Multi-state and Time-varying Covariate Survival Models for Multiple Knee Contractures in People with Cerebral Palsy

Maiya Tebäck

Bachelor's thesis
2024:K2

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Multi-state and Time-varying Covariate Survival Models for Multiple Knee Contractures in People with Cerebral Palsy

A Bachelor's Degree Project in Mathematical Statistics at Lund University

Maiya Tebäck

February 5, 2024

Populärvetenskaplig sammanfattning

Cerebral pares är ett samlingsnamn för en grupp funktionsnedsättningar som påverkar rörelseförmågan, och är den vanligaste orsaken till rörelsehinder hos barn. Cerebral pares orsakas av en hjärnskada som inträffar under fosterstadiet, vid födelsen, eller före två års ålder. Musklerna hos personer med cerebral pares har en mindre storlek än hos andra, vilket till viss del kan kompenseras av längre senor. När musklerna blir för korta, och senorna inte kan kompensera tillräckligt, kan det leda till så kallade kontrakturer i olika leder, såsom knä- och fotleder. Vid en knäkontraktur kan inte benet sträckas fullt ut, så att det blir permanent böjt. Detta in sin tur kan leda till smärta, och påverka bland annat förmågan att gå och stå.
I Sverige finns det ett uppföljningsprogram för personer med cerebral pares (CPUP). När det startades 1994 i de södra delarna av landet inkluderades endast barn, men sedan 2005 är det ett nationellt kvalitetsregister, och sedan 2009 inkluderas även vuxna. Programmet inkluderar bland annat mätningar rörande förekomsten av kontrakturer.
Detta kandidatarbetes syfte var att analysera data från CPUP, för att undersöka huruvida förekomsten av en knäkontraktur på det ena benet kan påverka risken att också utveckla en knäkontraktur på det andra benet. Flera möjliga matematiska modeller för att skildra situationen undersöktes. Först beskrivs den bakomliggande teorin, varefter de olika modellerna tillämpas på datan.
På grund av de olika förenklande antaganden som måste göras, finnes alla modeller ha stora begränsningar, och resultaten av undersökningarna bör ses mer som första indikatorer samt inspiration till mer forskning. Förslag på hur modellerna skulle kunna förbättras och tas vidare ges också. Preliminärt tycks risken öka för ytterligare knäkontraktur när en redan har skett, men innan detta tas som fakta bör detta undersökas vidare.
Det är samtidigt värt att nämna, att skulle resultaten bekräftas i vidare forskning, så skulle det ytterligare framhäva vikten av förebyggande åtgärder, då vad undersökningarna här antyder är en slags snöbollseffekt. Även om effekten inte skulle visa sig, eller till lika stor grad som här, så lär det ändå vara gynnsamt att undersöka och applicera förebyggande åtgärder, som är säkra och effektiva. Trots allt, om de förebyggande åtgärderna utförs för båda benen, så får det direkt effekt, även om den ytterligare effekt av att förebygga kontraktur på respektive andra ben som antyds av modellerna inte skulle visa sig hålla vid vidare studier.

Abstract

Cerebral palsy is an umbrella term for a group of neurological disorders, affecting motor function, movement and posture. Contractures, restrictions in the range of motion of joints, are a common problem affecting people living with cerebral palsy. In the present thesis, the effect of having a contracture on one knee, on the hazard for developing a contracture on the other knee, is explored. This is done through various kinds of survival analysis, incorporating multi-state modeling and time-varying covariates respectively. While all models seem to suggest an increased hazard, direct interpretation is cautioned against, since all models suffer from various simplifications and flaws. Suggestions for further research, and improvements to the models, are also given.

Contents

1 Introduction **9**

2 Theory **11**
2.1 Survival Analysis 11
2.1.1 Censoring 11
2.1.2 The Survival Function 12
2.1.3 Proportional Hazards 13
2.2 Multi-state Survival Analysis 23
2.2.1 Non-Parametric Estimation 24
2.2.2 Proportional Hazards 26
2.2.3 Panel-Type Data 27

3 Analysis of data from the CPUP **31**
3.1 Background 31
3.2 The data set 33
3.3 Multi-state analysis 35
3.3.1 Models using the msm package 36
3.3.2 Models using the survival package 39
3.4 Time-varying covariate models 42
3.4.1 Time until the first time both knees have a contracture . . 43
3.4.2 Time to contracture on one knee, given status on the other knee 44

4 Discussion **49**

A Graphs for model evaluation of the msm models **53**

B Extra figures for the survival package multi-state models **57**

C Extra figures for the time-varying covariate models **61**

Chapter 1

Introduction

Cerebral palsy (CP) is an umbrella term for a group of neurological disorders, affecting motor function, movement and posture, and is the most common motor disability in children [1]. A common problem affecting people living with cerebral palsy is so-called contractures, which occur because the muscles become shortened, and restrict range of motion in a joint they are attached to.
The Swedish cerebral palsy follow-up program (CPUP) is a certified national health care registry, originally only for children living with cerebral palsy, but since 2009 also including adults. Regularly scheduled clinical examinations are performed as part of the program, the data of which is stored in the registry. These data include information about the range of motion of several joints, including the knee joints.
In the doctoral thesis by Cloodt [1], data from the CPUP was used to study the time from a first to second contracture in the same leg, based on for which joint the first contracture occurred. In the present thesis, the aim is to study the data from a different angle. Precisely, to try to model the potential effect of having a contracture on the knee joint of one leg, on the risk of developing a contracture on the knee joint of the other leg.
Several different approaches to the problem are explored in what follows, all unified by originating from the field of survival analysis. Survival analysis is the study of data pertaining to the time to some event — the event in question here being some variant of a contracture on one (or two) knee(s). Some of the underlying theory of survival analysis, including for the problem valuable extensions into multi-state models, is presented in chapter 2. Then, in chapter 3, further background on cerebral palsy and the data set used is presented, followed by four different attempts at analysing the data. Throughout, and summarised and expanded upon in chapter 4, the models are critiqued, and potential ways of improving them suggested.

Chapter 2

Theory

Survival analysis, broadly speaking, is the study of the amount of time until some event happens. This event can, in principle, be any type of event — progression of a disease, the arrival of spring, the breakdown of a machine — but as the name suggests, death started out as and remains an important event of interest. Often, the exact event time is not known — observations can be censored, or truncated. Two main approaches in survival analysis is the study of the survival distribution itself, often non-parametrically, as well as the study of factors influencing the survival time, or more accurately, affecting the associated hazard function.
Since the inception of the field, numerous extensions and refinements have emerged, including models of competing risks, where the time to any one of multiple outcomes is studied, as well as multi-state models, allowing for more complex sequences of events. Methods have also been developed to allow for different modes of censoring, as well as time-varying covariates, among other examples. In the present chapter, we will begin with introducing some concepts of classical survival analysis, before moving on to multi-state models.

2.1 Survival Analysis

2.1.1 Censoring

To study the time to an event, we would like to have, for each subject under observation, a clear starting time, and the time when the event happens. Often, we do not have such complete data. Hosmer, Lemeshow and May [2] point out that there are two reasons for this kind of incompleteness, namely censoring and truncation. As they explain, censoring is due to factors that are individual and random for each subject, while truncation is due to study design. There are three kinds of censoring: *right censoring*, where the event has not yet occurred at the

final observation time; *left censoring*, where the event has already occurred before observation begins; and *interval censoring*, where the event is only known to have occurred between two time points. Right censoring is the most common, and will as such be our primary focus for introducing the concepts, but we will return later to interval censored data, as that is a more accurate model for the data set at hand. In the present section, we will draw upon Moore [3] and [2] for notation and definitions, and refer to these sources for a more detailed account.
We have, in right censored data, two underlying random variables: T^* for the *time to event*, and C for the *time to censoring*. For an individual, we can only observe one of these times, and we create a new random variable T such that $T = \min(T^*, C)$. We also know $\lambda = \mathbb{1}_{\{T^* \leq C\}}$, where $\mathbb{1}$ is the indicator random variable. Thus, λ is 1 whenever an event has occurred during the time under observation, 0 if not. Censoring can be further classified into types, as in [3]; most importantly, to not introduce bias, the censoring mechanism needs to be independent from the event process itself.

2.1.2 The Survival Function

Fundamental to survival analysis is the *survival function* itself:

$$S(t) = \mathbb{P}(T > t) \quad \text{for } 0 < t < \infty. \tag{2.1.1}$$

It is the probability that an event happens later than time t. Note that $S(t) = 1 - F(t)$, where $F(t) = \mathbb{P}(T \leq t)$ is the *cumulative distribution function* of the random variable for time T, and as such, $F(t) + S(t) = 1$ for all $t \geq 0$. In survival analysis, this cumulative distribution function is often called the *cumulative risk function* [3]. Further, we have the *hazard function*

$$h(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t < T < t + \delta t | T > t)}{\delta t}, \tag{2.1.2}$$

the probability that an event happens in an infinitesimal interval after time t, given that the survival time is greater than t, divided by the length of the interval. Assuming that the underlying time random variable is absolutely continuous, we have $f(t)$, the *probability density function* corresponding to $F(t)$, that is

$$f(t) = \frac{d}{dt} F(t) = -\frac{d}{dt} S(t). \tag{2.1.3}$$

Then, the hazard function is related to the survival function as

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{\mathbb{P}(t < T < t + \delta t | T > t)}{\delta t} = \lim_{\delta t \to 0} \frac{\mathbb{P}(t < T < t + \delta t, T > t)}{\delta t \mathbb{P}(T > t)} \\ &= \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)}. \end{aligned} \tag{2.1.4}$$

From the hazard function, we can then construct the *cumulative hazard function*

$$H(t) = \int_0^t h(u)\,du. \tag{2.1.5}$$

We can then in turn express the survival function as

$$S(t) = \exp\big[-H(t)\big]. \tag{2.1.6}$$

The most widely used non-parametric estimator for the survival function is the *Kaplan–Meier estimator*, first presented by Kaplan and Meier [4]. We briefly state its form here, and refer to [2] [3] [4] for more details. Say that we are studying n individuals, for whom we have the observed times t_i, $i = 1, \ldots, n$ of instances of the random variable T from above, and say that we observe m events. Let $t_{(i)}$ for $i = 1, \ldots, m$ be the *ordered* observed survival times. Then the Kaplan–Meier estimator of the survival function in (2.1.1) is

$$\widehat{S}(t) = \prod_{t_{(i)} \leq t} \frac{n_{(i)} - d_{(i)}}{n_{(i)}}, \tag{2.1.7}$$

where $n_{(i)}$ is the number of subjects at risk of an event at time $t_{(i)}$, and $d_{(i)}$ the number of individuals experiencing an event at that time.
Another non-parametric estimator of the survival function, which will become very important especially as we move on to the multi-state part of the theory, is the *Nelson–Aalen estimator*. Relying on the assumption that the time random variable T is absolutely continuous, it is derived through first deriving an estimator for the cumulative hazard function (2.1.5), and then using its relation to the survival function in (2.1.6). While details of the counting process approach to deriving the Nelson–Aalen estimator are not given directly in [2], several references are provided on page 59. Briefly, the Nelson–Aalen estimator of the cumulative hazard is

$$\widetilde{H}(t) = \sum_{t_{(i)} \leq t} \frac{d_{(i)}}{n_{(i)}}, \tag{2.1.8}$$

and the corresponding Nelson–Aalen estimator of the survival function is thus

$$\widetilde{S}(t) = \exp\big[-\widetilde{H}(t)\big]. \tag{2.1.9}$$

2.1.3 Proportional Hazards

Suppose that we have a set of covariates $\mathbf{x} = (x_1, \ldots, x_p)$ for each subject under study, and that we want to investigate how these influence survival. For reasons discussed at more length in [2], the hazard function (2.1.2) is often chosen as the

subject of regression modeling in the survival setting. As also discussed in [3], the primary interest might not be the survival or hazard functions themselves, but rather how they differ between groups with different values of the covariates. As pointed out in [2], in that case we might not need a full parametric model of the hazard function, but rather a semi-parametric one might suffice. The most prominent such model was proposed by Cox [5] in 1972, and relies on the assumption that hazard functions are proportional to one another, with a hazard ratio that is constant with respect to time. It is as such often called the *Cox proportional hazards model*. Notation in what follows is inspired by [5] [2] [3]. According to this model, we can write

$$h(t, \mathbf{x}, \boldsymbol{\beta}) = r(\mathbf{x}, \boldsymbol{\beta}) h_0(t) = \exp(\mathbf{x}\boldsymbol{\beta}) h_0(t), \tag{2.1.10}$$

where $\mathbf{x}$ is a row vector as before, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a column vector of p unknown parameters, and $h_0(t)$ is an unknown, baseline hazard function. Notice that, in contrast to in linear regression, we have no β_0 or "intercept" term; the baseline hazard function could be seen as fulfilling such a role. The form of the function for how the hazard changes as the covariates change, $r(\mathbf{x}, \boldsymbol{\beta}) = \exp(\mathbf{x}\boldsymbol{\beta})$, is the one used by Cox [5] and also the most widely used since then according to [2], although in theory other known functions could be used. Note that, as promised, the hazard ratios between hazards with parameters $\mathbf{x}_1$ and $\mathbf{x}_0$ take the time-independent form

$$\mathrm{HR}(t, \mathbf{x}_1, \mathbf{x}_0) = \frac{\exp(\mathbf{x}_1\boldsymbol{\beta}) h_0(t)}{\exp(\mathbf{x}_0\boldsymbol{\beta}) h_0(t)} = \frac{\exp(\mathbf{x}_1\boldsymbol{\beta})}{\exp(\mathbf{x}_0\boldsymbol{\beta})} = \exp\big((\mathbf{x}_1 - \mathbf{x}_0)\boldsymbol{\beta}\big).$$

As in linear regression, we would like to use maximum likelihood estimation to estimate the parameters. We first need a likelihood function for the situation at hand. We assume that the time random variable T is uniformly continuous, so that the probability density (2.1.3) exists. If $\lambda_i = 1$ for observation i, we have observed the exact time of event t_i. Since the probability density function is the probability, per unit of time, that an event happens in a neighbourhood of time t_i, we use $f(t_i; \boldsymbol{\beta}; \mathbf{x}_i)$ as that observation's contribution to the likelihood. If on the other hand $\lambda_j = 0$ for observation j, we know only that the survival time is greater than t_j, which is exactly what the survival function represents; for such an observation we thus use $S(t_j; \boldsymbol{\beta}; \mathbf{x}_j)$ as the contribution to the likelihood. If we assume further that all observations are independent, the likelihood for full maximum likelihood estimation would be, also using the relationship between

probability density, hazard and survival functions (2.1.4):

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} f(t_i;\boldsymbol{\beta};\mathbf{x}_i)^{\lambda_i} S(t_i;\boldsymbol{\beta};\mathbf{x}_i)^{1-\lambda_i} \\ &= \prod_{i=1}^{n} \Big[h(t_i;\boldsymbol{\beta};\mathbf{x}_i) S(t_i;\boldsymbol{\beta};\mathbf{x}_i) \Big]^{\lambda_i} S(t_i;\boldsymbol{\beta};\mathbf{x}_i)^{1-\lambda_i} \\ &= \prod_{i=1}^{n} h(t_i;\boldsymbol{\beta};\mathbf{x}_i)^{\lambda_i} S(t_i;\boldsymbol{\beta};\mathbf{x}_i). \end{aligned} \tag{2.1.11}$$

Since the logarithm is a non-decreasing function, we can attempt to maximize the log-likelihood rather than the likelihood function itself, in order to obtain estimates of our parameters. Before moving on, we also establish explicitly how the survival function depends on the parameters, using its relationship to the cumulative hazard (2.1.6) and the proportional hazards model (2.1.10). We get

$$\begin{aligned} S(t;\boldsymbol{\beta};\mathbf{x}) &= \exp\Big[-H(t;\boldsymbol{\beta};x) \Big] = \exp\Big[-\int_0^t h(u;\boldsymbol{\beta};\mathbf{x})\, du \Big] \\ &= \exp\Big[-\int_0^t h_0(u) \exp(\mathbf{x}\boldsymbol{\beta})\, du \Big] = \exp\Big[-\exp(\mathbf{x}\boldsymbol{\beta}) \int_0^t h_0(u)\, du \Big] \\ &= \exp\Big[-\exp(\mathbf{x}\boldsymbol{\beta}) H_0(t) \Big] = \Big\{ \exp\Big[-H_0(t) \Big] \Big\}^{\exp(\mathbf{x}\boldsymbol{\beta})} = \Big[S_0(t) \Big]^{\exp(\mathbf{x}\boldsymbol{\beta})}, \end{aligned}$$

where $H_0(t) = \int_0^t h_0(u)\, du$ is the baseline cumulative hazard, and $S_0(t) = \exp\big[-H_0(t)\big]$ is the baseline survival function. Then, taking the logarithm of the likelihood (2.1.11) gives

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \lambda_i \log\Big[h(t_i;\boldsymbol{\beta};\mathbf{x}_i) \Big] + \log\Big[S(t_i;\boldsymbol{\beta};\mathbf{x}_i) \Big] \\ &= \sum_{i=1}^{n} \left(\lambda_i \log\Big[h_0(t_i) \exp(\mathbf{x}_i\boldsymbol{\beta}) \Big] + \log\Big\{ \Big[S_0(t_i) \Big]^{\exp(\mathbf{x}_i\boldsymbol{\beta})} \Big\} \right) \\ &= \sum_{i=1}^{n} \Big\{ \lambda_i \log\Big[h_0(t_i) \Big] + \lambda_i \mathbf{x}_i\boldsymbol{\beta} + \exp(\mathbf{x}_i\boldsymbol{\beta}) \log\Big[S_0(t_i) \Big] \Big\}. \end{aligned} \tag{2.1.12}$$

Maximizing the log-likelihood (2.1.12) would require taking into account also the baseline hazard function and the baseline survival function, in addition to the parameters of interest. This would not only be hard, but according to [2] is even impossible. In [5], Cox proposed using instead a partial, or in his words conditional, likelihood. As [2] points out, proofs came later of the fact that the estimators derived from maximizing this partial likelihood have the same properties

as usual full maximum likelihood estimators have; here, we take this for granted and refer to other sources for proofs (see e.g. the references on pp. 74–75 in [2]). We assume that among n observations, we have m events, and further assume that there are no tied event times. Following Cox [5] we argue conditionally on the ordered observed event times $t_{(i)}$ for $i = 1, \ldots, m$. Let $R[t_{(i)}]$ be the set of individuals at risk at time $t_{(i)}$. These are all individuals who have not yet had an event, nor have been censored. Then, for the event at time $t_{(i)}$, conditionally on the associated risk set, the probability that an event happens for the individual as observed, is

$$\begin{aligned} \frac{h_{(i)}\left[t_{(i)}; \boldsymbol{\beta}; \mathbf{x}_{(i)}\right]}{\sum_{j \in R[t_{(i)}]} h_{(j)}\left[t_{(i)}; \boldsymbol{\beta}; \mathbf{x}_{(j)}\right]} &= \frac{h_0[t_{(i)}] \exp[\mathbf{x}_{(i)}\boldsymbol{\beta}]}{\sum_{j \in R[t_{(i)}]} h_0[t_{(i)}] \exp[\mathbf{x}_{(j)}\boldsymbol{\beta}]} \\ &= \frac{\exp[\mathbf{x}_{(i)}\boldsymbol{\beta}]}{\sum_{j \in R[t_{(i)}]} \exp[\mathbf{x}_{(j)}\boldsymbol{\beta}]}. \end{aligned} \tag{2.1.13}$$

The partial likelihood is then the product of the terms in (2.1.13), for $i = 1, \ldots, m$:

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp[\mathbf{x}_{(i)}\boldsymbol{\beta}]}{\sum_{j \in R[t_{(i)}]} \exp[\mathbf{x}_{(j)}\boldsymbol{\beta}]}, \tag{2.1.14}$$

and, taking the logarithm of (2.1.14), the partial log-likelihood is thus

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{m} \mathbf{x}_{(i)}\boldsymbol{\beta} - \sum_{i=1}^{m} \log\Big\{ \sum_{j \in R[t_{(i)}]} \exp[\mathbf{x}_{(j)}\boldsymbol{\beta}] \Big\}. \tag{2.1.15}$$

In order to find the values of $\boldsymbol{\beta}$ that maximize the partial log-likelihood, we take the derivative of (2.1.15) with respect to a specific parameter β_k, which we do for all $k = 1, \ldots, p$. We then solve for β_k when the derivative equals 0. For later use, we call this first partial derivative of the partial log-likelihood the *score function* with respect to the kth parameter:

$$U_k(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{m} \left[x_{(ki)} - A_{(ki)}(\boldsymbol{\beta}) \right], \tag{2.1.16}$$

where

$$A_{(ki)}(\boldsymbol{\beta}) = \frac{\sum_{j \in R[t_{(i)}]} x_{kj} \exp\left[\mathbf{x}_{(j)}\boldsymbol{\beta}\right]}{\sum_{j \in R[t_{(i)}]} \exp[\mathbf{x}_{(j)}\boldsymbol{\beta}]} \tag{2.1.17}$$

can be seen as an exponentially weighted *average* of the covariate x_k, taken over the finite risk set $R[t_{(i)}]$. To get the variances of our parameters (as well as covariances), we need the *information matrix*. The (k,l)th entry of this information matrix is the

negative of the second partial derivative of the partial log-likelihood (2.1.15) with respect to the kth and lth parameters,

$$\mathbf{I}_{kl}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_k \partial\beta_l} = \sum_{i=1}^{m} C_{(kli)}(\boldsymbol{\beta}). \tag{2.1.18}$$

Here, the terms in the sum represent the covariance of covariates x_k and x_l, using exponential weights as for the averages, and take the form

$$C_{(kli)}(\boldsymbol{\beta}) = \frac{\sum_{j\in R[t_{(i)}]} x_{kj} x_{lj} \exp[\mathbf{x}_{(j)}\boldsymbol{\beta}]}{\sum_{j\in R[t_{(i)}]} \exp[\mathbf{x}_{(j)}\boldsymbol{\beta}]} - A_{(ki)}(\boldsymbol{\beta}) A_{(li)}\boldsymbol{\beta}. \tag{2.1.19}$$

Finally, the estimated covariance matrix of the estimators of the parameters, is the inverse of the information matrix with entries (2.1.18), evaluated at the values of the estimators:

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \left[\mathbf{I}(\widehat{\boldsymbol{\beta}})\right]^{-1}. \tag{2.1.20}$$

Time-Varying Covariates

It is also possible to incorporate *time-varying* covariates into a proportional hazards model. A time-varying covariate is a covariate such that its value will be different at different time points — for example, it could represent having had a certain type of surgery, or be the value of some medical measurement. As emphasised in the manual for using time dependent covariates in the survival package for R [6], as well as in [3] and [2], one has to be careful when using covariates that vary with time, so that we do not, for example, by mistake use a future value of a covariate. This could also happen if a time-varying covariate is treated as fixed from the beginning, i.e. encoding the variable for having had a surgery as a baseline covariate, when in reality it might happen later in study time. For a greater quantity of, and more detailed, examples, see the above sources.
Further, as pointed out in [3], using time-varying covariates creates *internal left-truncated data*. Left-truncation is explained in [2] as delayed entry, i.e. a subject enters the study at a later time than the defined starting point. This happens with time-varying covariates as a subject's data will be split at each time the covariates change values. Essentially, we may view their data as coming from two different subjects, one with the covariate values before the split, who might have been observed from the start of the study, and one who enters the study only at the time of the covariate change. That the left-truncation is internal simply means that it is due to the nature of the covariates, not due to the observation scheme (which could then be called *external* left-truncation). This is illustrated in figure 2.1.

Let $\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))$ be the set of covariates for a subject under observation. Note that this notation still works with covariate values that do not vary

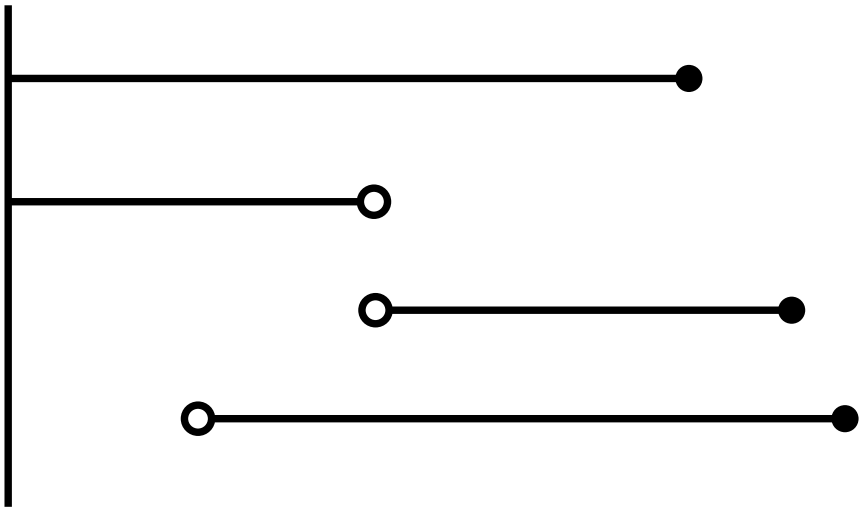

Figure 2.1: Examples of individual observations in the time-varying covariate setting. Here, the left-hand vertical line represents the defined starting time. From above, we have first an individual who we start observing at the defined starting time, and whose covariate values do not change, and is observed until an event occurs (the black dot). Next, we have an individual whose covariate values change part-way during observation. This is represented by using two lines to represent the same individual, split at the time when the covariates change (the hollow dots on both lines). Because an event is observed for this individual, their second line ends in a black dot. This is an example of internal left-truncation. Finally, we have an individual who experiences external left-truncation, and simply enters observation later than the defined starting time, and who experiences an event at the end.

with time; if the kth covariate is constant over time as earlier, then we simply have $x_k(t) = x_k(0) = x_k$, for all $t \geq 0$. Then the Cox proportional hazards model in (2.1.10) generalises to

$$h\left[t, \mathbf{x}(t), \boldsymbol{\beta}\right] = r\left[\mathbf{x}(t), \boldsymbol{\beta}\right] h_0(t) = \exp\left[\mathbf{x}(t)\boldsymbol{\beta}\right] h_0(t) \tag{2.1.21}$$

and the partial likelihood in (2.1.14) generalises to

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left\{\mathbf{x}_{(i)}[t_{(i)}]\boldsymbol{\beta}\right\}}{\sum_{j \in R[t_{(i)}]} \exp\left\{\mathbf{x}_{(j)}[t_{(i)}]\boldsymbol{\beta}\right\}}. \tag{2.1.22}$$

Note, especially, that in the sum, not only the sum itself, but each individual term, (may) have to be recalculated at each event time i, since the values of the covariates (can) change with time.

Hypothesis Tests

As in any form of regression modeling, we would like to test the null hypothesis that the parameters are equal to zero, all at once as well as individually. In general, let $\boldsymbol{\beta}$ be a vector of p parameters. Then if we want to test $H_0 : \boldsymbol{\beta} = \mathbf{0}$, there are three main tests in use. These are the Wald test, the score test, and the likelihood ratio test. Under the null hypothesis, assumptions of the proportional hazards model itself, and given "enough" uncensored observations, all corresponding test statistics asymptotically follow chi-square distributions with p degrees of freedom — for an extended account, and references to more rigorous details, see [2], pp. 77–85. For hypothesis tests pertaining to individual covariates, "perhaps the most commonly used test" [3] is the Wald test. The likelihood ratio test is, on the other hand, according to [2] the preferred test if the tests are in disagreement, as well as for the multivariate setting since the other tests involve extensive matrix calculations.
The *Wald test statistic* is, slightly informally, the square of the vector of estimators, multiplied by the inverse of their estimated covariance matrix (i.e. the information matrix (2.1.18) evaluated at the values of the estimators). For single-parameter hypothesis tests, sometimes the Wald test is performed with the test statistic being the single estimator, divided by the square root of its estimated variance (its *standard error*), which is then distributed asymptotically as a standard normal random variable, under the null hypothesis. For the general case, we write

$$\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{I}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\beta}}. \tag{2.1.23}$$

The *score test statistic* is unique in that it does not use the estimated values of the parameters. Instead, it uses the (vector of) the score function (2.1.16) and the

information matrix (2.1.18), both evaluated at the value of the parameters under the null hypothesis, $\boldsymbol{\beta} = \mathbf{0}$. In the single-parameter hypothesis test, the statistic is sometimes given as the score function divided by the square root of the information, evaluated at $\beta = 0$, which is then asymptotically distributed as a standard normal random variable. In the general case, it is given by

$$\mathbf{U}^{\mathrm{T}}(\mathbf{0})\big[\mathbf{I}(\mathbf{0})\big]^{-1}\mathbf{U}(\mathbf{0}). \tag{2.1.24}$$

Finally, the *likelihood ratio test statistic*, or more accurately the *partial log*-likelihood ratio test statistic, is simply given as twice the difference between two partial log-likelihoods as in (2.1.15), evaluated at $\widehat{\boldsymbol{\beta}}$ and $\mathbf{0}$ respectively:

$$2\big[l(\widehat{\boldsymbol{\beta}}) - l(\mathbf{0})\big]. \tag{2.1.25}$$

The likelihood ratio test can also be used to compare two nested models, with parameters respectively $\boldsymbol{\beta}_{full}$ and $\boldsymbol{\beta}_{reduced}$, where all the parameters in the reduced model are also contained in the full model. Then the test statistic analogous to (2.1.25),

$$2\big[l(\widehat{\boldsymbol{\beta}}_{full}) - l(\widehat{\boldsymbol{\beta}}_{reduced})\big],$$

is asymptotically distributed as a chi-square random variable with $p_{full} - p_{reduced}$ degrees of freedom.
As is standard practice, we will consider a p-value of 0.05 or below to be adequate for statistical significance, allowing us to reject the null hypothesis under question.

Model Evaluation

Since everything so far has relied on the assumptions of the model being true, it is of utmost importance to attempt to check whether these truly hold. To this end, various kinds of residuals have been developed. As pointed out in [2], defining residuals is not as straightforward in the survival setting as in linear or logistic regression. The true value of the "outcome", survival time, is often not known due to censoring. Further, the fitted model does *not* provide an estimate of the mean of the outcome variable. As such, there is no immediate analogue to the usual observed–versus–predicted residual. We here mention a few residuals discussed at more length in [2] and [3]. For the following, assume that we have p covariates, and n independent observations of time, covariates and censoring indicators $(t_i, \mathbf{x}_i, \lambda_i)$. The *Schoenfeld residuals* can be seen as individual contributions to the derivative of the partial log-likelihood, i.e. the score function (2.1.16):

$$\widehat{r}_{ik} = \lambda_i\big[x_{ki} - \widehat{A}_{ki}(\widehat{\boldsymbol{\beta}})\big], \tag{2.1.26}$$

where

$$\widehat{A}_{ki}(\widehat{\boldsymbol{\beta}}) = \frac{\sum_{j \in R(t_i)} x_{kj} \exp(\mathbf{x}_j \widehat{\boldsymbol{\beta}})}{\sum_{j \in R(t_i)} \exp(\mathbf{x}_j \widehat{\boldsymbol{\beta}})} \tag{2.1.27}$$

is the *estimator of* the exponentially weighted average of the covariate x_k taken over the risk set $R(t_i)$ as in (2.1.17), or in the terms of [2], "the estimator of the risk set conditional mean of the covariate" (p. 171). Note that they are covariate-specific. Since the $\widehat{\boldsymbol{\beta}}$ are calculated as the values for which the score function is zero, the Schoenfeld residuals sum up to 0 over all individuals $i = 1, \ldots, n$. Further, since the partial likelihood did not include censored observations, the Schoenfeld residuals are often said to be undefined for all i such that $\lambda_i = 0$. If $\widehat{\mathbf{r}}_i = (\widehat{r}_{i1}, \ldots, \widehat{r}_{ip})$ is the vector of Schoenfeld residuals for individual i, then the *scaled Schoenfeld residuals* are given as

$$\widehat{\mathbf{r}}_i^* = \left[\widehat{\mathrm{Var}}(\widehat{\mathbf{r}}_i)\right]^{-1} \widehat{\mathbf{r}}_i \approx m \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) \widehat{\mathbf{r}}_i, \tag{2.1.28}$$

most often using the approximation of the inverse of the estimated variance of the estimated Schoenfeld residuals as given, where m is the observed number of events.

The Schoenfeld residuals are used to test the proportional hazards assumption, essential to the entire model. Taking the logarithm of the model in (2.1.10), we get

$$\log\left[h(t, \mathbf{x}, \boldsymbol{\beta})\right] = \log[h_0(t)] + \mathbf{x}\boldsymbol{\beta}. \tag{2.1.29}$$

Thus, using (2.1.29) we can view $\mathbf{x}\boldsymbol{\beta}$ as a linear predictor and the model as a function of time. It is clear that these log-hazards should have a constant difference between them over time, given fixed, different covariate values, if the model is correct. If not, and the covariates change with time in a particular way, we write, for covariate k,

$$\beta_k(t) = \beta_k + \gamma_k g_k(t), \tag{2.1.30}$$

where $g_k(t)$ is some specified function of time, and γ_k is a coefficient. Then it turns out that the scaled Schoenfeld residuals have an expected value approximately equal to the time-varying part of (2.1.30) [2] (see also [7]):

$$\mathbb{E}\left[r_k^*(t)\right] \approx \gamma_k g_k(t). \tag{2.1.31}$$

If the proportional hazards model holds, γ_k would be 0 for all $k = 1, \ldots, p$. Plotting the scaled Schoenfeld residuals, plus the estimated parameters $\widehat{\beta}_k$, versus time, should then yield a horizontal line. If the proportional hazards model does not hold, then the plot should hint at the form of the time-dependent parameter $\beta_k(t)$. [2] suggest that these plots can be hard to interpret, and departures from proportionality hard to see, and as such recommend using formal tests. These can be performed for

specific functions $g(t)$, where some common ones are $g(t) = \log(t)$, $g(t) = \widehat{S}_{KM}(t)$ (the Kaplan–Meier estimator of the survival function (2.1.7)), $g(t) = \text{rank}(t)$ or $g(t) = t$. These are simply performed by adding the term $x_j g_j(t)$ to a proportional hazards model, and using any of the partial log-likelihood ratio (2.1.25), score (2.1.24) or Wald (2.1.23) test statistics from the subsection on hypothesis tests 2.1.3. Note that, as discussed in the subsection on time-varying covariates 2.1.3, the partial likelihood becomes much more complicated when we have a time-varying interaction added to the model.
For a more complete account of the other residuals, we refer to [2] and [3]. Briefly, the *martingale residuals* are derived from the counting process approach to survival analysis, and are given, for all individuals $i = 1, \ldots, n$, as

$$\widehat{M}_i = \lambda_i - \widehat{H}_0(t_i) \exp(\mathbf{x}_i \widehat{\boldsymbol{\beta}}), \tag{2.1.32}$$

where $\widehat{H}_0(t_i)$ is an estimator of the baseline cumulative hazard function, as detailed in [2] (pp. 87–90). Out of all the survival analysis residuals, these resemble most the difference between an observed and expected value of the model, and can be used much like residuals in linear regression. Plotted versus individual covariates, they can reveal discrepancies in the model, and specifically if used with a null model, the functional forms of (continuous) covariates, in which case a transformation is necessary. If the model is correct, the martingale residuals sum to zero, $-\infty < \widehat{M}_i \leq 1$, and $\mathbb{E}(M_i) = 0$.
The *score residuals* are derived also from the counting process approach, and involve re-expressing the score function (2.1.16) in such a way that we have

$$U_k(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{n} L_{ik}. \tag{2.1.33}$$

Then the estimates of each term L_{ik} in (2.1.33) are the score residuals for individual i and covariate k, the expression of which is rather complex, but specified as (6.16) on page 176 of [2]. They also appear in a scaled form, such that $\widehat{\mathbf{L}}_i^* = \widehat{\text{Var}}(\widehat{\boldsymbol{\beta}})\widehat{\mathbf{L}}_i$. The score residuals function somewhat like leverage residuals in linear regression, in that they can help identify subjects with unusual covariate values, whereas the scaled score residuals work like Cook's distance, in that they indicate a subject's influence on a particular parameter value. Through a transformation, the score residuals can yield the so-called *dfbeta* residuals, which are approximations of the change in the value of the estimate of a parameter, and its value if that individual observation was dropped; the same scaling as for the score residuals themselves yield the standardised *dfbetas*.

2.2 Multi-state Survival Analysis

Having some baseline knowledge of general survival analysis, we are now ready to move on to the multi-state case. The material in this section relies heavily on [8] [9] [10]. In a multi-state model, we are not just studying the time to one event. Rather, we can study the time to one of several events — also called a *competing risks* model; or we can study a sequence of transitions between states — for example, progression of some disease through several states; we can also study a situation where it is possible to transition backwards between states — for example modeling recovery of a disease, or repair of a machine; or indeed any combination of the above. Even the ordinary survival setting can be viewed as a special case of a multi-state model, with two states and one forwards transition.
Let $\Lambda(t)$ be the state an individual is in at time t. Note that, if we in the ordinary survival setting label the state for "no event has happened yet" as 0, and label the state for "an event has happened" as 1, then the censoring indicator λ corresponds to $\Lambda(t)$, evaluated at the time $t = T$ where the individual was last observed. In general, $\Lambda(t)$ can take any of the values in the finite state space $\{0, \dots, J-1\}$, if we have J possible states in the model.
Analogous to the hazard function (2.1.2) from the ordinary survival analysis setting, we here have *transition hazards* $q_{lj}[t, \mathbf{x}(t)]$ for going from one state l to another state j, for $l, j \in \{0, \dots, J-1\}$. These may depend on the current time t, and/or a set of (possibly time-varying) covariates $\mathbf{x}(t)$. On the other hand, in the multi-state setting we usually rely on the *Markov assumption*, that the transition hazards *only* depend on these and the current state, i.e. that $q_{lj}\left[t, \mathbf{x}(t), \mathcal{F}_t\right] = q_{lj}[t, \mathbf{x}(t)]$ is independent of the history of the process up until time t, denoted $\mathcal{F}_t$ (which is, more accurately, the σ-algebra generated by the the history of the process [8]). For the moment, we will ignore the covariates, until we come back to them later when we speak about proportional hazards models. Similar to for the hazards, these transition hazards can be thought of as instantaneous risks of transitioning from one state l to another state $j \neq l$, for all states $l, j \in \{0, \dots, J-1\}$:

$$q_{lj}\left[t, \mathbf{x}(t)\right] = \lim_{\delta t \to 0} \frac{\mathbb{P}\left[\Lambda(t+\delta t) = j | \Lambda(t) = l\right]}{\delta t}. \tag{2.2.1}$$

Note that, as stated above, we have a *time-inhomogeneous* model, i.e. the values of the transition hazards depend on the current time t. We will later return to the case where we assume that they are independent of time, which will be needed for panel-type data. For now, we assume that we have continuously observed data, and as such that we know exactly when transitions take place. Then, we do not have to make the restrictive assumption of time-homogeneity. In the present case, it then makes sense to speak of cumulative transition hazards, which are, analogously to

the ordinary cumulative hazard function (2.1.5), given as

$$Q_{lj}(t) = \int_0^t q_{lj}(u)\,du \tag{2.2.2}$$

for all $l, j \in \{0, \ldots, J-1\}$ such that $l \neq j$. We can collect these cumulative transition hazards in a matrix $\mathbf{Q}(t)$, where we define the diagonal elements such that rows sum to zero:

$$Q_{ll}(t) = -\sum_{j \neq l} Q_{lj}(t). \tag{2.2.3}$$

We will also need to consider the *transition probability matrix* $\mathcal{P}(u, t+u)$, the (l, j)th entry of which, for all $l, j \in \{0, \ldots, J-1\}$, is the probability of being in state j at time $t+u$, given that the current state at time u is state l. We write

$$\mathcal{P}_{lj}(u, t+u) = \mathbb{P}\big[\Lambda(t+u) = j \,|\, \Lambda(u) = l\big]. \tag{2.2.4}$$

2.2.1 Non-Parametric Estimation

As promised in section 2.1.2, the Nelson–Aalen estimator shows up and plays a very important role in non-parametric estimation in the multi-state setting. To formulate it, we need some further notation. Say that we are studying n individuals. Then we have n multi-state processes $\Lambda_i(t)$, for all $i = 1, \ldots, n$ and $t \geq 0$, which can take values in the state space such that $\Lambda_i(t) \in \{0, \ldots, J-1\}$. We assume that the observed processes are independent replicates of the same process, *conditionally* on their initial states $\Lambda_i(0)$.

In a multi-state setting, the question of which individuals are "at risk" of a transition at any point of time is more complex than in the ordinary survival setting. Depending on what transitions are allowed, one individual can enter and exit any particular risk set potentially any number of times. We define an at-risk indicator for transitions *out of* a state l, for all states $l \in \{0, \ldots, J-1\}$ and all individuals $i = 1, \ldots, n$, as

$$N_{l;i}(t) = \mathbb{1}\big\{\Lambda_i(t-) = l, L_i < t \leq C_i\big\} \tag{2.2.5}$$

where $t-$ indicates the time immediately before time t, C_i is the right-censoring time for individual i, and L_i is the left-truncation time for individual i. In the multi-state setting, we encounter left-truncation naturally due to the nature of the process; an individual is not at risk for a transition *from* a particular state, until they enter the corresponding state. As such, we can view the entry time into a state as delayed entry or internal left-truncation, as discussed briefly in section 2.1.3 on time-varying covariates. Since any one individual can only be in any one state at any one time, and as such only at risk for transitions out of one state at a time, there are no dependency concerns. Note, however, that we might have several

left-truncation times for an individual i, if they enter the same state multiple times (after having spent time in another state in between). We also let $D_{lj;i}(t)$ be the number of *direct* transitions from state l to state $j \neq l$ for individual i, in the time interval $[0,t]$.
Next, we sum the at-risk indicators and direct transition counters over all individuals, such that we have the total number of individuals at risk for transitions from state l at time t as $N_l(t) = \sum_{i=1}^{n} N_{l;i}(t)$ and the total number of observed direct l to j transitions up until time t as $D_{lj}(t) = \sum_{i=1}^{n} D_{lj;i}(t)$. Then we define increments in direct transitions as $\Delta D_{lj}(t) = D_{lj}(t) - D_{lj}(t-)$ — this is thus precisely the number of transitions observed from state l to state j exactly at time t. Then, the multi-state Nelson–Aalen estimators for the cumulative transition hazards are:

$$\widetilde{Q}_{lj}(t) = \sum_{u \leq t} \frac{\Delta D_{lj}(u)}{N_l(u)}, \tag{2.2.6}$$

where the sum is over all observed transition times u in the time interval $[0,t]$. Note the similarity between the multi-state (2.2.6) and ordinary (2.1.8) Nelson–Aalen estimators. The number of individuals at risk are now split between several (rather than two — an event has happened or not) states at any one time, but is otherwise similar; and the number of events d now correspond to number of transitions between two specific states in ΔD_{lj}. Its form is motivated informally in [8] on page 178. Essentially, it is the sum of estimated hazard increments, with as fine a partition of time as we can get, given the observed data.
From the Nelson–Aalen estimators of the cumulative transition hazards (2.2.6) we can get the so-called *Aalen–Johansen estimator*, or *empirical transition matrix*, of the matrix of transition probabilities. Let us say that we have M transitions (between any pair of states) in the interval $(u, t+u]$. Further, define the observed increments in the cumulative transition hazards as $\Delta Q_{lj}[t_{(m)}] = Q_{lj}[t_{(m)}] - Q_{lj}[t_{(m-1)}]$, and let I be the $J \times J$ identity matrix. Note that, because of the definition of $Q_{ll}(t)$ (2.2.3), and from how we calculated the increments in the Nelson–Aalen estimator (2.2.6), all rows in $\mathrm{I} + \Delta Q[t_{(m)}]$ sum to 1, and can thus be seen as a (transition) probability matrix, where the (l,j)th entry is $\mathbb{P}\{\Lambda[t_{(m)}] = j | \Lambda[t_{(m-1)}] = l\}$. Finally, using the Markov assumption, we get that an estimator of the matrix of transition probabilities is

$$\widetilde{\mathcal{P}}(u, t+u) = \prod_{m=1}^{M} \left\{ \mathrm{I} + \Delta \widetilde{\mathbf{Q}}[t_{(m)}] \right\}. \tag{2.2.7}$$

Note that, unlike the estimators for the survival function in section 2.1.2, the Aalen–Johansen estimators of transition probabilities are *conditional* probabilities. Thus, if we would want to compare the transition probabilities from, say, state 0

and state 1 respectively, to another state 2, we can only do so for specific starting and ending times. One way to still get an idea of the differences between the transition probabilities is to select several starting times u, and comparing the transition probabilities up to some final common time t conditional on the state occupied at time u. This is referred to in [8] as "the 'landmark method'", with further references and an example application on page 187.

2.2.2 Proportional Hazards

Suppose, similarly to in the ordinary survival setting in subsection 2.1.3, that we have sets of covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ for each subject $i = 1, \ldots, n$, with column vectors of regression coefficients $\boldsymbol{\beta}_{lj} = (\beta_{lj;1}, \ldots, \beta_{lj;p})^{\mathrm{T}}$ for each transition from a state l to another state j. Assume that, as in the previous subsection 2.2.1, that for each individual i we have a multi-state process $\Lambda_i(t)$, and that these are *conditionally* independent, given baseline covariate values and initial states. Assume also that right-censoring and left-truncation are independent of the processes. Then the multi-state proportional hazards model gives, analogously to the model in (2.1.10), but for all individual transitions, that

$$q_{lj;i}[t, \mathbf{x}_i, \boldsymbol{\beta}_{lj}] = \exp[\mathbf{x}_i \boldsymbol{\beta}_{lj}] q_{lj;0}(t) \tag{2.2.8}$$

for all $l, j \in \{0, \ldots, J-1\}$, such that $l \neq j$, and $i = 1, \ldots, n$, where $q_{lj;0}(t)$ is the (unspecified) baseline transition hazard for the l to j transition. We also recall from the previous subsection the at-risk indicators $N_{l;i}(t)$ and the increments in observed direct transitions $\Delta D_{lj;i}$. To allow for potentially shared coefficients between different transitions, we can reformulate such that we get all transition coefficients in one mutual vector $\boldsymbol{\beta}$, instead having transition-specific covariate variables, $x_{lj;i}$ (where a component of a variable corresponding to a coefficient that only applies for another transition simply is 0). Then, the multi-state analogue to the partial likelihood (2.1.14) is given by

$$L_p(\boldsymbol{\beta}) = \prod_t \prod_{i=1}^{n} \prod_{l=0}^{J-1} \prod_{j \neq l} \left[\frac{\exp(\mathbf{x}_{lj;i} \boldsymbol{\beta})}{\sum_{i=1}^{n} \exp(\mathbf{x}_{lj;i} \boldsymbol{\beta}) N_{l;i}(t)} \right]^{\Delta D_{lj;i}(t)}, \tag{2.2.9}$$

where the first product is taken over all observed transition times t. Since the risk sets are more complicated in the multi-state setting, as well as the fact that one individual might very well experience multiple transitions, we here take the products and sums over all individuals, and adjust which individuals are included in each factor by the values of $\Delta D_{lj;i(t)}$ and $N_{l;i}(t)$ respectively.
Maximizing (2.2.9) then gives us partial maximum likelihood estimators of the parameters in $\boldsymbol{\beta}$, with similar properties as those derived in the ordinary survival

case. As such, similar hypothesis tests as those in 2.1.3 apply, although specific formulae for, say, covariance estimators become increasingly complex. Similar model evaluation tools can also be used, albeit in also slightly modified form.

2.2.3 Panel-Type Data

Often, especially in medical settings, we do not in fact observe individuals continuously, as was previously assumed. Instead, we often have *panel-type data*, where individuals are only observed at a finite number of times — say, the times of doctor visits where various measurements might be taken. In this case, we only know that a transition has happened in the *interval* between the last and current visit — the data is *interval-censored*. Further, since we are in a multi-state setting, we also cannot necessarily rule out that other transitions have happened in between, or indeed, that if a subject remains in the same state as last time, they might have transitioned to another state and later returned. As such, due to this lack of data, in addition to the assumptions made in 2.2, we assume that the process is *time-homogeneous*, and the transition hazards independent of time t. This is the assumption made in [9], which forms the basis for this section. We may still have, possibly time-varying, covariates though. For fixed covariates, notice that the cumulative transition hazards from (2.2.2) simplify to $Q_{lj}(t) = tq_{lj}$. Further, if $d\mathbf{Q}(t) = \mathbf{Q}(t) - \mathbf{Q}(t-)$ is the instantaneous change in the matrix of cumulative transition hazards, then in the time-homogeneous case, $dQ_{lj}(t) = q_{lj}$, for all $t \geq 0$, and all $l, j \in \{0, \ldots, J-1\}$ (using the definition (2.2.3) for when $l = j$).
It is now a theoretically simple (if practically difficult) matter to calculate the transition probability matrix given by (2.2.4). Since $d\mathbf{Q}(t) = d\mathbf{Q}$ is in fact constant over any interval of time $(u, t+u]$, we may write $\mathcal{P}(u, t+u) = \mathcal{P}(t)$. The Kolmogorov forward equations for the transition probability matrix, for any positive value of the time t, can then be solved by a matrix exponential:

$$\mathcal{P}(t) = \mathrm{Exp}(td\mathbf{Q}) = \sum_{k=0}^{\infty} (td\mathbf{Q})^k / k! = \mathrm{I} + (td\mathbf{Q}) + (td\mathbf{Q})^2/2! + (td\mathbf{Q})^3/3! + \cdots \tag{2.2.10}$$

where the term $(td\mathbf{Q})^0$ is defined as the identity matrix I with the same dimensions as $d\mathbf{Q}$, and further terms are defined by multiple *matrix* products.
Say that we have n individuals, and M_i observation times for individual i. We assume that the observation times in themselves do not give information about the value of the observation. This is discussed in more length in [9], where examples of non-informative observation times are given as times fixed in advance, chosen independently of the states, or where the *next* time is based on the *current* state.

Then a full likelihood is

$$L(\mathbf{Q}) = \prod_{i=1}^{n} \prod_{m=1}^{M_i - 1} \mathcal{P}_{\Lambda(t_{i,m})\Lambda(t_{i,m+1})}(t_{i,m+1} - t_{i,m}), \tag{2.2.11}$$

where each term is the entry of the transition probability matrix, at the $\Lambda(t_{i,m})$th row and $\Lambda(t_{i,m+1})$th column, evaluated for time $t = t_{i,m+1} - t_{i,m}$, and $t_{i,m}$ is the mth observation time for the ith individual. The corresponding log-likelihood can then be maximized, to arrive at estimates of first the logarithm of, and consequently the transition hazards q_{lj} themselves.
In a proportional hazards model with covariates, we have

$$q_{lj;i}[\mathbf{x}_i(t_m), \boldsymbol{\beta}_{lj}] = \exp[\mathbf{x}_i(t_m)\boldsymbol{\beta}_{lj}]q_{lj;0} \tag{2.2.12}$$

for all transitions $l, j \in \{0, \ldots, J-1\}$, individuals $i = 1, \ldots, n$ and observation times $m = 1, \ldots, M_i$. Because of the interval-censored data, we have to still use the full likelihood, and maximize over *both* the baseline transition hazards $q_{lj;0}$ and coefficients $\boldsymbol{\beta}_{lj}$. If covariates are time-varying, it is important to use, for each term in the likelihood, their values at the *first* observation times in each interval.

Model Evaluation

As pointed out in the R package manual [9], especially "the Markov property and homogeneity of transition rates, both between individuals and through time, can be restrictive assumptions" (p. 19). As such, they suggest some approaches to model evaluation in the multi-state, interval-censored case. There are two graphical approaches mentioned in [9]. One approach is to compare the predictions of entry times into a particular state, with non-parametric estimates. However, this only works for entry into absorbing states, from which it is impossible to exit. Another approach is to compare observed and expected prevalence of states, at a series of times. However, if not all individuals are observed at these same times, it relies on approximations and interpolations and may then be unreliable.
A formal goodness–of–fit test is also mentioned, comparing observed and expected *transitions* between the different pairs of states. This is done for a series of transition starting times, transition time intervals, and covariate categories, and summarised in a Pearson-type contingency table test statistic. Under the null hypothesis that the model does fit the data well, this test statistic follows a complex distribution, which in simpler cases can be approximated as χ^2. Generally, a parametric bootstrap procedure is used. In any case, if the p-value is under a predetermined threshold, then we can reject the hypothesis that the model fits the data well.

Score residuals are also available in the msm package for R, and mentioned in [9], for assessing an individual's influence on the maximized likelihood.

Chapter 3

Analysis of data from the CPUP

3.1 Background

In this chapter, we will use the theory from chapter 2 to analyse data from the Swedish *cerebral palsy follow-up program* (CPUP). In this introductory presentation of the data, we rely on the doctoral thesis by Cloodt [1], to which we also refer for more details.

Cerebral palsy (CP) is an umbrella term for a group of permanent but non-degenerative neurological disorders, affecting motor function, movement and posture, caused by injury to the developing brain during pregnancy, at birth, or during the first two years of life. It is the most common motor disability in children. According to [1] (pp. 17–18), "[t]he most common definition of CP today is from Rosenbaum and colleagues from 2006; 'Cerebral palsy (CP) describes a group of permanent disorders of the development of movement and posture, causing activity limitations, that are attributed to non-progressive disturbances that occurred in the developing fetal or infant brain. The motor disorders of cerebral palsy are often accompanied by disturbances of sensation, perception, cognition, communication, and behaviour, epilepsy, and by secondary musculoskeletal problems'".

Cerebral palsy can be divided into various subtypes, of which spastic is the most common, and also the only subtype which appears in the data set used in this analysis. Within the spastic subtype, it can be described as unilateral or bilateral; again, bilateral is the only subtype in the current data set.

There is also a measure "describing the child's self-initiated mobility [...], and the use of assistive devices" [1] (p. 20), called the *Gross Motor Function Classification System* (GMFCS). This measure is stable over time, and is divided into five levels. Here, "[l]evel I describes the highest level of function and level V the lowest" [1] (p. 20).

For people living with cerebral palsy, the skeletal muscles have a reduced size.

They tend to be shorter, and have a lesser thickness and cross-sectional area. Tendons, on the other hand, are longer. Greater reduction in muscle size correlates with higher GMFCS levels, and thus lower levels of motor function.
Specifically, a common problem are *contractures*, which arise from a permanent shortening of the muscle–tendon unit as the soft tissues lose elasticity, leading to reduced range of motion in a specific joint. Contractures limit the ability to move freely, which in turn lead to decreased activity and participation levels. Knee contractures will be the primary focus of this analysis, but technically any joint can be affected, and ankle contractures will be briefly considered. In knee contractures, the knee is prevented from extending fully, leading to a permanently flexed state. This can both directly and through posture asymmetries lead to pain, as well as affect the ability to stand, increase the risk of scoliosis, and lead to a change in gait pattern. The affected gait can in turn lead to decreases in step length and walking speed, as well as increased fatigue and energy costs of movement. A knee contracture of $-10°$ extension or worse is cited by [1] as possibly having a large impact on the development of so-called "crouch gait", the altered gait pattern most associated with knee contractures (pp. 24–25). As such, this will be used as a cut-off point in the present analysis for defining what counts as a knee contracture, as done in [1]. Several surgical and non-operative treatments for knee contractures exist, although the evidence for non-operative treatments is limited, and there is still a risk of recurrence with surgical treatments [1] (pp. 25–26).
The Swedish cerebral palsy follow-up program (CPUP) started in the southern parts of the country in 1994, and expanded, eventually becoming a certified national health care registry in 2005, also including adults starting in 2009. In the CPUP, regular clinical examinations are performed, with the frequency being guided by the GMFCS level and age, varying from at least once every two years, up to twice per year. During these examinations, various data are collected, including measurements of range of motion, which are what will be used in the present analysis.
In one of the articles in the doctoral thesis [1], the time from first to second contracture in the same leg was studied, based on GMFCS level and where the first contracture had happened. There, legs were analysed separately, not taking into account that most people would have two legs. In the present analysis, the primary aim was to study if having a contracture on the knee of one leg would affect the risk of developing a contracture on the knee of the other leg. This was done in an exploratory fashion, using several different models and methods.

3.2 The data set

Throughout this degree project, R version 4.2.1 [11] was used for the analysis of data. In this section, we introduce the data set used. Some adjustments to the original data were necessary for this analysis, some notes for which follow here.
The original data set comes from the CPUP, and consists of 38293 observations, from 3542 legs, for 1775 individuals; as can be seen, we do not always have data from both legs for all individuals. Aside from coded identifiers for specific legs and individuals, respectively, the observations also include data for if a specific observation corresponds to a left or right leg; the date of observation, the first and last dates the individual was examined; the number of days since the first observation; the number of total observations; the date of birth; age in years and days; the subtype of cerebral palsy (only spastic bilateral in the data set); the maximally recorded level of GMFCS; and knee and foot status, measured in degrees of extension in range of motion for knee joints and degrees of dorsiflexion for ankle joints (feet).
Counting the numbers of legs, individuals and observations separately for each level of GMFCS as in table 3.1, we see that the group at level II is very small, consisting only of data from 16 individuals. As such, in much of the analysis, this group was combined with the group at level I.
It was also observed that for 1347 and 1529 observations respectively, data were

GMFCS_max	Legs	Individuals	Observations
I	1882	941	17696
II	32	16	272
III	470	235	6004
IV	548	277	7170
V	610	306	7151

Table 3.1: The number of legs, individuals, and observations for each GMFCS level

missing for the status variables for knees and feet. It was decided that missing values would be replaced with the last observed value. For the first observation, the last observed value was defined to be 0. After this, contracture variables were defined to be 1 if the status variables were $\leq -10°$, and 0 otherwise.
Since cerebral palsy is, as stated in section 3.1, often present from or even before birth, date of birth was chosen as the beginning time point. However, as the first recorded observation occurred later (between 30 days and 5 years, with a median of slightly above 2 years), the data is left-truncated. Further, as observations were made only at a predetermined, finite number of times, as stated in section 3.1, and as such fit the description of panel-type data in subsection 2.2.3, the data is

interval-censored. Importantly, since the times were decided based on the GMFCS level as judged at the previous visit, they fit the description of non-informative observation times.
In order to analyse the effect of having a contracture on the knee of one leg on the knee of the other leg, the data was first split into two parts, one containing data for all left legs, one containing data for all right legs. For multi-state type analysis, these were then joined together, in such a way that each line contained observations for both legs, at this point deleting observations for which data was only available for one leg. This led to a loss of 6 individuals at GMFCS level IV, and 2 at level V. We then create a new variable, counting how many knees have a contracture at any point in time, by adding the left and right leg knee contracture variables — this will be the state variable of interest in the multi-state analyses. It is seen in the data that knee status can improve, such that a knee can recover from contracture by our definition of the cut-off point. Thus, backwards transitions should be allowed in the model. We also reason that it is unlikely that contracture develops simultaneously in both legs, and thus define only direct transitions between adjacent states as possible. This is the model seen in figure 3.1.
For a more standard kind of survival analysis, we transform the data prepared for

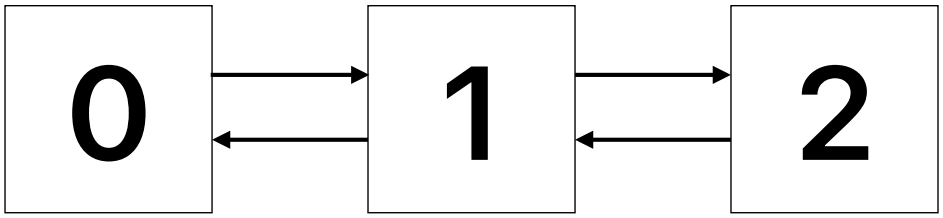

Figure 3.1: The multi-state model. The number of each state corresponds to the number of knees with a contracture, that is, extension of $-10°$ or lower. Transitions are allowed between adjacent states only, and recovery is modeled.

multi-state analysis. It is suggested in [8] Section 11 (pp. 211–225) that one can interpret a multi-state model as a joint model for a time-dependent covariate process and a time-to-event process. In this case, the event has to be an absorbing state in the original model. This is not quite the case in the data set, as recovery from both knees being affected by a contracture is seen. What is done, is to treat the event of interest as the *first* time both knees are affected, which will then be absorbing. New variables are created, one that will be used as a time-varying covariate, representing one knee having a contracture, another for when both knees are affected. Referring to the model in figure 3.1, the time-varying covariate corresponds to state 1, and

the event of interest to the first time state 2 is reached. Then, the data are filtered such that observations are only retained for an individual up until the first time both knees are affected.

3.3 Multi-state analysis

We begin analysing the data from a multi-state perspective. This uses the modified data set mentioned in section 3.2. First, we get an idea of the age ranges represented in the data, by extracting the first and last observations for each person in the data set, respectively. We make histograms for the number of individuals at each age, shown in years, and present these in figure 3.2. The medians and means for ages

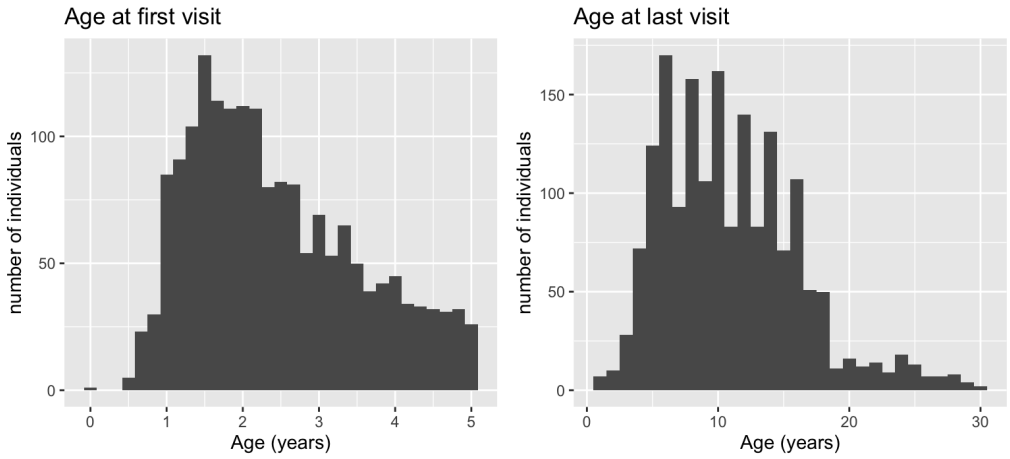

Figure 3.2: Histograms showing the number of individuals at each age in the data set, for the first and last visits, respectively.

at first and last visits were seen to be 2.17 and 2.43, and 10.18 and 11.01 years, respectively. As such, any models constructed from the data should be interpreted with caution, if at all, outside of these bounds, especially if extrapolating beyond the first quantile for first visits, 1.54 years, and the third quantile for last visits, 14.16 years. Similar data was also produced for follow-up times, but as these can be mostly inferred from the above, we omit them here.

3.3.1 Models using the msm package

Since the data set is interval-censored, we begin our analysis by treating it as such. We use the msm package for R, as detailed in [9], and the theory of which we summarised in section 2.2.3 on panel-type data. Since the package requires all individuals to have at least two observations, we filter out all individuals with only one observation. This led to a loss of 11 individuals at GMFCS level I, 3 at level III, 2 at level IV, and 6 at level V. We use the model in figure 3.1. In the data, since we do not have exact event transition times, we observe transitions between states 0 and 2. In fact, we observe transitions as in table 3.2. Since estimations are made

from—to	0	1	2
0	13326	521	446
1	314	356	276
2	214	179	1712

Table 3.2: Observed transitions in the multi-state data.

based on the transition probability matrix, and not the transition hazards directly, it still makes sense to in the model only allow direct transitions between adjacent states.
Estimation was carried out using three different optimisation methods, which all gave very similar results, so only those for the default method are reported. First, a null model without covariates was fit. Age in years was used as the time variable. The estimated transition hazards for allowed transitions, are then given, with approximate 95% confidence intervals, as in table 3.3. Using the qratio.msm

Transition	Estimated hazard
0–1	0.12 (0.11, 0.13)
1–0	0.99 (0.90, 1.10)
1–2	1.22 (1.10, 1.35)
2–1	0.38 (0.34, 0.43)

Table 3.3: msm null model. Estimated transition hazards are given together with approximate 95% confidence intervals within parentheses.

function, estimates of ratios between transition hazards can be calculated, together with 95% confidence intervals. With this, we see that the 1–2 transition hazard is about ten times greater than the 0–1 transition hazard, or $10.40\ (9.11, 11.87)$ times. At the same time, the hazard for returning to baseline is almost as large as the one for advancing, but slightly less, with a ratio of $0.82\ (0.71, 0.94)$. This would suggest that indeed, the hazard for getting a contracture on a second knee,

is increased compared to getting a first contracture, given that a contracture has already occurred — if the model is accurate.
A proportional hazards model as in (2.2.12) was also fit with the maximum GMFCS level as a covariate. As GMFCS level is best described as a factor variable, dummy variables were created for non-baseline levels, as is practice in standard linear regression. GMFCS level I was chosen as a natural baseline, as well as because it has the greatest number of individuals, as seen in table 3.1. Because of the low number of individuals in group II, the confidence intervals for its estimated hazard ratios were very large, when initially estimated. Thus, as discussed above, group II was combined with group I, into a group I-II, which was then chosen as baseline. Baseline estimated transition hazards, as calculated for level I-II, together with estimated hazard ratios for level III, IV and V were then given as in table 3.4. Again,

Transition	Level I-II hazard	Level III HR	Level IV HR	Level V HR
0–1	0.036 (0.030, 0.042)	4.45 (3.51, 5.64)	6.63 (5.32, 8.25)	10.74 (8.69, 13.28)
1–0	0.92 (0.75, 1.15)	1.26 (0.91, 1.75)	1.08 (0.80, 1.45)	1.15 (0.87, 1.53)
1–2	0.53 (0.39, 0.72)	4.10 (2.67, 6.30)	2.71 (1.89, 3.89)	2.52 (1.78, 3.58)
2–1	0.56 (0.40, 0.80)	1.20 (0.75, 1.92)	0.51 (0.33, 0.77)	0.67 (0.45, 0.99)

Table 3.4: msm model with covariates. Baseline estimated transition hazards, as well as estimated hazard ratios for other GMFCS levels, are given, together with approximate 95% confidence intervals. GMFCS level I-II, created by combining all observations at levels I and II, is used as baseline.

we see clearly that the hazard is greater for the second knee getting a contracture, if one is already present, in the level I-II group, although the magnitude is different to before. It is also worth noting that the hazard for recovery is greater than the hazard for progression in the baseline group. As one might expect, the hazard ratios for getting more contractures, from baseline or from one already present, are above one with 95% confidence for all higher levels of GMFCS. The hazard ratios for backwards transitions are not significantly different from 1, though (since 1 is included in the 95% confidence intervals, we can say that they are not statistically significantly different from 1, and have p-values greater than 0.05), except for transition 2–1 for levels IV and V, where it is significantly reduced, i.e. lowered hazard of getting better once a contracture has occurred, as compared with level I-II. A likelihood ratio test, with 12 degrees of freedom (because we have three dummy variables, representing the non-baseline GMFCS levels, for each of the four transition hazards), gives a likelihood ratio test statistic of 1008.019, which is statistically significant for any reasonable p-value, and certainly for 0.05, indicating that the model is improved with the covariate(s) added.
Estimated transition hazards, and not just hazard ratios, could also be calculated

for GMFCS levels III, IV and V. Other helpful summary tables could be provided to give information about the models, but as these, as well as the results presented so far, depend on the model assumptions being true to be accurate, we examine these beforehand.

As there are no absorbing states in the model, we cannot use the first approach mentioned in the subsection on model evaluation for panel-type data 2.2.3. The second approach is also dubious, since the exact ages of observation likely differ between the studied individuals. If nevertheless making plots representing observed and expected prevalence of states (to be found in appendix A, figures A.1–A.5), the prevalence of state 0 seems to be underestimated, and of state 2 overestimated consistently, except for the baseline group, GMFCS level I-II, in the model with covariates, where the relationship is reversed. We turn to the formal goodness–of–fit test. To make the contingency tables less sparse and improve the χ^2 approximation, only two groups each of transition starting times and transition time intervals are chosen, and two covariate groups, for the model with covariates. The test statistics are then given by the pearson.msm function as 955.7862 for the null model and 891.7974 for the model with covariates, both leading to p-values so small that R only shows them as 0, and thus are highly statistically significant. As such, neither model actually provides an adequate overall fit to the data, confirming what was seen in the plots.

This should perhaps not come as a surprise, as especially some of the assumptions in the model seem quite questionable. Not least the assumption that the process is time-homogeneous. In an attempt to remedy this, the msm package allows for fitting a model with piece-wise constant hazards. To this end, an attempt was made with piece-wise null and covariate models, letting the (baseline) hazard vary across four intervals of time: [0,5), [5,10), [10,15), and [15,∞). Goodness–of–fit tests yield test statistics of 281.966 and 237.4049 for the null and covariate models respectively, but this is still high enough to reject the null hypothesis of a good fit, and they are therefore not elaborated further on. Another possibility is that the Markov assumption is too restrictive. Maybe, for example, it would be more realistic if one incorporated time-varying covariates corresponding to how many times an individual has been in a certain state, or the time since entry into the current state, for example. It could also be that perhaps, constructing a model with all transitions allowed, i.e. allowing for directly going from no knees having a contracture, to both having a contracture at the same time, would in fact be more realistic, despite the assertion earlier that we in the model should only allow transitions between adjacent states. Further, the proportional hazards assumption maybe holds some of the blame for the bad fit of the covariate models; this would not explain, however, the bad fit of the null models. It could also be the case that we lack covariates that would be valuable to include. We could also consider letting the coefficients vary with time.

Score residuals were also calculated for the null model and the model with covariates, and plotted as shown in figures A.6 and A.7 respectively in appendix A. Examining data from the individuals with unusually high score values (above 0.5 in the null model, above 2 in the model with covariates) does not reveal anything particularly unusual, expect maybe that they all have a decently high number of total visits. If that is the reason that their score residuals are high, it would make sense, as more observations should lead to greater contribution to the likelihood. Possibly, one could consider re-fitting the model omitting these individuals from analysis, especially if a more granular analysis would reveal that the data is unusual in some other way, although even if unusual it would likely still be medically feasible, and no real reason to omit the data could then be given.
Since the fit is unquestionably inadequate, we do not provide more summary information about or derived from the model. While the above adjustments could be valuable avenues to explore in order to improve it, we turn here instead to exploring other ways entirely of modeling the data.

3.3.2 Models using the survival package

The survival package for R, commonly used in standard survival analysis, also provides functionality for handling multi-state models, as detailed in [10]. While the package does not support interval-censored, panel-type data, "[i]f subjects reliably come in at regular intervals then the difference between the two results can be small" [10], where the msm package estimates *occurrence* of progression, whereas the survival package estimates *observation* of progression. Bearing this in mind, we attempt to model the data as a multi-state model using the survival package. Because all transitions are then taken to happen exactly when they are observed, we have to allow direct transitions between states 0 and 2. As such, we have to modify the multi-state model as compared to the model in figure 3.1, to the model shown in figure 3.3.

Recalling the Aalen–Johansen estimator (2.2.7), if we are interested in the probability of being in a certain state at time t, we can multiply the estimated matrix of conditional transition probabilities for the interval $(0,t]$, $\widetilde{\mathcal{P}}(0,t)$, with the initial distribution of states at time 0, $p(0)$, to get a probability-in-state vector at time t as

$$\widetilde{p}(t) = p(0)\widetilde{\mathcal{P}}(0,t). \tag{3.3.1}$$

The survfit function in the survival package can calculate these Aalen–Johansen estimators of probability-in-state (3.3.1) simultaneously for all states, which is crucial to making the estimates accurate, as discussed further in [10]. The Aalen–Johansen estimators for the model in figure 3.3 are given in figure 3.4. Since the

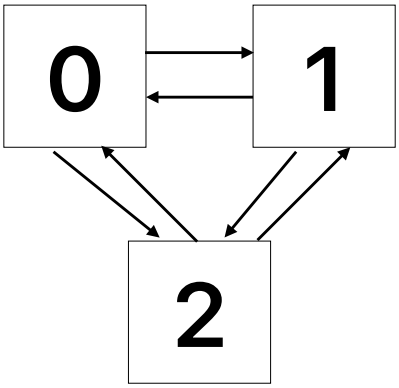

Figure 3.3: The modified multi-state model used in 3.3.2. The number of each state corresponds to the number of knees with a contracture, that is, extension of $-10°$ or lower. Transitions are allowed between all states, and recovery is modeled.

probabilities sum up to 1 at any given point of time, the curve representing the unaffected state 0 is omitted. If covariates are passed to the survfit function, similar estimators can be made for different subgroups. This was done with GMFCS level as covariate, leading to four dummy variables for the levels above I. These are presented in appendix B as figures B.1–B.5. As can be seen by the confidence bars in the plots, in the aggregate case the confidence intervals remain relatively small up until around 15 years, which fits well with the assessment in section 3.3, figure 3.2, that we have the most data up until slightly before 15 years of age. Anything before around 1.5–2 years of age should also be interpreted with caution, as mentioned in the same section. Here, the assumption has been made that everyone starts out in state 0 at birth, even though we do not have proper data until the age of first visit for any individual.

Looking at the estimators for the subgroups, we see a clear trend of increasing probability of having one or two knee contractures for each level, for any specific time. We also see in figure B.2 clearly that having the individual observations for the subgroup with level II on their own will not provide any reasonable data, thus further motivating the choice in the previous section and later in this section of making the combined group I-II. Also unsurprisingly, since we have the most individuals in the subgroup with level I, its confidence intervals are the smallest, as seen in figure B.1. Interestingly, for group I, we also see that the probabilities of being in the states for having a contracture on one or two knees respectively, are approximately equal. Otherwise, the probability of being in the state with a contracture on both knees is consistently higher than the probability of being in the state with a contracture on one knee. This might suggest either that contractures

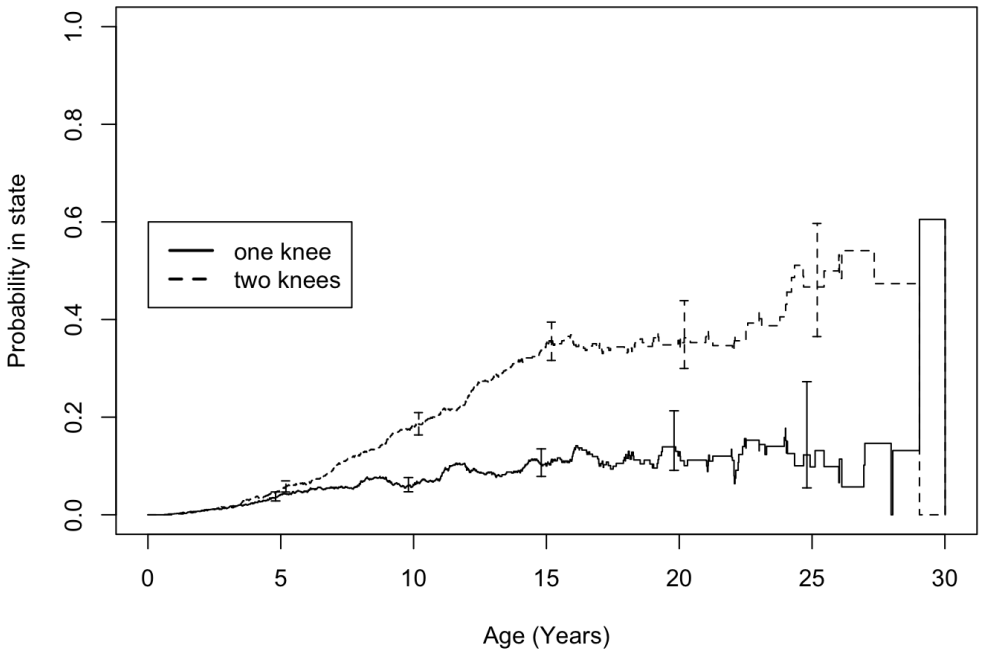

Figure 3.4: The Aalen–Johansen estimators of probability-in-state for the multi-state model in figure 3.3. This represents the estimated probability at any point of time of being in states 1 or 2, for one knee or two knees having a contracture. 95% confidence intervals are represented as bars for times 5, 10, 15, 20, 25 years.

are more likely to appear in pairs, or that once one knee has a contracture, the risk is then increased for contracture on the other knee.

We turn now to a proportional hazards model, as in (2.2.8). We use the GMFCS level as sole covariate, which leads to three dummy variables for levels III, IV and V, compared to the baseline combined group I-II. Since we model nine possible "transitions", if we include those for remaining in the same state, we get 27 total estimates of coefficients. Most hazard ratios, except some for remaining in state and for recovery, are found to be significantly different from 1 on significance level 0.05, with values similar to those found for the msm model and reported in table 3.4. An overall likelihood ratio test also finds a statistically significant difference from a null model. Hazard ratios are given for forwards transitions, together with 95% confidence intervals, in table 3.5. This suggests, like for the msm model, that forward transition hazards all increase with increased GMFCS levels, which is unsurprising.

Transition	Level III HR	Level IV HR	Level V HR
0–1	3.52 (2.60, 4.76)	5.77 (4.38, 7.59)	9.64 (7.48, 12.42)
0–2	10.08 (7.02, 14.47)	15.24 (10.73, 21.65)	23.31 (16.75, 32.45)
1–2	3.13 (1.92, 5.10)	3.01 (1.85, 4.91)	2.77 (1.72, 4.45)

Table 3.5: survival package multi-state model with covariates. Estimated hazard ratios for GMFCS levels III, IV and V are given, together with approximate 95% confidence intervals. GMFCS level I-II, created by combining all observations at levels I and II, is used as baseline.

Testing the proportional hazards assumption through the use of scaled Schoenfeld residuals (2.1.28), with formal tests of the null hypothesis that the γ_k as in (2.1.31) are equal to zero, performed separately for all transitions, is easily done with the cox.zph function. Both Kaplan-Meier and rank transformations of time suggest rejecting the null hypothesis of proportionality for transitions 0–0 and 2–0, whereas using $g(t) = t$ suggests rejecting proportionality for the 0–0 and 0–1 transitions. The 0–0 transition is not interesting, and should not really be in the model in the first place; it appears here because the data set still contains observations where no change has been observed since the last visit. Transforming the data set in some way could probably remedy this. Alternatively, there are ways of forcing coefficients to remain at a value of 0, and thus hazard ratios to be 1. That the 2–0 transition might not have proportional hazards is perhaps also less of an issue, if we are more interested in modeling the effect on forwards transitions. The 0–1 transition is of more concern for the same reason, and one could consider letting the coefficients corresponding to this transition vary with time. Further, if we truly only want to model the effect of the covariates on the forwards transition hazards, we could exclude these from the model. This was attempted, and yielded the same results as in table 3.5. However, now all transformations of time indicate non-proportionality for the 0–1 transition, so exploring time-varying coefficients there would be a natural next step. Here, we instead turn to other models for the data, bringing us back to studying if and how having a contracture on one knee might affect the probability of getting a contracture on the other.

3.4 Time-varying covariate models

Here, we consider two approaches to analysing the data using time-varying covariates. We will use data on that one knee has a contracture as a time-varying covariate, to model either the time until both knees first have a contracture, or the first time the other knee has a contracture.

3.4.1 Time until the first time both knees have a contracture

We begin with the approach mentioned in section 3.2, where we model time until the first time both knees have a contracture, including a separate, time-varying covariate for when one knee has a contracture. This will then essentially be an ordinary survival model. Phrased in another way, the single event of interest is the first time when both knees have a contracture, i.e. the first time state 2 is reached in the model shown in figure 3.1, or perhaps more accurately in figure 3.3, since we here have to allow direct 0–2 transitions. The time-varying covariate which we will include is then equal to 1 whenever we are in state 1, and 0 otherwise. Since we cannot predict probabilities in a meaningful way, lacking an analysis of the 0–1 transition, we go directly to a proportional hazards model, as in the model (2.1.21). Here, the baseline hazard function $h_0(t)$ corresponds to the baseline transition hazard $q_{02;0}(t)$ in the multi-state model. We make the assumption that the 0–2 and 1–2 transition hazards are proportional. Further, since the coxph function cannot handle interval-censored data, we have to ignore this fact for this analysis.
Then, a cox proportional hazards model with sole covariate being the one-knee contracture status, we get that the hazard ratio is, with 95% confidence interval, 8.054 (6.576, 9.865). Likelihood ratio, score and Wald tests are all significant, with p-values smaller than is shown by R, and as such certainly less than 0.05. Interestingly, this value is indeed close to the ratio between the 1–2 and 0–2 hazards given in table 3.3, for the msm null model where hazards were assumed to be constant over time. Adding level of GMFCS to the model, after combining groups I and II as before, gives hazard ratios as in table 3.6. Of particular note is that

Covariate	Hazard ratio (95% confidence interval)
one-knee status	3.502 (2.832, 4.33)
GMFCS III	7.608 (5.520, 10.49)
GMFCS IV	10.449 (7.719, 14.14)
GMFCS V	14.195 (10.529, 19.14)

Table 3.6: Hazard ratios for the model of time until the first time both knees have a contracture. The one-knee status covariate is defined to be 1 whenever the multi-state process in figure 3.3 is in state 1, 0 otherwise. The other covariates represent GMFCS level, and are dummy variables created for the overall factor variable, and are compared to the baseline group I-II.

the hazard ratio for the one-knee status was more than halved, when GMFCS level was added to the model. This suggests that at least some of the associated

increased risk of getting a contracture on both knees, given that one knee already has one, is explained by the GMFCS level being generally higher for those who get a contracture on one knee in the first place. However, all hypothesis tests are still highly statistically significant, so it seems that we can quite reasonably reject the null hypothesis of no effect. As the likelihood ratio test for comparing the model without and with GMFCS level is also highly statistically significant on any reasonable level, we can also conclude that it does indeed provide valuable information, as indicated already be the change in the hazard ratio for the one-knee status.
Testing the proportional hazards assumption gives non-significant p-values, for both the one-knee status variable, as well as GMFCS level, except for the rank transformation of time for the one-knee status variable, in the model with both covariates. One could possibly then consider letting the one-knee status coefficient vary with time; although the test is non-significant when considering the model with one-knee status alone.
As no covariates are continuous, we omit the martingale residuals. The dfbeta and dfbetas residuals are plotted against individual identifiers, separately for each covariate. This is shown in figures C.1 and C.2 in appendix C. None are particularly large, indicating that no individual has had a particularly large influence on the estimates.
As such, this model might seem quite decent. However, it still depends on the, wrong, assumption that the data contains exact transition times, and is not interval-censored, as it in fact is. It also ignores that recovery is highly possible in the model. Further, it should be no surprise that having a contracture on one knee, should increase the hazard for getting a contracture on both knees, as compared to going directly from none to two knees with a contracture at once — after all, this skips over an intermediary step, and while not seen in the model, everyone has to essentially pass through this. Thus, the only thing we can really say from the model, is that if we have observed a contracture on one knee, the hazard for that person returning next time with a contracture on both knees is increased, compared to if they had no contractures at all; which does not quite sound surprising at all. As such, we consider one final way of modeling the data.

3.4.2 Time to contracture on one knee, given status on the other knee

Finally, we attempt to model the time to contracture knee by knee, incorporating a time-varying covariate to keep track of the status of the other knee. Restructuring the data for this analysis is somewhat complex, but briefly, the data was separated

into left and right legs, then combined twice, creating a status variable for the other knee, defined as the status of the left knee when the right leg was taken as the primary leg for that data set, and vice versa. Importantly, the status of the other knee was taken as its status at the *last* visit, for the data line corresponding to the current visit; otherwise, we would essentially predict using future data. The data was then added together again, creating a data set consisting of observations for all legs, with the status of the other knee as a variable available. Then, we chose to model only the time until the first contracture on each knee, so as to simplify the model into a more ordinary survival setting, and thus filtered out data pertaining to visits after the first time a particular leg was observed to have a contracture.

Standard Kaplan–Meier type estimates of the baseline survival function can be calculated using this data, as well as separate estimates for the different GMFCS levels. Plots are shown in appendix C, figures C.3 and C.4. However, we cannot directly compare similar plots split by the status of the other knee, precisely because it is time-varying. Attempting to do this straightforwardly would only compare individuals who start out with a contracture on the other knee with those who do not, at whatever starting time is chosen. One could use the landmark method mentioned in section 2.2.1, but we choose here to instead turn directly to the proportional hazards model, as in the previous section 3.4.1.

For the proportional hazards model as in (2.1.21), we have four covariates which we can examine the potential effects of. These are the GMFCS level, status of the knee on the other leg, as previously used, as well as the status of the foot on the same and other leg as compared to the leg of the current knee. We follow the suggestions from the section on purposeful selection of covariates as described in [2] pp. 133–141. Both this and automatic stepwise selection using the Akaike Information Criterion (AIC) yield the same final main effects model, though stepwise selection using the Bayesian Information Criterion (BIC) yield a slightly reduced model.

The model preferred by AIC and the purposeful selection procedure includes the covariates for GMFCS level, other knee status, and foot status on the same leg, whereas BIC prefers to leave out the foot status. As a likelihood ratio test between the models gives a test statistic of 7.6422 with 1 degree of freedom, and as such a p-value of 0.005702, and the difference is thus statistically significant on significance level 0.01, and thus also on level 0.05, we choose to here present only the model with all three covariates. Note, however, that the variable for the status of the foot on the other leg was left out. The hazard ratios for this main effects model is given in table 3.7. We have, as before, combined groups with GMFCS levels I-II, to function as baseline for the GMFCS level dummy variables. When kept separate, the coefficient for GMFCS level II was found to have very large confidence intervals, and to be highly non-significant (presumably mostly because of lack of data in group II) so we present only the results using the combined

group. All coefficients were found to be significantly different from zero (and thus

Covariate	Hazard ratio (95% confidence interval)
GMFCS III	5.557 (4.596, 6.719)
GMFCS IV	8.284 (6.946, 9.881)
GMFCS V	12.378 (10.419, 14.705)
other knee status	2.627 (2.185, 3.158)
foot status	1.429 (1.123, 1.819)

Table 3.7: The main effects model for contracture status on one knee, given status of the other knee. This includes covariates for GMFCS level, other knee status, and foot status on the same leg.

yield hazard ratios different from 1) using all different hypothesis tests, as well as likelihood ratio tests between the final model and those containing a subset of the covariates, on significance level 0.05 or below. All show an increasing effect on the hazard.

Interaction models were also considered. Since we do observe an effect of GMFCS level on one knee, it is reasonable to assume that it should also have an effect on the contracture status of the other knee, and it would also be reasonable to assume an effect on the foot status. Thus, the interactions between GMFCS level and other knee status and foot status, respectively, were added to the main effect model. A likelihood ratio test between this model and the main effects model gave a test statistic of 25.379, with 6 degrees of freedom, yielding a p-value of 0.0002905, and thus suggests a significant difference between the two models, on significance level 0.05 (or indeed less, i.e. 0.001). As such, we present the hazard ratios for this model in table 3.8. Of particular note is that, while coefficients corresponding to different GMFCS levels are almost unchanged, the baseline coefficients for the status of the other knee, as well as for the foot status, have become significantly larger, with no overlap in confidence intervals as compared to the previous estimates in the main effects model in table 3.7. This suggests a greater effect of having a contracture on the other knee, or same foot, for the lower levels of GMFCS I-II. Since this group experiences fewer contractures overall, it seems reasonable for this to be the case — if contractures on the other knee and same foot have an effect at all, they should have more effect in a scenario where contractures are more rare. Note that a hazard ratio below 1, leading to a decreased hazard, is shown for the interactions, with groups IV and V (with a non-significant hazard ratio for the interactions with group III), further supporting the observation — the higher GMFCS level explains already some of the effect of a contracture on the other knee (or same foot), since it increased the hazard for it occurring in the first place, so the effect of having a contracture is attenuated by also being at a higher GMFCS level.

Covariate	Hazard ratio (95% confidence interval)
GMFCS III	5.8320 (4.7548, 7.1533)
GMFCS IV	9.0843 (7.5189, 10.9755)
GMFCS V	14.2228 (11.8343, 17.0934)
other knee status	6.7773 (3.9759, 11.5526)
foot status	3.3344 (1.9360, 5.7429)
GMFCS III*other knee status	0.5416 (0.2683, 1.0933)
GMFCS IV*other knee status	0.4071 (0.2198, 0.7540)
GMFCS V*other knee status	0.2933 (0.1615, 0.5327)
GMFCS III*foot status	0.5159 (0.2356, 1.1297)
GMFCS IV*foot status	0.3296 (0.1587, 0.6847)
GMFCS V*foot status	0.3556 (0.1829, 0.6914)

Table 3.8: The interactions model for contracture status on one knee, given status of the other knee. This includes covariates for GMFCS level, other knee status, and foot status on the same leg, as well as the interactions between GMFCS level and the other two covariates separately.

We continue by checking the proportional hazards assumption. The coefficient for foot status and its interaction with GMFCS level shows non-significant p-values for all transformations of time. However, the other knee status, GMFCS level, and their interaction show significance for non-proportionality for Kaplan–Meier and identity transformations of time, and other knee status and the interaction (but not GMFCS level on its own) for the rank transformation. To improve the model, one could thus consider letting the coefficients vary with time.
We also consider briefly the martingale and dfbeta residuals for both models, main effects and with interaction. Plots are shown in appendix C. The martingale residuals, plotted versus GMFCS level as in figure C.5, are relatively evenly distributed with means around zero, except for a slightly lower value discovered for the higher levels, suggesting that there might be a slight discrepancy here in the model. Plotted against the status of the other knee as in figure C.6 and against the foot status on the same leg as in figure C.7, the residuals seem to have a mean of roughly 0 for when the covariates are zero, with maybe slightly below-zero means for when the covariates are equal to 1. None of the dfbeta residuals, as plotted for the main effects model in figure C.8 and the model with interactions in figure C.9, have particularly worrying values. Very similar results are shown for standardised dfbeta residuals, and we omit the plots.
As such, the model shows significant room for improvement, probably mostly due to the fact that the hazards seem to be non-proportional. As such, the given hazard ratios should be regarded with caution, as they at best give time-averaged hazard

ratios. Also, it should be remembered that the model has been highly simplified and adjusted, not least because we here have not modeled recovery, and only modeled time until the first knee contracture, as well as ignored the interval-censoring of the data. Further, it would probably be a good idea to incorporate some sort of measure to compensate for the fact that the same person has two legs. This can for example be done using so-called frailty models, as discussed in e.g. [3] section 9.1 (pp. 113–120).

Chapter 4

Discussion

Taking the models at face value seems to suggest a statistically significant increase in the hazard of developing a contracture on the second knee, if a contracture has already occurred on one knee. That all models seem to suggest this, either through the values of the transition hazards in section 3.3.1, the Aalen–Johansen curves in 3.3.2, or the greater-than-one hazard ratios in sections 3.4.1 and 3.4.2, might seem to strengthen this conclusion.

However, as seen in the respective chapters, all models face significant problems. The model in section 3.3.1 is the only one which can account for the interval-censoring which is present in the data set, but according to formal goodness-of-fit tests, it does not provide a good fit to the data. This is likely because several other heavy assumptions have to be made in order to treat the interval-censored data as such, including assuming that transition hazards are constant over time. The Markov assumption that the future of the model is independent of its past, given the current state only, is also highly restrictive, something that is shared with the model in section 3.3.2. Choices made of which transitions to model could also be questioned, and perhaps allowing for more would have yielded better results.

All other models are immediately faced with the problem that they treat the data as not being interval-censored, when it clearly is. As such, at best, the time until a contracture is *observed* is modeled, not the time until it actually happens. This also forces us in section 3.3.2 to model all transitions as possible, which might be questioned, but for the opposite reason as for the model in section 3.3.1, namely that too many might have been included. Further, the models in sections 3.4.1 and 3.4.2 only model the time until the first time contractures are observed, which ignores the recovery which is clearly observed in the other two models.

When including covariates, all models suffer in lesser or greater extent in that the assumptions made of proportional hazards might be incorrect, for one or more covariates. The model in section 3.4.1 might turn out to be completely trivial, because of the way it attempts to model a process happening on two knees together

as one, treating a clear intermediate step as a covariate. This only makes sense in the first place because of the interval-censored data causing us to observe direct transitions between having a contracture on no knees to on both knees, whereas had we had exact times for the data, this presumably would not happen. The model in section 3.4.2 suffers almost the reverse problem of modeling different legs on the same person separately, without attempting to compensate for this.

As such, it is clear that there are several possible avenues of improving the models. Time-varying covariates could be introduced into the multi-state models, to provide some data on the past of the model, and thus loosen the Markov assumption. Time-varying coefficients could be allowed, to loosen the proportional hazards assumption. Frailty effects, or similar, could be introduced to more closely connect the data from legs from the same person in the last model where legs are treated separately. The transitions allowed could be examined in the multi-state, especially msm, models, and for which transitions we model covariates could also be questioned.

Using other ways of transforming the data set could also be an option, which would allow for use with other packages in R, which could potentially yield better results, or allow for other model evaluation tools. Given the relative lack of observations at higher ages, this data could also be considered for exclusion, until more has accrued as the CPUP continues. Along this line, other kinds of observations could perhaps be considered for collection, and further inclusion in the data set, leading to the possibility of including other covariates into the models.

Other kinds of survival models could also be considered, as well as other kinds of models entirely. Perhaps even new models, or ways of handling existing ones, could be developed mathematically, and be implemented into software such as R. Also, there might be models and approaches already in existence mathematically, which have not yet been developed the software for using.

In closing, drawing concrete conclusions from the models presented is discouraged, due to the various problems present. However, they can hopefully serve as inspiration for further research and modeling efforts, be that through presenting examples of models to develop further and improve upon, or as approaches to avoid. Cautiously, even if the models here are wrong, that all of them suggested an increased risk of a knee contracture on the second leg if a knee contracture is already present on one leg, could be taken as a prompt to look into this potential relationship closer and with better models. If this relationship holds up in further studies, it would emphasize the importance of trying to prevent contractures from occurring in the first place. Developing and practicing such preventive measures, that are safe and effective, could perhaps then even be taken as a suggestion from the current thesis, despite the flaws in the exact models.

Bibliography

[1] E. Cloodt, "Knee contracture in children with cerebral palsy," Doctoral Thesis (compilation), Dept. of Clin. Sci., Lund Univ., Lund, 2022.

[2] D. W. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*, 2nd ed. John Wiley & Sons, Inc., 2008.

[3] D. F. Moore, *Applied Survival Analysis Using R*. Applied Survival Analysis Using R, 2016.

[4] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *J. Amer. Statist. Assoc.*, vol. 53, no. 282, pp. 457–481, Jun. 1958.

[5] D. Cox, "Regression models and life-tables," *J. Roy. Statist. Soc. Ser. B*, vol. 32, no. 2, pp. 187–220, 1972.

[6] T. Therneau, C. Crowson, and E. Atkinson, *Using time dependent covariates and time dependent coefficients in the cox model*, Online, R package version 3.5-7, 2023. [Online]. Available: `https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf`.

[7] P. M. Grambsch and T. M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, vol. 81, no. 3, pp. 515–526, 1994, ISSN: 00063444. [Online]. Available: `http://www.jstor.org/stable/2337123` (visited on 01/01/2024).

[8] J. Beyersmann, A. Allignol, and M. Schumacher, *Competing Risks and Multistate Models with R*. Springer New York, 2011.

[9] C. Jackson, "Multi-state models for panel data: The msm package for R," *J. Stat. Soft.*, vol. 38, no. 8, pp. 1–29, Jan. 2011. DOI: `10.18637/jss.v038.i08`.

[10] T. Therneau, C. Crowson, and E. Atkinson, *Multi-state models and competing risks*, Online, R package version 3.5-7, 2023. [Online]. Available: `https://cran.r-project.org/web/packages/survival/vignettes/compete.pdf`.

[11] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: `https://www.R-project.org/`.

Appendix A

Graphs for model evaluation of the msm models

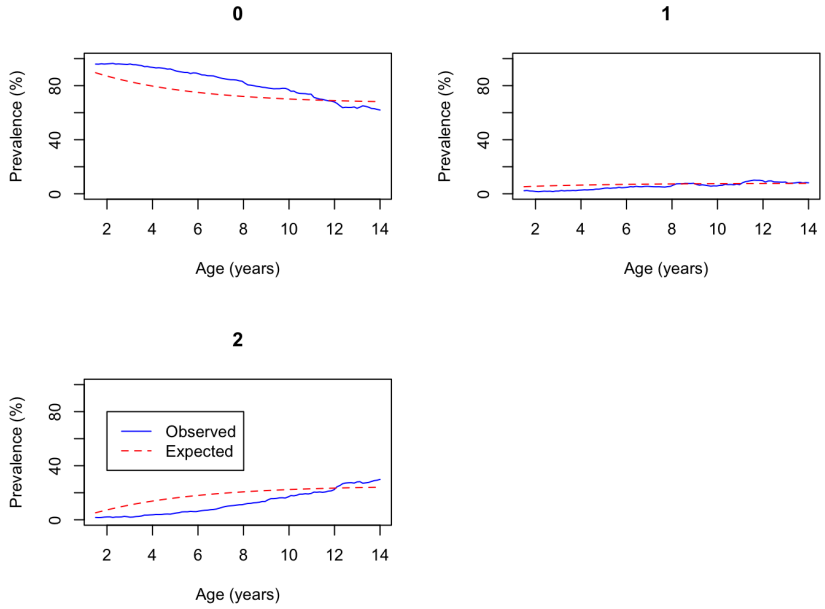

Figure A.1: Observed and expected prevalence of states in the msm null model.

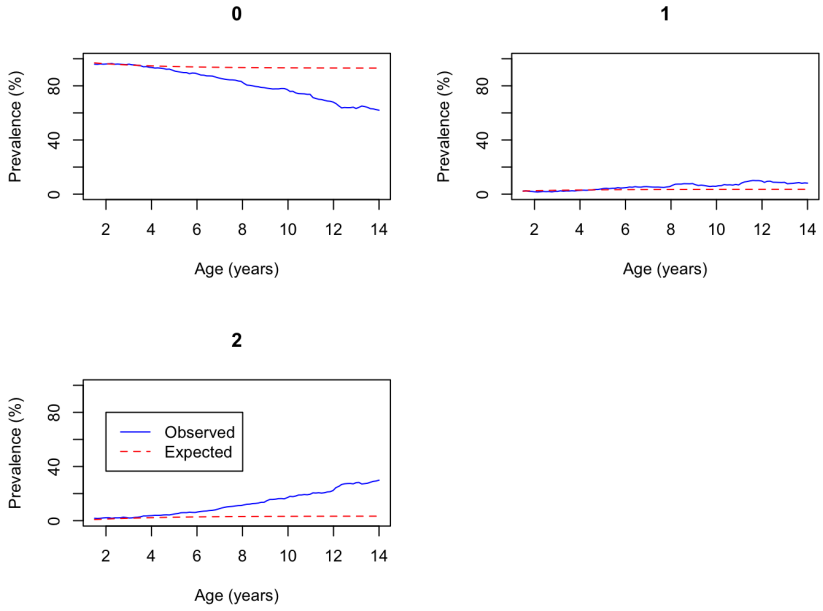

Figure A.2: Observed and expected prevalence of states for GMFCS levels I-II in the msm model with covariates.

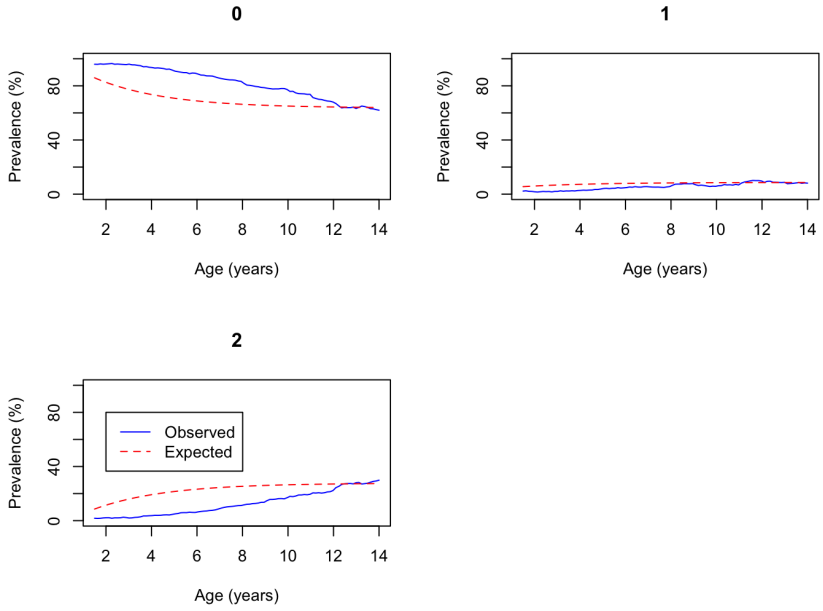

Figure A.3: Observed and expected prevalence of states for GMFCS level III in the msm model with covariates.

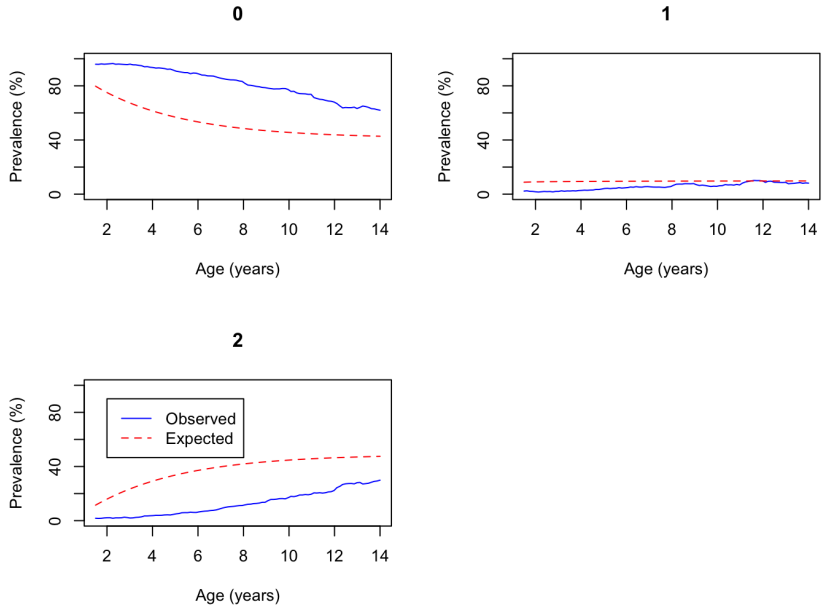

Figure A.4: Observed and expected prevalence of states for GMFCS level IV in the msm model with covariates.

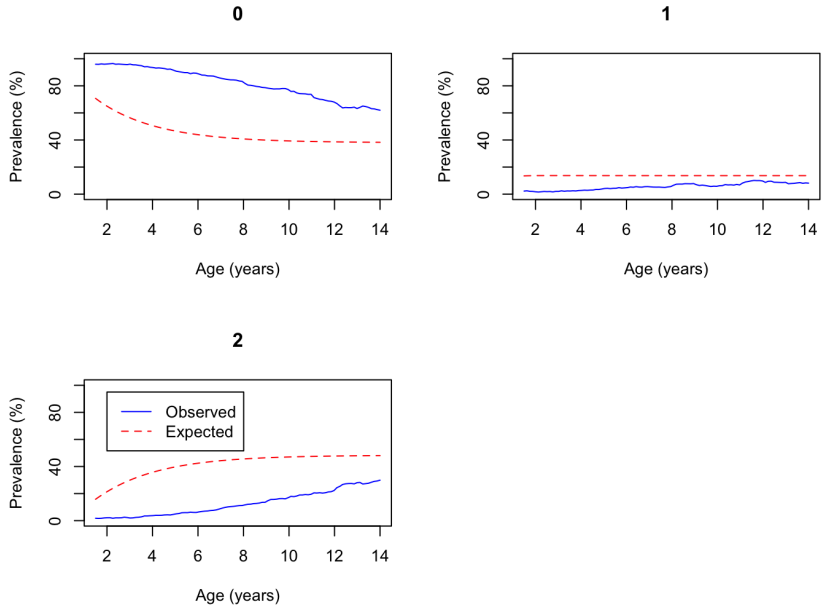

Figure A.5: Observed and expected prevalence of states for GMFCS level V in the msm model with covariates.

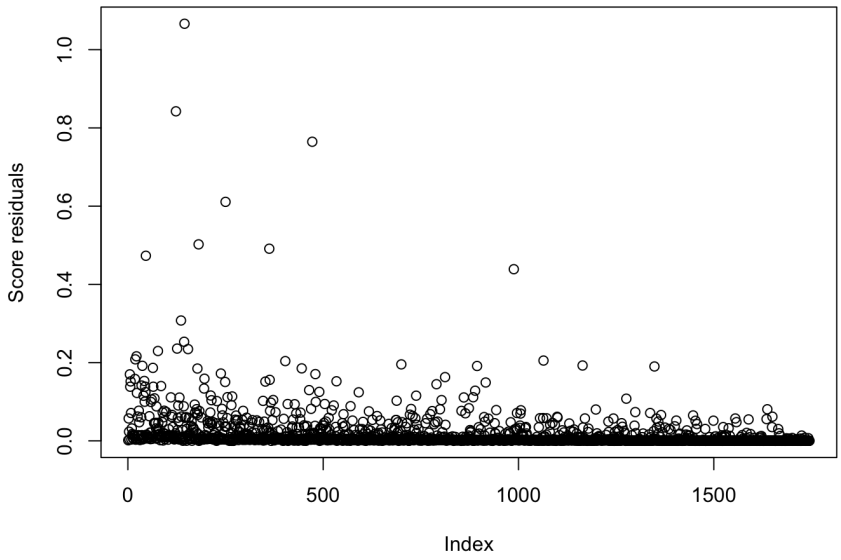

Figure A.6: Score residuals for the msm null model.

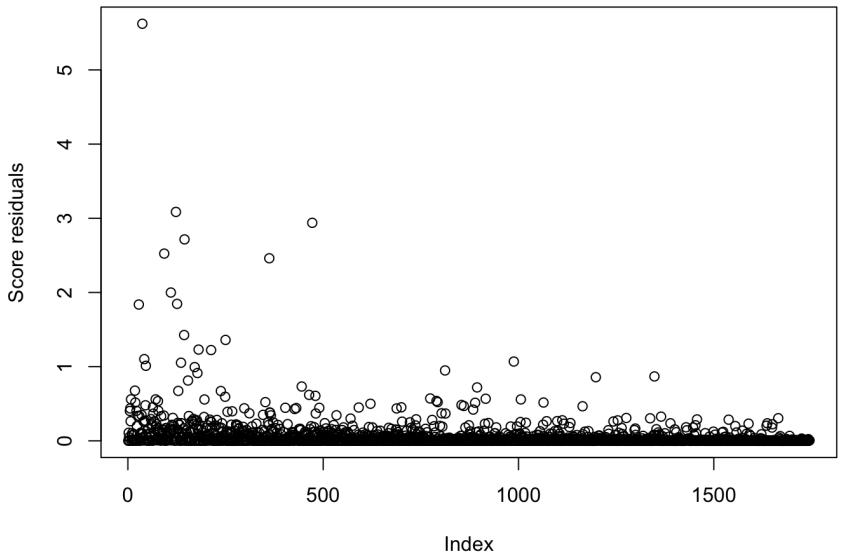

Figure A.7: Score residuals for the msm model with covariates.

Appendix B

Extra figures for the survival package multi-state models

The following figures B.1–B.5 plot the Aalen–Johansen estimators of probability-in-state (3.3.1) for the multi-state model in figure 3.3, forsubgroups with different levels of GMFCS. This represents the estimated probability at any point of time of being in states 1 or 2, for one knee or two knees having a contracture. 95% confidence intervals are represented as bars for times 5, 10, 15, 20, 25 years.

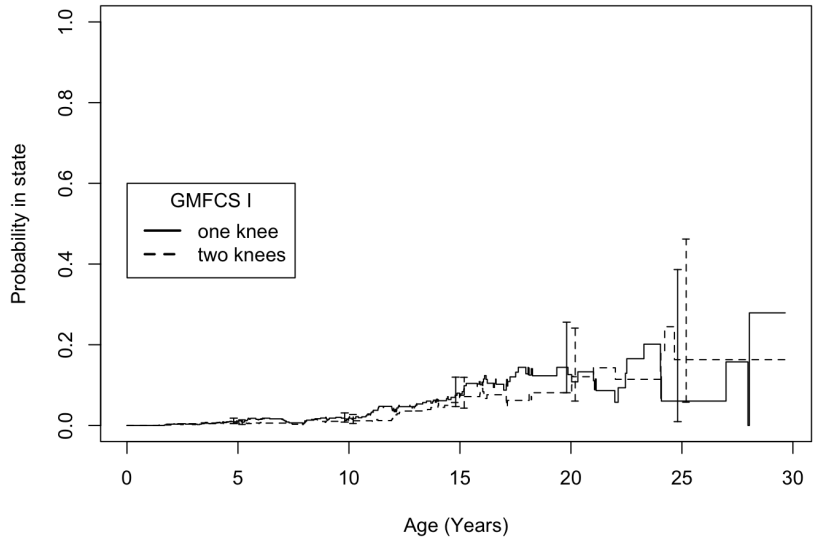

Figure B.1: The Aalen–Johansen estimators of probability-in-state for the subgroup with GMFCS level I.

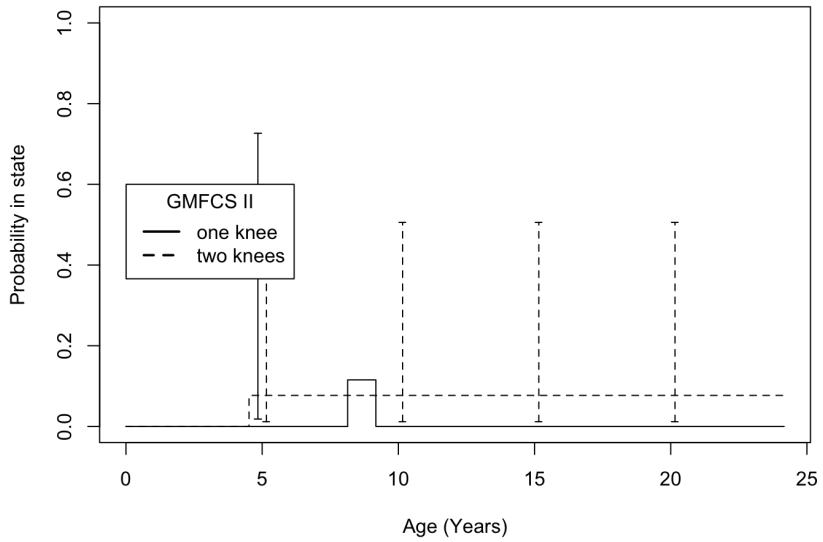

Figure B.2: The Aalen–Johansen estimators of probability-in-state for the subgroup with GMFCS level II.

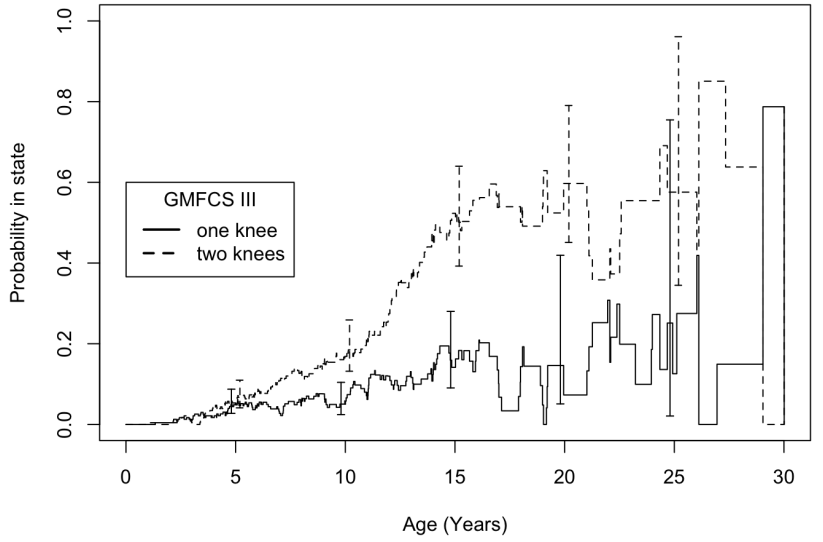

Figure B.3: The Aalen–Johansen estimators of probability-in-state for the subgroup with GMFCS level III.

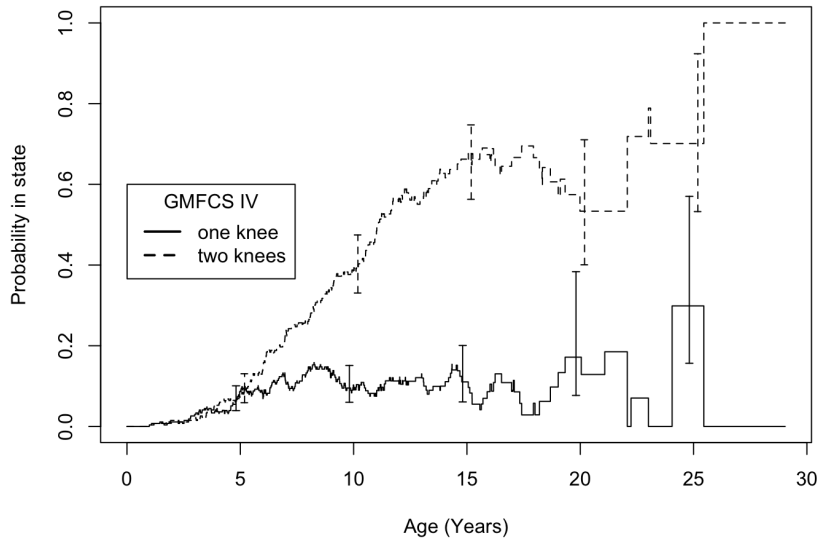

Figure B.4: The Aalen–Johansen estimators of probability-in-state for the subgroup with GMFCS level IV.

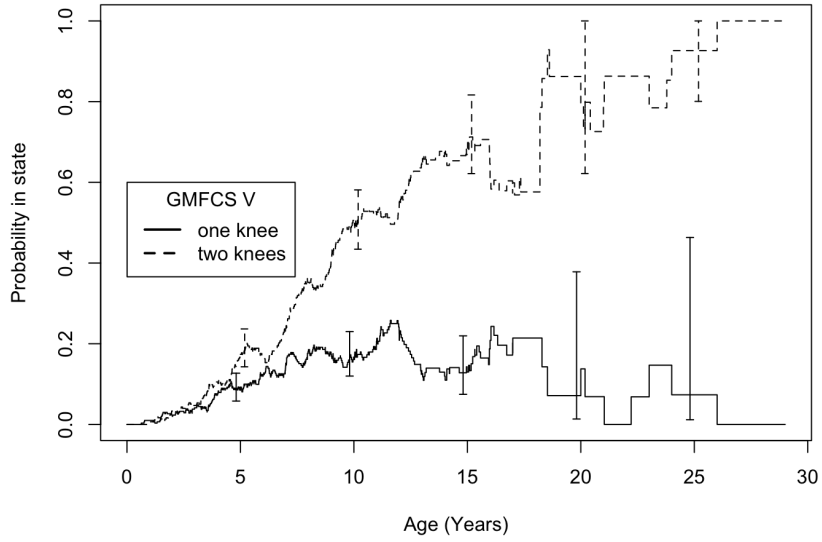

Figure B.5: The Aalen–Johansen estimators of probability-in-state for the subgroup with GMFCS level V.

Appendix C

Extra figures for the time-varying covariate models

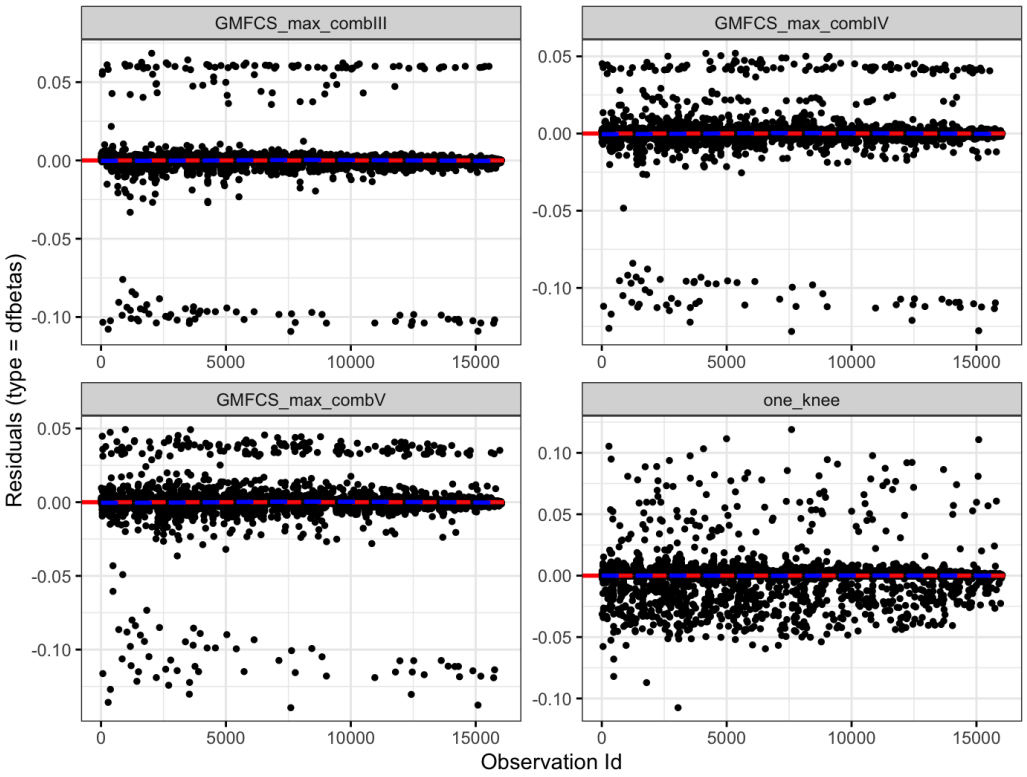

Figure C.1: dfbeta residuals, plotted separately for each covariate (different levels of GMFCS, and one-knee status), against individual ids. None are particularly large.

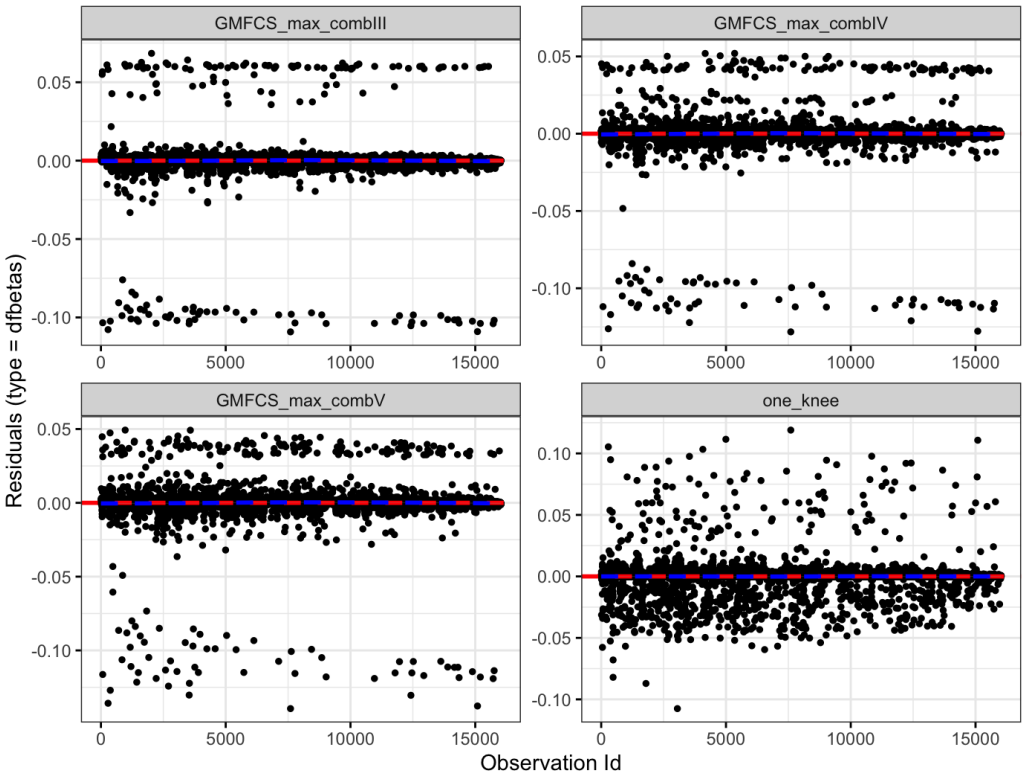

Figure C.2: dfbetas, standardised dfbeta, residuals, plotted separately for each covariate (different levels of GMFCS, and one-knee status), against individual ids. None are particularly large.

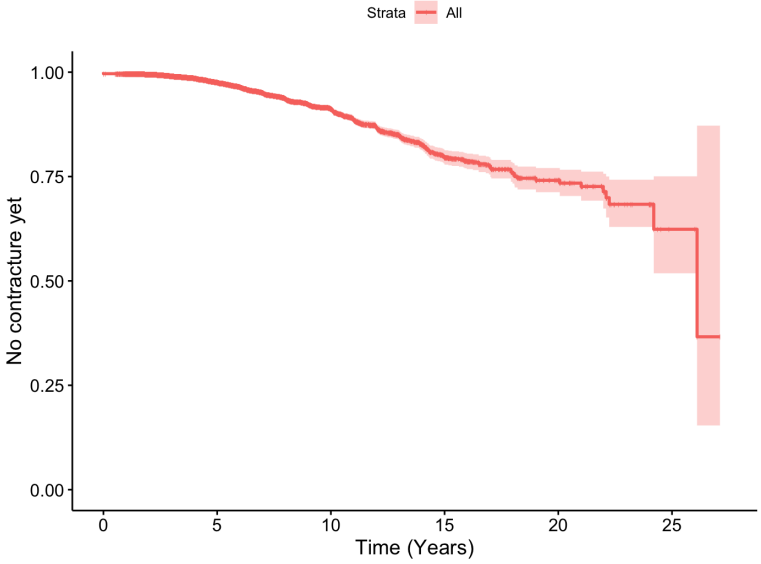

Figure C.3: Kaplan–Meier estimate of the time until first knee contracture, separately for all legs, with shaded 95% confidence interval. This does not take into account that each person has two legs.

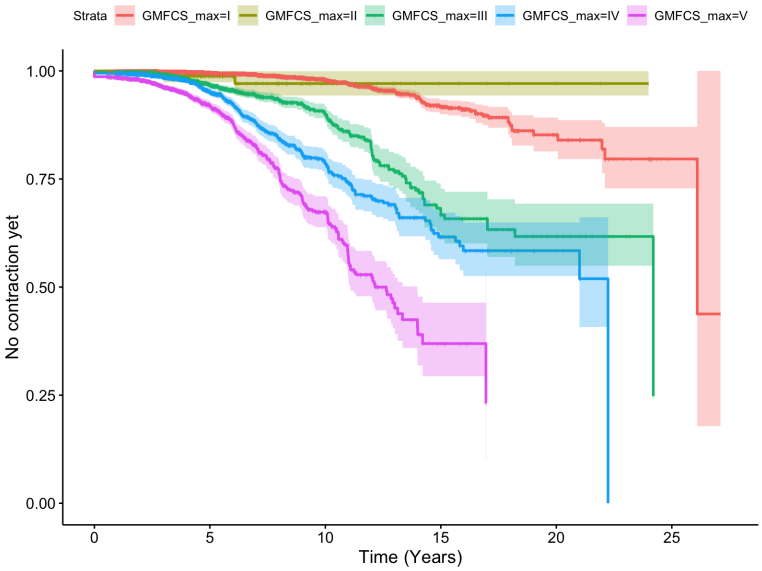

Figure C.4: Kaplan–Meier estimates of the time until first knee contracture, separately for all legs, with shaded 95% confidence intervals. This is done separately for the different levels of GMFCS. This does not take into account that each person has two legs.

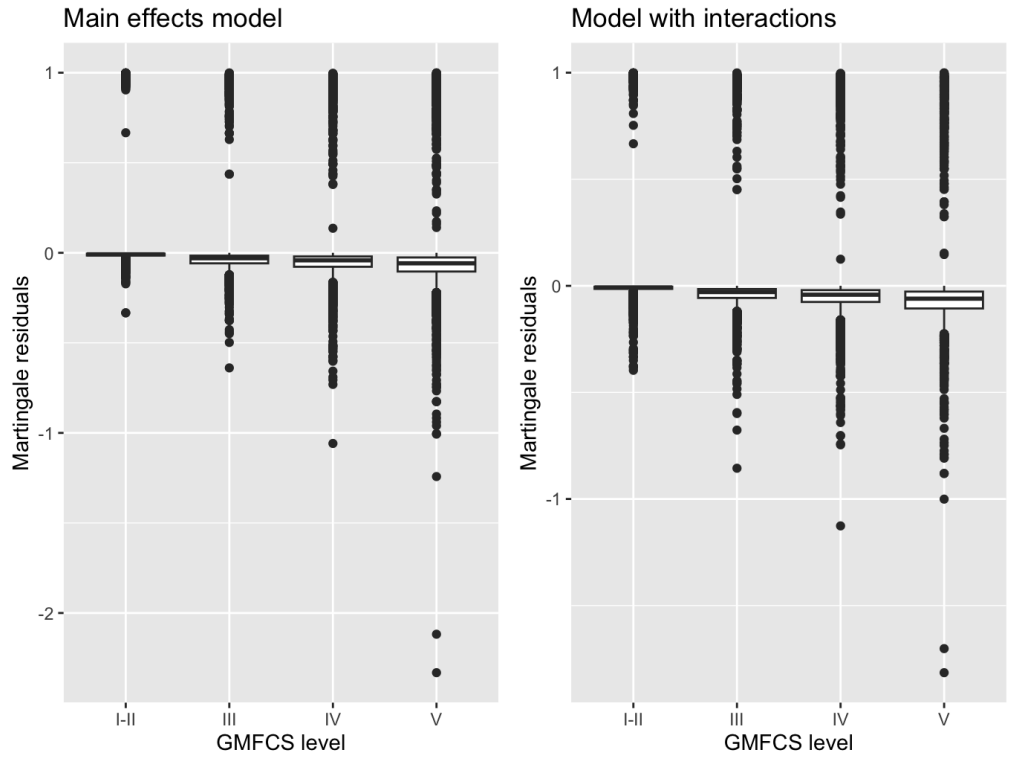

Figure C.5: Martingale residuals for the main effects and interactions model, plotted against GMFCS level, in the scenario when studying the time to knee contracture given the status of the other knee.

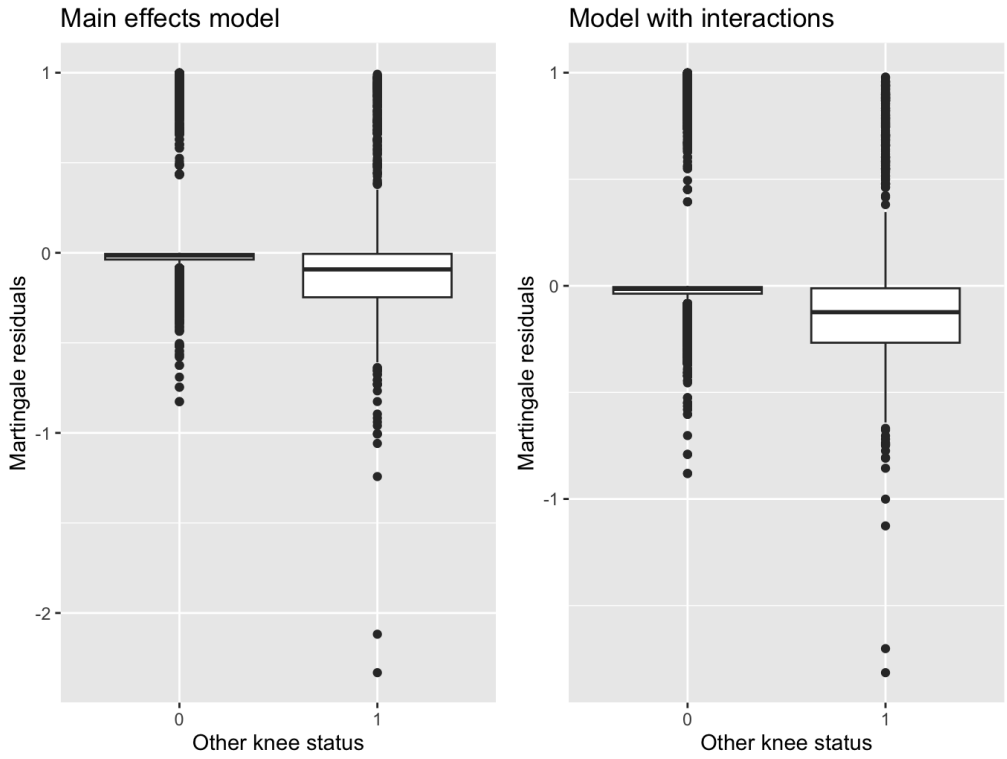

Figure C.6: Martingale residuals for the main effects and interactions model, plotted against the status of the other knee, in the scenario when studying the time to knee contracture given the status of the other knee.

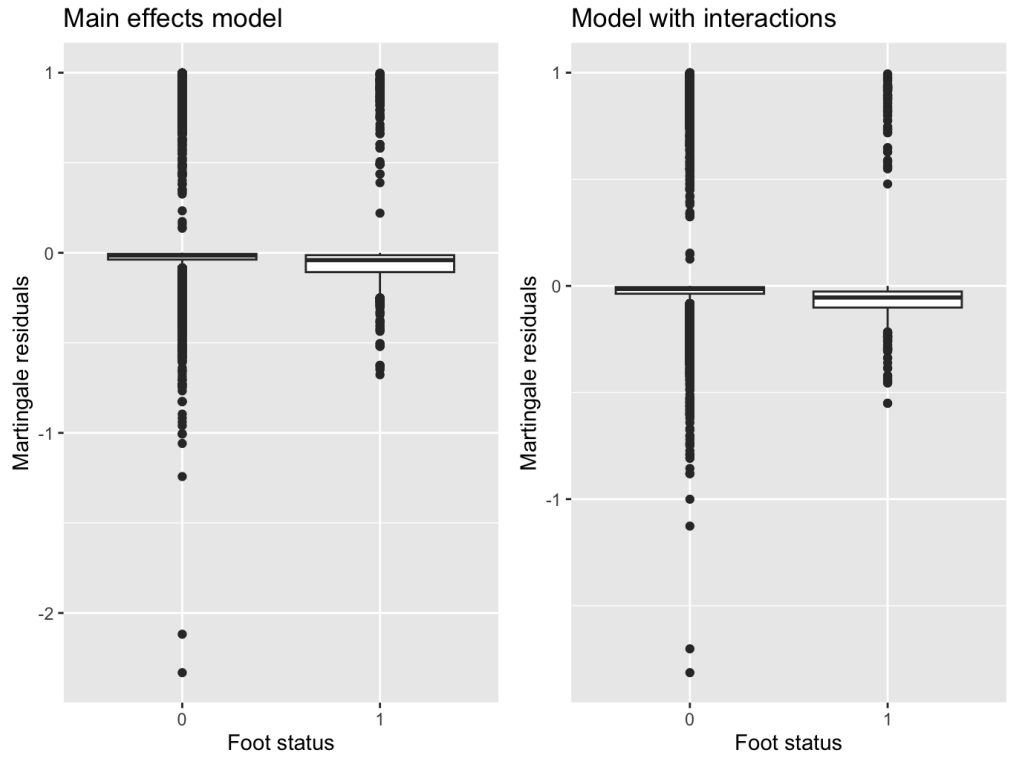

Figure C.7: Martingale residuals for the main effects and interactions model, plotted against the foot status, in the scenario when studying the time to knee contracture given the status of the other knee.

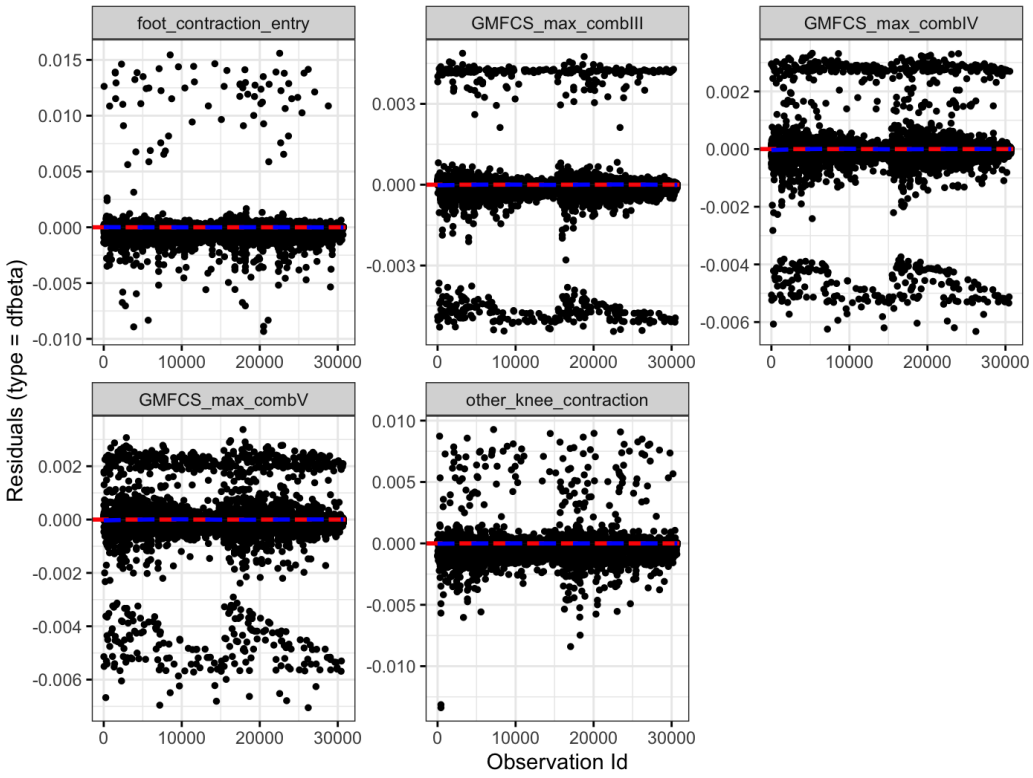

Figure C.8: dfbeta residuals for the main effects model, in the scenario when studying the time to knee contracture given the status of the other knee.

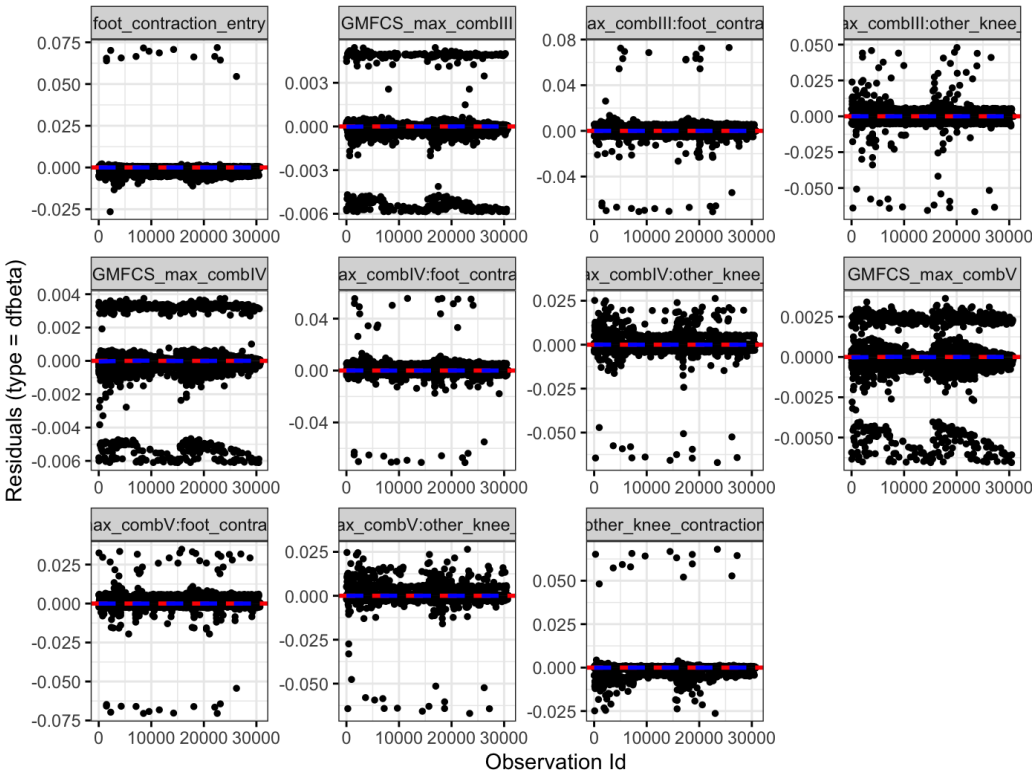

Figure C.9: dfbeta residuals for the model with interactions, in the scenario when studying the time to knee contracture given the status of the other knee.

Bachelor's Theses in Mathematical Sciences 2024:K2
ISSN 1654-6229

LUNFMS-4072-2024

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/