The inception selection effect of diagnosis in a German long term care portfolio

Claudia Czado Susanne Gschlößl *

July 8, 2003

Abstract

In this paper we quantify the inception selection effect of diagnosis in a large German long term care (LTC) portfolio. First we are interested in modeling transition intensities, which will then be used in a multistate model set up to estimate transition probabilities. Finally we use these probability estimates as the basis for premium calculations. For the estimation of transition intensities we use semiparametric hazard models introduced by Cox (1972) allowing the inclusion of diagnosis as explanatory variable. Using modern model diagnostics we build a statistical model for the transition intensities and show that the resulting transition probability estimates including diagnosis perform better than when diagnosis is neglected. To quantify the inception selection effect of diagnosis we show how these improved transition probability estimates affect the premiums in an LTC insurance contract. In particular for younger age groups higher premiums are obtained when the diagnoses are taken into account compared to a model which disregards diagnosis. This demonstrates the actuarial need for allowing for an inception selection effect of diagnosis.

Keywords: semiparametric hazard model, survival analysis, long term care insurance, multistate model, inception selection

*Both at Center of Mathematical Sciences, Munich University of Technology, Boltzmannstr.3, D-85747 Garching, Germany, email: cczado@ma.tum.de, susanne@ma.tum.de, http://www.ma.tum.de/m4/

1 Introduction

In this paper a statistical and actuarial analysis of long term care (LTC) insurance data is conducted. Several authors have dealt with this topic. Levikson and Mizrahi (1994) consider Markovian multi-state models for pricing LTC insurance contracts. For this, they use transition probabilities which depend only on age and on the health of the insured persons. Premiums are then determined by using backward induction methods for given transition probabilities. Jones and Willmot (1993) present a stochastic multi-state model to analyse future requirements and costs in long-term care. Individuals are supposed to enter LTC according to a non-homogeneous Poisson process, while transitions among different care levels are only specified by assuming fixed known transition probabilities. They derive the distribution of the number of individuals requiring care at each level at an arbitrary future time.

Our aim is the modeling of transition intensities between states. Czado and Rudolph (2002) examined part of an LTC-claim portfolio of a German health insurance using a Cox proportional hazard model. They have shown that besides age of the claimant and time spent in LTC, also factors like gender, severness of the claim and type of care have a significant influence on survival. We want to analyse the same data, taking the diagnoses which led to LTC into account which are additionally given in the data. The main purpose of this paper is to investigate the effects a neglect of the information given by the diagnosis has on the transition intensities and probabilities. We will show that the inclusion of this information leads to more realistic transition rates and probabilities. In particular we show that estimated mortality probabilities including diagnosis are closer to the observed empirical mortality probabilities.

Finally, we want to investigate the inception selection effect of diagnosis on premiums calculated in a LTC insurance product, where annuities are paid depending on type of care and care level when long term care is required. Although the overall influence of the diagnoses on the premiums is rather slight, they should be taken into account when pricing LTC premiums for clients between 20 - 40 years. Without the information of the diagnoses premiums seem to be underestimated for this group.

The paper is organized as follows. In Section 2 an introduction to Cox's semiparametric model is given. The estimation of the transition intensities to death is given in Section 3. An important point is the assessment of model fit which is presented in Section 4. We use fractional polynomials proposed by Royston and Altman (1994) and an exponential approach to model the influence of continuous covariates on the transition intensities. The assumption of proportional hazards is checked using scaled Schoenfeld residuals (see Grambsch and Therneau (1994)). In Section 5 an estimation of transition probabilities is conducted using the estimated hazard rates as transition rates in a multiple state model. An actuarial application including premium calculations for specific LTC contracts is given in Section 6. A summary and discussion complete the paper.

2 Cox's semiparametric hazard model

Cox's semiparametric hazard model (Cox 1972) is a standard tool for modeling survival data. The data is given in form of triplets $(T_j, \delta_j, \mathbf{Z}_j), j = 1, ..., n$, allowing for censoring. Here, the observation time $T_j = \min(X_j, C_j)$ of individual j takes the minimum value of the survival time X_j or the subject specific censoring time C_j. The indicator defined as

$$\delta_j = I(X_j \leq C_j) = \begin{cases} 1 & \text{event observed for subject j} \\ 0 & \text{censored} \end{cases}$$

denotes if the event of interest, death for instance, has been observed or if individual j is censored at time C_j. $\mathbf{Z}_j(t) \in \mathbb{R}^p$ is the vector of covariates for the j-th individual which may depend on time. Under the semiparametric hazard model the hazard function $\lambda(t|\mathbf{Z}(t))$ has the form

$$\lambda(t \,|\, \mathbf{Z}(t)) = \lambda_0(t) \exp[\boldsymbol{\beta}' \mathbf{Z}(t)], \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of unknown regression coefficients and the baseline hazard $\lambda_0(t)$ is an arbitrary function of time. The original model of Cox (1972) excluded time dependency of the covariate, i.e. $\mathbf{Z}(t) = \mathbf{Z}$. In this case the proportional hazards assumption has to hold, i.e. the hazard ratio for two individuals with covariate vectors $\mathbf{Z}$ and $\mathbf{Z}^*$, respectively

$$\frac{\lambda(t \,|\, \mathbf{Z})}{\lambda(t \,|\, \mathbf{Z}^*)} = \frac{\lambda_0(t) \exp[\boldsymbol{\beta}' \mathbf{Z}]}{\lambda_0(t) \exp[\boldsymbol{\beta}' \mathbf{Z}^*]} = \exp[\sum_{k=1}^{p} \beta_k (\mathbf{Z}_k - \mathbf{Z}_k^*)] \tag{2.2}$$

is independent of time. For time-varying covariates $\mathbf{Z}(t)$ this ratio is not independent of time, but for any two given values of a covariate the relative hazard in (2.2) is still determined by a time independent coefficient $\boldsymbol{\beta}$. Parameter estimation of $\boldsymbol{\beta}$ is done using Cox's partial likelihood (Cox 1975), a method which allows estimation without knowing the baseline hazard. Estimation of the cumulative hazard function $\Lambda_0(t) := \int_0^t \lambda_0(s) ds$ is achieved by Breslow's estimator (Breslow 1974). For this let $t_1 < t_2 < ... < t_D$ be the observed death times and the Breslow estimator is now given by

$$\hat{\Lambda}_0(t) = \sum_{t_i \leq t} \frac{d_i}{\sum_{j \in R(t_i)} \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j]}, \tag{2.3}$$

where d_i is the number of events at time t_i and $R(t_i)$ is the risk set at time t_i, i.e. the set of subjects that is still under study at time just prior to t_i.

3 Data Analysis for Compulsory Long Term Care Insurance

The data was recorded between April 1, 1995 and December 31, 1998. In 1995 the German government introduced compulsory long term care (LTC) insurance. This required part of the German welfare system paid benefits for home care since April 1, 1995. Starting July 1, 1996, the benefits were extended to care in a nursing home as well. For 5042 claimants, 3175 female and 1867 male, information about age, gender, severity and type of care (at home or in a nursing home) are available. There are three different levels of severity, which are roughly defined as follows:

- **Level 1**: considerable need of long-term care
- **Level 2**: severe need of long-term care
- **Level 3**: extreme need of long-term care

For further details on the exact definitions of these levels of severity see Czado and Rudolph (2002). In addition, the diagnoses which led to LTC are known. Table 1 contains a short description of the covariates considered in the model. The care status may change over time. If a change occurs at time t we refer to this time as a event time. Transitions between care levels as well as transitions between type of care are possible.

One aim of this paper is to investigate the effects of the diagnoses which lead to LTC on the hazard function. It is important to note that only 45.5 % of the claimants are recorded with a single diagnosis. The occurrence of multiple diagnoses, mainly double and triple diagnoses, is very common (see Table 2). This fact has to be taken into account in the modeling. There are 11 different diagnoses recorded in the data set, the seven main diagnoses are listed in Table 3 together with the percentage of a single diagnosis. The occurrence of all combinations of double diagnoses and the three main groups of triple diagnoses can be found in Tables 4 and 5, respectively.

Covariate	Description	Values
$Z_{Age}(t)$	age of claimant when a	
	state transition occurs at event time t	0 - 108 years
Z_{Sex}	gender	1 = female, 0 = male
$Z_{nh}(t)$	nursing home care indicator	1 = care in a nursing home,
	at event time t	0 = care at home at event time t
$Z_{Level2}(t)$	indicator for Level 2 at event time t	1 = care at level 2 at event time t, 0 = otherwise
$Z_{Level3}(t)$	indicator for Level 3 at event time t	1 = care at level 3 at event time t, 0 = otherwise
$Z_{Diagnosis\,i}$	diagnosis which led to LTC	1 = diagnosis i, 0 = otherwise

Table 1: Description of available covariates in the LTC data set

Number of diagnoses	1	2	3	4	5	6
Number of claimants	2296	1830	708	178	27	3
Percentage	45.55	36.30	14.04	3.53	0.54	0.05

Table 2: Number of diagnoses causing LTC and their relative frequency

Diagnosis	Number of claimants with (multiple diagnoses included)	Number of single diagnosis	Percentage
Tumor	694	276*	39.8
Psychosis	1254	394*	31.4
Heart attack	1922	378*	19.7
Stroke	1044	309*	29.6
Arthritis	534	85*	15.9
Lung disease	93	16*	12.9
Dementia	2151	587*	27.3
Bone disease	1015	189*	18.6
Others	226	66	29.2

Table 3: Frequency of diagnoses (multiple diagnoses included) and percentage of single diagnoses (* indicates diagnosis later considered in the analysis)

3.1 Analysis of the Survival of LTC Claimants

We used a Cox semiparametric hazard model, where possible covariates are all diagnoses as well as the remaining covariates listed in Table 1. Significant covariates are filtered out by

	Psychosis	Heart	Stroke	Arthritis	Lung	Dementia	Bone disease	Others
Tumor	52*	65*	46	7	4	54*	36	4
Psychosis		109*	84*	22	3	141*	45	23
Heart			105*	72*	12	351*	128	17
Stroke				8	1	135*	28	10
Arthritis					1	66*	31	5
Lung						6	4	0
Dementia							134	15
Bone disease								5

Table 4: Frequency of combinations of double diagnoses (* indicates diagnosis combination later considered in the analysis)

Diagnoses	Frequency
Psychosis, Heart attack and Dementia	74*
Heart attack, Stroke and Dementia	56*
Heart attack, Arthritis and Dementia	54*

Table 5: Frequency of the most important combinations of triple diagnoses (* indicates diagnosis combination later considered in the analysis)

using partial log-likelihood ratio tests and Akaike's information criterion (AIC) (Akaike 1973). Interactions are considered as well. Details of the model selection are given in Gschlößl (2002) (pp. 54-63). As final model for the hazard rate the following was chosen:

$$
\begin{aligned}
\lambda(t) \;=\; & \lambda_0(t)\exp[\beta_1 \mathbf{Z}_{Age}(t) + \beta_2 \mathbf{Z}_{Sex} + \beta_3 \mathbf{Z}_{nh}(t) + \beta_4 \mathbf{Z}_{Level2}(t) + \beta_5 \mathbf{Z}_{Level3}(t) \\
+ & \beta_6 \mathbf{Z}_{Sex} \times \mathbf{Z}_{nh}(t) + \beta_7 \mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Level2}(t) + \beta_8 \mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Level3}(t) \\
+ & \beta_9 \mathbf{Z}_{Dementia} + \beta_{10} \mathbf{Z}_{Stroke} + \beta_{11} \mathbf{Z}_{Psychosis} + \beta_{12} \mathbf{Z}_{Tumor} + \beta_{13} \mathbf{Z}_{Heart} + \beta_{14} \mathbf{Z}_{Lung} \\
+ & \beta_{15} \mathbf{Z}_{Arthritis} + \beta_{16} \mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Tumor} + \beta_{17} \mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Arthritis} \\
+ & \beta_{18} \mathbf{Z}_{Sex} \times \mathbf{Z}_{Tumor} + \beta_{19} \mathbf{Z}_{Sex} \times \mathbf{Z}_{Psychosis} + \beta_{20} \mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Psychosis} \\
+ & \beta_{21} \mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Stroke} + \beta_{22} \mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Lung} + \beta_{23} \mathbf{Z}_{Tumor} \times \mathbf{Z}_{Heart} \\
+ & \beta_{24} \mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Heart} + \beta_{25} \mathbf{Z}_{Level2}(t) \times \mathbf{Z}_{Heart} + \beta_{26} \mathbf{Z}_{Level3}(t) \times \mathbf{Z}_{Heart} \qquad (3.1)
\end{aligned}
$$

$$+ \quad \beta_{27}\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Level2}(t) + \beta_{28}\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Level3}(t) + \beta_{29}\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Psychosis}$$

$$+ \quad \beta_{30}\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Dementia} + \beta_{31}\mathbf{Z}_{Psychosis} \times \mathbf{Z}_{Stroke} + \beta_{32}\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Heart}].$$

Since numerous interactions are present in Model (3.1), the interpretation is not an easy task. Therefore, the multipliers $\exp[\boldsymbol{\beta}'\mathbf{Z}]$ are plotted for several groups of claimants in Figure 1. A higher care level results in a higher risk of mortality, whereas women seem to have a lower risk to die. In most of the groups claimants with care level 1 and 2 in nursing homes (thin lines) have a lower life expectancy than claimants who receive care at home. For claimants with care level 3 however, the type of care doesn't play a very decisive role. Women with care level 3 even have a lower mortality risk when living in a nursing home (except for women with lung diseases). Note, that tumors clearly reduce the expected lifetime.

4 Assessing the Model Adequacy

We now want to assess the fit of Model (3.1). There are two assumptions to check. The functional form of continuous covariates and the proportional hazards assumption.

4.1 Functional Form

Under the Cox semiparametric hazard model continuous covariates are linear in the log-hazard. Otherwise adequate transformations have to be found. We check this assumption using martingale residuals. The martingale residual for the j-th individual is defined as

$$\hat{M}_j(t) = N_j(t) - \int_0^t Y_j(s) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}_j(s)] d\hat{\Lambda}_0(s), \quad j = 1, .., n, \tag{4.2}$$

where $N_j(t)$ is a counting process denoting the number of events up to time t and $Y_j(t) = I\{T_j \geq t\}$ indicates whether individual j is still under study at time t. A smoothed plot of the martingale residuals (see Therneau, Grambsch, and Fleming (1990)) against the variable of interest should be a straight line, otherwise the plot indicates the correct shape of the covariate. In Model (3.1), age is the only continuous covariate included. We have plotted martingale residuals of age for a variety of groups with single and double diagnoses (plots are not shown here, but contained in Gschlößl (2002)). Most of the plots clearly reveal a nonlinear functional form of age and an obvious difference in the shape of single diagnoses and the same diagnoses in combination with another one is observed. Just men with psychosis, psychosis and heart attack or psychosis and

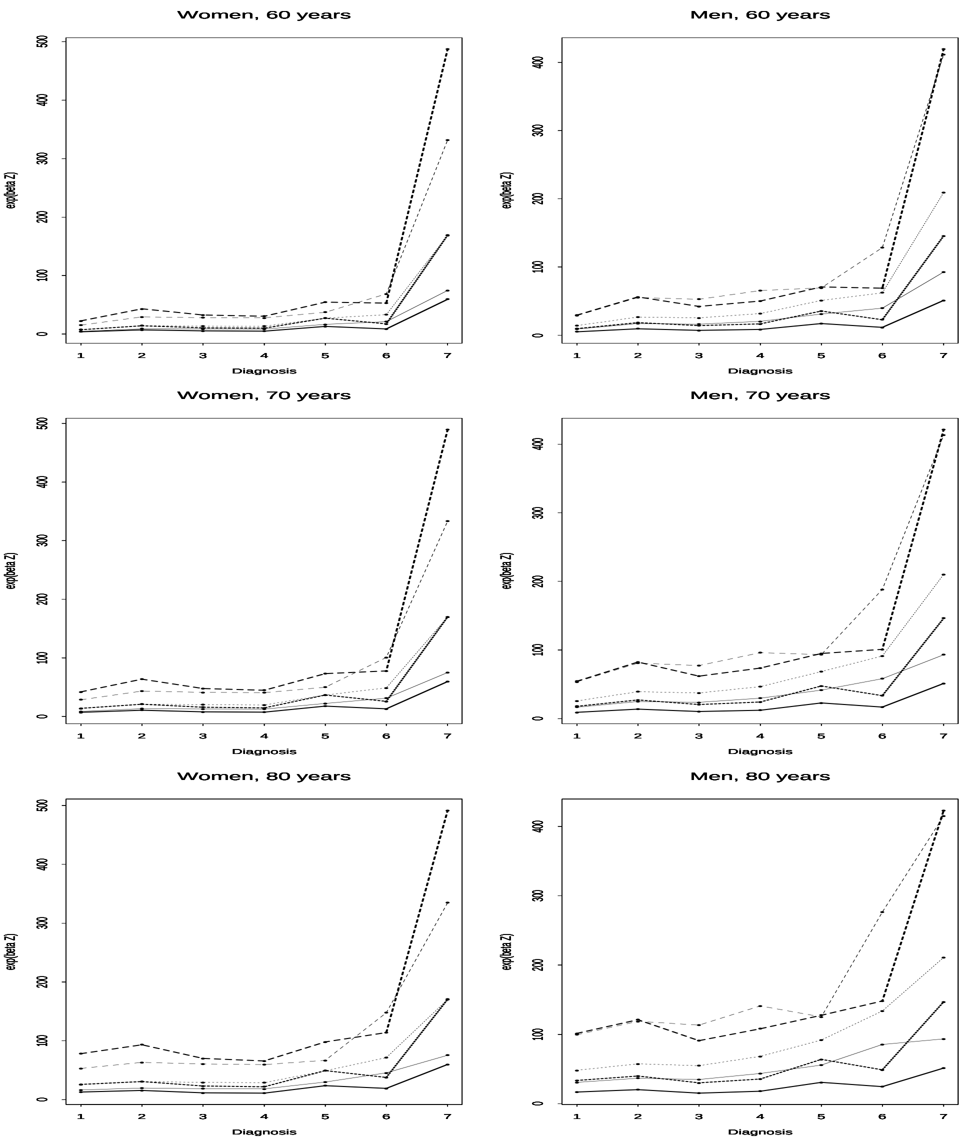

Figure 1: Estimated multipliers $\exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}]$ for women and men with single diagnoses in Model (3.1) (Legend is given in Table 6)

x-axis	Diagnosis	Line types	Care Level	Type of Care
1	Arthritis	—	Level 1	Nursing home
2	Dementia	—	Level 1	Home
3	Stroke	- - -	Level 2	Nursing home
4	Psychosis	- - -	Level 2	Home
5	Heart	– –	Level 3	Nursing home
6	Lung	– –	Level 3	Home
7	Tumor			

Table 6: Legend to Figure 1

dementia may be summarized to one group of claimants due to their similar functional form. In the following this group will be referred to as

$$\mathbf{Z}_{PHD} = \begin{cases} 1 & \text{member of this group} \\ 0 & \text{otherwise} \end{cases} .$$

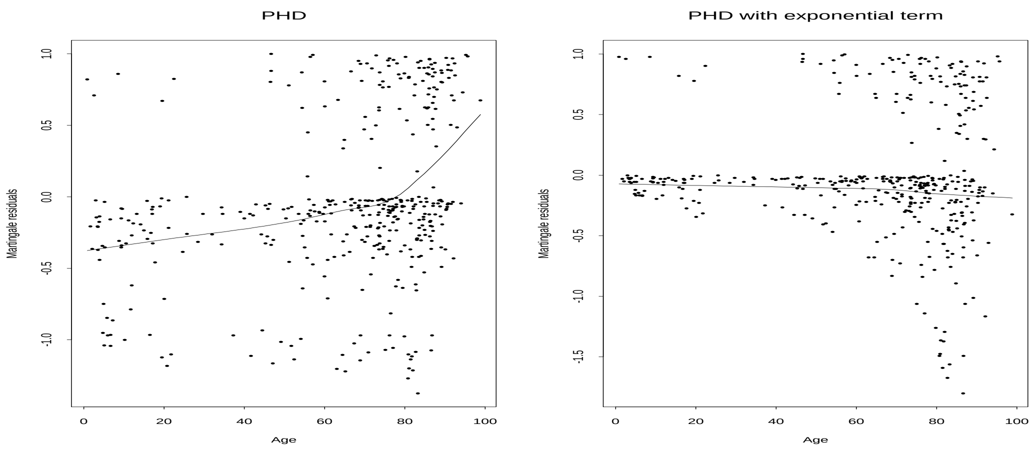

Figure 2: Martingale residuals for men with psychosis, heart attack and dementia in model (3.1) and with exponential term (4.3)

For all the other groups, age should be modeled separately. The martingale residuals for the group PHD plotted in the left panel of Figure 2 indicate that an exponential function might be appropriate, whereas the correct functional form for women with tumor (see left panel of Figure 3) might be given by a quadratic function. Fractional polynomials (see Royston and Altman (1994)) which include quadratic shapes can be used here. We will check an exponential fit first. Therefore the interaction

$$\exp(c \cdot \mathbf{Z}_{Age}(t)) \times \mathbf{Z}_{Group} \tag{4.3}$$

is added to Model (3.1), where $\mathbf{Z}_{Group}$ is an indicator function for the considered group of claimants. Apart from Group PHD, we only consider the diagnoses groups indicated by a $*$ in Table 3, 4 and 5, this means, single, double and triple diagnoses groups are modeled separately. The remaining claimants are summarized in the group "others". The covariate $\mathbf{Z}_{Age}(t)$ in Model (3.1) is replaced by the interaction $\mathbf{Z}_{Age}(t) \times (1 - \mathbf{Z}_{Group})$ to guarantee that only the age of the group of interest is modeled in a nonlinear way. The constant c in (4.3) varies from 0.002 to 0.102, for groups with a concave shape the negative of these values is included as well. Since the optimal value of c is determined by maximizing the log-likelihood with respect to c, using a grid search for the optimum, we consider two degrees of freedom for the new interaction term (4.3)- one for the regression coefficient and one for c. A significant improvement of the fit can be achieved for women with tumor ($\hat{c} = -0.1020$) and the group PHD ($\hat{c} = 0.002$). Here, the partial log-likelihood test results in a p-value of $3.24 \cdot 10^{-3}$ and $1.7 \cdot 10^{-2}$, respectively.

In a similar way we use fractional polynomials to find adequate transformations for age. A fractional polynomial of degree m for a continuous covariate x is given by

$$\Phi_m(x,p) = \boldsymbol{\beta}_0 + \sum_{j=1}^{m} \beta_j x^{p_j},$$

where m is an integer, β_j are regression coefficients and $p_1 \leq ... \leq p_k$ are any real valued exponents. $p_j = 0$ corresponds to the logarithm of x, i.e. $x^0 = \ln(x)$. Fractional polynomials allow for a variety of different functional shapes and in most data sets fractional polynomials of degree one and two are sufficient.

Again, in Model (3.1) $\mathbf{Z}_{Age}(t)$ is replaced by $\mathbf{Z}_{Age}(t) \times (1 - \mathbf{Z}_{Group})$ and the interaction

$$\mathbf{Z}_{Age}^{p_1}(t) \times \mathbf{Z}_{Group}$$

for m=1 and

$$(\mathbf{Z}_{Age}^{p_1}(t) + \mathbf{Z}_{Age}^{p_2}(t)) \times \mathbf{Z}_{Group}$$

for m=2 are added to the model, respectively. As proposed by Royston and Altman (1994) we restrict the values for p_1 and p_2 to the set $\wp = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Again, we achieve significant results for the same two groups as before using exponential forms. For women with tumor age is modeled best by the fractional polynomial $\mathbf{Z}_{Age}^{-1}(t)$, for the group PHD the functional form $\mathbf{Z}_{Age}^{-2}(t) + \mathbf{Z}_{Age}(t)$ seems to be adequate. The p-values of the corresponding log-likelihood ratio tests compared to a linear modeling $\mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Group}$ are 0.011 and 0.07, respectively. Again, these are based on two degrees of freedom. Comparing the fractional polynomial approach to the exponential approach we get similar values of the log-likelihood for the goup PHD. Since the exponential approach uses two degrees of freedom less, this one seems to be more appropriate. To check the achieved improvement of the fit, we plot again the martingale residuals versus age in Model (3.1) containing the exponential term for group PHD in the right panel of Figure 2. The plot is now linear which indicates that we have found an appropriate transformation. For women with tumor both approaches led to similar results as well (see Gschlößl (2002), p.80-81). A further look of the corresponding martingale plot in the left panel of Figure 3 shows, that the plot is almost a straight line up from 40 years. The functional form is mainly determined by a few observations up to about fifteen years. Therefore, we build two separate groups, girls up to 15 years and women over 15 years with tumor using the following indicators

$$\mathbf{Z}_{Tumor,G} = \begin{cases} 1 & \text{if} \quad \text{female}, \leq 15 \text{ years, Tumor} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{Z}_{Tumor,W} = \begin{cases} 1 & \text{if} \quad \text{female}, > 15 \text{ years, Tumor} \\ 0 & \text{otherwise} \end{cases}$$

A separate martingale plot for both groups (see the middle and right panel of Figure 3) clearly shows, that there is no need of modeling age for women with tumor. For the girls we haven't got enough observations to make a statement. The interaction $\mathbf{Z}_{Age} \times \mathbf{Z}_{Tumor,W}$ is not significant, thus, our model now has the following hazard function

$$\begin{aligned} \lambda(t) &= \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}(\text{Model (3.1) without } \mathbf{Z}_{Age}(t))] \\ &\times \exp[\beta_1 \mathbf{Z}_{Age}(t) \times (1 - \mathbf{Z}_{Tumor,W}) \times (1 - \mathbf{Z}_{PHD}) \\ &+ \beta_{33} \mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Tumor,G} + \beta_{34} \exp(0.022 \cdot \mathbf{Z}_{Age}(t)) \times \mathbf{Z}_{PHD}] \quad . \end{aligned} \tag{4.4}$$

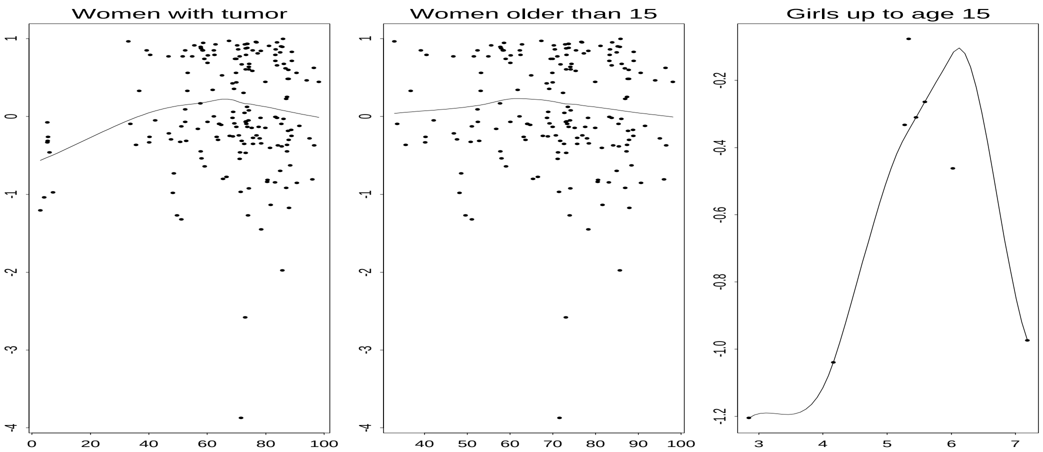

Figure 3: Martingale residuals for women with all ages and separated for women older than fifteen up to age 15 with tumor

4.2 Assessing the proportional hazards assumption

The second assumption to check is the proportional hazards assumption (2.2), i.e. if there is a need to allow for time dependent coefficients. A graphical check can be done by using scaled Schoenfeld residuals (see Grambsch and Therneau (1994) for further details). A plot of the scaled Schoenfeld residuals against time reveals the change of the coefficients with time, a constant therefore indicates no time dependency. In Figure 4 these plots are presented for several covariates in Model (3.1) and Model (4.4), respectively. For the time-varying covariates $\mathbf{Z}_{nh}(t)$, $\mathbf{Z}_{Level2}(t)$ and $\mathbf{Z}_{Level3}(t)$ constant values are assumed. A significant change over time can be recorded in almost all covariates. However, this dependency seems to be present mainly during the first 900 days of LTC, which is indicated through the vertical line. Afterwards most of the plots are almost constant. We therefore split our data in observations up to 900 days in LTC and observations longer than 900 days in LTC and fit two separate models. Again the functional form is allowed to include exponential and fractional polynomial terms. However when the data set is split these give no significant improvement and as a result we get the following models

$$\begin{aligned}
\lambda(t) &= \lambda_0(t)\exp[\beta_1\mathbf{Z}_{Age}(t)\times\mathbf{Z}_{Tumor,G}+\beta_2\mathbf{Z}_{Age}(t)\times(1-\mathbf{Z}_{Tumor,W})+\beta_3\mathbf{Z}_{Sex}\\
&+ \beta_4\mathbf{Z}_{nh}(t)+\beta_5\mathbf{Z}_{Level2}(t)+\beta_6\mathbf{Z}_{Level3}(t)+\beta_7\mathbf{Z}_{Tumor}+\beta_8\mathbf{Z}_{Dementia}+\beta_9\mathbf{Z}_{Heart}\\
&+ \beta_{10}\mathbf{Z}_{Psychosis}+\beta_{11}\mathbf{Z}_{Stroke}+\beta_{12}\mathbf{Z}_{Arthritis}+\beta_{13}\mathbf{Z}_{Lung}\\
&+ \beta_{14}\mathbf{Z}_{Age}(t)\times(1-\mathbf{Z}_{Tumor,W})\times\mathbf{Z}_{Tumor}+\beta_{15}\mathbf{Z}_{Age}(t)\times(1-\mathbf{Z}_{Tumor,W})\times\mathbf{Z}_{Arthritis}\\
&+ \beta_{16}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Stroke}+\beta_{17}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Level2}(t)+\beta_{18}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Level3}(t)\\
&+ \beta_{19}\mathbf{Z}_{Sex}\times\mathbf{Z}_{Tumor}+\beta_{20}\mathbf{Z}_{Level2}(t)\times\mathbf{Z}_{Heart}+\beta_{21}\mathbf{Z}_{Level3}(t)\times\mathbf{Z}_{Heart}\\
&+ \beta_{22}\mathbf{Z}_{Psychosis}\times\mathbf{Z}_{Stroke}+\beta_{23}\mathbf{Z}_{Tumor}\times\mathbf{Z}_{Heart}+\beta_{24}\mathbf{Z}_{Stroke}\times\mathbf{Z}_{Arthritis}]
\end{aligned}\tag{4.5}$$

for LTC durations up to 900 days and

$$\begin{aligned}
\lambda(t) &= \lambda_0(t)\exp[\beta_1\mathbf{Z}_{Age}(t)\times(1-\mathbf{Z}_{Tumor,W})+\beta_2\mathbf{Z}_{Sex}+\beta_3\mathbf{Z}_{nh}(t)+\beta_4\mathbf{Z}_{Level2}(t)\\
&+ \beta_5\mathbf{Z}_{Level3}(t)+\beta_6\mathbf{Z}_{Tumor}+\beta_7\mathbf{Z}_{Heart}+\beta_8\mathbf{Z}_{Psychosis}+\beta_9\mathbf{Z}_{Stroke}\\
&+ \beta_{10}\log(\mathbf{Z}_{Age}(t))\times\mathbf{Z}_{Tumor,W}+\beta_{11}\mathbf{Z}_{Age}(t)\times(1-\mathbf{Z}_{Tumor,W})\times\mathbf{Z}_{Level2}\\
&+ \beta_{12}\mathbf{Z}_{Age}(t)\times(1-\mathbf{Z}_{Tumor,W})\times\mathbf{Z}_{Level3}(t)+\beta_{13}\mathbf{Z}_{Sex}\times\mathbf{Z}_{nh}(t)\\
&+ \beta_{14}\mathbf{Z}_{Sex}\times\mathbf{Z}_{Psychosis}+\beta_{15}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Level2}(t)+\beta_{16}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Level3}(t)\\
&+ \beta_{17}\mathbf{Z}_{Psychosis}\times\mathbf{Z}_{Stroke}+\beta_{18}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Psychosis}+\beta_{19}\mathbf{Z}_{Level2}(t)\times\mathbf{Z}_{Heart}\\
&+ \beta_{20}\mathbf{Z}_{Level3}(t)\times\mathbf{Z}_{Heart}]
\end{aligned}\tag{4.6}$$

for LTC durations longer than 900 days. A test for time dependency, implemented in the Splus-routine zph, for Model (4.6) results in a single p-value greater than 0.08 for all covariates.

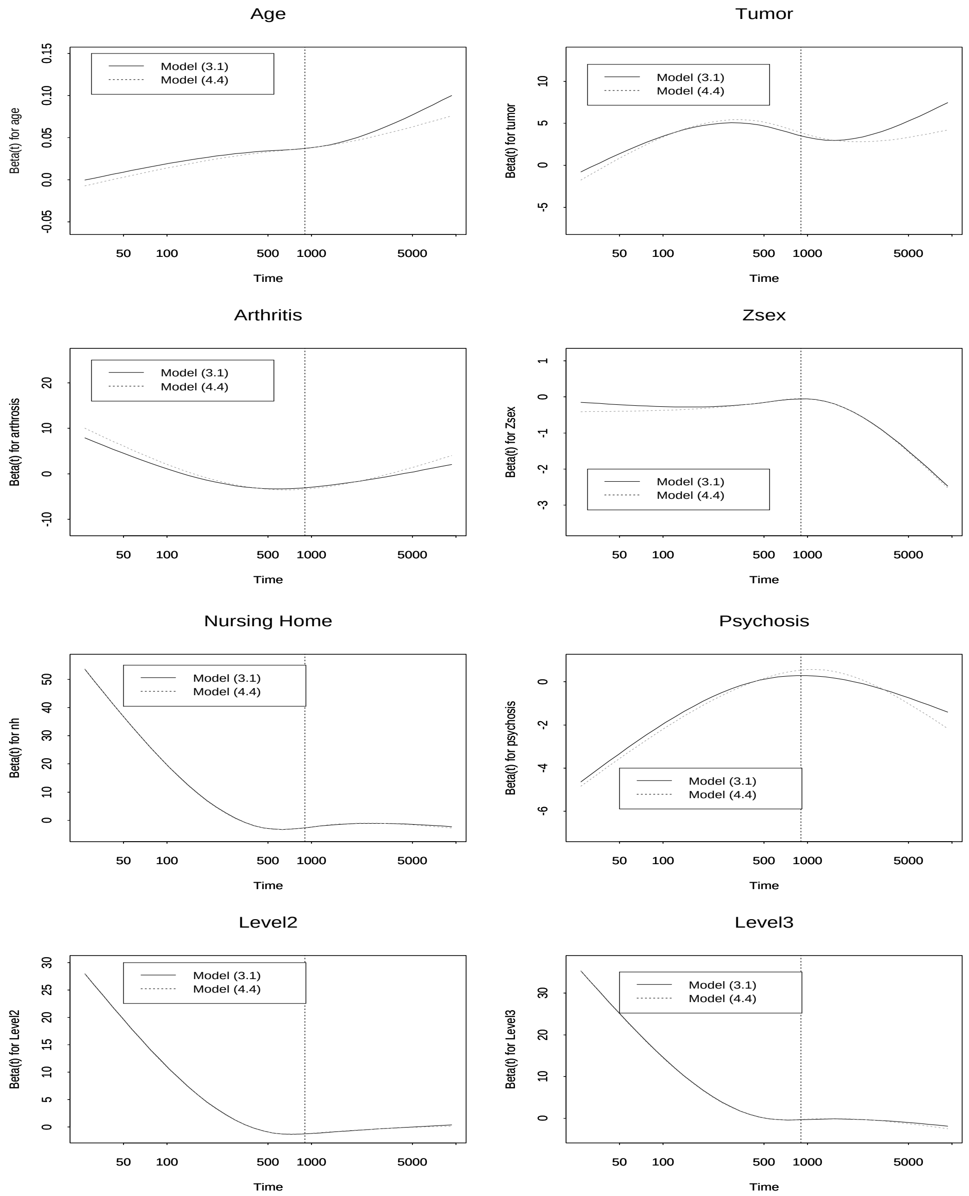

Figure 4: Scaled Schoenfeld residuals against time for Models (3.1) and (4.4)

Therefore, there are no more significant time dependent covariates in the model. For the Model (4.5), the proportional hazards assumption holds except for $\mathbf{Z}_{Level2}$ (p-value: 0.0043), $\mathbf{Z}_{Level3}$ (p-value: 0.0026) and $\mathbf{Z}_{Dementia}$ (p-value: 0.0016). Thus, the Models (4.5) and (4.6), based on the split data, led to a significant improvement in comparison to Model (4.4).

Up to now only transition intensities to death have been considered. In a similar way a semi-parametric hazard model can be used to model transitions between care at home and nursing home and transitions between the different levels of care. These transitions are illustrated in the multi-state models given in Figures 5 and 6. The results of the modeling of these transitions won't be shown in this paper, for details see Gschlößl (2002), pp.93-97.

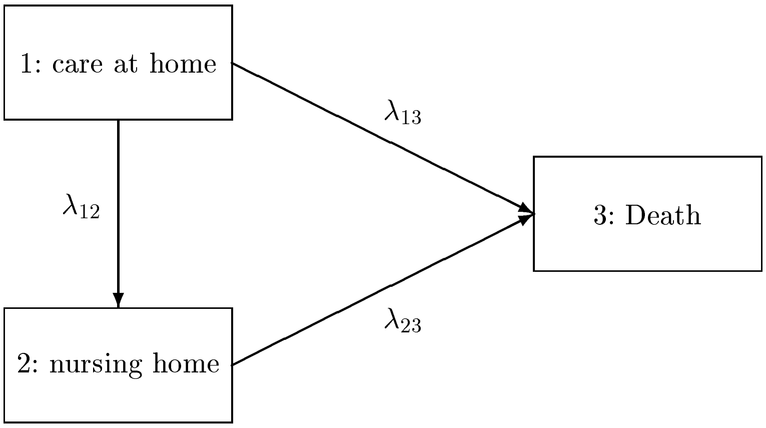

Figure 5: State transitions between type of care and transitions to death

Note, that in Figure 5 besides transitions to death only transitions from care at home to nursing home are taken into account while in Figure 6 only transitions to higher care levels are considered. Since in our data transitions from a higher care status to a lower care status are very rare events, the remaining transitions intensities are taken to be zero.

5 Estimation of transition probabilities

Having modeled transition intensities, we now want to estimate transition probabilities. Based on these probabilities insurance companies calculate their rates. In particular, we consider the model illustrated in Figure 6 here. In the multi-state model presented in Figure 5 probabilities can be estimated analogous. Estimates of one-year transition probabilities from state i to state j in dependency of Age x, Sex s, type of care c (where c = nh or c = cah for care at home), group k and LTC duration d of the claimant can be assessed by

$$\begin{aligned}\hat{p}_{ij}(x,s,d,c,k) = \quad & \sum_{d \leq t_k < d+1} \Big\{ \hat{\lambda}_{ij}(t_k,\, \mathbf{Z}_{Age}(t_k) = x, \mathbf{Z}_{Sex} = s, \mathbf{Z}_{nh} = c, \mathbf{Z}_{Group} = k) \\ & \prod_{d \leq t_l < t_k} \quad (1 - \hat{\lambda}_{ij}(t_l,\, \mathbf{Z}_{Age}(t_l) = x, \mathbf{Z}_{Sex} = s, \mathbf{Z}_{nh} = c, \mathbf{Z}_{Group} = k)) \Big\}. \end{aligned} \quad (5.7)$$

Here, k takes the values 1,...,22, denoting the groups of diagnoses indicated by a $*$ in Tables 3, 4 und 5. This formula is based on the Kaplan-Meier-estimator $\hat{S}(t) = \prod_{t_j \leq t}(1 - \hat{\lambda}(t_j))$ for the survival function $S(t) = P(T > t)$. We obtain the estimated hazard functions $\hat{\lambda}_{ij}(t, \mathbf{Z})$ using

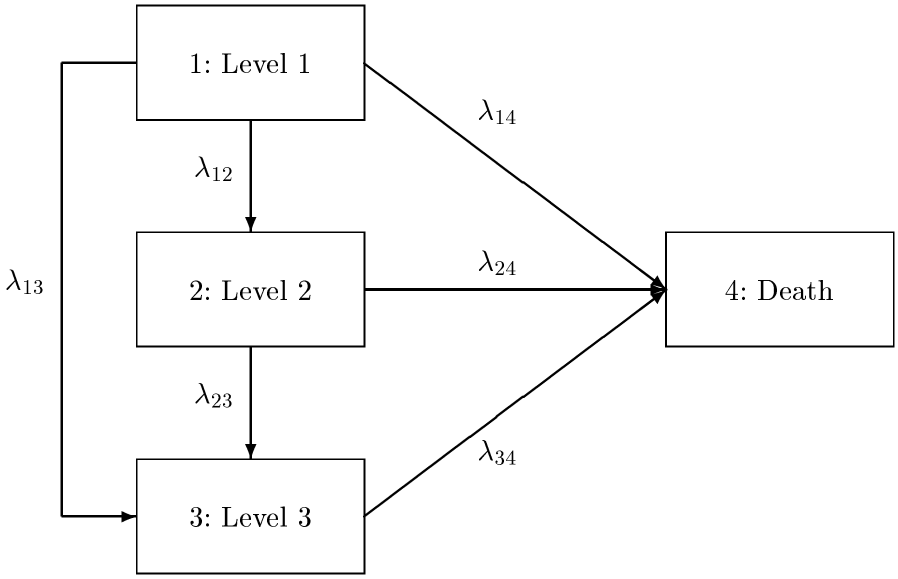

Figure 6: State transitions between levels and transitions to death

Breslow's estimator (2.3). Since we have included the diagnoses and type of care (at home or in nursing homes) in our model we also get probabilities in dependency of these covariates which insurance companies ignore in their calculations. To eliminate this dependency we calculate a weighted mean over the considered groups of diagnoses as follows

$$\begin{aligned} \hat{p}_{i,j}(x,s,d) &= \frac{1}{n^i_{x,s,d,cah} + n^i_{x,s,d,nh}} \Big\{ \sum_{k=1}^{22} [n^i_{x,s,d,cah,k} \hat{p}_{i,j}(x,s,d,\text{cah, k}) \\ &+ n^i_{x,s,d,nh,k} \hat{p}_{i,j}(x,s,d,\text{nh, k})] \Big\}, \end{aligned} \tag{5.8}$$

where $n^i_{x,s,d,cah}$ and $n^i_{x,s,d,nh}$ give the number of observations in state i in Figure 6 with age between x-5 and x+5 years, sex s, duration d, who receive care at home (cah) or in nursing homes (nh), respectively. Here, $n^i_{x,s,d,cah,k}$ and $n^i_{x,s,d,nh,k}, k = 1, ..., 22$ denote the corresponding total numbers separated by the 22 groups of diagnoses. We now concentrate on transitions to state 4, corresponding to death. We estimate these probabilities for Model (4.5) and Model (4.6) as well as for the best model without diagnoses

$$\lambda(t) = \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{Age}(t) + \beta_2 \mathbf{Z}_{Sex} + \beta_3 \mathbf{Z}_{nh}(t) + \beta_4 \mathbf{Z}_{Level2}(t) + \beta_5 \mathbf{Z}_{Level3}(t)$$

$$
\begin{aligned}
&+ \quad \beta_6 \mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Sex} + \beta_7 \mathbf{Z}_{Age}(t) \times \mathbf{Z}_{nh}(t) + \beta_8 \mathbf{Z}_{Sex} \times \mathbf{Z}_{nh}(t) \\
&+ \quad \beta_9 \mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Level2}(t) + \beta_{10} \mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Level3}(t)]
\end{aligned} \tag{5.9}
$$

(see Rudolph (2000), p.78). Here again, we consider a weighted mean over the probabilities similar to (5.8), but this time only averaged over the type of care. In addition, we determine the empirical mortality probabilities given by $\hat{p}^e_{i,4}(x, s, d) = \frac{p}{n}$, where p is the number of deaths and n is the number of observations with age between x-5 and x+5 years, sex s and LTC duration d. A graphical comparison of the mortality probabilities in the first year of LTC is given in Figure 7. It is obvious that the probabilities based on Model (4.5) including diagnoses are closer to the empirical ones. See for example the groups of age 55 to 64 years (x-axis = 1) or women in Level 3 (y-axis = 6). For the second and third year of LTC the difference between the models decreases (see corresponding plots in Gschlößl (2002), pp.102-106). Therefore, the diagnoses seem to have an important influence on survival particularly during the first year. In Model (4.5) there are 7 diagnoses included, whereas Model (4.6) for longer durations than 900 days only contains four significant diagnoses. In addition, the coefficients of the diagnoses, that are not presented here, take higher values during the first 900 days, i.e. have a bigger impact on survival. In Table 8, for the first three years of LTC three different measures of deviances between the empirical mortality probabilities and estimated probabilities based on the models with and without diagnoses are given. For each year we consider a weighted sum $\sum_{i=1}^{24} n_i |\hat{p}_{i4} - \hat{p}^e_{i4}|$ of absolute values, a weighted sum of squares $\sum_{i=1}^{24} n_i (\hat{p}_{i4} - \hat{p}^e_{i4})^2$ and a weighted sum of log odds $\sum_{i=1}^{24} n_i (\log(\frac{\hat{p}_{i4}}{1-\hat{p}_{i4}}) - \log(\frac{\hat{p}^e_{i4}}{1-\hat{p}^e_{i4}}))^2$. The sum is always taken over the 24 groups of claimants, divided by gender, age (4 groups) and care level, which are plotted in Figure 7. Here, $\hat{p}_{i4}$ and $\hat{p}^e_{i4}$ denote the estimated and empirical mortality probabilities for claimant group i, n_i is the number of observations in the corresponding group. Thus, the reliability of the empirical probabilities is taken into account. All of the three measures of deviances lead to qualitatively similar results. In the first year of LTC the probabilities based on the Models (4.4) and (4.5) including diagnoses are clearly closer to the empirical ones than those based on Model (5.9) without diagnoses. The same is observed in the second year, although the difference between the models is decreasing. Further, we note that the splitting of the data additionally improved the resulting estimated probabilities. In the third year of LTC Model (4.4) based on the whole data still shows a smaller error than Model (5.9) without diagnoses. However, the split Models (4.5) and (4.6) give the worst results here. Since the data have been split after 900 days, for this year the probabilities are based on both of the models. The sudden change in the underlying model could be a reason for the observed aggravation.

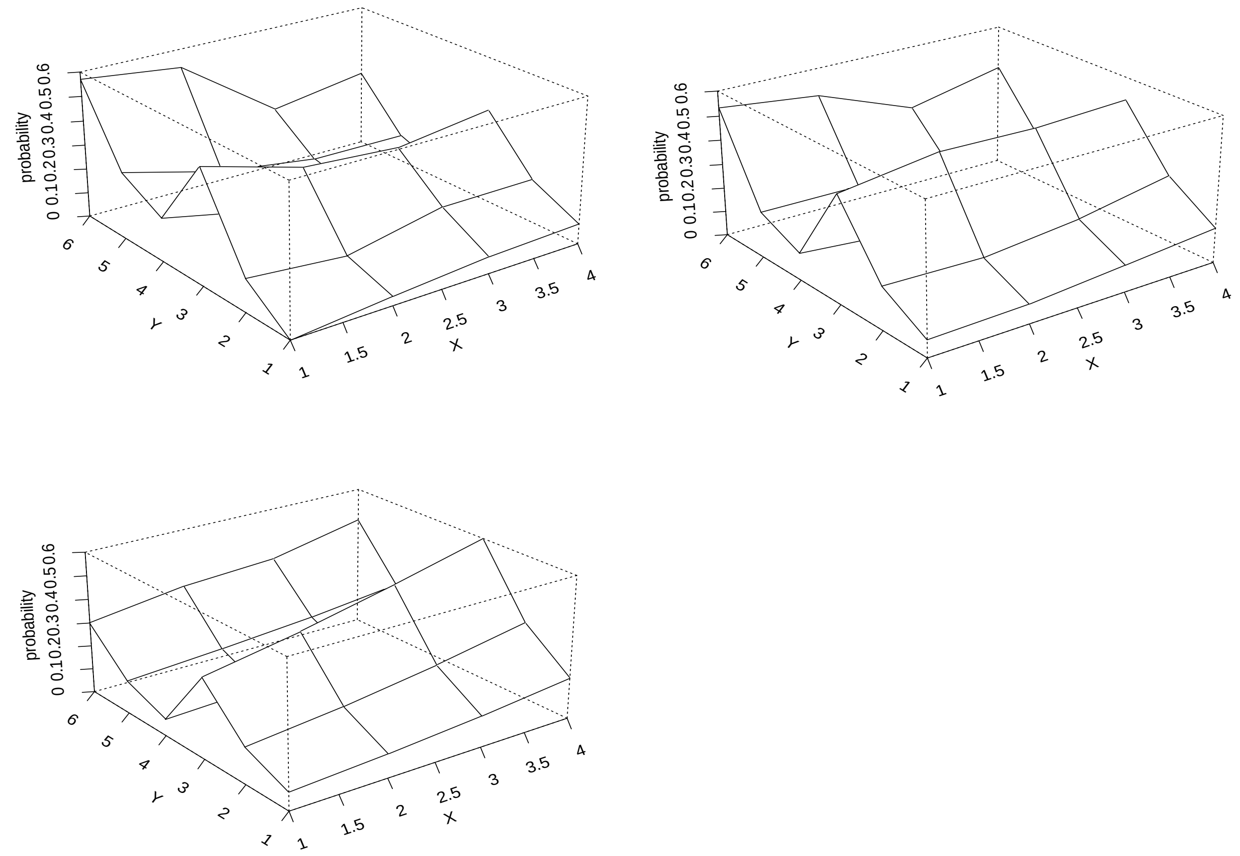

Figure 7: Mortality probabilities during the first year of LTC (Legend see Table 7)

top left	empirical probabilities $\hat{p}^e_{i4}(x, s, d = 1), i = 1, 2, 3$
top right	$\hat{p}_{i4}(x, s, d = 1)$ based on Model (4.5)
bottom left	$\hat{p}_{i4}(x, s, d = 1)$ based on Model (5.9)

x-axis		y-axis	
1	55-64 years	1	Male, Level 1
2	65-74 years	2	Male, Level 2
3	75-84 years	3	Male, Level 3
4	85-94 years	4	Female, Level 1
		5	Female, Level 2
		6	Female, Level 3

Table 7: Legend to Figure 7

	Model (5.9)	Models (4.5) and (4.6)	Model (4.4)
$\sum_{i=1}^{24} n_i \lvert \hat{p}_{i4} - \hat{p}^e_{i4} \rvert$			
first year of LTC	301.018	198.942	222.869
second year of LTC	208.797	170.519	177.453
third year of LTC	135.589	154.542	129.524
$\sum_{i=1}^{24} n_i (\hat{p}_{i4} - \hat{p}^e_{i4})^2$			
first year of LTC	35.43	16.69	19.89
second year of LTC	21.28	16.05	16.14
third year of LTC	13.30	18.87	12.34
$\sum_{i=1}^{24} n_i (\log(\frac{\hat{p}_{i4}}{1-\hat{p}_{i4}}) - \log(\frac{\hat{p}^e_{i4}}{1-\hat{p}^e_{i4}}))^2$			
first year of LTC	3162.51	2110.49	2317.91
second year of LTC	1038.63	730.35	779.46
third year of LTC	651.03	687.55	554.75

Table 8: Several measures of deviances between the empirical mortality probabilities $\hat{p}^e_{i4}$ and the estimated mortality probabilities $\hat{p}_{i4}$ based on Model (5.9) without diagnoses and Models (4.5), (4.6) and (4.4) including diagnoses

6 Inception Selection effect of diagnosis on LTC premiums

Using the estimated one-year transition probabilities now premiums can be calculated. We will base our calculations on the multi-state model shown in Figure 5, extended by the additional state "healthy" since transition probabilities from state "healthy" to the remaining states enter into the calculations as well. This model is illustrated in Figure 8. Since in the data only information about the transitions in the dotted box is given, we have to use external sources for the transitions out of state 1, further details are given below.

We have chosen this model here for its simpler structure in comparison to the model including transitions between care levels. Similar the model presented in Figure 6 could be used only requiring higher computational efforts. In addition we take the estimated transition intensities from care at home to nursing home to be more reliable than the estimated transition intensities between care levels because those are based only on about half as much data.

To model the insured risk we define a time continuous Markov Process $S : T \times \Omega \to \{1, .., n\}$ with finite states $1, .., n$, interval $T = [0, \tau)$ from policy begin to policy end and underlying

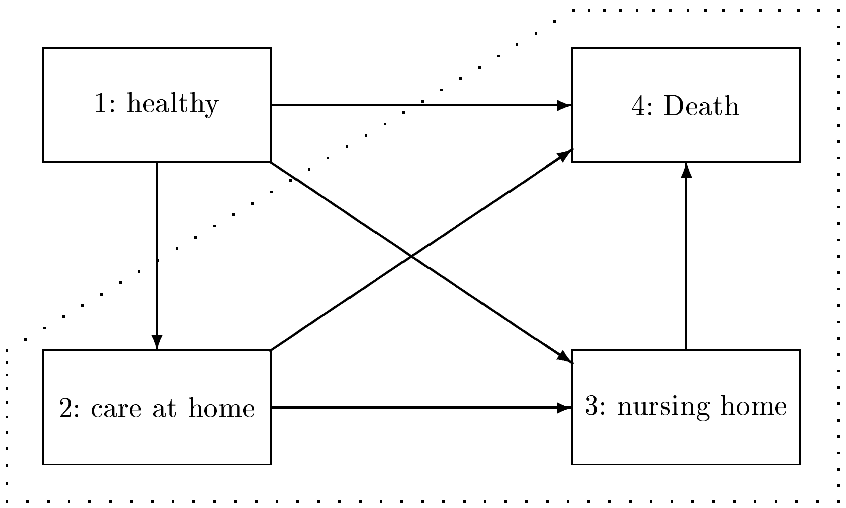

Figure 8: State transitions in LTC insurance

probability space Ω. Consider the following transition probabilities

$$P_{ij}(z,t) := P(S(t) = j | S(z) = i) \quad \forall z \leq t,\, z,t \in T, \quad i,j \in \{1,..,n\}.$$

The corresponding transition intensities are defined by

$$\lambda_{ij}(t) = \lim_{dt \to 0} \frac{P_{ij}(t, t+dt)}{dt}$$

and

$$\lambda_i(t) := \sum_{j \neq i} \lambda_{ij}(t).$$

Then the forward Kolmogorov-differential equations (e.g. Karlin and Taylor (1981))

$$\frac{d}{dt} P_{ij}(z,t) = \sum_{k: k \neq j} P_{ik}(z,t) \lambda_{kj}(t) - P_{ij}(z,t) \lambda_j(t) \tag{6.10}$$

give a relation between transition intensities and probabilities. In particular for the multi-state model presented in Figure 8 these differential equations are given by

$$\begin{aligned}
\frac{d}{dt} P_{11}(z,t) &= -P_{11}(z,t)(\lambda_{12}(t) + \lambda_{13}(t) + \lambda_{14}(t)), \\
\frac{d}{dt} P_{12}(z,t) &= P_{11}(z,t)\lambda_{12}(t) - P_{12}(z,t)(\lambda_{23}(t) + \lambda_{24}(t)), \\
\frac{d}{dt} P_{13}(z,t) &= P_{11}(z,t)\lambda_{13}(t) + P_{12}(z,t)\lambda_{23}(t) - P_{13}(z,t)\lambda_{34}(t), \\
\frac{d}{dt} P_{14}(z,t) &= P_{11}(z,t)\lambda_{14}(t) + P_{12}(z,t)\lambda_{24}(t) + P_{13}(z,t)\lambda_{34}(t)
\end{aligned} \tag{6.11}$$

with solution

$$
\begin{aligned}
P_{11}(z,t) &= \exp\Big(-\int_z^t [\lambda_{12}(u)+\lambda_{13}(u)+\lambda_{14}(u)]du\Big), \quad P_{22}(z,t)=\exp\Big(-\int_z^t [\lambda_{23}(u)+\lambda_{24}(u)]du\Big),\\
P_{33}(z,t) &= \exp\Big(-\int_z^t \lambda_{34}(u)du\Big), \quad P_{34}(z,t)=1-P_{33}(z,t),\\
P_{23}(z,t) &= \int_z^t P_{22}(z,u)\lambda_{23}(u)P_{33}(u,t)du, \quad P_{24}(z,t)=1-P_{22}(z,t)-P_{23}(z,t),\\
P_{12}(z,t) &= \int_z^t P_{11}(z,u)\lambda_{12}(u)P_{22}(u,t)du,\\
P_{13}(z,t) &= \int_z^t [P_{11}(z,u)\lambda_{13}(u)+P_{12}(z,u)\lambda_{23}(u)P_{33}(u,t)]du,\\
P_{14}(z,t) &= 1-P_{11}(z,t)-P_{12}(z,t)-P_{13}(z,t).
\end{aligned}
\tag{6.12}
$$

Since premium calculation is based on one-year transition probabilities we use a discretized version of (6.12). In particular, the transition intensities are replaced by one-year transition probabilities $p_{ij}(n) = P(S(n+1) = j|S(n) = 1)$ according to (5.7). The dependence of the diagnoses is eliminated using a weighted mean over the 22 groups of diagnoses similar to (5.8). The dependence of age, gender, LTC duration and care level is still included, for notational convenience however the probabilities are only given in dependence of insurance duration n.

We assume the following payments:

- π: annual premium, paid by the insured while the risk is in state 1
- b_i: annuity paid by the insurer while the risk is in state i
- c_{ij}: a lump sum paid by the insurer when a state transition from state i to state j occurs

Then the actuarial values of the expected benefits at the beginning of the insurance contract are given as follows: for an individual with entry age x when a lump sum c_{1j} is payable at the moment of a change from healthy (state 1) to state 2 or 3 the actuarial value $B_{1,c_{1j}}(0), j \in \{2,3\}$ is given by

$$B_{1,c_{1j}}(0) = \sum_{i=0}^{\omega-x-1} P_{11}(0,i)p_{1j}(i)v^i c_{1j}.$$

Here ω denotes the actuarial end-age, i.e. $P(T > \omega) := 0$ and $v = e^{-\delta}$ where δ is the force of interest which is taken $\delta = 0.035$ in our calculations. That's simply the probability that in year i a transition from healthy to state j occurs multiplied by the discounted value of the benefit and summed up over all years. For an annuity b_j payable while an insured person is in state

$j \in \{2,3\}$ the actuarial value $B_{1,b_j}(0)$ is given by

$$B_{1,b_j} = \sum_{i=0}^{\omega-x-1} P_{1j}(0,i)v^i b_j.$$

Here, the probability of an insured person being in state j in year i is multiplied by the discounted value of the annuity. For a detailed derivation of these actuarial values see Rudolph (2000) and Czado and Rudolph (2002). The actuarial values $B_{1,c_{1j}}, B_{1,b_j}, j \in \{2,3\}$ are calculated separately for the three care levels. Then a mean weighted by the proportion of each care level in the data is taken. Based on the equivalence principle at policy begin, i.e. the sum of all expected benefits $\sum_{j=2}^{3}(B_{1,b_j}(0) + B_{1,c_{1j}}(0))$ at time 0 has to equal the sum of all expected premiums $P_{1,\pi}(0) = \sum_{i=0}^{\omega-x-1} P_{11}(0,i)v^i\pi$ at time 0, finally the premium π can be calculated.
A computer program in C was written to calculate these premiums based on the mortality intensities resulting from the semiparametric hazard models including diagnoses (4.5), (4.6) and without diagnoses (5.9). The transitions intensities from state 2 to state 3 are modeled as well with and without diagnoses, for details see Gschlößl (2002). We want to compare the results to the premiums resulting from the LTC-plan "PLTC" offered by a German health insurer. In this plan the insured persons contract a fixed amount serving as a daily cash allowance. If an insured person receives care at home it is paid 25 % of this allowance in level 1, 50 % in level 2 and 75 % in level 3. In the case of care in nursing homes 100 % are paid. Therefore, input parameters of the C program are annuities depending on type of care and care level. Since our data contain no information about transitions from healthy to death as well as from healthy to care needing, we use Bavarian life tables (1986-1988) and LTC incidence rates of custodial insurance, Japan, for those transition probabilities. In Table 9 the estimated premiums based on the semiparametric hazard models including and without diagnoses are compared to the premiums based on the plan "PLTC" for an daily allowance of 10 Euro. The same results are presented graphically in Figure 9. First of all note, that the premiums by the plan "PLTC" can only be taken as a rough benchmark, since our calculations are based on probabilities from several sources and the underlying interest rate of the plan "PLTC" is not known exactly.

The increase in the calculated premiums according to age is quite similar to the "PLTC" premiums. However, in our calculations women have higher premiums in relation to men which indicates that the risk for women might be underestimated in reality. Premiums based on the information of the diagnoses are slightly higher as without this information (see Figure 9). In particular, for men premiums including diagnoses are up to 6 % percent higher, for women up to

Age	Premiums based on semiparametric hazard model				Premium offered by	
	including diagnoses		without diagnoses		German health insurer	
	Female	Male	Female	Male	Female	Male
20	2.79	1.60	2.74	1.51	1.74	1.18
25	3.46	1.98	3.40	1.87	2.38	1.52
30	4.33	2.48	4.26	2.34	3.23	2.03
35	5.48	3.15	5.40	2.97	4.28	2.69
40	7.02	4.05	6.91	3.82	5.59	3.51
45	9.03	5.21	8.92	4.94	7.27	4.59
50	11.71	6.75	11.61	6.42	9.48	6.05
55	15.37	8.79	15.25	8.43	12.43	8.04
60	20.35	11.51	20.21	11.17	16.49	10.84
65	27.15	15.15	26.99	14.92	21.53	14.40
70	36.10	20.05	36.16	20.01	29.69	20.09

Table 9: Premiums calculated based on a model with and without diagnoses in comparison to premiums offered by a German health insurer for an daily allowance of 10 Euro

3.8 % (see Table 10). Further the diagnoses affect the premiums mainly for younger age groups, for increasing age the influence of the diagnoses on premiums is diminishing. Hence, without the consideration of diagnoses the risk seems to be underestimated for younger age groups and the premiums offered by a German health insurer might be too low. This demonstrates the need to account for inception selection effect of diagnosis.

Age	20	25	30	35	40	45	50	55	60	65	70
Men	1.060	1.059	1.056	1.057	1.055	1.051	1.047	1.038	1.025	1.010	0.996
Women	1.037	1.038	1.014	1.037	1.036	1.033	1.029	1.029	1.028	1.026	1.019

Table 10: Ratio of premiums based on the model including diagnoses and the model without diagnoses

7 Summary and Discussion

The main issue of this paper was to quantify the inception selection effect of diagnosis on LTC insurance premiums when information on case specific diagnoses are available in a large port-

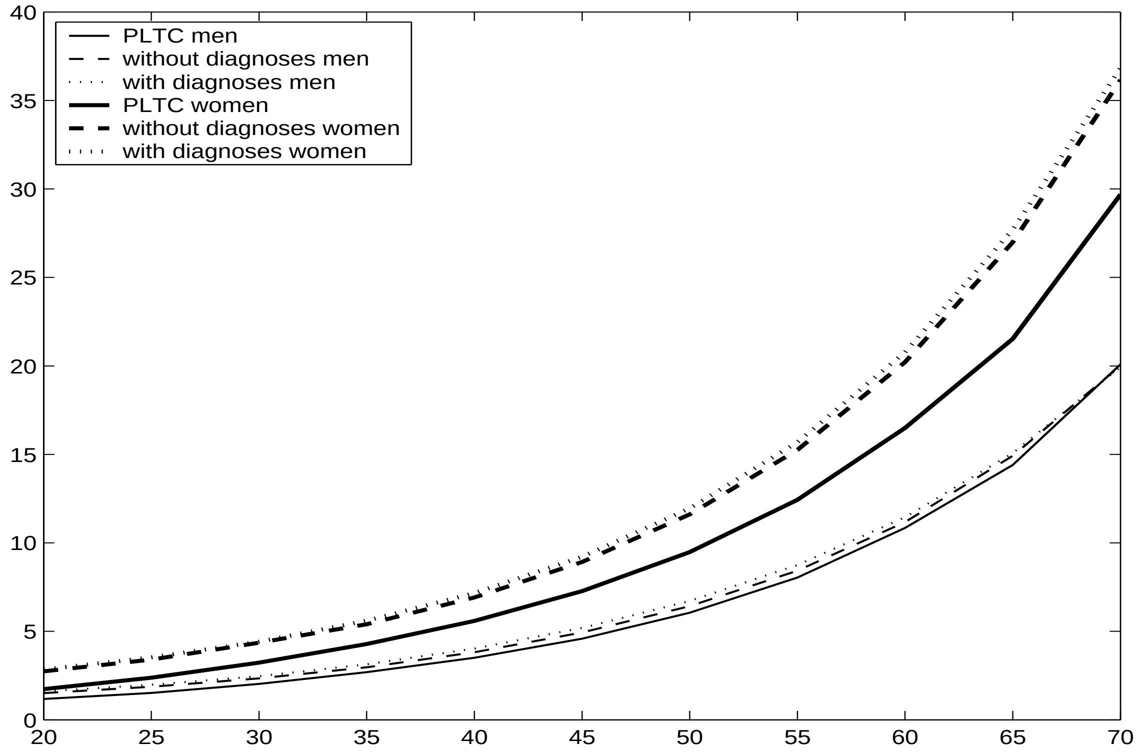

Figure 9: Calculated premiums based on two semiparametric hazard models and for the plan "PLTC" of a German health insurer. Thick lines: women, thin lines: men

folio. For this we first have to allow for diagnosis specific effects in the estimation of transition intensities and probabilities. In the second step these transition intensities and probabilities are utilized to price LTC premiums. For the modeling of diagnosis effects on estimated transition intensities and probabilities the problem of multiple diagnoses had to be considered. Modern model diagnostics for the semiparametric hazard model were used. Violations of the assumptions of the model could be detected using graphical methods. An appropriate modeling of the functional form of continuous covariates was achieved by using fractional polynomials and exponential functions. By splitting the data set, most of the time dependency present in the data could be deleted. To obtain transition probabilities independent of the diagnoses a weighted mean over the different groups of diagnoses was used. We have shown that the inclusion of diagnoses leads to a significant improvement over a model without diagnoses. The transition probabilities based on the model including diagnoses give more realistic estimates, particularly in the first year of LTC. Although the influence of the diagnoses is diminishing with time, their effects do not vanish completely. After diagnosis specific transition probabilities have been estimated and validated, we show how LTC premiums can be calculated when annuities based on type of care and level of care are paid to the insured requiring LTC. For this LTC product this gives higher premiums for younger age groups when diagnoses are taken into account. Although the difference to the premiums without diagnoses decreases with age, still slightly higher values are observed. This indicates that insurance companies might underestimate their risk if the diagnoses are neglected. Thus insurance companies should be encouraged to investigate if their offered premiums are sufficient for their portfolio to cover the actual losses observed through a retrospective analysis using the methods given in this paper. The need for analyses which study specific inception selection effects is well recognized. For example Macdonald and Pritchard (2001) study the effect of Alzheimer's Disease on LTC. Since genetic testing for this disease is available they can also investigate the adverse selection effect, which arises when carriers of the disease buy LTC insurance at a high rate. In contrast our analysis does not include such effects. However our approach is more general, since it includes diagnoses where the preposition of a disease yielding to LTC cannot be tested at entry into the insurance contract. In summary, we like to stress the need to understand the risks at younger ages when issuing LTC insurance (see for example Weltz (2002)) and therefore we believe even a moderate improvement in assessing the risk in younger age groups is important.

Acknowledgement

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

References

Akaike, H. (1973). Information Theory and a Extension of the Maximum Likelihood Principle. *2nd International Symposium of Information Theory and Control*, 267–281.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics 30*, 89–99.

Cox, D. (1975). Partial likelihood. *Biometrika 62*, 269–276.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Stat. Society B 34*, 187–220.

Czado, C. and F. Rudolph (2002). Application of survival analysis methods to long term care insurance. *Insurance: Mathematics and Economics 31*, 395–413.

Grambsch, P. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrica 81*, 515–526.

Gschlößl, S. (2002). Neuere statistische Methoden in der Pflegeversicherung. Diplomarbeit, Technische Universität München.

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

Karlin, S. and H. Taylor (1981). *A second course in stochastic processes.* New York: Academic Press.

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

Macdonald, A. S. and D. J. Pritchard (2001). Genetics, alzheimer's disease and long-term care insurance. *North American Actuarial Journal 5:2*, 54–78.

Royston, P. and D. Altman (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics 43*, 429–467.

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

Therneau, T. M., P. M. Grambsch, and T. R. Fleming (1990). Martingale-based residuals for survival models. *Biometrika 1*, 147–160.

Weltz, S. A. (2002). Understanding the risks at the younger ages. *Long-Term Care News, December 2002*, Issue No.7.