CHAPTER 1
Introduction

Statistical analysis aims to find patterns in data and to increase understanding of such data. The beauty of statistics is that it can be applied to a great variety of fields to answer relevant research questions. The particular field of statistics called survival analysis is where the topics of this thesis find their place. Survival analysis deals with life-time data. In this type of data the time from a specific starting point until an event of interest occurs are recorded. In medical research for example, time from diagnosis of disease until death could be studied. What characterizes life-time data, also called survival data, is that it is generally incomplete. Some individuals in the data might not have experienced the event of interest at the end of the study period or have dropped out of the study before the event has occurred. These data are called *right-censored*. The event time is unknown, it is known however, that the event had not occurred before the last observation time. To handle this particular type of missing data, and other similar types, special methodology is necessary summarized under the term survival analysis. Even though survival analysis is relevant to a large number of applications, the works in this thesis are all motivated by medical research. For this reason, examples are given in the context of clinical research.

Survival analysis is used by clinicians to identify risk factors associated with the occurrence of a clinical event of interest. For example in cancer research, clinicians use survival models to investigate if a patient's age, sex, tumor size, and other variables are associated to the risk of death. To describe the evolution of disease complex mathematical models are required. Patients may experience several disease related events in different orders. *Multi-state models* can be applied in such context. Another extension of survival models is to add a random effect, also called *frailty*. Frailty terms are used to model unobserved covariates which might have an effect on the event of interest. In all studies not all relevant patient or disease characteristics can be collected and therefore the survival model is incomplete. Random effects quantify the so called unobserved heterogeneity resulting from an incomplete model.

Survival models may be used to investigate the effect of risk factors on clinical events of interest and to predict survival probabilities. Such predictions inform both patients and clinicians of a patient's prognosis and may help in the shared decision making process. Prediction models are available for a variety of diseases and there is a demand for more and more sophisticated models. Ordinary prediction models are often limited to a single prediction time point. This means that predictions can only be made at a particular time, such as at time of diagnosis of disease. When a

patient comes back for a follow-up visit, such models are not able to provide accurate predictions. A patient may experience disease related events over time which are not taken into account by a model that considers only risk factors known at diagnosis or at start of treatment. *Dynamic prediction* models provide updated predictions from different time points during follow-up. They are able to include updated information as it becomes available. A simple idea to create dynamic prediction models is through the *landmarking* approach. Predictions are made from a chosen landmark time point by using a subset of the data consisting of patients still alive at that time. Multiple landmark times can be chosen to make predictions from different time points during follow-up.

The remainder of this chapter introduces basic concepts of survival analysis as well as more complex models that are used in this thesis. The following Section provides an introduction to survival analysis and explains simple survival models. Section 1.2, 1.3, 1.4, 1.5 introduce frailty models, competing risks models, multi-state models, and dynamic prediction models, respectively. In Section 1.6 and 1.7 the C-index and multiple imputation are explained, respectively. Section 1.8 and 1.9 introduce the motivating soft tissue sarcoma data set and the developed prediction tool, respectively. The last Section gives an overview of the remaining chapters of this thesis.

§1.1 Introduction to survival analysis

The concepts and definitions of this Section are introduced as in Klein and Moeschberger [92], which is referred to for further reading.

Survival analysis developed from the need to analyse life-time data. The structure of such data can be of different kind and often motivates the development of new methods. A first step in understanding survival concepts is in understanding the data it has to deal with.

The subject of study is the event time T. In medical research, T could represent the time from diagnosis until death. The event time for an individual may not be observed, if he dropped out of the study early, or the study ended before the event of interest occurred, or another event occurred. Denote by C the right censoring time for an individual. This is the last time a subject was observed in the study. The information observed for an individual is $\tilde{T} = \min(T, C)$, the minimum between right-censoring and event time, and $\delta = \mathbb{1}(T \leq C)$, the event time indicator. $\delta = 1$, if the event time was observed and $\delta = 0$, if it was not. Survival models for right-censored data assume that the event time T and the right-censoring time C are independent, sometimes conditional on covariates.

For other types of events, the exact event time cannot be observed directly. In cancer care for example, after removal of the tumor a patient attends scheduled follow-up visits where he is screened for recurrence of disease. If recurrence is found it is only known that it had occurred between the last negative screening and the first positive screening. The time until recurrence is *interval-censored*.

To study the distribution of the survival time T different parameters are studied. The most prominent function of interest is the survival function $S(t) = P(T > t)$,

which at time t is equal to the probability of being event-free at time t. The survival function is usually modelled by the hazard function

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

The hazard function at time t is equal to the conditional probability of experiencing the event in the next instant conditional on being event-free just before time t. The survival function can be defined in terms of the hazard function,

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right),$$

and it can be estimated nonparametrically with the *Kaplan-Meier estimator* [88]. Let $t_1 < t_2 < ... < t_D$ be the ordered event times, d_i the number of events at time t_i, and Y_i denote the number of individuals at risk at time t_i. The Kaplan-Meier estimator of the survival function is given as follows

$$\hat{S}(t) = \begin{cases} 1, & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right), & \text{if } t_1 \leq t. \end{cases} \tag{1.1.1}$$

An example of survival data set and the corresponding Kaplan-Meier curve are shown in Figure 1.1. Subject 2 and 10 are right-censored. The Kaplan-Meier curve changes at event times and remains unchanged at censoring times. The censoring times however affect the size of the jump the curve makes.

The effect of a covariate vector $\mathbf{Z}$ is most commonly modelled with the Cox proportional hazards model [44] in which the hazard is defined as

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) \exp(\beta^T \mathbf{Z}),$$

where $\lambda_0(t)$ is the baseline hazard and β is the vector of regression coefficients. In the Cox model, the effect of covariates is assumed to be multiplicative on the nonparametric baseline hazard. Let $t_1 < t_2 < ... < t_D$ be the ordered event times, $\mathbf{Z}_{(i)}$ the covariates of the individual who experiences the event at time t_i, $\mathbf{Z}_j$ the covariates of individual j, and $R(t_i)$ denote the set of individuals still at risk at time t_i. The vector of regression coefficients β is estimated, assuming all event times are distinct, by maximising the partial likelihood

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^T \mathbf{Z}_{(i)})}{\sum_{j \in R(t_i)} \exp(\beta^T \mathbf{Z}_j)},$$

and the baseline hazard $\lambda_0(t)$ can then be computed using the Breslow estimator [34].

The covariates $\mathbf{Z}$ discussed so far are time-fixed and known at the time origin. However covariates can also change over time, like blood values which are repeatedly measured. Let $\mathbf{Z}(t)$ be a vector of time-dependent covariates, whose values change over time. The Cox model with time-dependent covariates is defined as

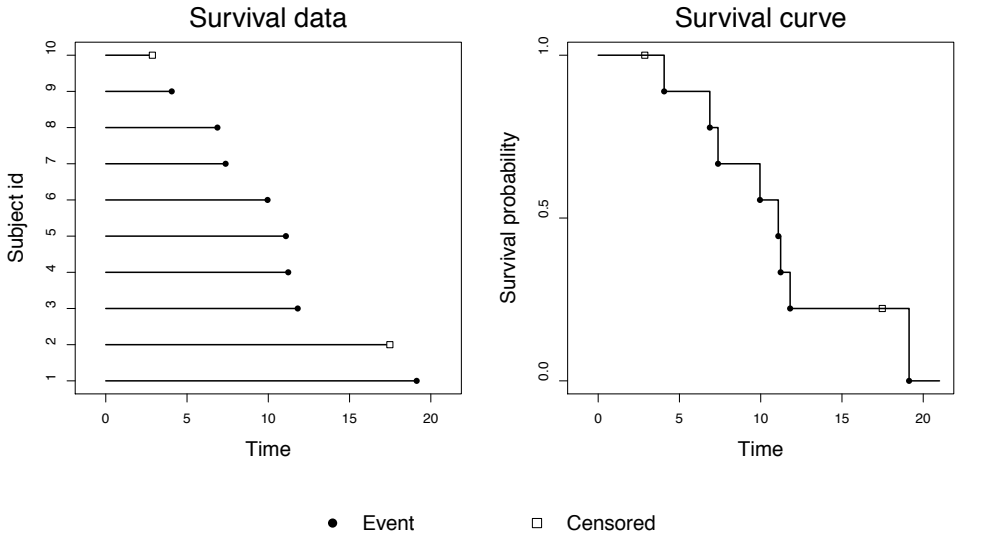

Figure 1.1: Left panel: survival data for 10 subjects. Right panel: the corresponding conditional Kaplan-Meier survival estimate.

$$\lambda(t \mid \mathbf{Z}(t)) = \lambda_0(t) \exp(\beta^T \mathbf{Z}(t)).$$

The partial likelihood for this Cox model is defined analogously to the model with only time-fixed covariates. Let $\mathbf{Z}_{(i)}(t_i)$ be the covariate vector at time t_i of the individual who experiences the event at time t_i, $\mathbf{Z}_j(t)$ the covariate vector of individual j at time t, and $R(t_i)$ the set of individuals still at risk at time t_i. Again assuming all event times are distinct, the vector of regression coefficients β is estimated by maximising the partial likelihood

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta^T \mathbf{Z}_{(i)}(t))}{\sum_{j \in R(t_i)} \exp(\beta^T \mathbf{Z}_j(t))}.$$

§1.2 Frailty models

Survival regression models aim to explain the differences of survival times between individuals using covariate information. If the model is perfectly specified, the remaining variation reflects the randomness of the event time, conditional on the covariate values. However, often not all relevant covariates can be included in the model. The variation of survival time accounted for the missing covariates in the model is called unobserved heterogeneity. The effect of unobserved heterogeneity on the event time is called frailty [157].

In survival analysis, frailty can be modelled by a random effect included in a survival model. The variance of the random effect is a measure of the amount of unobserved heterogeneity. The frailty variable can be chosen subject specific or it can be shared for clusters of individuals. In univariate frailty models, a subject specific frailty models unobserved heterogeneity on the individual level. In multivariate frailty models, a shared frailty variable is used for a cluster of individuals which models unobserved heterogeneity on the cluster level.

The random frailty variable can be incorporated in a survival model with a multiplicative effect on the hazard. The cluster i specific frailty W_i has a multiplicative effect on the hazard,

$$\lambda(t \mid W_i) = W_i \lambda_0(t),$$

where $\lambda_0(t)$ is the baseline hazard. Often $E(W_i) = 1$, then $Var(W_i)$ describes the extent of unobserved heterogeneity.A univariate frailty model has cluster size equal to 1. In this case, the estimated frailty variance represents the unobserved heterogeneity between individuals. For cluster size bigger than 1 the estimated frailty variance represents the unobserved heterogeneity between clusters. The effect of a covariate vector $\mathbf{Z}$ can be modelled by using a Cox model with frailty term

$$\lambda(t \mid W_i, \mathbf{Z}) = W_i \lambda_0(t) \exp(\beta^T \mathbf{Z}),$$

where $\lambda_0(t)$ is the baseline hazard and β are the regression coefficients. The frailty terms W_i are iid random variables with a specific distribution. The gamma distribution is a popular choice as frailty distribution due to its mathematical properties. An additional assumption of the frailty model is that censoring does not depend on the frailty [109].

§1.3 Competing risks models

In some applications, more than one type of terminal event is possible, such as in the study of different causes of death. A competing risks model is described by a starting state in which individuals are event-free and several end states, also referred to as causes of failure, see Figure 1.2.

The survival data of an individual has a different structure. Let $T_1, T_2, ..., T_J$ be the event times of J competing events and C the independent right-censoring time. For an individual, only the minimum of the first event or right-censoring time $T = \min(C, T_1, T_2, ..., T_J)$ is observed together with an indicator $\delta = 0, 1, ..., J$ indicating the cause of failure or censoring ($\delta = 0$).

A fundamental concept used in competing risks analysis is the *cause-specific* hazard,

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, \delta = j \mid T \geq t)}{\Delta t},$$

where $j, (j = 1, ..., J)$ is one of the competing events. The cause-specific hazard represents the conditional probability of experiencing event j in the next instant

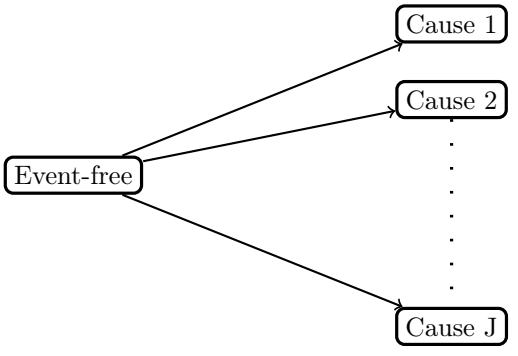

Figure 1.2: Competing risks model with J causes of failure.

given still event-free just before time t. Define the cumulative cause-specific hazard by

$$\Lambda_j(t) = \int_0^t \lambda_j(u)du.$$

A Cox model can be used to model the cause-specific hazard together with the effect of covariates $\mathbf{Z}$,

$$\lambda_j(t \mid \mathbf{Z}) = \lambda_{j0}(t) \exp(\beta_j^T \mathbf{Z}),$$

where $\lambda_{j0}(t)$ and β_j are the cause-specific baseline hazard and the regression coefficients, respectively.

A quantity of interest, in particular in applications with competing risks, is the cumulative incidence function. This function corresponds to the probability of experiencing an event j before time t,

$$I_j(t) = P(T \leq t, \delta = j) = \int_0^t \lambda_j(u)S(u)du,$$

where $S(u) = P(T > u)$ is the probability of being event-free at time u. In this context the survival function depends on all cause-specific hazards,

$$S(t) = \exp\left(-\sum_{j=1}^{J} \Lambda_j(t)\right),$$

where $\Lambda_j(t)$ is the cause-specific cumulative hazard at time t. The cumulative incidence function therefore not only depends on cause j but also on the cause-specific hazards of all the other causes.

In the competing risks setting, the cause-specific cumulative incidences are often the quantities of interest to answer questions such as, what is the probability of a recurrence of disease within a certain time frame. The Cox model can be used

to model covariate effects on the cause-specific hazards, however the effects on the cause-specific cumulative incidences are not straightforward, since they depend on all other cause-specific hazards simultaneously. Another approach to model the effect of risk factors was introduced by Fine and Gray [57]. The covariate effects are modelled directly on the cause-specific cumulative incidence through the *subdistribution hazard*,

$$\bar{\lambda}_j(t) = -\frac{d\log(1 - I_j(t))}{dt}.$$

To model covariate effects a proportional hazards model analogous to the Cox model was proposed,

$$\bar{\lambda}_j(t \mid \mathbf{Z}) = \bar{\lambda}_{j0}(t)\exp(\beta_j^T \mathbf{Z}).$$

This model can be estimated with a partial likelihood approach like the Cox model. The regression coefficients from Fine and Gray's model have an intuitive interpretation because they are regressed on the cause-specific cumulative incidence directly and therefore can be easily interpreted clinically.

§1.4 Multi-state models

Competing risks models extend standard survival models by adding more end states. Multi-state models allow multiple end points as well as transition states [119]. Figure 1.3, represents a particular multi-state model referred to as illness-death model. An individual starts in state 0, he can then move to state 1, which can represent a disease he may experience and subsequently move to state 2, death.

The transitions from state i to state j are modelled by the *transition hazard*,

$$\lambda_{ij}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

where T denotes the time of reaching state j from state i. The types and times of the occurrence of events of an individual define his path through the multi-state model. A common assumption, is that the multi-state model is a Markov model. Given the present state and the event history (the trajectory through the multi-state model so

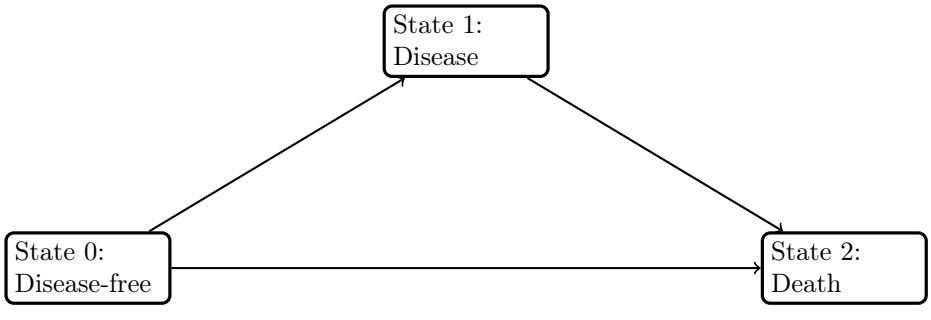

Figure 1.3: Illness-death model.

far) the next state to move to as well as the time of transition only depends on the present state.

To model the effect of covariates on the transition hazards in a Markov model the Cox model can be used,

$$\lambda_{ij}(t \mid \mathbf{Z}) = \lambda_{ij0}(t) \exp(\beta_{ij}^T \mathbf{Z}),$$

where $\lambda_{ij0}(t)$ and β_{ij} are the transition-specific baseline hazard and regression coefficients, respectively. Here the baseline hazards are transition-specific. One could however choose a subset of baseline transition hazards to be proportional to each other. For example in Figure 1.3, the two transitions towards state 2 (the death state) could be assumed proportional. This model would be similar to a single end point Cox model with the disease state as a binary (0, 1) time-dependent covariate. The multi-state model however would have the additional benefit of simultaneously modelling the rate of occurrence of the time-dependent covariate.

Similar to the competing risks setting the covariate effects are difficult to interpret. The effects act on the transition hazard which is the conditional probability of moving to state j at time t, given in state i just before t. From those transition hazards however, clinically more relevant and accessible quantities can be computed, for example, probabilities of future events. Particularly, the conditional probabilities of future events given a patients event history and covariate information $\mathbf{Z}$. In a Markov multi-state model instead of the entire event history only the current state is of relevance. For example in cancer care, given that a patient with particular characteristics is recurrence-free one year past surgery, the probability of getting a recurrence within the next 5 years may be of interest. Estimating these probabilities, also referred to as making predictions, is very relevant to patients and clinicians and can be used in the shared decision making process. The probabilities are expressed as $P_{ij}(u, t \mid \mathbf{Z})$, which represents the conditional probability of being in state j at time t given that the subject is in state i at time u. These probabilities can be computed from the transition hazards [119].

§1.5 Dynamic prediction

The previously discussed survival models may be used to model disease progression and to find risk factors for events of interest. They are also used to predict probabilities of future events. Prediction models are gaining popularity in the medical field, where they are used to inform clinician and patient about a patient's future prognosis. Survival estimates can help in the shared decision making process between patient and clinician. Many prediction models use baseline covariates; covariates measured before the time origin, such as time of diagnosis or time of treatment.

Some disease markers however, are measured and updated during follow-up. For example, blood values could be measured regularly, or recurrence of disease could be diagnosed during follow-up. Updating predictions over time with new information is defined as dynamic prediction. Dynamic prediction models can provide survival predictions at different time points, using all available information at that time point.

Among the models previously discussed only multi-state models are able to use updated information, if it can be expressed by an intermediate state. For example, the information of disease recurrence (yes vs. no) could be modelled by a state in a multi-state model.

Another approach to dynamic prediction is to use a landmark model [13, 149, 151]. Landmarking was introduced to avoid a common mistake in the analysis of tumor response. Patients were treated with chemotherapy and then followed to see whether the tumor responded to treatment. A common malpractice was to make two groups of patients, responders and nonresponders and to compare their survival times from start of treatment. This is known in literature as *immortal time bias*. The groups are compared from time of treatment, the status of response however, is not yet observed at that time. In this situation information from the future is used. An individual needs to survive long enough for a response to develop and to be observed. Individuals who die early and did not have time to develop a response yet are automatically grouped into the nonresponse group, giving this group an unfair disadvantage. On the other hand, responders must first survive long enough to become a responder and have therefore an unfair survival advantage. In this grouping scheme, responders are immortal until their time of response, hence the name immortal time bias.

A solution proposed by [13] is to evaluate the effect of tumor response with a landmark model. The idea is to choose a specific time point t_{LM} called *landmark* as new time origin for the analysis. At the landmark time the response of patients is evaluated and can be treated as a baseline covariate. The outcome is survival from landmark time and the interpretation of the response variable is response before landmark time. For this analysis only patients who are still in the risk set, meaning alive and in follow-up, are considered. Patients who developed a response before the landmark will be grouped in the response group. Patients who did not have a response before the landmark are grouped in the nonresponse group. Note that patients in the nonresponse group may develop a response later on. Figure 1.4 shows 3 landmark data sets and their corresponding Kaplan-Meier curves. The figures for landmark at time 0 show the original data and corresponding Kaplan-Meier survival curve. At time $t = 0$ no patient has developed a tumor response therefor all patients are grouped in the nonresponse group. The data for landmark at time 5 show that only individuals still in follow-up at that time are selected and that only individuals who developed a response before time 5 are grouped into the response group. The other individuals are grouped in the nonresponse group, even though they may develop a response later on.

In [151] it was suggested to use landmark models to make dynamic predictions. The Cox model is applied to the landmark data set to make predictions from the landmark time t_{LM} up until a prediction horizon $t_{LM} + w$, defined as

$$\lambda(t \mid \mathbf{Z}, t_{LM}, w) = \lambda_0(t \mid t_{LM}, w) \exp(\beta_{LM}^T \mathbf{Z}), \qquad t_{LM} \leq t \leq t_{LM} + w,$$

where $\lambda_0(t \mid t_{LM}, w)$ is the baseline hazard for landmark t_{LM} and prediction window w, β_{LM} are the landmark-specific regression coefficients, and $\mathbf{Z}$ are regression coefficients. The Cox landmark model can be used to predict survival probabilities from the landmark time, using the updated information of response status of a patient.

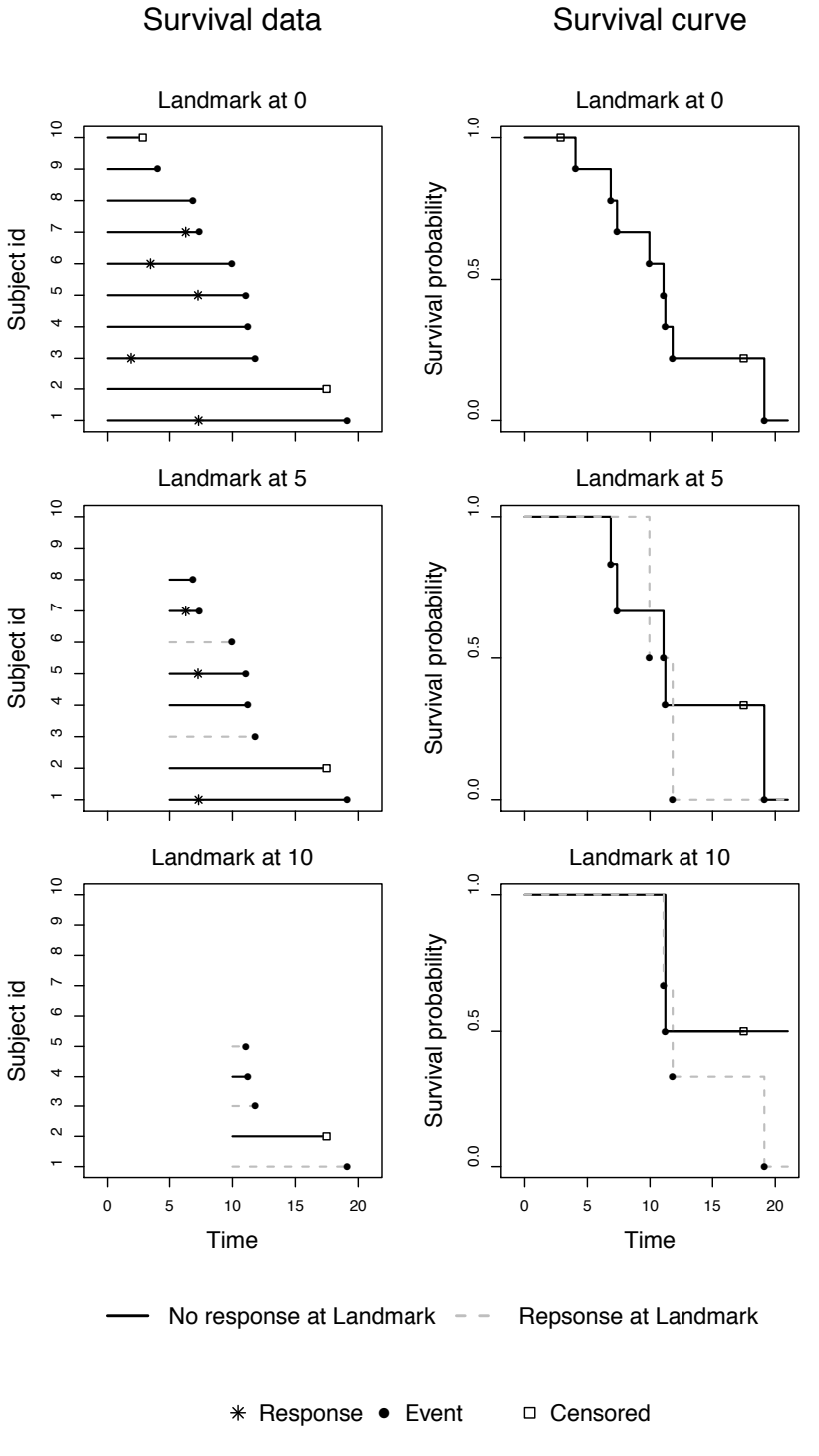

Figure 1.4: Left panel: landmark data sets. Right panel: the corresponding Kaplan-Meier survival estimate.

to make predictions from multiple landmarks one can choose to create a different landmark model for each landmark time, or combine them to a landmark *supermodel* [151].

§1.6 AUC and C-index

Survival predictions are important to patients and clinicians. More and more prediction models are becoming available for a variety of diseases. Before such a prediction model is published its clinical importance must be investigated. Does the model provide accurate predictions? To validate a prediction model different prediction aspects are considered [150].

Existing methods, such as sensitivity and specificity, were extended to survival analysis by [79], [80], and [171]. Sensitivity and specificity were originally defined for a binary outcome B, where $B = 1$ is considered a 'case' and $B = 0$ is considered a 'control'. For a covariate X and a classification criterion c a simple prediction rule is to predict individuals to be cases if $X > c$ and otherwise controls. The correct classification rates, $\text{sensitivity}(c) = P(X > c \mid B = 1)$ and $\text{specificity}(c) = P(X \leq c \mid B = 0)$, summarize the accuracy of this classification rule.

A graphical summary which illustrates the whole range of sensitivity and specificity for different values of c is the Receiver Operation Characteristic (ROC) curve which plots sensitivity against 1-specificity, illustrating the difference of the marker distribution between cases and controls. In case the marker distributions are the same, the ROC curve lies on the 45 degree line, which indicates that the marker does not contribute in distinguishing cases from controls. A summarizing measure of concordance between marker and outcome, which can be used to measure predictive accuracy is the Area Under the ROC Curve (AUC).

Heagerty and Zheng (2005) [80] extended the concepts of sensitivity and specificity to survival analysis by defining time-specific 'cases' and 'controls'. They give several different definitions, where particularly their *incident cases* and *dynamic controls* is of interest. At time t incident cases are those individuals that experience the event at time t and dynamic controls are those individuals who survive beyond time t. The time-specific AUC based on this definition of cases and controls is a time-specific measure of discrimination.

Discrimination refers to how well a model can distinguish between high and low risk individuals. A prediction model discriminates well if it predicts high risk for individuals who experience the event earlier and lower risk for individuals who experience the event late or not at all during the follow-up.

A weighted average of the time-specific incident/dynamic AUC coincides with Harrel's *concordance index* (C-index), which is a popular measure of model discrimination [149, 78]. It was originally defined as the proportion of evaluable ordered pairs for which prediction and outcome are concordant. Ordered pairs are individuals (i, j) where the observation time of individual i is shorter or equal to the observation time of individual j. An ordered pair is evaluable if it was observed that i experienced the event of interest before j. Pairs in which i was censored before j experienced the event or was censored are not evaluable, since it is unknown who experienced the

event first. Concordant pairs are pairs in which the prediction model predicted higher risk for the individual experiencing the event of interest earlier. In the case that both patients have the same predicted risk 0.5 is added to the count of concordant pairs instead of 1. The C-index is computed by dividing the number of concordant pairs by the number of evaluable pairs. The values of the C-index are between 0 and 1, where a value of 0.5 indicates no discriminative ability.

The incident/dynamic AUC measures the ability of the model to discriminate at a particular time t. The C-index measures the overall ability of the model to discriminate between individuals. Considering dynamic predictions, one might be interested in the ability of the model to discriminate within a time window $[t, t + w]$. A dynamic C-index can be obtained by computing the proportion of ordered pairs for which prediction and outcome are concordant, only for individuals who are at risk at time t and considering event times censored at time $t + w$ [149]. For predictions made at time t, this index measures the discriminative ability of the model within a time window $[t, t + w]$.

In the competing risk setting, [165] proposed a different concordance index by defining evaluable and concordant patient pairs differently. For the event of interest, they define an ordered pair as evaluable if the first patient experiences the event at a time at which the second patient is still at risk. The risk set at time t in this case is made of patients who did not yet experience any event and are still in follow-up and those individuals who experienced a competing event. Furthermore, an ordered evaluable pair is defined as concordant if the first patient to experience the event of interest has a higher predicted risk than the other patient. In case that both patients have the same predicted risk 0.5 is added instead of 1 to the count of concordant pairs.

§1.7 Multiple imputation

Most statistical methods cannot be applied when missing information are present in the data. By default many statistical programs will remove observations with missing values and analyse only complete observations. This approach reduces the amount of subjects and therefor the power of the statistical tests and can lead to biased results in some cases [100].

Multiple imputation is a general approach to handling missing data which uses all available information, even for subjects with missing values [125, 126]. The method increases statistical power and reduces bias compared to a complete case analysis. The idea behind imputation of missing values is to generate likely values for the missing values to create a complete data set. Multiple imputation uses an imputation model to generate multiple complete data sets. For these data sets the observed values are the same, however the missing values are different. Statistical methods can be applied to each complete data set and results can be pooled using Rubin's rule [125].

The concept of multiple imputation was introduced by Rubin [125]. The idea is to draw m values for each missing value from the posterior predictive distribution of the missing values under a Bayesian model for the data and the missing-data mechanism. From the resulting m complete data sets m complete-data statistics $\hat{Q}_1, ..., \hat{Q}_m$ and

the corresponding variance-covariance $U_1, ..., U_m$ can be computed and combined. The estimate from an analysis using multiple imputation can be computed by averaging the complete-data estimates

$$\bar{Q} = \frac{1}{m} \sum_{l=1}^{m} \hat{Q}_l.$$

The variance-covariance of $\bar{Q}$ is equal to

$$T = \bar{U} + \frac{m+1}{m} B,$$

where

$$\bar{U} = \frac{1}{m} \sum_{l=1}^{m} U_l,$$

and

$$B = \frac{1}{m-1} \sum_{l=1}^{m} (\hat{Q}_l - \bar{Q})(\hat{Q}_l - \bar{Q})^T.$$

The terms $\bar{U}$ and B correspond to the within-imputation variability and the between-imputation variability.

Throughout this thesis the package Amelia II was used to generate multiple imputations [82]. The assumptions for the imputation model in Amelia II are that the data are missing at random and that the complete data are multivariate normal. Missing at random means that the distribution of missingness only depends on the observed data. The multivariate normal distribution may only crudely approximate the true data distribution, however there is evidence that it works well even for categorical or mixed data [82, 129, 130]. The imputation models employed in this thesis included all variables used in the analysis together with the event status. Categorical variables were modelled as such, using the `noms` option of the `amelia` function and age and size were modelled using a square root transformation, so that no negative values could be imputed. For categorical covariates `amelia` determined the number of categories p and substituted them with $p-1$ binary variables to specify each category. These variables were treated as if they were continuous variables and missing information received continuous imputations. Those were scaled into probabilities for each category and from the resulting multinomial distribution one category was drawn so that the original multinomial variable is reconstructed.

The number of imputations was chosen to be equal 5 in Chapter 6 and 8 and 10 in Chapter 7.

§1.8 Soft tissue sarcoma data

Soft tissue sarcomas (STS) are a rare type of cancer that make up approximately 1% of all adult cancers [137]. It begins in the bodies connecting tissues, such as muscle, fat, blood vessels, nerves, tendons and the lining of joints [3]. It can appear anywhere in the body but most commonly in the extremities (about 60% of STS cases) [41]. The standard treatment for primary STS patients is surgical removal of the tumor and potentially (neo)adjuvant radiotherapy or chemotherapy [56].

Figure 1.5 shows the different disease states a patient may follow. After surgery a patient can remain disease free, or develop local recurrence (LR) or distant metastasis (DM), or die. A LR is diagnosed if evidence of tumor at the previously treated tumor bed is found, while DM is diagnosed if spread of tumor is found at another location.

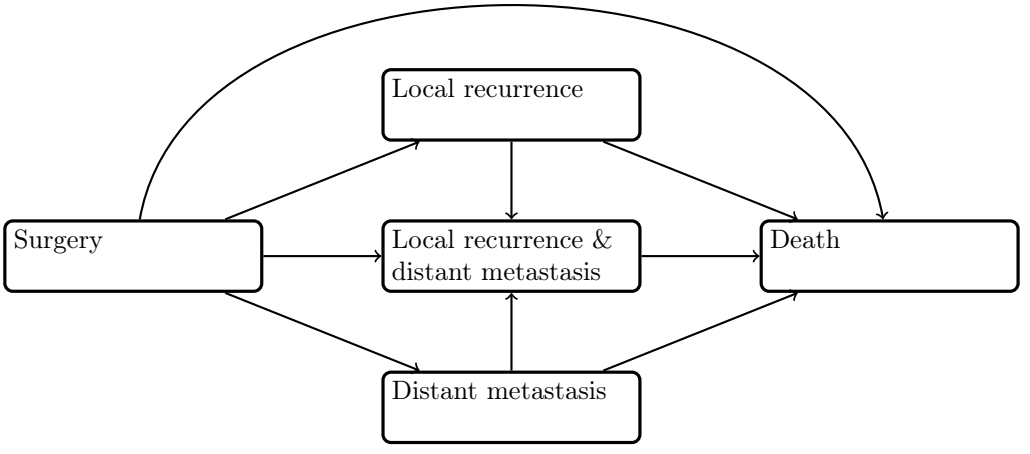

Figure 1.5: Soft tissue sarcoma data.

Because of the rarity of STS, many studies conducted have been subject to small sample size and large heterogeneity in the study population [53, 102]. Some prognostic factors for survival, such as histology, grade, depth and size were generally recognized [83, 117, 169, 53, 73, 146, 140, 98, 139, 29]. The effect of surgical margin and LR however, was long unclear [140, 106, 164, 74, 110, 102, 111]. Surgical margin is the amount of healthy tissue that is removed surrounding the tumor during primary surgery. Generally, it is desired to remove the tumor surrounded by healthy tissue, but this can be challenging depending on the tumor location. The effect of surgical margin has been of great interest, because of the effect on the quality of life after surgery.

The lack of an established prognostic profile for STS patients motivated a group of researchers to start an interdisciplinary project between the Leiden University Medical Center and the Mathematical Institute of Leiden University in 2016. The aim was to collect STS data on an international scale and to develop statistical models

to investigate the effect of risk factors on clinical outcomes, with particular interest in surgical margin, as well as to provide reliable survival predictions for patients. The funding granted by the Dutch Cancer Society (DCS) - KWF Kankerbestrijding allowed this project to become reality. The work in this thesis is the result of this collaboration.

Clinical data was collected retrospectively over the period of this project by contacting tertiary centers treating high-grade STS patients of the extremities. Patients were selected based on histological subtype, if they were treated surgically with curative intent from 2000 on, and if they had high-grade disease (as defined by FNCLCC larger than grade 2 [145]). All the collaborating centers adhered to the guidelines of the European Society for Medical Oncology for follow-up [56]. Chapter 4 is the first published article resulting from this collaboration. For each Chapters 5, 6, 7 more data has been added.

§1.9 Personalised sarcoma care app

For clinical decision making a patients prognosis always has played an important role. Reliable survival predictions are an important information to clinicians to consider in the patient care. Nowadays particularly in cancer care, the clinical community embraces the concept of shared decision making. In the shared decision making approach, the patient is involved in the choice of treatment. An important information, for this process are a patient's survival predictions. Prediction models have become popular in the clinical world [1, 6, 7]. Their increase and availability for various disease reflects the demand.

To support the shared decision making process, a prognostic prediction model, the PERsonalised SARcoma Care (PERSARC) model, for patients with high-grade STS of the extremities was developed in Chapter 5 [20]. It predicts from time of surgery a patient's probability of developing LR and survival. The model was internally validated by means of discrimination and calibration. To make predictions accessible to clinicians a mobile application was developed, which is available in the Apple and Google Play store [4, 5]. In the app patient specific characteristics can be entered and predictions are returned, see Figure 1.6 for illustration.

A group of researchers from Leiden University Medical Center was granted funding from the Dutch Cancer Society (KWF) to implement shared decision making for high-grade soft tissue sarcoma patients in the Netherlands. The goal is to ensure that soft tissue sarcoma patients receive personalised care, in which risks and benefits of treatment options and patient preferences are balanced. Part of the implementation strategy is the introduction of the PERSARC app to clinical practice.

An updated version of the PERSARC app is expected to be released in 2020. PERSARC version 2.0 will be able to provide dynamic predictions of survival as described in Chapter 7. Predictions will be made taking LR and DM events, that occur during follow-up, into account.

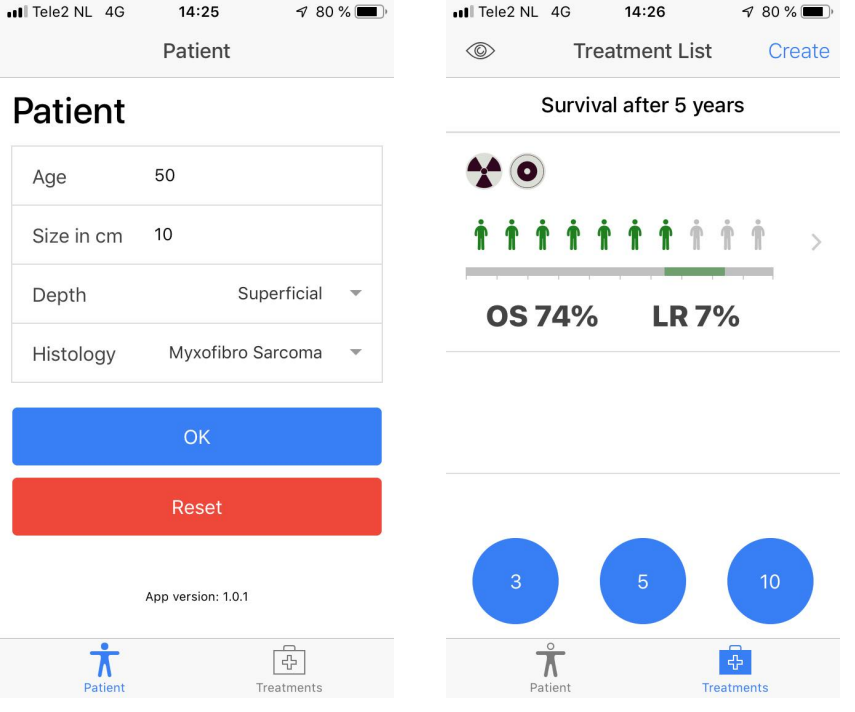

Figure 1.6: PERSARC app. Left panel: patient characteristics interface, with information of a 50 year old patient, with a 10 cm superficial tumor of type myxofibro sarcoma. Right panel: his prediction of survival and probability of recurrence within 5 years, with radiotherapy treatment and a margin of 0.1-2mm.

§1.10 Outline of thesis

The main objective of this thesis was to develop clinically relevant survival models for patients with high-grade STS of the extremities, in particular the development and validation of prediction models for use in clinical practice. The interdisciplinary collaboration between the Mathematical Institute of Leiden University and the Leiden University Medical Center resulted in important contributions to the care of STS patients. Each chapter is briefly summarized below.

In Chapter 2 [17] a novel frailty model is proposed for multi-center data with two competing risks. Frailty variables are used to model unobserved heterogeneity on the hospital level; they could be interpreted as the "hospital effect" on the competing events. The frailty model developed models the hospital effects on the competing events to be correlated within each hospital.

In Chapter 3, which is to be submitted to *Statistics in Medicine* [16], the effect of interval censoring is studied on the predicted accuracy of a binary disease marker. The motivation comes from cancer care. After surgery a patient is regularly screened for LR and DM. Once a recurrence is diagnosed , however, it is only known that it occurred between the last negative and the first positive screening. In this chapter we investigate through simulations how the assessment of predictive accuracy of recurrence is affected by the intermittent screening process.

Chapter 4 [21] was the first in a series of publications based on the growing soft tissue sarcoma data set. A data set of 687 patients was analysed with a multi-state model. The effect of risk factors on LR and DM/Death were studied, with particular interest in the effect of surgical margin.

Chapter 5 [20] is the continuation of the STS project, with a data set of 766 patients. Prediction models for survival and probability of local recurrence were developed and implemented in the PERSARC mobile application. The models are internally validated in terms of calibration and discrimination.

In Chapter 6 [19] a dynamic prediction model based on 2232 STS patients was developed. A landmark supermodel was used to provide predictions of additional 5-year survival from different prediction time points during follow-up. Disease related events, LR and DM, are used to update predictions over time and covariates were investigated for time-varying effects. The model was internally validated.

In Chapter 7, which is to be submitted to *Surgical Oncology* [18], the previously developed dynamic prediction model for STS patients is updated and externally validated. The updated model is based on 3826 patients and now includes grade as additional covariate in the model. It was externally validated using a cohort of 1111 patients and it is implemented in the updated PERSARC mobile application.

In Chapter 8 [14] a multi-state model was developed for 982 Ewing sarcoma patients. Adverse events in the multi-state model were LR, DM of the lungs, DM at other locations, and death. The effect of risk factors was studied with particular interest in surgical margins, histological response, and radiotherapy.

In Chapter 9 previous chapters are put in broader perspective and future research directions are suggested.