Czado, Gschlößl:

Modeling of transition intensities and probabilities in a German long term care portfolio with known diagnosis

Sonderforschungsbereich 386, Paper 302 (2002)

Online unter: http://epub.ub.uni-muenchen.de/

Modeling of transition intensities and probabilities in a German long term care portfolio with known diagnosis

Claudia Czado Susanne Gschlößl *

November 26, 2002

Abstract

In this paper a semiparametric hazard model introduced by Cox (1972) is used to model transitions intensities for a long term care (LTC) data set. The main focus is the inclusion of the diagnoses which led to LTC as explanatory variables. Modern model diagnostic techniques are applied to check the model assumptions. Fractional Polynomials proposed by Royston and Altman (1994) are used to model the functional form of continuous covariates. Time dependency is examined graphically by using scaled Schoenfeld residuals (see Grambsch and Therneau (1994)). It is shown that the inclusion of diagnoses significantly improves the estimated transition probabilities on which premiums are based. As an alternative approach a piecewise exponential model is fitted and compared to the semiparametric hazard model.

Keywords: semiparametric hazard model, survival analysis, long term care insurance, fractional polynomials, piecewise exponential model

*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 an 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 better transition rates and probabilities. Apart from using a semiparametric hazard model, we follow the approach of a piecewise exponential model given by Holford (1980) and Laird and Olivier (1981). They show an equivalence between a piecewise exponential model and a Poisson model. Haberman and Pitacco (1999) use a similar approach to model transition intensities for permanent health insurance data considering the factors age and duration of sickness only. 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. A comparison to piecewise exponential models 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.

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

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.

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 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:

	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)

$$
\begin{aligned}
\lambda(t) &= \lambda_0(t)\exp[\beta_1\mathbf{Z}_{Age}(t) + \beta_2\mathbf{Z}_{Sex} + \beta_3\mathbf{Z}_{nh}(t) + \beta_4\mathbf{Z}_{Level2}(t) + \beta_5\mathbf{Z}_{Level3}(t) \\
&+ \beta_6\mathbf{Z}_{Sex}\times\mathbf{Z}_{nh}(t) + \beta_7\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Level2}(t) + \beta_8\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Level3}(t) \\
&+ \beta_9\mathbf{Z}_{Dementia} + \beta_{10}\mathbf{Z}_{Stroke} + \beta_{11}\mathbf{Z}_{Psychosis} + \beta_{12}\mathbf{Z}_{Tumor} + \beta_{13}\mathbf{Z}_{Heart} + \beta_{14}\mathbf{Z}_{Lung} \\
&+ \beta_{15}\mathbf{Z}_{Arthritis} + \beta_{16}\mathbf{Z}_{Age}(t)\times\mathbf{Z}_{Tumor} + \beta_{17}\mathbf{Z}_{Age}(t)\times\mathbf{Z}_{Arthritis} \\
&+ \beta_{18}\mathbf{Z}_{Sex}\times\mathbf{Z}_{Tumor} + \beta_{19}\mathbf{Z}_{Sex}\times\mathbf{Z}_{Psychosis} + \beta_{20}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Psychosis} \\
&+ \beta_{21}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Stroke} + \beta_{22}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Lung} + \beta_{23}\mathbf{Z}_{Tumor}\times\mathbf{Z}_{Heart} \\
&+ \beta_{24}\mathbf{Z}_{Age}(t)\times\mathbf{Z}_{Heart} + \beta_{25}\mathbf{Z}_{Level2}(t)\times\mathbf{Z}_{Heart} + \beta_{26}\mathbf{Z}_{Level3}(t)\times\mathbf{Z}_{Heart} \qquad (3.1) \\
&+ \beta_{27}\mathbf{Z}_{Tumor}\times\mathbf{Z}_{Level2}(t) + \beta_{28}\mathbf{Z}_{Tumor}\times\mathbf{Z}_{Level3}(t) + \beta_{29}\mathbf{Z}_{Tumor}\times\mathbf{Z}_{Psychosis} \\
&+ \beta_{30}\mathbf{Z}_{Tumor}\times\mathbf{Z}_{Dementia} + \beta_{31}\mathbf{Z}_{Psychosis}\times\mathbf{Z}_{Stroke} + \beta_{32}\mathbf{Z}_{Tumor}\times\mathbf{Z}_{Age}(t)\times\mathbf{Z}_{Heart}].
\end{aligned}
$$

The estimated regression coefficients of Model (3.1), the corresponding Wald statistics $W = \frac{\hat{\beta}}{SE(\hat{\beta})}$ and the resulting p-values are presented in Table 6. Note, that tumors clearly reduce the expected lifetime (estimated coefficient = 3.9). A higher care level results in a higher risk of mortality, whereas women seem to have a lower risk to die. However, the interpretation of Table 6 is not an easy task, since numerous interactions present in Model (3.1) must be taken into account. Therefore, the multipliers $\exp[\boldsymbol{\beta}'\mathbf{Z}]$ are plotted for several groups of claimants in Figure 1. 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).

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. In Figure 2 martingale residuals of age are presented for a variety of groups with single and double diagnoses. First of all, most of the plots clearly show a nonlinear functional form of age. Further, note the obvious difference in the shape of single diagnoses and the same diagnoses in combination with another one. Just men with psychosis, psychosis and heart attack or psychosis and dementia may be summarized

Covariate	$\hat{\beta}$	$\exp(\hat{\beta})$	$\mathrm{SE}(\hat{\beta})$	W	p-Value
$\mathbf{Z}_{Age}(t)$	0.0384	1.0391	0.0031	12.200	$< 10^{-16}$
$\mathbf{Z}_{Tumor}$	3.9070	49.7497	0.3670	10.589	$< 10^{-16}$
$\mathbf{Z}_{Arthritis}$	-2.1288	0.1190	0.9061	-2.349	$1.9 \cdot 10^{-2}$
$\mathbf{Z}_{Sex}$	-0.2617	0.7698	0.0726	-3.603	$3.1 \cdot 10^{-4}$
$\mathbf{Z}_{nh}(t)$	0.5954	1.8137	0.1529	3.895	$9.8 \cdot 10^{-5}$
$\mathbf{Z}_{Psychosis}$	-0.1757	0.8389	0.0976	-1.801	$7.2 \cdot 10^{-2}$
$\mathbf{Z}_{Level2}(t)$	0.6822	1.9782	0.1136	6.003	$1.9 \cdot 10^{-9}$
$\mathbf{Z}_{Level3}(t)$	1.7890	5.9836	0.1113	16.077	$< 10^{-16}$
$\mathbf{Z}_{Stroke}$	-0.3485	0.7058	0.0773	-4.510	$6.5 \cdot 10^{-6}$
$\mathbf{Z}_{Lung}$	0.1401	1.1504	0.1660	0.844	$4.0 \cdot 10^{-1}$
$\mathbf{Z}_{Heart}$	1.0692	2.9130	0.4779	2.237	$2.5 \cdot 10^{-2}$
$\mathbf{Z}_{Dementia}$	-0.0616	0.9403	0.0527	-1.169	$2.4 \cdot 10^{-1}$
$\mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Tumor}$	-0.0379	0.9628	0.0044	-8.614	$< 10^{-16}$
$\mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Arthritis}$	0.0236	1.0239	0.0103	2.306	$2.1 \cdot 10^{-2}$
$\mathbf{Z}_{Sex} \times \mathbf{Z}_{nh}(t)$	-0.3648	0.6943	0.1019	-3.581	$3.4 \cdot 10^{-4}$
$\mathbf{Z}_{Sex} \times \mathbf{Z}_{Tumor}$	0.4116	1.5093	0.1129	3.645	$2.7 \cdot 10^{-4}$
$\mathbf{Z}_{Sex} \times \mathbf{Z}_{Psychosis}$	-0.2348	0.7907	0.1103	-2.128	$3.3 \cdot 10^{-2}$
$\mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Level2}(t)$	-0.2329	0.7922	0.1529	-1.524	$1.3 \cdot 10^{-1}$
$\mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Level3}(t)$	-0.6148	0.5408	0.1484	-4.142	$3.4 \cdot 10^{-5}$
$\mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Psychosis}$	0.2861	1.3313	0.1142	2.506	$1.2 \cdot 10^{-2}$
$\mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Stroke}$	0.2403	1.2717	0.1174	2.047	$4.1 \cdot 10^{-2}$
$\mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Lung}$	0.6426	1.9013	0.3036	2.117	$3.4 \cdot 10^{-2}$
$\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Heart}$	-3.8934	0.0204	1.1845	-3.286	$1.0 \cdot 10^{-3}$
$\mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Heart}$	-0.0089	0.9911	0.0055	-1.620	$1.1 \cdot 10^{-1}$
$\mathbf{Z}_{Level2}(t) \times \mathbf{Z}_{Heart}$	0.0496	1.0508	0.1350	0.367	$7.1 \cdot 10^{-1}$
$\mathbf{Z}_{Level3}(t) \times \mathbf{Z}_{Heart}$	-0.3673	0.6926	0.1317	-2.788	$5.3 \cdot 10^{-3}$
$\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Level2}(t)$	0.3657	1.4415	0.1697	2.155	$3.1 \cdot 10^{-2}$
$\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Level3}(t)$	0.3166	1.3724	0.1656	1.911	$5.6 \cdot 10^{-2}$
$\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Psychosis}$	-0.3499	0.7048	0.1536	-2.279	$2.3 \cdot 10^{-2}$
$\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Dementia}$	-0.3117	0.7322	0.1345	-2.318	$2.0 \cdot 10^{-2}$
$\mathbf{Z}_{Psychosis} \times \mathbf{Z}_{Stroke}$	0.5159	1.6751	0.1274	4.051	$5.1 \cdot 10^{-5}$
$\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Age}(t) \times \mathbf{Z}_{Heart}$	0.0408	1.0417	0.0143	2.863	$4.2 \cdot 10^{-3}$

Table 6: Estimated Regression Coefficients, together with estimated standard errors (SE), the corresponding Wald statistic and p-value in Model (3.1)

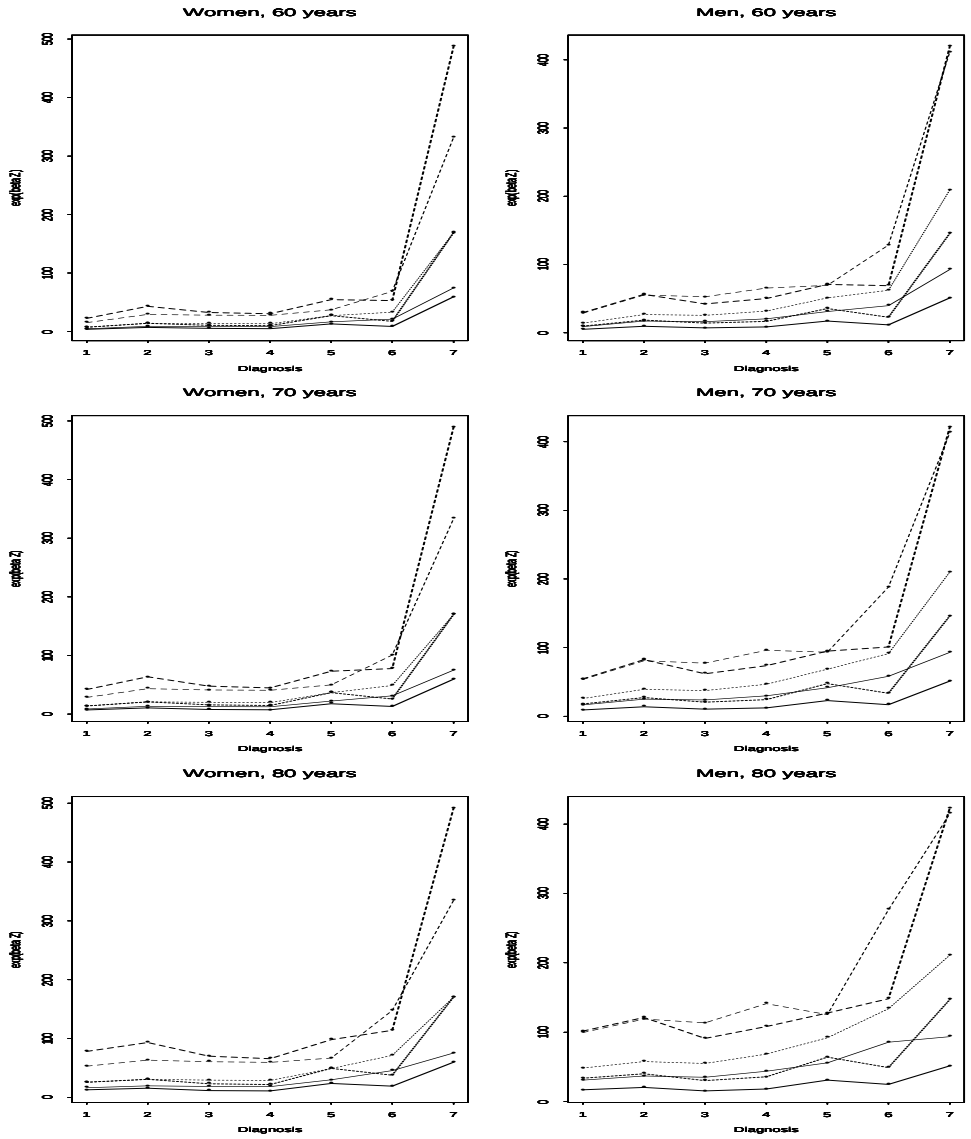

Figure 1: Estimated multipliers $\exp[\hat{\beta}'\mathbf{Z}]$ for women and men with single diagnoses in Model (3.1) (Legend is given in Table 7)

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 7: Legend to Figure 1

to one group of claimants due to their similar functional form (see Figure 2, last row). In the following this group will be referred to as

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

For all the other groups, age should be modeled separately. In the group PHD an exponential function might be appropriate, whereas the correct functional form might be given by quadratic functions for claimants with tumor or women with psychosis and heart attack for example. 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 (claimants with tumor as single diagnosis) the negative of this 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. $x^0 = \ln(x)$. By using repeated exponents $p_i = ... = p_k, i < k \leq m$ combinations of $\ln x$ can be incorporated.

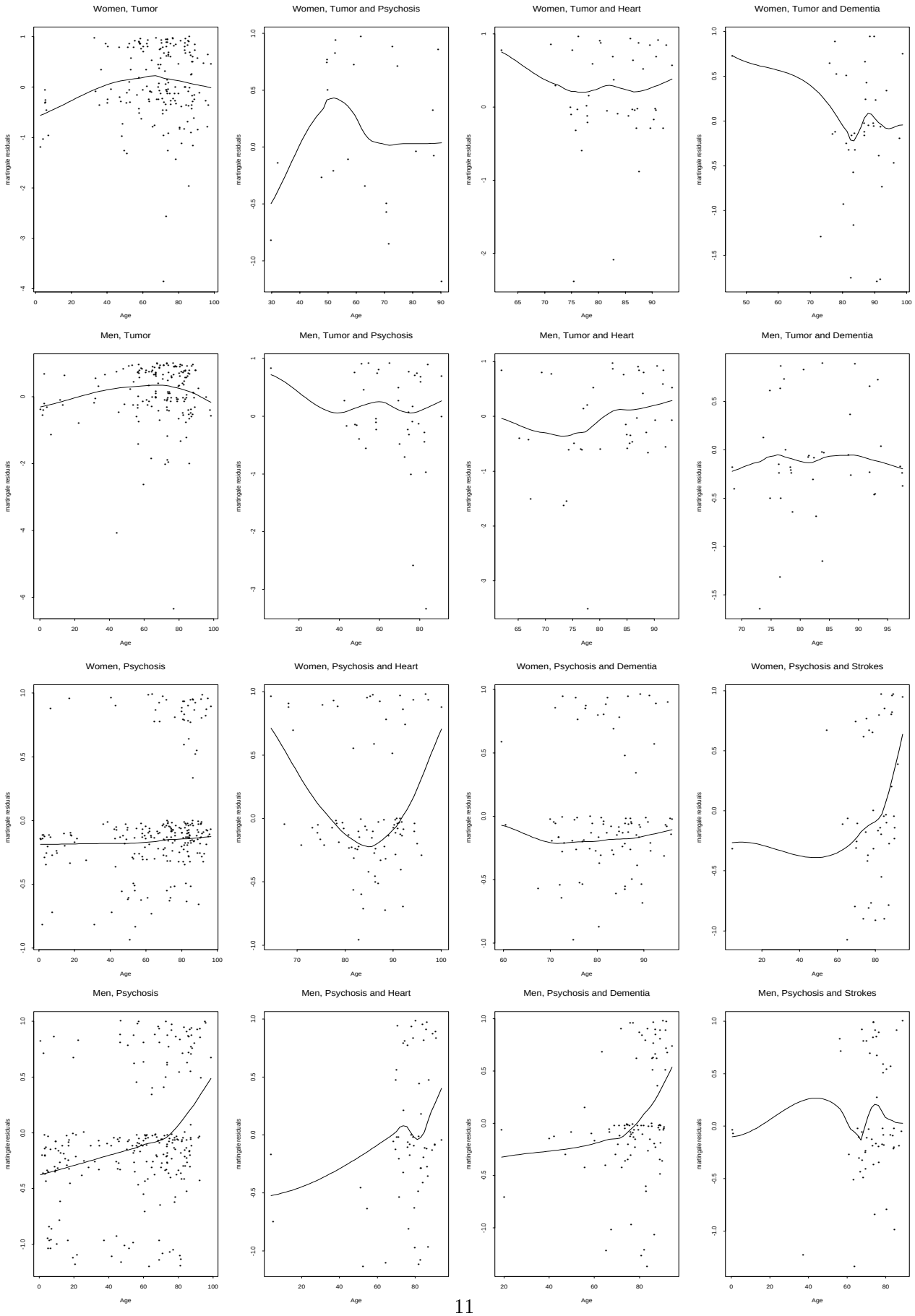

Figure 2: Martingale residuals for single and double diagnoses with tumors und psychosis

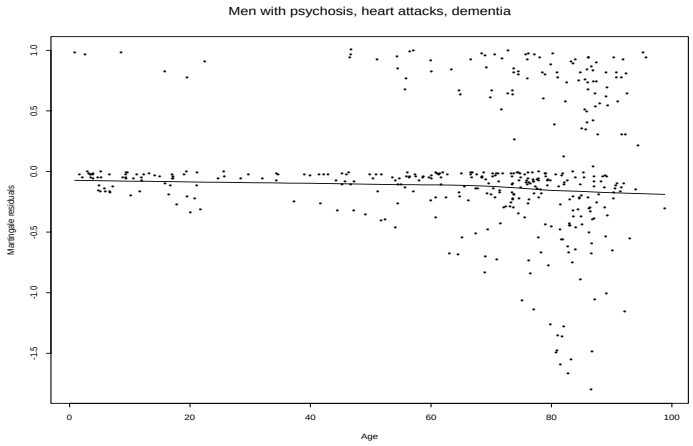

Figure 3: Martingale residuals in the model with exponential term (4.3) for men with psychosis, heart attack and dementia

For example, if $m = 6$ and $p = (-1, 0, 0.5, 2, 2, 2)$ the associate fractional polynom is defined as follows

$$\Phi_6(x, p) = \beta_0 + \beta_1 \frac{1}{x} + \beta_2 \ln x + \beta_3 \sqrt{x} + \beta_4 x^2 + \beta_5 x^2 \ln x + \beta_6 x^2 (\ln x)^2.$$

Thus, fractional polynomials allow for a variety of different functional shapes. Ordinary polynomials are included, if only intergers are chosen as exponents. In general, fractional polynomials of degree one and two are sufficient in most data sets.

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 (see Tables 8 and 9). 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 Figure 3. 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 Figure 2 (first column on the left hand side) 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 Figure 4) 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

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

	Log-Likelihood	$2 \cdot \chi^2_{LR}$ vs. linear	p-Value	exponents
linear (1 df)	-14168.1			1
m=1 (2 df)	-14164.86	3.24	0.011	-1
m=2 (4 df)	-14163.44	4.66	0.02	-1, -1

Table 8: Fractional polynomials for age of women 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

	Log-Likelihood	$2 \cdot \chi^2_{LR}$ vs. linear	p-Value	exponents
linear (1 df)	-14170.6			1
m=1 (2 df)	-14169.48	2.24	0.13	2
m=2 (4 df)	-14167.09	7.02	0.07	-2,1

Table 9: Fractional polynomials for age of men with psychosis, heart attack or dementia

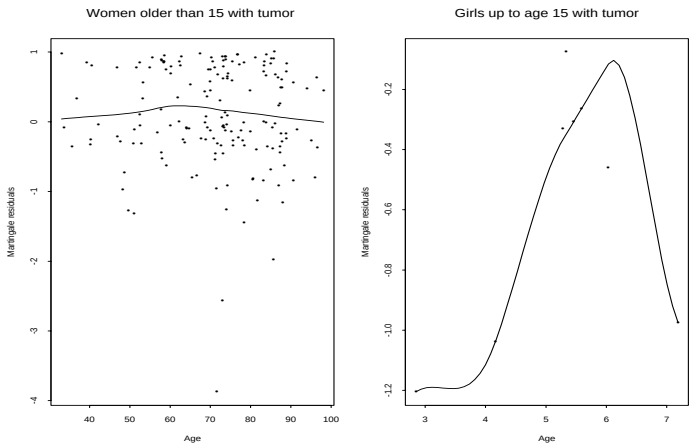

Figure 4: Martingale residuals for women older than fifteen and girls up to age 15 with tumor

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 5 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 all 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 modeled via the exponential and fractional polynomial approach. As a result we get the following models

$$\begin{aligned}\lambda(t) &= \lambda_0(t)\exp[\beta_1\mathbf{Z}_{Age}(t)\times\mathbf{Z}_{Tumor,G}+\beta_2\mathbf{Z}_{Age}(t)\times(1-\mathbf{Z}_{Tumor,W})+\beta_3\mathbf{Z}_{Sex}\\ &+ \beta_4\mathbf{Z}_{nh}(t)+\beta_5\mathbf{Z}_{Level2}(t)+\beta_6\mathbf{Z}_{Level3}(t)+\beta_7\mathbf{Z}_{Tumor}+\beta_8\mathbf{Z}_{Dementia}+\beta_9\mathbf{Z}_{Heart}\\ &+ \beta_{10}\mathbf{Z}_{Psychosis}+\beta_{11}\mathbf{Z}_{Stroke}+\beta_{12}\mathbf{Z}_{Arthritis}+\beta_{13}\mathbf{Z}_{Lung}\end{aligned}$$

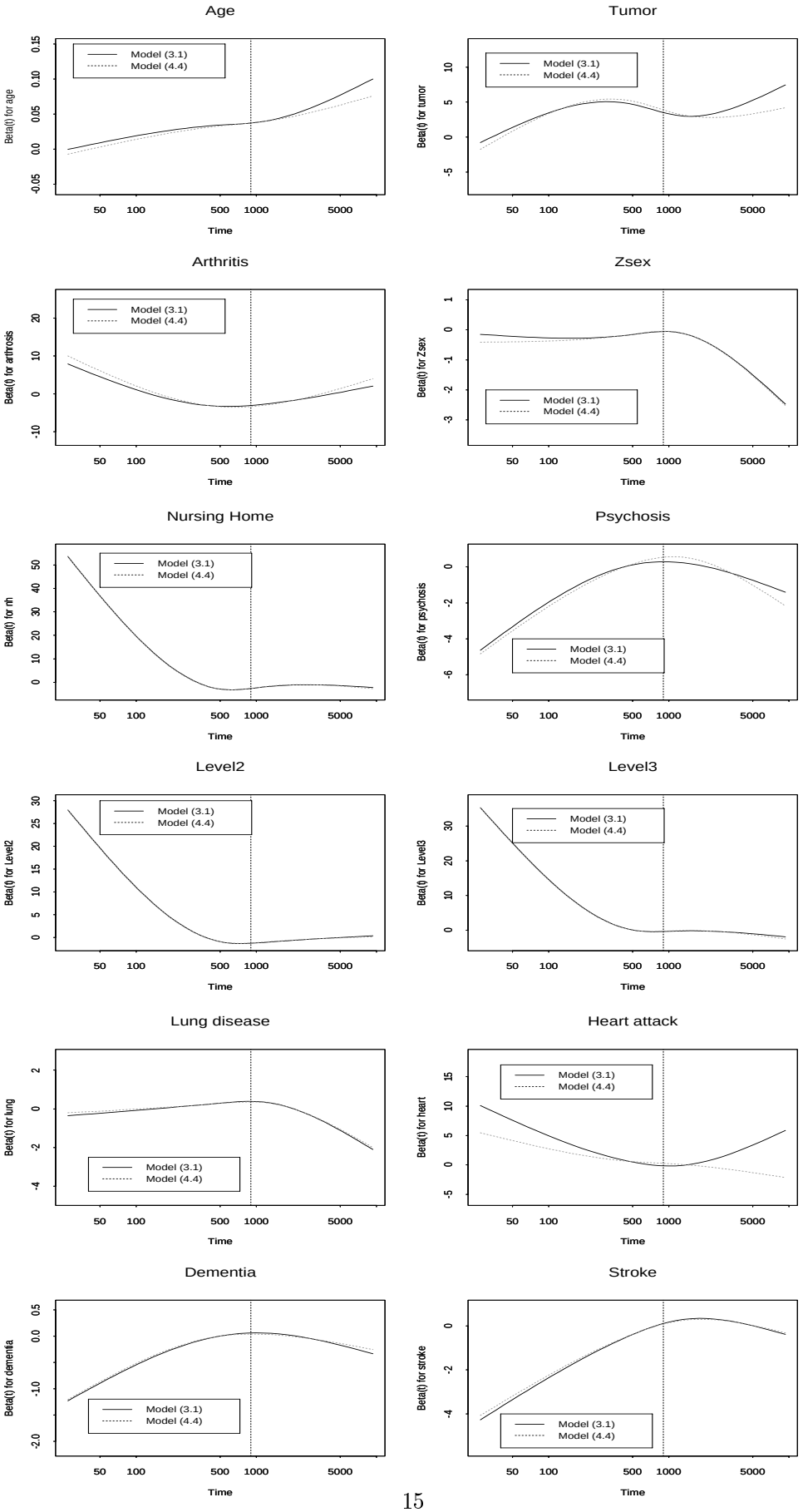

Figure 5: Scaled Schoenfeld residuals against time for Models (3.1) and (4.4)

$$
\begin{aligned}
&+ \quad \beta_{14}\mathbf{Z}_{Age}(t) \times (1 - \mathbf{Z}_{Tumor,W}) \times \mathbf{Z}_{Tumor} + \beta_{15}\mathbf{Z}_{Age}(t) \times (1 - \mathbf{Z}_{Tumor,W}) \times \mathbf{Z}_{Arthritis} \\
&+ \quad \beta_{16}\mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Stroke} + \beta_{17}\mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Level2}(t) + \beta_{18}\mathbf{Z}_{nh}(t) \times \mathbf{Z}_{Level3}(t) \\
&+ \quad \beta_{19}\mathbf{Z}_{Sex} \times \mathbf{Z}_{Tumor} + \beta_{20}\mathbf{Z}_{Level2}(t) \times \mathbf{Z}_{Heart} + \beta_{21}\mathbf{Z}_{Level3}(t) \times \mathbf{Z}_{Heart} \\
&+ \quad \beta_{22}\mathbf{Z}_{Psychosis} \times \mathbf{Z}_{Stroke} + \beta_{23}\mathbf{Z}_{Tumor} \times \mathbf{Z}_{Heart} + \beta_{24}\mathbf{Z}_{Stroke} \times \mathbf{Z}_{Arthritis}] \qquad (4.5)
\end{aligned}
$$

for LTC durations up to 900 days and

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

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. 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).

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. The transition rates $\lambda_{i4}, i = 1, 2, 3$ have already been examined in the previous section. The modeling of the remaining intensities can be done in a similar way, but won't be shown in this paper (for details see Gschlößl (2002), pp.93-97). 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

$$
\hat{p}_{ij}(x, s, d, c, k) = \quad \sum_{d \le t_k < d+1} \quad \Big\{ \hat{\lambda}_{ij}(t_k,\, \mathbf{Z}_{Age}(t_k) = x, \mathbf{Z}_{Sex} = s, \mathbf{Z}_{nh} = c, \mathbf{Z}_{Group} = k)
$$

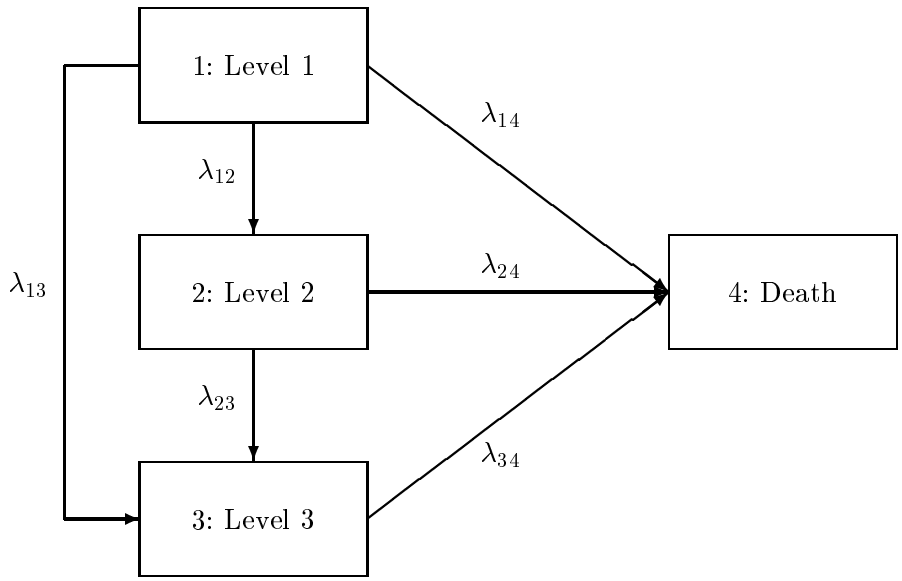

Figure 6: State transitions between levels and transitions to death

$$\prod_{d\le t_l<t_k} \quad (1-\hat{\lambda}_{ij}(t_l,\, \mathbf{Z}_{Age}(t_l)=x, \mathbf{Z}_{Sex}=s, \mathbf{Z}_{nh}=c, \mathbf{Z}_{Group}=k))\Big\}. \quad (5.7)$$

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 $\hat{S}(t)=\prod_{t_j\le t}(1-\hat{\lambda}(t_j))$ for the survival function $S(t)=P(T>t)$. We obtain the estimated hazard functions $\hat{\lambda}_{ij}(t,\mathbf{Z})$ using 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

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

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

$$\begin{aligned}\lambda(t) &= \lambda_0(t)\exp[\beta_1\mathbf{Z}_{Age}(t) + \beta_2\mathbf{Z}_{Sex} + \beta_3\mathbf{Z}_{nh}(t) + \beta_4\mathbf{Z}_{Level2}(t) + \beta_5\mathbf{Z}_{Level3}(t) \\ &+ \beta_6\mathbf{Z}_{Age}(t)\times\mathbf{Z}_{Sex} + \beta_7\mathbf{Z}_{Age}(t)\times\mathbf{Z}_{nh}(t) + \beta_8\mathbf{Z}_{Sex}\times\mathbf{Z}_{nh}(t) \\ &+ \beta_9\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Level2}(t) + \beta_{10}\mathbf{Z}_{nh}(t)\times\mathbf{Z}_{Level3}(t)] \qquad (5.9)\end{aligned}$$

(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 11, 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.

6 Comparison to piecewise exponential models

An alternative approach to the semiparametric hazard model (2.1) is the piecewise exponential model. The piecewise exponential model (see for example Laird and Olivier (1981) or Holford (1980)) is a semiparametric hazard model of the form (2.1) with the restriction, that the baseline hazard is constant during certain time intervals with limits $0 = \tau_0 < \tau_1 < ... < \tau_J = \infty$, i.e. $\lambda_0(t) = \lambda_j$ for $t \in [\tau_{j-1}, \tau_j)$. Therefore, the hazard rate for individual i in interval j is given by

$$\lambda_{ij} = \lambda_j \exp[\boldsymbol{\beta}'\mathbf{Z}_i]. \tag{6.10}$$

Let

$$d_i = \begin{cases} 1 & \text{individual i dies} \\ 0 & \text{otherwise} \end{cases}$$

be the death indicator for individual i. Creating pseudo observations, i.e. a separate observation for each individual per interval, we define for individual i in interval $[\tau_{j-1}, \tau_j)$ the exposure time

$$t_{ij} = \begin{cases} \tau_j - \tau_{j-1}, & T_i > \tau_j \\ T_i - \tau_{j-1}, & \tau_{j-1} < T_i < \tau_j \\ 0, & T_i < \tau_{j-1} \end{cases}$$

where T_i is the observation time for individual i and the death indicator

$$d_{ij} = \begin{cases} 1 & \text{individual i dies in interval j} \\ 0 & \text{otherwise} \end{cases} . \tag{6.11}$$

Holford (1980) and Laird and Olivier (1981) show that the piecewise exponential model and a Poisson regression model $d_{ij} \sim Pois(t_{ij}\lambda_{ij})$ have proportional likelihoods. This implies that the maximum likelihood estimates of the regression coefficients $\boldsymbol{\beta}$ are the same in both models. Using the logarithm as link function we get

$$\log(t_{ij}\lambda_{ij}) = \log t_{ij} + \log \lambda_j + \boldsymbol{\beta}'\mathbf{Z}_i = \log t_{ij} + \alpha + \alpha_j + \boldsymbol{\beta}'\mathbf{Z}_i,$$

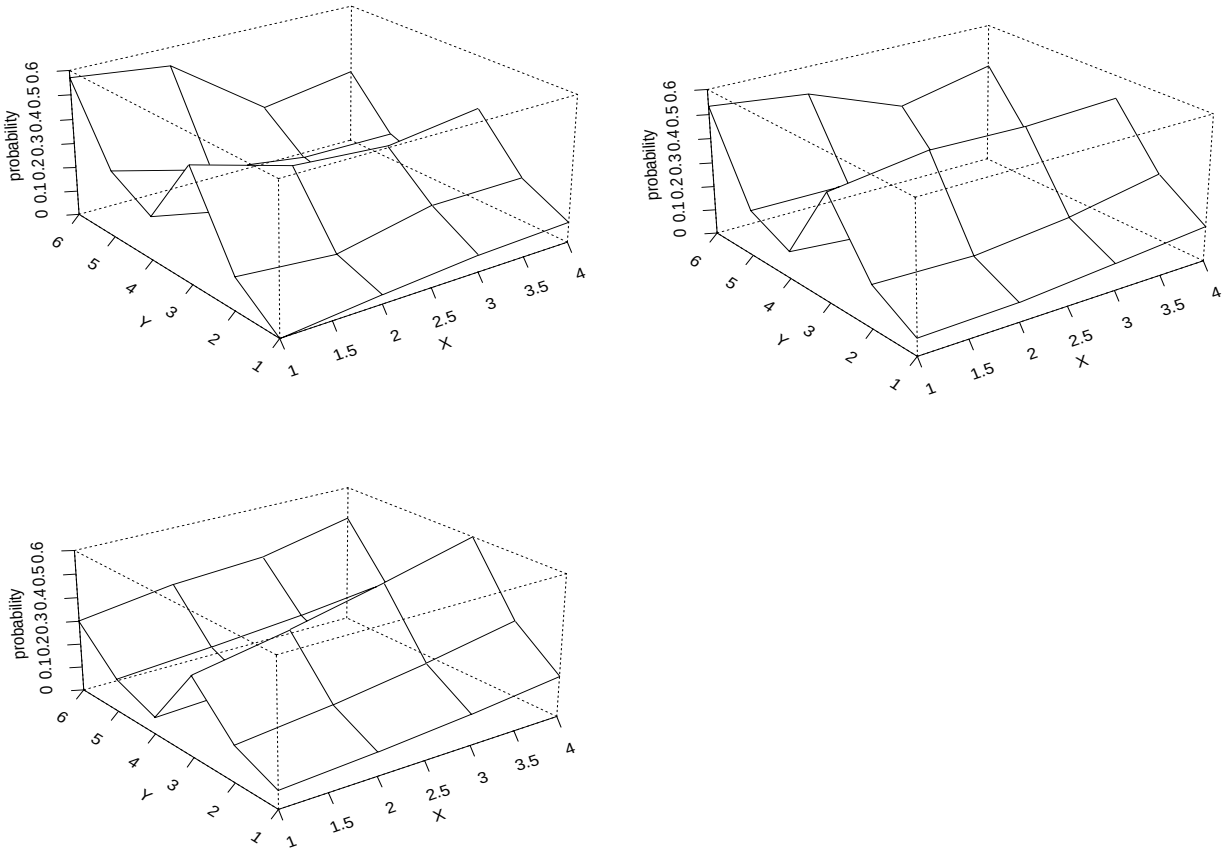

Figure 7: Mortality probabilities during the first year of LTC (Legend see Table 10)

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 10: 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 11: 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

where α is an intercept and α_j is the effect of time interval $[\tau_{j-1}, \tau_j)$. Thus, $\log t_{ij}$ enters the model as an offset. Since the creation of pseudo observations d_{ij} in (6.11) increases substantially the number of observations we only use data from the transition from Level 1 to death for our comparison of the semiparametric hazard model (2.1) to the piecewise exponential model (6.10). This data include 1528 observations, after restructing the data the number of observations has risen up to 2829. In particular, we consider the following time intervals

$$[0, 301),\ [301, 501),\ [501, 831),\ [831, 1370),\ [1370, \infty)\,\text{days}. \tag{6.12}$$

Each interval contains an approximately equal number of observations. For computational simplicity we group the data and thus use age as a factor, divided by quartiles as follows

$$[0, 75),\ [75, 83),\ [83, 88),\ [88, 108)\,\text{years}.$$

We choose the class [0,75) years as reference category. Time is treated as a factor as well with the indicator functions

$$\mathbf{Z}_{timej} = \begin{cases} 1 & \text{interval j} \\ 0 & \text{otherwise} \end{cases} \qquad , j = 1, .., 5$$

where the first interval serves as reference category, i.e. $\alpha_1 = 0$. Without including diagnoses we get the following log-hazard function for the transition from Level 1 to death

$$\begin{aligned}
\log \lambda_j &= \alpha + \alpha_j \mathbf{Z}_{timej} + \beta_1 \mathbf{Z}_{Sex} + \beta_2 \mathbf{Z}_{nh}(t) + \beta_3 \mathbf{Z}_{75-83years} + \beta_4 \mathbf{Z}_{83-88years} \\
&+ \beta_5 \mathbf{Z}_{88-108years} + \beta_6 (\mathbf{Z}_{75-83years} \times \mathbf{Z}_{nh}(t)) + \beta_7 (\mathbf{Z}_{83-88years} \times \mathbf{Z}_{nh}(t)) \\
&+ \beta_8 (\mathbf{Z}_{88-108years} \times \mathbf{Z}_{nh}(t)) + \beta_{j9} (\mathbf{Z}_{75-83years} \times \mathbf{Z}_{timej}) + \beta_{j10} (\mathbf{Z}_{83-88years} \times \mathbf{Z}_{timej}) \\
&+ \beta_{j11} (\mathbf{Z}_{88-108years} \times \mathbf{Z}_{timej}), \quad j = 1, .., 5.
\end{aligned} \tag{6.13}$$

Note, that age is modeled time dependent here through the inclusion of interactions between age and time. For comparison we fit the transition from Level 1 to Death as well with a semi-parametric hazard model. The hazard rate is given as follows

$$\begin{aligned}
\lambda(t) &= \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{75-83years} + \beta_2 \mathbf{Z}_{83-88years} + \beta_3 \mathbf{Z}_{88-108years} \\
&+ \beta_4 \mathbf{Z}_{nh}(t) + \beta_5 \mathbf{Z}_{Sex} + \beta_6 (\mathbf{Z}_{75-83years} \times \mathbf{Z}_{nh}(t)) \\
&+ \beta_7 (\mathbf{Z}_{83-88years} \times \mathbf{Z}_{nh}(t)) + \beta_8 (\mathbf{Z}_{88-108years} \times \mathbf{Z}_{nh}(t))]
\end{aligned} \tag{6.14}$$

To allow for a more appropriate comparison we fit as well a piecewise exponential model containing the same covariates as in Model (6.14), i.e. no time dependent modeling of age:

$$\begin{aligned}
\log \lambda_j &= \alpha + \alpha_j \mathbf{Z}_{timej} + \beta_1 \mathbf{Z}_{Sex} + \beta_2 \mathbf{Z}_{nh}(t) + \beta_3 \mathbf{Z}_{75-83years} \\
&+ \beta_4 \mathbf{Z}_{83-88years} + \beta_5 \mathbf{Z}_{88-108years} + \beta_6 (\mathbf{Z}_{75-83years} \times \mathbf{Z}_{nh}(t)) \\
&+ \beta_7 (\mathbf{Z}_{83-88years} \times \mathbf{Z}_{nh}(t)) + \beta_8 (\mathbf{Z}_{88-108years} \times \mathbf{Z}_{nh}(t)).
\end{aligned} \tag{6.15}$$

The resulting estimated hazard rates in interval j of the three models (6.13) - (6.15) and the empirical hazard rates

$$\lambda_{ij}^e := \sum_{\tau_{j-1} \le t_l < \tau_j} \frac{d_i(t_l)}{r_i(t_l)},$$

where $d_i(t)$ and $r_i(t)$ are the number of deaths respective number individuals in the risk set at time t, are plotted in Figure 8 for the first 4 time intervals. To estimate the baseline hazards of the piecewise exponential models, we need the estimated coefficients $\hat{\alpha}_j$ as well as the intercept $\hat{\alpha}$ and the length of each interval. Therefore, the baseline hazards of the piecewise exponential model can be estimated as follows:

$$\hat{\lambda}_j^{pe} := (\tau_j - \tau_{j-1}) \cdot \exp[\hat{\alpha} + \hat{\alpha}_j], \qquad j = 1, .., 5.$$

In the semiparametric hazard model the baseline hazard for interval j is obtained by

$$\hat{\lambda}_j^s := \sum_{\tau_{j-1} \leq t_i < \tau_j} \hat{\lambda}_0(t_i),$$

where $\hat{\lambda}_0(t_i)$ is the baseline hazard at death time t_i, estimated by Breslow's estimator (2.3) without covariates, i.e. $\mathbf{Z} = \mathbf{0}$. Thus, the estimated hazard rates for an individual with covariates $\mathbf{Z}_i$ in interval j are given by

$$\hat{\lambda}_{ij}^s := \hat{\lambda}_j^s \exp[\hat{\boldsymbol{\beta}}^{\boldsymbol{s}\prime}\mathbf{Z}_i] \qquad \text{and} \qquad \hat{\lambda}_{ij}^{pe} := \hat{\lambda}_j^{pe} \exp[\hat{\boldsymbol{\beta}}^{\boldsymbol{pe}\prime}\mathbf{Z}_i]$$

for the semiparametric hazard model and piecewise exponential model, respectively.

For the last time interval (not shown, but given in Gschlößl (2002), p.116) we see that models tend to overestimate the hazard. However the overestimation is more severe for the piecewise exponential models (6.13) and (6.15). Similar results were observed for men (see Gschlößl (2002), p.120) as well as for the transitions from Level 2 and Level 3 to Death, respectively.

6.1 Comparison

A comparison of the two approaches is not straightforward. First of all, the similarity between the semiparametric hazard model (6.14) and the corresponding piecewise exponential model (6.15) is obvious as expected. While the hazards based on those models are increasing with time, as assumed by the model, we can see nonlinear hazards for some groups in the piecewise exponential model (6.13), as for example women in nursing homes with age between 75 and 83 years (y-axis = 2). This is due to the separate modeling of the time intervals which allow for possible time dependency. A similar time dependency can be observed for the empirical hazards. However, in time intervals 3 and 4 the piecewise exponential model (6.13) assumes considerably higher mortality rates compared to the empirical ones. Therefore, the semiparametric hazard model (6.14) seems to give a better fit. In Figure 9 a smoothed plot of the estimated baseline hazards at each death time in Model (6.14) is given for the first four time intervals. The vertical lines indicate the limits of the intervals. Clearly, we can see that the assumption of piecewise constant baseline hazards does not hold here. All together we can state that the restriction to piecewise constant baseline hazards reduces the flexibility of the piecewise exponential model. The quality of the model heavily depends of the choice of the underlying time intervals. Another partition of time assuring a more homogeneous allocation of the observations may have led to a different fit. In addition, the intervals would have to be chosen annual if we wanted to estimate

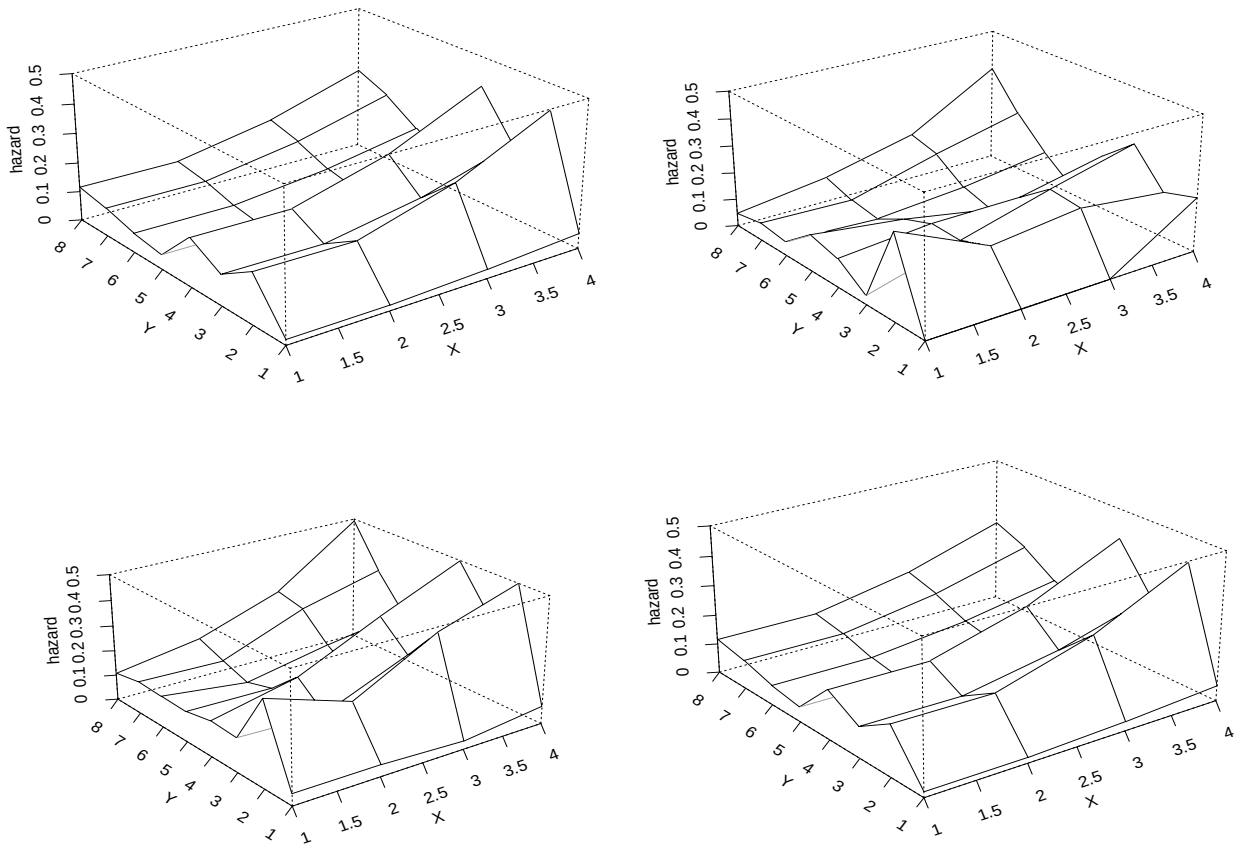

Figure 8: Hazard rates for women during the first four time intervals (Legend see Table 12)

top left	$\hat{\lambda}^s_{ij}$ based on Model (6.14)
top right	empirical hazards $\hat{\lambda}^e_{ij}$
bottom left	$\hat{\lambda}^{pe}_{ij}$ based on Model (6.13)
bottom right	$\hat{\lambda}^{pe}_{ij}$ based on Model (6.15)

x-axis		y-axis	
1	[0,301)	1	nh, -75 years
		2	nh, 75-83 years
2	[301,500)	3	nh, 83-88 years
		4	nh, 88-108 years
3	[501,831)	5	cah, -75 years
		6	cah, 75-83 years
4	[831,1370)	7	cah, 83-88 years
		8	cah, 88-108 years

Table 12: Legend to Figure 8

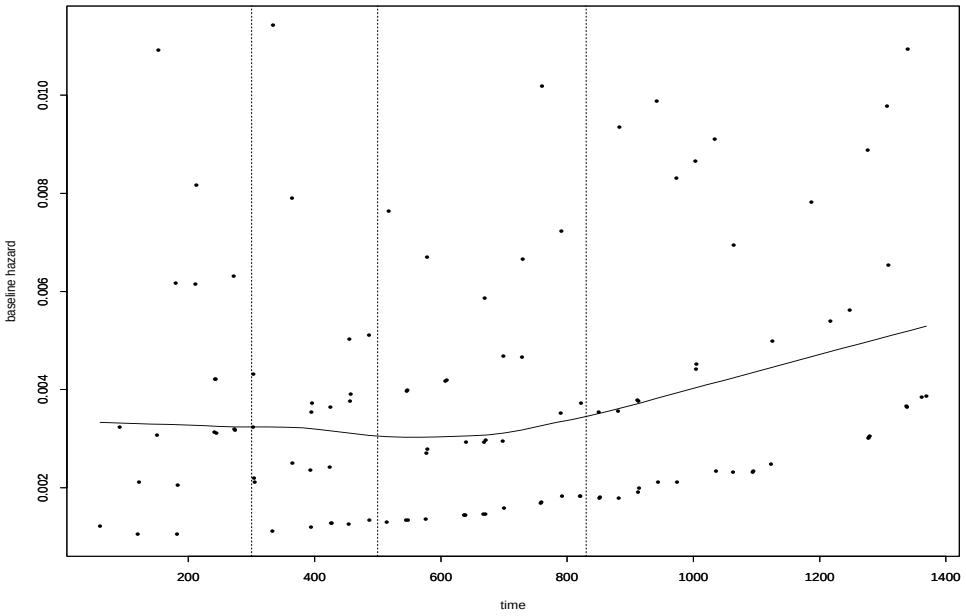

Figure 9: Estimated baseline hazards in Model (6.14) for the first four time intervals

one-year transition probabilities. However, a certain convenience of the piecewise exponential model is the easy inclusion of time dependency. Nonlinear baseline hazards are possible as well as the modeling of time dependent coefficients by simple inclusion of interactions with time. Yet, here again we have to note that the goodness of fit depends on the choice of the time intervals.

7 Summary and Discussion

The main issue of this paper was the inclusion of diagnoses as additional factors to model transition intensities. In particular the problem of multiple diagnoses was considered. 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 yield to 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. Since insurance companies calculate their premiums based on the estimated transition probabilities, one can assume that the quality of the premiums heavily depends on the underlying model. Thus, we suggest to use a model including diagnoses to achieve more reasonable calculations. 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. A piecewise exponential model might be an alternative to the semiparametric hazard model. However, blowing up the data is needed and there is no standard software available. Though the modeling of time dependency is an easy task in the piecewise exponential approach, this model has to be used with caution. The main problem is the appropriate choice of the underlying time intervals in between which the baseline hazard is assumed to be constant.

Acknowledgement

The first author was supported by Sonderforschungsbereich 386 *statistische Analyse Diskreter Strukturen* 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. *to appear in Insurance: Mathematics and Economics*.

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.

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

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics 36*, 299–305.

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

Laird, N. and D. Olivier (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association 76*, 231–240.

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

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.