Statistical Analysis of Familial Aggregation of Adverse Outcomes

Luise Cederkvist Kristiansen

Kongens Lyngby 2012
IMM-MSc-2012-13

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
IMM-MSc-2012-13

Summary

In survival analysis, the survival times of the subjects in the study population are generally assumed to be statistically independent, conditional on the covariate information. However, situations where the survival times are correlated due to a natural clustering of the study subjects may arise.

In this study, different statistical methods for analysis of clustered survival data are evaluated and compared using data from a Danish register-based family study of the psychological effects of exposure to childhood cancer. In addition to assessing the effect of exposure to childhood cancer whilst coping with familial clustering, two of the presented models are applied in order to estimate familial correlation of ages at onset. The models show that individuals diagnosed with cancer and individuals with a family history of admission have an increased hazard rate. Furthermore, a significant correlation of age at onset within families is identified.

Of the models presented, the shared gamma frailty Cox proportional hazards model and the Clayton-Oakes copula model with the marginal Cox proportional hazards model as margin are the most applicable in this study.

Resumé

I overlevelsesanalyse antages det generelt, at observationernes overlevelsestider er statistisk uafhængige betinget af kovariaterne. Der kan imidlertid opstå situationer, hvor overlevelsestiderne er korrelerede pga. en naturlig gruppering af data.

I dette projekt evalueres og sammenlignes forskellige statistiske metoder til analyse af korreleret overlevelsdata vha. data fra et dansk registerbaseret familiestudie af de psykologiske senfølger af eksponering for børnecancer. Udover at estimere effekten af eksponering for børnecancer, alt imens der tages højde for, at data er grupperet, kan to af de præsenterede modeller bruges til at estimere korrelationen mellem overlevelsestiderne indenfor en familie. Modellerne viser, at individer, der diagnosticeres med kræft, samt individer med tidligere indlæggelser i familien har en øget hazard rate. Endvidere ses det, at der er en signifikant korrelation mellem overlevelsestiderne indenfor en familie.

De mest anvendelige modeller er i dette projekt shared gamma frailty Cox proportional hazards modellen og Clayton Oakes copula modellen med den marginale Cox proportional hazards model som margin.

Preface

This thesis was prepared at the Department of Informatics and Mathematical Modelling (IMM), the Technical University of Denmark and at the Department of Statistics, Bioinformatics and Registry (SBR), the Danish Cancer Society in partial fulfillment of the requirements for acquiring a Master of Science in Engineering. The thesis was supervised by Per Bruun Brockhoff, IMM, and co-supervised by Klaus Kaae Andersen and Kirsten Frederiksen, SBR.

The thesis deals with different statistical methods for analysis of clustered survival data. The main focus is on semi-parametric methods, though also parametric methods are presented. The statistical methods are explored using three small data sets and then applied to data from a Danish register-based family study of the psychological late effects of exposure to childhood cancer.

Copenhagen, February 2012

Luise Cederkvist Kristiansen

Acknowledgements

First of all, I would like to thank my supervisors Per Bruun Brockhoff, Klaus Kaae Andersen and Kirsten Frederiksen. They have been great in different ways and have throughout this project offered valuable guidance and support. I have truly appreciated my weekly meetings with Klaus and Kirsten, and I am grateful of their commitment to my project.

I would also like to thank Lasse Wegener Lund from the Danish Cancer Society for letting me use his data in this study.

Finally, it would like to thank my parents for giving me shelter the last couple of weekends and my friend Kathrine Grell for encouraging me and for proofreading my thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In survival analysis, the survival times of the subjects in the study population are generally assumed to be statistically independent, conditional on the covariate information. However, situations where the survival times are correlated due to a natural clustering of the study subjects may arise and can occur for different kinds of data [23, 36]. Simple examples where independence between survival times cannot be assumed are the lifetimes of related individuals, e.g. twins, or time between recurrent events.

Dependence between survival times may be considered a nuisance of survival data. In other applications the correlation is of primary interest. For example, in family studies, correlation of age at onset is typically of main interest and considered as evidence of familial aggregation [1]. Familial aggregation may be attributed to unobserved genetic and environmental factors, which are shared by the members of a family, and it may be important for understanding the etiology of many common diseases including cancers and psychiatric disorders.

A commonly used and very general approach to the modelling of clustered survival data is to assume that there is an unobserved risk factor, a so-called frailty, which is shared by all subjects in a cluster, see e.g. [22, 24, 35, 37, 43, 49]. This is similar to classic linear regression, where a cluster effect is typically modelled as a random effect. In the classic linear regression model, the mean of the response variable is unaltered by the random effect because of the linear structure of the

model [36, 39], however, in the shared frailty model the covariate effects are specified conditionally on the frailty and are thus to be interpreted on cluster level, i.e. within clusters.

Another approach is to apply a marginal model, see e.g. [18, 19, 25, 47, 56], where the survival times are compared across clusters and the covariate effects may be interpreted on population level. In the marginal model, the covariate effects are modelled without taking the clustering of subjects into account, however, the standard errors of the estimates are subsequently adjusted for correlation of survival times.

As mentioned, the association between survival times may be considered a nuisance or an interesting aspect of survival data. The shared frailty model can readily be used to obtain a measure of the dependence between survival times of subjects in a cluster, whereas the marginal model cannot. Yet, the marginal model may be used in combination with a copula model in order to obtain a measure of dependence [2, 16, 42, 50]. Copula models can link population survival functions to generate the joint survival function and in the process estimate the dependence between the population survival functions [5, 42]. By using the combined approach, an estimate of the degree of dependence between the clustered survival times are obtained together with covariate effects that may be interpreted on population level.

1.1 Objectives

The objectives of this study, is to evaluate and compare different statistical methods for analysis of clustered survival data. This includes the shared frailty model and the marginal model in combination with the copula model. In addition, simpler methods will be studied. Focus is on semi-parametric methods, however their parametric counterparts are also presented, though not applied.

The statistical methods will be evaluated and compared using data from a Danish register-based family study of the psychological late effects of exposure to childhood cancer. The purpose of the family study is to investigate how childhood cancer survivors and their parents and siblings are affected later in life with regard to psychological outcomes. In this study, focus is on the childhood cancer survivors and their siblings. In addition to assessing the effect of exposure to childhood cancer whilst coping with familial clustering, the statistical methods are applied in order to estimate correlation of age at onset of psychological disorders within families.

Before the statistical methods are applied to the large data set from the register-based family study, they are explored using three smaller data sets, which are available through the statistical software R [48].

1.2 Overview of thesis

Chapter 2 will start with a brief introduction to survival analysis, where the basic terminology will be covered. Hereafter, the different statistical methods, which are applied in this study, will be presented. Although, focus in this study is on semi-parametric survival models, their parametric counterparts are also presented. In Chapter 3, the data analysed in this study will be presented and it will be described how the statistical analyses are conducted. The results from the statistical analyses are presented in Chapter 4. In Chapter 5, the applied statistical models are discussed based on the results. This will be followed by suggestions for further work. Finally, a summary of the results and conclusions is given in Chapter 6.

Chapter 2

Theory

In this chapter, the different statistical methods, which are applied in this study, will be described. First, a brief introduction to survival analysis is given and then the proportional hazards model is presented. Finally, different statistical methods for analysis of clustered survival data are presented. The methods are all based on the proportional hazards model. Two of the methods presented, may be applied for estimation of the degree of dependence between the clustered survival data.

2.1 Survival analysis

Survival analysis is statistical analysis of data, where the response of interest is time from a well-defined origin to the occurrence of an event of interest [36]. A key feature, which makes survival analysis different from other areas in statistics is that survival data is usually censored [7]. Censoring occurs when the exact survival time (time until event of interest) is unknown. The survival time may be unknown because the subjects have withdrawn from the study or in some other way been lost to follow-up, e.g. moved to another country. For these subjects, the survival time is at least until withdrawal or last contact. Subjects for whom no event of interest has occurred at the end of the study are also

censored. Their survival times are at least until the end of the study. The type of censoring described here is called right censoring and is the most common form of censoring in survival data [27] though other censoring schemes exist, e.g. left censoring which is applied when the event of interest occurs prior to a certain time t or interval censoring which is applied if the event of interest occurs between times t_a and t_b. Censoring is assumed to be statistically independent of the survival time.

In addition to censoring, survival data can be truncated. Truncation is concerned with the entry of subjects into a study. When the survival data is truncated, only subjects with survival times within a certain interval, e.g. $[T_L, T_R]$ are observed. Left truncation occurs when subjects for whom the event of interest either has occurred before some truncation threshold T_L, i.e. $T < T_L$, or is known never to occur, are excluded from the study. Right truncation occur when subjects for whom the event of interest has occurred after some truncation threshold T_R, i.e. $T_R < T$, are excluded from the study [29].

2.1.1 Terminology

The information of interest for a subject is contained in the pair (T, δ), where T is the survival time until the event of interest or censoring, and δ is the censoring indicator. The survival time T is a random variable and is equal to or greater than 0. The indicator variable δ is equal to 1 if the event of interest has occurred and 0 otherwise

$$T \geq 0 \qquad \text{and} \qquad \delta = \begin{cases} 1 & \text{if event} \\ 0 & \text{if censored} \end{cases}$$

A visualisation of (T, δ) is given in Figure 2.1.

Three important and closely related functions in survival analysis are the probability density function, $f(t)$, the survival function, $S(t)$, and the hazard function, $h(t)$. Specifying one of three functions, specifies all three functions as there is a clearly defined relationship between them, i.e. the probability density function can be expressed as

$$f(t) = h(t)S(t) \tag{2.1}$$

The probability density function, $f(t)$, gives the unconditional event rate and is defined as

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t} \tag{2.2}$$

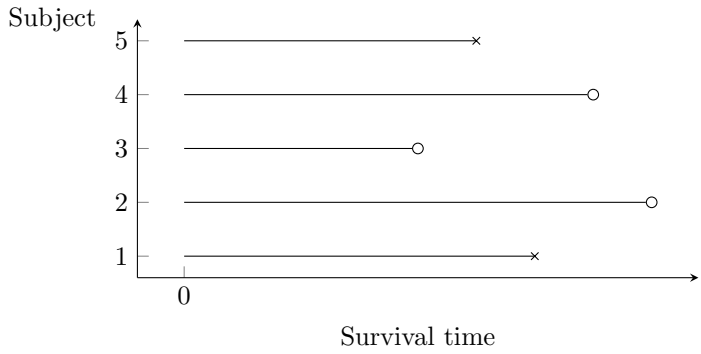

Figure 2.1: An example of survival times and censoring. The symbol × indicates event and the symbol ∘ indicates censoring.

The probability density function is non-negative and the integral of $f(t)$ from 0 to infinity is equal to 1. The corresponding cumulative distribution function is given as

$$F(t) = \Pr(T \leq t) = \int_0^t f(u)\,\mathrm{d}u \tag{2.3}$$

The survival function, $S(t)$ is defined as the probability of a subject surviving longer than time t

$$S(t) = \Pr(T > t) = 1 - F(t) = \int_t^\infty f(u)\,\mathrm{d}u \tag{2.4}$$

The survival function is non-increasing and theoretically equal to one for t equal to zero, and zero for t equal to infinity. An example of a hypothetical survival function is shown in Figure 2.2.

The survival function gives the cumulative survival and may be estimated using the non-parametric Kaplan-Meier method [23, 29]. If there are multiple events at the same time (ties), the Kaplan-Meier estimate is given as

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{m_i}{Y(t_i)}\right) \tag{2.5}$$

where $t_1, \ldots, t_e$ are the ordered event times, m_i is the number of events at t_i, and $Y(t_i)$ is the number of subjects at risk immediately before t_i. The Kaplan-Meier function is right continuous decreasing step function, that changes at each

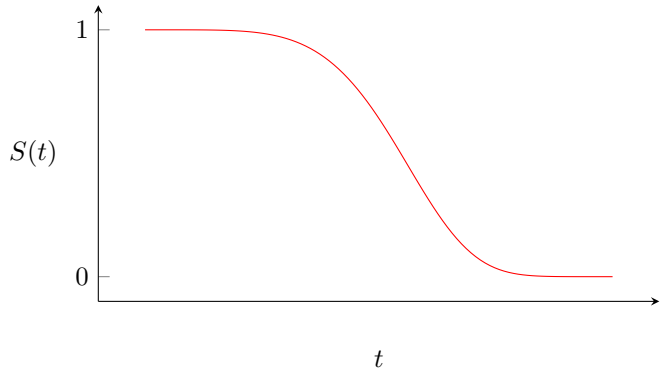

Figure 2.2: An example of a hypothetical survival function.

event time. Utilising, that $S(t) = 1 - F(t)$ (2.4), the Kaplan-Meier method can also be used to estimate the cumulative incidence.

The hazard function, $h(t)$, gives the instantaneous and conditional event rate and is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2.6}$$

The hazard function is non-negative and it typically progresses according to the event being studied. For example, if the event of interest is dying after having received complicated and risky surgery, the hazard function is most likely decreasing, as the risk of dying from the surgery will decrease as time goes by. Conversely, the hazard function is most likely increasing if the event of interest is dying after being diagnosed with a fatal illness.

2.2 Proportional hazards models

In survival analysis as well as in many other areas of statistics, the goal is to obtain some measure of effect describing the relationship between given covariates and a given outcome. In survival analysis, the outcome of interest is time to event, and the effect of the covariates of interest is most often measured using the proportional hazards model [7], which is based on the hazard function. The proportional hazards model is given as

$$h(t|\boldsymbol{X}) = h_0(t) \exp(X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k) \tag{2.7}$$

where $h_0(t)$ is the baseline hazard function, $X_1, \ldots, X_k$ are the covariates, and $\beta_1, \ldots, \beta_k$ are the covariate effects. The covariate effects act multiplicatively on and thereby scale the baseline hazard function, which is common to all subjects. As seen, the effects of the covariates are independent of time, and thus assumed to be the same at all values of t. The covariate effects are additive and linear for the log hazard

$$\log(h(t|\boldsymbol{X})) = \log(h_0(t)) + X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k \tag{2.8}$$

The baseline hazard function $h_0(t)$ can be assumed to have a particular parametric form, i.e. have the survival times follow some distribution, or left unspecified. Commonly used distributions for the survival times are the Weibull distribution, the exponential distribution (which is a special case of the Weibull distribution), and the log-logistic distribution [27]. If the baseline hazard is left unspecified, the proportional hazards model is semi-parametric. In the following sections, the parametric Weibull proportional hazards model and the semi-parametric Cox proportional hazards model are introduced.

2.2.1 Weibull proportional hazards model

The Weibull model is the most widely used parametric survival model [27]. If the survival times are assumed to be Weibull distributed, $T \sim W(\lambda, \rho)$, then the hazard function is given as

$$h(t) = \lambda\rho t^{\rho-1}, \quad \rho > 0 \text{ and } \lambda > 0 \tag{2.9}$$

The shape of the hazard function is determined by the shape parameter ρ, which is typically held fixed. As illustrated in Figure 2.3, if $\rho > 1$, the hazard function is increasing with time; if $\rho < 1$, the hazard function is decreasing with time. If $\rho = 1$, the hazard function is constant and the Weibull model is reduced to the exponential model $h(t) = \lambda$.

The parameter λ is called the scale parameter and influences the statistical dispersion of the underlying probability distribution. If λ is large, the distribution will be more spread out and conversely if λ is small, the distribution will be more concentrated.

The Weibull proportional hazards model is defined by reparameterising the scale parameter λ

$$h(t|\boldsymbol{X}) = \exp(\beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k)\rho t^{\rho-1} \tag{2.10}$$

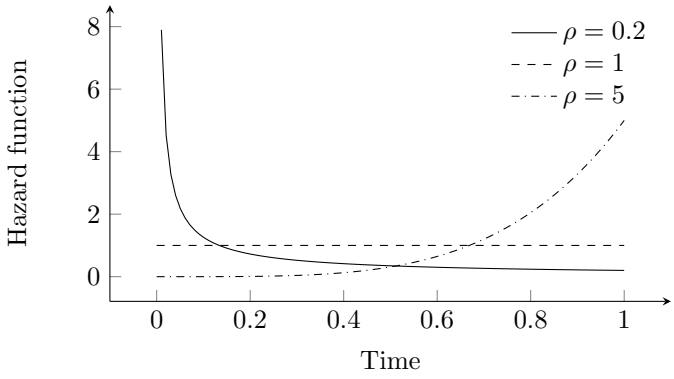

Figure 2.3: Hazard functions for Weibull distributed survival times. Hazard functions with different values for ρ are depicted with $\lambda = 1$.

The Weibull proportional hazards model can also be expressed as in (2.7)

$$h(t|\boldsymbol{X}) = \exp(\beta_0)\rho t^{\rho-1} \exp(X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k) \tag{2.11}$$

where $\exp(\beta_0)\rho t^{\rho-1}$ may be regarded as $h_0(t)$.

2.2.1.1 Estimation of the Weibull proportional hazards model

As the Weibull proportional hazards model is fully parametric, the model parameters are estimated by maximising the likelihood function. Right censored survival data consists of a combination of subjects that experience an event and subjects that are right censored, and as a result [7], the likelihood function for a sample with n subjects is given by

$$\mathcal{L}_n(\boldsymbol{\alpha}|\boldsymbol{T}, \boldsymbol{\delta}) = \prod_{i=1}^{n} (f(T_i))^{\delta_i} (S(T_i))^{1-\delta_i} \tag{2.12}$$

where $\boldsymbol{\alpha}$ is the parameter vector of interest. Utilising that the probability density function, $f(t)$, can be expressed as a product of the survival function and the hazard function (2.1), the likelihood function can be rewritten as

$$\mathcal{L}_n(\boldsymbol{\alpha}|\boldsymbol{T}, \boldsymbol{\delta}) = \prod_{i=1}^{n} (h(T_i))^{\delta_i} S(T_i) \tag{2.13}$$

For Weibull distributed event times, the likelihood function is

$$\mathcal{L}_n(\boldsymbol{\beta}, \rho|\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{X}) = \prod_{i=1}^{n} (\rho T_i^{\rho-1} \exp(X_i\boldsymbol{\beta}))^{\delta_i} \exp(-T_i^{\rho} \exp(X_i\boldsymbol{\beta})) \tag{2.14}$$

where $X_i = (1, X_{i1}, \ldots, X_{ik})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$. The log-likelihood function $\ell_n(\boldsymbol{\alpha}) = \log \mathcal{L}_n(\boldsymbol{\alpha})$, is typically easier to work with than the likelihood function itself, and it is therefore used in the maximisation process. It makes no difference since maximising the log-likelihood gives the same estimates as maximising the likelihood. The log-likelihood function is maximised in an iteratively manner.

2.2.2 Cox proportional hazards model

The baseline hazard function in the proportional hazards model (2.7) can be left unspecified, which will result in a semi-parametric proportional hazards model. The most widely used semi-parametric survival model is the Cox proportional hazards model, which is given as

$$h(t|\boldsymbol{X}) = h_0(t) \exp(X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k) \tag{2.15}$$

The model is semi-parametric, because while no knowledge of the baseline hazard function, $h_0(t)$, is required, the covariates still enter the model linearly on the log hazard scale.

2.2.2.1 Estimation of the Cox proportional hazards model

Since the Cox proportional hazards model is semi-parametric, the model parameters cannot be estimated using the likelihood function in (2.13). Instead, one must use the method of partial likelihood, developed by David R. Cox in 1972 [6]. The partial likelihood function is given by

$$\mathcal{L}_n(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{T}) = \prod_{i=1}^{r} \frac{\exp(X_i\boldsymbol{\beta})}{\sum_{T_j \geq T_i} \exp(X_j\boldsymbol{\beta})} \tag{2.16}$$

where r is the number of events, $X_i = (X_{i1}, \ldots, X_{ik})$, and $X_j = (X_{j1}, \ldots, X_{jk})$. The likelihood is called partial, as only the probabilities of subjects that experience an event are considered [27]. The estimates obtained by the partial

likelihood are consistent and asymptotic normal [9, 12], and found by maximising the partial log-likelihood in an iteratively manner. The partial likelihood is valid when no subjects have the same event time (no ties). If this is not the case, the Efron approximation [8, 52] to the partial likelihood may be applied.

2.2.3 Interpretation of the hazard ratio

The covariate effects obtained by the proportional hazards model are interpreted by means of the hazard ratios. The hazard ratio is given as the ratio of the hazard rates of two subjects with different levels of the covariate in question. The hazard ratio (HR) is given as

$$\text{HR} = \frac{h(t|X_i)}{h(t|X_j)} = \frac{h_0(t)\exp(X_i\boldsymbol{\beta})}{h_0(t)\exp(X_j\boldsymbol{\beta})} = \frac{\exp(X_i\boldsymbol{\beta})}{\exp(X_j\boldsymbol{\beta})} = \exp((X_i - X_j)\boldsymbol{\beta}) \tag{2.17}$$

As seen, the hazard ratio is independent of time t and thus constant, i.e. the hazard rates are proportional. If there is only one covariate X, which is binary and $X_i = 1$ and $X_j = 0$, the hazard ratio is given as

$$\text{HR} = \frac{h(t|X_i = 1)}{h(t|X_j = 0)} = \frac{h_0(t)\exp(\beta)}{h_0(t)} = \exp(\beta) \tag{2.18}$$

If the covariate is continuous rather than categorical, the hazard ratio states the effect of increasing the level of the covariate by one unit. The interpretation of the hazard ratio is the same for parametric and semi-parametric proportional hazards models. The hazard ratio acts multiplicative on the baseline hazard, thus a hazard ratio of 0.5 reduces the hazard rate by 50%, while a hazard ratio of 1.5 increases the hazard rate by 50%. If the hazard ratio is equal to one, the covariate in question has no effect on the hazard rate. The significance of the covariate effects can be evaluated using hypothesis testing (typically based on the Wald statistic) and a confidence interval for the hazard ratio is also easily constructed.

2.2.4 Checking model assumptions

A key assumption of the proportional hazards model is proportional hazards, which means that the hazard ratio of two subjects is constant and the covariate effects are independent of time. The appropriateness of the proportional hazards assumption may be evaluated using different approaches:

- Log-log survival curves
- Checking the Scoenfeld residuals
- Interaction of covariates with time

Log-log survival curves It can be shown that

$$\begin{aligned} S(t) &= \exp\Big(-\int_0^t h(u|\boldsymbol{X})\,\mathrm{d}u\Big) \\ &= \exp\Big(-\int_0^t h_0(u)\exp(\boldsymbol{X}\boldsymbol{\beta})\,\mathrm{d}u\Big) \\ &= \exp(-H_0(t)\exp(\boldsymbol{X}\boldsymbol{\beta})) \end{aligned} \tag{2.19}$$

where $H_0(t) = \int_0^t h_0(u)\,\mathrm{d}u$. In consequence

$$\begin{aligned} \log(S(t)) &= -H_0(t)\exp(\boldsymbol{X}\boldsymbol{\beta}) \quad \Leftrightarrow \\ \log(-\log(S(t))) &= \log(-H_0(t)) + \boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{2.20}$$

Thus, the proportional hazards assumption may be evaluated by visualising the log-log survival curves of the different levels of the covariates. The survival curves may be estimated using the non-parametric Kaplan-Meier method [23, 29]. If the covariates are continuous, they will need to be categorised into an appropriate number of groups. If the log-log survival curves are approximately parallel when plotted on the log-log scale, the proportional hazards assumption is satisfied.

Checking the Schoenfeld residuals Another way of checking the proportional hazards assumption is by means of the scaled Schoenfeld residuals. It can be shown, that if the proportional hazards assumption for a given covariate holds, then the Schoenfeld residuals for that covariate will be independent of the survival time. The method is elaborated by Therneau and Grambsch in [20] and [52].

Interaction of covariates with time The proportional hazards assumption can also be evaluated by extending the proportional hazards model and including an interaction term involving the covariate being assessed and some function of time [27]. The proportional hazards assumption is then evaluated by testing the significance of the interaction term. The significance may be tested using a Wald test or a likelihood ratio test.

Weibull assumption

In addition to the proportional hazards assumption, the Weibull proportional hazards model is based on the assumption that the survival times are Weibull distributed. This assumption can also be evaluated by means of the log-log survival curves. The Weibull survival function is given by

$$S(t) = \exp(-\lambda t^{\rho}) \tag{2.21}$$

And thus

$$\log(-\log(S(t))) = \log(\lambda) + \rho \log(t) \tag{2.22}$$

From (2.22), it is seen that the $\log(-\log(S(t)))$ is a linear function of $\log(t)$. This means, that if the log-log survival curves are reasonably straight, the Weibull assumption holds. If the log-log survival curves are parallel but not straight, it means the proportional hazards assumption holds, but the Weibull assumption does not and vice versa. If the Weibull assumption holds and the proportional hazards assumption does not, this indicates that the shape parameter ρ cannot assumed to be constant in the Weibull proportional hazards model (2.10)[27]. In Kleinbaum and Klein (2005) a method for modeling ρ is presented.

Linearity of continuous covariates

The linearity of continuous covariates may be checked by visual inspection of the exposure-response relationship between the covariate in question and the log relative hazard or by adding higher-order terms of the covariate and checking their significance. Furthermore, the linearity may be checked by comparing the model fit to that of a more flexible model, e.g. a spline model [52].

2.3 Analysis of clustered survival data

In the following sections, different statistical methods for analysis of clustered survival data are presented. Only methods based on the proportional hazards model will be considered. First, the fixed effects model and the semi-parametric stratified model will be presented. The two methods are computationally simple, but have some major drawbacks [7]. Then, the shared frailty model will be presented. In the shared frailty model all subjects in a cluster are assumed to share a random cluster effect, which impacts the interpretation of the hazard ratio. Finally, the marginal model and the copula model are presented. The marginal model is an independence working model, which means that the estimates are obtained by assuming all subjects are independent. The estimated parameter variance of the estimates are subsequently adjusted according to the correlation between subjects. The copula model can be used to combine the marginal survival functions of different subjects in a cluster and thereby generate the joint survival function.

2.3.1 Fixed effects model

The clustering of data may be modeled by introducing a fixed effect for each cluster in the proportional hazards model

$$h_{ij}(t|X_{ij}) = h_0(t)\exp(X_{ij}\boldsymbol{\beta} + c_i) \tag{2.23}$$

where $h_{ij}(t)$ is the conditional hazard function for subject j in cluster i, c_i is the fixed effect for cluster i, and $X_{ij} = (X_{ij1}, \ldots, X_{ijk})$. The introduction of a fixed cluster effect results in a loss of degrees of freedom, and to avoid an overparameterised model, a restriction is necessary, e.g. $c_1 = 0$. As a result, the fixed effects of all other clusters are to be compared to cluster 1, which complicates the model interpretation. In addition, since the number of subjects in each cluster is likely to be limited, the standard errors of the fixed effects will be very large. Moreover, there may be some clusters with no events and only censored observations, where the hazard is difficult to estimate [7]. However, the fixed effect estimates and their standard errors are of secondary interest, as the fixed effect is introduced in order to adjust for the clustering and not to estimate the fixed cluster effects per se.

2.3.2 Semi-parametric stratified model

In the semi-parametric stratified model, each cluster is allowed to have its own unspecified baseline hazard

$$h_{ij}(t|X_{ij}) = h_{i0}(t)\exp(X_{ij}\boldsymbol{\beta}) \tag{2.24}$$

where $h_{ij}(t)$ is the conditional hazard function for subject j in cluster i, h_{i0} is the baseline hazard for cluster i, and $X_{ij} = (X_{ij1}, \ldots, X_{ijk})$. The partial likelihood function for the stratified model is

$$\mathcal{L}_n(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{\delta}) = \prod_{i=1}^{s}\prod_{j=1}^{n_i}\left(\frac{\exp(X_{ij}\boldsymbol{\beta})}{\sum_{T_{il}\geq T_{ij}}\exp(X_{il}\boldsymbol{\beta})}\right)^{\delta_{ij}} \tag{2.25}$$

where s is the number of clusters, n_i is the number of subjects in the ith cluster. A cluster only contributes to the partial likelihood if an event for a subject occurs while at least one other subject is still at risk. Furthermore, a cluster where all subjects have the same covariate information do not contribute to the partial likelihood [7, 23].

2.3.3 Shared frailty model

Another way of managing clustering of data is by assuming that there is an unobserved risk factor, a so-called frailty, which is shared by all subjects in a cluster. The frailty accounts for the between-group variability and simultaneously induces a dependence within clusters [23]. The shared frailty model is defined as

$$h_{ij}(t|X_{ij}) = h_0(t)u_i\exp(X_{ij}\boldsymbol{\beta}) \tag{2.26}$$

where $h_{ij}(t|X_{ij})$ is the conditional hazard function for subject j in cluster i, and u_i is the frailty of cluster i. The frailty is considered random and constant over time and acts multiplicatively on the baseline hazard function. Given the values of the frailties, the subjects are assumed to be independent, thus the shared frailty model is a conditional independence model [23]. The frailty is random, because focus is not on each cluster as such, but on the population of clusters [23].

2.3.3.1 Choice of frailty distribution

The dependence structure in the clustered data is described by the frailties, which all follow the same distribution. The most common choice of distribution is the one-parameter gamma distribution [7, 23] with the density function

$$f_U(u) = \frac{u^{1/\theta-1}\exp(-u/\theta)}{\theta^{\frac{1}{\theta}}\Gamma(1/\theta)}, \qquad \theta > 0 \tag{2.27}$$

with mean 1 and variance θ, where the latter of the two provides information on the variability in the population of clusters. As θ approaches 0, there is no heterogeneity between clusters and no dependence between subjects in a cluster. The gamma distribution is mathematically convenient, as the gamma distributed frailties can be integrated out from the conditional likelihood, which is used to estimate the model parameters.

2.3.3.2 Estimation of the shared frailty model

The estimation of the model parameters in the shared frailty model depends on whether the baseline hazard is assumed to have a particular parametric form or is left unspecified. Just like any of the other model parameters, the significance of the parameter θ can be evaluated using a Wald or a likelihood ratio test. According to Therneau and Gramsch (2000) and Therneau *et al.* (2003), the likelihood ratio test is preferable.

Parametric baseline If the survival times are assumed to follow the Weibull distribution, the conditional likelihood for the ith cluster is according to (2.14) given as

$$\mathcal{L}_i(\boldsymbol{\beta}, \rho | T_i, \delta_i, X_i, u_i) = \prod_{j=1}^{n_i} (\rho T_{ij}^{\rho-1} u_i \exp(X_{ij}\boldsymbol{\beta}))^{\delta_{ij}} \exp(-T_{ij}^{\rho} u_i \exp(X_{ij}\boldsymbol{\beta})) \tag{2.28}$$

The gamma distributed frailties can be integrated out to obtain the marginal likelihood [7] given as

$$\mathcal{L}_{marg,i}(\boldsymbol{\beta}, \theta, \rho | T_i, \delta_i, X_i, e_i) = \frac{\Gamma(e_i + 1/\theta) \prod_{j=1}^{n_i} (\rho T_{ij}^{\rho-1} \exp(X_{ij}\boldsymbol{\beta}))^{\delta_{ij}}}{\left(1/\theta + \sum_{j=1}^{n_i} T_{ij}^{\rho} \exp(X_{ij}\boldsymbol{\beta})\right)^{1/\theta+d_i} \theta^{1/\theta}\Gamma(1/\theta)} \tag{2.29}$$

where e_i is the number of events in the ith cluster. By taking the logarithm of this expression and summing over all clusters, the marginal log-likelihood function is obtained [7], and the model parameters can be estimated by maximisation.

Unspecified baseline If the shared frailty model is semi-parametric and based on the Cox proportional hazards model, the model parameters can be estimated using either partial likelihood ideas in combination with the expectation-maximisation (EM) algorithm [26] or the penalised partial likelihood. If the frailties follow a gamma distribution, the two approaches lead to the same solution [7, 52, 53]. In this study, focus is on the penalised partial likelihood approach, which is implemented in the R function `coxph`.

The penalised partial log-likelihood function can be written as a sum of two parts

$$\ell_{ppl}(\boldsymbol{\zeta}, \boldsymbol{u}) = \ell_{part}(\boldsymbol{\beta}, \boldsymbol{u}) + \ell_{pen}(\boldsymbol{u}) \tag{2.30}$$

where $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \theta)$, and

$$\ell_{part}(\boldsymbol{\beta}, \boldsymbol{u}) = \log \prod_{i=1}^{s} \prod_{j=1}^{n_i} \left(\frac{u_i \exp(X_{ij}\boldsymbol{\beta})}{\sum_{T_{qw} > T_{ij}} u_q \exp(X_{qw}\boldsymbol{\beta})} \right)^{\delta_{ij}} \tag{2.31}$$

and

$$\ell_{pen}(\boldsymbol{u}) = \sum_{i=1}^{s} \log f_U(u_i) \tag{2.32}$$

The first part of the penalised log-likelihood, $\ell_{part}(\boldsymbol{\beta}, \boldsymbol{u})$, is the log of an usual Cox partial likelihood, where the frailty u_i is treated as fixed. The second part of the likelihood is a penalty term, which will have a large negative contribution if the random effect is very different from its mean. The maximisation of the penalised partial log-likelihood consists of an inner and an outer loop. The parameter θ is held fixed in the inner loop, where the penalised partial log-likelihood function (2.30) is maximised to obtain an estimate of $\boldsymbol{\beta}$ and $\boldsymbol{u}$. In the outer loop, maximisation of the observable log-likelihood is used to obtain an estimate of θ [7, 52, 53]. The observable log-likelihood is given as

$$\begin{aligned} \ell(\theta) =& \ell_{ppl} + \sum_{i=1}^{s} 1/\theta - (1/\theta + e_i)\log(1/\theta + e_i) + \log(1/\theta)/\theta \\ &+ \log\left(\frac{\Gamma(1/\theta + e_i)}{\Gamma(1/\theta)} \right) \end{aligned} \tag{2.33}$$

where e_i is the number of events in the ith cluster. The maximisation process is conducted in an iteratively manner and can be quite computationally expensive and time-consuming [53].

2.3.3.3 The frailties

The clustering can either be considered a nuisance or an interesting aspect of the survival data. If the clustering is considered a nuisance, the shared frailty model may be used as a method of variance reduction [23]. If the clustering is considered an interesting aspect, the distribution of the individual frailties according to different cluster traits may be explored [52], for example by plotting the individual frailties against cluster size.

The interpretation of the individual frailties is similar to the interpretation of the hazard ratio; subjects in a cluster i with frailty $u_i > 1$ are frail, meaning they have a higher risk of experiencing the event of interest and subjects in a cluster k with frailty $u_k < 1$ are strong, they have a lower risk.

2.3.3.4 Conditional hazard ratio

For the shared frailty model, the hazard ratio is conditioned on the same level of frailty

$$\text{HR} = \frac{h_0(t)u_i \exp(X_{ij}\boldsymbol{\beta})}{h_0(t)u_m \exp(X_{mk}\boldsymbol{\beta})} = \exp((X_{ij} - X_{mk})\boldsymbol{\beta}) \qquad \text{only if } u_i = u_m \quad (2.34)$$

where u_i is the frailty of cluster i and u_m is the frailty of cluster m. Thus, the hazard ratio cannot be interpreted at population level as the proportional hazard assumption is not satisfied for the unconditional hazards. In the shared frailty model, the relative risks are estimated within clusters.

2.3.4 Marginal model

The marginal model is a so-called independence working model (IWM), which means that all subjects are assumed to be independent despite of the clustering. I.e. a marginal Cox proportional hazards model is identical to the model (2.15) on page 11, and the model parameters are estimated in the same way. Although the estimation of the model parameters is conducted without taking

the clustering of subjects into account, the estimators are consistent under a reasonable set of conditions [7, 51]. However, the information matrix obtained by the IWM is not a consistent estimator of the asymptotic variance-covariance matrix. An approximation to the grouped jackknife estimator [31, 32] is applied to obtain the robust variance estimator

$$\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{S}(\hat{\boldsymbol{\beta}})\boldsymbol{S}^{T}(\hat{\boldsymbol{\beta}})\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}) \tag{2.35}$$

where $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{S}(\hat{\boldsymbol{\beta}})$ are the information matrix and the score vector, respectively, of the IWM for all observations. The expression of the robust variance estimator is equivalent to the sandwich estimator [57].

2.3.5 Copula model

The copula model can be used to combine the marginal survival functions of different subjects in a cluster and thereby generate the joint survival function. The joint survival function is given as

$$S(t_1, \ldots, t_{n_i}) = C_{\boldsymbol{\theta}}\{S_1(t_1), \ldots, S_{n_i}(t_{n_i})\}, \quad t_1, \ldots, t_{n_i} \geq 0 \tag{2.36}$$

where $C_{\boldsymbol{\theta}}(v_1, \ldots, v_{n_i})$ is a n_i-dimensional copula function with parameter vector $\boldsymbol{\theta}$, defined for $(v_1, \ldots, v_{n_i}) \in [0,1]^{n_i}$ and taking values in $[0,1]$. The copula model enables modeling of the dependence between the marginals expressed in the parameter $\boldsymbol{\theta}$ [2].

2.3.5.1 The Clayton-Oakes copula

The family of Archimedean copulas are very popular and most often applied to multivariate survival data [7]. An Archimedean copula has the form

$$C_{\theta}(v_1, \ldots, v_{n_i}) = \phi_{\theta}(\phi_{\theta}^{-1}(v_1) + \ldots + \phi_{\theta}^{-1}(v_{n_i})) \tag{2.37}$$

where ϕ is a decreasing function defined on $[0, \infty]$, taking values in $[0,1]$ and satisfying $\phi(0) = 1$. The Archimedean copula family is popular, because the copulas are easily derived and are capable of capturing many kinds of dependence. For more details on Archimedean copulas see Genest and MacKay (1986), Nelsen (2006) and Trivedi and Zimmer (2007).

The Clayton-Oakes copula [5, 42] from the family of Archimedean copulas is suitable for correlated survival times [55]. The Clayton-Oakes copula is given as

$$S(t_1,\ldots,t_{n_i}) = \{S_1^{-\theta}(t_1) + \ldots + S_{n_i}^{-\theta}(t_{n_i}) - (n_i - 1)\}^{-\frac{1}{\theta}}, \quad \theta > 0 \tag{2.38}$$

where the parameter θ measures the dependence between the marginal survival functions. As θ approaches 0, the marginal survival functions become independent, and for $\theta > 0$, the survival times are positively correlated. As a rule, the Clayton-Oakes copula cannot account for negative dependence [55], however, a bivariate Clayton-Oakes copula may be extended to present a negative dependence [11, 41]. In this case

$$S(t_1,t_2) = \left(\max\left\{S_1^{-\theta}(t_1) + S_2^{-\theta}(t_2) - 1\right\}, 0\right)^{-\frac{1}{\theta}}, \quad \theta \in [-1,\infty) \setminus \{0\} \tag{2.39}$$

An example of negative dependence is in transplantation studies, where it has been shown, that the longer an individual has to wait for a transplant, the shorter the survival time after the transplantation [58].

2.3.5.2 Estimation of the Clayton-Oakes copula model

The estimation of the model parameters in the Clayton-Oakes copula model is a two-stage procedure. The model parameters in the marginal model are estimated first and the variance of the parameter estimates adjusted by taking the clustering of the subjects into account. In the subsequent estimation of the association parameter θ, the estimates from the marginal model are regarded as fixed. The estimation of the association parameter θ depends on whether the baseline hazard is assumed to have a particular parametric form or is left unspecified [2, 50]. In the following, focus is on bivariate survival data for ease of notation, however all methods can be used for clusters of varying size. For bivariate survival data, the Clayton-Oakes copula is given by

$$S(t_1,t_2) = \left\{S_1^{-\theta}(t_1) + S_2^{-\theta}(t_2) - 1\right\}^{-\frac{1}{\theta}}, \qquad \theta > 0 \tag{2.40}$$

And the corresponding likelihood function

$$\begin{aligned}\mathcal{L} = &\prod_{i=1}^{s} \left(f(t_{i1},t_{i2})\right)^{\delta_{i1}\delta_{i2}} \left(-\frac{\partial S(t_{i1},t_{i2})}{\partial t_{i1}}\right)^{\delta_{i1}(1-\delta_{i2})} \\ &\times \left(-\frac{\partial S(t_{i1},t_{i2})}{\partial t_{i2}}\right)^{(1-\delta_{i1})\delta_{i2}} \left(S(t_{i1},t_{i2})\right)^{(1-\delta_{i1})(1-\delta_{i2})}\end{aligned} \tag{2.41}$$

Parametric baseline If the survival times are assumed to follow a given distribution, the association parameter θ is estimated by maximisation of the likelihood function given in (2.41), where the estimated parameters from the marginal model have been plugged in. If the baseline is parametric, the parameters $\boldsymbol{\beta}$ and θ may also be estimated simultaneously [17], but the computations quickly become very complicated [2].

Unspecified baseline If the baseline is left unspecified and the marginals are modeled by a Cox proportional hazards model, it is necessary to estimate the baseline hazard function $h_0(t)$ in addition to the model parameters $\boldsymbol{\beta}$ in order to estimate the association parameter θ. The baseline hazard function $h_0(t)$ is estimated by means of an Aalen-Breslow type estimator [2, 51]. Using the estimated parameters $\hat{\boldsymbol{\beta}}$ and the estimated baseline hazard function $\hat{h}_0(t)$ in the likelihood function (2.41), a pseudo log-likelihood is obtained. The association parameter is then estimated by maximisation of this pseudo log-likelihood [2, 15, 36]. The pseudo log-likelihood is given as

$$\ell(\theta) = \frac{1}{s}\left(\sum_{i=1}^{s} \int_0^{\tau} \log\left(1 + \theta^{-1} N_{\cdot i}(t-)\right) dN_{\cdot i}(t) + \sum_{i=1}^{s} \sum_{j=i}^{n_i} \theta^{-1} N_{ji}(\tau) \hat{G}_{ji} - \sum_{i=1}^{s} (\theta + N_{\cdot i}(\tau)) \log(\hat{R}_i(\theta)) \right) \tag{2.42}$$

where

$$\hat{G}_{ji} = \int_0^{\tau} Y_{ji}(t) \exp(X_{ji}\hat{\boldsymbol{\beta}})\, \mathrm{d}\hat{H}_0(t), \quad \hat{R}_i = 1 + \sum_{j=1}^{n_i} (\exp(\theta^{-1}\hat{G}_{ji}) - 1)$$

2.3.5.3 The Clayton-Oakes copula and the shared gamma frailty model

For bivariate survival data, the joint survival function derived from the shared frailty model described in Section 2.3.3 becomes [7]

$$S_f(t_1, t_2) = \left(S_{1,f}^{-\theta}(t_1) + S_{2,f}^{-\theta}(t_2) - 1 \right)^{-\frac{1}{\theta}} \tag{2.43}$$

This function looks very similar to the joint survival function of the Clayton-Oakes copula, which may written as

$$S_c(t_1, t_2) = \left(S_{1,c}^{-\theta}(t_1) + S_{2,c}^{-\theta}(t_2) - 1 \right)^{-\frac{1}{\theta}} \tag{2.44}$$

Indeed, the joint survival function derived from the shared frailty model is also a Clayton-Oakes copula. However, even though the functional forms of the joint survival functions are identical, the two models are not equivalent and do not lead to the same parameter estimates because the marginal survival functions are modelled differently [7, 17]. The difference is that the survival function $S_{i,f}(t_i)$ from the conditional shared frailty model is also a function of the parameter θ, whereas the survival function $S_{i,c}(t_i)$ from the Clayton-Oakes copula is not, cf. Section 2.3.3.2 and Section 2.3.5.2.

2.4 Measure of dependence

The shared frailty model and the Clayton-Oakes copula model described in Section 2.3.3 and in Section 2.3.5, respectively, can both be applied for estimation of the degree of dependence between clustered survival data. The shared frailty model estimates the variance θ_f of the frailties, and the copula model estimates the association parameter θ_c. The two parameters are somewhat similar due to the concordance of the two models (cf. previous section).

The parameters θ_f and θ_c can be considered as measures of correlation of the survival times within clusters. For both models applies that if θ approaches 0, there is no dependence between the survival times of the subjects in a cluster and if θ is greater than 0, the survival times are positively correlated. As mentioned, the bivariate Clayton-Oakes copula may be extended to present a negative dependence, in this case the survival times are negatively correlated if $\theta < 0$. For negatively correlated survival times, there is no frailty interpretation [23].

Just like any of the other model parameters, the significance of the parameter θ can be evaluated using a Wald or a likelihood ratio test. According to Therneau and Gramsch (2000) and Therneau *et al.* (2003), the likelihood ratio test is preferable.

2.4.0.4 Kendall's τ

The degree of dependence of bivariate survival data may be evaluated using Kendall's τ [10, 52], which gives a generalised measure of the correlation between the survival times. For the shared frailty model, Kendall's τ is given as

$$\tau = \frac{\theta_f}{(\theta_f + 2)}, \qquad \theta_f > 0 \tag{2.45}$$

As seen, Kendall's τ will be between 0 and 1. For the extended bivariate Clayton-Oakes copula, Kendall's τ is given by the same formula but defined on a wider interval

$$\tau = \frac{\theta_c}{(\theta_c + 2)}, \qquad \theta_c \in [-1, \infty) \setminus \{0\} \tag{2.46}$$

Here, Kendall's τ will be between -1 and 1.

Chapter 3

Materials & Methods

The purpose of this study is to investigate and compare different statistical methods for analysis of clustered survival data by means of data from a large Danish register-based family study of the psychological late effects of exposure to childhood cancer. First, however, the statistical methods are explored using three smaller data sets, which are available through the statistical software R [48]. In this chapter, the three small data sets as well as the data from the register-based family study are presented. In addition, it will be described how the statistical analyses have been conducted.

3.1 Data

The three small data sets, which are available through different R packages are described in the following section, after which the data from the register-based family study of the late effects of childhood cancer is presented.

3.1.1 Data sets available through R

The three data sets are

- The diabetic retinopathy data (`timereg` package)
- The kidney catheter data (`survival` package)
- The NCCTG lung cancer data (`survival` package)

3.1.1.1 Diabetic retinopathy data

The diabetic retinopathy data was collected in order to test the effect of laser treatment for delaying blindness in patients with diabetic retinopathy, which is a complication associated with diabetes. The data set available through R consists of the subset of 197 patients defined in Huster *et al.* (1989). The 197 patients all had laser treatment on a randomly selected eye, while the other eye was observed without treatment. The patients were then followed over several years for observation of blindness. The clusters in this data set are the 197 patients. In addition to the treatment variable, the variable defining juvenile versus adult onset of the disease (younger or older than 20, respectively) is included. The diabetic retinopathy data have previously been analysed, i.a., by Huster *et al.* (1989), Lee *et al.* (1992), Therneau and Grambsch (2000) and Martinussen and Scheike (2006).

The cumulative incidence of blindness for the two treatment groups and for patients with juvenile and adult onset, respectively, of the disease is shown in Figure 3.1. The overall incidence rates are listed in Table 3.1. Laser treatment seems to have a positive effect with regard to delaying blindness, which is most pronounced in patients with adult onset of the disease.

Table 3.1: Incidence rates by treatment group and disease onset (diabetic retinopathy data).

	Person-years	No.	Events	Rate per 10 person-years
No treatment, juvenile onset	316.96	114	51	1.61
No treatment, adult onset	213.61	83	50	2.34
Treatment, juvenile onset	346.85	114	36	1.04
Treatment, adult onset	291.79	83	18	0.62

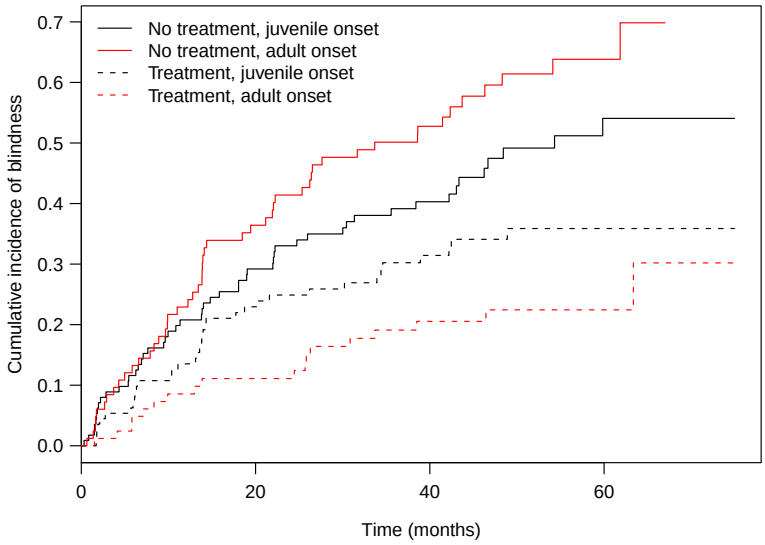

Figure 3.1: Cumulative incidence of blindness for treatment group and disease onset (diabetic retinopathy data).

3.1.1.2 Kidney catheter data

One of the most common complications in kidney patients using portable dialysis equipment is recurrent infection of the catheter, which is inserted in order to lead the blood from the patient to the dialysis equipment. When an infection occurs, the catheter is removed, the infection cleared, and the catheter reinserted [37]. The kidney data consists of time to recurrence of infection in 38 kidney patients. The variables included in the analysis are gender and age (continuous). The recurrence time is censored, if the catheter is removed for reasons other than infection. Two recurrence times (some of which may be censored) are measured for each patient, thus the clusters are the 38 patients. The kidney catheter data have previously been analysed, i.a., by McGilchrist and Aisbett (1991), Hougaard (2000), and Therneau and Grambsch (2000).

The cumulative incidence of recurrent infection for males and females is shown in Figure 3.2. The overall incidence rates are listed in Table 3.2. In addition, the overall incidence rates for defined age groups have been calculated. These are listed in Table 3.3. It looks like there is an increased risk of recurrent infection for males, while the effect of age is less clear.

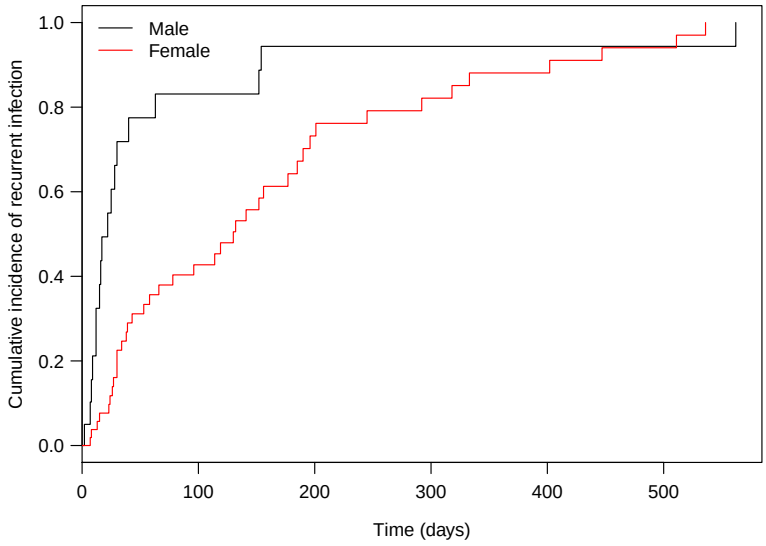

Figure 3.2: Cumulative incidence of recurrent infection for males and females (kidney catheter data).

Table 3.2: Incidence rates by gender (kidney catheter data).

	Person-years	No.	Events	Rate
Male	3.25	20	18	5.54
Female	17.90	56	40	2.23

Table 3.3: Incidence rates by age group (kidney catheter data).

	Person-years	No.	Events	Rate
10-19	3.18	8	7	2.20
20-29	0.41	4	4	9.68
30-39	4.17	10	8	1.92
40-49	6.13	20	12	1.96
50-59	5.11	24	19	3.72
60-69	2.14	10	8	3.74

3.1.1.3 NCCTG lung cancer data

The North Central Cancer Treatment Group (NCCTG) lung cancer data consists of data on 228 patients from a study of prognostic variables in advanced lung cancer [33]. The 228 patients were enrolled at 18 different institutions, which here represent the clusters. After enrollment in the study, patients were followed for observation of death. The variables included in the analysis are gender, age (continuous), and ECOG performance score estimated by the physicians. The ECOG performance score is a measure of the patients well-being ranging from 0 (good) to 5 (dead). None of the patients in this data set have an ECOG score above 3. Patients missing information on any of the variables have been removed, and the number of patients reduced to 226. The NCCTG lung cancer data have previously been analysed, i.a., by Loprinzi *et al.* (1994) and Therneau and Grambsch (2000).

In Figure 3.3, a bar plot visualising the number of patients enrolled at each institution is shown. The number of patients enrolled ranges from 2 to 36. The median number of patients is 10.5 and the interquartile range is 11.8. In Figure 3.4, the cumulative incidence of death for males and female is shown. The overall incidence rates for male and females are listed in Table 3.4. In addition, the overall incidence rates for defined age groups and the ECOG score have been calculated. These are listed in Table 3.5 and Table 3.6. It looks like there is an increased risk of death for males, and not surprisingly that the risk of death increases with age and ECOG score.

Table 3.4: Incidence rates by gender (NCCTG lung cancer data).

	Person-years	No.	Events	Rate per 10 person-years
Male	105.92	136	110	10.39
Female	83.52	90	53	6.35

Table 3.5: Incidence rates by age group (NCCTG lung cancer data).

	Person-years	No.	Events	Rate per 10 person-years
30-39	1.30	2	0	0.00
40-49	16.33	18	11	6.73
50-59	56.91	63	45	7.91
60-69	72.63	87	61	8.40
70-79	40.39	52	42	10.40
80-89	1.88	4	4	21.24

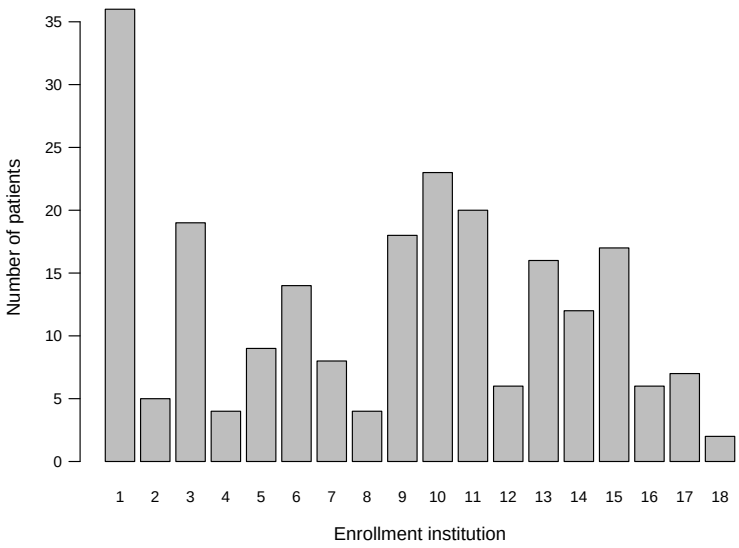

Figure 3.3: Number of patients at each enrollment institution (NCCTG lung cancer data).

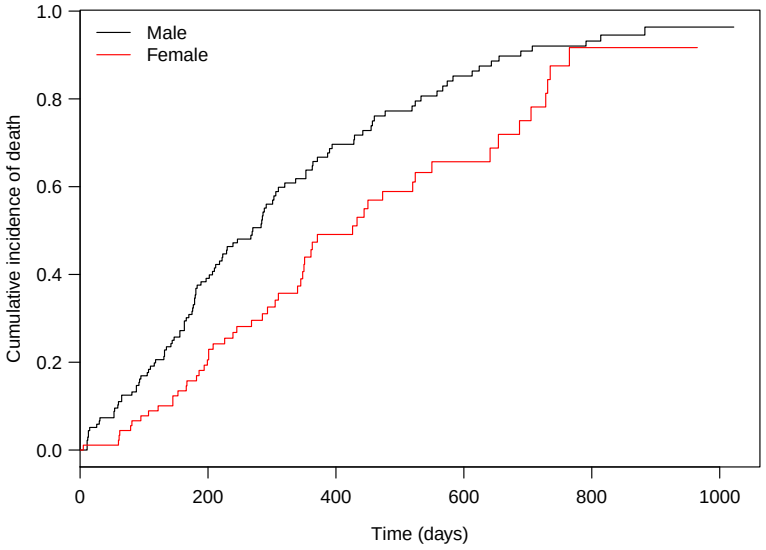

Figure 3.4: Cumulative incidence of death for males and females (NCCTG lung cancer data).

Table 3.6: Incidence rates by ECOG score (NCCTG lung cancer data).

ECOG score	Person-years	No.	Events	Rate per 10 person-years
0	60.69	63	37	6.10
1	97.28	113	82	8.43
2	31.14	49	43	13.81
3	0.32	1	1	30.95

3.1.2 Childhood cancer data

The childhood cancer data analysed in this study consists of data from a Danish register-based family study of psychological late effects in families exposed to childhood cancer. Research suggests that childhood cancer survivors, particular those with central nervous system tumors, have poor psychological health [21, 34, 60], and also siblings may suffer from from psychological distress [4]. The original cohort from the family study will be briefly described in the following section after which the analysed subset is presented.

3.1.2.1 Description of original cohort

The original cohort is based on 8561 children diagnosed with cancer in the period January 1 1975 to December 31 2009. The children were aged between 0 and 19 at diagnosis and identified through the Danish Cancer Register [13]. By means of the Danish Civil Registration System [44, 45], each of the children were matched on gender and age to twenty children without cancer (at date of diagnosis). In the following, the children with cancer and their match are named exposed and unexposed probands, respectively. The probands' (both exposed and unexposed) full siblings and half siblings born no later than December 31 2009 (end of study period) were identified based on the personal identification number of their parents through the Danish Civil Registration System and included in the cohort. The parents were also included in the cohort.

The Danish Civil Registration System was used to obtain date of death, disappearance and emigration (if any), and by linking the cohort to the Danish Psychiatric Central Research Register [40] admissions due to mental disorder were identified. The incidence admission was defined as any admission due to a mental disorder between inclusion date and December 31 2009 (end of study period). The inclusion date was date of diagnosis of the exposed proband. However, for unborn siblings (at date of diagnosis) inclusion date was date of birth.

Both individual and family-based left truncation were applied. Exposed and unexposed individuals admitted due to a mental disorder five years prior to the inclusion date were excluded from the study and their families marked. Furthermore, families where a family member (aged 0 to 19) had been diagnosed with cancer up to five years prior to the inclusion date, were excluded from the study. If the family in question was an exposed family, the twenty matched unexposed families were also excluded. The cohort was right censored in the event of death, disappearance, emigration or December 31 2009, whichever came first. In addition, unexposed families were right censored in the event of cancer in a family member aged 0 to 19 and hereafter entered as an exposed family.

3.1.2.2 Subset

The childhood cancer data analysed in this study consists of exposed and unexposed probands and their full brothers and sisters. The original twenty unexposed probands per exposed proband have been reduced to 5, which have been chosen so that the family size of the exposed and unexposed probands match. The family size is equal to the total number of identifiable family members; parents, full siblings and half siblings. Families with no full siblings have been excluded. The data consists of 56252 individuals in 24066 families. In Table 3.7, the distribution of individuals according to exposure group and relation is shown. In Table 3.8 and Table 3.9, the distribution of number of family members (sum of proband and full sibling(s)) and the distribution of number of events are shown. There were 3272 events.

Table 3.7: Distribution of individuals according to relation (childhood cancer data).

	Probands	Siblings	Total
Exposed	3987	5369	9356
Unexposed	20079	26817	46896

Table 3.8: Distribution of families according to number of family members (childhood cancer data).

	2	3	4	5
Exposed families	2662	1271	51	3
Unexposed families	13620	6184	271	4

In the statistical analysis, age is used as underlying time scale, and three vari-

Table 3.9: Distribution of families according to number of events (childhood cancer data).

	0	1	2	3
Exposed families	3485	469	32	1
Unexposed families	17614	2208	243	14

ables are included. The three variables are exposure group, relation (proband or sibling), and previous admission in family.

The cumulative incidence of admission for exposed and unexposed probands and siblings and for previous admission in family are shown in Figure 3.5 and Figure 3.6, respectively. The overall incidence rates are listed in Table 3.10 and Table 3.11. It looks like there is an increased risk of admission for exposed probands and for individuals with previous admission in the family.

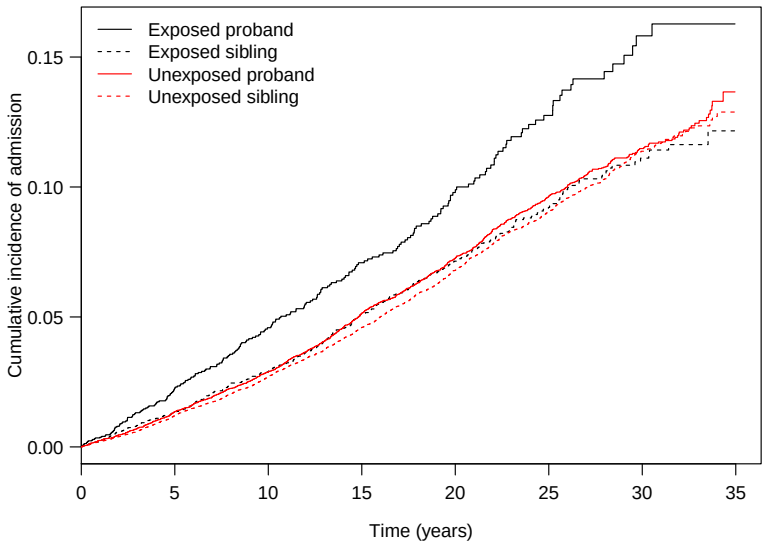

Figure 3.5: Cumulative incidence of admission for exposed and unexposed probands and siblings (childhood cancer data).

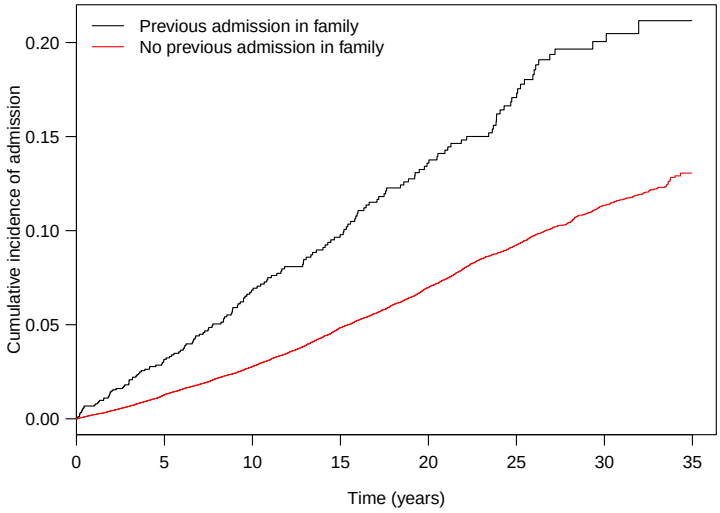

Figure 3.6: Cumulative incidence of admission for variable previous admission in family (childhood cancer data).

Table 3.10: Incidence rates by exposure group (childhood cancer data).

	Person-years	No.	Events	Rate per 10^3 person-years
Exposed proband	43236	3987	220	5.09
Exposed sibling	86513	5369	316	3.65
Unexposed proband	322557	20079	1205	3.74
Unexposed sibling	432550	26817	1531	3.54

Table 3.11: Incidence rates by previous admission in family (childhood cancer data).

	Person-years	No.	Events	Rate per 10^3 person-years
Previous admission	23494	1800	169	7.19
No previous admission	861362	54452	3103	3.60

3.2 Statistical analysis

The statistical analyses in this study have all been conducted using the statistical software R version 2.14.1 [48] by means of the two packages `survival` and `timereg`. As mentioned, only semi-parametric models based on the Cox proportional hazards model have been fitted to the data.

The function `coxph` in the package `survival` has been applied to fit the unadjusted Cox proportional hazards model, the fixed effects model, the stratified model, and the shared frailty model. The shared frailty model are fitted using the penalised partial likelihood approach [53].

The proportional hazard assumption has been tested by means of the Schoenfeld residuals using the function `cox.zph` also from the package `survival`. The linearity assumption has been tested by means of restricted cubic splines using the function `cph` in the `rms` package.

The function `cox.aalen` in the package `timereg` has been applied to fit the marginal Cox proportional hazards model, which is the first step of the estimation of the copula model. The function `two.stage` also from the package `timereg` has been applied to the second and final step of the estimation of the copula model.

With regard to the significance of the parameters θ_f and θ_c from the shared frailty model and the copula model, respectively, they are evaluated using a Wald test. The reason for this is, that it has not been possible to get an estimate of the log likelihood from the copula model using the `timereg` package.

The R code can be found in Appendix A.

Chapter 4

Results

In this chapter, the results of the statistical analyses are presented. First, the results from the analyses of the three data sets available through the statistical software R will be presented. Hereafter, the results from the analyses of the data from the Danish register-based family study of the psychological late effects of exposure to childhood cancer are presented.

4.1 Data sets available through R

4.1.1 Diabetic retinopathy data

The survival models that have been fitted to the diabetic retinopathy data are the Cox proportional hazards model unadjusted for any patient effect, the Cox proportional hazards model stratified by patient, the shared gamma frailty Cox proportional hazards model, and the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin.

It was not possible to fit a Cox proportional hazards model with patient as a fixed effect, as several patients did not experience blindness on either of their eyes. Thus the fixed effects model did not converge. Furthermore, the variable disease

onset in the diabetic retinopathy data was nested within patient, and thus it was not possible to estimate the effect of this variable using the Cox proportional hazards model stratified by patient. Therefore, in the Cox proportional hazards model stratified by patient, the only variable included is treatment.

4.1.1.1 Unadjusted Cox proportional hazards model

The Cox proportional hazard model unadjusted for any patient effect was first fitted to the diabetic retinopathy data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.1, where the term interaction covers the effect modification of adult onset of the disease on treatment. As seen, the effect of laser treatment and onset of disease are slightly insignificant based on a 5% significance level, while the interaction between treatment and adult onset of the disease is significant. The hazard ratios of the treatment groups with juvenile onset and adult onset, respectively, are shown in Table 4.2 and visualised in Figure 4.1. Generally seen, laser treatment reduces the hazard rate of blindness. The effect is most pronounced in patients with adult onset of the disease, where the hazard rate is reduced by approximately 60%.

Table 4.1: Estimates from unadjusted Cox proportional hazards model (diabetic retinopathy data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
No treatment			1.000	
Treatment	-0.425	0.218	0.654 (0.427-1.002)	0.051
Juvenile onset			1.000	
Adult onset	0.341	0.199	1.407 (0.952-2.079)	0.087
Interaction	-0.846	0.351	0.429 (0.216-0.853)	0.016

Table 4.2: Hazard ratios from unadjusted Cox proportional hazards model (diabetic retinopathy data).

	HR	95% CI
No treatment, juvenile onset	1.000	
No treatment, adult onset	1.407	0.952-2.079
Treatment, juvenile onset	0.654	0.427-1.002
Treatment, adult onset	0.395	0.230-0.676

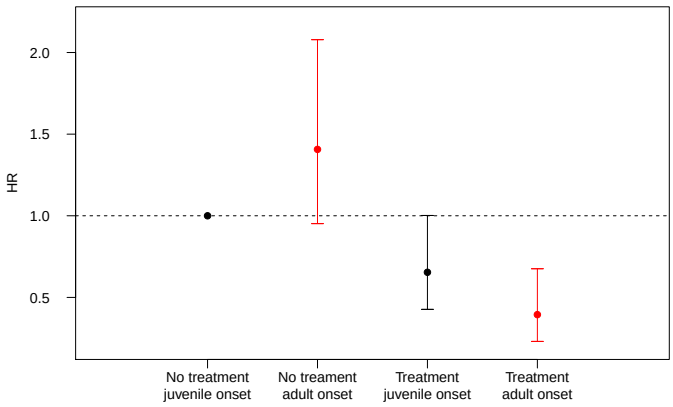

Figure 4.1: Visualisation of results from Cox proportional hazards model unadjusted for any patient effect (diabetic retinopathy data).

The proportional hazards assumption has been tested by means of the Schoenfeld residuals. There is no evidence against proportionality.

4.1.1.2 Stratified Cox proportional hazards model

Then, the Cox proportional hazard model stratified by patient was fitted to the data. The only variable included in the model was treatment group. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.3. The effect of laser treatment is highly significant (p-value $1.3 \cdot 10^{-6}$) and reduces the hazard rate of blindness by approximately 64%. As the unadjusted Cox proprotional hazards model shows, that the interaction between laser treatment and disease onset is significant, stratification by patient is not a satisfactory solution, when disease onset is nested within patient.

Table 4.3: Estimates from Cox proportional hazards model stratified by patient (diabetic retinopathy data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
No treatment			1.000	
Treatment	-1.030	0.213	0.357 (0.235-0.542)	$1.3 \cdot 10^{-6}$

4.1.1.3 Shared gamma frailty Cox proportional hazards model

Then, the Cox proportional hazards model with shared gamma-distributed frailties was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.4. As in the unadjusted model, the interaction between treatment and adult onset of the disease is significant (p-value $6.4{\cdot}10^{-3}$) and reduces the hazard rate of blindness considerably. The hazard ratios of the treatment groups with juvenile onset and adult onset, respectively, are shown in Table 4.5. The estimated effects are slightly different than the estimates from the Cox proportional hazards model unadjusted for any patient effect. However, the effects estimated in the shared gamma frailty Cox proportional hazards model are to be interpreted on patient level.

The estimate of the variance of the frailties θ_f is 0.93 corresponding to Kendall's τ_f equal to 0.32, thus there is on average a positive correlation of 0.32 between the time to blindness for the eyes of a patient. The random effect is significant with a p-value of $9.8{\cdot}10^{-3}$. The estimated individual frailties are visualised in a histogram in Figure 4.2.

Table 4.4: Estimates from shared gamma frailty Cox proportional hazards model (diabetic retinopathy data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
No treatment			1.000	
Treatment	-0.505	0.225	0.603 (0.388-0.938)	0.025
Juvenile onset			1.000	
Adult onset	0.397	0.259	1.488 (0.895-2.472)	0.130
Interaction	-0.986	0.362	0.373 (0.184-0.758)	$6.4{\cdot}10^{-3}$

Table 4.5: Hazard ratios from shared gamma frailty Cox proportional hazards model (diabetic retinopathy data).

	HR	95% CI
No treatment, juvenile onset	1.000	
No treatment, adult onset	1.488	0.895-2.472
Treatment, juvenile onset	0.603	0.388-0.938
Treatment, adult onset	0.335	0.180-0.623

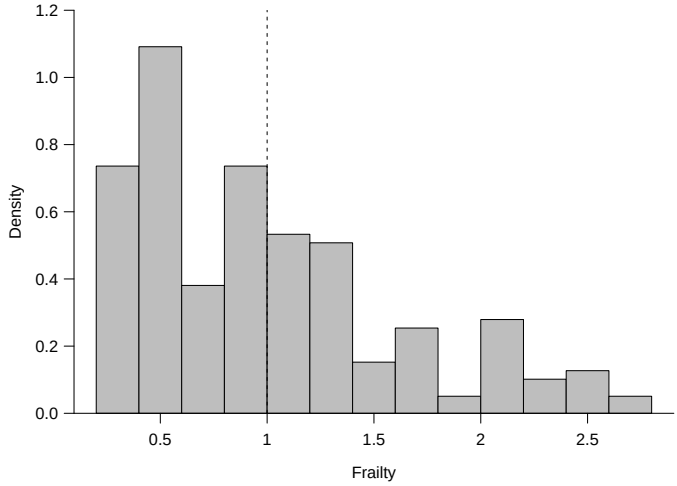

Figure 4.2: Histogram of estimated frailties (diabetic retinopathy data).

4.1.1.4 Clayton-Oakes copula model

Finally, the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.6. Again, the interaction between treatment and adult onset of disease is significant and reduces the hazard rate of blindness considerably. The hazard ratios of the treatment groups with juvenile onset and adult onset, respectively, are shown in Table 4.7. The estimated effects are identical to the estimates from the Cox proportional hazards model unadjusted for any patient effect. This is not surprising, since the effects in the marginal model are estimated using the IWM approach. The robust standard errors of the estimates are smaller, especially for the variable laser treament and the interaction term. This makes perfectly good sense, as both levels of the treatment variable are tested on each patient, i.e. the variable treatment is balanced within patients [52].

The association parameter θ_c estimated by the copula model is 1.07 corresponding to Kendall's τ_c 0.35, thus there is on average a positive correlation of 0.35 between the time to blindness for the eyes of a patient. The association parameter is significant with a p-value of $3.6 \cdot 10^{-3}$. Note, that the estimated association parameter θ_c is very similar to the estimated variance of the frailties θ_f in the shared frailty model.

Table 4.6: Estimates from margin of Clayton-Oakes copula model (diabetic retinopathy data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
No treatment			1.000	
Treatment	-0.425	0.185	0.654 (0.455-0.940)	0.022
Juvenile onset			1.000	
Adult onset	0.341	0.196	1.407 (0.958-2.064)	0.081
Interaction	-0.846	0.304	0.429 (0.237-0.778)	$5.3{\cdot}10^{-3}$

Table 4.7: Hazard ratios from margin of Clayton-Oakes copula model (diabetic retinopathy data).

	HR	95% CI
No treatment, juvenile onset	1.000	
No treatment, adult onset	1.407	0.958-2.064
Treatment, juvenile onset	0.654	0.455-0.940
Treatment, adult onset	0.395	0.242-0.644

4.1.1.5 Summary of results

The estimated covariate effects, standard errors, hazard ratios and p-values from the survival models fitted to the diabetic retinopathy data are summarised in Table 4.8. The stratified model is omitted, since it is not comparable to the other fitted models.

Table 4.8: Summary of results (diabetic retinopathy data).

		$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR	p-value	Additional parameters
Unadjusted model	Treatment	-0.425	0.218	0.654	0.051	
	Adult onset	0.341	0.199	1.407	0.087	
	Interaction	-0.846	0.351	0.429	0.016	
Shared frailty model	Treatment	-0.505	0.225	0.603	0.025	$\hat{\theta}_f = 0.93$ and Kendall's $\tau_f = 0.32$
	Adult onset	0.397	0.259	1.488	0.130	(p-value $9.8{\cdot}10^{-3}$)
	Interaction	-0.986	0.362	0.373	$6.4{\cdot}10^{-3}$	
Copula model	Treatment	-0.425	0.185	0.654	0.022	$\hat{\theta}_c = 1.07$ and Kendall's $\tau_c = 0.35$
	Adult onset	0.341	0.196	1.407	0.081	(p-value $3.6{\cdot}10^{-3}$)
	Interaction	-0.846	0.304	0.429	$5.3{\cdot}10^{-3}$	

4.1.2 Kidney catheter data

The survival models that have been fitted to the kidney catheter data are the Cox proportional hazards model unadjusted for any patient effect, the shared gamma frailty Cox proportional hazards model, and the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin.

It was not possible to fit a Cox proportional hazards model with patient as a fixed effect to the kidney data. Several patients did not experience recurrence of infection, and thus the fixed effects model did not converge. Furthermore, it was not possible to fit a Cox proportional hazards model stratified by patient to the kidney data. The variable gender was both nested within patients and the variable age only varied for 11 out of the 38 patients and only by one year.

4.1.2.1 Unadjusted Cox proportional hazards model

The Cox proportional hazard model unadjusted for any patient effect was first fitted to the kidney catheter data. There were no significant interaction between the variable gender and the variable age (tested by means of likelihood ratio tests). The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.9. As seen, the effect of gender is significant based on a 5% significance level (p-value $5.5{\cdot}10^{-3}$). The hazard rate of females is approximately 70% lower than the hazard rate of males. The effect of age is insignificant.

Table 4.9: Estimates from unadjusted Cox proportional hazards model (kidney catheter data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Male			1.000	
Female	-0.829	0.299	0.436 (0.243-0.784)	$5.5{\cdot}10^{-3}$
Age (per 10 years)	0.020	0.092	1.021 (0.851-1.223)	0.826

The proportional hazards assumption have been tested by means of the Schoenfeld residuals. With regard to the variable gender there is evidence against proportionality. However, this is because of an outlier, which will be elaborated in a moment. The linearity assumption of the variable age has been evaluated by means of a restricted cubic spline. In Figure 4.3, the relationship of age and the log relative hazard is shown. It looks okay.

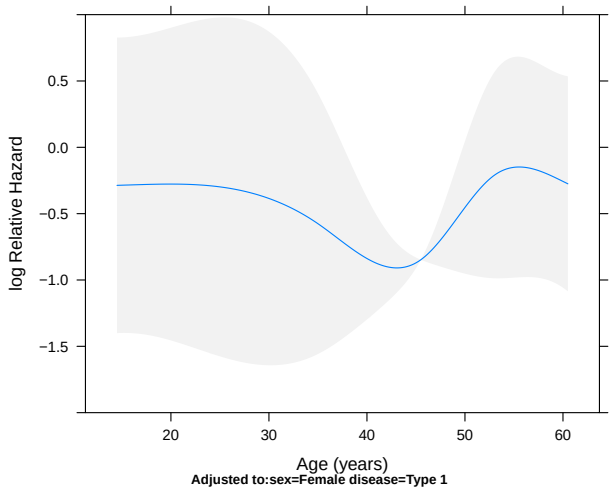

Figure 4.3: Relationship between age and incidence recurrent infection (kidney catheter data).

4.1.2.2 Shared gamma frailty Cox proportional hazards model

Then, the Cox proportional hazards model with shared gamma-distributed frailties was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.10. In this model, the effect of gender is significant (p-value $5.7 \cdot 10^{-4}$). The event rate of females is approximately 80% lower than the hazard rate of males. The estimated effects are different than the estimates from the Cox proportional hazards model unadjusted for any patient effect, especially the effect of gender. However, the effects estimated in the shared gamma frailty Cox proportional hazards model are to be interpreted on patient level.

The estimate of the variance of the frailties θ_f is 0.41 corresponding to Kendall's τ_f equal to 0.17, thus there is on average a positive correlation of 0.17 between the infection recurrence times. The random effect is significant with a p-value of 0.04. The estimated individual frailties are visualised in a histogram in Figure 4.4.

Table 4.10: Estimates from shared gamma frailty Cox proportional hazards model (kidney catheter data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Male			1.000	
Female	-1.587	0.461	0.204 (0.083-0.504)	$5.7{\cdot}10^{-4}$
Age (per 10 years)	0.052	0.119	1.054 (0.835-1.331)	0.660

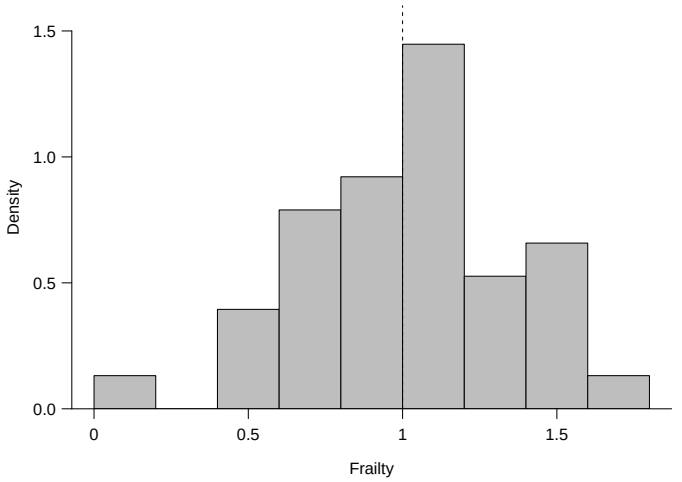

Figure 4.4: Histogram of estimated frailties (kidney catheter data).

4.1.2.3 Clayton-Oakes copula model

Finally, the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.11. In this model, the effect of both gender and age are insignificant (p-values 0.082 and 0.832, respectively). The association parameter θ_c estimated by copula model is 0.20 corresponding to Kendall's τ_c 0.09, thus there is on average a small positive correlation of 0.09 between the infection recurrence times. The association parameter is on the border of significance with a p-value of 0.056.

Not surprisingly, the estimated effects are approximately the same as the estimates from the Cox proportional hazards model unadjusted for any patient effect. The small discrepancies may be explained by the fact that the models are fitted using functions from two different R packages. The estimated robust standard error is smaller for the effect of age and larger for the effect of gender.

As emphasised by Therneau and Grambsch (2000), there is an outlier in the kidney catheter data. The patient with identification number 21, a 46-year-old male, had recurrence of infection at 152 and 562 days. As there are only 10 men in the study and their median time to recurrence of infection is 19.5 days, the male outlier most likely influences the effect and standard error of gender and the degree of dependence observed in the data in the model unadjusted for any patient effect and in the shared frailty model, respectively. However, because the variance of the estimates in the marginal Cox proportional hazards model are calculated using an approximation to the grouped jackknife technique, the influence of the outlier is reduced in Clayton-Oakes copula with the marginal Cox proportional hazards model as margins. Note, that the estimated association parameter θ_c is much smaller than the estimated variance of the frailties θ_f in the shared frailty model and insignificant.

Table 4.11: Estimates from margin of Clayton-Oakes copula model (kidney catheter data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Male			1.000	
Female	-0.843	0.485	0.430 (0.166-1.114)	0.082
Age (per 10 years)	0.017	0.082	1.018 (0.867-1.195)	0.832

The analyses presented here have been repeated without the patient with identification number 21. The parameter estimates from these analyses are summarised in Table B.1 on page 86 in Appendix B. Without this patient, neither the variance of frailties or the association parameter are significant. However, gender is significant in all models. Thus, if the patient with identification number 21 is included in the analysis, the frailty in the shared frailty Cox proportional hazards model will account for the specific characteristics of this patients, while the Clayton-Oakes copula model with marginal Cox proportional hazards model as margin will compensate with the association parameter and by increasing the robust standard errors.

4.1.2.4 Summary of results

The estimated covariate effects, standard errors, hazard ratios and p-values from the survival models fitted to the kidney catheter data are summarised in Table 4.12.

Table 4.12: Summary of results (kidney catheter data).

		$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR	p-value	Additional parameters
Unadjusted model	Female	-0.829	0.299	0.436	$5.5{\cdot}10^{-3}$	
	Age (per 10 years)	0.020	0.092	1.021	0.826	
Shared frailty model	Female	-1.587	0.461	0.204	$5.7{\cdot}10^{-4}$	$\theta_f = 0.41$ and Kendall's $\tau_f = 0.17$
	Age (per 10 years)	0.052	0.119	1.054	0.660	(p-value 0.040)
Copula model	Female	-0.843	0.485	0.430	0.082	$\theta_c = 0.20$ and Kendall's $\tau_c = 0.09$
	Age (per 10 years)	0.017	0.082	1.018	0.832	(p-value 0.056)

4.1.3 NCCTG lung cancer data

The survival models that have been fitted to the NCCTG lung cancer data are the Cox proportional hazards model unadjusted for any institution effect, the Cox proportional hazards model with institution as a fixed effect, the Cox proportional hazards model stratified by institution, the shared gamma frailty Cox proportional hazards model, and the Clayton-Oakes copula model with the marginal Cox proportional hazards model as margin.

4.1.3.1 Unadjusted Cox proportional hazards model

The Cox proportional hazard model unadjusted for any patient effect was first fitted to the NCCTG lung cancer data. There were no significant interaction between any of the included variables (tested by means of likelihood ratio tests). The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.13. As seen, the effect of gender and ECOG score are significant (p-values $9.3{\cdot}10^{-4}$ and $4.0{\cdot}10^{-5}$, respectively). The hazard rate of females are 43% lower than the hazard rate of males, while the hazard rate increases with increasing ECOG score. The effect of age is insignificant.

Table 4.13: Estimates from unadjusted Cox proprotional hazards model (NCCTG lung cancer data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Male			1.000	
Female	-0.557	0.168	0.573 (0.412-0.797)	$9.3{\cdot}10^{-4}$
Age (per 5 years)	0.056	0.046	1.058 (0.966-1.158)	0.225
ECOG score	0.469	0.114	1.599 (1.278-2.000)	$4.0{\cdot}10^{-5}$

The proportional hazards assumption has been tested by means of the Schoenfeld residuals. There is no evidence against proportionality. The linearity assumption has been tested by restricted cubic splines. In Figure 4.5 and Figure 4.6, the relationship between the log relative hazard and age and ECOG score, respectively, are shown. The relationships are linear.

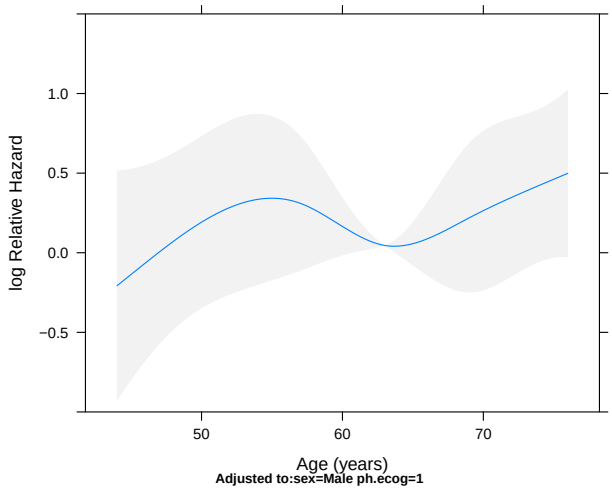

Figure 4.5: Relationship between age and incidence death (NCCTG lung cancer data).

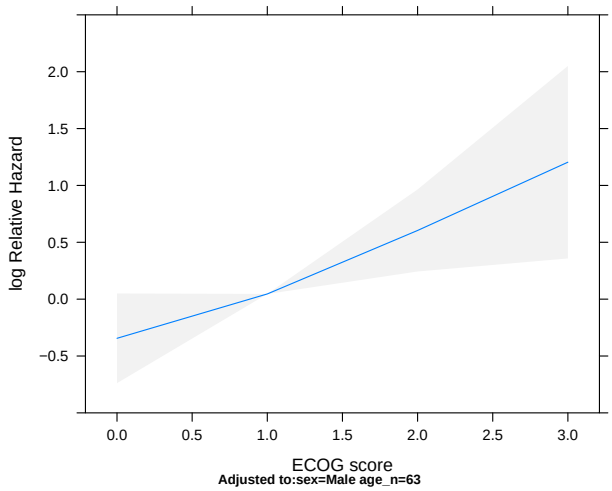

Figure 4.6: Relationship between ECOG score and incidence death (NCCTG lung cancer data).

4.1.3.2 Fixed effects Cox proportional hazards model

Then, the Cox proportional hazards model with enrollment institution as a fixed effect was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.14. As seen, the effects of gender and ECOG score are significant (p-values $1.0{\cdot}10^{-3}$ and $3.4{\cdot}10^{-6}$, respectively), while the effects of age and institution are not. Compared to the unadjusted Cox proportional hazards model, the estimates change a bit, when institution is included as a fixed effect. The overall significance of the institution variable has been tested in a likelihood ratio test and found to be insignificant (p-value 0.29). Some of the standard errors of the institution effects are quite large, since only a few patients are enrolled at these institutions.

Table 4.14: Estimates from Cox proportional model with institution as fixed effect (NCCTG lung cancer data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Male			1.000	
Female	-0.571	0.174	0.565 (0.402-0.795)	$1.0{\cdot}10^{-3}$
Age (per 5 years)	0.049	0.049	1.051 (0.955-1.156)	0.312
ECOG score	0.604	0.130	1.829 (1.418-2.359)	$3.4{\cdot}10^{-6}$
Institution 1			1.000	
Institution 2	0.530	0.546	1.699 (0.583-4.951)	0.331
Institution 3	-0.317	0.325	0.728 (0.386-1.376)	0.329
Institution 4	-0.464	0.539	0.629 (0.218-1.809)	0.390
Institution 5	0.034	0.456	1.034 (0.423-2.526)	0.941
Institution 6	0.029	0.348	1.030 (0.520-2.037)	0.933
Institution 7	-0.612	0.462	0.542 (0.219-1.342)	0.186
Institution 8	0.495	0.543	1.640 (0.566-4.754)	0.362
Institution 9	-0.547	0.360	0.578 (0.286-1.172)	0.128
Institution 10	-0.119	0.308	0.887 (0.485-1.623)	0.698
Institution 11	-0.522	0.351	0.593 (0.298-1.180)	0.137
Institution 12	-0.508	0.544	0.602 (0.207-1.746)	0.350
Institution 13	-0.718	0.362	0.488 (0.240-0.992)	0.047
Institution 14	0.334	0.376	1.397 (0.669-2.917)	0.374
Institution 15	-0.790	0.347	0.454 (0.230-0.896)	0.023
Institution 16	-0.940	0.740	0.391 (0.092-1.665)	0.204
Institution 17	-0.349	0.740	0.705 (0.165-3.006)	0.637
Institution 18	0.472	1.030	1.603 (0.213-12.071)	0.647

4.1.3.3 Stratified Cox proportional hazards model

Then, the Cox proportional hazards model stratified by enrollment institution was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.15. As before, the effect of gender and ECOG score are significant (p-values $2.6{\cdot}10^{-3}$ and $1.5{\cdot}10^{-5}$, respectively). The hazard rate of females is approximately 42% lower than the hazard rate of males, while the hazard rate increases 82% with increasing ECOG score. The effect of age is insignificant. The estimates are a bit different from the estimates from the unadjusted Cox proportional hazards model.

Table 4.15: Estimates from stratified Cox proportional hazards model (NCCTG lung cancer data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Male			1.000	
Female	-0.547	0.182	0.578 (0.405-0.826)	$2.6{\cdot}10^{-3}$
Age (per 5 years)	0.048	0.051	1.049 (0.948-1.160)	0.353
ECOG score	0.597	0.138	1.817 (1.387-2.381)	$1.5{\cdot}10^{-5}$

4.1.3.4 Shared gamma frailty Cox proportional hazards model

Then, the Cox proportional hazards model with shared gamma-distributed frailties was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.16. The estimated covariate effects and their standard error are practically identical to the estimates from the Cox proportional hazards model unadjusted for any patient effect. The explanation is straightforward; the estimate of the variance of the frailties θ_f is very small $8.4{\cdot}10^{-3}$ and insignificant (p-value 0.27). Thus, there is no random effect of patient and the shared frailty model is reduced to the unadjusted Cox proportional hazards model. The estimated individual frailties are visualised in a histogram in Figure 4.7. As seen, they are distributed closely around 1.

Table 4.16: Estimates from shared gamma frailty Cox proportional hazards model (NCCTG lung cancer data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Male			1.000	
Female	-0.557	0.168	0.573 (0.412-0.797)	$9.4 \cdot 10^{-4}$
Age (per 5 years)	0.056	0.046	1.058 (0.965-1.159)	0.225
ECOG score	0.481	0.116	1.618 (1.290-2.029)	$3.1 \cdot 10^{-5}$

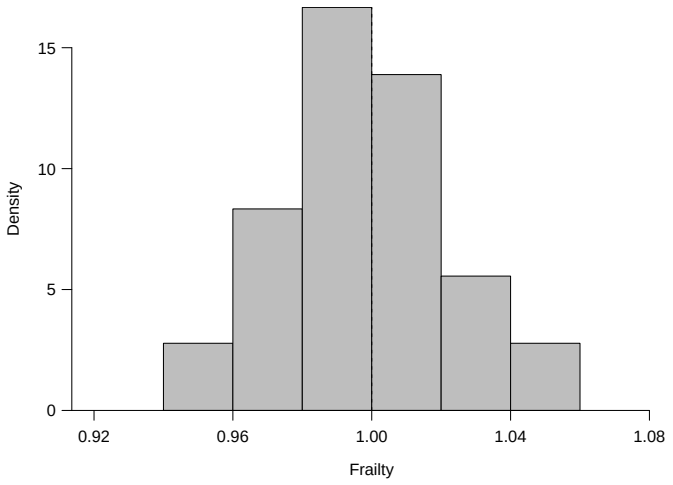

Figure 4.7: Histogram of estimated frailties (NCCTG lung cancer data).

4.1.3.5 Clayton-Oakes copula model

Finally, the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.17. Again, the effect of gender and ECOG score are significant (p-values $7.1{\cdot}10^{-7}$ and $6.4{\cdot}10^{-5}$, respectively), while the effect of age is insignificant. The hazard rate of females are 43% lower than the hazard rate of males, while the hazard rate increases with increasing ECOG score. The effect of age is insignificant. Compared to the estimates from the unadjusted Cox proportional hazards model, the robust standard errors of the estimates are smaller, except for the ECOG score.

The association parameter θ_c estimated by copula model is very small $3.7{\cdot}10^{-3}$ and significant (p-value $<2{\cdot}10^{-16}$. It is approximately half as big as the estimated variance of the frailties in the shared frailty model.

Table 4.17: Estimates from margin of Clayton-Oakes copula model (NCCTG lung cancer data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Male			1.000	
Female	-0.556	0.112	0.574 (0.460-0.714)	$7.1{\cdot}10^{-7}$
Age (per 5 years)	0.056	0.035	1.058 (0.988-1.132)	0.105
ECOG score	0.469	0.117	1.598 (1.270-2.012)	$6.4{\cdot}10^{-5}$

4.1.3.6 Summary of results

The estimated covariate effects, standard errors, hazard ratios and p-values from the survival models fitted to the NCCTG lung cancer data are summarised in Table 4.18.

Table 4.18: Summary of results (NCCTG lung cancer data).

		$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR	p-value	Additional parameters
Unadjusted model	Female	-0.557	0.168	0.573	$9.3{\cdot}10^{-4}$	
	Age (per 5 years)	0.056	0.046	1.058	0.225	
	ECOG score	0.469	0.114	1.599	$4.0{\cdot}10^{-5}$	
Fixed effects model	Female	-0.571	0.174	0.565	$1.0{\cdot}10^{-3}$	
	Age (per 5 years)	0.049	0.049	1.051	0.312	
	ECOG score	0.604	0.130	1.829	$3.4{\cdot}10^{-6}$	
	Institution 2	0.530	0.546	1.699	0.331	
	...					
	Institution 18	0.472	1.030	1.603	0.647	
Stratified model	Female	-0.547	0.182	0.578	$2.6{\cdot}10^{-3}$	
	Age (per 5 years)	0.048	0.051	1.049	0.353	
	ECOG score	0.597	0.138	1.817	$1.5{\cdot}10^{-5}$	
Shared frailty model	Female	-0.557	0.168	0.573	$9.4{\cdot}10^{-4}$	$\theta_f = 8.5{\cdot}10^{-3}$
	Age (per 5 years)	0.056	0.046	1.058	0.225	(p-value 0.27)
	ECOG score	0.481	0.116	1.618	$3.1{\cdot}10^{-5}$	
Copula model	Female	-0.556	0.112	0.574	$7.1{\cdot}10^{-7}$	$\theta_c = 3.7{\cdot}10^{-3}$
	Age (per 5 years)	0.056	0.035	1.058	0.105	(p-value $<2{\cdot}10^{-16}$)
	ECOG score	0.469	0.117	1.598	$6.4{\cdot}10^{-5}$	

4.2 Childhood cancer data

In this section, the results from the statistical analyses of the childhood cancer data are presented. The number of families in the childhood cancer data is 24066, thus to fit Cox proportional hazards model with family as a fixed effect is extremely computationally expensive and not possible in R. Therefore, this model has not been fitted to the childhood cancer data. The variable of interest in the childhood cancer data is exposure group, however all members in a family are in the same exposure group and therefore it does not make sense to fit the Cox proportional hazards model stratified by family to this data. This being so, the survival models that have been fitted to this data are the Cox proportional hazards model unadjusted for any family effect, the shared gamma frailty Cox proportional hazards model, and the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin.

4.2.1 Unadjusted Cox proportional hazards model

The Cox proportional hazard model unadjusted for any family effect was first fitted to the childhood cancer data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.19, where the term interaction covers the interaction between no exposure and sibling. As seen, all variables are significant. The hazard ratios of the different groups are shown in Table 4.20 and visualised in Figure 4.8. Siblings and unexposed individuals have a lower hazard rate of admission due to a mental disorder than childhood cancer survivors (exposed probands).

Table 4.19: Estimates from unadjusted Cox proportional hazards model (childhood cancer data).

	$\hat{\beta}$	se	HR (95% CI)	p-value
Exposed			1.000	
Unexposed	-0.335	0.073	0.715 (0.619-0.826)	$4.9{\cdot}10^{-6}$
Proband			1.000	
Sibling	-0.328	0.088	0.720 (0.606-0.855)	$1.9{\cdot}10^{-4}$
Interaction	0.290	0.096	1.336 (1.107-1.612)	$2.5{\cdot}10^{-3}$
No previous admission			1.000	
Previous admission	0.664	0.079	1.942 (1.664-2.268)	$<2{\cdot}10^{-16}$

Table 4.20: Hazard ratios from unadjusted Cox proportional hazards model (childhood cancer data).

	HR	95% CI
Exposed proband	1.000	
Exposed sibling	0.720	0.606-0.855
Unexposed proband	0.715	0.619-0.826
Unexposed sibling	0.688	0.597-0.793

The proportional hazards assumption of the included variables have been tested by means of the Schoenfeld residuals and their correlation with time. There is no evidence against proportionality.

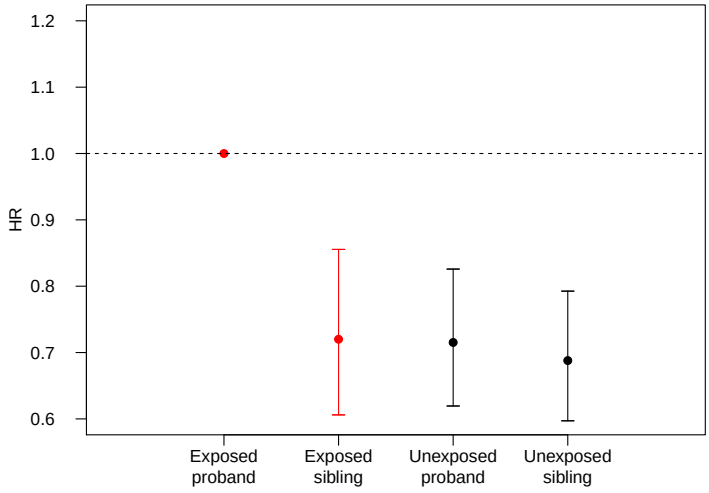

Figure 4.8: Visualisation of results from Cox proportional hazards model unadjusted for any patient effect (childhood cancer data).

4.2.2 Shared gamma frailty Cox proportional hazards model

Then, the Cox proportional hazards model with shared gamma-distributed frailties was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.21, where the term interaction covers the interaction be-

tween no exposure and sibling. As in the unadjusted model, the interaction term and all the variables are significant. The estimated effects are slightly different than the estimates from the Cox proportional hazards model unadjusted for any family effect. The hazard ratios of the different groups are shown in Table 4.22 and visualised in Figure 4.9. Siblings and unexposed individuals have a lower hazard rate of admission due to a mental disorder than childhood cancer survivors (exposed probands).

The estimate of the variance of the frailties θ_f is 1.31 and significant (p-value $4.1{\cdot}10^{-3}$). The frailties are visualised in a histogram in Figure 4.10. In Figure 4.11, the frailties have been plotted against number of events in family. As seen, there is a clear relationship between the frailty and the number of admissions in a family. Families with multiple events are more frail than families with no or only a single event.

Table 4.21: Estimates from shared gamma frailty Cox proportional hazards model (childhood cancer data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Exposed			1.000	
Unexposed	-0.349	0.078	0.706 (0.605-0.822)	$8.3{\cdot}10^{-6}$
Proband			1.000	
Sibling	-0.340	0.090	0.712 (0.597-0.848)	$1.5{\cdot}10^{-4}$
Interaction	0.302	0.098	1.353 (1.117-1.638)	$2.0{\cdot}10^{-3}$
No previous admission			1.000	
Previous admission	0.749	0.096	2.114 (1.751-2.552)	$6.6{\cdot}10^{-15}$

Table 4.22: Hazard ratios from shared gamma frailty Cox proportional hazards model (childhood cancer data).

	HR	95% CI
Exposed proband	1.000	
Exposed sibling	0.712	0.597-0.848
Unexposed proband	0.706	0.605-0.822
Unexposed sibling	0.679	0.584-0.790

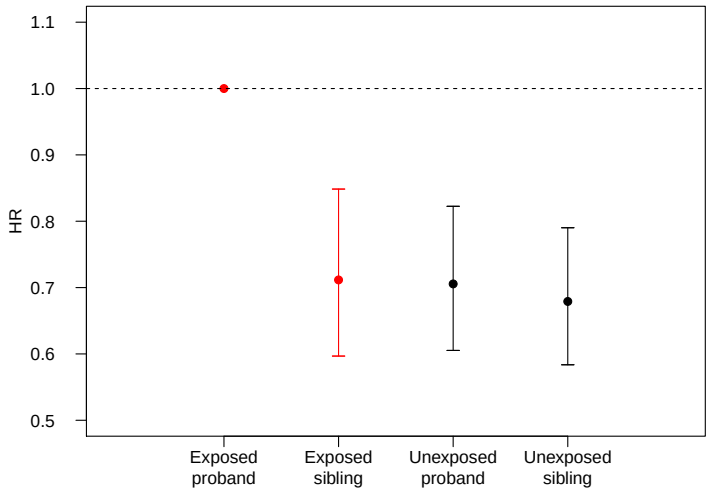

Figure 4.9: Visualisation of results from shared gamma frailty Cox proportional hazards model (childhood cancer data).

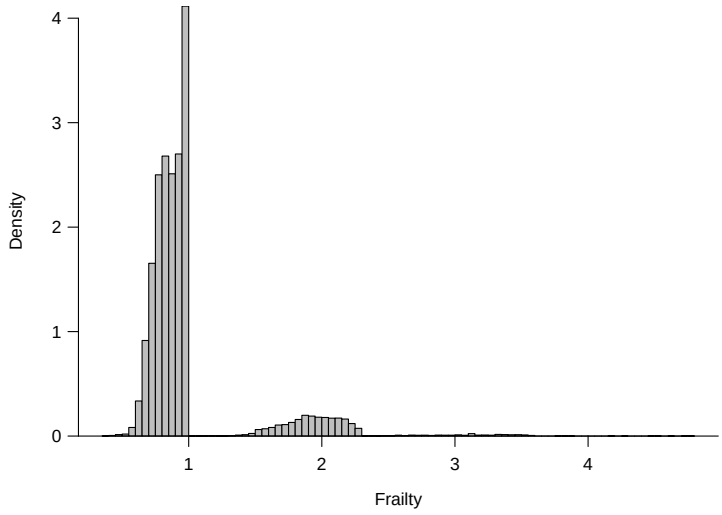

Figure 4.10: Histogram of estimated frailties (childhood cancer data).

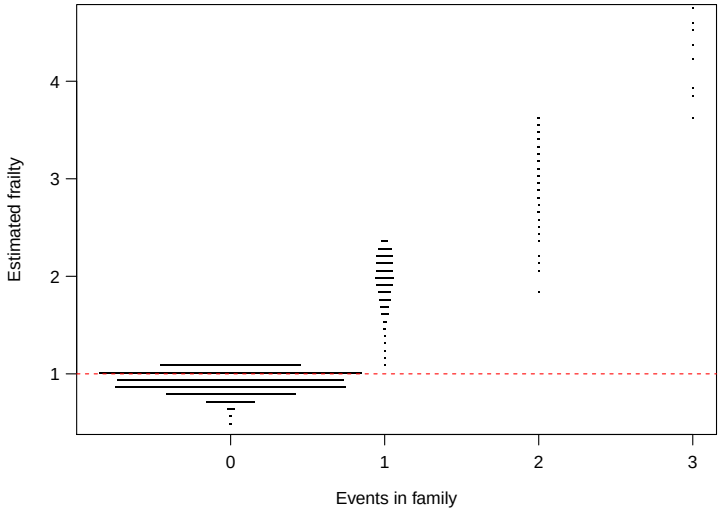

Figure 4.11: Estimated frailty plotted against number of events in family (childhood cancer data).

4.2.3 Clayton-Oakes copula model

Finally, the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin was fitted to the data. The estimated covariate effects, their standard errors (se), the corresponding hazard ratios (HR) and 95% confidence intervals (CI) are shown in Table 4.23, where the term interaction covers the interaction between no exposure and sibling. As seen, the estimates are approximately identical to the estimates from the Cox proportional hazards model unadjusted for any family effect, however, the estimated robust standard errors of the covariate effects are extremely small, which cause all effects to be very significant. This will be elaborated in the discussion. The hazard ratios of the different groups are shown in Table 4.24 and visualised in Figure 4.12. Siblings and unexposed individuals have a lower hazard rate of admission due to a mental disorder than childhood cancer survivors (exposed probands).

The association parameter θ_c estimated by copula model is 1.42 and significant (p-value $<2{\cdot}10^{-16}$). Note, that the estimated association parameter θ_c is of the same magnitude as the estimated variance of the frailties θ_f from the shared gamma frailty model.

Table 4.23: Estimates from margin of Clayton-Oakes copula model (childhood cancer data).

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR (95% CI)	p-value
Exposed			1.000	
Unexposed	-0.335	0.001	0.715 (0.713-0.717)	$<2{\cdot}10^{-16}$
Proband			1.000	
Sibling	-0.328	0.013	0.720 (0.702-0.739)	$<2{\cdot}10^{-16}$
Interaction	0.290	0.014	1.336 (1.299-1.373)	$<2{\cdot}10^{-16}$
No previous admission			1.000	
Previous admission	0.664	0.008	1.942 (1.913-1.973)	$<2{\cdot}10^{-16}$

Table 4.24: Hazard ratios from margin of Clayton-Oakes copula (childhood cancer data).

	HR	95% CI
Exposed proband	1.000	
Exposed sibling	0.720	0.702-0.739
Unexposed proband	0.715	0.713-0.717
Unexposed sibling	0.688	0.680-0.696

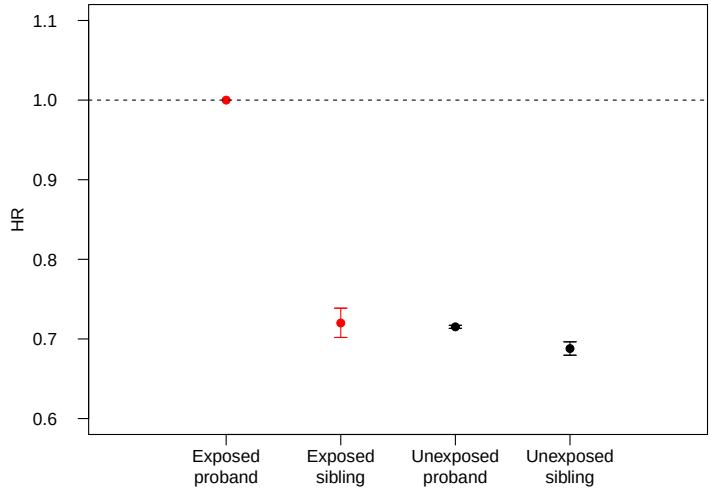

Figure 4.12: Visualisation of results from margin of Clayton-Oakes copula (childhood cancer data).

4.2.4 Summary of results

The estimated covariate effects, standard errors, hazard ratios and p-values from the survival models fitted to the childhood cancer data are summarised in Table 4.25.

Table 4.25: Summary of results (childhood cancer data).

		$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR	p-value	Additional parameters
Unadjusted model	Unexposed	-0.335	0.073	0.715	$4.9{\cdot}10^{-6}$	
	Sibling	-0.328	0.088	0.720	$1.9{\cdot}10^{-4}$	
	Interaction	0.290	0.096	1.336	$2.5{\cdot}10^{-3}$	
	Previous admission	0.664	0.079	1.942	$<2{\cdot}10^{-16}$	
Shared frailty model	Unexposed	-0.349	0.078	0.706	$8.3{\cdot}10^{-6}$	$\theta_f = 1.31$
	Sibling	-0.340	0.090	0.712	$1.5{\cdot}10^{-4}$	(p-value $4.1{\cdot}10^{-3}$)
	Interaction	0.302	0.098	1.353	$2.0{\cdot}10^{-3}$	
	Previous admission	0.749	0.096	2.114	$6.6{\cdot}10^{-15}$	
Copula model	Unexposed	-0.335	0.001	0.715	$<2{\cdot}10^{-16}$	$\theta_c = 1.42$
	Sibling	-0.328	0.013	0.720	$<2{\cdot}10^{-16}$	(p-value $<2{\cdot}10^{-16}$)
	Interaction	0.290	0.014	1.336	$<2{\cdot}10^{-16}$	
	Previous admission	0.664	0.008	1.942	$<2{\cdot}10^{-16}$	

Chapter 5

Discussion

In this chapter, the statistical methods will be discussed based on the results presented in the previous chapter. In addition, extensions of the applied methods will be presented and suggestions for further work will be given.

5.1 Data sets available through R

As illustrated by the three data examples, natural clustering of study subjects arises in different situations and for different kinds of data.

When cluster sizes are small relative to the number of clusters, introduction of a fixed effect for each cluster is generally not a good solution, and in some situations it is not even possible. With regard to the data sets available through R, it was only possible to fit the Cox proportional hazards model with cluster as a fixed effect to one out of the three; the NCCTG lung cancer data. In the diabetic retinopathy data and the kidney catheter data, there were clusters without any events, where it was not possible to estimate a cluster effect, and thus the model did not converge. With regard to the model fitted to the NCCTG lung cancer data, the standard errors for some of the cluster effects were very large because of a limited number of subjects in these clusters. Although, data was adjusted for

clustering, the introduction of a fixed cluster effect did not contribute valuable information to the analysis but reduced the degrees of freedom.

Stratification by cluster allows each cluster to have its own unspecified baseline. This approach is normally used to accommodate covariates that do not satisfy the proportional hazard assumption [27]. Although this model is more flexible than the fixed effects model [7], it is inefficient, when the covariates of interest are nested within the clusters as was the case with both the diabetic retinopathy data and the kidney catheter data. When the clustering is on a higher level as in the NCTTG lung cancer data, i.e. institution versus patient, stratification by cluster is a reasonable solution. With regard to the NCCTG lung cancer data, the effect estimates were similar to that of the Cox proportional hazards model unadjusted for any cluster effect, although the standard errors were somewhat higher. This may be explained by the fact that a cluster only contributes to the partial likelihood if an event is observed, while at least one other subject is still at risk [7, 23].

It was possible to fit the shared frailty model to all three data sets, however the variance of the frailty parameter was only significant in the diabetic retinopathy data and in the kidney catheter data. The estimated effects were slightly different from the estimates from the model unadjusted for any patient effect, however, this is not surprising, as the covariate effects are conditioned on the frailty and have to be interpreted on cluster level. For the NCCTG lung cancer data, the variance of the frailty was approximately zero, and the shared frailty model was practically reduced to the model unadjusted for any cluster effect.

The Clayton-Oakes copula modula with the marginal Cox proportional hazards model as margin was like the shared frailty model fitted to all three data sets. As the marginal Cox proportional hazards model is an IWM, the effect estimates were identical to the estimates from the Cox proportional hazards model unadjusted for any cluster effect, while the standard errors varied. With regard to the diabetic retinopathy data, the estimated association parameter was significant and very similar to the variance of the frailty parameter in the shared frailty model. There was an outlier in the kidney catheter data, which affected both the shared frailty model and the copula model. The characteristics of the outlier was accounted for by the frailty in the shared frailty model, while the Clayton-Oakes copula model with marginal Cox proportional hazards model as margin compensated by means of the association parameter and the robust standard errors, see Table 4.12 and Table B.1 on page 49 and 86, respectively.

5.2 Childhood cancer data

Three different statistical models were applied in the analysis of the data from the Danish register-based study of incident admission due to mental disorder in families exposed to childhood cancer. The three models were

- The Cox proportional hazards model unadjusted for any family effect
- The shared gamma frailty Cox proportional hazards model
- The Clayton-Oakes copula with the marginal Cox proportional hazards model as margin

All three models showed that individuals diagnosed with cancer and individuals with a family history of admission due to mental disorders have an increased hazard rate. Having a sister or brother diagnosed with cancer did not increase the hazard rate.

5.2.1 Degree of dependence

The variance of the frailty parameter in the shared gamma frailty Cox proportional hazards model was significant, thus the event times within a family was positively correlated. As seen in Figure 4.11 on page 61, there was a clear relationship between the number of events and the frailty of a family; families with multiple events were more frail than families with no or only a single event. This is not surprising, as the individual frailties may be considered as unobserved risk factors. The association parameter estimated in the Clayton-Oakes copula model was also significant, and somewhat similar to the variance of the frailty parameter.

5.2.2 Robust standard errors

The robust standard errors of the estimated covariate effects in the Clayton-Oakes copula model with the marginal Cox proportional hazards model as margin, were extremely small, which caused all variables to be significant, see Table 5.1. The marginal model was fitted using the function `cox.aalen` in the `timereg` package. The marginal Cox proportional hazards model has subsequently been fitted using the function `coxph` in the `survival` package. The

robust standard errors of the marginal model when fitted using this function are larger and more reasonable, see Table 5.2.

Table 5.1: Estimates from Cox proportional hazards model fitted using the function `cox.aalen` in the `timereg` package and with age as timescale.

	$\hat{\beta}$	se$\left(\hat{\beta}\right)$	p-value
Unexposed	-0.335	0.001	$<2{\cdot}10^{-16}$
Sibling	-0.328	0.013	$<2{\cdot}10^{-16}$
Interaction	0.290	0.014	$<2{\cdot}10^{-16}$
Previous admission	0.664	0.008	$<2{\cdot}10^{-16}$

Table 5.2: Estimates from Cox proportional hazards model fitted using the function `coxph` in the `survival` package and with age as timescale.

	$\hat{\beta}$	se$\left(\hat{\beta}\right)$	p-value
Unexposed	-0.335	0.073	$4.7{\cdot}10^{-6}$
Sibling	-0.328	0.086	$1.4{\cdot}10^{-4}$
Interaction	0.290	0.094	$2.0{\cdot}10^{-3}$
Previous admission	0.664	0.083	$1.67{\cdot}10^{-15}$

In the analysis of the three data examples, the standard errors estimated by the Cox proportional hazards model unadjusted for any cluster effects and the robust standard errors estimated by the margin of the Clayton-Oakes copula were generally of the same magnitude. In these analyses, time on study was used as timescale as opposed to age as in the analysis of the childhood cancer data. Generally, it is recommended to use age as timescale, as age is a probable confounding variable in most epidemiological studies and because of delayed entry [28, 54]. To test if it is the timescale, that causes problems, the marginal Cox proportional hazards model have been fitted using the function `cox.aalen` from the `timereg` package and the function `coxph` from the `survival` package, respectively, with time on study as timescale and age at entry included as a covariate. The estimated covariate effects and robust standard errors (se) are listed in Table 5.3 and Table 5.4. As seen, there are much more concordance between the estimated robust standard errors of the two functions. It appears, that there is a problem with the function `cox.aalen`, when age is used as timescale. These discrepancies should be investigated further.

Table 5.3: Estimates from Cox proportional hazards model fitted using the function `cox.aalen` in the `timereg` package and with time on study as timescale.

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	p-value
Unexposed	-0.353	0.074	$1.9{\cdot}10^{-6}$
Sibling	-0.369	0.087	$2.3{\cdot}10^{-5}$
Interaction	0.299	0.095	$1.6{\cdot}10^{-3}$
Previous admission	0.703	0.086	$2.2{\cdot}10^{-16}$
Age	0.003	0.003	0.284

Table 5.4: Estimates from Cox proportional hazards model fitted using the function `coxph` in the `survival` package and with time on study as timescale.

	$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	p-value
Unexposed	-0.353	0.074	$1.6{\cdot}10^{-6}$
Sibling	-0.371	0.086	$1.7{\cdot}10^{-5}$
Interaction	0.314	0.094	$8.3{\cdot}10^{-4}$
Previous admission	0.709	0.084	$<2{\cdot}10^{-16}$
Age	0.003	0.003	0.267

5.3 Discussion of models

In this study, the shared gamma frailty Cox proportional hazards model and the Clayton-Oakes copula model with the marginal Cox proportional hazards model as margin were the most applicable. It was possible to fit the two models to all data sets, and in addition they could be applied in order to obtain a measure of the degree of dependence within clusters (whether significant or not). The Cox proportional hazards model with cluster as a fixed effect and the Cox proportional hazards model stratified by cluster did not perform well because of small cluster sizes and nested covariates, respectively. Yet, in other situations they may be satisfactory for analysis of clustered data.

In the analysis of the diabetic retinopathy data and the childhood cancer data, the estimated variance of the frailty parameters in the shared frailty Cox proportional hazards model were similar to the estimated association parameter in the Clayton-Oakes copula model. In the analysis of the two other data sets, the two models were not completely consistent with regard to the estimated degree of dependence within cluster. However, in these data sets the degree of dependence within clusters was (seemingly) quite small. The interpretation of

the estimated degree of dependence is easiest for bivariate survival data, where there is a clear connection between the variance of the frailty parameters and the association parameter, respectively, and Kendall's τ.

An advantage of the shared gamma frailty Cox proportional hazards model is that it has a natural random effects interpretation and that the individual frailties are estimated and may be subsequently explored. The latter may provide insight into what makes a family frail and for example lead to inclusion of additional explanatory covariates. There is no frailty interpretation of negative dependence [23], whereas the bivariate Clayton-Oakes copula model with the marginal Cox proportional hazards model as margin may be extended to allow for negative dependence.

There are great concordance between the shared gamma frailty Cox proportional hazards model and the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin. The joint survival functions of the two models have the same functional form, yet the two models are not equivalent and do not lead to the same parameter estimates, because the marginal survival functions are modelled differently [7, 17]. The parameter estimates obtained by the shared frailty model are conditioned on the frailty and thus to be interpreted on cluster level, while the parameter estimates obtained by the copula model may be interpreted on population level. In this study, the parameter estimates from the two models are somewhat similar and tell the same story, i.e. the practical implications of the cluster versus population level interpretation are minor. However, this may not always be the case. If the practical implications were large, a population level interpretation and thus the copula model would most likely be preferred to the shared frailty model.

5.4 Checking the adequacy of the model

Checking the adequacy of the shared gamma frailty Cox proportional hazards model and the Clayton-Oakes copula with the Cox proportional hazard model as margin, respectively, has not been the focus of this study, but deserves some attention and further work. As suggested by Hougaard (2000) and Andersen (2005), the models may be checked by comparing the fit to that of a larger model, here e.g. the power variance frailty model [23] and power variance copula, respectively [2]. In addition, the shared frailty gamma model may be checked by evaluating the conditional mean of the frailty variable θ as a function of time. Ideally, it should fluctuate around one [14, 23, 49]. For more details on checking the model adequacy, please see Shih and Louis (1995a), Glidden (1999), and Hougaard (2000).

5.5 Extensions

In reality, all subjects in a cluster do not necessarily have the same degree of dependence. For example if the cluster is a family with members on different levels, e.g. parents and children, the degree of dependence between children will be different from the degree of dependence between the parents, which will again be different from the degree of dependence between a parent and a child. Thus, the shared frailty model and the copula model are in some situations too simple. However, several extensions of the models exists. E.g. the additive, multiplicative and hierarchical frailty models [7, 23, 38, 46, 59] and hierarchical copula models [1, 3]. It is evident, that these models may with advantage be applied to the data from the register-based family study, where there are family members on different levels, which cannot be assumed to have the same degree of dependence.

Chapter 6

Conclusion

In this study, different statistical methods for analysis of clustered survival data have been evaluated and compared using data from a Danish register-based family study of the psychological effects of exposure to childhood cancer. It has been investigated how childhood cancer survivors and their full siblings are affected later in life with regard to psychological outcomes. The methods that have been applied to the data are the shared gamma frailty Cox proportional hazards model and the Clayton-Oakes copula model with the marginal Cox proportional hazards model as margin. In addition to assessing the effect of exposure to childhood cancer whilst coping with familial clustering, the shared frailty model and the copula model were applied in order to estimate familial correlation of ages at onset of psychological disorders.

It was not possible to fit the Cox proportional hazards model with cluster as a fixed effect and the Cox proportional hazards model stratified by cluster to the data. Initially, all methods were explored using three smaller data sets, and generally seen, these two methods did not perform well because of small cluster sizes and nested covariates, among other things. Yet, in other situations they may be satisfactory for analysis of clustered data.

The applied models showed, that individuals diagnosed with cancer and individuals with a family history of admissions due to mental disorders have an increased hazard rate. Having a sister or brother diagnosed with cancer does

not increase the hazard rate. A significant correlation of age at onset of psychological disorders within families was identified by both the shared gamma frailty Cox proportional hazards model and the Clayton-Oakes copula model.

An advantage of the shared gamma frailty Cox proportional hazards model is that it has a natural random effects interpretation and that the individual frailties are estimated and may be subsequently explored. The latter may provide insight into what makes a family frail and for example lead to inclusion of additional explanatory covariates. There is no frailty interpretation of negative dependence, whereas the bivariate Clayton-Oakes copula model with the marginal Cox proportional hazards model as margin may be extended to allow for negative dependence. However, there are great concordance between the shared gamma frailty Cox proportional hazards model and the Clayton-Oakes copula with the marginal Cox proportional hazards model as margin. The main difference between the two models, is that the estimated effects from the shared frailty model are to be interpreted conditional on the frailty, i.e. within the same cluster, while the estimated effects from the margin of the copula model may be interpreted on population level. In this study, the parameter estimates from the two models are somewhat similar and tell the same story, i.e. the practical implications of the cluster versus population level interpretation are minor. However, this may not always be the case. If the practical implications were large, a population level interpretation and thus the copula model would most likely be preferred to the shared frailty model.

Suggestions for further work include checking the model adequacy and extensions of both the shared frailty model and the copula model.

Appendix A

R Code

A.1 R code for data examples

Diabetic retinopathy data

```
#### Diabetic retinopathy data ####

# Loading libraries
library(survival); library(timereg);

# Reading data
data(diabetes)

# New factor
diabetes$new <- ifelse(diabetes$treat==0 & diabetes$adult
    ==1, 1, 0)
diabetes$new <- ifelse(diabetes$treat==0 & diabetes$adult
    ==2, 2, diabetes$new)
diabetes$new <- ifelse(diabetes$treat==1 & diabetes$adult
    ==1, 3, diabetes$new)
diabetes$new <- ifelse(diabetes$treat==1 & diabetes$adult
    ==2, 4, diabetes$new)
diabetes$new <- as.factor(diabetes$new)
```

```
levels(diabetes$new) <- c("No treatment, juvenile onset","No
    treatment, adult onset","Treatment, juvenile onset","
    Treatment, adult onset")

# Factors and levels
diabetes$treat <- as.factor(diabetes$treat)
diabetes$adult <- as.factor(diabetes$adult)
levels(diabetes$treat) <- c("No treatment","Treatment")
levels(diabetes$adult) <- c("Younger than 20","Older than 20
    ")

## Cumulative incidence ##
fit1 <- survfit(Surv(diabetes$time, status==1)~ diabetes$new
    , data=diabetes)
plot(fit1, fun="event", mark.time=F, conf.int=F, col=1:2,
    lty=c(1,1,2,2), xlab="Time (months)", ylab="Cumulative
    incidence of blindness")

## Crude rates ##
dia <- pyears(Surv(time, status==1)~ new, data=diabetes,
    data.frame=T, scale=12)$data
dia$rate <- round(dia$event/dia$pyears, 2)

## Unadjusted Cox proportional hazards model ##
u <- coxph(Surv(time, status==1)~ treat*adult, data=diabetes
    )
summary(u)

# Proportional hazards test
ph <- cox.zph(u, transform="km")
ph

## Cox proportional hazards model stratified by patient ##
s <- coxph(Surv(time, status==1)~ treat + strata(id), data=
    diabetes)
summary(s)

## Shared frailty Cox proportionals hazard model ##
sf <- coxph(Surv(time, status==1)~ treat*adult + frailty(id)
    , data=diabetes)
summary(sf)

## Copula model with Cox proportional hazards model as
    marginal ##

# Step 1: Marginal model
```

```
m <- cox.aalen(Surv(time, status==1)~ prop(treat)*prop(adult
    ) + cluster(id),
   data=diabetes, clusters=diabetes$id, robust=1, max.clust=
       NULL)
summary(m)

# Step 2: Estimation of association parameter
c <- two.stage(m, data=diabetes, theta=0.99, detail=0, Nit
    =40)
summary(c)
```

Kidney catheter data

```
#### Kidney catheter data ####

# Loading libraries
library(survival); library(timereg); library(rms);

# Reading data
data(kidney)

# Factors and levels
kidney$sex <- as.factor(kidney$sex)
levels(kidney$sex) <- c("Male","Female")
kidney$agegroup <- cut(kidney$age, seq(9,69,10), c("10-19","
    20-29","30-39","40-49","50-59","60-69"))

# To get effect of age per 10 years
kidney$age <- kidney$age/10

## Cumulative incidence ##
fit1 <- survfit(Surv(kidney$time,status==1)~ sex, data=
    kidney)
plot(fit1, fun="event", mark.time=F, conf.int=F, col=1:2,
    lty=1, xlab="Time (days)", ylab="Cumulative incidence of
    recurrent infection")

## Incidence rates ##
kid1 <- pyears(Surv(time, status==1)~ sex, data=kidney, data
    .frame=T, scale=365.25)$data
kid1$rate <- round(kid1$event/kid1$pyears, 2)

kid2 <- pyears(Surv(time, status==1)~ agegroup, data=kidney,
     data.frame=T, scale=365.25)$data
kid2$rate <- round(kid2$event/kid2$pyears, 2)
```

```
## Unadjusted Cox proportional hazards model ##
u <- coxph(Surv(time, status==1)~ sex + age, data=kidney)
summary(u)

# Proportional hazard test
ph <- cox.zph(u, transform="km")
ph

# Checking linearity
kidney$age_n <- kidney$age*10

dd <- datadist(kidney)
options(datadist='dd')

up <- cph(Surv(time, status==1) ~ sex + rcs(age_n), data =
    kidney, x=T, y=T)
p <- Predict(up, age_n)
plot(p)

## Shared frailty Cox proportional hazards model ##
sf <- coxph(Surv(time, status==1)~ sex + age + frailty(id),
    data=kidney)
summary(sf)

## Copula model with Cox proportional hazards model as
    marginal ##

# Step 1: Marginal model
m <- cox.aalen(Surv(time, status==1)~ prop(sex) + prop(age)
    + cluster(id), data=kidney, clusters=kidney$id, robust=1,
     max.clust=NULL)
summary(m)

# Step 2: Estimation of association parameter
c <- two.stage(m, data=kidney, theta=0.99, detail=0, Nit=40)
summary(c)
```

NCCTG lung cancer data

```
#### NCCTG lung cancer data ####

# Loading libraries
library(survival); library(timereg); library(rms);
```

```
# Reading data
data(lung)

# Removing observations with missing values
lung <- lung[-which(is.na(lung$inst)|is.na(lung$ph.ecog)), ]

# Factors and levels
lung$sex <- as.factor(lung$sex)
levels(lung$sex) <- c("Male","Female")
lung$inst <- as.factor(lung$inst)
levels(lung$inst) <- c(1:nlevels(lung$inst))
lung$agegroup <- cut(lung$age, seq(29,89,10), c("30-39","
    40-49","50-59","60-69","70-79","80-89"))

# To get effect of age per 5 years
lung$age <- lung$age/5

## Barplot ##
counts <- table(lung$inst)
barplot(counts, xlab="Enrollment institution", ylab="Number
    of patients")

## Cumulative incidence ##
fit1 <- survfit(Surv(time,status==2)~ sex, data=lung)
plot(fit1, fun="event",mark.time=F,conf.int=F,col=1:2,lty=1,
    xlab="Time (days)", ylab="Cumulative incidence of death")

## Incidence rates ##
lung1 <- pyears(Surv(time,status==2)~ sex, data=lung, data.
    frame=T, scale=365.25)$data
lung1$rate <- round(lung1$event/lung1$pyears, 2)

lung2 <- pyears(Surv(time,status==2)~ agegroup, data=lung,
    data.frame=T, scale=365.25)$data
lung2$rate <- round(lung2$event/lung2$pyears, 2)

lung3 <- pyears(Surv(time,status==2)~ as.factor(ph.ecog),
    data=lung, data.frame=T, scale=365.25)$data
lung3$rate <- round(lung3$event/lung3$pyears, 2)

## Unadjusted Cox proportional hazards model ##
u <- coxph(Surv(time, status==2)~ sex + age + ph.ecog, data=
    lung)
summary(u)

# Proportional hazards test
ph <- cox.zph(u, transform="km")
```

```
ph

# Checking linearity
lung$age_n <- lung$age*5

dd <- datadist(lung)
options(datadist='dd')

up <- cph(Surv(time, status==2) ~ sex + rcs(age_n) + rcs(ph.
    ecog), data = lung, x=T, y=T)
p1 <- Predict(up, age_n)
plot(p1)
p2 <- Predict(up, ph.ecog)
plot(p2)

## Cox proportional hazards model with institution as fixed
    effect ##
f <- coxph(Surv(time, status==2)~ sex + age + ph.ecog + as.
    factor(inst), data = lung)
summary(f)

anova(f,u)

## Cox proportional hazard model stratified by institution
    ##
s <- coxph(Surv(time, status==2)~ sex + age + ph.ecog +
    strata(inst), data = lung)
summary(s)

## Shared frailty Cox proportional hazards model ##
sf <- coxph(Surv(time, status==2)~ sex + age + ph.ecog +
    frailty(inst), data=lung)
summary(sf)

## Copula model with Cox proportional hazards model as
    marginal ##

lung$inst <- as.numeric(as.character(lung$inst))

# Step 1: Marginal model
m <- cox.aalen(Surv(time, status==2)~ prop(sex) + prop(age)
    + prop(ph.ecog) + cluster(inst), data=lung, clusters=lung
    $inst, robust=1, max.clust=NULL)
summary(m)

# Step 2: Estimation of association parameter
```

```
c <- two.stage(m, data=lung, theta=0.99, detail=0, Nit=40,
    step=0.1)
summary(c)
```

A.2 R code for childhood cancer data

```
#### Childhood cancer data ####

# Loading libraries
library(cmprsk); library(rms); library(reshape); library(
    survival); library(timereg)

# Reading data
cohort <- read.table("cohortthesis_070212.txt", sep="\t",
    header=T)

# Correcting pnr
cohort$pnr <- as.character(cohort$pnr)
cohort$pnr <- ifelse(nchar(cohort$pnr)==9, paste(0, cohort$
    pnr, sep=""), cohort$pnr)

# Factors
cohort$R_ID <- as.factor(cohort$R_ID)
cohort$F_ID <- as.factor(cohort$F_ID)
cohort$kF_ID <- as.factor(cohort$kF_ID)
cohort$exposed <- as.factor(cohort$exposed)

# Dates
cohort$fdato <- as.Date(cohort$fdato, "%Y-%m-%d")
cohort$date_in <- as.Date(cohort$date_in, "%Y-%m-%d")

## Cumulative incidence ##
cohort$time <- cohort$age_out - cohort$age_in

fit1 <- survfit(Surv(cohort$time,event==1)~ R_ID, data=
    cohort)
fit2 <- survfit(Surv(cohort$time,event==1)~ histx, data=
    cohort)

plot(fit1, fun="event",mark.time=F,conf.int=F,col=c(1,1,2,2)
    ,lty=c(1,2,1,2),xlab="Time (years)", ylab="Cumulative
    incidence of admission")
plot(fit2, fun="event",mark.time=F,conf.int=F,col=1:2,lty=1,
    xlab="Time (years)", ylab="Cumulative incidence of
```

```
    admission")


## Incidence rates ##
child <- pyears(Surv(time,event==1)~ R_ID, data=cohort,data.
    frame=T,scale=1)$data
child$rate <- round(1000*child$event/child$pyears,2)

child2 <- pyears(Surv(time,event==1)~ histx, data=cohort,
    data.frame=T,scale=1)$data
child2$rate <- round(1000*child2$event/child2$pyears,2)

cohort$exposed <- relevel(cohort$exposed, ref="1")

## Unadjusted Cox proportional hazards model ##
u <- coxph(Surv(age_in, age_out, event==1) ~ exposed*rel +
    hist, data=cohort)
summary(u)

# Proportional hazard test
ph <- cox.zph(u, transform="km")
ph

## Shared frailty Cox proportional hazards model ##
sf <- coxph(Surv(age_in, age_out, event==1) ~ exposed*rel +
    hist + frailty(newF_ID), data = cohort)
summary(sf)

## Copula model with Cox proportional hazards model as
    marginal ##

# Step 1: Marginal model
m <- cox.aalen(Surv(age_in, age_out, event==1)~ prop(exposed
    )*prop(rel) + prop(hist) + cluster(newF_ID), clusters=
    cohort$newF_ID, data=cohort, max.clust=NULL, robust=1)


# Step 2: Estimation of association parameter
c <- two.stage(m, data=cohort, theta=0.99, detail=0, Nit=40)
summary(c)

## Marginal model fitting using coxph ##
m_coxph <- coxph(Surv(age_in, age_out, event==1)~ exposed*
    rel + hist + cluster(newF_ID), data=cohort, robust=T)
summary(m_coxph)
```

```
## Marginal models with time on study as timescale ##

cohort$age <- as.numeric(cohort$date_in - cohort$fdato)/
    365.25
cohort$time <- as.numeric(cohort$age_out - cohort$age_in)

m_time.cox.aalen <- cox.aalen(Surv(time, event==1)~ prop(
    exposed)*prop(rel) + prop(hist) + prop(age) + cluster(
    newF_ID), clusters=cohort$newF_ID, data=cohort, max.clust
    =NULL, robust=1)
summary(m_time.cox.aalen)


m_time.coxph <- coxph(Surv(time, event==1)~ exposed*rel +
    hist + age + cluster(newF_ID), data=cohort, robust=T)
summary(m_time.coxph)
```

Appendix B

Additional Results

B.1 Kidney catheter data

Here, additional results for the kidney catheter data are presented. The statistical analyses presented in Chapter 4 have been repeated without the patient with identification number 21. The results if these analyses are summarised in Table B.1.

Table B.1: Additional results (kidney catheter data).

		$\hat{\beta}$	$\text{se}\left(\hat{\beta}\right)$	HR	p-value	Additional parameters
Unadjusted model	Female	-1.666	0.332	0.189	$5.1{\cdot}10^{-7}$	
	Age (per 10 years)	0.063	0.090	1.065	0.484	
Shared frailty model	Female	-1.666	0.332	0.189	$5.1{\cdot}10^{-7}$	$\theta_f = 5{\cdot}10^{-7}$ and Kendall's $\tau_f = 2.5{\cdot}10^{-7}$
	Age (per 10 years)	0.063	0.090	1.065	0.484	(p-value 0.93)
Copula model	Female	-1.682	0.350	0.186	$1.6{\cdot}10^{-6}$	θ_c = -0.04 and Kendall's τ_c = -0.02
	Age (per 10 years)	0.059	0.089	1.061	0.502	(p-value 0.87)

Bibliography

[1] Elisabeth W. Andersen. Composite likelihood and two-stage estimation in family studies. *Biostatistics*, 5:15–30, 2004.

[2] Elisabeth W. Andersen. Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis*, 11:333–350, 2005.

[3] Karen J. Bandeen-Roche and Liang K.-Y. Modelling failure-time associations in data with multiple levels of clustering. *Biometrika*, 83:29–39, 1996.

[4] David Buchbinder, Jacqueline Casillas, Kevin R. Krull, Pam Goodman, Wendy Leisenring, Christopher Recktilis, Melissa A. Alderfer, Leslie L. Robison, Gregory T. Armstrong, Alicia Kunin-Batson, Margaret Stuber, and Lonnie K. Zelzter. Psychological outcomes of siblings of cancer survivors: A report from the Childhood Cancer Survior Study. *Psycho-Oncology*, 20:1259–1268, 2011.

[5] D.G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151, 1978.

[6] David R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

[7] Luc Duchateau and Paul Janssen. *The Frailty Model*. Springer, 2008.

[8] Bradley Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72:557–565, 1977.

[9] Thomas R. Fleming and David P. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc., 1991.

[10] Christian Genest and Jock MacKay. The joy of copulas: Bivariate distributions with uniform margins. *The American Statistician*, 40:280–283, 1986.

[11] Christian Genest and R. Jock MacKay. Copules achimédiennes et familles de lois bidimensionnelles dont les marges sont données. *The Canadian Journal of Statistics*, 14:145–159, 1986.

[12] Richard D. Gill. Understanding Cox's regression model: A martingale approach. *Journal of the American Statistical Association*, 79:441–447, 1984.

[13] Marianne L. Gjerstorff. The Danish Cancer Registry. *Scandinavian Journal of Public Health*, 39:42–45, 2011.

[14] David V. Glidden. Checking the adequacy of the gamma frailty model for multivariate failure times. *Biometrika*, 86:381–393, 1999.

[15] David V. Glidden. A two-stage estimator of the dependence parameter for the Clayton-Oakes model. *Lifetime Data Analysis*, 6:141–156, 2000.

[16] David V. Glidden and Steven G. Self. Semiparametric likelihood estimation in the Clayton-Oakes failure time model. *Scandinavian Journal of Statistics*, 26:363–372, 1999.

[17] Klara Goethals, Paul Jannsen, and Luc Duchateau. Frailty models and copulas: Similarities and differences. *Journal of Applied Statistics*, 35:1071–1079, 2008.

[18] Lynn R. Goldin, Ruth M. Pfeiffer, Gloria Gridley, Mitchell H. Gail, Xinjun Li, Lene Mellemkjaer, Jørgen H. Olsen, Kari Hemminki, and Martha S. Linet. Familial aggregation of Hodgkin lymphoma and related tumors. *Cancer*, 100:1902–1908, 2004.

[19] Lynn R. Goldin, Ruth M. Pfeiffer, and Kari Hemminki. Familial risk of lymphoproliferative tumors in families of patients with chronic lymphocytic leukemia: Results from the Swedish family-cancer database. *Blood*, 104:1850–1854, 2004.

[20] Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526, 1994.

[21] James G. Gurney, Kevin R. Krull, Nina Kadan-Lottick, H. Stacy Nicholson, Paul C. Nathan, Brad Zebrack, Jean M. Tersak, and Kirsten K. Ness. Social outcomes in the Childhood Cancer Survior Study cohort. *Journal of Clinical Oncology*, 27:2390–2395, 2009.

[22] Philip Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1:255–273, 1995.

[23] Philip Hougaard. *Analysis of Multivariate Survival Data*. Springer, 2000.

[24] Philip Hougaard, Bent Harvald, and Niels V. Holm. Measuring similarities between the lifetimes of adult Danish twins born between 1881-1930. *Journal of the American Statistical Association*, 87:17–24, 1992.

[25] William J. Huster, Ron Brookmeyer, and Steven G. Self. Modelling paired survival data with covariates. *Biometrics*, 45:145–156, 1989.

[26] John P. Klein. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48:795–806, 1992.

[27] David G. Kleinbaum and Mitchel Klein. *Survival Analysis - A Self-Learning Text*. Springer, second edition, 2005.

[28] Edward L. Korn, Barry I. Graubard, and Douglas Midthune. Time-to-event analysis of longitudinal follow-up of a survey: Choice of time-scale. *American Journal of Epidemiology*, 145:72–80, 1997.

[29] Elisa T. Lee and John W. Wang. *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, Inc., third edition, 2003.

[30] E.W. Lee, L.J. Wei, and D.A. Amato. Cox-type regression analysis for large number of small groups of correlated failure time observations. In J.P. Klein and P.K. Goel, editors, *Survival Analysis, State of the Art*, pages 237–247. Kluwer, 1992.

[31] Stuart R. Lipsitz, Keith B.G. Dear, and Lueping Zhao. Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics*, 50:842–846, 1994.

[32] Stuart R. Lipsitz and Michael Parzen. A jackknife estimator of variance for Cox regression for correlated survival data. *Biometrics*, 52:291–298, 1996.

[33] C.L. Loprinzi, J.A. Laurie, H.S. Wieand, J.E. Krook, P.J. Novotny, J.W. Kugler, J. Bartel, M. Law, M. Bateman, and N.E. Klatt. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, 12:601–607, 1994.

[34] Lasse W. Lund, Kjeld Schmiegelow, Catherine Rechnitzer, and Christoffer Johansen. A systematic review of studies on psychological late effects of childhood cancer: Structures of society and methodological pitfalls may challenge conclusions. *Pediatric Blood & Cancer*, 56:532–543, 2011.

[35] Wendy Mack, Bryan Langholz, and Duncan C. Thomas. Survival models for familial aggregation of cancer. *Environmental Health Perspectives*, 87:27–35, 1990.

[36] Torben Martinussen and Thomas H. Scheike. *Dynamic Regression Models for Survival Data*. Springer, 2006.

[37] C.A. McGilchrist and C.W. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47:461–466, 1991.

[38] Tron A. Moger, Marion Haugen, Benjamin H.K. Yip, Håkon K. Gjessing, and Ørnulf Borgan. A hierarchical frailty model applied to two generation melanoma data. *Lifetime Data Analysis*, 17:445–460, 2011.

[39] Douglas M. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, Inc., seventh edition, 2009.

[40] Ole Mors, Gurli P. Perto, and Preben B. Mortensen. The Danish Psychiatric Central Research Register. *Scandinavian Journal of Public Health*, 39:54–57, 2011.

[41] Roger B. Nelsen. *An Introduction to Copulas*. Springer, second edition, 2006.

[42] David Oakes. A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B*, 44:414–422, 1982.

[43] David Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84:487–493, 1989.

[44] Carsten B. Pedersen. The Danish Civil Registration System. *Scandinavian Journal of Public Health*, 39:22–25, 2011.

[45] Carsten B. Pedersen, Heine Gøtzsche, Jørgen Ø. Møller, and Preben B. Mortensen. The Danish Civil Registration System. *Danish Medical Bulletin*, 53:441–449, 2006.

[46] Jørgen Holm Petersen. A additive frailty model for correlated life times. *Biometrics*, 54:646–661, 1998.

[47] R.M. Pfeiffer, L.R. Goldin, N. Chatterjee, S. Daugherty, K. Hemminki, D. Pee, L.I. X, and M.H. Gail. Methods for testing familial aggregation of diseases in population-based samples: Application to Hodgkin lymphoma in Swedish registry data. *Annals of Human Genetics*, 68:498–508, 2004.

[48] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[49] Joanna H. Shih and Thomas A. Louis. Assessing gamma frailty models for clustered failure time data. *Lifetime Data Analysis*, 1:205–220, 1995a.

[50] Joanna H. Shih and Thomas A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384–1399, 1995b.

[51] Charles F. Spiekerman and D.Y. Lin. Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*, 93:1164–1175, 1998.

[52] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data - Extending the Cox Model*. Springer, 2000.

[53] Terry M. Therneau, Patricia M. Grambsch, and V. Shane Pankratz. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12:156–175, 2003.

[54] Anne C.M. Thiébaut and Jacques Bénichou. Choice of time-sclae in Cox's model analysis of epidemiologic cohort data: A simulation study. *Statistics in Medicine*, 23:3803–3820, 2004.

[55] Pravin K. Trivedi and David M. Zimmer. *Copula Modeling: An Introduction for Practitioners*. Now Pub, 2007.

[56] L.J. Wei, D.Y. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84:1065–1073, 1989.

[57] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.

[58] Andreas Wienke. *Frailty Models in Survival Analysis*. Chapmann & Hall, 2011.

[59] Anatoli I. Yashin, James W. Vaupel, and Ivan A. Iachine. Correlated individual frailty: An advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies*, 5:145–159, 1995.

[60] Lonnie K. Zeltzer, Christopher Recklitis, David Buchbinder, Bradley Zebrack, Jacqueline Casillas, Jennie C.I. Tsao, Qian Lu, and Kevin Krull. Psychological status in childhood cancer survivors: A report from the Childhood Cancer Survior Study. *Journal of Clinical Oncology*, 27:2396–2404, 2009.