never smoked	1	ref	$6.4{\cdot}10^{-15}$	1	ref	$4.62{\cdot}10^{-9}$
former smk	2.47	(1.62 , 3.78)	$2.96{\cdot}10^{-5}$	1.67	(1.08 , 2.58)	0.02
1-9 cigar	4.01	(2.52 , 6.37)	$4.41{\cdot}10^{-9}$	3.37	(2.11 , 5.36)	$3.19{\cdot}10^{-7}$
10-19 cigar	3.86	(2.54 , 5.86)	$2.42{\cdot}10^{-10}$	2.99	(1.95 , 4.57)	$4.40{\cdot}10^{-7}$
20+ cigar	4.76	(2.90 , 7.80)	$6.40{\cdot}10^{-10}$	2.94	(1.76 , 4.90)	$3.66{\cdot}10^{-5}$
pipe-cigar	6.99	(3.61 , 13.54)	$8.11{\cdot}10^{-9}$	4.05	(2.05 , 8.01)	$5.58{\cdot}10^{-5}$
county OPP-L	1	ref	0.01	1	ref	0.01
county S&F	0.95	(0.69 , 1.30)	0.73	0.96	(0.70 , 1.33)	0.82
county F	1.57	(1.17 , 2.10)	0.00	1.51	(1.12 , 2.04)	0.01
bmi	1.05	(1.02 , 1.09)	0.00			
smk strt	0.96	(0.94 , 0.98)	0.00			

Table 2.4: Uni-variable and multi-variables Cox regression estimates for death from cardiovascular.

2.3.4 Cox proportional hazard model for death from other medical causes

Table 2.5 shows the uni-variable and multi-variables analysis for the death from other medical causes. The blood pressure is significant and related to increase the hazard of death from the death from other medical causes in both, the uni-variable and the multi-variables analysis. Hazard of death from other medical causes is effected by the smoking's level 20+ cigarettes per day, and the pipe smokers. These two are significant in the uni-variable and the multi-variables model. Smoking-start age is related to increase the hazard of death from other medical causes in , but it is not significant in the multi-variable model.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
		Uni-variable			Multi-variables	
sex	1.04	(0.65 , 1.69)	0.86			
sbp	1.02	(1.01 , 1.03)	0.01	1.02	(1.00 , 1.03)	0.00
never smoked	1	ref	0.10			
former smk	1.14	(0.57 , 2.25)	0.72			
1-9 cigar	1.50	(0.65 , 3.45)	0.34			
10-19 cigar	1.58	(0.78 , 3.14)	0.19			
20+ cigar	2.29	(1.00 , 5.27)	0.05			
pipe-cigar	1.47	(1.47 , 12.82)	0.01			
county OPP-L	1	ref	0.60			
county S&F	0.79	(0.66 , 1.26)	0.42			
county F	0.78	(0.81 , 1.57)	0.43			
bmi	0.94	(0.88 , 1.02)	0.15			
smk strt	0.92	(0.88 , 0.97)	$2{\cdot}10^{-4}$			

Table 2.5: Uni-variable and multi-variables Cox regression estimates for death from other medical causes.

2.3.5 Cox proportional hazard model for death from alcohol abuse

The estimations of Cox proportional hazard regression model for alcohol abuse mortality are shown in table 2.6. The risk of death from alcohol abuse is significant and highly related with the sex in the uni-variable and multi-variable model. The hazard ratio of 0.26 indicates that being female reduces the hazard of death by 74% compared to being male. The smoking-start age is significant in the uni-variable model, but not in multi-variable model. The smoking is significant for the all level except *former smoked* and *pipe* smoker levels in the uni-varaible model, but in the multi-variable model only smoking 10 - 19 is significant. The smoking-start age is significant and related to death from alcohol abuse in the uni-variable model.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
		Uni-variable			Multi-variables	
sex	0.26	(0.14 , 0.48)	$2.52 \cdot 10^{-5}$	0.26	(0.14 , 0.51)	$6.39 \cdot 10^{-5}$
sbp	1.01	(1.00 , 1.03)	0.03			
never smoked	1	ref	0.01	1	ref	0.02
former smk	0.96	(0.41 , 2.24)	0.92	0.63	(0.26 , 1.49)	0.29
1-9 cigar	2.38	(1.02 , 5.56)	0.05	1.97	(0.84 , 4.63)	0.12
10-19 cigar	2.85	(1.42 , 5.74)	0.00	2.08	(1.02 , 4.23)	0.04
20+ cigar	3.42	(1.46 , 8.00)	0.01	2.01	(0.84 , 4.81)	0.12
pipe-cigar	1.59	(0.21 , 12.16)	0.66	0.81	(0.10 , 6.27)	0.84
county OPP-L	1	ref	0.27			
county S&F	1.29	(0.70 , 2.36)	0.41			
county F	1.66	(0.90 , 3.04)	0.10			
bmi	1.05	(0.98 , 1.12)	0.17			
smk strt	0.94	(0.90 , 0.99)	0.01			

Table 2.6: Uni-variable and multi-variables Cox regression estimates for death from alcohol abuse.

Chapter 3

3 Competing risks analysis

Cox proportional hazard models and Kaplan–Meier estimates of survival curves are widely used to assess the effects of some given covariates on the survival time and to describe the survival tendency, respectively. These two statistical methods are appropriate when we deal with one type of event, for example death, regardless of its cause. A specific situation appears when interest is in a particular cause of failure, whereas some different causes are present. These other causes alter the probability of occurrence of the event of interest from the predetermined cause. Hence, estimating the survival probability of a specific cause treating the other causes as censoring, which are present at the time, will underestimate the survival probability, and this the case of *competing risk*s.

In this chapter we would introduce the problem of competing risks which is the main topic of this work. The chapter starts with giving a brief definition for the problem of competing risks, then the alternative methods which are more appropriate than the traditional methods, such as Kaplan-Meier and Cox regression model, to estimate the survival times and effect of some covariates in presence of competing risks. We will introduce two of these methods. Firstly, we will present a method that replaces Kaplan-Meier, and one of the appropriate estimates of the failure probabilities, namely, the cumulative incidence plots. Then we will present Fine-Gray model, in order to show and discuses the estimates which we got after conducting these two methods on the previously mentioned *(in chapter 2)* data of the Norwegian mortality in three counties in the following chapter.

3.1 The problem of competing risks

In survival analysis, as we previously discussed, we aim to estimate time elapsed from a certain time point to occurrence of a certain type of event *(event of interest)*. But sometimes, specially in medical's studies, the subjects can be exposed to experience more than one type of event *(failure)*, or to experience a particular event *(failure)* from more than one cause. For example in transplant studies, if the aim was to estimate time to relapse, then death of patient from transplant is an another event that competes and can hinder occurrence of the event of the interest (relapse) [19].

Competing risks problem appears in the situations where subjects are exposed

to experience more than one type of event. These events compete with the event of the interest, and occurrence of anyone of these events prevent the occurrence of event of the interest. The competing risks problem concern, as well, the situations where subjects are exposed to experience a single event of interest *(failure)* from more than one cause of failure, and occurrence of one of these other failure causes prevent and hinder occurrence of the event due to a particular and predetermined cause of failure *(cause of interest)* [22]. In the situation of the data that we have already previously introduced in section 2.1, data of the Norwegian mortality in three counties, all inhabits were exposed to risk of experiencing the death from one of four different causes of death that compete with each other. There was only one event of interest, which is death, but there were more than one cause *(risk)* of death *(the event)*. Assume that we were interesting in estimating the survival time of cancer mortality. Here, death due to anyone of the other three causes, *cardiovascular, alchol abuse and other medical cause*s, will prevent occurrence of death *(the event)* from cancer. Hence, considering the probabilities of survival of the other causes in parallel with the probability of survival of the cause of interest is inevitable. Figure 3.1 represents a graphical model of the competing risks problem for the data of the Norwegian mortality in three counties with individuals initial state which is alive (did not experience the event of the interest yet), and the four competing end points(death from cancer, cardiovascular, alcohol abuse or from other medical causes).

The competing risks problem is not only limited to the medical field. It can be an issue in many other fields of studies concerning survival data analysis. In the industrial reliability, for instance, if the interest was failure of a particular component in a system that could disrupt the system, whereas failure of some other different components in the same system which could also disrupt the system was possible can be counted as a competing risks problem [22].

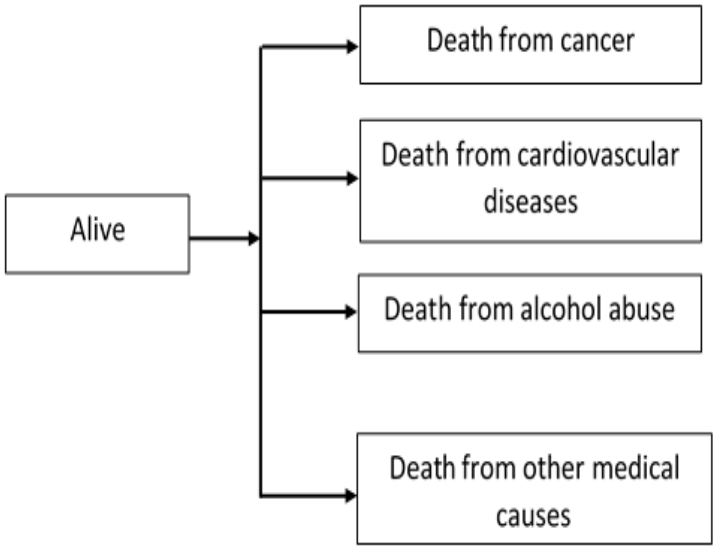

Figure 3.1: A graphical model of competing risks problem for the data set of the Norwegian mortality in three counties due to four death causes

Using conventional survival analysis approaches, such as the Kaplan-Meier etimate, in presence of these competing events (events in case of Norwegian mortality data are death from one of the four causes) will not precisely estimate the survival probability of event of the interest from the predetermined failure cause because Kaplan-Meier method estimates survival probability of only one type of event, for instance death from cancer, and treats death due to the competing risk (death due to cardiovascular, alcohol abuse and other medical causes) as censored [19]. If we estimate the survival function for one cause of failure using Kaplan-Meier and let death from other causes be treated as censoring we assume that the other causes of death are independent from the cause of interest (which might not be realistic) and we estimate the survival probability in a hypothetical world where the cause of interest is the only cause of death.

That is why it was imperative to propose some alternative statistical methods that could replace the traditional approaches, such as Kaplan-Meier, and give better estimates for the survival probability. During the last two decades, it has been proposed some methods to analyze survival data in presence of competing risks[19]. In the following two subsections we will introduce two of these methods.

3.1.1 Cumulative incidence function

In competing risk analysis two quantities are important the *cumulative incidence function (CIF)*, and *cause specific hazard function*. In contrast to the Kaplan-Meier method, the *cumulative incidence function* estimates probability of occurrence of an event of interest taking into account probability of occurrence of competing risks (other events). In competing risk observations a subject is exposed to experience an event from a set of different causes. Therefore the *cause specific hazard* function is a key concept in the competing risk analysis. This function gives hazard of failing due to a distinct cause of failure while other causes are present. Let $\varepsilon \in (1,2,..,K)$ be causes of failure (assuming there are K observable causes of failure), then the probability of failure due to the k^{th} cause at time t for subjects who have not yet experienced any event at that time is defined by [22]:

$$h_k(t) = lim_{\triangle t \to 0} \frac{P(t \leq T < (t+\triangle t), \varepsilon = k/T \geq t)}{\triangle t} \tag{3.1}$$

where T is the time of failure and $k = 1,2,...,K$.

Equation (3.1) is known by *the cause specific hazard function*. Then, cumulative cause specific hazard of the k^{th} cause of failure is given by :

$$H_k(t) = \int_0^t h_k(s)\, ds$$

Then define

$$S_k(t) = e^{(-H_k(t))} \tag{3.2}$$

$$S(t) = e^{\left(-\Sigma_{k=1}^{K} H_k(t)\right)} \tag{3.3}$$

$S_k(t)$ in equation (3.2), is the survival probability of cause k of failure which is the probability of having not failed from the k^{th} cause of failure at time t, and equation (3.3) describes the overall survival probability which is the probability of having not failed from any of the K causes of failure at time t.

The *cumulative incidence function* for the k^{th} cause of failure $I_k(t)$ is the probability of failing due to the k^{th} cause of failure prior or at time t.

$$I_k(t) = Prob(T \leq t, \varepsilon = k)$$

The *cumulative incidence function* can be expressed using the cause specific hazard as the follow:

$$I_k(t) = \int_0^t S(s)\, h_k(s)\, ds \tag{3.4}$$

Methods, such as *Kaplan-Meier*, which treat censoring due to competing event as non-informative censoring overestimate the failure probability compared to methods that take into account the competing event. Hence, they might give misleading estimates when the probability of the competing cause is high. To explain that let 1 denotes death from cancer as the event of interest, and 2 denotes death due to cardiovascular as the competing event. Then the cumulative hazard functions for these two causes are $H_1(t)$, $H_2(t)$, respectively. The survival function at a distinct time t considering only the event of interest is $S_1(t) = e^{-H_1(t)} \geq e^{-H_1(t)-H_2(t)} =$ the overall survival function taking into account the hazard of competing event *(death due to cardiovascular)* [26].

In the presence of competing risks, the Kaplan-Meier method estimates a different quantity, $S_k(t)$ which is survival probability in the hypothetical world where cause k is the only cause of failure, and treats events from other causes than cause k of interest as censored, whereas the cumulative incidence function take into account the fact that the failure might occur from other causes. In addition, even if we want to estimate $S_k(t)$ using Kaplan-Meier and assuming non-informative censoring might lead to wrong results as the non-informative censoring might be wrong. The Kaplan-Meier estimates the cumulative probability of occurring of event k prior and at time t by [22].

$$1 - S_k(t) = \int_0^t S_k(s)\, h_k(t) \tag{3.5}$$

Since $S(t) \leq S_k(t)$, then

$$\int_0^t S_k(s)\, h_k(s) \geq \int_0^t S(s)\, h_k(s) = I_k \tag{3.6}$$

From equation (3.5), this implies that $I_k \leq 1 - S_k$. The equality holds when there is no competing hazards (e.i $\sum_{j=1, j\neq k}^{K} H_j(t) = 0$). This implies that Kaplan-Meier

method, in the presence of competing risk, overestimates the probability of failure due to cause k.

The *cumulative incidence functio*n is known, as well, by many other names such as *crude cumulative incidence function* and *sub-distribution function* [22].

To estimate the *cumulative incidence functio*n, let $t_1 < t_2 < .... < t_N$ be the distinct time points at which any cause of failure occur. Assume the number of subjects failing at time t_j by failure cause k is given by d_{kj}, then the total number of failures from any cause of the K causes of failure at t_j could be given by $d_j = \sum_{k=1}^{K} d_{kj}$. Now, as in equation (1.4), let $R\left(t_j\right)$ gives the number of individuals at risk, which is the number of individuals which have not failed from any cause of failure before time t_j *(risk set).* From equation (1.4) the survival function regardless the cause of failure is estimated by Kaplan-Meier by:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{R\left(t_j\right)}\right)$$

The cause specific hazard, equation (3.1), gives the conditional probability of failing from cause k at time t_j given still alive just before time t_j is:

$$h_k\left(t_j\right) = Prob(T = t_j, \varepsilon = k / T > t_{j-1})$$

This probability can be estimated by:

$$\hat{h_k}\left(t_j\right) = \frac{d_{kj}}{R\left(t_j\right)} \tag{3.7}$$

Then the survival function can be written as:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \sum_{k=1}^{K} \hat{h_k}\left(t_j\right)\right)$$

The portability of failing due to cause k at time t_j, *unconditional probability*, $P_k\left(t_j\right) = Prob\left(T = t_j, \varepsilon = k\right)$ is estimated by the product of the cause specific and the survival probability at t_j as:

$$\widehat{P_k}\left(t_j\right) = \hat{h_k}\left(t_j\right) \cdot \hat{S}\left(t_{j-1}\right) \tag{3.8}$$

Sum of equation (3.8) over all time points $t_j \leq t$, estimates the cumulative incidence of cause k

$$\hat{I}_k(t) = \sum_{t_j \leq t} \widehat{P}_k(t_j) \tag{3.9}$$

3.1.2 Fine-Gray model

To estimate the effect of some given covariates on hazard rate we introduced the Cox model in subsection 1.6. In equation (1.2) we defined the hazard function, generally, when a subject is exposed to fail due to only one type of event (one case of failure) and there is no any other competing event. In the presence of competing risk the hazard function is represented by the cause specific hazard function (equation (3.1)). In subsection 2.3 we applied the Cox model on the Norwegian mortality in three counties data set (which contains competing risk observations), to each competing event (death cause), individually. The results been shown by tables 2.2, 2.3, 2.4, 2.5 and 2.6 are the estimated effects of the given covariates on each of the four *cause specific hazard function* (death from cancer, cardiovascular diseases, alcohol abuse and other medical causes). The problem with the cause specific hazard function is it treats competing events as censoring events. The Cox hazard model estimates the effect of some given covariates on the *cause specific hazard function*. Therefore the main problem with the Cox model appear when we want to use the result to model the effect of the covariates on the CIF. Hence, in the presence of competing risk using the Cox model may lead to biased results [28]. However, Cox model remains one of the best and widely used method in the survival analysis and it is appropriate when the aim is to estimate the effect of the covariates on the cause-specific hazard [29]. The hazard function has one-one correspondence to the CIF only when there is no competing risk, therefore the hazard ratio gives the risk of the study subjects, whereas in the presence of competing risk, there is no one-one correspondence between the hazard foundation and CIF therefore the effect of the covariates on hazard can not directly be linked to its effect on CIF [22, 26]. Let $H_1(t)$,$H_2(t)$,...$H_K(t)$ denote K cause specific hazard functions for K causes of failure, then from equation (3.4) of the cumulative incidence function for the k^{th} cause of failure:

$$I_k(t) = \int_0^t S(s)\, h_k(s)\, ds = \int_0^t e^{-(H_1(s)+H_2(s)+\ldots+H_K(s))} h_k(s)\, ds$$

Hence, $I_k(t)$ depends on all the cause specific hazards.

Fine and Gray [30] introduced a method to link covariates to the cumulative incidence function CIF. The subdistribution hazard function proposed by Fine and Gray to estimate the covariates effect on the CIF in the presence of competing risk is defined by [22, 30] :

$$h_k(t) = lim_{\triangle t \to 0} \frac{P(t \leq T < (t + \triangle t), \varepsilon = k/T \geq t \cup (T < t \cap \varepsilon \neq k))}{\triangle t} \tag{3.10}$$

Which gives the risk of failure due to the k^{th}cause in individuals who have not yet experienced an event from k^{th} cause (individuals who experienced the competing event will be included in the risk set)

Chapter 4

4 Estimating cumulative incidence function and Fine-Gray model

In this chapter we will estimate and present cumulative incidence plots after performing the cumulative incidence method on the data of Norwegian mortality in three Norwegian counties. We will make comparisons between these estimates (the cumulative incidence estimates) and the Kaplan-Meier estimates which have been presented in chapter 2. Then, we will present estimates of Fine-Gray model and compare these estimates with Cox model estimates which we already have presented in the chapter 2.

4.1 Cumulative incidence

This section starts with presenting the cumulative incidence for each gender followed by presenting the cumulative incidence for each category of smoking grade, which has six levels, and then the cumulative incidence for each of the three counties.

4.1.1 The cumulative incidence curves by gender

Figure 4.1 illustrates the cumulative incidence curves of death from four causes of death among Norwegian males and females aged 40-70 years old. The highest incidence of death among males is death from cardiovascular diseases. By turning 70 years old, the estimated incidence of death from cardiovascular for males is 11.95%, while Kaplan-Meier method estimated the probability of dying from cardiovascular for males to be 12.96% (i.e. a relative change 0.08). Cumulative incidence of death from cardiovascular diseases for females is about 4.2%, whereas using Kaplan-Meier method gave about 4.4% (i.e. a relative change 0.06). The difference in the death from cancer between gender is smaller compared to the death from cardiovascular. The estimated incidence of death from cancer are 8.6% and 6.3% for males and females, respectively. The Kaplan-Meier estimated the risk of death from cancer by 9.62% and 6.53 %, respectively, for males and females (i.e. a relative change 0.10 for males and 0.03 for females). The difference in the estimated risk of death from cancer between the two methods, Kapan-Meier

and cumulative incidences, is very small for females. It is because of the risk of death from the other three competing causes is very low. It is the same for risk of death from other medical causes (relative change for male is 0.10 and 0.06 for female), and alcohol abuse (relative change for male is 0.09 and 0.05 for female) as the difference between Kaplan-Meier and the cumulative incidence is small. How ever the relative change for other medical causes and alcohol abuse is a little bit bigger than the relative change for cancer and cardiovascular, because here the risk of death from the sum of competing causes is a little bit higher.

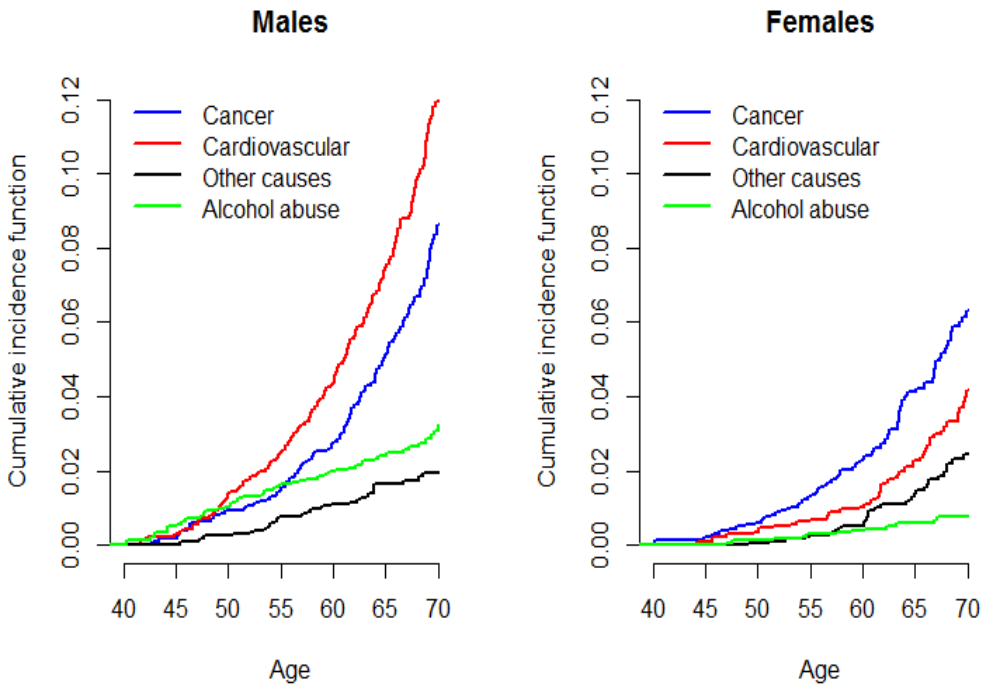

Figure 4.1: The cumulative incidence curves for the four causes of death by gender

4.1.2 The cumulative incidence curves by smoking habits

Figure 4.2 represents the cumulative incidence curves for the four causes of death by smoking habits. Bad smoking habits implies worse prognosis in survival probabilities of cancer and cardiovascular diseases than in survival probabilities of alcohol abuse and other medical causes. In the following we will present esti-

mates of the cumulative incidence (for individuals who have reached 70 years old) and make a comparison between these estimates and those which were previously estimated by Kaplan-Meier.

There is a small difference between the two methods in estimating of cardiovascular and cancer death for individuals who were smoking 1-9 cigarettes per day. The estimated cumulative incidence for death from cardiovascular at age 70 is 12.23%, while Kaplan-Meier was 13.16% giving a relative change of 0.07. For death from cancer, the cumulative incidence is 9.2% while the Kaplan-Meier estimated it by 10.4% which gives a relative change of 0.12. The cumulative incidence for death from other medical causes is estimated by 0.02%, whereas the Kaplan-Meier estimated it by 0.03 (i.e. a relative change 0.13). For alcohol abuse, the cumulative incidence is 3.96%, while Kaplan-Meier estimated it by 4.54% (i.e. a relative change 0.13). Clearly, the relative change between the Kaplan-Meier and the cumulative incidence for other medical causes and for alcohol abuse are bigger compared the relative change for cancer and for cardiovascular. It is the same for the smoking group *10-19 cigarettes* since the relative change between the two methods are 0.11, 0.09, 0.13 and 0.13 for cancer, cardiovascular, other medical causes and alcohol abuse.

For individuals who were smoking more than 20 cigarettes per day, the estimated cumulative incidence for death from cancer and cardiovascular are, respectively, 10.4% and 12.43%. Risk of death estimated by Kaplan-Meier for these two causes were, respectively, 16.07% and 15.03%. This means the relative change between the two methods for this group of smoking is 0.35 for death from cancer, and 0.17 for death from cardiovascular. In other hand, the relative change for the death from other medical causes is 0.30 (cumulative incidence estimated by 2.07% and Kaplan-Meier estimate was 3.07%), and the relative change for death from alcohol abuse is 0.1 (cumulative incidence estimated by 3.96% and Kaplan-Meier estimate was 4.38%).

The difference between death from cancer and from cardiovascular is, obviously, bigger among pipe smokers compared to the difference of death from these two causes among the others smoking group. In other hand death from other medical causes is more related to pipe smoking than the others smoking groups. The cumulative incidence risk of death from cancer and cardiovascular are, respectively, 11.04% (relative change of 0.17) and 19.23% (relative change of .1). For other medical causes the cumulative incidence is 6.12% (relative change of 0.15), and for alcohol abuse it is 1.42% (relative change of 0.15).

Though Kaplan-Meier and the cumulative incidence did not give big differences, however, the relative difference for the most unlikely causes of death, death

from other medical causes or from alcohol abuse, is mostly bigger than the relative difference for the most likely causes of death, death from cancer or from cardiovascular.

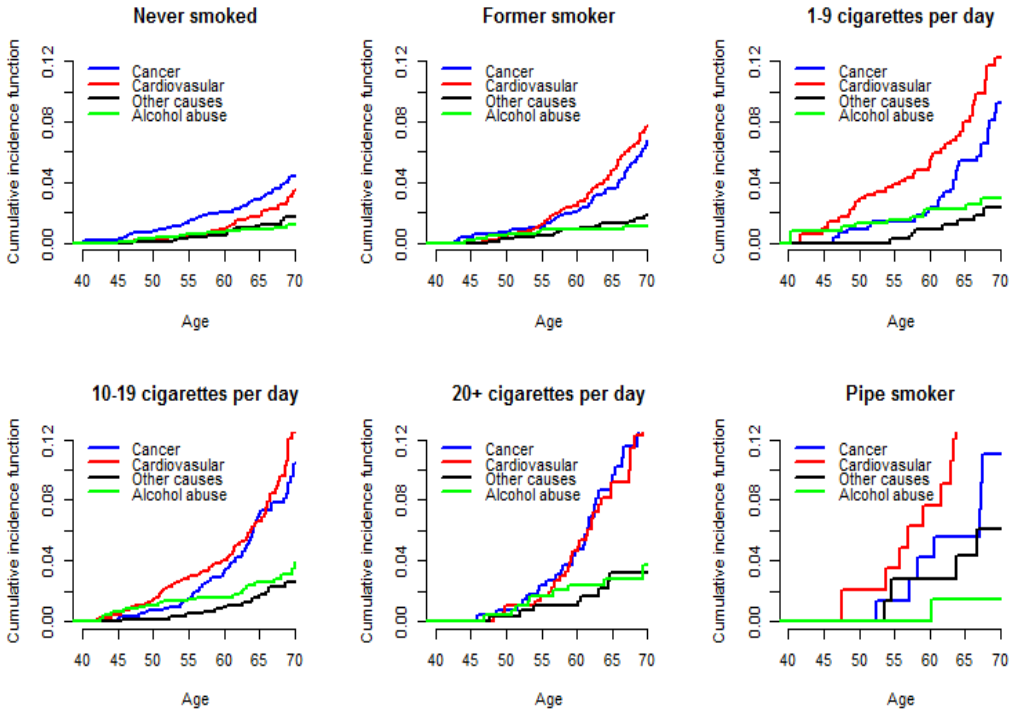

Figure 4.2: The cumulative incidence curves for the four causes of death by smoking habits

4.1.3 The cumulative incidence curves by county

Figure 4.3 illustrates the cumulative incidence curves by county. After turning 70 years old, the estimated cumulative incidence risk of death from cancer for individuals who were living in Oppdal county is 7.26% (relative change of 0.06), for individuals who were living in Sogn og Fjordane county is about 7.33% (relative change of 0.07) and for individuals who were living in Finmark is 8.27% (relative change of 0.09).

The cumulative incidence risk of death from cardiovascular for individuals (turned 70 years old) who were living in Oppdal county is 7.59% (relative change

of 0.04), for those who were living in Sogn og Fjordane county is about 7.52% (relative change of 0.04) and for individuals who were living in Finmark is 10.76% (relative change of 0.07).

For individuals who were living in Oppdal county the estimated cumulative incidence risk of death from alcohol abuse is 1.66% (relative change of 0.07), for those who were living in Sogn og Fjordane county is 2.40% (relative change of 0.06) and for individuals who were living in Finmark is 2.57% (relative change of 0.08).

Probability of death from other medical causes (for all ages) is the smallest in the three counties except for individuals who were turned 63-70 years old and were living in Oppdal county. The cumulative incidence risk of death from other medical causes for individuals who were living in Oppdal, Sogn og Fjord and Finmark are 2.71%(relative change of 0.1), 1.70% (relative change of 0.05) and 1.74% (relative change of 0.10), respectively.

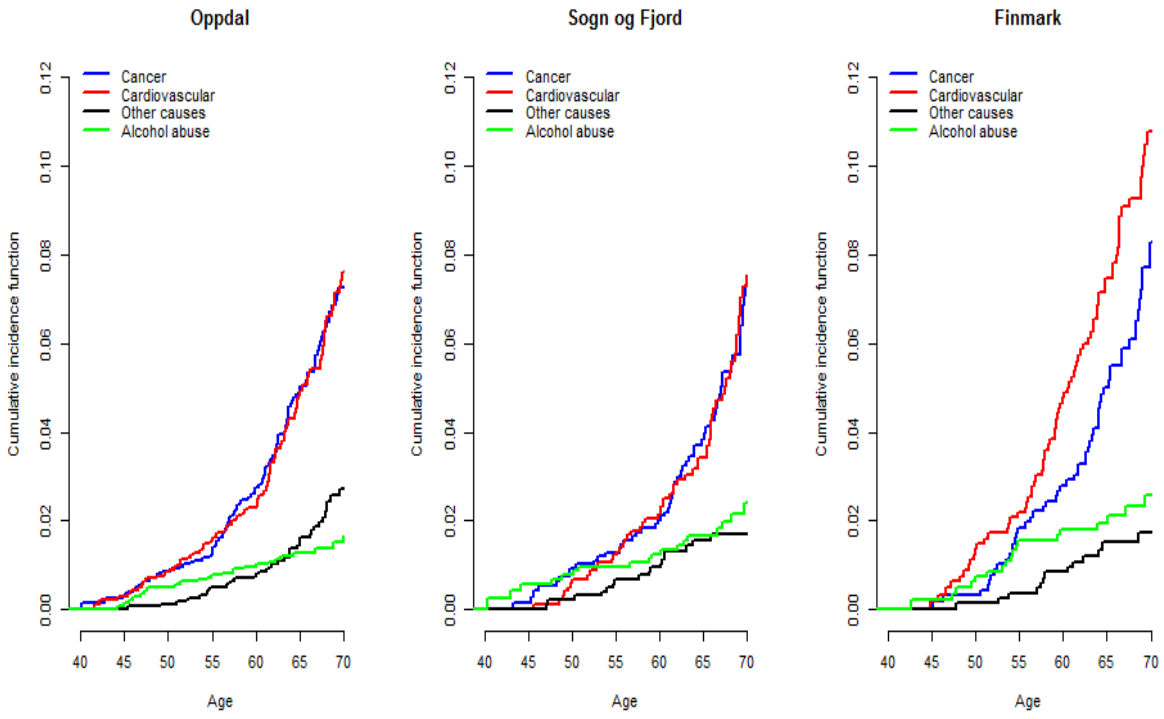

Figure 4.3: The cumulative incidence curves for the four causes of death by county.

4.2 Fine-Gray model

This section start with presenting uni-variable and multi-variable Fine-Gray estimates for death from cancer, then for death from cardiovascular followed by uni-variable and multi-variable fine-Gray estimates for death from others medical causes and for death from alcohol abuse.

4.2.1 Fine-Gray model for death from cancer

Table 4.1 shows the uni-variable and multi-variable Fine-Gray estimates for the death from cancer. In the uni-variable analysis the covariates sex, smoking grade and age when started smoking are significant, and smoking grade is the only significant covariate in the multi-variable analysis. It is the same results that have been estimated by Cox model for the same cause of death, 2.3, where sex, smoking grade and age when started smoking were significant, and smoking grade was the only significant covariate in the multi-variable analysis as well.

There are no big differences in the the hazard ratio estimated by Cox model and by Fine-Gray model. For the uni-variable Cox model analysis the hazard ratio of the covariate sex were 0.71, whereas the hazard ratio for the significant levels of smoking grade were 2.16, 2.54, 3.82 and 3.06 for the levels 1-9 cigarettes, 10-19 cigarettes, 20+ cigarettes and pipe smoker, respectively. For the covariate *age when started smoking* Cox model estimated the hazard ratio by 0.95 and there is no big different between this value and the hazard ratio estimated by Fine-Gray model. Similarly to the uni-variable Cox model analysis the *smoking grade* is the only significant covariate in the uni-variable Fine -Gray model.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
		Uni-variable			Multi-variables	
sex	0.75	(0.57 , 0.98)	0.03			
sbp	1.00	(1.00 , 1.01)	0.27			
never smoked	1	ref	$1{\cdot}10^{-6}$	1	ref	$1{\cdot}10^{-6}$
former smk	1.49	(0.99 , 2.24)	0.06	1.49	(0.99 , 2.24)	0.06
1-9 cigar	2.09	(1.30 , 3.36)	0.00	2.09	(1.30 , 3.36)	0.00
10-19 cigar	2.27	(1.54 , 3.38)	$3.54{\cdot}10^{-5}$	2.27	(1.54 , 3.38)	$3.54{\cdot}10^{-5}$
20+ cigar	3.36	(2.13 , 5.30)	$1.82{\cdot}10^{-7}$	3.36	(2.13 , 5.30)	$1.82{\cdot}10^{-7}$
pipe-cigar	3.27	(1.47 , 7.26)	0.00	3.27	(1.47 , 7.26)	0.00
county OPP-L	1	ref	0.7			
county S&F	0.88	(0.66 , 1.26)	0.45			
county F	1.03	(0.81 , 1.57)	0.87			
bmi	1.01	(0.97 , 1.05)	0.75			
smk strt	0.96	(0.94 , 0.99)	0.00			

Table 4.1: Fine-Gray model for death from cancer

4.2.2 Fine-Gray model for death from cardiovascular disease

Table 4.2 shows the uni-variable and multi-variable Fine-Gray model analysis for the death from cardiovascular. In the uni-variable analysis the all covariates are statically significant, except the level of county *county S&F* (Sogn og Fjordane) exactly as the same as the uni-variable Cox model for death from cardiovascular 2.4. There are noa big difference in the hazard ratios for Fine-Gray model compared to the hazard ratios for the corresponding covariates estimated by the Cox model. However, the difference between the two models in estimating of the hazard ratio for the covariate smoking grade (of the two levels *1-19 cigarettes* and *20+ cigarettes*) is a little bit bigger compared to the difference between the others covariates. It is exactly the same with the multi-variable model that there are no concrete and clear differences between the two model in estimating the hazard ratios of and, again, the biggest difference is in the hazard ratio for the covariate smoking grade (of the two levels *1-19 cigarettes* and *20+ cigarettes*). On the contrary to Cox model the covariate body mass is significant in the multi-variable analysis.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
	Uni-variable			Multi-variables		
sex	0.31	(0.23 , 0.42)	$3.07 \cdot 10^{-14}$	0.39	(0.28 , 0.53)	$7.08 \cdot 10^{-9}$
sbp	1.03	(1.02 , 1.03)	$2 \cdot 10^{-16}$	1.03	(1.02 , 1.03)	$2 \cdot 10^{-16}$
never smoked	1	ref	$6 \cdot 10^{-14}$	1	ref	$13.35 \cdot 10^{-8}$
former smk	2.52	(1.65 , 3.86)	$1.93 \cdot 10^{-5}$	1.72	(1.12 , 2.66)	0.01
1-9 cigar	4.01	(2.52 , 6.38)	$4.31 \cdot 10^{-9}$	1.22	(2.11 , 5.39)	$3.58 \cdot 10^{-7}$
10-19 cigar	3.56	(2.35 , 5.40)	$2.56 \cdot 10^{-9}$	2.84	(1.85 , 4.36)	$1.64 \cdot 10^{-6}$
20+ cigar	4.23	(2.58 , 6.94)	$1.07 \cdot 10^{-8}$	2.71	(1.63 , 4.54)	0.00
pipe-cigar	7.72	(3.99 , 14.95)	$1.34 \cdot 10^{-9}$	4.29	(2.13 , 8.67)	$4.53 \cdot 10^{-5}$
county OPP-L	1	ref	0.01	1	ref	0.03
county S&F	0.95	(0.69 , 1.32)	0.78	0.95	(0.69 , 1.32)	0.77
county F	1.57	(1.17 , 2.12)	0.00	1.46	(1.08 , 1.98)	0.02
bmi	1.06	(1.03 , 1.09)	0.00	1.04	(1.00 , 1.08)	0.03
smk strt	0.97	(0.95 , 0.99)	0.00			

Table 4.2: Uni-variable and multi-variable Fine-Gray model for death from cardiovascular disease.

4.2.3 Fine-Gray model for death from other medical causes

Fine-Gray estimates for death from other medical causes are shown in table 4.3. Comparing these estimates to the Cox model estimates for death from the same cause of death (death from other medical causes) shown in table 2.5 one can observe some substantial differences regarding the significance of some covariates. Sex, blood pressure (*sbp*), smoking grade and age when started smoking (*smk strt*) are significant in the uni-variable model, whereas in the Cox model sex and smoking grade were not. In the multi-variable analysis, in contrast with the Cox multi-variable analysis, smoking grade is significant, and the hazard ratio of blood pressure, which is significant, does not differ much from the hazard ratio estimated by the Cox model. There are not big differences in the hazard rates estimated by the Cox model and the Fine-Gray model, and it could be because of the very low mortality rate due to other medical causes.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
		Uni-variable			Multi-variables	
sex	0.27	(0.14 , 0.50)	$4.09 \cdot 10^{-5}$			
sbp	1.01	(1.00 , 1.03)	0.04	1.01	(1.00 , 1.03)	0.04
never smoked	1	ref	0.01	1	ref	0.01
former smk	2.26	(0.40 , 2.21)	0.89	0.93	(0.40 , 2.16)	0.86
1-9 cigar	2.09	(0.97 , 5.30)	0.06	2.22	(0.95 , 5.19)	0.07
10-19 cigar	0.97	(1.32 , 5.33)	0.00	2.65	(1.32 , 5.33)	00
20+ cigar	1.13	(1.33 , 7.26)	0.01	3.11	(1.33 , 7.28)	0.00
pipe-cigar	0.42	(0.20 , 11.67)	0.68	1.48	(0.19 , 11.35)	0.70
county OPP-L	1	ref	0.7			
county S&F	1.24	(0.68 , 2.29)	0.49			
county F	1.51	0.81 , 2.80)	0.19			
bmi	1.05	(0.98 , 1.11)	0.17			
smk strt	0.92	(0.90 , 0.99)	0.02			

Table 4.3: Uni-variable and multi-variable Fine-Gray model for death from other medical causes.

4.2.4 Fine-Gray model for death from alcohol abuse

Table 4.4 shows the uni-variable and the multi-variable Fine-Gray model estimations for death from alcohol abuse. The covariate sex is, obviously, not significant in both the uni-variable and multi-variable analysis whereas the uni-variable and multi-variable analysis of Cox model for death from alcohol abuse, table 2.6 , was estimated it as significant. The blood pressure (*sbp*) is the only covariate that is significant in both, the uni-variable and the multi-variable analysis, but it was not in the multi-variable analysis of Cox model and the difference in the hazard ratio (uni-variable analysis) estimated by the Cox model and Fine-Gray model is not too big. In contrast to the Cox model, the smoking grade is not significant neither in the uni-variable nor in the multi-variable analysis. Age when started smoking (*smk strt*) is only significant in the uni-variable analysis similarly to Cox model estimate and the difference in the hazard ratio between the model is, as will, not big.

Covariate	$exp(\hat{\beta})$	(95% CI)	$P-value$	$exp(\hat{\beta})$	(95% CI)	$P-value$
		Uni-variable			Multi-variables	
sex	1.10	(0.68 , 1.77)	0.69			
sbp	1.02	(1.00 , 1.03)	0.00	1.02	(1.00 , 1.03)	0.00
never smoked	1	ref	0.20			
former smk	1.13	(0.57 , 2.25)	0.72			
1-9 cigar	1.45	(0.63 , 3.33)	0.38			
10-19 cigar	1.43	(0.72 , 2.85)	0.30			
20+ cigar	2.01	(0.87 , 4.62)	0.10			
pipe-cigar	4.49	(1.52 , 13.27)	0.00			
county OPP-L	1	ref	0.6			
county S&F	0.78	(0.43 , 1.38)	0.39			
county F	0.75	(0.40 , 1.42)	0.38			
bmi	0.95	(0.88 , 1.03)	0.20			
smk strt	0.93	(0.88 , 0.98)	0.00			

Table 4.4: Uni-variable and multi-variable Fine-Gray model for alcohol abuse

The overall number of deaths in the Norwegian mortality in three counties data set, *the data which we worked on in this thesis*, is very small where only 586 of 4000 individuals were died. Therefore the difference between the two model in estimating the covariates effect is not big.

5 Conclusion

Competing risks are common in the analysis of survival data. Ignoring to compute correctly for competing events can result in counteractive consequences such as underestimation of the survival probability and imprecise estimate of the magnitude of effects of some given covariates on the occurrence of event of the interest. The overall number of deaths in the Norwegian mortality in three counties data set is very small. The number of deaths from cancer, cardiovascular, other medical causes and alcohol abuse is 217, 240, 68 and 61, respectively, which means 586 of 4000 individuals died, and 3414 individuals were truly censored. One can argue that it is, definitely, the reason behind there are no big difference between survival functions estimated by the Kaplan-Meier estimator and by the cumulative incidence function. However, the relative change between these two methods is, mostly, bigger for death from other medical causes and from alcohol abuse

than for death from cancer and from cardiovascular, and it is because the first two causes of death was the most unlikely cause of death while the later two causes were the most likely cause of death. This means death from cancer or from cardiovascular were high competing risk for death from other medical causes or from alcohol abuse, and this is the case where the difference between Kaplan-Meier estimates and the cumulative incidence function is bigger as the hazard of failure from the competing risk is higher.

The Cox model and Fine-Gray model were used to estimate the effect of the covariates on the cause specific hazard and the subdistribution hazard, respectively, but again the difference in the effect of those covariates on the cause specific hazard rates estimated by Cox model and the corresponding effect of the same covariates on the subdistribution estimated by Fine-Gray is not big, on contrary to many studies made and have been published by many statisticians and showed notable difference in the hazard rates which have been estimated by the two models for the same covariates. However, there are some differences between the two models in terms of the significance of some given covariates as some of these covariates were not statistically significant in the Cox model, but they were statistically significant in the Fine-Gray model.

We refer to an important post by Paul Allison [31] who recommended not to use Fine-Gray's subdistribution method for causal analysis of competing risks after he found out that for any competing risks H_1 and H_2, covariate that increases the cause specific hazard of event H_1will appear to decrease the subdistribution hazard for event H_2.

References

[1] Goel, Manish Kumar, Pardeep Khanna, and Jugal Kishore. "Understanding survival analysis: Kaplan-Meier estimate." International journal of Ayurveda research 1.4 (2010): 274.

[2] Kaplan, Edward L., and Paul Meier. "Nonparametric estimation from incomplete observations." Journal of the American statistical association 53.282 (1958): 457-481.

[3] Cox, David Roxbee, and David Oakes. "Analysis of survival data (Vol. 21)." Chapman & Hall/CRC 1 (1984): 984.

[4] Rosner, B. Fundamentals of Biostatistics. 4th ed. Belmont, California: Duxbury Press(1995).

[5] Broström, Göran. Event history analysis with R. CRC Press, 2012.

[6] Altman, Douglas G. "Analysis of Survival times.In:Practical statistics for Medical research." London (UK): Chapman and Hall (1992): 365–93.

[7] Stevenson, Mark, and I. V. A. B. S. EpiCentre. "An introduction to survival analysis." EpiCentre, IVABS, Massey University (2009).

[8] "Cox 'Proportional Hazards' Regression." StatsDirect. Retrieved from https://statsdirect.com/help/Default.htm#survival_analysis/cox_regression.htm. Web. 1 Oct. 2018.

[9] Kleinbaum, David G., and Mitchel Klein. Survival analysis. Vol. 3. New York: Springer, 2010.

[10] Fabsic, Peter, E. Vakhrushev, and Kevin Zemmer. "The Cox proportional hazard model and its characteristics." (2011).

[11] Saurav De. Partial Likelihood and Cox Proportional Hazard Model. e-PG Pathshala. Retrieved from https://epgp.inflibnet.ac.in/epgpdata/uploads/epgp_content/statistics/05._statistical_inference_iii_____/23._partial_likelihood_and_cox_proportional_hazard_model/et/9587_et_module_17.pdf .(accessed 10 Nov. 2018).

[12] Dobson, Annette J and Barnett, Adrian G An introduction to generalized linear models (3rd ed). CRC Press, Boca Raton, 2008.

[13] ”Cox Regression.” NCSS. Retrieved from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Cox_Regression.pdf. (accessed 15 Nov. 2018).

[14] Efird JT. Sinusoidal cox regression-a rare cancer example. Cancer Inform. 2010;9:265-79. Published 2010 Nov 28. doi:10.4137/CIN.S6202

[15] Cox, David R. "Regression models and life-tables." Journal of the Royal Statistical Society: Series B (Methodological) 34.2 (1972): 187-202.

[16] Cox, David R. "Partial likelihood." Biometrika 62.2 (1975): 269-276.

[17] Aalen, Odd, Ornulf Borgan, and Hakon Gjessing. Survival and event history analysis: a process point of view. Springer Science & Business Media, 2008.

[18] Data,"Norwegian population mortality." Web. http://folk.uio.no/borgan/abg-2008/data/causes_death.tx

[19] Scrucca, L., A. Santucci, and F. Aversa. "Regression modeling of competing risk using R: an in depth guide for clinicians." Bone marrow transplantation 45.9 (2010): 1388.

[20] Ebeling, Charles E. An Introdution to Reliability and Maintainability Engineering. Waveland, 2010.

[21] Der, Geoff, and Brian Everitt. Basic statistics using sas enterprise guide: A primer. SAS Institute, 2007.

[22] Putter, Hein, Marta Fiocco, and Ronald B. Geskus. "Tutorial in biostatistics: competing risks and multi-state models." Statistics in medicine 26.11 (2007): 2389-2430.

[23] Hess, Kenneth R. "Graphical methods for assessing violations of the proportional hazards assumption in Cox regression." Statistics in medicine 14.15 (1995): 1707-1723.

[24] How can I perform the likelihood ratio and wald test in STATA. UCLA: Statistical Consulting Group. from https://stats.idre.ucla.edu/stata/faq/how-can-i-perform-the-likelihood-ratio-wald-and-lagrange-multiplier-score-test-in-stata/ (accessed 22 Nov. 2018).

[25] Meloun, Milan, and Jiri Militky. Statistical data analysis: A practical guide. Woodhead Publishing, Limited, 2011.

[26] Zhang, Zhongheng. "Survival analysis in the presence of competing risks." Annals of translational medicine 5.3 (2017).

[27] Ranganathan P, Pramesh CS. Censoring in survival analysis: Potential for bias. Perspect Clin Res. 2012;3(1):40. doi:10.4103/2229-3485.92307

[28] Gooley, Ted A., et al. "Estimation of failure probabilities in the presence of competing risks: new representations of old estimators." Statistics in medicine 18.6 (1999): 695-706.

[29] Gillam, Marianne H., et al. "Different competing risks models applied to data from the Australian Orthopaedic Association National Joint Replacement Registry." Acta orthopaedica 82.5 (2011): 513-520.

[30] Fine, Jason P., and Robert J. Gray. "A proportional hazards model for the subdistribution of a competing risk." Journal of the American statistical association 94.446 (1999): 496-509.

[31] Paul Allison. For Causal Analysis of Competing Risks, Don't Use Fine & Gray's Subdistribution Method. Statistical Horizons. Retrieved from https://statisticalhorizons.com/for-causal-analysis-of-competing-risks . (accessed 5 June. 2019).