Calculation of LTC Premiums based on direct estimates of transition probabilities

Florian Helms*, Claudia Czado†and Susanne Gschlößl‡
Technische Universität München, Zentrum Mathematik

Abstract

In this paper we model the life-history of LTC patients using a Markovian multi-state model in order to calculate premiums for a given LTC-plan. Instead of estimating the transition intensities in this model we use the approach suggested by Andersen et al. (2003) for a direct estimation of the transition probabilities. Based on the Aalen-Johansen estimator, an almost unbiased estimator for the transition matrix of a Markovian multi-state model, we calculate so-called pseudo-values, known from Jackknife methods. Further, we assume that the relationship between these pseudo-values and the covariates of our data are given by a GLM with the logit as link-function. Since the GLMs do not allow for correlation between successive observations we use instead the "Generalized Estimating Equations" (GEEs) to estimate the parameters of our regression model. The approach is illustrated using a representative sample from a German LTC portfolio.

Keywords: Markovian Multi-State Model, Transition Probabilities, Aalen-Johansen Estimator, Pseudo-Values, GLM, GEE, LTC, Premium,

1 Introduction

The problem on how to incorporate increasing life expectancy in social welfare systems has become more urgent during the last decades. When looking for example at pension and health care systems in most industrialized countries it seems obvious that necessary adjustments have not been made. In particular the compressed morbidity hypothesis confronts the expanded morbidity hypothesis. The first one claims that increasing life expectancy is due to a decreasing morbidity at all ages, whereas the latter states that gained years of life expectancy are entirely spent in illness. Reality is probably somewhere in between. But not only newborn children have a higher life expectancy today, also the life expectancy of older people has been increased. The numbers from

*email: florian.helms@web.de
†email: cczado@ma.tum.de
‡email: susanne@ma.tum.de

the Federal Statistical Office in Germany from the "Abbreviated Mortality Table for 1999/2001" show that a 60-year old male can still expect to live for another 19.5 years, a female of the same age for another 23.7 years.

In addition to this higher life expectancy the so often quoted demographic development changes the life of older people, as well (see CoE (2003)). As a result of a more individualized society and a different family structure people tend to live more frequently alone. For example at older age they might require some kind of external assistance to manage the tasks of daily life. To pay for this external assistance, in most industrialized countries some kind of long term care insurance (LTCI) was established and added to the social welfare system. Insurance companies also have developed full or additional insurance cover for long term care (LTC).

In Germany compulsory LTCI was introduced in 1995. The leading idea was, that "LTCI follows health insurance". The main difference between LTCI and health insurance is the duration of the period care is needed. According to German law (see §14 of the 11^{th} Sozialgesetzbuch), LTC-beneficiaries are "persons, who on account of a physical, mental or psychic illness or disability are in considerable or even more serious need of care for usual and regular recurring activities of daily living on a continuing base, presumably for at least six months."

In 1995, persons in the public health insurance system were given insurance coverage through the compulsory LTCI-system, paying 1.7% of their income. Persons with private health insurance were also insured in a private LTCI-system without any underwriting, paying premiums according to actuarial rates. In both LTCI-systems the same benefits are paid. However insurance companies also offer additional coverage for public or private insured. As a consequence of the change in the social welfare laws private insurance companies had to take over a large claim portfolio right at the start of the compulsory LTCI.

We used for our analysis a representative random sample of such a portfolio from the private compulsory LTCI-system. In general the benefits depend on the place of care and on the level of care measuring the severeness of assistance needed. External assistance might be provided at home or in a nursing home. The severeness of assistance needed is expressed in terms of level of care: People in "Level 1" are in considerable need of care, people in "Level 2" in serious need of care and in "Level 3" in extreme need of care.

The above situation of different levels and places of care can be modeled by a three-state Markovian model, with states "Care at home", "Care in a nursing home" and "Death". Different lives are observed over time and their transitions between these states are recorded. The transition between different states can be expressed in terms of transition probabilities, that is the probability that a person with age z lives in state g at time t and transfers to state h within the next year. A representation of the transitions in terms of transition intensities is also possible and used in Cox's proportional hazard-model (see Czado and Rudolph (2002)) under the competing risk assumption (see Andersen et al. (2001)). Other approaches, for example Levikson and Mizrahi (1994) determine transition probabilities in a Markovian multi-state model, that depend on age and health-status of the beneficiaries, whereas Jones and Willmot (1993) assume that people become LTC-patients according to a non-homogeneous Poisson processes and transition probabilities between different levels of care are fixed and known. This leads to a distribution function of the number of lives requiring care at a certain level at an arbitrary future time.

For actuarial purposes transition probabilities are needed. In the case of Cox's proportional hazard-model the transition probabilities are calculated from the transition intensities, that have to sum up to zero over the different states, using a relationship given by the set of Kolmogorov forward differential equations (see Haberman and Pitacco (1999)). Thus the transition probabilities are complex non-linear functions of the intensity regression coefficients.

In this context Andersen et al. (2003) developed a method that models the transition probabilities directly. This method calculates pseudo-values based on the Aalen-Johansen estimator, an almost unbiased estimator of the transition matrix of an Markovian multi-state model. These pseudo-values are then used in Generalized Estimating Equations (GEEs), that take, in contrast to Maximum Likelihood Estimation correlation between observations into account, to estimate the parameters of the model.

The goal of this paper is to introduce the approach of Andersen et al. (2003) to actuarial scientists and to demonstrate that this method can be successfully implemented for calculating LTC premiums. For this we provide a complete description of all statistical and actuarial tools necessary and illustrate its application to a representative random sample from the German private compulsory LTCI-system by deriving the necessary transition probabilities in order to calculate insurance premiums required for a given LTC-plan.

The paper is organized as follows: In Section 2 we introduce an appropriate insurance model and show how LTC-premiums are calculated in this model using actuarial values. To obtain these actuarial values we need to estimate the transition probabilities from our data. Thus we define in Section 3 the Aalen-Johansen estimator, a non-parametric and almost unbiased estimator of the transition matrix of a Markovian multi-state model.

Since the Aalen-Johansen estimator does not allow for covariates such as sex and age of claimants and only generates one outcome for a given set of data, we need pseudo-values to generate data required for a regression analysis and to construct a relationship between the Aalen-Johansen estimator and covariates associated with the observations. Therefore we introduce pseudo-values in Section 4 and explain their use in generalized linear models (GLMs), which are popular extensions of linear models. They are discussed for example in McCullagh and Nelder (1989) and Fahrmeir and Tutz (1994). The pseudo-values are calculated at different points in time involving the same observations. Therefore the assumption of independence required for GLMs does no longer hold and we have to introduce the generalized estimating equations (GEEs) in Section 5 that take correlation between observations into account.

In Section 6 we apply the methods discussed to the LTCI data set: We calculate the Aalen-Johansen estimator of a three-state model, derive the pseudo-values and thus generate the data for a regression analysis using GEEs, where we specify the logit as link function in a linear model. With the estimates obtained from this regression analysis we calculate finally the one-year transition probabilities of our insurance model. These transition probabilities are then used to calculate the actuarial values for a given LTC-plan and derive the necessary premiums. A summary with a comparison of our premiums with premiums offered by a German health insurer completes our analysis.

2 Markovian Multi-State Insurance Models

We use a Markovian three-state model with states corresponding to the places of care according to the German compulsory LTCI-system, that is with states "Care at home", "Care in a nursing home" and "Death" given in Figure 1:

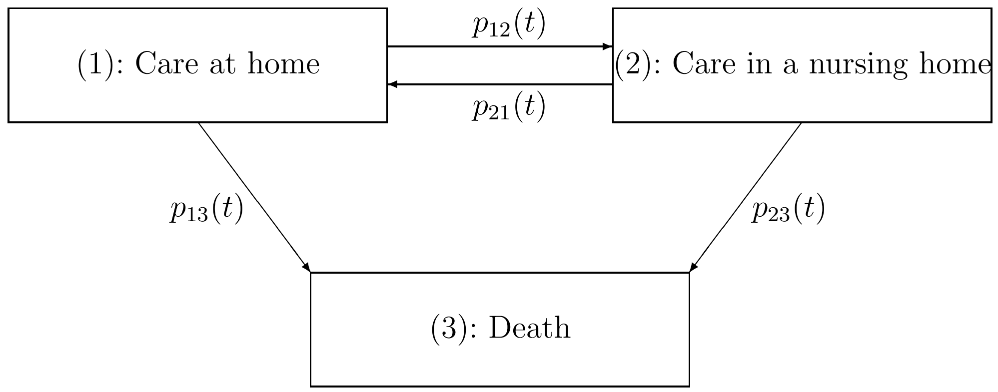

Figure 1: Markovian three-state model for LTC

The quantity $p_{gh}(t)$ denotes the one-year transition probability for a transition from state g to state h. We excluded transitions from state 2 to 1, consequently $p_{21}(t) := 0$ for all t since a transition from state 2 to 1 was observed very rarely, and we consider state 3 as an absorbing state for natural reasons. Thus $p_{31}(t)$ and $p_{32}(t)$ are zero, as well.

This three-state model only accounts for individuals that already qualified for one of the three levels of care. If we extend our model to the situation of an insurance company, we have to add the state "Active" to our model. In Figure 2 we added this state. For the area within the dotted line we were able to calculate the necessary transition probabilities using our data.

For transition probabilities from outside this area additional information is necessary, such as incidence rates for LTC and mortality rates for active lives, to derive all transition probabilities that are needed to calculate actuarial values and thus the premiums required for a given LTC-plan. For this we used incidence rates from "Custodial Insurance, Japan" and mortality rates from the "Bavarian life tables 1986-1988" (see e.g. Rudolph (2000) Appendix C.1 and C.2). Note that we excluded transitions from state 2 to 1 and from state 1 and 2 to state 0 since such transitions occurred very rarely and also transitions out of state 3 for obvious reasons.

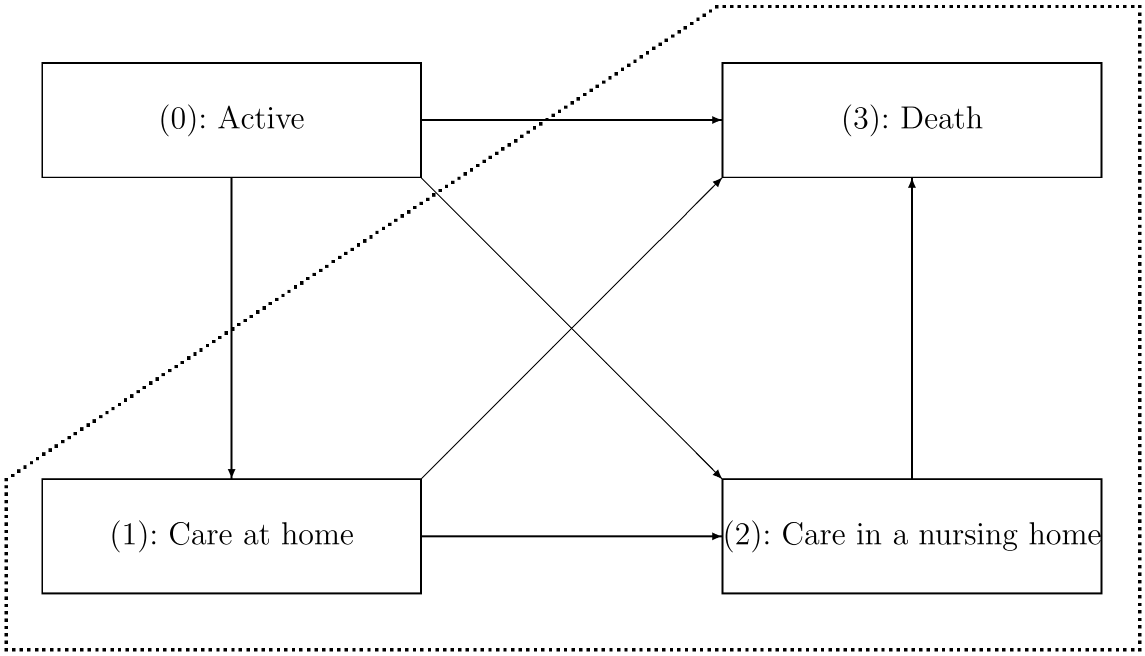

Figure 2: Markovian four-state Model for LTCI

We use the extended Markovian four-state model given in Figure 2 as the insurance model in which premiums will be calculated. Thus the life-history of an individual will be described by a time-continuous Markov process $S(t)$ with state space $\mathcal{S} = 0, 1, 2, 3$.

The transition probability of this time-continuous Markov process $S(t)$ is defined for all $0 \leq t \leq u$ and g, $h \in \mathcal{S}$ as

$$p_{gh}(t, u) := P(S(u) = h | S(t) = g).$$

The corresponding transition intensity is given by

$$\lambda_{gh}(t) := \lim_{dt\to 0} \frac{p_{gh}(t, t+dt)}{dt}.$$

Further we define the one-year transition probability as

$$p_{gh}(t) := p_{gh}(t, t+1).$$

In a time-discrete model one can use these one-year transition probabilities to calculate any transition probability $p_{gh}(t,u)$. Haberman and Pitacco (1999) showed this using a special case of the Chapman-Kolmogorov equations:

$$p_{gh}(t,u) = \sum_{k\in\mathcal{S}} p_{gk}(t, t+1) \cdot p_{kh}(t+1, u)$$

From an insurance point of view one distinguishes between inflows (e.g the premiums paid by the insured) and outflows (e.g. annuity benefits or lump sums paid by the insurer). Generally the following types of premiums and benefits are possible. Further details and examples can also be found in Haberman and Pitacco (1999):

- a continuous premium at a rate $p_g(t)$ at time t, if $S(t) = g$;
- a continuous annuity benefit at rate $b_h(t)$ at time t, if $S(t) = h$;
- a lump sum $c_{gh}(t)$, if at time t a transition occurs from state g to state h;

We denote by $p_g(t)dt$ the premium amount and by $b_h(t)dt$ the benefit amount paid out in the infinitesimal interval $[t, t + dt)$, respectively.

Since these cash-flows occur at different points in time we use the concept of random present values to compare inflows with outflows, that is we deflate inflows and outflows to the present time using the factor $v = \exp\{-\delta\}$ from the compound interest model, where the force of interest δ is assumed to be deterministic and constant.

Consider a continuous premium at rate $p_h(u)$ at time u, if $S(u) = h$. As already mentioned, $p_h(u)du$ is the premium amount paid out in the infinitesimal interval $[u, u + du)$. The random present value of this premium at time t is given by

$$Y_t^{p_h}(u, u + du) := v^{u-t} I_{\{S(u)=h\}} p_h(u)du.$$

The same continuous premium paid over the time interval $[u_1, u_2)$, with $t \leq u_1 < u_2$, has the following random present value at time t:

$$Y_t^{p_h}(u_1, u_2) := \int_{u_1}^{u_2} v^{u-t} I_{\{S(u)=h\}} p_h(u)du.$$

Consider a continuous annuity benefit at a rate $b_h(u)$ at time u, if $S(u) = h$. Again, $b_h(u)du$ is the benefit amount paid out over the infinitesimal interval $[u, u + du)$. The random present value of this benefit at time t is given by

$$Y_t^{b_h}(u, u + du) := v^{u-t} I_{\{S(u)=h\}} b_h(u)du.$$

The same continuous annuity benefit on the time interval $[u_1, u_2)$, with $t \leq u_1 < u_2$, has the following random present value at time t:

$$Y_t^{b_h}(u_1, u_2) := \int_{u_1}^{u_2} v^{u-t} I_{\{S(u)=h\}} b_h(u)du.$$

Consider a lump sum $c_{hk}(u)$, paid just after time u, if a transition from state h to k occurs at time u. The random present value of this lump sum at time t is given by

$$Y_t^{c_{hk}}(u) := v^{u-t} I_{\{S(u-)=h,\, S(u)=k\}} c_{hk}(u).$$

In the following these random present values will be used to calculate the actuarial values, which are the basic tool to determine premiums and reserves in actuarial sciences.

Actuarial values are expected present values. In addition to the financial structure of random present values we need now a probabilistic structure, as well. Thus we use at this point the assumption that the life-history can be modeled as a time-continuous Markov chain. Further we suppose that the risk is in state g at time t, that is $S(t) = g$, and define the actuarial values as a conditioning event. Actuarial values are therefore conditional expected present values. Following Czado and Rudolph (2002) we define:

Definition 2.1 (Actuarial values) *Actuarial values are expected present values. Assuming that the insured risk is in state g at time t, then the actuarial values are given as conditional expectations of the random present values, that is*

- $E[Y_t(u)|\, S(t) = g]$ *for lump sum payments*
- $E[Y_t(u, u + du)|\, S(t) = g]$ *for annuities*

In the following we are going to specify the actuarial values for the random present values introduced earlier:

The actuarial value of the continuous premium at rate $p_h(u)$ at time u, if $S(u) = h$, is

$$E\left[Y_t^{p_h}(u, u + du)|\, S(t) = g\right] = v^{u-t} p_{gh}(t, u) p_h(u)du.$$

Further we define for $u_1 < u_2$

$$E\left[Y_t^{p_h}(u_1,u_2)\middle|\, S(t)=g\right] = \int_{u_1}^{u_2} v^{u-t} p_{gh}(t,u) p_h(u) du.$$

In the interval $[t,n]$ we have the following cumulative premium function denoted by $\mathcal{P}_i(t,n)$:

$$\mathcal{P}_g(t,n) = \int_t^n v^{u-t} \sum_{h\in S} p_{gh}(t,u) p_h(u) du.$$

The actuarial value of the continuous annuity benefit at rate $b_h(u)$ at time u, if $S(u)=h$, is

$$E\left[Y_t^{b_h}(u,u+du)\middle|\, S(t)=g\right] = v^{u-t} p_{gh}(t,u) b_h(u) du$$

and for $u_1 < u_2$ we define

$$E\left[Y_t^{b_h}(u_1,u_2)\middle|\, S(t)=g\right] = \int_{u_1}^{u_2} v^{u-t} p_{gh}(t,u) b_h(u) du.$$

The actuarial value of a lump sum c_{hk} paid just after time u, if a transition from state h to k occurs at time u, is

$$E\left[Y_t^{c_{hk}}(u)\middle|\, S(t)=g\right] = v^{u-t} p_{gh}(t,u) \lambda_{hk}(u) c_{hk}(u)$$

and correspondingly for $u_1 < u_2$

$$E\left[Y_t^{c_{hk}}(u_1,u_2)\middle|\, S(t)=g\right] = \int_{u_1}^{u_2} v^{u-t} p_{gh}(t,u) \lambda_{hk} c_{hk}(u) du.$$

Taking all the benefits defined above together we obtain the cumulative benefit function $\mathcal{B}_g(t,n)$, that is

$$\begin{aligned}
\mathcal{B}_g(t,n) &= \int_t^n v^{u-t} \sum_{h\in S} p_{gh}(t,u) b_h(u) du \\
&+ \int_t^n v^{u-t} \sum_{h\in S} \sum_{k:k\neq h} p_{gh}(t,u) \lambda_{hk}(u) c_{hk}(u) du.
\end{aligned}$$

The principle of equivalence states, as mentioned by Czado and Rudolph (2002), that the expected amount of premiums has to be equal to the expected amount of benefits. This means that at the time when the policy is issued, the actuarial value of the benefits, that are paid under this contract, has to be the same as the actuarial value of the premiums, that are received by the insurer:

Definition 2.2 (The Principle of Equivalence) *For an insured risk with policy end at n and initial state $S(0) = 0$, the equivalence principle is satisfied if and only if*

$$\mathcal{P}_0(0,n) = \mathcal{B}_0(0,n)$$

Clearly this relationship might be satisfied by an infinity of premium functions. The so-called "funding condition" is a further constraint. At any time during the insurance contract is in force we require that

$$\mathcal{B}_{S(t)}(t,n) \geq \mathcal{P}_{S(t)}(t,n)$$

Since the principle of equivalence only has to be fulfilled at policy begin, we are able to construct insurance contracts with increasing, decreasing or level premiums according to given laws or customers needs.

Assume an insured pays a premium π in state 0, i.e. "Active", and receives a lump sum c_{0h} for a transition to state h and an annuity b_h in state h until death, where h might be 1 or 2, i.e. "Care at home" or "Care in a nursing home". The discretized version of above actuarial values can be obtained correspondingly to the time-continuous case, i.e. we have to calculate the expectation of the random present values. This is nothing else than the sum over the values of the outcome multiplied by the probability of this outcome happening. Note that $P_{0h}(0,t)$ is the probability to be at time 0 in state 0 and to transfer to state h at time t, whereas $p_{0h}(t)$ is the probability to transfer from state 0 to state h in the interval $(t, t+1]$. As seen above, we can express $P_{0h}(0,t)$ in terms of the one-year transition probabilities $p_{0h}(t)$, $h \in \mathcal{S}$ (see Haberman and Pitacco (1999)). We obtain the following actuarial values for a z-year old insured:

$$\begin{aligned}
P_{0,\pi} &:= \sum_{t=0}^{\omega-z-1} P_{00}(0,t) v^t \pi; \\
B_{0,c_{0h}} &:= \sum_{t=0}^{\omega-z-1} P_{00}(0,t) p_{0h}(t) v^t c_{0h}; \\
B_{0,b_h} &:= \sum_{t=0}^{\omega-z-1} P_{0h}(0,t) v^t b_h
\end{aligned}$$

where ω is the limiting age, i.e. the probability to survive beyond ω is assumed to be zero. Outflows occur for transitions to state 1 and 2. Therefore we have to add up the actuarial values of these outflows and obtain, using the principle of equivalence, the following equation:

$$P_{0,\pi} = \sum_{t=0}^{\omega-z-1} P_{00}(0,t) v^t \pi = \sum_{h=1}^{2} B_{0,c_{0h}} + \sum_{h=1}^{2} B_{0,b_h}$$

We simply factor π out and divide the right hand side by the rest of the sum giving us the premium to be charged for any values of b_h and c_{0h}. For any LTC-plan we can calculate the necessary LTC-premiums using actuarial values and the principle of equivalence if we know the financial and probabilistic structure. Thus we need to estimate the one-year transition probabilities for a z-year old individual that has been LTC-claimant for t years out of our set of data, derive the actuarial values and calculate the necessary premium.

3 Aalen-Johansen Estimator

As mentioned before we need a estimator for the transition matrix of a Markovian multi-state model, in our case a three-state model with states "Care at home", "Care in a nursing home" and "Death". The following notations will be used:

- $t_1 < t_2 < \dots$ are times, where transitions are observed.
- $d_{gh}^{(j)}$ is the number of lives, that transfer from state g to state h at time t_j.
- $r_g^{(j)}$ is the number of lives in state g, alive and uncensored just prior to time t_j.

If only one transition is observed at all $t_j's$, the Aalen-Johansen estimator, a non-parametric estimator for the transition matrix of a Markovian multi-state model $P(t,u) = (P_{gh}(t,u))_{g,h \in \mathcal{S}}$, is according to Aalen and Johansen (1978) defined as

$$\hat{\mathbf{P}}(t,u) := \prod_{j:t<t_j\leq u} \begin{pmatrix} 1 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 1-\frac{d_{gh}^{(j)}}{r_g^{(j)}} & \dots & \frac{d_{gh}^{(j)}}{r_g^{(j)}} & \dots & 0 \\ 0 & \dots & \dots & 1 & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & 1 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 1 & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 1 \end{pmatrix} \tag{1}$$

assuming the only transition at time t_j occurs from state g to state h. This estimator is a product of matrices at each time a transition is observed. Element (g,g) is equal to $1 - d_{gh}^{(j)}/r_g^{(j)}$ and element (g,h) is equal to $d_{gh}^{(j)}/r_g^{(j)}$, which is the number of lives that transferred from state g to h at time t_j divided by the number of lives in state g just prior to time t_j; all other diagonal elements are equal to one, whereas other off-diagonal elements equal to zero.

If we take a two-state model with states "Alive" and "Death" the Aalen-Johansen estimator reduces to the well-known Kaplan-Meier estimator (Kaplan and Meier (1958)).

It is clear from the definition that $\hat{P}(t,u)$ is a stochastic matrix $\forall t,u$ and one can see that the Chapman-Kolmogorov equations hold, as well. So all requirements on the transition probabilities of a Markovian multi-state model are fulfilled. If more than one transition occurs or transitions to different states at the same time the entries of this matrix have to be modified accordingly (see Helms (2003)).

Further one can show, using the theory of counting processes, more precisely Duhamel's equation, that the Aalen-Johansen estimator is almost unbiased, which is of importance in the following. It is even an unbiased estimator if the probability that $r_g^{(j)} = 0$ is equal to zero for all times t_j and states g. Given a large sample of observations this is usually the case and we refer in the following to the Aalen-Johansen estimator as an unbiased estimator. The necessary theory and a proof of this result can be found in Andersen et al. (1993).

The Aalen-Johansen estimator is calculated using the life-history of all lives, but does not take their associated covariates into account, and we receive only a single estimate of the transition matrix. Thus we need a way to generate the data required for a regression analysis and construct a relationship between the Aalen-Johansen estimator and the covariates of each live under observation. This can be done using pseudo-values.

4 Pseudo-Values

We assume we are given a sample of n observations $\mathbf{x} = (x_1, \ldots, x_n)$, e.g. the claim-history of LTC-patients, and an estimator $\hat{\theta} = s(\mathbf{x})$, e.g. the Aalen-Johansen estimator; according to Efron and Tibshirani (1993) we define the i^{th}-jackknife sample of $\mathbf{x}$ as

$$\mathbf{x}_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

This is the same data set as $\mathbf{x}$, but with the i^{th}-observation removed. The i^{th}-jackknife replication of $\hat{\theta}$, the so-called "leave-one-out" estimator, is defined as

$$\hat{\theta}_{-i} := s(\mathbf{x}_{-i})$$

This is nothing else than the estimator $\hat{\theta}$ based on the i^{th}-jackknife sample; in other words, we calculate the estimator $\hat{\theta}$ without the i^{th}-observation. The i^{th}-pseudo-value of $\hat{\theta}$ is defined as

$$\tilde{\theta}_i := \hat{\theta} + (n-1) \cdot (\hat{\theta} - \hat{\theta}_{-i}) = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-i} \tag{2}$$

Jackknife is usually used to detect outliers and check the bias and precision of an estimator: For a good estimator $\hat{\theta}$ one would expect that the i^{th}-jackknife replication of $\hat{\theta}$ is roughly the same as $\hat{\theta}$ and consequently equal to the i^{th}-pseudo-value. Andersen et al. (2001) suggested to go now a step further and perform a regression analysis using these pseudo-values $\tilde{\theta}_i$: For n observations, we are able to calculate n pseudo-values and thus generate data required for a regression analysis. But still, there is nothing which links the i^{th}-pseudo-value with the covariates of any observation. How this link can be provided is explained in the following:

Let x_i be the realization of a random variable X_i, e.g. the claim-history of the i^{th}-LTC-patient. We assume that the X_i's are independent and identically distributed with expectation θ, and that an unbiased estimator $\hat{\theta}$ is available for θ, e.g. the Aalen-Johansen estimator, that is $E[\hat{\theta}] = \theta$.

Given i.i.d. covariates $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})^T$, which might be age, sex, level of care etc. with distribution function C, we can write:

$$\theta = E[X_i] = E[E[X_i|\mathbf{Z}_i]] = \int_0^{\infty} E[X_i|\mathbf{Z}_i = \mathbf{z}]dC(\mathbf{z})$$

We define

$$\theta_i := E[X_i|\mathbf{Z}_i = \mathbf{z}_i],$$

where $\mathbf{z}_i$, $i = 1, \dots, n$ are the observed covariate values. Estimating C, the distribution of the i.i.d. covariates $\mathbf{Z}_i$, by its empirical distribution $\hat{C}$ the parameter θ can be interpreted as the simple average of the θ_i's since

$$\theta \approx \int_0^{\infty} E[X_i|\mathbf{Z}_i = \mathbf{z}]d\hat{C}(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} \theta_i.$$

Since $E[\hat{\theta}] = \theta$ we have $E[\hat{\theta}] \approx \frac{1}{n}\sum_{i=1}^{n} \theta_i$. The same holds true if we remove one observation since the data are assumed i.i.d. and we can go through the above steps with the i^{th}-jackknife sample and it follows that

$$E\left[\hat{\theta}_{-i}\right] \approx \frac{1}{n-1}\sum_{j \neq i} \theta_j.$$

Since the data is only available given the covariates, we need a link between the estimator $\hat{\theta}$ and the quantity $\theta_i = E\left[X_i|\,\mathbf{Z}_i = \mathbf{z}_i\right]$. Therefore Andersen et al. (2001) defined the quantity $\tilde{\theta}_i$ as the summary statistic $\hat{\theta}$ based on the entire sample modified in the direction given by the "leave-one-out estimator" $(\hat{\theta} - \hat{\theta}_{-i})$, that we defined as the pseudo-values above. Using the above results it follows that

$$E[\tilde{\theta}_i] = E[n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-i}] \approx \sum_j \theta_j - \sum_{j \neq i} \theta_j = \theta_i.$$

Where does this lead to? The expectation of $\tilde{\theta}_i$ is approximately equal to θ_i, that takes, in contrast to θ, the covariates $\mathbf{Z}_i = \mathbf{z}_i$ into account. In other words we match the covariates of the i^{th}-observation with the i^{th}-pseudo-value. This establishes a relationship between the covariates and the pseudo-values. For our data we use the Aalen-Johansen estimator of the transition matrix from time t to u based on the whole data set as well as on the data set with the i^{th}-observation removed to construct the i^{th}-pseudo-value of the transition matrix from time t to u. The i^{th}-pseudo-value can then be matched to covariates of the i^{th}-claimant. We assume that the relationship between the pseudo-values $\tilde{\theta}_i$ and the covariates is, for example, given by a GLM for $\tilde{\theta}_i$ with link function $g(\cdot)$, see McCullagh and Nelder (1989), that is

$$g(\tilde{\theta}_i) = \mathbf{Z}_i^T \boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is the parameter vector to be estimated. For our data we choose the logit as link-function and normal errors, i.e.

$$\tilde{\theta}_i = \frac{\exp\{\mathbf{Z}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{Z}_i^T \boldsymbol{\beta}\}} + \varepsilon_i \qquad \text{with} \qquad \varepsilon_i \sim N(0, \sigma^2) \qquad \text{i.i.d.}$$

This leads us to a regression model for the transition probabilities over the whole period transitions are observed. But we are instead more interested in the change of the transition probabilities over time, since we need the one-year transition probabilities to calculate the necessary actuarial values. Therefore we have to extend above model to a multivariate one: We consider a series of time-points $t_0, \ldots, t_k$ and define $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{\theta}}(t_0), \ldots, \hat{\boldsymbol{\theta}}(t_k))$ calculating the pseudo-values analogue to (2) as

$$\tilde{\boldsymbol{\theta}}_{il} := n \cdot \hat{\boldsymbol{\theta}}(t_l) - (n-1) \cdot \hat{\boldsymbol{\theta}}_{-i}(t_l) \qquad i = 1, \ldots, n \qquad l = 0, \ldots, k \qquad (3)$$

For our purpose we use the Aalen-Johansen estimator (1) and define the pseudo-values for each element of the transition matrix:

$$\hat{\boldsymbol{\theta}}^{(gh)}(t) := \hat{\mathbf{P}}_{gh}(t, t+1) := \hat{\mathbf{P}}_{gh}(t) \qquad t = 0, \ldots, T$$

giving us $\hat{\boldsymbol{\theta}}^{(gh)} = (\hat{\boldsymbol{\theta}}^{(gh)}(0), \ldots, \hat{\boldsymbol{\theta}}^{(gh)}(T))$. We calculate the Aalen-Johansen estimator based on the entire sample and the "leave-one-out estimator" for all $i = 1, \ldots, n$ and $t = 0, \ldots, T$ giving us the pseudo-values $\tilde{\mathbf{P}}_{(i,gh)}(t)$ as defined in (3), i.e. $\tilde{\mathbf{P}}_{(i,gh)}(t) = n \cdot \hat{\mathbf{P}}_{(gh)}(t) - (n-1) \cdot \hat{\mathbf{P}}_{(-i,gh)}(t)$. Since the Aalen-Johansen estimator is a matrix, we obtain for each observation i one matrix of pseudo-values at each time-point t, $t = 0, \ldots, T$ giving us nT pseudo-value matrices, that is

$$\tilde{\mathbf{P}}_{(i)}(t) := (\tilde{\mathbf{P}}_{(i,gh)}(t), g, h \in \mathcal{S}). \qquad (4)$$

We assume now a regression model for each element of $\tilde{\mathbf{P}}_{(i)}(t)$ on $\mathbf{Z}_i$ to quantify the effect of the covariates on the transition probabilities. To perform this regression we assume that the relationship between $\tilde{P}_{i,gh}(t)$ and the covariates is given by a GLM with link function $g(\cdot)$, i.e.

$$g(\tilde{P}_{i,gh}(t)) = \alpha_t^{(gh)} + \mathbf{Z}_i^T \boldsymbol{\beta}^{(gh)}.$$

As covariates we include an intercept term, "Age", "Sex", "Level of care 2", "Level of care 3" and "Duration of Care", denoted by Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}, Z_{i5} and $\alpha_t^{(gh)}$. This means that $Z_{i1} = 1 \; \forall i = 1, \ldots, n$ and Z_{i2} denotes the age at transition time t.

Note that the covariates "Sex", "Level of care 2" and "Level of care 3" and "Duration of care" are factors, which will be dummy coded, i.e.

$$\begin{aligned}
Z_{i3} &:= \begin{cases} 0 & \text{if the individual is female} \\ 1 & \text{if the individual is male} \end{cases} \\
Z_{i4} &:= \begin{cases} 0 & \text{if the individual is in level of care 1 or 3} \\ 1 & \text{if the individual is in level of care 2} \end{cases} \\
Z_{i5} &:= \begin{cases} 0 & \text{if the individual is in level of care 1 or 2} \\ 1 & \text{if the individual is in level of care 3} \end{cases} \\
\alpha_t^{(gh)} &:= \begin{cases} 1 & \text{if the duration is in the interval } (t, t+1] \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Note that the Aalen-Johansen estimator is a matrix. Therefore we perform a regression analysis for each element of this matrix. Our model for the pseudo-values of the transition probability from state g to state h is the following:

$$\tilde{P}_{i,gh}(t) = \frac{\exp\{\alpha_t^{(gh)} + \mathbf{Z}_i^T \boldsymbol{\beta}^{(gh)}\}}{1 + \exp\{\alpha_t^{(gh)} + \mathbf{Z}_i^T \boldsymbol{\beta}^{(gh)}\}} + \varepsilon_{i,gh}(t) \qquad i = 1, \ldots, n \quad t = 0, \ldots, T \tag{5}$$

where $\varepsilon_{i,gh} \sim N(0, \sigma^2_{(gh)})$. Further $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}, Z_{i5})^T$ is the vector of covariates and $\boldsymbol{\beta}^{(gh)} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T$ the vector of parameters to be estimated. Define

$$\begin{aligned}
\eta_{i,gh}(t) &:= \alpha_t^{(gh)} + \mathbf{Z}_i^T \boldsymbol{\beta}^{(gh)} \in \mathbb{R} \\
\mu_{i,gh}(t) &:= \frac{\exp\{\alpha_t^{(gh)} + \mathbf{Z}_i^T \boldsymbol{\beta}^{(gh)}\}}{1 + \exp\{\alpha_t^{(gh)} + \mathbf{Z}_i^T \boldsymbol{\beta}^{(gh)}\}} = \frac{\exp\{\eta_{i,gh}(t)\}}{1 + \exp\{\eta_{i,gh}(t)\}} \in [0, 1].
\end{aligned}$$

Even though $\hat{P}_{gh}(t)$ and $\hat{P}_{-i,gh}(t)$ are estimated transition probabilities, i.e. restricted to the interval $(0, 1)$, the corresponding pseudo-value $\tilde{P}_{i,gh}(t) = n \cdot \hat{P}_{gh}(t) - (n-1) \cdot \hat{P}_{-i,gh}(t)$ does not need to satisfy this restriction for finite

n. Therefore we assume a normal error distribution for the pseudo-values and specify the link function with the logit $g(p) = \log\{\frac{p}{1-p}\}$ and the variance function as constant, that is

$$\tilde{P}_{i,gh}(t) \sim N(\mu_{i,gh}(t), \sigma^2_{(gh)}) \qquad i = 1, \ldots, n, \quad g, h \in \mathbf{S}$$

where $\mu_{i,gh}(t)$ is the value of the inverse of the link function evaluated at $\eta_{i,gh}(t)$.

The problem at hand now is a problem of longitudinal data analysis: Since the same lives are observed at different time points, correlation between the pseudo-values occurs in our model. Pseudo-values of different individuals can still be assumed to be independent, but the pseudo-values from the same observation at different times are obviously dependent. Therefore we use generalized estimating equations (GEEs) instead of standard Maximum Likelihood estimation in GLMs.

5 Parameter Estimation

In the last section we showed how to generate the data for a regression analysis using pseudo-values and provided reasons that allow us to match the i^{th}-pseudo-value with the covariates of the i^{th}-observation. In particular we proposed to use the GLM (5) to model the relationship between pseudo-values and their corresponding covariates. Since the pseudo-values of an individual at different time-points are correlated we have to allow for this correlation when parameters are estimated.

Generalizing Maximum Likelihood (ML) estimation leads us to Maximum Quasi-Likelihood (MQL) estimation . In the following we present this more general method since the further extension from MQL estimation to GEEs becomes then even more clear. Also, for a distribution from the exponential family, as in our case, the log-likelihood function is identical to the quasi-likelihood function and we obtain the same parameter estimates in ML and QML estimation anyway. For further details see Wedderburn (1974).

In contrast to the ML approach, where the whole distribution function of the outcomes y_i's is specified, for the MQL approach only the link- and variance function are specified. Wedderburn (1974) defined for each observation the quasi-likelihood function, denoted by $K(y_i, \mu_i)$, by the relationship

$$\frac{\partial K(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{v_i} \tag{6}$$

where $\mu_i := E[y_i]$ and $v_i := Var[y_i]$. The latter can be expressed by the variance function $k(\cdot)$ evaluated at μ_i and divided by the dispersion parameter

ϕ: $v_i = k(\mu_i)/\phi$. Equation (6) is analogue to the situation of a GLM using the ML approach, where the first order derivative of the log-likelihood function with respect to μ_i is given by

$$\frac{\partial l(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{v_i}$$

Both, ML estimation in GLMs and MQL estimation, can be performed by solving the set of the so called score equations for GLMs or the score-like equations for Quasi-Likelihood, respectively for a regression parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. The latter has the following form:

$$SL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial K(y_i, \mu_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial K(y_i, \mu_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \mu_i)\, v_i^{-1} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mathbf{0} \tag{7}$$

Using the Fisher Scoring method, one can show that the score equations for GLMs and the score-like equations for Quasi-Likelihood can be solved using an iterative weighted least-squares algorithm (see Wedderburn (1974)).

However, both approaches rely on the assumption of independence between observations. For longitudinal data this assumption still holds for the outcomes of different observations but not for the outcomes of the same observation at different points in time. Therefore we need a model that takes correlation into account. The GEE approach, introduced by Liang and Zeger (1986) and Zeger and Liang (1986) can be seen as the extension of Quasi-Likelihood to longitudinal data.

For this we observe for each subject i the response y_{it} at times $t = 1, \ldots, T$. In addition we have the corresponding covariate vectors $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itp})^T$ available. Further we define $\mu_{it} = E[y_{it}]$ and $v_{it} = Var[y_{it}]$ assuming that they exist. In vector notation we have the subject specific response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT})^T$ and the subject specific design matrix

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \ldots \\ \mathbf{x}_{iT} \end{pmatrix} = \begin{pmatrix} x_{i11} \ldots x_{i1p} \\ \ldots\ldots\ldots \\ x_{iT1} \ldots x_{iTp} \end{pmatrix} \quad \in \mathbb{R}^{T \times p}$$

For $T = 1$ we are in the situation of quasi-likelihood as sketched above. Additionally, if we specify the distribution of y_{it} as a distribution from the exponential family we recover the situation of the classical likelihood.

We assume now a similar regression setup as in GLMs. This means that the relationship between the linear predictor $\eta_{it} := \mathbf{x}_{it}\boldsymbol{\beta} = x_{it1}\beta_1 + \ldots + x_{itp}\beta_p$ and μ_{it} is given by the link function $g(\cdot)$:

$$g(\mu_{it}) = \eta_{it} = \mathbf{x}_{it}\boldsymbol{\beta} \qquad \text{or} \qquad \mu_{it} = g^{-1}(\eta_{it})$$

The variance v_{it} is defined as the variance function $k(\cdot)$ evaluated at μ_{it} and divided by the quantity ϕ, the so-called scale parameter, that is

$$v_{it} := k(\mu_{it})/\phi$$

This leads us for each observation to the following derivative of the quasi-likelihood function $K(y_{it}, \mu_{it})$ with respect to the mean function μ_{it}, that is

$$\frac{\partial K(y_{it}, \mu_{it})}{\partial \mu_{it}} = \frac{y_{it} - \mu_{it}}{v_{it}}$$

For each $\mathbf{y}_i$ we define the same $T \times T$ correlation matrix $R(\boldsymbol{\alpha})$, that is fully parametrized by a $s \times 1$ correlation parameter vector $\boldsymbol{\alpha}$. We only require $R(\boldsymbol{\alpha})$ to be a correlation matrix and call it the "working" correlation matrix, since we do not expect it to be correctly specified, though we want consistent estimates and consistent variances of these estimates. We define V_i using the $T \times T$ diagonal matrix $A_i := diag(\mathbf{v}_i \phi)$ with $\mathbf{v}_i = (v_{i1}, \ldots, v_{iT})^T$ as

$$V_i := A_i^{\frac{1}{2}} R(\boldsymbol{\alpha}) A_i^{\frac{1}{2}} / \phi. \tag{8}$$

If $R(\boldsymbol{\alpha})$ is the true correlation matrix for $\mathbf{y}_i$, the matrix V_i is equal to the true covariance matrix for $\mathbf{y}_i$, that is $V_i = Cov[\mathbf{y}_i]$, since

$$Cov[\mathbf{y}_i] = Var[\mathbf{y}_i]^{\frac{1}{2}} Corr[\mathbf{y}_i] Var[\mathbf{y}_i]^{\frac{1}{2}} = diag(\mathbf{v}_i \phi)^{\frac{1}{2}} R(\boldsymbol{\alpha}) diag(\mathbf{v}_i \phi)^{\frac{1}{2}} / \phi = V_i.$$

In addition to $\boldsymbol{\beta}$ and ϕ we have now also to estimate $\boldsymbol{\alpha}$. To obtain the GEE estimates of $\boldsymbol{\beta}$ we solve the score equations

$$U_G(\boldsymbol{\beta}) := \sum_{i=1}^{n} \frac{\partial l(\mathbf{y}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l}{\partial \boldsymbol{\mu}_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^T V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{9}$$

Further we define

$$D_{it} := \frac{\partial \mu_{it}}{\partial \boldsymbol{\beta}} \qquad \text{and} \qquad D_i := (D_{i1}, \ldots, D_{ik})^T = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \in \mathbb{R}^{k \times p}. \tag{10}$$

With (10) the score equations in (9) can be rewritten as

$$U_G(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i^T V_i^{-1} S_i = \sum_{i=1}^{n} U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{0}, \tag{11}$$

where $U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) := D_i^T V_i^{-1} S_i$ and $S_i = \mathbf{y}_i - \boldsymbol{\mu}_i$. If only one observation is available for each subject, that is $T = 1$, this equation becomes identical to the score-like equation (7) obtained for quasi-likelihood.

In contrast to the quasi-likelihood approach from Wedderburn (1974), the matrix V_i in the function $U_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ depends for each i not only on the parameter

$\boldsymbol{\beta}$ but also on the parameter $\boldsymbol{\alpha}$. A next step is now to replace $\boldsymbol{\alpha}$ and ϕ by any $n^{\frac{1}{2}}$-consistent estimators. Recall that an estimator $\hat{\theta}_n$ for θ is $n^{\frac{1}{2}}$-consistent if $\forall \eta > 0\ \exists$ a constant $k(\eta)$ and an integer $n(\eta)$ such that $P(|n^{-\frac{1}{2}}(\hat{\theta}_n - \theta)| \leq k(\eta)) \geq 1 - \eta\ \forall n > n(\eta)$. Assuming that $\boldsymbol{\beta}$ and ϕ are known we denote the estimator for $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi) := \hat{\boldsymbol{\alpha}}(Y, \boldsymbol{\beta}, \phi)$. Given $\boldsymbol{\beta}$ we take $\hat{\phi}(\boldsymbol{\beta}) := \hat{\phi}(Y, \boldsymbol{\beta})$ as estimator for ϕ. We insert these $n^{\frac{1}{2}}$-consistent estimators in (11) and obtain:

$$U_G(\boldsymbol{\beta}) \approx \sum_{i=1}^{n} U_i\left(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}))\right) = \mathbf{0}$$

We define the GEE estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_G$, as the solution of this equation. Under mild regularity conditions, we have that

$$n^{\frac{1}{2}}\left(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}\right) \xrightarrow{\mathcal{L}} \mathbf{Z} \qquad \text{as } n \to \infty \tag{12}$$

where $\mathbf{Z}$ has a $N_p(\mathbf{0}, V_G)$ distribution. Further

$$V_G = \lim_{n\to\infty} n \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} Cov\left[\mathbf{y}_i\right] V_i^{-1} D_i\right) \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}.$$

Here $N_p(\boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. A sketch of the proof of this result can be found in Liang and Zeger (1986) while Helms (2003) provides further details. The GEE estimates of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ are determined by iterating between a iterative weighted least-squares algorithm based on Fisher-Scoring for β and moment estimation for $\boldsymbol{\alpha}$ and ϕ. To estimate the values of $\boldsymbol{\alpha}$ and ϕ, Liang and

Zeger (1986) use Pearson residuals, that can be calculated in each step of the iteration given the current value $\hat{\boldsymbol{\beta}}_C$ for $\hat{\boldsymbol{\beta}}_G$ as

$$\hat{r}_{it} := \frac{y_{it} - \hat{\mu}_{it}}{k(\hat{\mu}_{it})^{\frac{1}{2}}}$$

where we calculate $\hat{\mu}_{it} := g^{-1}(x_{it}\hat{\boldsymbol{\beta}}_C)$ with the current value for $\hat{\boldsymbol{\beta}}_G$. Let N be again the number of all observations, that is $N = nT$, we can get a new estimate of ϕ by

$$\hat{\phi}^{-1} = \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{\hat{r}_{it}^2}{N - p},$$

which is the longitudinal analogue of the familiar Pearson statistic (Zeger and Liang 1986). This can be seen as follows:

$$E\left[\sum_{i=1}^{n} \sum_{t=1}^{T} \frac{(y_{it} - \hat{\mu}_{it})^2}{k(\hat{\mu}_{it})/\phi}\right] = E\left[\sum_{i=1}^{n} \sum_{t=1}^{T} \frac{(y_{it} - \hat{\mu}_{it})^2}{\hat{Var}[y_{it}]}\right] \approx N - p$$

$$\Rightarrow \qquad \phi^{-1} \approx \frac{1}{N-p} E\left[\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{(y_{it}-\hat{\mu}_{it})^2}{k(\hat{\mu}_{it})}\right] \approx \sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\hat{r}_{it}^2}{N-p}$$

The estimation of the parameter $\boldsymbol{\alpha}$ depends on the correlation structure selected for the working correlation matrix. Liang and Zeger (1986) presented five different types of correlation structure: An "independent", "exchangeable", "unstructured", "autoregressive (AR-I)" and "one-dependent" working correlation. In the following we are going to define these correlation structures and give corresponding estimators for them:

When the working correlation matrix $R(\boldsymbol{\alpha})$ is chosen to be the identity matrix, we do not allow for any correlation between observations even if we measure the same observations at different points in time, i.e. we ignore the dependency present in the data. In this case estimation of the correlation matrix is unnecessary, since the correlation matrix is fixed to the identity matrix.

If we choose the working correlation matrix as "exchangeable", the correlation between different observations within a time series is the same regardless of the distance in time. In particular we assume

$$Corr[y_{it}, y_{it'}] = \alpha \qquad \forall \quad t \neq t'.$$

An estimator for α corresponding to this correlation structure is given by

$$\hat{\alpha} = \frac{\hat{\phi}}{n}\sum_{i=1}^{n}\sum_{t>t'}^{T}\hat{r}_{it}\hat{r}_{it'} / \left(\frac{1}{2}\cdot T\cdot(T-1)-p\right).$$

In the "unstructured" case, a totally unspecified working correlation matrix R is used. We have to estimate all $\frac{1}{2}\cdot T\cdot(T-1)$ correlations. This can be done by setting

$$\hat{R} := \frac{\hat{\phi}}{n}\sum_{i=1}^{n} A_i^{-\frac{1}{2}} S_i S_i^T A_i^{-\frac{1}{2}},$$

where $A_i := diag(\mathbf{v}_i\hat{\phi})$ and $S_i = \mathbf{y}_i - \boldsymbol{\mu}_i$.

The "autoregressive (AR-I)" working correlation is nothing else than the correlation structure of a continuous first-order autoregressive process (AR-I). This means that observations with the same distance in time have the same correlation and the correlation decreases polynomially as the distance increases. In particular we have

$$Corr[y_{it}, y_{it'}] = \alpha^{|t-t'|}.$$

Since in the case of an "autoregressive (AR-I)" correlation structure $E[\hat{r}_{it}\hat{r}_{it'}] \approx \alpha^{|t-t'|}$, Liang and Zeger (1986) suggested to estimate the parameter α by the slope estimate obtained from the regression of $\log\{\hat{r}_{it}\hat{r}_{it'}\}$ on $\log\{|t-t'|\}$ for $i = 1, \ldots, n$, $t = 1, \ldots, T$.

In contrast to the polynomial decrease of the correlations in the "autoregressive (AR-I)" case is the "dependent" correlation structure. Observations with the same distance do still have the same correlation, but for each distance a separate value, not necessarily decreasing, is assumed, i.e

$$Corr[y_{it}, y_{it'}] = \alpha_{|t-t'|} \qquad \text{if } t \neq t'.$$

A special case of the "dependent" working correlation matrix is the "one-dependent" structure. This is equivalent to the correlation structure of a stationary Markov process of degree one, i.e.

$$Corr[y_{it}, y_{it'}] \ := \ \begin{cases} \alpha_{|t-t'|} & \text{if } t \leq 1 \\ 0 & \text{if } t > 1 \end{cases}$$

The "one-dependent" correlation structure has $T-1$ parameters, that can be estimated by

$$\hat{\alpha}_t := \frac{\hat{\phi}}{n-p} \sum_{i=1}^{n} \hat{r}_{i,t}\hat{r}_{i,t+1}.$$

In the case where $\alpha_t = \alpha$ for all $t = 1, \ldots, T-1$ we can estimate the overall α by

$$\hat{\alpha} = \frac{1}{T-1} \sum_{i=1}^{T-1} \hat{\alpha}_t.$$

Since $\boldsymbol{\beta}_G$ and V_G are robust to the choice of the correlation structure (see Liang and Zeger (1986)) we obtain using the asymptotic normality unbiased estimates even if the correlation structure is misspecified. Clearly, if we choose the working correlation matrix close to the true correlation the estimates will be more efficient. For details on simulation studies using different correlation structures and misspecified correlation structures see Liang and Zeger (1986).

A program written by Mark X. Norleans allows for five correlation structures for the GEE estimation. They include the "independent", "exchangeable", which is called "compoundsymmetric", the "unstructured", "autoregressive

(AR-I)" and "dependent" correlation structure. GEE parameter estimation of $\boldsymbol{\alpha}$ and ϕ can be facilitated in this program. The program is designed for the statistical software program Splus can be obtained from http://lib.stat.cmu.edu/.

Another statistical software program called "Oswald" is developed by the Statistics Group at the University of Lancaster and can be obtained from http://www.maths.lancs.ac.uk/Software/Oswald/ as a Splus library. Here additional correlation structures are possible: Beside the above mentioned correlation structures one can choose between a "stationary Markov process" or a "non-stationary Markov process" structure of degree "Mv", where "Mv" is a quantity to be specified. Further a fixed user-specified matrix "R" can be used as well as the correlation structure of an autoregressive process of degree "Mv".

6 Application to LTC-Data

Now all necessary tools are given to estimate transition probabilities from a set of data containing the claim-history of LTC-patients directly. In summary we proceed as follows:

First calculate the non-parametric Aalen-Johansen estimates $\hat{\mathbf{P}}(t, t+1)$ for $t = 0, \ldots, T$ of the transition probability matrices ignoring the covariate information defined in (1). In the next step generate the i^{th} pseudo-value matrices $\tilde{\mathbf{P}}_{(i)}(t)$ defined in (4) for $i = 1, \ldots, n$. Assume that element by element the GLM given in (5) holds for $\tilde{\mathbf{P}}_{(i)}(t)$. To adjust for the correlation present among the pseudo-values $\tilde{\mathbf{P}}_{i,gh}(t)$, $t = 0, \ldots, T$, $i = 1, \ldots, n$, $g, h \in \mathcal{S}$ we apply now the GEE estimation approach described in Section 4.

For a specified working correlation matrix $\mathrm{R}(\boldsymbol{\alpha})$ and $\phi = \sigma^2$ we estimate $\boldsymbol{\beta}$, α and σ^2 by GEE. For choosing $\mathrm{R}(\boldsymbol{\alpha})$ we first fitted an "unstructured" working correlation. Based on the correlation matrix estimated in this approach we decided then which correlation structure would represent this estimated correlation matrix best and performed another regression using this new correlation structure as working correlation matrix.

Since for a transition from state 1 to 2, that is from "Care at home" to "Care in a nursing home", the values in the "unstructured" correlation matrix were quite small we chose the "independence" working correlation matrix, i.e. it is fixed to the identity matrix and does not need to be estimated. In contrast to a transition from state 1 to 2 we obtained for a transition from state 1 to 3 larger estimated correlations using the "unstructured" working correlation matrix. Since the correlations tend to decrease when the distance in time increases we considered an "autoregressive (AR-I)" correlation structure sufficient for this transition probability. In the case of a transition from state 2 to 3 we also decided to use the "autoregressive (AR-I)" working correlation matrix, since a similar behavior in the correlation estimate of the "unstructured" working correlation matrix can be observed.

In Table 1 we summarized the estimates obtained from the GEE approach for all possible transitions of the model shown in Figure 1, using above mentioned working correlations. The corresponding p-values are based on the asymptotic normality result presented in Section 5. In particular standard error estimates of the regression parameter estimates are based on an estimate of V_G given in (12).

working correlation	Independence		AR-I		AR-I	
transition from state g to h	$g = 1,\ h = 2$		$g = 1,\ h = 3$		$g = 2,\ h = 3$	
$\hat{\beta}_j^{(gh)}$	Values	Pr($\lvert t\rvert >$)	Values	Pr($\lvert t\rvert >$)	Values	Pr($\lvert t\rvert >$)
Intercept	-5.31	0.00	-2.16	0.00	7.40	0.00
Age	0.02	0.05	0.02	0.02	-0.06	0.00
Sex	0.66	0.00	0.36	0.07	1.51	0.00
Level of care = 2	-1.37	0.00	-0.69	0.00	-1.96	0.00
Level of care = 3	-1.38	0.00	-0.53	0.02	-2.27	0.00
$\hat{\alpha}_t^{(gh)}$						
Duration of care = 1	0.35	0.30	0.02	0.71	0.34	0.19
Duration of care = 2	0.40	0.25	0.18	0.01	-1.08	0.00
Duration of care = 3	0.90	0.01	0.88	0.00	-0.37	0.19
Duration of care = 4	1.24	0.00	0.99	0.00	-1.68	0.00
Duration of care = 5	0.10	0.85	0.74	0.00	-1.96	0.00
Duration of care = 6	1.53	0.00	0.41	0.01	-1.06	0.01
Duration of care = 7	-0.11	0.88	0.24	0.18	-0.65	0.16
Duration of care = 8	0.25	0.60	0.82	0.00	-0.09	0.88
Duration of care = 9	0.80	0.11	1.19	0.00	-2.28	0.00
Duration of care = 10	1.26	0.00	-0.48	0.12	-0.03	0.97

Table 1: Parameter estimates of Model 5 and their p-values for all transitions of the Markovian three-state model for LTC in Figure 1 using the GEE approach

Estimated transition probabilities can now be derived from these estimates for any given set of covariates simply by calculating for person i

$$\hat{P}_{gh}^{(i)}(t) := \frac{\exp\{\hat{\alpha}_t^{(gh)} + \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}^{(gh)}\}}{1 + \exp\{\hat{\alpha}_t^{(gh)} + \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}^{(gh)}\}}.$$

Generally the probability for a transition from state 1 to 2 or 3 increases with age, whereas it decreases for a transition from state 2 to 3, but very slightly. The values of all three estimated transition probabilities are higher for males than the values for females as indicated by a significant positive value of the estimated regression coefficient corresponding to the covariate "Sex".

Transitions out of state "Care at home" are more likely to happen to individuals in "Level 1", whereas the values for individuals in "Level 2" and "Level 3" are nearly the same. This can be understood looking at the regression coefficient estimates for "Level of care 2" and "Level of care 3" that are nearly the same for these transition probabilities. The probability of dying in state "Care at home" decreases from "Level 1" to "Level 2" before it increases to "Level 3" but does not reach the value from "Level 1" again. In contrast, for a transition from state 2 to 3 one observes decreasing transition probabilities for an increase in the severeness of care needed.

The "Duration of Care" causes rising transition probabilities for transitions out of state 1 until a duration of four years, then a decrease can be observed for three years until they increase again. For the transition from state "Care in a nursing home" to "Death" the values are very close together for all durations and only small changes occur. With these estimates we are finally able to

calculate the premiums for the LTC-plan "PET" sold by a German insurer. According to this LTC-plan the insured receives a certain allowance depending on the level of care needed. This is for "Care at home" 25% in "Level 1", 50% in "Level 2" and 75% in "Level 3" and for "Care in a nursing home" 100% of the allowance. Thus, the c_{0h}'s are zero and in the case of "Care at home" $b_h = 1 - 0.25 * (4 - h)$, where $h \in \{1, 2, 3\}$, for an unit allowance and in the case of "Care in a nursing home" $b_h = 1$, $h \in \{1, 2, 3\}$.

For the calculation of the premiums we use a modified version of a C-program, which needs the benefits, interest rate and transition probabilities as input. For details see Rudolph (2000). The results for the LTC plan "PET" obtained for a 10 EUR daily allowance and an interest rate of 3.5 % can be found in the left columns of Table 2. The right columns show premiums offered by a German health insurer.

We observe higher premiums when a GEE estimation approach is used compared to the commercial premiums, but we can see that the behavior with respect to age is similar as well as the proportion between males and females, see Figure 3. It also should be noted that the incidence rates and mortality rates for "Active" individuals include administrative costs, whereas the transition probabilities do not. Therefore a direct comparison between the calculated premiums using GEEs and the commercial premiums might not be sensible. Further, the LTC-definition in different countries, such as Japan and Germany, varies and therefore country-specific incidence rates might be necessary.

Age	Premium based on GEEs		Premium offered by German health insurer	
	Female	Male	Female	Male
20	04.94	03.85	02.12	01.70
25	06.14	04.80	02.92	02.33
30	07.70	06.05	03.90	03.10
35	09.76	07.72	05.05	04.01
40	12.51	09.97	06.44	05.13
45	16.14	12.95	08.16	06.52
50	21.01	16.95	10.39	08.36
55	27.58	22.40	13.32	10.86
60	36.48	29.88	17.31	14.40
65	48.53	40.06	22.01	18.84
70	64.55	53.62	29.04	25.71

Table 2: Comparison of Premiums for a 10 EUR daily allowance

7 Summary and Discussion

This paper gives a detailed introduction to a method proposed by Andersen et al. (2003) for estimating transition probabilities directly and its application to a large German LTC portfolio. In particular the necessary actuarial setup for calculating premiums based on a Markovian multi-state model is provided. For the required transition probabilities pseudo values of the Aalen-Johansen transition matrix estimators which are specific to the claim history of a LTC patient are generated. These are linked to patient specific covariates in a longitudinal GLM with normal errors and a logistic link function. The parameters of this longitudinal GLM are estimated using a GEE approach accounting for correlation within the claim history of a patient. These provide finally the required transition probability estimates for calculating LTC premiums. Diverse statistical tools ranging from survival analysis, jackknifing methods, GLMs to GEE estimation for longitudinal regression models are introduced and discussed to give a basis for understanding how this method for estimating transition probabilities directly is working.

Even though Andersen et al. (2003) have investigated the validity of their method through simulation some further points are worthwhile to be addressed in further research. More precisely, in our case there are no methods available at the moment to examine the goodness-of-fit or confirm the choice of link function in the longitudinal GLM used. The choice of time-points might also influence the results, but in our case the time-points are given, since we need one-year transition probabilities for the calculation of actuarial values. Further, more precise estimates might be obtained if the correlation matrix is chosen close to the true one.

An alternative to GEE estimation in longitudinal GLMs is to use a Bayesian

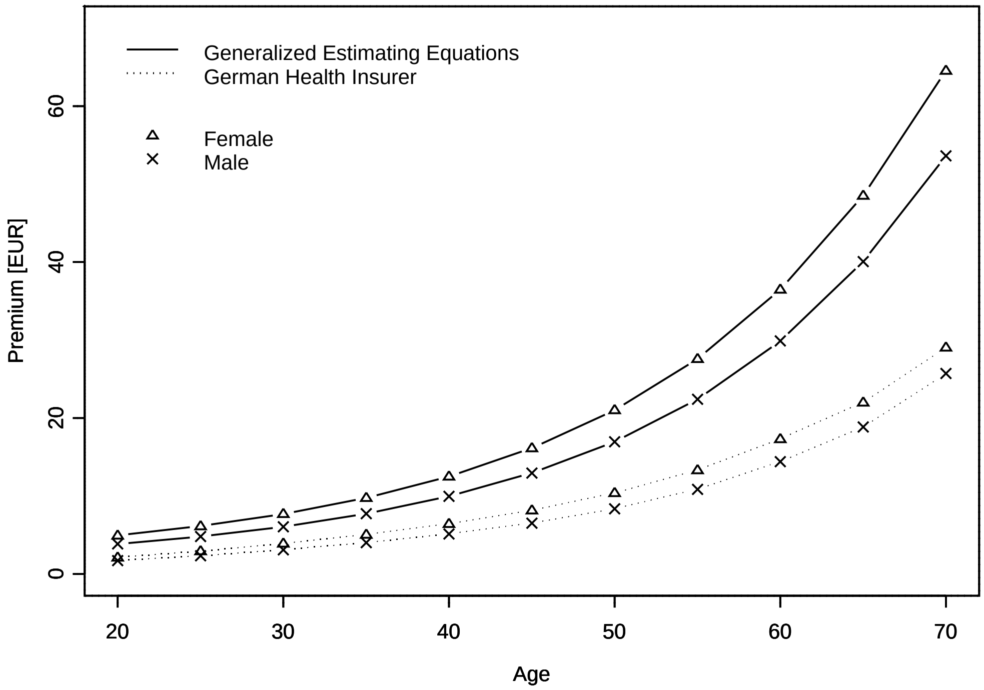

Figure 3: Comparison of Premiums

approach for estimation. This will require the use of Markov Chain Monte Carlo (MCMC) methods (see for example Gilks et al. (1996) and Gamerman (1990)). Bayesian model selection criteria such as Bayes factors (see Kaas and Raftery (1995)) and the deviance information criteria (DIC) by Spiegelhalter et al. (2002) can then be applied to assess goodness of fit and the choice of link function. This approach will be pursued in future research.

Acknowledgements

The second author was supported by Sonderforschungsbereich 386 *Statistische Analyse Diskreter Strukturen*, while the third author was supported by a doctoral fellowship within the Graduiertenkolleg *Angewandte Algorithmische Mathematik*, both sponsored by the *Deutsche Forschungsgemeinschaft.*

References

Aalen, O. O. and S. Johansen (1978). An Empirical Transition Matrix for Non-homogeneous Markov Chains based on Censored Observations. *Scand J Statist 5*, 141–150.

Andersen, P. K., S. Z. Abildstrom, and S. Rosthøj (2001). Competing risks as a multi-state model. *Research Reports 2001 http://www.pubhealth.ku.dk/bsa/publ-e.htm*.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Andersen, P. K., J. P. Klein, and S. Rosthøj (2001). From Summary Statistics to Generalized Linear Models for Pseudo-Observations; with Applications to Multi-State Models. *Research Reports 2001 http://www.pubhealth.ku.dk/bsa/publ-e.htm*.

Andersen, P. K., J. P. Klein, and S. Rosthøj (2003). Generalised Linear Models for correlated Pseudo-Observations, with Applications to Multi-State Models. *Biometrika 90, 1*, pp. 15–27.

CoE (2003). Recent demographic developments in europe 2002. *www.coe.int*.

Czado, C. and F. Rudolph (2002). Application of Survival Analysis Methods to Long Term Care Insurance. *Insurance: Mathematics and Economics 31*, 395–413.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modelling Based On Generalized Linear Models*. New York: Springer.

Gamerman, D. (1990). *Stochastic simulation for Bayesian inference*. London: Champman & Hall.

Gilks, W. R., W. S. Richardson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Haberman, S. and E. Pitacco (1999). *Actuarial Models for Disability Insurance*. London: Chapman & Hall.

Helms, F. (2003). Estimating LTC-Premiums using GEEs for Peudo-Values. *Technische Universität München Diplomarbeit*.

Jones, B. L. and G. E. Willmot (1993). An open group long-term care model. *Scand. Actuarial J. 2*, 161–172.

Kaas, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kaplan, E. L. and P. Meier (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association 53, 282*, 457–481.

Levikson, B. and G. Mizrahi (1994). Pricing long term care insurance contracts. *Insurance: Mathematics and Economics 14*, 1–18.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika 73, 1*, 13–221.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models 2nd ed.* London: Chapman and Hall.

Rudolph, F. (2000). Anwendungen der Überlebenzeitanalyse in der Pflegeversicherung. *Technische Universität München Diplomarbeit*.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *J. Roy. Stat. Soc. B. 64*, 583–639.

Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika 61, 3*, 439–447.

Zeger, S. L. and K.-Y. Liang (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics 42, 1*, 121–130.