Multi-state models for late effects in childhood cancer survivors

Kadriye Kaplan

Kongens Lyngby 2012
IMM-M.Sc.

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Summary

This thesis deals with statistical methods and their applications for describing diabetes-related morbidity and cause-specific mortality in the Nordic childhood cancer survivors. The purpose in the study is to analyze these outcomes in the survivors compared to the general population both separately and jointly by using multi-state models.

The data is provided by a Nordic childhood cancer study, the Adult Life after Childhood Cancer in Scandinavia (ALiCCS). It encompasses information about 24936 childhood cancer survivors who are diagnosed during 1943 to 2008 in Denmark and Sweden and are matched with 124663 controls on gender and date of birth.

In order to study statistical methods for analyzing the morbidity and mortality outcomes, the performing process is divided into two parts. In the first part of the analysis the standard two-state models are considered separately for each outcome whereas in the second part of the analysis more advanced multi-state models are constructed for describing the morbidity and mortality outcomes jointly. Both analyses are based on the ordinary as well as extended Cox models. The obtained results from the analyses are summarized and discussed.

Both the standard two-state models and multi-state models have shown some similar results in the univariate analyses. These models have revealed that the childhood cancer survivors are associated with higher risk of experiencing both morbidity and mortality outcome when compared to the general population. In addition to this, multi-state analysis has shown that the childhood cancer survivors were more likely to die if they have developed diabetes than the other way

around. Furthermore, it is found that the occurrence of diabetes has increased the risk of death in the study participants.

Resumé

Dette kandidatspeciale omhandler de statistiske metoder og deres anvendelser i analysen af morbiditetsudfaldet diabetes og mortalitetsudfald hos nordiske børnecancer-overlevere. Formålet med dette studie er at undersøge udfaldene hos overlevere sammenlignet med den generelle population både separat ved hjælp af "two-state" modeller og fælles ved hjælp af "multi-state" modeller.

Data er leveret af et nordisk børnecancer studie, the Adult Life after Childhood Cancer in Scandinavia (ALiCCS). Dette omfatter information om 24936 overlevere af børnecancer, som er diagnosisteret i løbet af 1943-2008 og er blevet matchet med 124663 kontroller på køn og alder.

For at studere de statistiske metoder i analysen af morbiditets- og mortalitetsudfald er arbejdsprocessen opdelt i to dele. I den første del af analysen er de almindelige "two-state" modeller opstillet separat for hvert udfald. Den anden del af analysen omfatter mere avancerede multi-state modeller som er opbygget for at beskrive morbiditets- og mortalitetsudfaldet fælles. Begge analyser er baseret på ordinære såvel som udvidede Cox modeller.

Både "two-state" og "multi-state" modeller har vist nogenlunde de samme resultater i de univariate analyser. Disse modeller har vist, at overlevere af børnecancer er associeret med en højere risiko for morbiditet og mortalitet sammenlignet med den generelle population. Derudover viste "multi-state" modellerne, at overlevere af børnecancer var mere tilbøjelige til at dø, hvis de havde udviklet diabetes. Endvidere har det vist sig at forekomsten af diabetes har forøget dødelighedsrisikoen hos studie populationen.

Preface

This thesis was prepared at Department of Informatics and Mathematical Modelling at the Technical University of Denmark in partial fulfillment of the requirements for acquiring a Master of Science degree (M.Sc.) in engineering.

The thesis deals with statistical methods that are applied to a survival data provided by the Danish Cancer Society. The main focus is on analyzing the diabetes-related morbidity and cause-specific mortality in the Nordic childhood cancer survivors both separately and jointly by means of multi-state models. The analyses are based on the ordinary as well as extended Cox regression models.

I would like to thank my supervisor Klaus Kaae Andersen for giving the opportunity to work on this thesis, for his guidance and support throughout the project. I also want to thank Per Bruun Brockhoff for his guidance and advice during the course of the thesis. Furthermore, I would like to thank the Danish Cancer Society for providing data and for contributing a pleasant work environment. All these made it possible to carry out this project. Last but not least, I would like to thank my family especially Fatma and Rifat Kaplan for their encouragement, patient, love and care. I could not have done this without you.

Kongens Lyngby, March 2012

Kadriye Kaplan

Acronym Table

Acronym	Term
ALiCCS	the Adult Life after Childhood Cancer in Scandinavia
CCS	Childhood cancer survivors
CCSS	Childhood cancer survivor study
CNS	Central nervous system
ALL	Acute lymphoblastic leukemia
CPH	Cox proportional hazards
HR	Hazard ratio

Contents

Chapter 1

Introduction

The development of effective treatments for childhood cancer has resulted in almost 80% affected children and adolescents to become long-term survivors [39]. Consequently, the improvement in the curative therapies has been accompanied by a variety of long-term sequelae, such as impairment of growth and development, reproductive difficulties, chronic late morbidity, second cancers, increased mortality and psychosocial and familial problems [47].

Most side effects of cancer treatments occur during or just after treatment and disappear a short time later, whereas long-term late effects do not become clinically apparent until decades after completion of cancer treatment. Research has shown that a high burden of morbidity among childhood cancer survivors is contributed by late effects, with about two-thirds developing at least one chronic health condition and at least one-third experiencing severe or life-threatening complications during adulthood [70][19][39][26].

Most late effects are caused by cancer treatments such as chemotherapy, radiation therapy and bone marrow/stem cell transplantation. The risk of developing late effects depends on several factors as; the type and the location of cancer, the area of the body treated, the type and dose of treatment, the child's age at diagnosis and treatment, the child's gender, genetics and family history, and whether other health problems existed before the cancer diagnosis [63]. The underlying principle of cancer treatments is to destroy fast-growing cells, such

as cancer cells. Drugs used in the chemotherapy interferes the process of cell division by damaging proteins involved or DNA itself, causing the cancer cells to die. In a child normal healthy cells in the body grow and divide quickly, too. These cells can be attacked by drugs used in the therapy and can be damaged drastically. Radiation therapy, on the other hand, uses high-energy rays to kill fast-growing cells. As with chemotherapy, radiation therapy can damage normal heathy cells along with the cancer cells and cause late effects [53].

Examples of late effects of childhood cancer treatment include cardiomyopathy, diabetes, hemorrhage, hypertension as the most important. The main late effect focused in this thesis is diabetes. It is reported that childhood cancer survivors of Acute lymphoblastic leukemia (ALL) have an increased prevalence of obesity and insulin resistance and therefore may be at risk for experiencing diabetes [71][27]. The disease is considered as being a metabolic syndrome, which is highly associated with a increased risk of cardiovascular events and mortality. There exist little information about long-term cardiovascular outcomes in the existing literature, however, an analysis of data from a Childhood Cancer Survivor Study has shown that the standardized mortality ratio for cardiac-related deaths was 8.2 [95% CI, 6.4-10.4] among long-term survivors of childhood cancer [31][2].

Although the need for long-term follow-up of childhood cancer survivors are well recognized, the study of long-term morbidity is still relatively new. This phenomenon is stated by a childhood cancer survivor as: "We are kind of an unknown element of society." (Christy, 2009)[69]. It is of importance to advance knowledge about the morbidity that follows the treatment of childhood cancer and its contribution to early mortality. This initiative is taken by a Nordic childhood cancer study, the Adult Life after Childhood Cancer in Scandinavia (ALiCCS) which utilizes resources including investigation of data from population-based registries and large-scale cohort studies. The main purpose is to compare the morbidity-specific incidence and cause specific mortality of the childhood cancer cohort to age- and gender-matched general population cohort. Due to its unique and high quality data ALiCCS will provide a better understanding of the occurrence and risks for late cancer treatment-related effects.

Survival analysis is the most appropriated statistical analysis technique used for describing time to event data. The event is typically death, but the term is also used for other events, like the occurrence of a disease. Traditionally, it is assumed that only a single event occurs for each subject, however, multiple events relax that assumption. In the illness-death process, often more than one type of event is involved.

The standard approach to analyze late effects is to consider morbidity and mortality outcomes separately by means of survival analysis and the Cox Propor-

tional Hazards (CPH) model. One concern, however, is that the intermediate event, occurrence of diabetes, may significantly change the risk of the event of interest (death) to occur. Often, one is also interested in what happens after occurrence of the intermediate event. The main objective of this thesis is to study statistical methods for describing the morbidity outcome, diabetes and cause specific mortality jointly rather than separately, and investigate if this yields different results and new insight when compared to the traditional analysis of late effects. Based on the extended Cox model, the event history data analysis is performed using multi-state models. The significance of the models are tested in the statistical computing program R version 2.13.1 with the threshold for statistical significance $\alpha = 5\%$. An estimate is said to be statistically significant if the p -value of the estimate is less than 5%.

The thesis consists of seven chapters. The present chapter (Chapter 1) gives an introduction to the issue. Chapter 2 describes the content of the data provided by ALiCCS and introduces exclusion criteria that is applied to the data. Chapter 3 presents a descriptive analysis of subject characteristics of the cohort and the crude estimates of the mortality and morbidity rates. Chapter 4 gives an overview of the statistical theory applied in the analyses. An introduction to the standard two-state survival analysis based on the Cox Proportional Hazards model is presented followed by an overview of more complicated multi-state analysis. The results from the two-state analysis obtained by applying presented statistical methods are gathered in Chapter 5 and the results from the multi-state analysis are given in Chapter 6. A conclusion and a discussion of the obtained results are presented in Chapter 7. All relevant figures and codes are appended in Appendix B and C, respectively.

Chapter 2

Study cohort

In this chapter the content of the data provided by ALiCCS is described. An introduction to the data followed by exclusion criteria that is applied to the data is introduced and illustrated by figures. The content of the final data is summarized in a table.

The study is based on the data obtained from a Nordic childhood cancer study, the Adult Life after Childhood Cancer in Scandinavia (ALiCCS) which was established at the end of 2009. The participating parties in the research project consist of Danish Cancer Society Research Center, Aarhus Universitethospital, Swedish Cancer Registry, Skaane Univesity Hospital, Cancer Registry of Norway, Finish Cancer Registry, Turku University Hospital, Icelandic Cancer Registry and Lund University [48].

For the research project, a complete, population-based series of children and adolescents in the Nordic countries (Denmark and Sweden) in whom cancer was diagnosed during the period 1943 to 2008 is established. Some clinical case-control studies of childhood cancer survivors (CCS), nested in the cohort is then set up in order to investigate late effects and its treatments in CCS. The cohort study is conducted to compare the morbidity-specific incidence and cause-specific mortality of the childhood cancer cohort to the general population cohort. CCS are identified and followed-up individually in national register.

Follow-up of CCS are achieved through the civil registration systems and information on subsequent disease is obtained by using large-scale record linkage techniques with national outcome registers. Registration of cancer diagnoses is obtained from Cancer Registries, registration of vital status is obtained from Civil Registries and registration of diagnosis types is obtained from National Board of Health [47].

The data set used in this thesis consists of two sub data sets; one collected in Sweden and the other in Denmark. It encompasses information about 149599 study participants from both countries. The cohort of childhood cancer cases is established with a combined comparison cohort consisting of 124663 controls taken at random from the general population of Denmark and Sweden. The study comprises 24936 childhood cancer survivors, of whom 9859 are diagnosed in Denmark during the period 1943 to 2008 and 15077 are diagnosed in Sweden during the period 1958 to 2008.

In order to ensure that exposed and unexposed to cancer are similar in variables that might confound a relationship that is being studied, each child exposed to cancer is matched with 5 unexposed participants from the study cohort. Unexposed were matched on characteristics as gender and date of birth (within 1 year). By means of matching it is ensured that the difference between exposed and unexposed are not a result of differences in the matching variables. Thus, comparison of exposed and unexposed can be done knowing that the effect of these variables are automatically adjusted for.

2.1 Data description

The type of late effects initially selected for this cohort study is cardiovascular and pulmonary diseases such as diabetes, cardiomyopathy, infarct, hypertension and hemorrhage. However, due to time limitation the analysis is restricted to investigate only the effect of occurrence of diabetes. The data set holds information about a specific id number, gender, date of birth, date of diagnosis, type of the diagnosis, status, last date of follow-up, date of different late effects and country. The variable `status` describes the different person civil registration status code given in Table 2.1 [65].

In order to perform a survival analysis a censoring variable must be defined by an indicator; 1 for the subjects that experience the event of interest and 0 for the subjects that are *censored*. A study participant is said to be censored if the person is lost to follow-up during the observation period or the person has still not experienced any event at the end of the study. The event of interest in

Denmark	Sweden	Description
01	1	registered with residence in Danish/Swedish population register.
60	-	changed civil registration number by amendment of date of birth and gender.
70	-	disappeared
80	8	emigrated
90	9	dead (dead or dead as emigrated or disappeared)

Table 2.1: *Person civil registration codes for Denmark and Sweden.*

this study can be specified as experiencing diabetes and being dead with and without experiencing diabetes, these are competing events. A status variable for the event of interest death is formed by labeling the status codes 90 and 9 in Table 2.1 as 1 and others as 0 for censoring. For the other events of interest, a status variable is similarly generated and added in the data set.

Different time scales such as duration, calendar time or age can be used for survival analysis. However, age is chosen to be the time scale in this analysis. Thus the patients are started observation at whatever age they are on the date of diagnosis, i.e. enter the risk set. For each event of interest age at entry and age at exit of the study is figured out by using the given dates in the data set. For some exposed the diagnosis is registered before their date of birth and thus negative age-entry values are recorded. In order to solve this problem, the negative values are set to zeros. Since the unexposed cohort consists of participants that are not diagnosed cancer at inclusion, their age at diagnosis do not exist. In order to make an appropriate analysis, the age of the unexposed at diagnosis is set to the average age of the exposed at diagnosis.

2.1.1 Exclusion Criteria

In this study some exclusion criteria are applied to the data. These criteria provide requirements as to who may or may not participate in the study. It is required that the event of interest occurs during the follow-up time or possibly later and not before study start. Furthermore, the study is considered to be progressive meaning that all events occur in consecutive order. The observation period starts with diagnosing cancer and ends by occurrence of event of interest (death), censoring or end of follow-up time. An illustration of some possible event observations for the cohort from Denmark is depicted in Figure 2.1.

Each line in the figure runs from the entry to follow-up until either death or

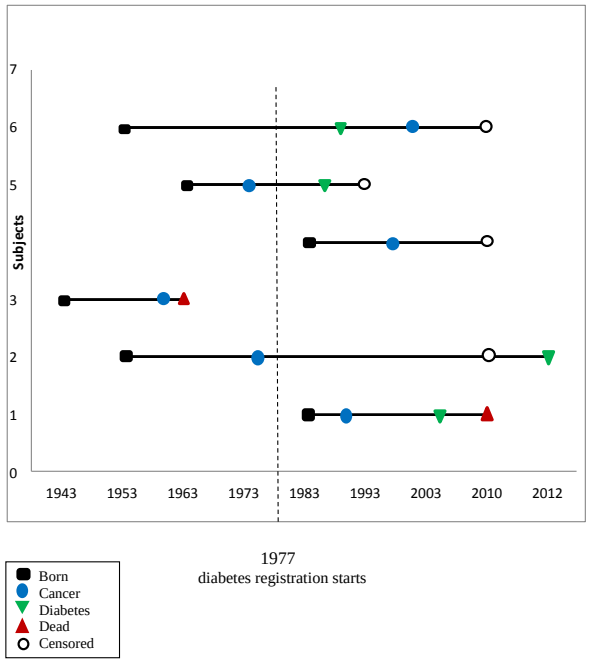

Figure 2.1: *An illustration of follow-up experience of childhood cancer survivors. The registration of diabetes in Denmark starts in 1977 which is highlighted with a dashed line.*

censoring. It may be clarified here that the registration of diabetes cases in Denmark is started in January 1, 1977 and in Sweden started in January 1, 1984. This means that even if a subject has had diabetes before these dates, the event, diabetes is first registered after it is really occurred. Consequently, age at diagnosis for subjects diagnosed before start of diabetes registration cannot be chosen as age at entry into the study, since the actual study starts in 1977 for subjects from Denmark and in 1984 for subjects from Sweden. As it is seen in Figure 2.1, subjects that end the study before start of diabetes registration, cannot be included in the study, here subject 3. If a participant experiences diabetes before diagnosis registration which is the case for subject 6 in the figure, the person is excluded from the study as well.

Figure 2.2 demonstrates how the study-entry of subjects that are diagnosed before start of diabetes registration in Denmark, is handled. Age of these subjects are figured out for 1977 and considered as their age at entry into the study. For subjects diagnosed after 1977, their age at diagnosis is considered to be their age

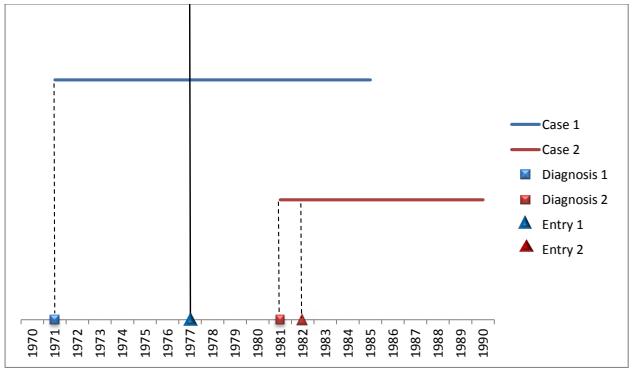

Figure 2.2: *A demonstration of how study-entry of subjects collected from Denmark is handled.*

at entry. This procedure is also applied to subjects from Sweden with diabetes registration started in 1984.

As cancer is detected and treated, some of the side treatment effects can occur and disappear within 1 year after diagnosis [75]. These type of effects should be excluded from the study so that late effects are the only risk factors left in the analysis. For this purpose an additional year is added to the age at entry. This is also illustrated in Figure 2.2.

The application of the exclusion criteria on the data set is shown in a flowchart in Figure 2.3. The criteria given in figure outline who may be considered for the study and who is excluded from consideration. As mentioned before, the process is considered to be progressive so that the occurrence of diabetes can only happen after cancer diagnosis. It appears in the figure that 40 childhood cancer survivors have had diabetes before registration of cancer and therefore they are excluded from the study. It is also seen that 5109 exposed that is diagnosed before diabetes registration and has an exit date before 1977 or 1984 is excluded from the study. The final cohort consists of 142426 participants of whom 19776 are exposed and 122650 are unexposed.

After these exclusions the final data set is constructed including only the relevant variables for the analysis. An overview of the variables divided in 3 categories is visualized in Figure 2.4. Gender, age at diagnosis, calendar year at diagnosis and country are listed as the demographical variables. The category denoted cohort includes groups consisting of exposed and unexposed and the final category, event of interest includes the events and their corresponding variables status, age at entry and age at exit.

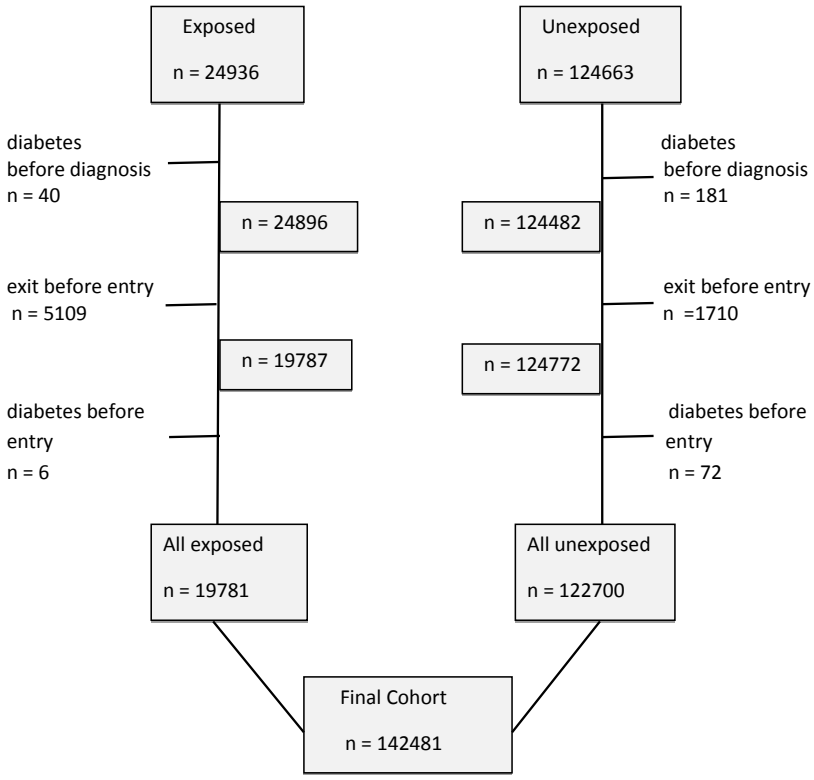

Figure 2.3: *A flowchart showing the order and the number of exclusions applied on the data set.*

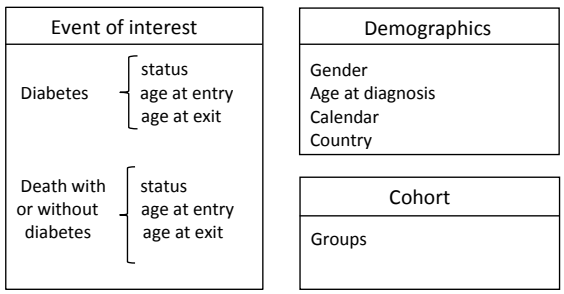

Figure 2.4: *An overview of all the relevant variables included in the final data set.*

Chapter 3

Descriptive Data Analysis

This chapter presents a descriptive analysis of subject characteristics of the ALLiCS childhood cancer survivor cohort. The results of the analysis are illustrated by tables and figures. Overall crude estimates of rates are given as a measure of risk of mortality and morbidity outcome in the follow-up period. Finally, the crude estimates of cumulative incidence of outcomes in the childhood cancer survivors are presented and compared to the general population.

The final data set used in the analysis consists of 142481 eligible subjects, whom have been followed during the years January 1943 to January 2010. Table 3.1 provides subject characteristics of the 19781 exposed and 122700 unexposed to childhood cancer included in the cohort study. As a result of the matching, exposed and unexposed are similar with regard to gender and age at diagnosis.

As it appears in the table, 39% of exposed is collected from Denmark, whereas 61% is collected from Sweden. There is almost a similar distribution of unexposed regarding to the countries. Since the data from Sweden consists of a large sample size, the cohort from Sweden represents a major part of the total. The distribution of the gender is fairly equal in the cohort, but a slight overweight of boys is distinct in both exposed and unexposed; 54% of exposed and 55% of unexposed are boys. As it is mentioned in the previous chapter, age of unexposed at diagnosis is set to the average age of exposed at diagnosis. The average

age of exposed at diagnosis is approximately 10 for both countries. Therefore, the distribution of age at diagnosis is well divided among exposed, whereas the contribution of age of unexposed at diagnosis is only given for one interval in the total. Nearly 30% of exposed has an age in the interval [0-5] and 32% of them has an age in the interval [15-20]. The distribution of exposed does not seem to differ significantly regarding to age groups [5-10] and [10-15].

	Exposed	%	Unexposed	%	Total (%)
Number	19781	13.88	122700	86.12	142481
Country					
Denmark	7699	38.92	48000	39.12	55699 (39.09)
Sweden	12082	61.08	74700	60.88	86782 (60.91)
Gender					
boy	10661	53.90	66998	54.60	77659 (54.51)
girl	9120	46.11	55702	45.40	64822 (54.60)
Age					
$0-5$	6006	30.36	-	-	6006 (4.22)
$5-10$	3532	17.86	-	-	3532 (2.48)
$10-15$	3934	19.89	-	-	126634 (88.88)*
$15-20$	6309	31.89	-	-	6309 (4.43)
Calendar					
$1943-1960$	823	4.16	4252	3.47	5075 (3.56)
$1960-1974$	2808	14.20	25930	21.13	28738 (20.17)
$1974-2010$	16150	81.64	92518	75.40	108668 (76.27)
Mean (sd)	Exposed	sd	Unexposed	sd	Total (sd)
Age	10.22	(6.41)	10.12	(4.06)	10.14 (4.43)
Calendar	1988.18	(13.85)	1986.36	(13.83)	1986.61 (13.85)

* Note that unexposed are not diagnosed, but their age at diagnosis is set to the average age of exposed at diagnosis which is approx. 10.

Table 3.1: *Subject characteristics of the ALLiCS childhood cancer survivor cohort. The number of exposed and unexposed for a given characteristic is listed together with the percentage contribution.*

By looking at the distribution of the cohort with regard to calendar year, it is observed that the major part (82%) of exposed is diagnosed in [1974-2010]. The variables; age at diagnosis and calendar year are presented both as categorical and numerical in the table. The reason for this is that the variables are considered being qualitative as well as quantitative in the analysis. It is due to statistical considerations that will be presented in the following chapter.

Using the values given in the table, a χ^2-test of statistical significance for bi-

variate tabular analysis is performed. The p-value for the hypothesis of no difference between exposed and unexposed according to calendar year at diagnosis is > 0.001 indicating that there is a strong evidence of a difference between exposed and unexposed. With regard to other explanatory variables; country and gender, the difference between exposed and unexposed does not seem to be significantly high.

A plot of number of exposed and unexposed in Denmark and Sweden is displayed in Figure 3.1. As it appears here, the number of exposed in both countries is increasing during follow-up time. A significant difference between the distribution of exposed in Denmark and Sweden is ascertainable. As a result of the matching, the number of unexposed is approx. 5 times the number of exposed.

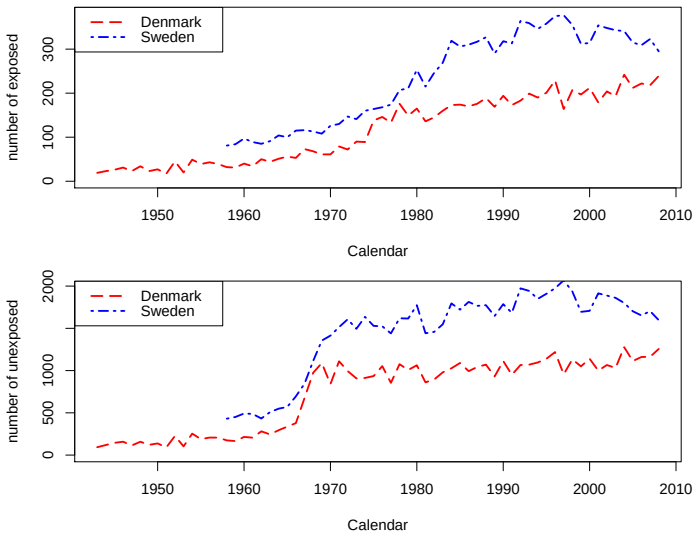

Figure 3.1: *The number of exposed (top) and unexposed (bottom) in Denmark and Sweden with regard to calendar year.*

Figure 3.2 shows the number of diabetes events among exposed and unexposed in both countries. The number of events in both childhood cancer and reference population seems to be constant until 1995 and after that a positive shift in the number of events is observable. This shift can be due to the delay in diabetes registration especially in Sweden and also due to increase in incidence of diabetes in the recent years. It appears in the figure that there is an overweight of diabetes events in reference population in both countries and occurrence of diabetes in Sweden is fairly high due to large sample size.

In order to visualize the follow-up trajectories of the cohort only for diabetes

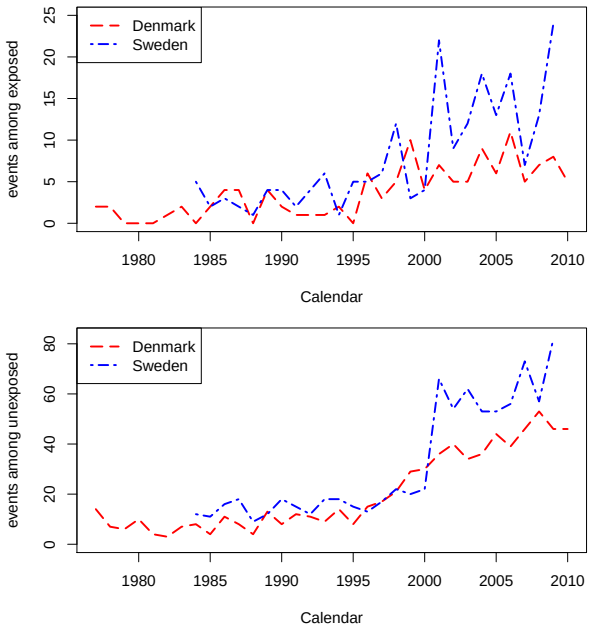

Figure 3.2: *The number of diabetes events among exposed (top) and unexposed (bottom) in Denmark and Sweden with regard to calendar year.*

events a Lexis diagram is set up. Figure 3.3 shows a small segment of the follow-up trajectories taken randomly from the data set. Each line in the Lexis diagram represents the follow-up of a single participant from entry to exit on two time scales; age and calendar time which are given in the same units (years). Exit status is denoted by a filled circle for the participants who experience diabetes and by an unfilled circle for the participants who are disease-free or censored.

In the previous chapter it is implied that diagnosis type for childhood cancer survivors will not be included in the analysis in order to reduce the number of variables that will be estimated. However, it of interest to see which type of diagnosis the exposed were associated with when they have developed diabetes. A table of number of exposed who has experienced diabetes and their corresponding diagnosis type is given in Table 3.2. It is obvious that most of the exposed were diagnosed with Central Nervous System (CNS) tumors and Leukemia when they have developed diabetes.

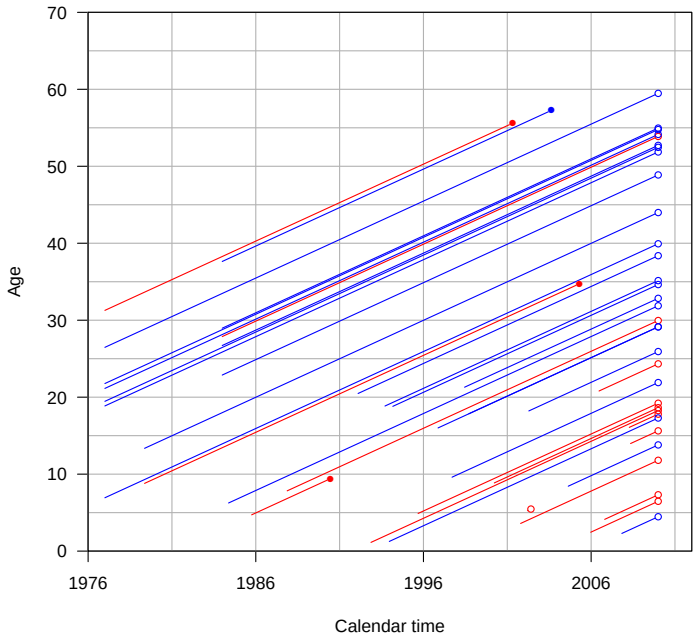

Figure 3.3: *Lexis diagram showing a small segment of the trajectories of exposed (red) and unexposed (blue) in the study.*

3.1 Crude estimates

Overall crude estimates of rates are recorded for measuring the risk of mortality and morbidity outcome in the follow-up period. It should be mentioned that when estimating the mortality rates, the participants that experience diabetes is censored and the time in which they are under risk is included in the risk set. Similarly, the estimation of diabetes rates is obtained by censoring those that do not experience diabetes and the estimation of mortality rates after experiencing diabetes is obtained by only considering subjects died after developing diabetes.

Incidence rates which are a measure of the mortality and morbidity occurrence per unit time is given in Table 3.3. It is seen that the exposed contributes 289552.48 person-years in the estimation of mortality rates without experiencing diabetes during the follow-up period. This is less than the total possible

Diagnosis type	experience diabetes
Carcinomas	40
CNS	94
Germ cell and other	26
Hepatic tumors	1
Leukemia	51
Lymphomas	40
Malignant bone tumors	10
Other and unspecified	3
Renal tumors	27
Retinoblastoma	9
Soft tissue sarcomas	18
Symp. NST	10

Table 3.2: *Number of exposed with diabetes and their corresponding diagnosis type.*

person-time since people who are censored before the end of the follow-up period stopped contributing person-time at the time of the event. The number of events is 3503; thus, the incidence rate per 1000 person years for the exposed is $(3503/289552.48) \cdot 1000 = 12.10$ meaning that 12 mortality events would be expected for 1000 persons observed for 1 year. Among 19781 survivors, 1.7% have diabetes and among 122700 unexposed, 1.2% have diabetes. The risk of experiencing diabetes among exposed is 1.14 per 1000 person-years which is quite high compared to incidence rate of unexposed (IR: 0.68). The exposed are $1.14/0.68 = 1.68$ times more likely to experience diabetes than unexposed. The mortality rate after experiencing diabetes among exposed is 26.03 and among unexposed is 12.29 per 1000 person-years indicating that the risk of dying with diabetes as a childhood cancer is 2.1 times the risk of dying with diabetes as a reference population.

In figure 3.4, crude estimates of cumulative incidence of mortality and morbidity in childhood cancer survivors are presented and compared to the general population.The first plot (top-left) in the figure shows the expected proportion of individuals that die without experiencing diabetes over the course of time. As seen in the plot, the expected mortality among exposed is increasing very steep in the first 5 years, but as time progresses, the slope of the increase remains constant. The figure clearly demonstrates that the incidence of mortality in childhood cancer survivors are higher compared to the general population.

The second plot (top-right) shows the cumulative incidence of diabetes in the study cohort. As it appears here, the expected proportion of individuals that develop diabetes is remarkable higher among exposed compared to the general

Mortality rates without diabetes				
Groups	p-years	n	event	rate per 1000 p-years
exposed	289552.48	19781	3503	12.10
unexposed	2245608.09	122700	2510	1.12
Diabetes rates				
Groups	p-years	n	event	rate per 1000 p-years
exposed	289552.48	19781	329	1.14
unexposed	2245608.09	122700	1517	0.68
Mortality rates with diabetes				
Groups	p-years	n	event	rate per 1000 p-years
exposed	2459.01	322	64	26.03
unexposed	12205.68	1465	150	12.29

Table 3.3: *Incidence rates of mortality and morbidity outcomes per 1000 person years.*

population. The cumulative incidence of diabetes in the exposed is 2.5% at 25 years after diagnosis whereas it is 1.2% in the unexposed. The cumulative incidence of mortality after experiencing diabetes in the cohort is illustrated in the last plot. By considering the magnitude of the slope of the cumulative incidence curves for both groups, one may note that both exposed and unexposed are associated with an increasing cumulative incidence of mortality after experiencing diabetes. Of particular interest is the comparison of the first and last plot suggesting that the occurrence of diabetes is relatively increased the cumulative incidence of mortality for both groups. In the first 5 year after diagnosis, the cumulative incidence of mortality before and after experiencing diabetes among exposed seems to be the same, but after 5 years the cumulative incidence of mortality after experiencing diabetes is much more higher than the cumulative incidence of mortality shown in the first plot. This difference is more visible for the cumulative incidence of mortality for the general population. It might be mentioned that the incidence rates shown in the figure are left truncated at 1 year after diagnosis, reflecting the eligibility entry criteria for the cohort.

Cumulative incidence function plotted in Figure 3.4 gives test statistics and p-values for comparing the sub-distribution for mortality and morbidity across groups. Test statistic for all three event of interest returns a p-value < 0.001 which verifies that the hypothesis of no difference between groups in each sub-distribution are rejected [22].

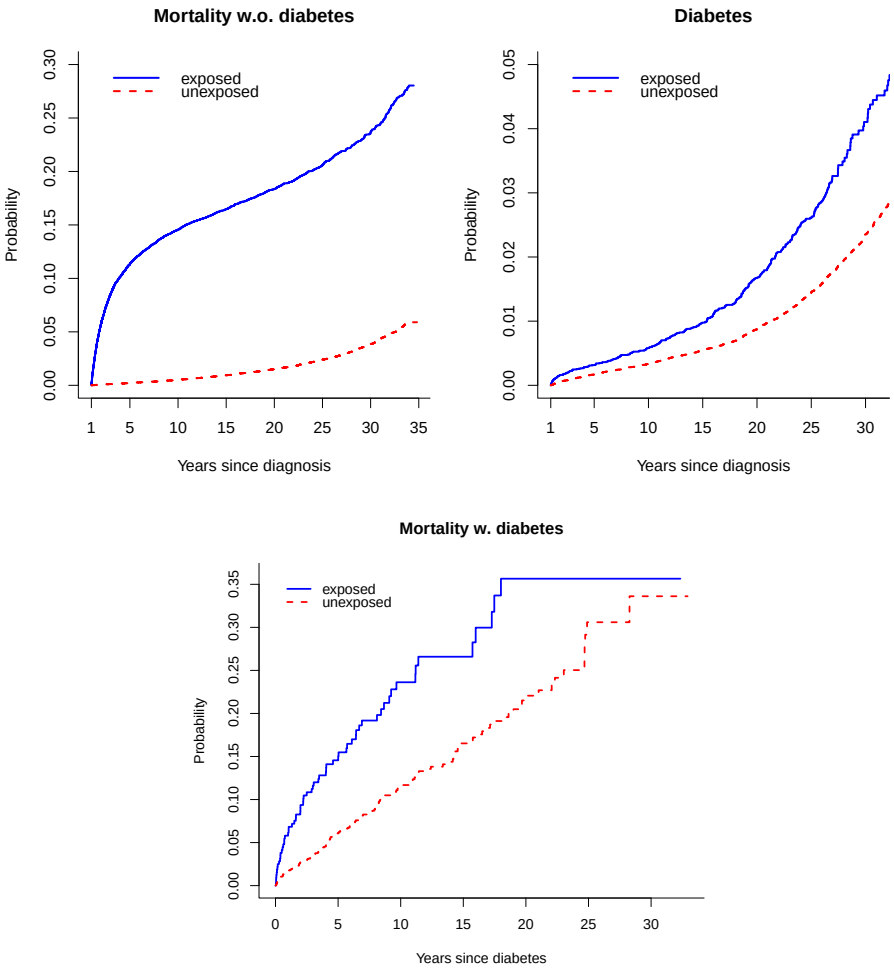

Figure 3.4: *Cumulative incidence of mortality and morbidity in survivors of childhood cancer compared to the general population.*

Chapter 4

Methodology

The cohort formed in ALLICS allows for time to event analysis, and therefore survival analysis is the most appreciated statistical analysis technique for this. In this chapter a short theoretical introduction to survival analysis is given. First, standard two-state analysis based on the Cox proportional Hazards model followed by assessment of model assumptions and extension of the Cox model is presented. Then, more advanced multi-state models that form an extension to the standard survival analysis is introduced by extended Cox models, transition probabilities and software packages that is developed for multi-state models. Finally, model assumptions made for this study is summarized.

4.1 Survival analysis

By considering the analysis on time to event data with possibly censoring it is assumed that each individual i has an event time t_i and a censoring time c_i. Then the observed time is given as $x_i = \min(t_i, c_i)$ along with $\delta_i = I(t_i \leq c_i)$ meaning that whether an event is observed ($\delta_i = 1$) or not ($\delta_i = 0$). In the data set the event times and censoring times are considered as being a random sample $(X_1, C_1), ..., (X_n, C_n)$ drawn from a survival distribution $X_i \sim S$, with $S(t) = Prob(T > t)$. For T denoting the survival time, the survival function, $S(t)$ is defined as the probability of survival to time t after entry time [23].

For the model analysis it is assumed that the event time distribution and the censoring distribution are independent conditioned on the covariates included in the model. The distribution of the lifetime T can be stated by means of the hazard function defined at any time point t as the probability of failure (experiencing the event) within a short time interval, given that the individual was alive at the beginning of the time interval. That is,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{4.1}$$

The definition of the hazard function implies that

$$\lambda(t) = \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t} = -\frac{\mathrm{d} \log S(t)}{\mathrm{d}\, t} \tag{4.2}$$

The cumulative hazard function is defined by

$$\Lambda(t) = \int_0^t \lambda(s)\, \mathrm{d}\, s \quad \text{where } 0 \leq s \leq t \tag{4.3}$$

and the estimation of hazard rate is based on the cumulative hazard instead of hazard function itself, since it is easier to estimate cumulative distribution function than probability density function [1]. The *Nelson-Aalen* estimator is most commonly used non-parametric estimator of the cumulative hazard function,

$$\hat{\Lambda}\,(t) = \sum_{t_i \leq t} \frac{N\,(t_i)}{Y\,(t_i)} \tag{4.4}$$

where $N(t_i)$ is the number of events at time t for each subject i under observation and $Y(t_i)$ is the number of subjects under observation and at risk at time t_i [45].

The survival function can be represented in terms of the cumulative hazard as,

$$S(t) = \exp(-\Lambda(t)) \tag{4.5}$$

In fact, if $S(t)$ is known, the corresponding hazard function can be derived, and vice varsa [23].

Figure 4.1 illustrates a model block symbolizing time to event data. Survival analysis offers several regression models for estimating the hazard rate $\lambda_i(t)$ at time t at which study participants experience the event of interest. It is of interest to determine whether the hazard rate differs across groups, e.g. cohort groups i, which can be examined using survival models by means of relative risks.

Several methods are established for obtaining such estimate, anyway, one of the well-recognized statistical technique for analyzing survival data is the Cox Propotional Hazard model.

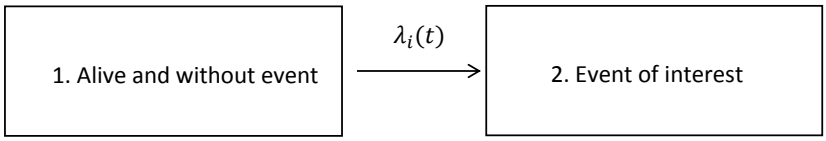

Figure 4.1: *Model block representing a standard survival model. At inclusion the individuals in the study are alive and free of the event of interest. As time passes the individuals in each group i may experience the event of interest with the hazard rate given by $\lambda_i(t)$ e.g. a morbidity-specific or a cause-specific mortality rate.*

4.2 The Cox Proportional-Hazards Model

Cox regression is a method for investigating the effect of covariates on the time of a specified event and deals with the censorings as well as delayed entry. The regression assumes that the effects of the predictor variables upon survival are constant over time and are additive in the log scale.

Under the proportional hazards model, the hazard function for the failure time T associated with possibly covariates is

$$\lambda_i(t|\mathbf{Z}) = \lambda_0(t)\exp(\beta_1 Z_{i1} + \beta_2 Z_{i2} + ... + \beta_p Z_{ip}) \tag{4.6}$$

where $\mathbf{Z}$ is the covariate vector, $\beta_1, \beta_2, ..., \beta_p$ are unknown regression parameters and $\lambda_0(t)$ is an unspecified baseline hazard function, thus the model is termed semi-parametric. Z_{i1} is the covariate value for covariate 1 for individual i, etc. The model is called semi-parametric since the functional form of the baseline hazard is not given. The hazard function is assumed to be proportional for different values of $\mathbf{Z}$, so that the regression coefficient β_m is interpreted as the change in the relative risk on a logarithmic scale when the covariate Z_m is increased by one unit, while all other covariates are kept fixed. The relative risk is given as $\exp(\beta_m)$, that is the ratio of hazards between e.g. two compared subjects [25].

Assuming all event times are distinct, the regression parameters are found by maximising the partial likelihood,

$$L(\beta) = \prod_{i=1}^{d} \frac{\exp(\beta Z_i)}{\sum_{j \in R(t_i)} \exp(\beta Z_j)} \tag{4.7}$$

where $R(t_i)$ is the risk set at death time t_i [1]. The partial likelihood function

[1] Note that βZ is used as a notation for $\sum_{k=1}^{p} \beta_k \times Z_k$.

is a product, over the event times, of a quotient that compares the hazard of the individual with the event at t_i to the hazard of all the individuals at risk at t_i. The function depends only on the order in which the events occur, not the times at which they occur. Thus, the baseline hazard function is not required [23].

The maximum partial likelihood estimator is found by solving the differentiated partial log-likelihood equal zero [32].

4.2.1 Test statistics

The standard asymptotic likelihood tests are also available for Cox partial likelihood to test hypotheses about β. The statistical significance of covariates i.e. the null hypothesis $H_0 : \beta = \beta_0$ can be tested by

- the patrial likelihood ratio test: $2(\log L(\hat{\beta}) - \log L(\beta_0))$
- the Wald test: $(\hat{\beta} - \beta_0)'\hat{\mathcal{I}}(\hat{\beta} - \beta_0)$
- the score test : $U'(\beta_0)\mathcal{I}(\beta_0)^{-1}U(\beta_0)$

where $\hat{\mathcal{I}} = \mathcal{I}(\hat{\beta})$ is the matrix of $\dfrac{\partial^2}{\partial^2\beta} \log L(\beta_0)$ and $U(\beta_0) = \dfrac{\partial}{\partial\beta} \log L(\beta_0)$.
All three tests have χ^2 distributions with p degrees of freedom [3][45].

4.2.2 Assessment of model assumptions

The accuracy of the Cox regression model may be effected by the violation of model assumptions. Assessment of model adequacy can be obtained by considering: linear relation between covariates and logarithm of hazard, proportional hazard assumption and possible need for time-varying covariates.

4.2.2.1 Functional form

For the Cox model it is assumed that the effect of a covariate on the hazard has a log-linear functional form. It is important to test for nonlinear effects since nonlinearity in the model can appear as nonproportional hazards. Diagnosing nonproportional hazards may suggest non-proportionality even though

the problem is due to the incorrect functional form for a coviariate. Hence, correct functional forms for covariates need to be determined and fitted before testing for nonproportional hazards [29].

It is suggested to examine residual plots for the model in question in order to detect a nonlinear patterns to effect for each continuous covariate. However, these plots can be misleading. If the variables included in the model are correlated, the plot may show a linear relationship when the true relationship is nonlinear [46][45].

Instead of analyzing residual plots, the functional form of the covariates can be modeled directly by spline fits and be tested by Wald test in order to decide whether the nonlinear effect should remain in the specification [29]. In this project the nonlinear covariates are detected by the method of spline fits and the non-linearity is avoided by categorizing these covariates.

4.2.2.2 Proportional hazard assumption

The proportional hazard assumption is violated when the effect of a given covariate changes over time. An evaluation of the proportional hazards assumption can be done by many numerical or graphical methods. The graphical diagnostics may be based on plot of log-minus-log survival functions, a plot of cumulative baseline hazards in different groups [6], a plot of the estimated cumulative hazard versus the number of failures [7], a smoothed plot of scaled Schoenfeld residuals versus time [42] etc. The interpretation of the graphical diagnostics can be arbitrary.

There are several numerical approaches for diagnosis of nonproportional hazards, such as including a time dependent covariate in the model [15], a test based on the scaled Schoenfeld residuals which is a difference between the observed and expected value of the covariate at each time [45][8]. It is shown by Grambsch and Therneau (1994)[21] that scaled Schoenfeld residuals can be utilized in assessing the proportional hazards assumption and hence it will be used as a numerical approach for detecting non-proportionality in the analyses.

The Schoenfeld residual expresses the difference between the covariate-value, x_i, for the individual i who died at time t_i and the expected value of the covariate for the risk set at t_i. The test is based on testing for a non-zero slope in a generalized linear regression model fitted for the scaled Schoenfeld residuals on functions of time. A non-zero slope indicates that the proportional hazard assumption is violated. In addition to performing the tests of non-zero slopes, the scaled Schoenfeld residuals can be graphed against a time variable and possible patterns

can be inspected [14].

Several methods can be used for dealing with violation of proportional hazard assumption, one of them is to construct a stratified Cox regression model. By stratification the baseline hazard is allowed to be different across categories called strata based on the value of one or more covariates. The Cox regression can be modified by the stratification of a covariate that does not satisfy the proportional hazards assumption. The coefficients of the remaining covariates are assumed to be constant across strata. Since it is not possible to examine the effect of the stratifying variable, this approach is not employed for dealing with non-proportionality of hazards, but stratification is used only for the covariates that are not of direct interest [17].

Instead, the Cox regression model may be extended to include time-varying variable by assuming that there may be a time dependency from a certain variable. The extension can be obtained by several methods such as inclusion of interaction time and the covariate that violate the proportional hazard assumption or assuming piecewise constant hazards for this variable.

In this project the piecewise constant hazards assumption is applied for corrections for nonproportional hazards. A piecewise constant time-varying hazards model assumes that the hazard is constant not over the whole range of time, but within certain specified intervals of time. An illustration of this approach is depicted in Figure 4.2 where the time axis is split into K intervals.

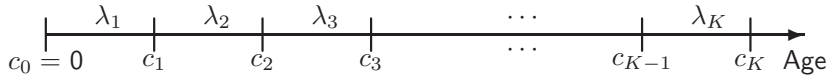

Figure 4.2: *An illustration of splitting time interval into K intervals by assuming piecewise constant but different hazard rates in each of the intervals.*

Thus the hazard rates are given as

$$\lambda(t) = \lambda_k \text{ for } t \in (c_{k-1}, c_k], \ k = 1, ..., K \tag{4.8}$$

The Cox model is then expanded to have time-varying covariates,

$$\lambda_i(t|\mathbf{Z}) = \lambda_{0h}(t) \exp\left(\beta_1 Z_{i1} + \beta_2 Z_{i2} + ... + \beta_{p-1} Z_{ip-1}(t) + \beta_p Z_{ip}(t)\right) \tag{4.9}$$

where λ_{0h} for $h = 1, ..., H$ is the baseline hazards in each of the H strata. The Equation 4.9 includes both stratification of the covariates which are not of direct interest and time-varying covariates which do not satisfy proportional hazard assumption. This model will be considered as the final model in the prospective analysis.

4.3 Multi state models

In order to achieve a more detailed information of the late effects of childhood cancer survivors, more sophisticated model than the model depicted in Figure 4.1 could be useful. Multi-state models form an extension to the survival model by dealing with several and competing events. In a multi-state model the subject may go through different disease states during study. The general well-known multi-state model is called *illness-death model* (Figure 4.3) in which participants start out in the initial state denoted state 1 as healthy, they may become ill and move to state 2 and then they may die in state 3. In multi-state models there are two type of states: absorbing state from which further transitions cannot occur (state 3) and transient state from which further transitions are allowed (state 1 and 2)[4].

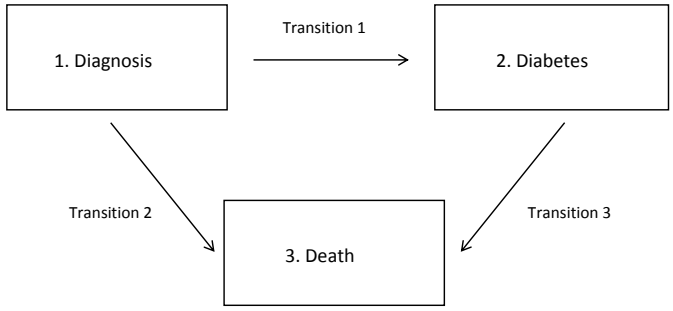

Figure 4.3: *Model block representing the illness-death model.*

An example of a multi-state model is illustrated in Figure 4.3. In this case the initial state is denoted diagnosis, as the state is entered at the moment of diagnosis for cancer or controls. The so-called 'illness' state is denoted diabetes and finally the absorbing state is denoted death as it is the the event of interest in the analysis. It is clear that some participants are censored before they reach an absorbing state. In this study the analysis is restricted to uni-directional multi-state models for which recurrent events are not possible and the transition times are recorded exactly.

4.3.1 Cox regression for multi-state analysis

As in the standard survival analysis the most commonly used semi-parametric model in multi-state survival analysis is the Cox regression model. By assuming

proportional hazards for each transition the hazard rate for an individual i, with time-fixed covariates $\mathbf{Z}_i$ is modeled as

$$\lambda_{hj}^{i}\left(t|\mathbf{Z}_i\right) = \lambda_{hj,0}\left(t\right)\exp\left(\beta_{hj}^{T}\mathbf{Z}_i\right) \tag{4.10}$$

where β_{hj} is a p-vector of regression coefficients for transition $h \to j$ and $\lambda_{hj,0}$ is the baseline hazard function of transition $h \to j$. Notation 4.10 specifies different covariate effects for the different transitions, as well as separate baseline transition hazards for each transition. The model is a natural extension of the Cox proportional hazard model to multi-state models. Thus by not assuming anything about the baseline hazards the model can be used for each of the transition hazards separately [4][33].

The model in Equation 4.10 is considered as a full model and can be tested and reduced to more parsimonious models in several ways:

1. The covariates have an identical effect for each transition and baseline hazards are different. For different transitions Equation 4.10 now simplifies to

$$\lambda_{hj}\left(t|\mathbf{Z}\right) = \lambda_{hj,0}\left(t\right)\exp\left(\beta^{T}\mathbf{Z}\right) \tag{4.11}$$

   where the covariates are time-fixed.

2. The covariates have different effect for each transition. It is assumed that some of the baseline hazards of model 4.10 are proportional, for instance if transitions $h \to j$ and $k \to l$ are proportional,

$$\lambda_{hj,0}\left(t\right) = \tilde{\delta}\lambda_{kl,0}\left(t\right) \tag{4.12}$$

For the last case it is often assumed that transitions going into the same state are proportional. The assumption is based on two reasons: more efficient use of the data and the fact that the hazard ratios of the covariates coding the different transitions into the same state can be interpreted as the effect of the occurrence of an intermediate event,e.g. illness [33]. This case can be modeled by means of a time-dependent covariate $\tilde{Z}(t)$ in the regression model 4.10. The covariate is introduced in order to distinguish between different transitions into the same state; $\tilde{Z}(t)$ is 0 if the subject has not yet experienced an intermediate event and 1 otherwise. The proportionality of transition rates is expressed by the coefficient $\tilde{\beta}$ of $\tilde{Z}(t)$: $\exp(\tilde{\beta}) = \tilde{\delta}$ [16].

It is noticeable that not only the models described above, but also any combination of common or different covariate effects across transitions and of stratified or proportional baseline hazards can be derived by Equation 4.10.

4.3.2 The illness-death model

It is desirable to change the state structure of the illness death model to a progressive model so that each state has only one possible transitions into it and the initial state has none. In this way it is ensured that the current state gives information about which states have been visited previously, and the order in which they have been visited, but the information about the times of the transitions are not given. The advantage of the progressive model is that the differential equations describing the transition probabilities can be simplified to integrals [25].

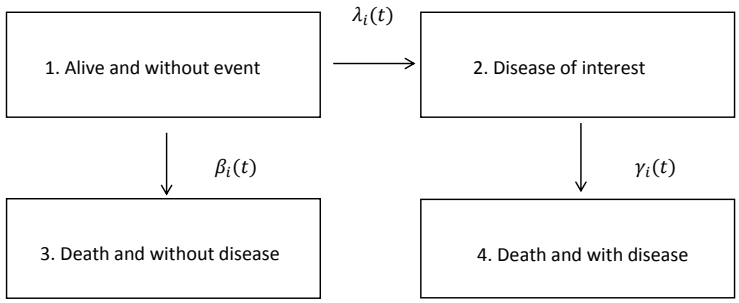

Figure 4.4: *Model block representing the progressive illness-death model.*

The progressive version of the illness-death model is shown in Figure 4.4. The model can be specified by the three transition intensities: the intensity of developing diabetes $\lambda_i(t)$, the death intensity without diabetes $\beta_i(t)$ and the death intensity with diabetes $\gamma_i(t)$.

4.3.3 Transition probabilities

For a multi-state model it is possible to estimate transition probabilities and make long-term predictions.

In order to formalize the notation of the multi state model, the states are denoted with $\mathbb{S} = 1, ..., S$ and a random process $X(t)$ taking values in $\mathbb{S}$ is chosen to describe the course of the model. The transition intensity or hazard rate λ_{hj} expressing the instantaneous risk of a transition from state h into state j at time

t is given by

$$\lambda_{hj}(t) = \lim_{\Delta t \to 0} \frac{P(X(t+\Delta t) = j | X(t) = h)}{\Delta t} \tag{4.13}$$

The process given in 4.13 is Markovian, since it presents the assumption that the future depends on the past only through the present, that is to say:

$$P(X(t+\Delta t) = j | X(t) = h, \{X(s), s < t\}) = P(X(t+\Delta t) = j | X(t) = h)$$

Hence the transition probability from state h to state g in the time interval $(s, t]$ is denoted by [16][4].

$$P_{hg}(s,t) = P(X(t) = g | X(s) = h) \tag{4.14}$$

The relationships between the rates and the probabilities are as follows. The probability of remaining in the first state given in Figure 4.4 in the time interval $[0, t]$ is

$$P_{11}\{\text{state 1 at } t\} = \exp\left(-\int_0^t \lambda(s) + \beta(s)\, ds\right) \tag{4.15}$$

that is, the contribution to the likelihood of a censoring at time t, i.e. survival function. The probability of dying without developing a disease is,

$$P_{13}\{\text{state 3 at } t\} = \int_0^t \beta(s) \exp\left(-\int_0^s \lambda(u) + \beta(u)\, du\right) ds \tag{4.16}$$

and the probability of developing a disease is,

$$P_{12}\{\text{state 2 at } t\} = \int_0^t \lambda(s) \exp\left(-\int_0^s \lambda(u) + \beta(u)\, du\right) \tag{4.17}$$

$$\times \exp\left(-\int_s^t \gamma(u)\, du\right) ds \tag{4.18}$$

Likewise, the probability of dying with the disease is,

$$P_{24}\{\text{state 4 at } t\} = 1 - P_{11}\{\text{state 1 at } t\} - P_{13}\{\text{state 3 at } t\} \tag{4.19}$$

$$- P_{12}\{\text{state 2 at } t\} \tag{4.20}$$

Given that the hazard rates are calculated, the transition probabilities can easily be estimated by means of these formulaes in practice by summing up over small intervals [11][9].

It might be mentioned that the Cox models described above are assumed to be Markov models. When there is not a time-dependent covariate in the model, the model is called homogeneous Markov model whereas if a time-dependent covariate is present the model is called semi-Markov model [25].

4.3.4 Multi-state models using R

In the recent years, multi-state models have gained popularity due to significant improvements in information technology that make it possible to record detailed information on clinical events on large numbers of patients. However, theoretical study and their application has been rather limited. Two reasons for this limitation are stated: the method is more advanced then standard survival analysis and a lack of good software for the analysis of multi-state models [41][16].

However, in recent years a number of software packages have been developed for the analysis of such models. In Table 4.1 a list of these packages and their contribution to the multi-state models is gathered. It should be noted that all of them is R packages, or objects within R packages.

Packages	
`timereg`	is devoted to competing risks models.
`msm`	is based on parametric models.
`p3state`	can be used for forward-going multi-state models with a single starting and end state. Do not rely on Markov assumption.
`etm`	estimates non-parametric general Markov multi-state models.
`Epi`	studies data with multiple time-scales for Cox model and Poisson model.
`mstate`	deals with non- and semi-parametric Cox models.

Table 4.1: *A list of available packages for analysis of multi-state models in R.*

The softwares have their own advantages and disadvantages. The common principle in these softwares is to make an appropriate data set that represent each subject by several observations. Since the main interest in this study is to construct semi-parametric models and time to event in data is observed exactly, the most appreciated softwares that can be used for constructing multi-state models are `Epi` and `mstate`. Both of the packages are tested for constructing multi-state models. Since `Epi` is considered to be more user-friendly and flexible, the results obtained from this package is presented in the thesis.

4.3.5 Model assumptions

The assumptions made for the multi-state model analysis are listed below.

1. Since the distribution of the survival times in this study are unknown and some prognostic covariates are available in the data set, semi-parametric models that allow analyzing the effect of the covariates will be used.

2. Age is chosen to be the time scale in this study. Hence, in the data set only right censoring and left truncation is allowed and all the event times are observed exactly.

3. A certain number of the covariates that may effect hazard rates are given in the data set. However, other covariates that may influence survival such as life style, smoking, health status and genetic risk factors are unknown and cannot be included in the analyses. These factors are ignored.

4. Since there is one process for each person, the given data set is a a longitudinal data. Thus it can be assumed that there is independence between different subjects and, possibly, dependence between the times to the event for the same subject [25].

Chapter 5

Results two-state models

The aim of this chapter is to analyze mortality and morbidity outcomes separately. This is investigated by means of two-state survival models specifically using Cox proportional Hazards model presented in Chapter 4, where time to event of interest is considered and all other events are censored. Furthermore, extended models with time-varying covariates are examined. In the first section the methods used for model verification and validation for the first analysis is presented. The results of the remaining analyses are given in the following section and the chapter is summarized with an overall conclusion.

The analysis conducted in this chapter is based on the standard survival analysis in which time to event of interest i.e. a morbidity-specific and a cause-specific mortality rate is investigated using survival models by means of relative risk. Figure 5.1 illustrates a model block that represents the first two-state standard survival model where there is a transition from the initial state; diagnosis to the absorbing state; dead. Since the event of interest is death without diabetes in this analysis, the participants that develop diabetes is censored, and the time in which they are under risk is included in the risk set. When there are more than one competing event in a data set and even though interest may focus on a single event, the analysis may be performed by censoring individuals at the time of the second event in order to obtain a valid inference [5].

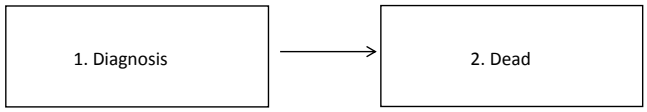

Figure 5.1: *Model block representing two-state standard survival model.*

When analysing mortality and morbidity outcomes separately, it is desirable to assess the goodness of fit for the Cox model constructed in each analysis. To clarify how the results of the models are examined for model verification and validation, a general example of the model assessment is given only for the first analysis. The assessment of model adequacy of the remaining analyses is skipped, but the results of the analyses are presented and interpreted.

5.1 Analysis of mortality rate

It is of interest to examine the mortality rate in childhood cancer survivors compared to the general population. A simple model is set up and then more detailed models are presented.

An univariate estimate of the model with only Groups as explanatory variable is performed at first. The general Cox proportional hazard model can be written as:

$$\lambda_i\left(t|X\right) = \lambda_0\left(t\right)\exp\left(\beta_1 x_{i1}\right)$$

where x_{i1} is Groups with the levels [unexposed, exposed] and β_1 is the effect of the covariate. The result obtained from the model is given in Table 5.1. The table includes information about the coefficient β_1, the risk $\exp(\beta_1)$, the standard deviation of the coefficient and the lower and the upper limit of a 95% confidence interval for the risk. It is obvious from the table that the exposed has a significantly higher mortality rate which is 10.3 times [95% CI, 9.8-10.9] higher compared to unexposed.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
Exposed	2.336	10.341	0.026	9.824	10.890	$<$ 2e-16

Table 5.1: *Results from the univariate Cox proportional hazards model with Groups as covariate. The hazard ratio is given as exp(coef).*

In addition to Groups, other confounders are tested for significance by using forward model selection approach proposed by Hosmer and Lemeshow [24]. The confounders are then included in the model in order to explain the variation in the mortality rate. Since the effect of the covariates Gender and Country is not of direct interest, the model is stratified by these. The multivariate Cox proportional model to apply to the data can be expressed as

$$\lambda_i\left(t|X\right) = \lambda_{0,k}\left(t\right)\exp\left(\beta^T X_i\right) \tag{5.1}$$

where $\lambda_{0,k}$ for $k = 1, ..., K$ is the baseline hazards in each of the K stratum. The covariates $X_i = [x_{i1}, x_{i2}, x_{i3}]$ are described as follows:

$$x_{i1} = \text{Groups} = \begin{cases} 0, & \text{if } i \text{ unexposed;} \\ 1, & \text{if } i \text{ exposed.} \end{cases}$$

$$x_{i2} = \text{Age}$$

$$x_{i3} = \text{Calendar}$$

where Age is the age at diagnosis and Calendar is the calendar year at diagnosis, both are quantitative variables. Since unexposed are not diagnosed, the age at diagnosis for this group is set to the average age of exposed at diagnosis.

Although the Cox model is semi-parametric that is to say no assumptions are made about the form of the baseline hazard, some of the important issues are needed to be considered before the results of the model can be safely applied.

5.1.1 Verifying model assumptions

It is desirable to check if the relationship between the dependent variable and the independent variables can be adequately described by the model. For this purpose two kinds of diagnostics will be considered: the functional form and the assumption of proportional hazards.

5.1.1.1 Functional form

One of the assumptions made for Cox proportional hazards model is that the effects of covariates are linear on the log risk scale. When this assumption is violated and the nonlinear effects in the Cox model is ignored, the results obtained from the model will lead to erroneous statistical conclusions. Incorrect functional form for a covariate can appear as nonproportional hazards. Hence, the functional form for the covariates needs to be determined before the proportional hazard assumption is examined. As mentioned in section 4.2.2.1, the method of smoothing spline fits are used for diagnosing nonlinearity.

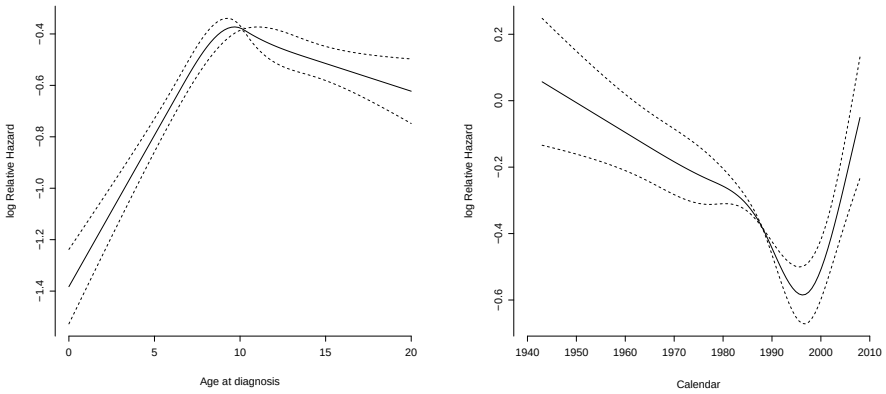

Figure 5.2: *Functional form of age at diagnosis (left) and calendar year at diagnosis (right) on log hazard of death. In both plots the thick lines represent the spline fit while the dashed lines represent 95% confidence bands for the fit.*

A Cox model with smoothing spline fits are estimated for the quantitative covariates after controlling for other confounders. A plot of the nonlinear effects of the covariates is displayed in Figure 5.2. Taking them one-by-one, it is seen that age at diagnosis appears to have an increasing effect on the hazard until a threshold, say age at diagnosis 9 is met. After this threshold is reached, the age at diagnosis performs a downward sloping effect. For calendar year at diagnosis a decreasing effect on the hazard is observable between 1940-1995. After 1995 the effect of calendar year appears to be strongly positive, but at the end it only reaches HR = 1 though.

It is obvious that the plots reveal evidence of nonlinearity in the effects. The correction of the nonlinear functional forms can be obtained by including splines in the Cox model or categorizing the nonlinear covariates. The latter is preferred to be applied in the model. Based on the plot given in Figure 5.2 age at diagnosis is categorized into 4 intervals: [0 - 5], [5 - 10], [10 - 15] and [15 - 20], and calendar year is categorized into 3 intervals: [1943 - 1960], [1960 - 1974] and [1974 - 2010].

5.1.1.2 Proportional hazard assumption

Before testing for nonproportional hazards for all covariates, the proportionality assumption of the model including only the covariate Groups is examined. There are a number of basic concepts to check if the explanatory variables analysed

satisfy the proportional hazard assumption of the model. One of the graphical diagnostics is based on plot of log minus log of the estimated survival. If the proportionality assumption is satisfied, the log minus log of estimated survival functions are supposed to be proportional curves.

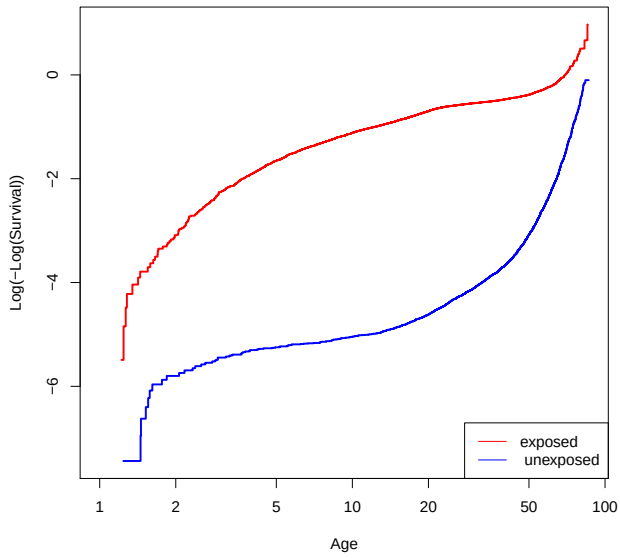

Figure 5.3: *Graphical check of proportionality assumption for groups. Log minus log plots for exposed and unexposed.*

As it appears in Figure 5.3 the curves are far from being proportional and thus the proportional hazards assumption for this covariate is violated. Several methods for dealing with violation of the proportional hazards assumption are presented in Chapter 4. However, one of the most appropriate methods is to include time-dependent covariates in the model [28]. Since groups seems to be a time-dependent covariate, the extended Cox model will be constructed for this. In order to do that it is assumed that the hazard ratio for exposed compared to unexposed is piecewise constant but different within some time intervals. The time scale is here divided into 6 pieces and the hazard ratios in each interval is calculated based on this assumption. The piecewise constant hazard model is

set up as in Equation 5.1 where the covariate Groups is given as

$$Groups = \begin{cases} \text{episode 0-5}, & \text{if } t \in (0;5]; \\ \text{episode 5-10}, & \text{if } t \in (5;10]; \\ \text{episode 10-15}, & \text{if } t \in (10;15]; \\ \text{episode 15-20}, & \text{if } t \in (15;20]; \\ \text{episode 20-25}, & \text{if } t \in (20;25]; \\ \text{episode 25+}, & \text{if } t > 25; \\ \text{unexposed}, & \end{cases} \tag{5.2}$$

The indicator functions that Cox model needs in order to estimate hazard ratios for groups for ith subject are:

$$\begin{aligned} x_{i1}(t) = \text{ episode 0-5} &= \begin{cases} 1, & \text{if } t \in (0,5]; \\ 0, & \text{otherwise.} \end{cases} \\ x_{i2}(t) = \text{ episode 5-10} &= \begin{cases} 1, & \text{if } t \in (5,10]; \\ 0, & \text{otherwise.} \end{cases} \\ &\vdots \\ x_{i6}(t) = \text{ episode 25+} &= \begin{cases} 1, & \text{if } t > 25; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

By this method the Cox model is extended to have time-varying covariates and thus the model becomes a piecewise constant hazard model. The hazard ratios of exposed based on the extended Cox model is estimated and plotted in Figure 5.4. The figure shows the hazard ratios and the corresponding 95% confidence intervals for Groups with unexposed as reference in each interval together with a line for HR = 1. As it is expected, the mortality rate in exposed is strongly high within the first 5 year after diagnosis compared to the unexposed and the rate decreases nearly linearly with increasing time since diagnosis. None of the hazard ratios fall under HR = 1.

The test for the proportional hazard assumption of the final model is performed using the scaled Schoenfeld residuals. The test statistic is based on tests of the proportional hazards assumption for each covariate, by correlating the corresponding set of scaled Schoenfeld residuals with a transformation of time [17]. The test for each covariate using the 'correlation with time' test is displayed in Table 5.2 with a corresponding correlation coefficient and a two-sided p-value. It appears that the correlation between residuals of all covariates and the transformed survival time is quit low. The p-values indicate that there is no evidence of non-proportional hazard for almost all covariates once the nonlinear functional forms for covariates have been taken into account and time varying effects are corrected. Furthermore, the piecewise constant hazard assumption holds for all episodes except episode 25+. It may be due to too few observations

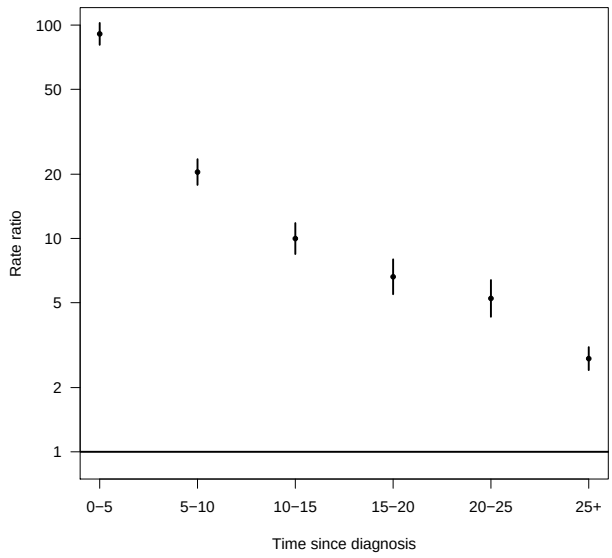

Figure 5.4: *Rate ratios and the corresponding* 95% *confidence intervals for Groups with non-exposed as reference in each interval.*

	ρ	χ^2	p -value
Age [5-10]	-0.014	1.343	0.246
Age [10-15]	-0.012	0.894	0.344
Age [15-20]	-0.016	1.595	0.207
Calendar [1960-1974]	-0.005	0.124	0.725
Calendar [1974-2010]	0.004	0.085	0.770
episode [0-5]	-0.015	1.218	0.270
episode [5-10]	-0.012	0.781	0.377
episode [10-15]	-0.010	0.601	0.438
episode [15-20]	-0.006	0.220	0.639
episode [20-25]	-0.008	0.333	0.564
episode 25+	-0.055	18.699	< 0.001

Table 5.2: *Test results of proportional hazards assumption.*

in the last interval and thus can be ignored. Overall, the final model fulfils the various assumptions and the model results can safely be interpreted.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p-value
Age[5-10]	0.334	1.397	0.064	1.231	1.584	< 0.001
Age[10-15]	0.253	1.288	0.070	1.123	1.477	< 0.001
Age[15-20]	0.173	1.189	0.068	1.041	1.360	0.011
Cal[1960-1974]	0.001	1.000	0.056	0.896	1.117	0.995
Cal[1974-2010]	-0.358	0.699	0.069	0.611	0.801	< 0.001
episode[0-5]	4.510	90.924	0.060	80.785	102.336	< 0.001
episode[5-10]	3.019	20.471	0.071	17.817	23.521	< 0.001
episode[10-15]	2.301	9.988	0.085	8.451	11.806	< 0.001
episode[15-20]	1.889	6.612	0.096	5.478	7.981	< 0.001
episode[20-25]	1.655	5.235	0.101	4.293	6.384	< 0.001
episode25+	1.006	2.734	0.063	2.415	3.096	< 0.001

Table 5.3: *Results from the extended multivariate Cox model constructed for analysing mortality rate.*

The result of the final extended multivariate Cox model is displayed in Table 5.3. In the model the reference variable for exposed is unexposed, the reference for age at diagnosis is the age group [0-5] and the reference for calendar year is the calendar year [1943-1960]. It appears that the effect of the age at diagnosis in the intervals [5-10], [10-15] and [15-20] on the mortality rate is statistically significant compared to the effect of age at diagnosis in the interval [0-5] after adjustment for the other explanatory variables in the model. A participant with an age at diagnosis [5-10] has 40% [95% CI, 23-58%] increased risk of death compared to a participant with an age at diagnosis in the interval [0-5]. The risk of mortality is decreasing with increasing age at diagnosis, but it is still significant. This is also reflected by the smoothing spline function shown in Figure 5.2. The model indicates that there is no difference in the mortality rate between the calendar years at diagnosis [1960-1974] and [1943-1960] which is denoted by a nonsignificant p-value and also by confidence intervals including hazard rate 1. A participant diagnosed in [1974-2010] has 30% [95% CI, 20-39%] less risk of death than a person diagnosed in [1943-1960] after adjustment for prognostic factors. As it is observed before, the risk of mortality in exposed is extremely high in the first 10 years after diagnosis compared to the unexposed. Although this risk decreases as time passes, it remains significant.

A childhood cancer survivor study conducted in 25 centers in the United States and one in Canada has shown some similar results. The cohort includes five-year survivors of childhood cancer diagnosed with cancer before age 21 years between 1970 and 1986. The study has shown that long-term survivors are at an 8.4-fold

increased risk of premature death when compared with an age-matched and gender-matched general population [70][38]. An other population-based study in the five Nordic countries (Denmark, Finland, Iceland, Norway, and Sweden) has assessed the risk of death in five-year childhood cancer survivors who were diagnosed with cancer before the age of 20 years between 1960 and 1989. It has been clearly demonstrated that the overall late mortality is significantly lower in survivors diagnosed from 1980 to 1989, compared with those diagnosed from 1960 to 1979 (hazard ratio, 0.61; [95% CI, 0.54 - 0.70])[36].

5.2 Results of the remaining analyses

In this section results of the other two analyses are given without a detailed assessment of model verification and validation. As in the previous section some simple models are set up based on the proportionality assumption and later on, extended models are used in order to explain the variation in morbidity rate and mortality rate after diabetes.

5.3 Analysis of morbidity rate

It is of interest to examine the diabetes-related morbidity in exposed and give an interpretation of the results. The first approach is to determine the effect of the exposed and unexposed to cancer on overall diabetes survival time. Since the event of interest is the occurrence of diabetes in this analysis, all other events (e.g. death) are censored. The results of the univariate proportional Cox model with only Groups as covariate is presented in Table 5.4.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
Exposed	0.498	1.646	0.060	1.461	1.854	<0.001

Table 5.4: *Results from the univariate Cox proportional hazards model constructed for analysing morbidity rate.*

As it appears in the table the diabetes-related morbidity rate in childhood cancer survivors is 65% [95%, CI 46-85%] higher compared to the general population. To examine how this rate changes by adjusting for other confounders the multivariate analysis is performed. When investigating the additivity assumption, it

is found that the functional form of the age at diagnosis and calendar year are nonlinear and therefore categorization of these are necessary. A test for proportional hazards assumption has shown that non of the variables have problems with proportionality implied by a nonsignificant p-value. Cox regression of diabetes survival time is stratified by gender and country. The result of the final model obtained is listed in Table 5.5.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p-value
Exposed	0.742	2.101	0.113	1.683	2.623	< 0.001
Age [5-10]	0.097	1.102	0.178	0.778	1.562	0.584
Age [10-15]	0.129	1.138	0.162	0.828	1.565	0.426
Age [15-20]	-0.359	0.698	0.154	0.516	0.945	0.020
Cal [1960-1974]	0.296	1.345	0.085	1.137	1.590	0.001
Cal [1974-2010]	0.394	1.483	0.104	1.209	1.820	< 0.001

Table 5.5: *Estimated regression coefficients with hazard ratios and* 95% *confidence intervals based on stratified proportional Cox model constructed for analysing morbidity rate.*

From the results it is seen that the estimated risk of experiencing diabetes in childhood cancer survivors is 2.1 [95% CI, 1.7-2.6] times the general population, holding other covariates constant. There is not a statistical significant difference between the effect of age at diagnosis [0-5] and [5-10],[10-15] on experiencing diabetes. Participants with an age at diagnosis [15-20] have an estimated 30% [94% CI, 10-49 %] decreased risk of developing diabetes compared to participants in age group [0-5]. It is reported that most of the late effects in childhood cancer survivors are caused by cancer treatments that damage quickly growing healthy cells [71]. Thus the older a childhood cancer survivor is at diagnosis, the lower risk of experiencing diabetes may be expected for the person in question. This expectation is reflected in the results. It can be commended that calendar year at diagnosis has a positive effect on the morbidity rate. The risk of developing diabetes for a person diagnosed in [1960-1974] is 35% [95% CI, 14-59%] higher and for a person diagnosed in [1974-2010] is 48% [95% CI, 21-82%] higher compared to a person diagnosed in [1943-1960] holding other covariates constant. The result can be due to late registration of diabetes events discussed in descriptive analysis in Chapter 3 or due to an increase in the diabetes events in the last 50 year. Researches has found an increasing prevalence of diabetes in the Danish population over the last decades [20]. In a study the prevalence of diabetes in 1995 - 2006 in Denmark is investigated and it is reported that the prevalence increased by 6% per year [9].

Parallels can be drawn between the recovered results and a childhood cancer survivor study conducted for characterizing the risk of morbidity among North American cohort of long-term survivors. Participants in the study were diagnosed before the age of 21 years from 1970 to 1986 and who were alive at least 5 years after their original diagnosis. It is found that survivors compared with siblings were 1.6 times as likely to have diabetes mellitus [95% CI, 1.2-2.2; p -value< .01] after adjustment for age at interview, sex, race/ethnicity, household income, and health insurance. Survivors of childhood cancer diagnosed before age 5 were 2.4 times more likely to report diabetes than those diagnosed in late adolescence (from ages 15 to 20)[34].

5.4 Analysis of mortality rate after developing diabetes

It is reported that cancer treatments such as radiation, chemotherapy, and biologic agents increase the risk of cardiovascular disease in survivors of childhood cancer; in fact, cardiovascular disease is the leading cause of non-cancer mortality in select cancers [71]. Researcher has found that childhood cancer survivors are at increased risk for diabetes, high cholesterol and high blood pressure all of which cause heart disease [70]. Consequently, it is of interest to examine the mortality rate among childhood cancer survivors after experiencing diabetes.

As in the previous section an univariate analysis is conducted by considering the effect of exposed alone. The fitted model is based on the proportionality assumption of the hazards and the result is given in Table 5.6. It appears that the exposed to cancer are 2.2 [95% CI, 1.6-2.9] times more likely to die after experiencing diabetes compared to unexposed with the disease.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
Exposed	0.772	2.164	0.150	1.612	2.905	< 0.001

Table 5.6: *Results from the univariate Cox proportional hazards model constructed for analysing mortality rate after experiencing diabetes.*

To explore the relations between the variables while simultaneously adjusting for all other variables that has influences on the outcome of interest mortality an multivariate model analysis is performed. As before the model assumptions

are checked in order to enhance the accuracy of the model. It is verified that the model satisfies neither the linearity assumption nor the proportional hazards assumption. The correction of non-linear functional forms of the covariates are obtained by categorizing the nonlinear covariates; age at diagnosis and calendar year and including these in the model. For assessing violations of the proportional hazard assumption, it is assumed that the variable Groups is a piecewise time-varying covariate and thus the hazard ratio is piecewise constant. In order to apply this assumption the time axis is divided into 3 pieces given as $[0; 5], [5; 10]$ and $[10; 15]$ and a piecewise constant Cox model is set up as in Equation 5.1. The covariate Groups is given as in Equation 5.2, though only for the introduced intervals. The results of the final extended Cox model stratified by gender and country is listed in Table 5.7.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p-value
Age[5-10]	0.308	1.360	0.556	0.458	4.043	0.580
Age[10-15]	1.130	3.095	0.454	1.270	7.543	0.013
Age[15-20]	1.085	2.960	0.459	1.204	7.277	0.018
Cal[1960-1974]	-0.195	0.823	0.191	0.566	1.196	0.307
Cal[1974-2010]	-0.249	0.780	0.298	0.434	1.399	0.404
episode[1-5]	2.869	17.627	0.283	10.116	30.714	< 0.001
episode[5-10]	2.535	12.619	0.338	6.500	24.496	< 0.001
episode[10-15]	1.773	5.887	0.618	1.753	19.771	0.004
episode15+	2.474	11.869	0.501	4.447	31.682	< 0.001

Table 5.7: *Results from the extended multivariate Cox model constructed for analysing mortality rate after experiencing diabetes.*

It is seen that the effect of all covariates on mortality rate is statistically significant except the effect of calendar year. Participants with an age at diagnosis in the interval [10-15] and [15-20] are associated with an increased risk of mortality after experiencing diabetes compared to the participants with an age at diagnosis in the interval [0-5] after adjustment for prognostic factors. The risk of mortality for a participant with an age at diagnosis in the interval [10-15] and [15-20] is nearly 3 times the risk for a participant with an age at diagnosis in the interval [0-5]. As in the analysis for mortality rate without developing diabetes, the exposed is associated with a strongly significant risk of mortality compared to general population after controlling for possible confounding of exposure effects.

Diabetes is considered as being a metabolic syndrome which is highly associated with cardiovascular events and mortality in childhood cancer survivors

[70]. Researchers have found that survivors are 8 times more likely to die from cardiovascular-related disease than the general population [18][40]. In a study the risk for disease- and treatment-associated late mortality of five-year survivors of childhood and adolescent cancer is investigated. Survival diagnosed between 1970 and 1986 is matched with the age-and sex- comparable US population. It is reported that survivors has 8.4 times higher mortality compared with the matched US population. Furthermore, it is found that deaths from pulmonary, cardiac, and other causes are relatively low during the [5-15] year interval, but increases are observed [15-30] years after diagnosis of the original cancer [35]. This result is consistent with the cumulative incidence of mortality after experiencing diabetes in Figure 3.4 presented in Chapter 3.

5.5 Conclusion

The mortality and morbidity outcomes in childhood cancer cohort compared to the general population cohort is analysed separately by means of extended Cox hazards model. The first analysis conducted for investigating mortality rates without experiencing diabetes has shown that childhood cancer survivors have an extreme high risk of death compared to the general population. The highest risks are observed in the first 10 years after diagnosis. It is found that age of participants at diagnosis has a significant effect on the mortality rate after adjustment for the other explanatory variables in the model. A participant diagnosed at the age in the interval [5-10], [10-15] and [15-20] has a higher risk of mortality than a participant diagnosed at an age in the interval [0-5] when holding other confounders constant. It is observed that diagnosing in calendar year [1974-2010] has a negative effect on the mortality rate compared to the calendar year [1943-1960].

The analysis of morbidity rate has shown that the estimated risk of developing diabetes in childhood cancer survivors is 2.1 [95% CI, 1.7-2.6] times the general population cohort after adjustment for prognostic factors. Comparison of the effect of age at diagnosis on morbidity has revealed that participants with an age at diagnosis [15-20] have an estimated 30% [95% CI, 10 -49%] decreased risk of developing diabetes compared to participants in age group [0-5]. Furthermore, it is found that calendar year at diagnosis has a positive effect on the morbidity rate.

The analysis performed for investigating mortality rates after experiencing diabetes provides a strong evidence for strongly high risk of mortality at exposed compared to unexposed. Calendar year at diagnosis does not seem to be a significant confounder for explaining the variation in the mortality rate whereas

age at diagnosis shows a significant effect on the mortality rate.

Chapter 6

Results multi-state models

Modeling survival time for each cause of mortality and each morbidity outcome has been considered separately so far. In some cases, however, it may be relevant to investigate outcomes jointly in order to give more biological insight into the late effects of the cancer survivors. In the present chapter, the application of multi-state models presented in Chapter 4 will therefore be assessed, first by an univariate analysis and then by a multivariate analysis. In both analyses several multi-state models will be constructed, and the statistical comparison of the models will be obtained by an anova test. Finally, a prediction of the transition probabilities will be made by using cumulative risks for each cause of death.

In order to perform analyses based on multi-state models, it is required that the allocation of follow-up time to states in the illness-death model, and timescales are done properly. To facilitate these required manipulations of data for multi-state models, a Lexis object is set up. The model structure is illustrated in a box diagram shown in Figure 6.1. The transitions; Tr. 1, Tr. 2 and Tr. 3 between states are denoted by arrows and the number of transitions are given on each arrow.

A summary of the Lexis object is displayed in Table 6.1. The transitions between entry state; diagnosis, intermediate state; diabetes and exit state; death with and without diabetes are given along with the rates of the transitions per 10

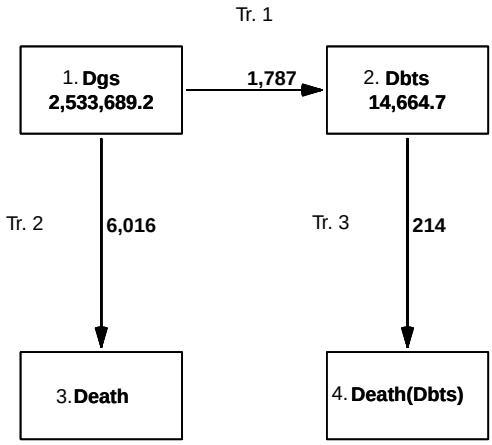

Figure 6.1: *Illustration of the transitions between states; diagnosis, diabetes, death with and without diabetes. The numbers on the arrows are the number of transitions, and the number in the boxes are person-years.*

```
   Transitions:
      To
 From      Dgs Dbts   Dth Dth(Dbts)  Records: Events: Risk time: Persons:
   Dgs  134677 1787  6016         0    142480    7803 2533689.21   142480
   Dbts      0 1573     0       214      1787     214   14664.69     1787
   Sum  134677 3360  6016       214    144267    8017 2548353.90   142480

 Rates:
      To
 From   Dgs     Dbts      Dth   Dth(Dbts)   Total
   Dgs    0    0.007    0.024           0   0.031
   Dbts   0        0        0       0.146   0.146
```

Table 6.1: *A summary of the Lexis object showing the transitions between entry and exit states and the transition rates which are multiplied by 10.*

person-years. It is observed that in this study 6016 out of 142480 participants have died without experiencing diabetes and 1787 participants have experienced diabetes of whom 214 are dead. Thus, among 142480 persons there are 1787 events, i.e. occurrence of diabetes during 2533689 person-years, corresponding to a rate of 0.007 events per 10 person-years. Similarly, the mortality rate without

experiencing diabetes is 2.4% per 10 person-years whereas the mortality rate after experiencing diabetes is 14.6% per 10 person-years.

6.1 Univariate analysis

In the previous chapter a separate analysis for each outcome with only groups as covariate was presented in order to achieve a basic understanding of the effect of the covariate. Similarly, an univariate analysis will now be conducted jointly for all transitions shown in Figure 6.1.

Each model analyzed in this chapter is based on the Cox proportional hazards model. It should be mentioned that the non-proportional hazards is not modeled by assuming piecewise constant hazards as in the previous chapter, instead a global hazard ratio for the covariate is taken into consideration.

6.1.1 Model I

Expecting the covariate Groups to have a different effect on each transition, the first model used for this analysis is given as,

$$\lambda_{hj}\left(t|\mathbf{Z}\right) = \lambda_{hj,0}\left(t\right)\exp\left(\beta_{hj}^{T}\mathbf{Z}\right) \tag{6.1}$$

where β_{hj} is the vector regression coefficients corresponding to the transition from state h into j. Equation 6.1 describes a stratified Cox model, in which each stratum represents one transition. Thus, it is assumed that there are different covariate effects for the different transitions and separate baseline transition hazards for each transition. In this case, $\mathbf{Z}$ is called a transition specific covariate. Considering just one basic covariate $\mathbf{Z}$, the three transition-specific covariate vectors $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{Z}_3$ are defined as $\mathbf{Z}_1 = (Z,0,0)^T$, $\mathbf{Z}_2 = (0,Z,0)^T$, $\mathbf{Z}_3 = (0,0,Z)^T$ and the regression vector $\beta = (\beta_1,\beta_2,\beta_3)^T$ has length 3 [16].

Considering Figure 6.1, the Cox proportional hazards model for each transition is given by

$$\lambda_{12}\left(t|Z\right) = \lambda_{12,0}\left(t\right)\exp\left(\beta_{12}Z\right) \tag{6.2}$$

for the transition going from state diagnosis to state diabetes,

$$\lambda_{13}\left(t|Z\right) = \lambda_{13,0}\left(t\right)\exp\left(\beta_{13}Z\right) \tag{6.3}$$

for the transition going from state diagnosis to state death and

$$\lambda_{24}(t|Z) = \lambda_{24,0}(t)\exp(\beta_{24}Z) \tag{6.4}$$

for the transition going from state diabetes to state death. Model I (Equation 6.1) can be reduced to more parsimonious models where some baseline hazards are assumed to be proportional or where some covariates have the same effect on several transition intensities [4].

Assuming that the two mortality rates; diagnosis → death and diabetes → death are proportional, the model given in Equation 6.1 can be reduced to

$$\lambda_{12}(t|Z) = \lambda_{12,0}(t)\exp(\beta_{12}Z) \tag{6.5}$$
$$\lambda_{24}(t|Z) = \lambda_{12,0}(t)\exp(\beta_{24}Z) \tag{6.6}$$

Note that only transitions going into the death states are considered. In this case there are different covariate effects for each transition with regression coefficients (β_{12}, β_{24}), but the transitions share a common baseline hazard; $\lambda_{12,0}$.

This model can additionally be reduced to a model where the effect of the covariates are assumed to be common for both transitions, i.e.

$$\lambda_{12}(t|Z) = \lambda_{12,0}(t)\exp(\beta Z) \tag{6.7}$$
$$\lambda_{24}(t|Z) = \lambda_{12,0}(t)\exp(\beta Z) \tag{6.8}$$

where β is a single, common effect of the covariate Z.

In the following results obtained from model I will be presented. Subsequently, the reduction of this model will further be explained in details and the results of these will be given.

6.1.1.1 Results

The results from model I is gathered in Table 6.2. T1 denotes the transition from state diagnosis to state diabetes, T2 denotes the transition from state diagnosis to state death and T3 denotes the transition from state diabetes to state death with diabetes. The effect of a covariate for a transition is given as the interaction with the transition in question. It is worth noticing that the results from the multi-state models are nearly similar to the results from the Cox models conducted separately for each transition.

In the analysis in the previous chapter it is seen that the childhood cancer survivors have 64.6% [95% CI, 46.1-85.4] increased risk of developing diabetes

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
T1:exposed	0.512	1.668	0.062	1.479	1.883	< 0.001
T2:exposed	2.335	10.334	0.026	9.817	10.879	< 0.001
T3:exposed	0.772	2.164	0.150	1.612	2.905	< 0.001

Table 6.2: *Results from the univariate Cox proportional hazards multi-state model assuming different covariate effects and different baseline hazards.*

compared to the general population. In the joint model, however, this risk is increased to 66.8% [95%, CI, 47.9-88.3]. A small deviation in the results from the two approaches for mortality rates without developing diabetes (transition 2) is similarly observable. But comparison of the results for transition 3 i.e. the mortality rate after developing diabetes shows an identical estimation of this rate.

6.1.2 Model II

In order to reduce model I to a model which is more parsimonious it is assumed that hazard rates going into the same state are proportional, e.g. in this case the two transitions into the death state, cf. Figure 6.1. This is equivalent to grouping the transition 2 and 3 and using the occurrence of the intermediate event, diabetes as a time dependent covariate.

Considering only this two transitions, Equation 6.1 implies for the first transition intensity,

$$\lambda_2 \left(t|\mathbf{Z}\right) = \lambda_{2,0} \left(t\right) \exp \left(\beta_2 \mathbf{Z}\right) \tag{6.9}$$

and similarly for transition 3,

$$\lambda_3 \left(t|\mathbf{Z}\right) = \lambda_{2,0} \left(t\right) \exp \left(\beta_3 \mathbf{Z} + \tilde{\beta}\right) \tag{6.10}$$

where $\tilde{\beta}$ is the coefficient of a time-dependent covariate $\tilde{Z}(t)$ that is an indicator of being in state, diabetes. The indicator is introduced in order to avoid a model for which rates are assumed to be identical. $\tilde{Z}(t)$ distinguishes between different transitions into the same state: $\tilde{Z}(t)$ is 0 if the participant has not yet experienced diabetes and 1 afterwards. The coefficient, $\tilde{\beta}$ is also an expression for the proportionality factor between the two baseline hazards.

Note that the same baseline hazard $\lambda_{2,0}$ for both transition 2 and 3 is used. The covariate vector, $\mathbf{Z}_2$ is defined either as $(Z, 0, 0)^T$ for transition 2 or as $(0, Z, 1)^T$ for transition 3, and the regression vector is given as $\beta = (\beta_2, \beta_3, \tilde{\beta})^T$. Hence, $\mathbf{Z}_2$ is now a time-dependent covariate; if t_{ds} is time of occurrence of diabetes, then $\mathbf{Z}_2 = (Z, 0, 0)^T$ for $t \leq t_{ds}$, and $\mathbf{Z}_2 = (0, Z, 1)^T$ for $t > t_{ds}$ [16].

6.1.2.1 Results

The results of the model, that only considers transition 2 and 3, and assumes different covariate effects but common baseline hazards, are listed in Table 6.3 The assumption of proportional hazards of the two mortality rates does not seem

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
T2:exposed	2.336	10.342	0.026	9.824	10.886	<0.001
T3:exposed	0.768	2.155	0.150	1.607	2.889	<0.001
"Dbts" TRUE	1.839	6.287	0.086	5.311	7.443	<0.001

Table 6.3: *Results from the univariate Cox proportional hazards multi-state model assuming different covariate effects but common baseline hazards.*

to change the results obtained from model I. The variable "Dbts" TRUE indicates whether or not diabetes has already occurred. Thus, the hazard ratio given for this variable expresses the ratio between participants that experience diabetes and those that do not experience diabetes. If the proportional hazard assumption is true, the mortality rate for the participants that experience diabetes is 6 times greater than the participants that do not experience diabetes.

6.1.3 Model III

Model III is based on the same model described above, but now it is assumed that covariate effects are the same for both sets of mortality rates.

In this case, the covariate vector is given as $\mathbf{Z}_2 = (Z, 0)^T$ for $t \leq t_{ds}$, and $\mathbf{Z}_2 = (Z, 1)^T$ for $t > t_{ds}$ with the corresponding regression vector $\beta = (\beta_c, \tilde{\beta})^T$.

It is also of interest to compare mortality rates pre and post diabetes occurrence among exposed and unexposed by assuming proportional transition rates. For

this purpose two additional models are conducted and the corresponding results are listed below.

6.1.3.1 Results

Table 6.4 represents the results obtained by assuming common, global covariate effect for the two transitions. Again, the difference between the transitions are described by an indicator of being in state diabetes. In general the exposed seems to have 9.8 [95 % CI, 9.3-10.3] times higher risk of death with and without developing diabetes compared to unexposed. The occurrence of diabetes increases the risk of death with 3-fold.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
exposed	2.284	9.813	0.026	9.331	10.319	< 0.001
"Dbts" TRUE	1.085	2.960	0.072	2.571	3.409	< 0.001

Table 6.4: *Results from the univariate Cox proportional hazards multi-state model assuming common covariate effects and common baseline hazards.*

The effect of occurrence of diabetes on mortality rate among exposed is shown in Table 6.5. It appears that a childhood cancer survivor with diabetes has 2.6 [95 % CI, 2.1-3.4] times higher risk of death compared to a survivor without diabetes.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
"Dbts" TRUE	0.972	2.644	0.129	2.050	3.410	<0.001

Table 6.5: *Results from the univariate Cox proportional hazards multi-state model assuming common baseline hazards for mortality rate among exposed.*

The effect of occurrence of diabetes on mortality rate among unexposed is shown in Table 6.6.

It is seen that an individual from the general population is 3.7 [95 % CI, 3.1-4.3] times more likely to die with diabetes compared to an individual without diabetes.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
"Dbts" TRUE	1.298	3.660	0.087	3.089	4.338	<0.001

Table 6.6: *Results from the univariate Cox proportional hazards multi-state model assuming common baseline hazards for mortality rate among unexposed.*

6.1.4 Comparison of the models

It is desirable to test whether the two mortality rates, transition 2 and 3 actually are proportional or not, i.e. whether the dependency on time is the same pre and post diabetes occurrence. The usual anova techniques cannot be used, since the applied models are semi-parametric. For these models the likelihoods only are concerned with the regression parameters and the baselines are profiled out. So in order to make a formal likelihood ratio test, the presented models are estimated by Poisson models which are fully parametric. The non-proportionality of rates in Poisson regression is formulated by assuming that the rates are constant in small intervals, but that the magnitude of rates follow some smooth function [13].

The non-proportionality assumption is expressed in a simple generalized Poisson regression model for the number of events in each transition and the natural logarithm of risk times as in the following;

$$log\left(\mu_{ij}\right) = \alpha + \beta \cdot Tr_{ij} + \lambda_j \cdot \left(Groups_{ij}, Tr_{ij}\right) + \gamma_j \cdot \left(ns\left(time_{ij}\right), Tr_{ij}\right) \quad (6.11)$$

for ith subject and jth transition. Here $ns()$ denotes the natural spline of the underlying time. A proportional model is similarly obtained by merely adjusting for time x state interaction.

The asymptotic likelihood ratio test has shown that there is a statistical significant difference between the constructed models with a p-value < 0.001. Hence, the assumption of proportional baseline hazards model does not hold i.e. the dependency on time is not the same pre and post diabetes occurrence. To visualize the shapes of the baseline hazards the mortality rates with and without diabetes is estimated from the Poisson model.

Figure 6.2 shows a plot of rates per 1000 person-years along with the estimated empirical rates. It is observable that the spline functions fit the empirical rates very well, but some uncertainty is obvious for empirical mortality rates with diabetes. The figure indicates that neither proportional hazards assumption

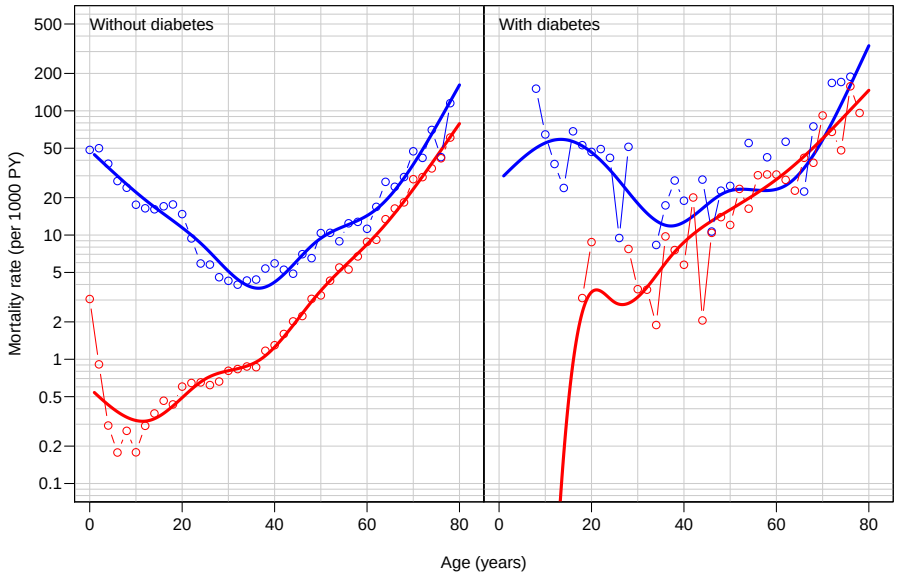

Figure 6.2: *Age-specific mortalities for each cause of death. Blue is exposed and red is unexposed.*

within groups nor between transitions are satisfied. Although the proportionality assumption is not fulfilled, the obtained results can be considered to be interpretable in the global plan.

6.1.5 Discussion

The mortality and morbidity outcomes in childhood cancer survivors are investigated jointly by means of univariate multi-state models so far. It is seen that the results from these outcomes obtained separately and jointly are nearly similar with deviations. As an advantage of multi-state models, the first model is reduced by assuming proportional mortality rates in participants with and without diabetes. This assumption has made it possible to estimate the effect of occurrence of diabetes in the cohort. Although the proportionality assumption is not satisfied, the results are considered to be acceptable as an average effect on the mortality rates. Hence, it is turned out that the occurrence of diabetes increases the risk of death with 3-fold.

6.2 Multivariate analysis

As in the previous chapter, other confounders in addition to Groups are included in the model given in Equation 6.1. The nonlinear form of the covariates age at diagnosis and calendar year are handled by categorizing. The models are stratified by Gender and Country.

6.2.1 Model I - results

The first model is fitted by assuming different baseline hazards with different covariates effects as in the previous section. The results obtained from the fit is listed in Table 6.7. In the model the reference variable for exposed is unexposed, the reference for age at diagnosis is the age group [0-5] and the reference for calendar year is the calendar year [1943-1960].

Comparison of the results obtained by separate and joint model is not possible for transition 2 and 3, since the assumption of piecewise constant hazard for the covariate Groups is not applied to the joint model. But the effect of the covariates on morbidity rates obtained separately cf. Table 5.5 are comparable to the results obtained jointly for transition 1. Some deviations between the results are observable. For instance, in the analysis conducted separately for morbidity rate it is found that a participant diagnosed in [1960-1974] has 35% [95% CI, 14-59 %] increased risk of developing diabetes compared to a participant diagnosed in [1943-1960]. The result from the multi-state model, on the other hand, has shown that the participant is 25% [95% CI, 6-49%] more likely to develop diabetes compared to a participant diagnosed in the reference year, holding other covariates constant.

Similar to the results found in univariate analysis, the childhood cancer survivor in this model has the highest risk of experiencing each transition compared to the general population. They are 12.8 [95 % CI, 11.8-13.9] times more likely to die without experiencing diabetes, and 2.1 [95% CI, 1.7-2.7] times more likely to experience diabetes and 2.4 [95% CI, 1.5-3.7] times more likely to die with diabetes compared to unexposed, holding other covariates constant. The effect of the age groups at diagnosis [5-10] and [10-15] is positive for each transition rate, but is only significant for transition 2 compared to the reference age at diagnosis. Age at diagnosis [15-20], on the other hand, has a negative significant effect on the morbidity rate, a significant positive effect on mortality rate without diabetes and an insignificant positive effect on mortality rate with diabetes, compared to the reference age at diagnosis. A participant with an age in the interval [5-10] has 9% [95% CI, -23-54 %] higher risk of developing diabetes

compared to a participant with an age at diagnosis [0-5]. But the risk does not appear to be significant. The same person has 65% higher risk of death without developing diabetes and 55% higher insignificant risk of death with developing diabetes compared to a reference participant after adjustment for other confounders. The effect of calendar years at diagnosis on transition rates is statistically significant relative to reference interval [1943-1960] except for transition 3. The effect is observed to be positive for transition 1 and 2. Hence, a participant diagnosed in [1974-2010] has 30% [95% CI, 6-60%] increased risk of developing diabetes compared to a participant diagnosed in the reference year,holding other covariates constant.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
T1:exposed	0.752	2.122	0.115	1.695	2.656	< 0.001
T2:exposed	2.550	12.810	0.043	11.773	13.938	< 0.001
T3:exposed	0.856	2.354	0.234	1.489	3.721	< 0.001
T1:Age [5-10]	0.084	1.088	0.178	0.767	1.542	0.637
T2:Age [5-10]	0.500	1.648	0.054	1.483	1.833	< 0.001
T3:Age [5-10]	0.439	1.551	0.543	0.535	4.494	0.419
T1:Age [10-15]	0.121	1.129	0.163	0.820	1.554	0.457
T2:Age [10-15]	0.536	1.710	0.055	1.536	1.902	< 0.001
T3:Age [10-15]	0.677	1.969	0.448	0.818	4.739	0.131
T1:Age [15-20]	-0.371	0.690	0.155	0.510	0.935	0.017
T2:Age [15-20]	0.471	1.602	0.052	1.448	1.773	< 0.001
T3:Age [15-20]	0.783	2.189	0.449	0.908	5.279	0.081
T1:Cal [1960-1974]	0.226	1.254	0.086	1.059	1.485	0.009
T2:Cal [1960-1974]	0.132	1.141	0.056	1.022	1.274	0.019
T3:Cal [1960-1974]	-0.138	0.871	0.217	0.570	1.332	0.524
T1:Cal [1974-2010]	0.265	1.303	0.105	1.060	1.602	0.012
T2:Cal [1974-2010]	0.072	1.074	0.066	0.943	1.224	0.282
T3:Cal [1974-2010]	-0.064	0.938	0.329	0.492	1.788	0.846

Table 6.7: *Results from the multivariate Cox proportional hazards multi-state model assuming different covariate effects and different baseline hazards.*

6.2.2 Model II - results

Model II is set up as before by assuming that transition rates going into death states are proportional. Furthermore, it is assumed that there are different

covariate effects on each transition. The model is constructed as in Equation 6.9 and 6.10 by considering only transition 2 and 3 and including all prognostic factors. Table 6.8 shows the results of Cox proportional hazards multi-state model. By comparing results from model I and model II with regard to transition 2 and 3, it is observed that the significant effects of almost all covariates from the two models differ with a small deviation. In model I the effect of calendar year is insignificant in transition 3, whereas it becomes slightly significant in model II. Note that occurrence of diabetes is also included in the model and here the hazard ratio represents the effect of experiencing diabetes on the rate of occurrence of the endpoint death. A participants developing diabetes has 5.3 times [95% CI, 2.1-13.3] increased risk of death compared to a participant that does not develop diabetes.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
T2:exposed	2.547	12.771	0.043	11.738	13.895	<0.001
T3:exposed	0.884	2.422	0.230	1.544	3.798	<0.001
T2:Age [5-10]	0.499	1.647	0.054	1.481	1.831	<0.001
T3:Age [5-10]	0.429	1.535	0.535	0.538	4.380	0.423
T2:Age [10-15]	0.535	1.707	0.054	1.534	1.899	<0.001
T3:Age [10-15]	0.508	1.662	0.436	0.707	3.907	0.244
T2:Age [15-20]	0.476	1.610	0.052	1.455	1.781	<0.001
T3:Age [15-20]	0.460	1.585	0.427	0.687	3.657	0.281
T2:Cal [1960-1974]	0.104	1.110	0.055	0.996	1.237	0.060
T3:Cal [1960-1974]	0.333	1.395	0.166	1.007	1.932	0.046
T2:Cal [1974-2010]	0.042	1.043	0.065	0.918	1.185	0.520
T3:Cal [1974-2010]	0.587	1.799	0.193	1.233	2.623	0.002
"Dbts" TRUE	1.673	5.329	0.468	2.129	13.336	<0.001

Table 6.8: *Results from the multivariate Cox proportional hazards multi-state model assuming different covariate effects but common baseline hazards.*

6.2.3 Model III - results

After combining baseline hazards of different transitions it may be relevant to joint parameters between transitions. Based on model II it is now assumed that a single effect estimate of the covariates is common for both sets of mortality rates. The results from the model is given in Table 6.9. Using global covariate itself, it is observed that childhood cancer survivors compared to the general

population are associated with increased risk of death. They are 11.6 [95%, CI, 10.6-12.6] times more likely to die, holding other covariates constant. The overall effect of age groups [5-10], [10-15] and [15-20] at diagnosis compared to age group [0-5] on mortality rates is significantly positive meaning that the risk of death for a participant with an age at diagnosis [5-10] is 1.7 times the risk for a participant with an age at diagnosis [0-5] and so forth. Calendar year at diagnosis, on the other hand, appears to have a negative effect on the mortality rates. A participant diagnosed in [1974-2010] has 14% [95%, CI, 3-24%] decreased risk of death compared to a participant diagnosed in [1943-1960] after adjustment of other prognostic factors. The effect of calendar year [1960-1974] on the mortality rate does not differ from the effect of calendar year [1943-1960]. The overall effect of the occurrence of diabetes on mortality rate is 2.9 times higher than if diabetes does not occur, holding all other covariates constant.

	coef	exp(coef)	se(coef)	lower 0.95	upper 0.95	p -value
exposed	2.447	11.555	0.042	10.636	12.553	< 0.001
Age [5-10]	0.506	1.658	0.054	1.493	1.842	< 0.001
Age [10-15]	0.494	1.638	0.054	1.474	1.821	< 0.001
Age [15-20]	0.443	1.558	0.051	1.410	1.722	< 0.001
Cal [1960-1974]	-0.059	0.943	0.051	0.853	1.043	0.255
Cal [1974-2010]	-0.149	0.862	0.063	0.762	0.974	0.017
"Dbts" TRUE	1.090	2.973	0.072	2.580	3.426	< 0.001

Table 6.9: *Results from the multivariate Cox proportional hazards multi-state model assuming common covariate effects and common baseline hazards.*

6.2.4 Comparison of the models

As before in order to test whether the two mortality rates, transition 2 and 3 actually are proportional, Poisson regressions using generalized linear models are fitted. The fit is based on the equation given in 6.11 in which age at diagnosis and calendar year is included. The models are tested by an anova test as previously. Comparison of the models fitted with and without a time x state interaction has shown a significant difference (p-value < 0.001) between the models meaning that the assumption of proportional mortality rates are violated. Although the transitions are time-dependent, the recovered results from model II and model III reflect an averaged effect of covariates on the mortality rates.

6.2.5 Discussion

A multi-state model including all confounders of interest is set up. It is realized that a comparison of morbidity outcome analyzed separately and jointly is possible. The results from the two methods has shown some small deviations. The reduction of multi-state models has made it possible to investigate the effect of occurrence of diabetes on mortality rates. Even though the proportionality assumption made under model reduction process is not satisfied, the results obtained from these models give some general idea about the effect of the covariates and the occurrence of diabetes on mortality rates. It appears from the multivariate Cox proportional hazards multi-state model assuming common covariate effects and common baseline hazards (Table 6.9) that the risk of mortality after occurrence of diabetes is 3 [95% CI,2.6-3.4] times risk of mortality without occurrence of diabetes. This recovered result is also reflected by cumulative incidence plots presented in Chapter 3.

In the same chapter number of exposed that have developed diabetes and their corresponding diagnosis type were also given. It was observed that most of the exposed with diabetes were associated with Central Nervous System (CNS) tumors and Leukemia. This founding is also supported by research which is reported that diabetes may be caused by CNS-involved leukemia [43] and a significant proportion of children with central nervous system (CNS) germ cell tumors (GCTs) was present with diabetes [68][37].

6.3 Prediction of transition probabilities

One of the advantages of multi-state models is the possibility to estimate transition probabilities by means of cumulative risks for each cause of death. The first approach is to determine the age-specific mortality rates and survival function, and then compute the cumulative probabilities of being dead from each of the causes before a given age. The mortality rates are modeled by Poisson regressions including natural splines.

Figure 6.3 shows the different rates along with the corresponding 95% confidence bands in the same frame relative to each other separately for exposed and unexposed. Since none of the participants in their younger age has died after developing diabetes, the confidence bands of mortality rate estimates with diabetes are not estimated well in this interval. It is apparent from the figure that the exposed has a quite higher mortality rates both with and without diabetes compared to unexposed. For both exposed and unexposed the mor-

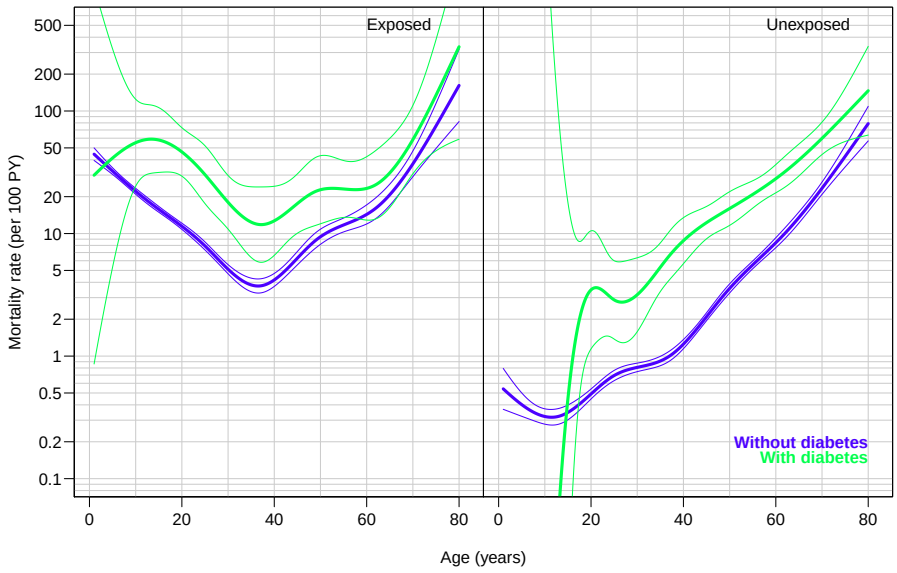

Figure 6.3: *Age-specific mortalities for each cause of death.*

tality rate with diabetes seems to be higher than the mortality rate without diabetes. The mortality rates without diabetes for unexposed is comparable to the mortality rates obtained from the Danish cause of death data analyzed by Bendix Carstensen [12]. The comparison of the rates has shown a strong similarity, which implies that the estimated rates can easily be used to compute the cumulative probabilities.

The cumulative risks or rather the probabilities that a participant ends up dead before or after developing diabetes is estimated by means of the mortality rates. The calculations are based on the theory presented in Chapter 4 for prediction of transition probabilities. The estimated transition probabilities are displayed in Figure 6.4. Note that the probabilities are conditioned on the number of participants that have died in the study. The upper panel in the figure illustrates cumulative risks for each cause for exposed and unexposed, whereas the lower panel shows predictions conditional on survival till age 20. Conditional survival is computed by using mortalities from age 20. It is estimated in order to show that the predicted cause of death patterns are calculated relevantly assuming that they apply throughout life. In all plots the lowest curve represents the survival function for the group in question.

The upper panel reflects that the fraction exposed dead with diabetes is higher

than the fraction exposed died without developing diabetes. A childhood cancer survivor with an age 40 in this study has approx. 55% probability of dying with diabetes and 30% probability of dying without diabetes. A nearly proportional distribution of mortality with and without diabetes is observable. Survivors aged from 0 to 20 has an decreasing survival reflected by a steep decrease in the survival function. The unexposed, on the other hand, shows sensemaking survival. Similar to exposed the fraction unexposed dead with diabetes is higher than the fraction died without diabetes. A 60 years old unexposed in this study is expected to live with 58% probability and die with 35% probability if the person is developed diabetes and the person is expected to die with 7% probability if the person is not developed diabetes.

By studying the lower panel, it is observed that conditional on survival to age 20 a childhood cancer survivor with an age 40 has approx. 58% probability of being alive, 34% probability of dying with diabetes and 8% probability of dying without diabetes.

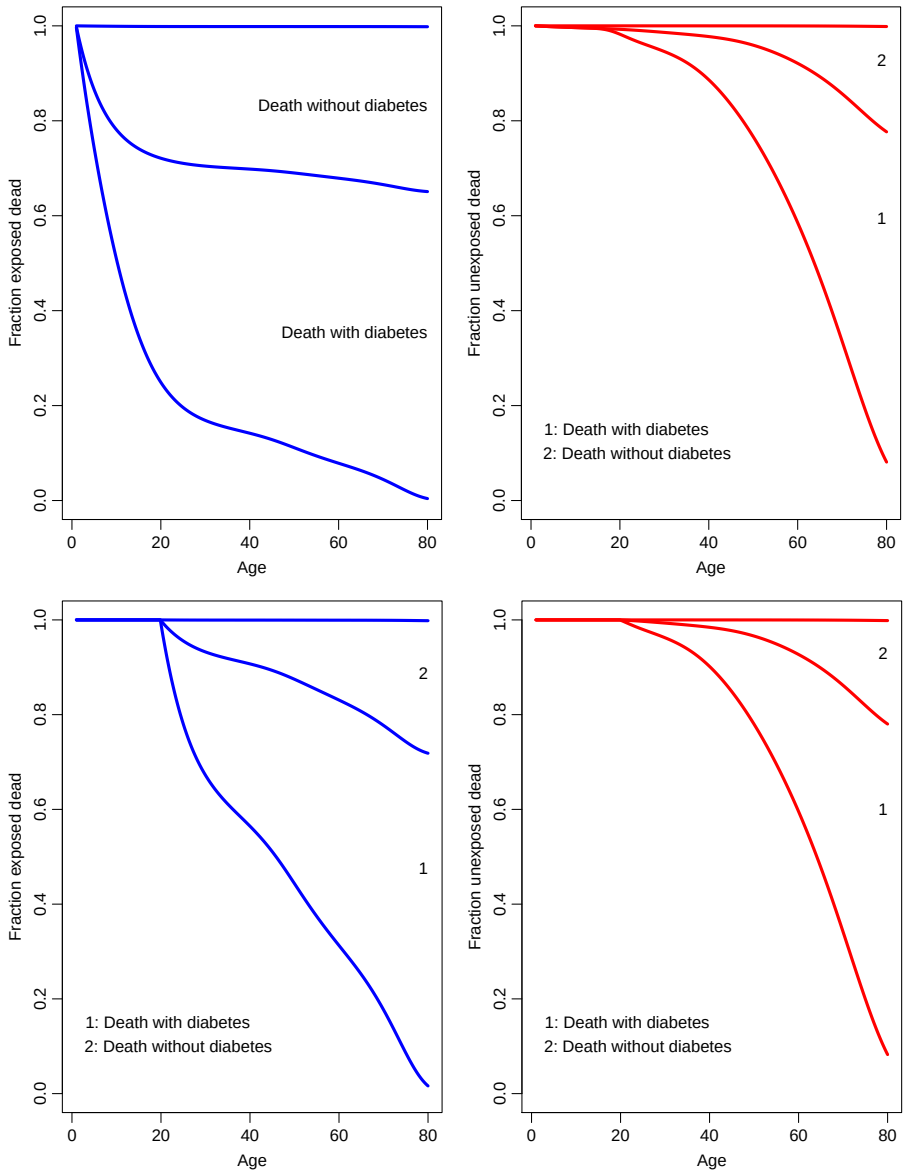

Figure 6.4: *Cumulative risk functions for exposed (blue) and unexposed (red). The lower panel shows predictions conditional on survival till age 20.*

6.4 Conclusion

In order to investigate the morbidity and mortality outcome jointly in childhood cancer cohort compared to the general population an univariate and a multivariate analysis are conducted respectively. In both analyses a Cox proportional hazards multi-state model is set up and reduced to some sub-models.

The univariate analysis has demonstrated some similar results compared to the results from outcomes obtained separately. From the univariate Cox proportional hazard multi-state model assuming different covariate effects and different baseline hazards, it is found that the childhood cancer survivor is 1.7 [95% CI, 1.5-1.8] times more likely to develop diabetes, 10.3 [95% CI, 9.8-10.9] times more likely to die without developing diabetes and 2.2 [95% CI, 1.6-2.9] times more likely to die with developing diabetes compared to the general population. The reduction of the multi-state model has made it possible to investigate the global effect of occurrence of diabetes by assuming that mortality rates in the participants with and without diabetes are proportional. It is turned out that a childhood cancer survivor is 2.6 [95 % CI, 2.1-3.4] times more likely to die if the survivor is developed diabetes than the other way around.

The multivariate analysis of multi-state models has shown that the risk of morbidity and mortality in childhood cancer survivors compared to the general population is quite similar to the results obtained from the univariate analysis. Furthermore, from the multivariate Cox proportional hazard multi-state model assuming different covariate effects and different baseline hazards, it is found that the age groups at diagnosis has a positive effect on each transition rate compared to age at diagnosis [0-5] except the age group [15-20] which has a negative effect on experiencing diabetes relative to age at diagnosis [0-5], holding other covariates constant. The effect of age groups at diagnosis compared to the reference group is statistically significant in explaining the transition rates for all transition except transition 3. In the same way, the effect of calendar year at diagnosis is observed to be positive and statistically significant compared to the reference year for all transition except for transition 3. After reducing the full model it is found from the multivariate Cox proportional hazards multi-state model assuming common covariate effects and common baseline hazards that the risk of mortality after occurrence of diabetes is 2.9 [95% CI,2.6-3.4] times the risk of mortality without occurrence of diabetes. It is, though, a global effect on the mortality rate.

The estimation of the transition probabilities is computed conditional on the number of participants that have died in the study. It is observed that both fraction exposed and unexposed dead with diabetes is higher than the fraction died without developing diabetes. For both groups a nearly proportional distri-

bution of cause of death is apparent.

Chapter 7

Conclusion and Discussion

7.1 Conclusion

The morbidity and mortality outcomes in the nordic childhood cancer cohort compared to the general population cohort is analyzed using two-state survival models as well as more advanced and sophisticated multi-state models. Both analyses are based on Cox Proportional hazards model.

The results from the univariate multi-state model with the aim to investigate the effect of the groups on different transition rates are similar to the results from the statistical standard analyses conducted separately, with small deviations. Both analyses have revealed that the childhood cancer survivors are associated with higher risk of experiencing both mortality and morbidity outcome when compared to the general population. The risk of mortality without developing diabetes, however, was highest among other outcome. The univariate Cox proportional hazards multi-state model assuming common baseline hazards for mortality rate among exposed has shown that a childhood cancer survivor with diabetes was 2.6 [95% CI, 2.1-3.4] times more likely to die compared to a survivor without diabetes.

The multivariate analysis conducted separately and jointly was only comparable with regard to the morbidity outcome. The comparison of the two approaches

has demonstrated similar results with small deviations. As in the univariate analysis, it is found that the risk of mortality and morbidity in childhood cancer cohort is notably higher compared to the general population cohort, after adjustment for other confounders. It is appeared that a participant that was diagnosed in his/her older age, i.e. in the interval [15-20] had 30% [95 % CI, 10-49%] decreased risk of developing diabetes compared to a participant with an age at diagnosis [0-5], holding other covariates constant. Apart from that, age at diagnosis seemed to have a positive statistically significant effect on morbidity rate and on mortality rate without developing diabetes compared to the reference age group. Furthermore, it is observed that neither age at diagnosis nor calendar year at diagnosis had a statistical significant effect on the risk of mortality after developing diabetes. It has become apparent that the effect of calendar year at diagnosis was positive on experiencing diabetes and on mortality rates without developing diabetes when compared to the reference calendar year.

By constructing multi-state models it has been possible to investigate the effect of occurrence of diabetes on mortality rates assuming proportional rates pre and post diabetes. Although the assumption was not satisfied, it is concluded that the results might be acceptable in explaining the average effect on the mortality rates. Hence, it is found from the multivariate Cox proportional hazards multistate model assuming common covariate effects and common baseline hazards that the effect of the occurrence of diabetes was 3 [95% CI, 2.6-3.4] times the effect of non-occurrence of diabetes on the mortality rates, holding other covariates constant. This result is supported partly by the cumulative incidence functions presented in Chapter 3 and partly by the prediction of the transition probabilities depicted in Figure 6.4.

Overall it can be concluded that the analysis of morbidity and mortality outcomes conducted jointly yields both similar and additional results when compared to the traditional analysis.

7.2 Discussion

It is an important public health issue to be able to quantify the elevated risk of late effects due to its treatment. Several statistical techniques can be utilized for this purpose. In this thesis the morbidity and mortality outcomes in childhood cancer survivors are analyzed both separately and jointly using multi-state models.

In the first part of the analysis two-state Cox proportional hazards models are constructed for morbidity and mortality outcomes separately in order to see how prognostic factors influence different phases of the illness/death process. Before applying the results of the models some of the important issues such as the assessment of the underlying model assumptions i.e. the functional form of the covariates and the assumption of proportional hazards are considered. The violation of linear functional form of the covariates and proportional hazards assumption are handled by categorizing non-linear covariates and assuming piecewise hazards, respectively. It might be noticed that since incorrect functional forms can appear as non-proportional hazards, the functional forms of the covariates are corrected before non-proportional hazards are diagnosed.

There are various approaches to deal with the violation of these assumptions. The non-linear form of a covariate could be incorporated into a Cox model by spline fits and be tested by Wald test in order to find out if the non-linear effect could remain in the specification [29]. To deal with violation of proportional hazards assumption one could choose to stratify the covariate that does not satisfy the proportional hazards assumption. But since it is not possible to examine the effect of the stratified covariate, stratification would not be a good option for the covariates that are of interest to be analyzed. An other approach could be to include log-time interaction with the covariate that violates the proportional hazards assumption.

In the second part of the analysis the multi-state models are set up based on the correction of nonlinear functional forms of the covariates diagnosed in the first part. These models are constructed in order to analyze different covariate effects on several transitions simultaneously and investigate if there is common covariate effect between transitions and if hazard rates going into same state (death) can assumed to be proportional. These models are also used for calculating transition probabilities and for investigating if the occurrence of a late effect changes the risk of the event of interest death to occur.

Both part of the analysis has shown some similar results. Both approaches have verified that the childhood cancer survivors were associated with increased risk of developing diabetes and dying with this late effect when compared to the

general population. In addition to the standard analyses, multi-state models have revealed that the occurrence of diabetes increased the risk of mortality. It is shown that the proportionality assumption of the mortality rates is not met, but the results are accepted to be interpretable as being a global effect on the mortality rates. The multi-state models based on the assumption of proportional baseline hazards is presented in order to demonstrate two advantages: one hazard can be estimated instead of two and the risk of occurrence of diabetes on mortality rate can be determined. The disadvantage of this approach is that it is still an assumption, and it does not always hold.

Given that the proportionality assumption is not made, one could still combine the covariate effects across transitions and estimate a global effect of the covariate. At this point it is up to the researcher if he/she will make use of the estimated global effect or the estimated transition-specific effects. This decision, however, depends partly on how many direct transitions and paths are between states and how many covariate effects are desired to be estimated. Modeling the effect of covariates for each transition separately leads to a large number of regression coefficients to be estimated and this can cause over-fitting, especially when transitions with few events are present. The problem can be solved by assuming equal covariate effects transitions or assuming zero-covariate effect for the event-poor transition [33]. In addition to that, in order to deal with the abundance of regression parameter to be estimated, the use of the reduced-rank techniques introduced in the paper [33] can be applied.

Multi-state models can be constructed by means of Cox proportional hazards models or Poisson regressions. There is not a difference between these two approaches. Poisson models enable smoothing of the effects of time-scales using standard regression tools and enable modeling of the interactions between timescales and other covariates. Cox proportional hazards models, on the other hand, are useful for clinical follow-up studies in which there is only one relevant timescale and the focus is on the effect of covariates than time. The primary interest in these studies is the survival function rather than the baseline hazard [10]. Cox models can be more desirable for multi-state models since the semi-parametric model specification allows flexible covariate structures such as frailties, time-dependent and transition-specific covariates.

Multi-state models can be useful tools in understanding and describing the disease progression and for prediction purposes. They allow for simultaneous analyses of several transitions and in this way a relative interpretation of hazards. They are flexible in varying, restricting covariate effects across transitions and combining baseline hazards of different transitions. Not only models analysed in Chapter 6, but also any combination of common and transition-specific covariates in stratified or proportional baseline hazards can be modeled. If one is interested in the effect of the duration in a state on hazard, this can merely

be included in the model. It is also possible to leave out covariates for one transition and include them for another transition [16]. But these possibilities make multi-state models complicated to deal with and the analysis of all possible combinations of model constructions is time-consuming.

Although an increasing interest in multi-state models is considered in the recent years, theoretical study and application of multi-state models has been limited to statistical journals. The reason for this limitation has been due to the complexity of the models and lack of good software [41]. However, a number of software packages have been developed in R for the analysis of multi-state models in the recent years. The principle in these softwares is to make an appropriate data set representing each individual by several observations. The softwares have all some limitations in practice as listed in Chapter 4. Since this study is dealt with a data set in which the event times are measured exactly i.e. all transitions are observed, and the main interest is to construct semi-parametric models, the most appreciated packages for analysis purposes are `Epi` and `mstate`. In this study both of the packages are used for constructing multi-state models for which similar results are obtained. But only results from Epi package is presented, since the package is considered as being user-friendly and more flexible. For the data that are interval censored, the package `msm` is recommended [23].

Multi-state models have many extensions. The standards are two-state models, competing risks, disability, bivariate and recurrent events models. These are used to model multivariate and multiple survival data. Furthermore, multi-state models consider all data as being longitudinal and therefore they are less useful for repeated observations [25].

7.3 Future work

In this thesis multi-state models including only one late effect in childhood cancer survivors are constructed. A broad range of other outcome information is available in the data obtained from ALiCCS study. Analysing these outcomes separately can lead to lost of information, since late effects are often correlated. It would be of interest to study the different late effects jointly by means of multi-state models. In this way it would be possible to quantify bias of effect estimates of covariates for childhood cancer when using traditional statistical methods and describe how late effects of treatment for childhood cancer are interrelated. An illustration of this approach is depicted in Figure 7.1.

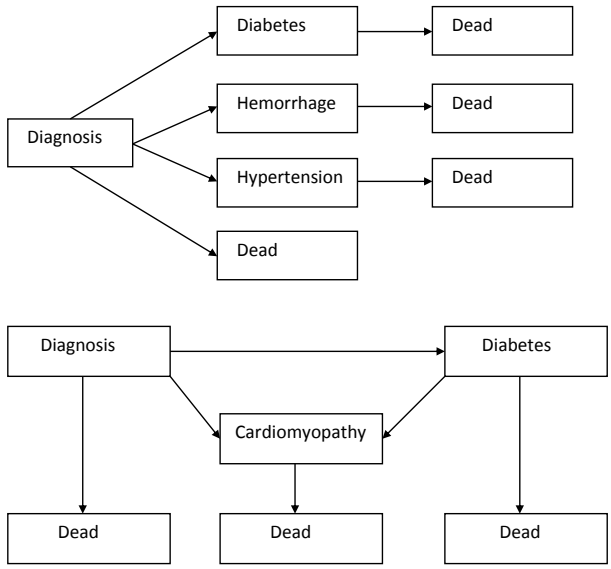

Figure 7.1: *Model blocks representing competing risk and multi-state models respectively.*

The first block demonstrates a competing risk model where the participants can experience different late effects independently and may die with or without a late effect. The second block, on the other hand, illustrates a multi-state model in which a participant can die with more than one late effect.

The analysis conducted in this thesis is restricted to uni-directional multi-state models for which recurrent events are not possible. However, in an illness/death model the reversibility, the transition back into the same state is possible. In this case disregarding recurrent events may yield to biases results in explaining the variety event rate of late effects, especially since late effects are correlated. A further work in analysing late effects could be to allow recurrent events in multi-state models so that the effect of first and recurrent morbidity could be estimated. A simple example of this is shown in Figure 7.2.

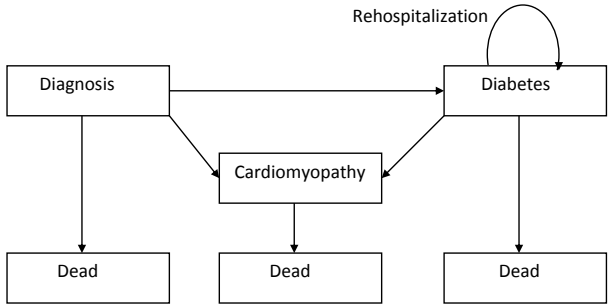

Figure 7.2: *Model block representing an illness-death model with recurrent events.*

An additional information given in the data set is the type of diagnosis for each childhood cancer survivor. It would be of interest to investigate the effect of different treatments for childhood cancer on late effects. Some other factors that may influence the occurrence of late effects, for example life style, health status, occupation, smoking and genetic risk factors is not taken into account in this study. It could be interesting to assess if these have an influence on the risk estimates of late effects and to quantify the precision of risk estimates of these.

Appendix A

Definitions

C

Cardiomyopathy: which means "heart muscle disease", is the deterioration of the function of the myocardium (i.e., the actual heart muscle) for any reason [51].

Cardiovascular and pulmonary diseases: are defined as any disorder that affects the heart or lungs' ability to function normally. There are a variety of different diseases and conditions which fit this description including endocarditis, heart attack, heart failure, chronic obstructive pulmonary disease and pulmonary stenosis [61].

Case-control studies: A case-control study is an analytical study which compares individuals who have a specific disease ("cases") with a group of individuals without the disease ("controls"). The proportion of each group having a history of a particular exposure or characteristic of interest is then compared. An association between the hypothesized exposure and the disease being studied will be reflected in a greater proportion of the cases being exposed. It is advantageous for the controls to come from the same population from which the cases were derived, to reduce the chance that some other difference between the groups is accounting for the difference in the exposure that is under investigation [52].

Censoring: occurs when some lifetimes are known to have have occurred only within certain intervals [30].

Competing events: refers to a situation where an individual is exposed to two or more causes of failure, and its eventual failure can be attributed exactly to only one [44].

Confounding: In statistics, a confounding variable (also confounding factor, hidden variable, lurking variable, a confound, or confounder) is an extraneous variable in a statistical model that correlates (positively or negatively) with both the dependent variable and the independent variable [55].

Cohort: In statistics and demography, a cohort is a group of subjects who have shared a particular time together during a particular time span. Cohorts may be tracked over extended periods in a cohort study [54].

Cox regression: Cox regression (or proportional hazards regression) is method for investigating the effect of several variables upon the time a specified event takes to happen. In the context of an outcome such as death this is known as Cox regression for survival analysis [56].

D

delayed entry (left truncation): occurs when subjects enter a study at a particular age (not necessarily the origin for the event of interest) and are followed from this delayed entry time until the event occurs or until the subject is censored [30].

Diabetes: is a group of metabolic diseases in which a person has high blood sugar, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced [57].

E

Epidemiology: is the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems [58].

G

Goodness-of-fit: The goodness of fit of a statistical model describes how well

it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question [59].

H

Hazard function: is a measure of the tendency to fail; the greater the value of the hazard function, the greater the probability of impending failure [60].

Hemorrhage: is the loss of blood or blood escape from the circulatory system [50].

Hypertension: Hypertension (HTN) or high blood pressure, sometimes arterial hypertension, is a chronic medical condition in which the blood pressure in the arteries is elevated [62].

L

Left censoring: the event of interest has already occurred for the individual before that person is observed in the study at the left censoring time [30].

Lexis diagram: is a two dimensional diagram that is used to represent events (such as births or deaths) that occur to individuals belonging to different cohorts [72].

Likelihood ratio test: In statistics, a likelihood ratio test is a statistical test used to compare the fit of two models, one of which (the null model) is a special case of the other (the alternative model) [64].

M

Metabolic syndrome: is a combination of medical disorders that, when occurring together, increase the risk of developing cardiovascular disease and diabetes [73].

R

Relative risk: In statistics and mathematical epidemiology, relative risk (RR) is the risk of an event (or of developing a disease) relative to exposure. Relative risk is a ratio of the probability of the event occurring in the exposed group versus a non-exposed group [66].

Right censoring: an event is observed only if it occurs prior to some prespecified time [30].

S

Semi-parametric model: in statistics a semiparametric model is a model that has parametric and nonparametric components [66].

Standardized mortality ratio: in epidemiology is the ratio of observed deaths to expected deaths, where expected deaths are calculated for a typical area with the same age and gender mix by looking at the death rates for different ages and genders in the larger population [74].

Survival analysis: Survival analysis is just another name for time to event analysis. The term survival analysis is used predominately in biomedical sciences where the interest is in observing time to death either of patients or of laboratory animals [49].

Survival function: is a property of any random variable that maps a set of events, usually associated with mortality or failure of some system, onto time. It captures the probability that the system will survive beyond a specified time [67].

Appendix B

Supplementary figures and tests

B.1 Cumulative incidence

In this appendix some supplementary figures and tests are presented. The crude estimates of cumulative incidence of mortality and morbidity with regard to country and gender is displayed in Figure B.1 and B.2, respectively.

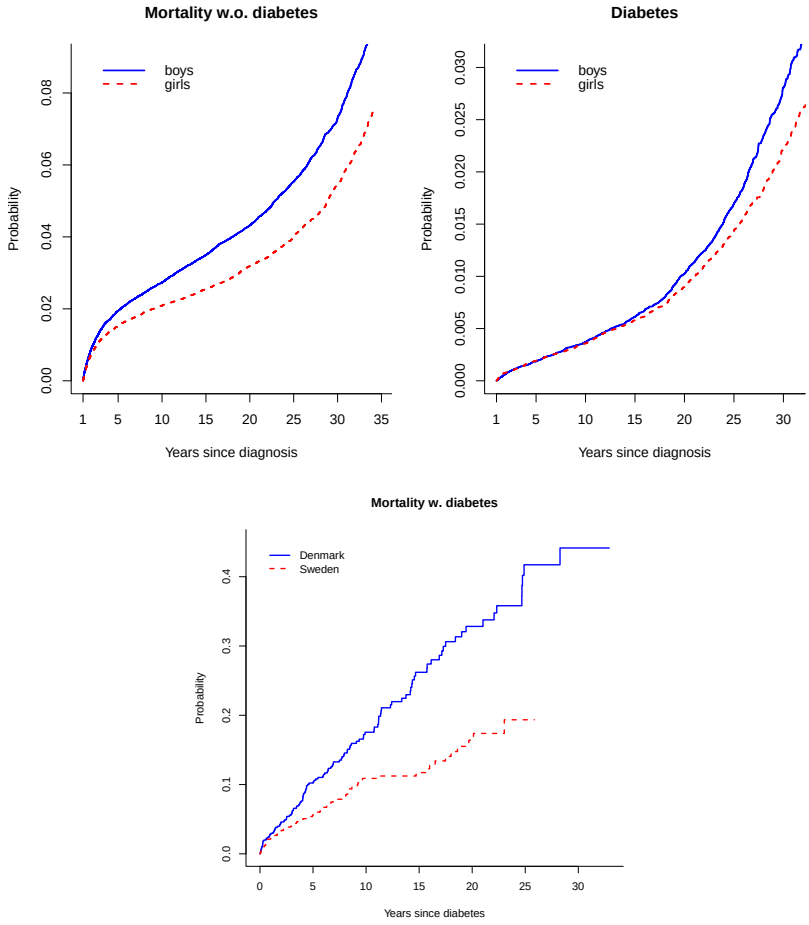

Figure B.1: *Cumulative incidence of mortality and morbidity with regard to country.*

B.2 Tests for two-state models

B.2.1 Morbidity rate

The functional form of the age at diagnosis and calendar year for morbidity rate analysis is diagnosed by using the method of smoothing. Figure B.3 shows the nonlinear effect of the covariates.

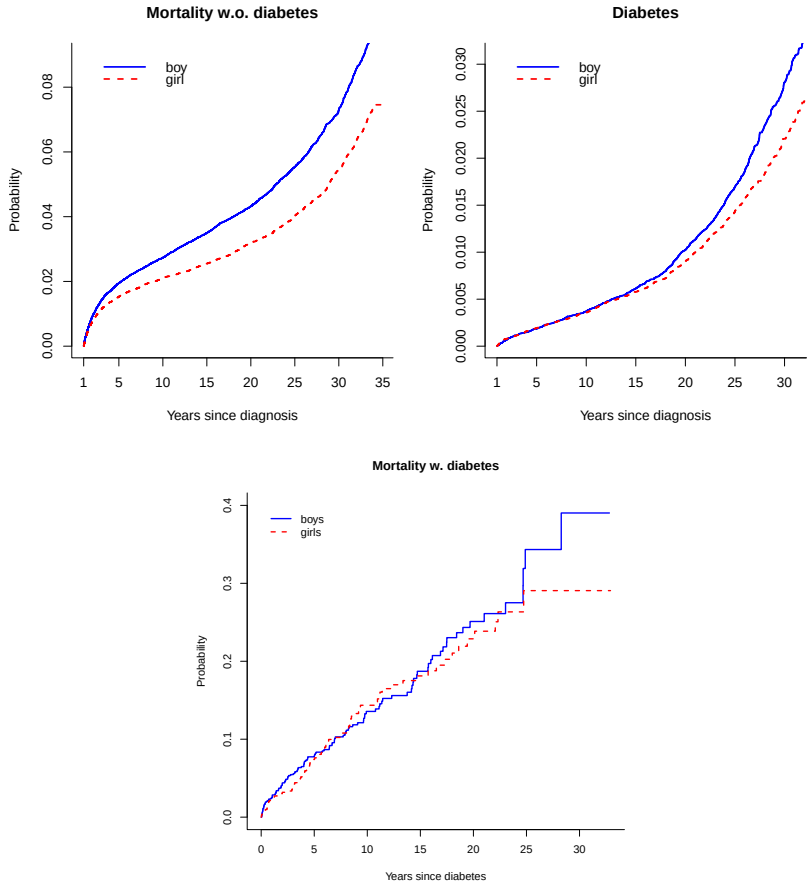

Figure B.2: *Cumulative incidence of mortality and morbidity with regard to gender.*

A test for the proportional hazards assumption of the final model for morbidity rate is gathered in Table B.1.

B.2.2 Mortality rate with diabetes

The nonlinear functional form of the covariates: age at diagnosis and calendar in the final two-state analysis is diagnosed and fitted by spline functions as seen in Figure B.4.

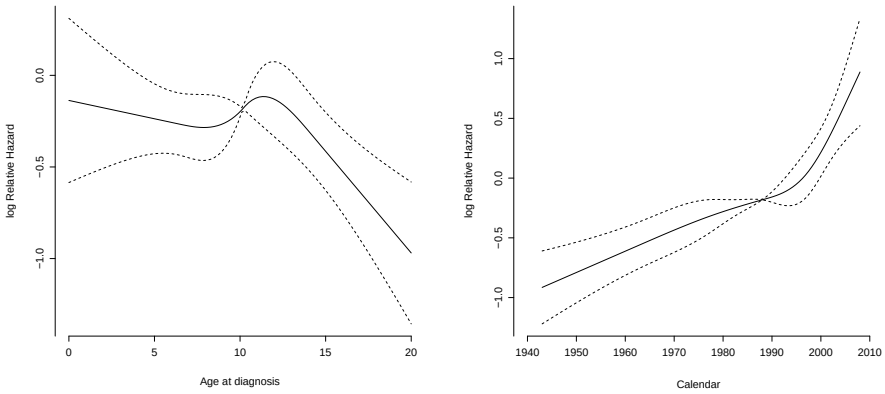

Figure B.3: *Functional form of age at diagnosis (left) and calendar year at diagnosis (right) on log hazard of morbidity. In both plots the thick lines represent the spline fit while the dashed lines represent 95% confidence bands for the fit.*

	ρ	χ^2	p -value
exposed	-0.02526	1.17575	0.2782
Age [5-10]	-0.02057	0.78898	0.3744
Age [10-15]	-0.02910	1.62180	0.2028
Age [15-20]	-0.04472	3.81239	0.0509
Calendar [1960-1974]	0.00131	0.00329	0.9543
Calendar [1974-2010]	0.01063	0.22025	0.6389

Table B.1: *Test results of proportional hazards assumptions.*

To assess violations of the proportional hazard assumption of the model analyzing mortality rate with diabates, it is assumed that the variable Groups is a piecewise time-varying covariate. The hazard ratios of exposed based on the extended Cox model is estimated and is shown in Figure B.5.

The test result of the proportional hazards assumption of the final model is displayed in Table B.2.

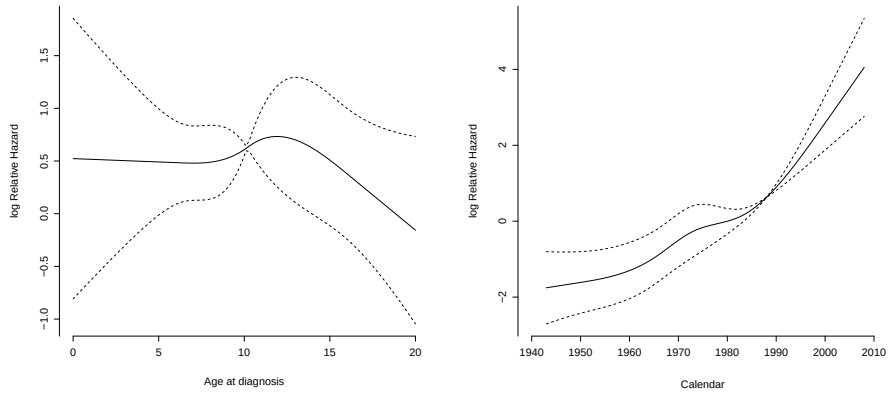

Figure B.4: *Functional form of age at diagnosis (left) and calendar year at diagnosis (right) on log hazard of death with diabetes.*

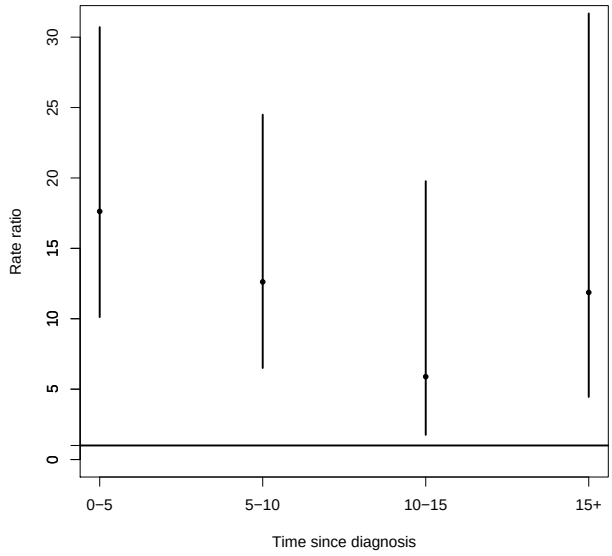

Figure B.5: *Rate ratios and the corresponding* 95% *confidence intervals for Groups with non-exposed as reference in each interval.*

	ρ	χ^2	p -value
Age [5-10]	0.0360	0.2569	0.6123
Age [10-15]	-0.0248	0.1292	0.7193
Age [15-20]	0.0112	0.0286	0.8658
Calendar [1960-1974]	-0.0126	0.0350	0.8515
Calendar [1974-2010]	-0.0576	0.6808	0.4093
episode [1-5]	-0.1497	4.7882	0.0287
episode [5-10]	-0.0745	1.1577	0.2819
episode [10-15]	-0.0708	0.9360	0.3333
episode 15+	-0.0606	0.7865	0.3752

Table B.2: *Test results of proportional hazards assumption.*

Appendix C

R programming

C.1 Preparation of data

```
# Data from Denmark
data<-read.csv("//filserv.cancer.dk/EPI/EPI-Brugere/kadriye/Thesis/R/
Datasets/data.csv", header=TRUE,sep = ";")

Groups<- rep(0,length(I))
Groups<- replace(data$kontrol_nr, data$kontrol_nr == 0, "exposed")
for(i in 1:5 ){
Groups<- replace(Groups, data$kontrol_nr == i, "unexposed")
}
data$Groups<-Groups

### Defining sex
data$sex<-replace(data$sex, data$sex==1, "boy")
data$sex<-replace(data$sex, data$sex==2, "girl")

### Defining status for event of interest: death
data$status<-replace(data$status, data$status!=90, 0)
data$status<-replace(data$status, data$status==90, 1)

### Defining the relevant dates
```

```
# first insert the missing values in the given dates
# for birth date
data$fsdato<-ifelse(nchar(data$fsdato)==7,
paste(0,data$fsdato, sep=""),
 data$fsdato )
# for diagnosis date
data$diagdato<-ifelse(nchar(data$diagdato)==7,
paste(0,data$diagdato, sep=""),
 data$diagdato )
# for status date
data$statdato<-ifelse(nchar(data$statdato)==7,
paste(0,data$statdato, sep=""),
 data$statdato )

# define date of diabetes registration
# note that a subject has a diabetes date if the person experiences
# diabetes, otherwise the date of diabetes registration is set to
# date of status.
data$diadato<-data$statdato
data$diadato[!is.na(data$dmA)]<-data$dmA[!is.na(data$dmA)]
data$diadato<-ifelse(nchar(data$diadato)==7,
paste(0,data$diadato, sep=""),data$diadato )

data$BirthDate<-as.Date(as.date(data$fsdato, order= "dmy"))
data$DiagDate<-as.Date(as.date(data$diagdato, order= "dmy"))
data$DiaDate<-as.Date(as.date(data$diadato, order= "dmy"))
data$ExitDate<-as.Date(as.date(data$statdato, order= "dmy"))

# Entry date
data$EntryDate<-c(rep(0,nrow(data)))
# Entry into the study happens one year after diagnosis.
data$EntryDate<-as.Date(cal.yr(data$DiagDate) +1)

# Entry date of the subjects that have a diagnosis date before 1977
cohortstart<-as.Date("01Jan1977", "%d%b%Y")
Agebefore<-which(data$EntryDate < cohortstart)
data$EntryDate[Agebefore]<-cohortstart

### Defining status variable for event: diabetes
data$EventDia<-rep(0,length(data$dmA))
data$EventDia[data$dmA!= "NA"]<-1

### Defining the relevant ages
data$AgeDiagnosis<-cal.yr(data$DiagDate)-cal.yr(data$BirthDate)
# set negative age values to zero.
data$AgeDiagnosis[which(data$AgeDiagnosis<0)]<-0
```

```
data$AgeExit<-cal.yr(data$ExitDate)-cal.yr(data$BirthDate)
agebline<- cal.yr(cohortstart)-cal.yr(data$BirthDate[Agebefore])
# Entry into the study happens one year after diagnosis.
data$AgeEntry<-data$AgeDiagnosis+1
data$AgeEntry[Agebefore]<-agebline
# age when experiencing diabetes
data$Dia.Age<- cal.yr(data$DiaDate)-cal.yr(data$BirthDate)

### time until experiencing diabetes
data$Dia.time<-data$Dia.Age-data$AgeEntry

### duration of the study
data$Dur<-data$AgeExit-data$AgeEntry

### Calendar year at diagnosis
data$Calendar<-substr(data$diagdato,5,8)

### Country
data$country<-rep("Denmark", nrow(data))

#### Exclusion ####
which(data$BirthDate>data$DiagDate)
# 196 subjects are diagnosed before they are born.
# Do not remove them. Set AgeEntry to 0.

which(data$BirthDate>data$DiaDate)
which(data$BirthDate>data$ExitDate)
rm1<-which(data$DiagDate>data$DiaDate)
data<-data[-rm1,]
which(data$DiagDate>data$ExitDate)
rm2<-which(data$EntryDate >data$ExitDate)
data<-data[-rm2,]
rm3<-which(data$EntryDate >data$DiaDate)
data<-data[-rm3,]
which(data$AgeDiagnosis<0)
which(data$AgeExit<0)
which(data$AgeEntry<0)
which(data$Dia.Age<0)
rm5<-which(data$Dur<0)
data<-data[-rm5,]
which(data$Dia.time<0)

# Find the average age of exposed at diagnosis
# and define it as age of unexposed at diagnosis.
avgAge<-mean(data$AgeDiagnosis[data$Groups == "exposed"],na.rm=T)
data$AgeDiagnosis[data$Groups == "unexposed"]<-avgAge
# year at diabetes registration
```

```
data$CalD<-substr(data$diadato,5,8)

### New Data Frame  ###
Data<-data.frame(Id =data$pnr, Groups=data$Groups, Gender = data$sex,
AgeDiagnosis = data$AgeDiagnosis, AgeEntry = data$AgeEntry,
Age.Dbts = data$Dia.Age,AgeExit= data$AgeExit,Ca.time = data$Dur,
D.Event = data$status, Dbts.time=data$Dia.time, Dbts.Event =
data$EventDia,Calendar = data$Calendar, Country = data$country,
Diagnosis = data$iccc, CalDia=data$CalD)
head(Data)
which(is.na(Data))
# no missing values
# Define Diagnosis types for controls as control.
Data$Diagnosis<-ifelse(Data$Groups =='unexposed',NA,Data$Diagnosis)
Data$Diagnosis <- as.factor(Data$Diagnosis)
levels(Data$Diagnosis)<-c("Leukemia","Lymphomas","CNS","Symp. NST",
"Retinoblastoma","Renal tumors", "Hepatic tumors","Malignant bone
 tumors", "Soft tissue sarcomas", "Germ cell and other","Carcinomas",
 "Other and unspecified")

Data$Diagnosis <- ifelse(is.na(Data$Diagnosis),"unexposed",
as.character(Data$Diagnosis))
Data$Diagnosis <- as.factor(Data$Diagnosis)
Data$Diagnosis <- relevel(Data$Diagnosis,ref="unexposed")

# Saving data set from Denmark
write.table(Data, "//filserv.cancer.dk/EPI/EPI-Brugere/kadriye/Thesis/R
/Datasets/Denmark.csv", sep = ";",dec=",",row.names=FALSE )

#########################################################################
# Data from Sweden is prepared in the same way and the final data set
# is merged.
```

C.2 Descriptive analysis

```
library(mstate);library(Epi);library(survival); library(xtable);
library(cmprsk)
#Categorizing age at diagnosis and calendar
Ndata$AgeDiag <- cut(Ndata$AgeDiagnosis,c(0,5,10,15,21),
c("0-5","5-10","10-15","15-20"))
Ndata$Cal <- cut(Ndata$Calendar,c(0,1960,1975,2012),
c("1943-1960","1960-1974","1974-2010"))
# distribution according groups
gr<-table(Ndata$Groups)
```

```
prop.table(gr)*100

# distribution according country
country<-table(Ndata$Country,Ndata$Groups)
prop.table(country,2)*100
prop.table(table(Ndata$Country))*100
chisq.test(country)
# p-value = 0.6007

# distribution according to gender
gen<-table(Ndata$Gender,Ndata$Groups)
prop.table(gen,2)*100
prop.table(table(Ndata$Gender))*100
chisq.test(gen)
# p-value = 0.06463

# distribution according to age at diagnosis
age<-table(Ndata$AgeDiag, Ndata$Groups)
prop.table(age,2)*100
prop.table(table(Ndata$AgeDiag))*100
chisq.test(age[,1])
# p-value < 2.2e-16

# as continuous variable:
agec<-table(Ndata$Groups, Ndata$AgeDiagnosis)
exp<-Ndata[Ndata$Groups=="exposed",]
mean(exp$AgeDiagnosis);sd(exp$AgeDiagnosis);
unexp<-Ndata[Ndata$Groups=="unexposed",]
mean(unexp$AgeDiagnosis);sd(unexp$AgeDiagnosis);
mean(Ndata$AgeDiagnosis);sd(Ndata$AgeDiagnosis);

# distribution according to calendar year
cal<-table(Ndata$Cal, Ndata$Groups)
prop.table(cal,2)*100
prop.table(table(Ndata$Cal))*100
chisq.test(cal)
# p-value < 2.2e-16
# as continuous variable:
mean(exp$Calendar);sd(exp$Calendar);
mean(unexp$Calendar);sd(unexp$Calendar);
mean(Ndata$Calendar);sd(Ndata$Calendar);

## additional plots
# number of exposed/unexposed in both countries
Dk<-Ndata[Ndata$Country=="Denmark",]
ca<-table(Dk$Calendar,Dk$Groups)
Se<-Ndata[Ndata$Country=="Sweden",]
```

```
caS<-table(Se$Calendar,Se$Groups)
x<-seq(1943,2008,1)
xx<-seq(1958,2008,1)
par(mfrow=c(2,1))
par(mar=c(4,4,1,3)+0.1)
plot(x,ca[,1],type="l",lty=5,col=2,lwd=2,ylim=c(0,380),xlab="Calendar",
 ylab = "number of exposed")
lines(xx,caS[,1],lty = 6,lwd=2, col = 4)
legend("topleft",c("Denmark", "Sweden"), lty = c(5,6),lwd=2,
 col = c(2,4))
par(mar=c(4,4,1,3)+0.1)
plot(x,ca[,2],type="l",lty=5,lwd=2,col=2,ylim=c(0,1980),xlab="Calendar",
 ylab = "number of unexposed")
lines(xx,caS[,2],lty = 6,lwd=2, col = 4)
legend("topleft",c("Denmark", "Sweden"), lty = c(5,6),lwd=2,
 col = c(2,4))

# number of event of interest: diabetes in both countries
# DK
new<-Ndata[Ndata$Dbts.Event=="1",]
dk<-new[new$Country=="Denmark",]
se<-new[new$Country=="Sweden",]
dbdk<-table(dk$CalD,dk$Groups)
dbse<-table(se$CalD,se$Groups)
x<-seq(1977,2010,1)
xx<-seq(1984,2009,1)
par(mfrow=c(2,1))
par(mar=c(4,4,1,3)+0.1)
plot(x,dbdk[,1],type="l",lty=5,col=2,ylim=c(0,25),lwd=2,xlab="Calendar",
 ylab = "events among exposed")
lines(xx,dbse[,1],lty = 6,lwd=2, col = 4)
legend("topleft",c("Denmark", "Sweden"), lty = c(5,6), lwd=2,
col = c(2,4))
par(mar=c(4,4,1,3)+0.1)
plot(x,dbdk[,2],type="l",lty=5,col=2,lwd=2,ylim=c(0,83),xlab="Calendar",
 ylab = "events among unexposed")
lines(xx,dbse[,2],lty = 6,lwd=2, col = 4)
legend("topleft",c("Denmark", "Sweden"), lty = c(5,6),lwd=2,
col = c(2,4))
######################### Transition 1  ###################################
# Mortality without diabetes
Ndata$Can.Event<-Ndata$D.Event
Ndata$Can.Event[which(Ndata$Dbts.Event==1)]<-0

######### Crude Estimates ##############
p1<-pyears(formula = Surv(AgeEntry,Age.Dbts,Can.Event)~Groups,
data = Ndata, data.frame= TRUE,scale=1)$data
```

```
p1$rate<-1000*p1$event/p1$pyears
########## Cumulative Incidence function ###########
ci<-cuminc(Ndata$Dbts.time, Ndata$Can.Event, Ndata$Groups, rho=0,
 cencode=0)
par(las=1)
plot(ci,ylim=c(0,0.3),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),
curvlab=c('exposed','unexposed'),xaxt="n")
axis(1,c(0,4,9,14,19,24,29,34),c("1","5","10","15","20","25","30","35"))
title("Mortality w.o. diabetes")

## Cumulative incidence curves regarding to country
ci1<-cuminc(Ndata$Dbts.time, Ndata$Can.Event, Ndata$Country, rho=0,
 cencode=0)
par(las=1)
plot(ci1,ylim=c(0,0.13),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),curvlab=c('Denmark',
'Sweden'),xaxt="n")
axis(1,c(0,4,9,14,19,24,29,34),c("1","5","10","15","20","25","30","35"))
title("Country")

## Cumulative incidence curves regarding to Gender
ci3<-cuminc(Ndata$Dbts.time, Ndata$Can.Event, Ndata$Gender, rho=0,
 cencode=0)
par(las=1)
plot(ci3,ylim=c(0,0.09),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),curvlab=c('boys',
'girls'),xaxt="n")
axis(1,c(0,4,9,14,19,24,29,34),c("1","5","10","15","20","25","30","35"))
title("Gender")
######################### Transition 2  ###################################
% Morbidity outcome
######### Crude Estimate ###############
p1T2<-pyears(formula = Surv(AgeEntry,Age.Dbts,Dbts.Event)~Groups,
 data = Ndata,
 data.frame= TRUE,scale=1)$data
p1T2$rate<-1000*p1T2$event/p1T2$pyears

########## Cumulative Incidence function ###########
ciT2<-cuminc(Ndata$Dbts.time, Ndata$Dbts.Event, Ndata$Groups, rho=0,
 cencode=0)
par(las=1)
plot(ciT2,ylim=c(0,0.05),xlim=c(0,30),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),
curvlab=c('exposed','unexposed'),xaxt="n")
axis(1,c(0,4,9,14,19,24,29,34),c("1","5","10","15","20","25","30","35"))
title("Diabetes")
```

```
## Cumulative incidence curves regarding to country
ci1T2<-cuminc(Ndata$Dbts.time, Ndata$Dbts.Event, Ndata$Country, rho=0,
 cencode=0)
par(las=1)
plot(ci1T2,ylim=c(0,0.026),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),curvlab=c('Denmark',
'Sweden'),xaxt="n")
axis(1,c(0,4,9,14,19,24,29,34),c("1","5","10","15","20","25","30","35"))
title("Country")

## Cumulative incidence curves regarding to Gender
ci3T2<-cuminc(Ndata$Dbts.time, Ndata$Dbts.Event, Ndata$Gender, rho=0,
cencode=0)
par(las=1)
plot(ci3T2,ylim=c(0,0.031),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),curvlab=
c('boys','girls'),xaxt="n")
axis(1,c(0,4,9,14,19,24,29,34),c("1","5","10","15","20","25","30","35"))
title("Gender")
######################### Transition 3  ###################################
#Mortality with diabetes
new<-Ndata[Ndata$Dbts.Event=="1",]

######### Crude Estimate ###############
p1T3<-pyears(formula = Surv(Age.Dbts,AgeExit,D.Event)~Groups,
data = new,
 data.frame= TRUE,scale=1)$data
p1T3$rate<-1000*p1T3$event/p1T3$pyears

########## Cumulative Incidence function ############
ciT3<-cuminc((new$Ca.time-new$Dbts.time), new$D.Event, new$Groups,
 rho=0, cencode=0)
par(las=1)
plot(ciT3,ylim=c(0,0.36),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),
curvlab=c('Childhood cancer','Reference population'),xaxt="n")
axis(1,c(0,4,9,14,19,24,29),c("1","5","10","15","20","25","30"))
title("Mortality w. diabetes")

## Cumulative incidence curves regarding to country
ci1T3<-cuminc((new$Ca.time-new$Dbts.time), new$D.Event, new$Country,
 rho=0, cencode=0)
par(las=1)
plot(ci1T3,ylim=c(0,0.45),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),curvlab=c('Denmark',
'Sweden'),xaxt="n")
```

```
axis(1,c(0,4,9,14,19,24,29,34),c("1","5","10","15","20","25","30","35"))
title("Country")

## Cumulative incidence curves regarding to Gender
ci3T3<-cuminc((new$Ca.time-new$Dbts.time), new$D.Event, new$Gender,
 rho=0, cencode=0)
par(las=1)
plot(ci3T3,ylim=c(0,0.40),lwd=2,
xlab='Years since diagnosis',col=c('blue','red'),curvlab=c('boys',
'girls'),xaxt="n")
axis(1,c(0,4,9,14,19,24,29,34),c("1","5","10","15","20","25","30","35"))
title("Gender")
############ Lexis diagram ##############
library(Epi)
temp1<-Ndata[Ndata$Groups=="exposed",]
temp2<-Ndata[Ndata$Groups!="exposed",]
new<-merge(temp1[runif(nrow(temp1))<0.001,],temp2[runif(nrow(temp2))
<0.0002,],all=T )
LL<-Lexis.diagram( age=c(0,70), date=c(1976,2012),
 entry.age=AgeEntry, exit.age=Age.Dbts, birth.date=cal.yr(BirthDate),
                    col.life=c("red","blue")[Groups],fail=
(Dbts.Event %in% 1), lwd.life=1,  cex.fail=0.8, col.fail=
c("red","blue")[Groups],pch.fail=c(1,16),data=new )
box()
```

C.3 Two state analysis

C.3.1 Transition 1- Analysis of morbidity rate

```
library(relsurv); library(survival);
library(xtable);library(Epi); library(Design);

## Univariate estimate
model1<- coxph(Surv(AgeEntry,Age.Dbts,Dbts.Event)~ Groups,
 data=Ndata)

## Multivariate estimate
model2<- coxph(Surv(AgeEntry,Age.Dbts,Dbts.Event)~
 Groups+AgeDiagnosis, data=Ndata)
model3<- coxph(Surv(AgeEntry,Age.Dbts,Dbts.Event)~
 Groups+AgeDiagnosis+Calendar, data=Ndata)
model4<- coxph(Surv(AgeEntry,Age.Dbts,Dbts.Event)~
 Groups+AgeDiagnosis+Calendar+ Gender, data=Ndata)
```

```
model5<- coxph(Surv(AgeEntry,Age.Dbts,Dbts.Event)~
Groups+AgeDiagnosis+Calendar+ Gender+Country, data=Ndata)
model6<- coxph(Surv(AgeEntry,Age.Dbts,Dbts.Event)~
 Groups+AgeDiagnosis+Calendar+ strata(Gender,Country),
 data=Ndata)

## check functional form
d <- datadist(Ndata)
options(datadist="d")
m1<-cph(Surv(AgeEntry,Age.Dbts,Dbts.Event)~Groups +
 rcs(AgeDiagnosis)+rcs(Calendar), data=Ndata)

plot(m1,AgeDiagnosis=NA,xlab="Age at diagnosis",
adj.subtitle=FALSE)
 # NA: use default range for predictor
plot(m1,Calendar=NA,adj.subtitle=FALSE)

# correct functional form by categorizing calender and
# agediagnosis
Ndata$AgeDiag <- cut(Ndata$AgeDiagnosis,c(0,5,10,15,20
),c("0-5","5-10","10-15","15-20"))
Ndata$Cal <- cut(Ndata$Calendar,c(0,1960,1975,2012)
,c("1943-1960","1960-1974","1974-2010"))
model9<- coxph(Surv(AgeEntry,Age.Dbts,Dbts.Event)~
 Groups + AgeDiag + Cal+ strata(Gender,Country), data=Ndata)
summary(model9)
## check proportionality
cox.zph(model9)
par(mfrow=c(2,2))
plot(cox.zph(model6))
```

C.3.2 Transition 2- Analysis of mortality rate

```
# censoring diabetes events
Ndata$Can.Event<-Ndata$D.Event
Ndata$Can.Event[which(Ndata$Dbts.Event==1)]<-0

## Univariate estimate
model1<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~ Groups,
 data=Ndata)
model2<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~
 AgeDiagnosis, data=Ndata)
model3<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~
 Calendar, data=Ndata)
summary(model1)
```

```
## Multivariate estimate
model4<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~
 Groups+AgeDiagnosis, data=Ndata)
model5<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~
 Groups+AgeDiagnosis+Calendar, data=Ndata)
model6<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~
 Groups+AgeDiagnosis+Calender+ Gender, data=Ndata)
model7<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~
 Groups+AgeDiagnosis+Calender+ Gender+Country, data=Ndata)
model8<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~
Groups+AgeDiagnosis+Calendar+ strata(Gender,Country), data=Ndata)
summary(model8)

## check functional form

d <- datadist(Ndata)
options(datadist="d")

m1<-cph(Surv(AgeEntry,Age.Dbts,Can.Event)~Groups +
 rcs(AgeDiagnosis)+rcs(Calendar), data=Ndata)
plot(m1,AgeDiagnosis=NA,xlab="Age at diagnosis",
adj.subtitle=FALSE)
 # NA: use default range for predictor
plot(m1,Calendar=NA,adj.subtitle=FALSE)


# including splines in the model
model9<- coxph(Surv(AgeEntry,Age.Dbts5,Can.Event)~
 Groups +rcs(AgeDiagnosis) +rcs(Calendar)
+ strata(Gender,Country), data=Ndata)
summary(model9)
cox.zph(model9)
# spline method does not correct for proportionality..
# correct functional form by categorizing calendar
# and age at diagnosis
Ndata$AgeDiag <- cut(Ndata$AgeDiagnosis,c(0,5,10,15,20),
c("0-5","5-10","10-15","15-20"))
Ndata$Cal <- cut(Ndata$Calendar,c(0,1960,1975,2012),
c("1943-1960","1960-1974","1974-2010"))
model9<- coxph(Surv(AgeEntry,Age.Dbts,Can.Event)~
 Groups + AgeDiag + Cal+ strata(Gender,Country),
data=Ndata)
summary(model9)
cox.zph(model9)

### check proportionality by log minus log plot
```

```
cfit<-survfit(Surv(AgeEntry,Age.Dbts,Can.Event)~
Groups,data =Ndata)
plot(cfit,mark.time =F,fun="cloglog",col=c(4,2),lwd=2,
xlab="Age", ylab="Log(-Log(Survival))")
legend ("bottomright",c("exposed", " unexposed "),
col =c(2 ,4) ,lty =1)

#### extend cox model by including time-varying covariates
temp1 <- subset(Ndata, Groups=="exposed")
temp2 <- subset(Ndata, Groups!="exposed")
temp1$time1 <- (temp1$AgeEntry-temp1$AgeDiagnosis)
temp1$time2 <-(temp1$Dbts.time+temp1$AgeEntry-temp1$AgeDiagnosis)

temp1 <- survsplit(temp1,cut=c(5,10,15,20,25),start="time1",
end="time2", event="Can.Event",episode="episode")

temp1$AgeEntry <- temp1$AgeDiagnosis + temp1$time1
temp1$Ca.time   <- temp1$time2 - temp1$time1
temp1 <- temp1[,-c(19,20)]
temp2$episode <- "unexposed"
nydatlat <- rbind(temp1,temp2)
nydatlat$episode <- as.factor(nydatlat$episode)
levels(nydatlat$episode)
levels(nydatlat$episode)[1:6]<-c("0-5","5-10",
"10-15","15-20","20-25","25+")
nydatlat$episode <- relevel(nydatlat$episode, ref="unexposed")

model10 <- coxph(Surv(AgeEntry,(AgeEntry+Ca.time),Can.Event==1)~
AgeDiag + Cal + episode + strata(Country,Gender), data=nydatlat)
rr.Dgs<- ci.lin(model10,subset=c("episode"),Exp=TRUE)[,5:7]

matplot( 1:nrow(rr.Dgs), rr.Dgs[,1],pch=20,log = "y",ylim=c(0.9,100),
las=1,ylab="Rate ratio",xlab="Time since diagnosis",frame.plot=T,
xaxt="n")

for(j in 1:nrow(rr.Dgs)){
lines(c(j,j),rr.Dgs[j,2:3],lwd=2)
}
abline(h=1,lwd=2)
axis(1,c(1,2,3,4,5,6),c("0-5","5-10","10-15","15-20","20-25","25+"))

## check proportionality
z<-cox.zph(model10)
```

C.3.3 Transition 3- Analysis of mortality rate with diabetes

```
new<-Ndata[Ndata$Dbts.Event==1,]

# An univariate analysis
model1 <- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Groups, data=new)
model2 <- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ AgeDiagnosis,
data=new)
model3 <- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Calendar,
data=new)
summary(model1)

# Multivariate analysis
model4<- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Groups+AgeDiagnosis,
 data=new)
summary(model4)
model5<- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Groups+
AgeDiagnosis+Calendar, data=new)
model6<- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Groups+
AgeDiagnosis+Calendar+ Gender, data=new)
model7<- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Groups
+AgeDiagnosis+Calendar+ Gender+Country, data=new)
model8<- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Groups+
AgeDiagnosis+Calendar+ strata(Gender,Country), data=new)
model9<- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Groups+
strata(Gender,Country), data=new)

## check functional form
d <- datadist(Ndata)
options(datadist="d")
m1<-cph(Surv(AgeEntry,AgeExit,D.Event)~Groups +
 rcs(AgeDiagnosis)+rcs(Calendar), data=new)
 plot(m1,AgeDiagnosis=NA,xlab="Age at diagnosis",
adj.subtitle=FALSE)
 # NA: use default range for predictor
plot(m1,Calendar=NA,adj.subtitle=FALSE)

# correct functional form by catogirazing calender and
# agediagnosis
new$AgeDiag <- cut(new$AgeDiagnosis,c(0,5,10,15,20),
c("0-5","5-10","10-15","15-20"))
new$Cal <- cut(new$Calendar,c(0,1960,1975,2012),
c("1943-1960","1960-1974","1974-2010"))
model9<- coxph(Surv(Age.Dbts,AgeExit,D.Event==1)~ Groups
+ AgeDiag + Cal+ strata(Gender,Country), data=new)
summary(model9)
```

```
cox.zph(model9)

## check proportionality
cox.zph(model8)
par(mfrow=c(2,2))
plot(cox.zph(model6))

######## extend cox model by including time-varying covariates
temp1 <- subset(new, Groups=="exposed")
temp2 <- subset(new, Groups!="exposed")
temp1$time1 <-0
temp1$time2 <- temp1$Ca.time-temp1$Dbts.time

temp1 <- survsplit(temp1,cut=c(5,10,15),start="time1",end="time2",
event="D.Event",episode="episode")

temp1$Age.Dbts <- temp1$Age.Dbts + temp1$time1
temp1$Ca.time   <- temp1$time2 - temp1$time1
temp1 <- temp1[,-c(18,19)]
temp2$episode <- "unexposed"
nydatlat <- rbind(temp1,temp2)
nydatlat$episode <- as.factor(nydatlat$episode)
levels(nydatlat$episode)
levels(nydatlat$episode)[1:4]<-c("1-5","5-10","10-15","15+")
#c("1-10","10-20","20-30","30+")
nydatlat$episode <- relevel(nydatlat$episode, ref="unexposed")

model4a <- coxph(Surv(Age.Dbts,(Age.Dbts+Ca.time),D.Event==1)~
AgeDiag+ Cal + episode + strata(Country,Gender), data=nydatlat)
cox.zph(model4a)

model4 <- coxph(Surv(Age.Dbts,(Age.Dbts+Ca.time),D.Event==1)~ episode +
strata(Country,Gender), data=nydatlat)
rr.DD<- ci.lin(model4a,subset=c("episode"),Exp=TRUE)[,5:7]

plot(rr.DD[,1],pch=20,ylim=c(0,31), ylab="Rate ratio",
xlab="Time since diagnosis",xaxt="n")
for(j in 1:nrow(rr.DD)){
lines(c(j,j),rr.DD[j,2:3],lwd=2)
}
abline(h=1,lwd=2)
axis(1,c(1,2,3,4),c("0-5","5-10","10-15","15+"))
axis(2,c(0,1,5,10,15),c(0,1,5,10,15))
```

C.4 Multi state analysis

C.4.1 Univariate estimates

```
library(Epi); library(survival);library(xtable);library(mstate);

dat  <-Lexis(exit=list(tft=(AgeExit-AgeEntry)), exit.status=
factor(D.Event,labels=c("Dgs","Death")),data=Ndata)
datr <-cutLexis(dat, cut = (dat$Age.Dbts-dat$AgeEntry), precursor.states
 = "Dgs",new.state="Dbts", new.scale="tfDbts", split.states=T)
summary(datr)
dats <- stack(datr)

# boxes
boxes.Lexis(datr, boxpos = list(x = c(20, 80, 20, 80),
 y = c(80, 80, 20, 20)), cex = 1.5, wmult = 1.5, hmult = 2.25,
 eq.wd = TRUE, eq.ht = TRUE, show.Y = TRUE, scale.Y = 1, digits.Y = 1,
 show.D = TRUE, scale.D = FALSE, digits.D = 0)
text(c(5,50,65),c(54,95,54),labels=c("Tr. 2","Tr. 1","Tr. 3"), lwd=1,
 cex=1.5)
 text(c(14,73,12.8,66.8),c(82.5,82.5,20,20.5),
 labels=c("1.","2.","3.","4."), lwd=3,lty=3, cex=1.5)

## different covariates, different baseline hazards
c1<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~
lex.Tr:Groups + strata(lex.Tr), data =dats, method = "breslow")

m1<- model.matrix(~lex.Tr:(Groups) , data = dats)
head(m1)
rm1<-grep(":Groupsunexposed", colnames(m1))
m1<-m1[,-c(1,rm1)]

c1<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~ m1 +
 strata(lex.Tr), data =cbind(dats,m1), method = "breslow")
summary(c1)
xtable(data.frame(summary(c1)$coef[,1:3],summary(c1)$conf.int[,3:4],
p.value=summary(c1)$coef[,5]),digits=3,caption="")

## different covariates but common baseline hazards
## assume proportional hazards for hazard rates going into the
## same state

dats1 <- dats[grep("->Death", dats$lex.Tr),]
dats1$lex.Tr<-factor(dats1$lex.Tr)
mm1<- model.matrix(~lex.Tr:(Groups) , data = dats1)
rm1<-grep(":Groupsunexposed", colnames(mm1))
```

```
mm1<-mm1[,-c(1,rm1)]

c2<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~ mm1+
I(lex.Cst == "Dbts"), data =cbind(dats1,mm1), method = "breslow")
summary(c2)
xtable(data.frame(summary(c2)$coef[,1:3],summary(c2)$conf.int[,3:4],
p.value=summary(c2)$coef[,5]),digits=3,caption="common covariates and
 common baseline hazards
")

## common covariates and common baseline hazards
c3<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~ Groups
+I(lex.Cst == "Dbts"), data =cbind(dats1), method = "breslow")
summary(c3)
xtable(data.frame(summary(c3)$coef[,1:3],summary(c3)$conf.int[,3:4],
p.value=summary(c3)$coef[,5]),digits=3,
caption="different covariates but common baseline hazards")

#The effect of occurrence of diabetes on mortality rate among exposed
c4<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~
I(lex.Cst == "Dbts"), data =subset(dats1,Groups=="exposed") ,
 method = "breslow")
summary(c4)
cox.zph(c4)

#The effect of occurrence of diabetes on mortality rate among unexposed
c5<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~
 I(lex.Cst == "Dbts"), data =subset(dats1,Groups=="unexposed") ,
  method = "breslow")
summary(c5)

########## checking proportionality of rates by Poisson modeling

datx <- splitLexis(datr, time.scale = "lex.dur", breaks =
c(0,1,2,3,4,seq(5,30,5)))
datxs <- stack(datx)
datxs <- datxs[grep("->Death", datxs$lex.Tr),]
datxs$lex.Tr <- factor(datxs$lex.Tr)

i.kn <- c(0.2, 0.5, 1, 1.5, 2, 8, 10,30)
b.kn <- c(0,32)
te<- ns(datxs$tft,knots=i.kn, Bo=b.kn)
mi <- glm(as.numeric(lex.Fail) ~ lex.Tr +  lex.Tr:ns(tft,knots=i.kn,
Bo=b.kn) + lex.Tr:Groups, family=poisson, offset=log(lex.dur)
,data=datxs)
ms <- glm(as.numeric(lex.Fail) ~ lex.Tr +  lex.Tr:ns(tft,knots=i.kn,
 Bo=b.kn) + Groups, family=poisson, offset=log(lex.dur)
```

```
 ,data=datxs)
mp <- glm(as.numeric(lex.Fail) ~ lex.Tr +  ns(tft,knots=i.kn, Bo=b.kn)
+ Groups, family=poisson, offset=log(lex.dur),data=datxs)
#memory.limit(size=2000)
anova(mi,ms,mp,mi,test="Chisq")
# significant difference, transitions are not proportional
```

C.4.2 Multivariate estimates

```
dat  <-Lexis(exit=list(tft=(AgeExit-AgeEntry)), exit.status=
factor(D.Event,labels=c("Dgs","Death")),data=Ndata)
datr <-cutLexis(dat, cut = (dat$Age.Dbts-dat$AgeEntry),
precursor.states = "Dgs",new.state="Dbts", new.scale="tfDbts",
 split.states=T)
dats <- stack(datr)

## different covariates, different baseline hazards
c1<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~
 lex.Tr:Groups + strata(lex.Tr), data =dats, method = "breslow")

dats$lex.Tr<-factor(dats$lex.Tr)
m1<- model.matrix(~lex.Tr:(Groups+AgeDiagnosis+Calendar) ,
data = dats)
rm1<-grep(":Groupsunexposed", colnames(m1))
m1<-m1[,-c(1,rm1)]

c1<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~ m1
+ strata(lex.Tr,Country,Gender), data =cbind(dats,m1),
method = "breslow")
summary(c1)

## use result from separate analysis for linear form
m2<- model.matrix(~lex.Tr:(Groups +AgeDiag+Cal) , data = dats)
rm2<-grep(":Groupsunexposed", colnames(m2))
m2<-m2[,-c(1,rm2)]

c2<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~m2 +
strata(lex.Tr,Country,Gender),data =cbind(dats,m2),
method = "breslow")
summary(c2)

## different covariates but common baseline hazards
## assume proportional hazards for hazard rates going into the
## same state..
```

```
dats1 <- dats[grep("->Death", dats$lex.Tr),]
dats1$lex.Tr<-factor(dats1$lex.Tr)
mm1<- model.matrix(~lex.Tr:(Groups+AgeDiag+Cal) , data = dats1)
rm1<-grep(":Groupsunexposed", colnames(mm1))
mm1<-mm1[,-c(1,rm1)]

c4<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~ mm1
+I(lex.Cst == "Dbts")+strata(Country,Gender),data =cbind(dats1,mm1),
 method = "breslow")
summary(c4)

## common covariates and common baseline hazards
c5<-coxph(Surv(tft+AgeEntry,tft+AgeEntry+lex.dur, lex.Fail) ~ Groups+
AgeDiag+Cal+I(lex.Cst == "Dbts"), data =cbind(dats1),
 method = "breslow")
summary(c5)

########## checking proportionality of rates by Poisson modeling

datx <- splitLexis(datr, time.scale = "lex.dur", breaks =
c(0,1,2,3,4,seq(5,30,5)))
datxs <- stack(datx)
datxs <- datxs[grep("->Death", datxs$lex.Tr),]
datxs$lex.Tr <- factor(datxs$lex.Tr)

i.kn <- c(0.2, 0.5, 1, 1.5, 2, 8, 10,30)
b.kn <- c(0,32)
te<- ns(datxs$tft,knots=i.kn, Bo=b.kn)
mi <- glm(as.numeric(lex.Fail) ~ lex.Tr +  lex.Tr:ns(tft,knots=i.kn,
 Bo=b.kn) + lex.Tr:(Groups+AgeDiag+Cal)
, family=poisson, offset=log(lex.dur),data=datxs)
ms <- glm(as.numeric(lex.Fail) ~ lex.Tr +  lex.Tr:ns(tft,knots=i.kn,
 Bo=b.kn) +Groups+AgeDiag+Cal, family=poisson, offset=log(lex.dur)
 ,data=datxs)
mp <- glm(as.numeric(lex.Fail) ~ lex.Tr +  ns(tft,knots=i.kn, Bo=b.kn)
+ Groups+AgeDiag+Cal, family=poisson, offset=log(lex.dur),data=datxs)
#memory.limit(size=2000)
anova(mi,ms,mp,mi,test="Chisq")
# significant difference, transitions are not proportional
```

C.4.3 Predictions

```
library(mstate);library(Epi)

Ndata$Can.Event<-Ndata$D.Event
```

```
Ndata$Can.Event[which(Ndata$Dbts.Event==1)]<-0

## Empirical rates
# mortality without diabetes
stepsize <- 2
maxage <- 80
minage <- 0
tempyr <- tcut(Ndata$AgeEntry,seq(minage,maxage,stepsize),
labels=as.character(seq(minage,maxage-stepsize,stepsize)))

datcan <- pyears(Surv(Dbts.time,Can.Event==1)~tempyr+Groups,dat=Ndata,
data.frame=T,scale=1)$data
datcan$tempyr<-as.numeric(as.character(datcan$tempyr))
datcan$rate <- 1000*datcan$event/datcan$pyears

# Mortality with diabetes
new<-Ndata[Ndata$Dbts.Event==1,]
rmm<-which((new$Ca.time-new$Dbts.time)<=0)
new<-new[-rmm,]

tempyr2 <- tcut(new$Age.Dbts,seq(minage,maxage,stepsize),
labels=as.character(seq(minage,maxage-stepsize,stepsize)))
datdbts <- pyears(Surv((Ca.time-Dbts.time),D.Event==1)~tempyr2+Groups,
dat=new,data.frame=T,scale=1)$data
datdbts$tempyr2<-as.numeric(as.character(datdbts$tempyr2))
datdbts$rate <- 1000*datdbts$event/datdbts$pyears

#a number of prespecied points for prediction
 a.Bo <- c(0,78) # boundary knots for age
 a.kn <- c(13,25,37,49,62) # internal knots for age
 a.int <- 1/20 # interval length for prediction of mortality by age
 a.pr <- seq(1,80,a.int) # prediction point for age-specific mortality

 CA <- ns( a.pr, knots=a.kn, Bo=a.Bo, intercept=TRUE )

e.can <- glm( event ~ ns(tempyr,knots=a.kn,Bo=a.Bo,i=T) - 1,
 offset=log(pyears),family=poisson,data = subset(datcan,Groups==
 "exposed") )
summary(e.can)

e.dbts <- glm(  event ~ ns(tempyr2,knots=a.kn,Bo=a.Bo,i=T) - 1,
offset=log(pyears),family=poisson,data = subset(datdbts,Groups==
"exposed"))
summary(e.dbts)

u.can <- glm( event ~ ns(tempyr,knots=a.kn,Bo=a.Bo,i=T) - 1,
offset=log(pyears),family=poisson,data = subset(datcan,Groups==
```

```
"unexposed") )
summary(u.can)

u.dbts <- glm(event ~ ns(tempyr2,knots=a.kn,Bo=a.Bo,i=T) - 1,
offset=log(pyears),family=poisson,data = subset(datdbts,
Groups=="unexposed"))
summary(u.dbts)

# compute rates for the prespecified points from Poisson models
er.can <- ci.lin( e.can, ctr.mat=cbind(CA), E=T )[,5:7]
er.dbts <- ci.lin( e.dbts, ctr.mat=cbind(CA), E=T )[,5:7]

 ur.can <- ci.lin( u.can, ctr.mat=cbind(CA), E=T )[,5:7]
 ur.dbts <- ci.lin( u.dbts, ctr.mat=cbind(CA), E=T )[,5:7]

## Plot of Mortality rate per 1000 person-years
yl <- c(0.1,500)
 yt <- as.vector( outer( c(1,2,5), 10^(-2:2), "*" ) )
 yg <- as.vector( outer( 1:9, 10^(-2:2), "*" ) )

par( mfrow=c(1,2), mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, las=1,
oma=c(4,4,1,1) )
plot( 1, 1, type="n", log="y", ylim=yl, yaxt="n", xlim=c(0,82) )
axis( side=2, at=yt, labels=formatC(yt) )
abline( v=seq(0,82,10), h=yg, col=gray(0.8) )
matlines( a.pr, cbind(er.can[,1],ur.can[,1])*1000,
type="l", lty=1, lwd=c(3,3),
col=rep(c("blue","red"),each=1) )
lines(datcan$tempyr[datcan$Groups=="exposed"],datcan$rate[datcan$Groups
=="exposed"],type="b",col="blue",lwd=1)
lines(datcan$tempyr[datcan$Groups!="exposed"],datcan$rate[datcan$Groups
!="exposed"],type="b",col="red",lwd=1)
 box()
 text( 0, 500, "Without diabetes", adj=0 )
 plot( 1, 1, type="n", log="y", ylim=yl, yaxt="n", xlim=c(0,82) )
 abline( v=seq(0,82,10), h=yg, col=gray(0.8) )
 matlines( a.pr, cbind(er.dbts[,1],ur.dbts[,1])*1000,
 type="l", lty=1, lwd=c(3,3),
 col=rep(c("blue","red"),each=1) )
lines(datdbts$tempyr[datdbts$Groups=="exposed"],
datdbts$rate[datdbts$Groups=="exposed"],type="b",col="blue")
lines(datdbts$tempyr[datdbts$Groups!="exposed"],
datdbts$rate[datdbts$Groups!="exposed"],type="b",col="red")
 text( 0, 500, "With diabetes", adj=0 )
 mtext( side=1, line=2.5, "Age (years)", outer=T )
 mtext( side=2, line=2.5, "Mortality rate (per 1000 PY)", outer=T,
  las=0 )
```

```
box()
x11()

# New plot: plot the different rates in the same frame to show them
# relative to each other separately for each group.

# tick marks on the y-axis
yt <- as.vector( outer( c(1,2,5), 10^(-2:2), "*" ) )
# horizontal grid lines
 yg <- as.vector( outer( 1:9, 10^(-2:2), "*" ) )

par( mfrow=c(1,2), mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, las=1,
 oma=c(4,4,1,1) )
 plot( 1, 1, type="n", log="y", ylim=yl, yaxt="n", xlim=c(0,82) )
 axis( side=2, at=yt, labels=formatC(yt) )
abline( v=seq(0,82,10), h=yg, col=gray(0.8) )
 matlines( a.pr, cbind(er.can,er.dbts)*1000,
 type="l", lty=1, lwd=c(3,1,1),
 col=rep(topo.colors(3),each=3) )
 box()
 text( 60, 500, "Exposed", adj=0 )
plot( 1, 1, type="n", log="y", ylim=yl, yaxt="n", xlim=c(0,82) )
 abline( v=seq(0,82,10), h=yg, col=gray(0.8) )
 matlines( a.pr, cbind(ur.can,ur.dbts)*1000,
 type="l", lty=1, lwd=c(3,1,1),
 col=rep(topo.colors(3),each=3) )
 box()
 text( 58, 500, "Unexposed", adj=0 )
text( 80, c(0.2,0.15), c("Without diabetes","With diabetes"),
 col=topo.colors(3), adj=1, font=2 )
 mtext( side=1, line=2.5, "Age (years)", outer=T )
 mtext( side=2, line=2.5, "Mortality rate (per 100 PY)", outer=T,
 las=0 )

#################Prediction:Cumulative risks #################
# survival functions for exposed and unexposed
 e.surv <- exp( -cumsum( (er.can[,1]+er.dbts[,1])*a.int ) )
 u.surv <- exp( -cumsum( (ur.can[,1]+ur.dbts[,1])*a.int ) )
 matplot( a.pr, cbind(e.surv,u.surv), type = "l", ylim=c(0,1),
 col=c("blue","red"), lwd=3, lty=1 )

#cumulative probabilities
e.pd.can <- cumsum( er.can[,1]*e.surv*a.int )
e.pd.dbts <- cumsum( er.dbts[,1]*e.surv*a.int )

u.pd.can <- cumsum( ur.can[,1]*u.surv*a.int )
u.pd.dbts <- cumsum( ur.dbts[,1]*u.surv*a.int )
```

```
 par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
 # exposed
 matplot( a.pr, cbind(e.surv,
 e.surv+e.pd.dbts,
 e.surv+e.pd.dbts+e.pd.can),
 type="l", lty=1, lwd=3,
 ylim=c(0,1), col="blue",
 xlab = "Age", ylab="Fraction exposed dead" )
 ll <- min( which(a.pr>70) )
 text( 80, (e.surv+e.pd.dbts/2)[ll], "Death with diabetes", adj=1 )
 text( 80, (e.surv+e.pd.dbts+e.pd.can/2)[ll], "Death without diabetes",
  adj=1 )
 # unexposed
 matplot( a.pr, cbind(u.surv,
 u.surv+u.pd.dbts,
 u.surv+u.pd.dbts+u.pd.can),
 type="l", lty=1, lwd=3,
 ylim=c(0,1), col="red",
 xlab = "Age", ylab="Fraction unexposed dead" )

ll <- min( which(a.pr>70) )
 text( 80, (u.surv+u.pd.dbts/2)[ll], "1", adj=1 )
 text( 80, (u.surv+u.pd.dbts+u.pd.can/2)[ll], "2", adj=1 )

 text( 40,0.15 , "1: Death with diabetes", adj=1 )
 text( 45,0.1 , "2: Death without diabetes", adj=1 )

# conditional survival
incl <- (a.pr>19.9)
 e.surv <- exp( -cumsum( (er.can[,1]+er.dbts[,1])*a.int*incl ) )
 u.surv <- exp( -cumsum( (ur.can[,1]+ur.dbts[,1])*a.int*incl ) )

 e.pd.can <- cumsum( er.can[,1]*e.surv*a.int*incl )
 e.pd.dbts <- cumsum( er.dbts[,1]*e.surv*a.int*incl )
  u.pd.can <- cumsum( ur.can[,1]*u.surv*a.int*incl )
 u.pd.dbts <- cumsum( ur.dbts[,1]*u.surv*a.int*incl )

x11()
par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
 # exposed
 matplot( a.pr, cbind(e.surv,
 e.surv+e.pd.dbts,
 e.surv+e.pd.dbts+e.pd.can),
 type="l", lty=1, lwd=3,
 ylim=c(0,1), col="blue",
 xlab = "Age", ylab="Fraction exposed dead" )
```

```
ll <- min( which(a.pr>70) )
text( 80, (e.surv+e.pd.dbts/2)[ll], "1", adj=1 )
text( 80, (e.surv+e.pd.dbts+e.pd.can/2)[ll], "2", adj=1 )
text( 40,0.15 , "1: Death with diabetes", adj=1 )
text( 45,0.1 , "2: Death without diabetes", adj=1 )

# unexposed
matplot( a.pr, cbind(u.surv,
u.surv+u.pd.dbts,
u.surv+u.pd.dbts+u.pd.can),
type="l", lty=1, lwd=3,
ylim=c(0,1), col="red",
xlab = "Age", ylab="Fraction unexposed dead" )
ll <- min( which(a.pr>70) )
text( 80, (u.surv+u.pd.dbts/2)[ll], "1", adj=1 )
text( 80, (u.surv+u.pd.dbts+u.pd.can/2)[ll], "2", adj=1 )

text( 40,0.15 , "1: Death with diabetes", adj=1 )
text( 45,0.1 , "2: Death without diabetes", adj=1 )
```

Bibliography

[1] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and Event History Analysis: A Process Point of View*. ISBN: 0387202870. Springer, first edition, August 2008.

[2] Mertens AC, Yasui Y, Neglia JP, Potter JD, Nesbit ME Jr, Ruccione K, Smithson WA, and Robison LL. Late mortality experience in five-year survivors of childhood and adolescent cancer: the childhood cancer survivor study. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 19(13):3163–72, July 2001.

[3] Per Kragh Andersen. *Statistical Models Based on Counting Processes*. ISBN 0387945199. Springer, second edition, January 1997.

[4] Per Kragh Andersen and Niels Keiding. Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115, April 2002.

[5] Per Kragh Andersen and Niels Keiding. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, August 2011.

[6] Per Kragh Andersen, Ørnulf Borgan, and Niels Keiding Richard Gill. Linear nonparametric tests for comparison of counting process with application to censored survival data. *International Statistical Review*, 50(3):219–258, December 1982.

[7] E. Arjas. A graphical method for assessing goodness ofit in cox's proportional hazards model. *Journal of the American Statistical Association*, 83:204–212, 1988.

[8] Nihal Ata and M.Tekin Sozer. Cox regression models with nonproportional hazards applied to lung cancer survival data. *Hacettepe Journal of Mathematics and Statistics*, 36(2):157–167, 2007.

[9] B. Carstensen, J. K. Kristensen, P. Ottosen, and K. Borch-Johnsen. The danish national diabetes register: trends in incidence, prevalence and mortality. *Diabetologia*, 51(12):2187–2196, December 2008.

[10] Bendix Carstensen. Demography and epidemiology: Practical use of the lexis diagram in the computer age. or: Who needs the cox-model anyway? *Annual meeting of Finnish Statistical Society*, December 2005.

[11] Bendix Carstensen. Practical aspects of multistate modelling: Representation, timescales and prediction. August 2009.

[12] Bendix Carstensen. Rates and (competing) risks: Example calculations using danish cause of death data. pages 1–28, 2011.

[13] Bendix Carstensen and Martyn Plummer. Using lexis objects for multistate models in R. *Journal of Statistical Software*, 38(6), January 2011.

[14] Xiao Chen. Score test of proportionality assumption for cox models. *Statistical Consulting Group UCLA*.

[15] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, 34.

[16] de Wreede LC, Fiocco M, and Putter H. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, 99(3):261–74, March 2010.

[17] John Fox. Cox proportional-hazards regression for survival data. February 2002.

[18] Vivian I. Franco, Jacqueline M. Henkel, Tracie L. Miller, and Steven E. Lipshultz. Cardiovascular effects in childhood cancer survivors treated with anthracyclines. *Cardiology Research and Practice*, 2011:13, February 2010.

[19] Maud M. Geenen, Mathilde C. Cardous-Ubbink, Leontien C. M. Kremer, Cor van den Bos, Helena J. H. van der Pal, Richard C. Heinen, Monique W. M. Jaspers, Caro C. E. Koning, Foppe Oldenburger, Nelia E. Langeveld, Augustinus A. M. Hart, Piet J. M. Bakker, Huib N. Caron, and Flora E. van Leeuwen. Medical assessment of adverse health outcomes in long-term survivors of childhood cancer. *Journal of the American Medical Association*, 297(24):2705–15, June 2007.

[20] Charlotte Glümer, Torben Jørgensen, and Knut Borch-Johnsen. Prevalences of diabetes and impaired glucose regulation in a danish population. *Diabetes care*, 26(8):2335–2340, August 2003.

[21] PM Grambsch and TM Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Bimetrika*, 81(3):515–526, August 1994.

[22] Robert J. Gray. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Statistics in Medicine*, 16(3):1141–1154, September 1988.

[23] Putter H, Fiocco M, and Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–430, May 2007.

[24] David W. Hosmer and Stanley Lemeshow. *Applied survival analysis - regression modeling of time to event.* ISBN 0471154105. John Wiley and Sons, first edition, January 1999.

[25] Philip Hougaard. *Analysis of Multivariate Survival Data.* ISBN 978-0-387-98873-3. Springer, first edition, August 2000.

[26] Melissa M. Hudson, Ann C. Mertens, Yutaka Yasui, Wendy Hobbie, Hegang Chen, James G. Gurney, Mark Yeazel, Christopher J. Recklitis, Neyssa Marina, Leslie R. Robison, and Kevin C. Oeffinger. Health status of adult long-term survivors of childhood cancer. *Journal of the American Medical Association*, 290(12):1583–1592, September 2003.

[27] Gurney JG., Ness KK., Sibley SD., O'Leary M., Dengel DR., Lee JM., Youngren NM., Glasser SP., and Baker KS. Metabolic syndrome and growth hormone deficiency in adult survivors of childhood acute lymphoblastic leukemia. *American Cancer Society*, 107(6):1303–12, September 2006.

[28] Frank E. Harrell Jr., Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, February 1996.

[29] Luke Keele. Covariate functional form in cox models. October 2005.

[30] John P. Klein and Merlvin L. Moeschberger. *Survival Analysis: Techniques for Censored adn Truncated Data.* ISBN 0387948295. Springer, second edition, 1997.

[31] Baker KS, Ness KK, Steinberger J, Carter A, Francisco L, Burns LJ, Sklar C, Forman S, Weisdorf D, Gurney JG, and Bhatia S. Diabetes, hypertension, and cardiovascular events in survivors of hematopoietic cell transplantation: a report from the bone marrow transplantation survivor study. *Blood Journal*, 109(4):1765–72, October 2006.

[32] D. Y. Lin. Cox regression analysis of multivariate failure time data: the marginal approach. *Statiscs in Medicine*, 13(21):2233–2247, November 1994.

[33] Fiocco M, Putter H, and van Houwelingen HC. Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in medicine*, 27(21):4340–4358, September 2008.

[34] Lillian R. Meacham, Charles A. Sklar, Suwen Li, Qi Liu, Nora Gimpel, Yutaka Yasui, John A. Whitton, Marilyn Stovall, Leslie L. Robison, and Kevin C. Oeffinger. Diabetes mellitus in long-term survivors of childhood cancer. *Arch Intern Med.*, 169(15):1381–1388, August 2009.

[35] Ann C. Mertens, Qi Liu, Joseph P. Neglia, Karen Wasilewski, Wendy Leisenring, Gregory T. Armstrong, Leslie L. Robison, and Yutaka Yasuil. Cause-specific late mortality among 5-year survivors of childhood cancer: The childhood cancer survivor study. *Journal of the National Cancer Institute*, 100(19):1368–1379, October 2008.

[36] Torgil R. Moller, Stanislaw Garwicz, Lotti Barlow, Jeanette Falck Winther, Eystein Glattre, Gudridur Olafsdottir, Jorgen H. Olsen, Roland Perfekt, Annukka Ritvanen, Risto Sankila, and Hrafn Tulinius. Decreasing late mortality among five-year survivors of cancer in childhood and adolescence: A population-based study in the nordic countries. *Journal of Clinical Oncology*, 19(13):3173–3181, July 2001.

[37] Jennings MT, Gelman R, and Hochberg F. Intracranial germ-cell tumors: natural history and pathogenesis. *Journal Neurosurg*, 63(2):155–67, August 1985.

[38] Paul C. Nathan, Mark L. Greenberg, Kirsten K. Ness, Melissa M. Hudson, Ann C. Mertens, Martin C. Mahoney, James G. Gurney, Sarah S. Donaldson, Wendy M. Leisenring, Leslie L. Robison, and Kevin C. Oeffinger. Medical care in long-term survivors of childhood cancer: A report from the childhood cancer survivor study. *Journal of Clinical Oncology*, 26(27), September 2008.

[39] K. C. Oeffinger, A. C. Mertens, and C. A. Sklar. Chronic health conditions in adult survivors of childhood cancer. *The New England Journal of Medicine*, 355(15):1572–1582, October 2006.

[40] K. C. Oeffinger, A. C. Mertens, and C. A. Sklar. Chronic health conditions in adult survivors of childhood cancer. *The New England Journal of Medicine*, 355(15):1572–1582, October 2006.

[41] Hein Putter. Special issue about competing risks and multi-state models. *Journal of Statistical Software*, 38(1), January 2011.

[42] David Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Oxford Journals*, 67(1):145–153, April 1979.

[43] Cindy L. Schwartz. *Survivors of Childhood and Adolescent Cancer: A Multidisciplinary Approach*. ISBN 9783540408406. Springer, Berlin, second edition, 2004.

[44] L Scrucca, A Santucci, and F Aversa. Competing risk analysis using ⁝ an easy guide for clinicians. *Bone Marrow Transplantation*, 40:381–387, June 2007.

[45] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. ISBN 978-0-387-98784-2. Springer, first edition, December 2000.

[46] Terry M Therneau, Patricia M Grambsch, and Thomas R Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, March 1990.

[47] Website. Adult life after childhood cancer in scandinavia. http://www.cancer.dk/aliccs/about+aliccs/, September 2011.

[48] Website. Aliccs-participating parties. http://www.cancer.dk/aliccs/participating+parties/, September 2011.

[49] Website. Backgorund for survival analysis, October 2011. http://www.ats.ucla.edu/stat/sas/seminars/sas_survival/default.htm.

[50] Website. Bleeding, September 2011. http://en.wikipedia.org/wiki/Bleeding.

[51] Website. Cardiomyopathy, September 2011. http://en.wikipedia.org/wiki/Cardiomyopathy.

[52] Website. Case-control studies, September 2011. http://www.ehib.org/faq.jsp?faq_key=34.

[53] Website. Children diagnosed with cancer: Late effects of cancer treatment. http://www.cancer.org/Treatment/ChildrenandCancer/WhenYourChildHasCancer/children-diagnosed-with-cancer-late-effects-of-cancer-treatment, September 2011.

[54] Website. Cohort, September 2011. http://en.wikipedia.org/wiki/Cohort_%28statistics%29.

[55] Website. Confounding, September 2011. http://en.wikipedia.org/wiki/Confounding.

[56] Website. Cox regression, September 2011. http://www.statsdirect.com/help/survival_analysis/cox_regression.htm.

[57] Website. Diabetes, September 2011. http://en.wikipedia.org/wiki/Diabetes_mellitus.

[58] Website. Epidemiology, September 2011. http://www.who.int/topics/epidemiology/en/.

[59] Website. Goodness of fit, October 2011. http://en.wikipedia.org/wiki/Goodness_of_fit.

[60] Website. Hazard function, October 2011. http://www.engineeredsoftware.com/nasa/hazard.htm.

[61] Website. Health risks of cardiovascular and pulmonary disease. http://www.livestrong.com/article/184181-health-risks-of-cardiovascular-pulmonary-disease/, September 2011.

[62] Website. Hypertension, September 2011. http://en.wikipedia.org/wiki/Hypertension.

[63] Website. Late effects of childhood cancer. http://www.cancer.net/patient/All+About+Cancer/Cancer.Net+Feature+Articles/After+Treatment+and+Survivorship/Late+Effects+of+Childhood+Cancer, October 2011.

[64] Website. Likelihood ratio test, October 2011. http://en.wikipedia.org/wiki/Likelihood-ratio_test.

[65] Website. Person civil registration codes for denmark. http://rep.oio.dk/cpr.dk/xml/schemas/core/2005/11/24/cpr_personcivilregistrationstatuscode.xsd.meta.xml, October 2011.

[66] Website. Semiparametric model, October 2011. http://en.wikipedia.org/wiki/Semiparametric_model.

[67] Website. Survival function, October 2011. http://en.wikipedia.org/wiki/Survival_function.

[68] Website. Childhood central nervous system germ cell tumors treatment. http://www.cancer.gov/cancertopics/pdq/treatment/childCNS-germ-cell/healthprofessional/page1/AllPages/Print#Reference9.1, February 2012.

[69] Website. Late effects of cancer treatment. http://csn.cancer.org/node/164654, January 2012.

[70] Website. Late effects of treatment for childhood cancer. http://journeyforward.org/general/late-effects-treatment-childhood-cancer-professional-pdq, February 2012.

[71] Website. Late effects of treatment for childhood cancer, February 2012. http://www.cancer.gov/cancertopics/pdq/treatment/lateeffects/HealthProfessional/page1/AllPages/Print.

[72] Website. Lexis diagram, January 2012. http://en.wikipedia.org/wiki/Lexis_diagram.

[73] Website. Metabolic syndrome, January 2012. http://en.wikipedia.org/wiki/Metabolic_syndrome.

[74] Website. Standardized mortality ratio, January 2012. http://en.wikipedia.org/wiki/Standardized_mortality_ratio.

[75] Website. What are the short and long term side effects of chemotherapy? http://www.chemocare.com/whatis/what_are_the_short_and_long_term.asp, January 2012.