Dynamic prediction in event history analysis

Grand, M.K.

3

Dynamic prediction of expected length of stay

In multi-state models the expected length of stay (ELOS) in a state is not a straightforward object to relate to covariates and the traditional approach has instead been to construct regression models for the transition intensities and calculate ELOS from these. The disadvantage of this approach is that the effect of covariates on the intensities is not easily translated into the effect on ELOS and it typically relies on the Markov assumption.

We propose to use pseudo-observations to construct regression models for ELOS, thereby allowing a direct interpretation of covariate effects, while at the same time avoiding the Markov assumption. For this approach, all we need is a non-parametric consistent estimator for ELOS. For every subject (and for every state of interest) a pseudo-observation is constructed and they are then used as outcome variables in the regression

model. We furthermore show how to construct longitudinal (pseudo-) data when combining the concept of pseudo-observations with landmarking. In doing so, covariates are allowed to be time-varying and we can investigate potential time-varying effects of the covariates. The models can be fitted using generalized estimating equations (GEE) and dependence between observations on the same subject are handled by applying the sandwich estimator. The method is illustrated using data from the US Health and Retirement Study where the impact of socioeconomic factors on ELOS in health and disability is explored. Finally, we investigate the performance of our approach under different degrees of left-truncation, non-Markovianity and right-censoring by means of simulation.

3.1 Introduction

Over the 20th century, from the 1920's onward, the life expectancy of humans has increased an incredible 2.5 years every decade[68]. The increase has been remarkably steady with no signs as yet that this trend is disappearing in the 21st century. Clearly this increased life expectancy will have a profound effect on modern society.

Among demographers there is a heavy debate, whether these additional life years are being spent in health or in disability. A distinction between life years spent in health and disability is crucial, both for the well-being of individuals and for health resources. An important question is then how background characteristics of individuals, such as gender and socio-economic status, and behavioral characteristics, like dietary habits and smoking, influence expected (remaining) life years spent in health and disability. In a paper studying the effects of these factors on healthy life expectancy and expected life in disability, Reuser et. al summarized the most striking behavioral effects as "Smoking kills, obesity disables"[75]. To contribute to this debate there is a need for methods to assess and model expected remaining life years in health and in disability for older people.

The typical approach used to address these questions, is to view this problem in the context of a multi-state model[72]. A reasonable multi-state model for the above healthy-

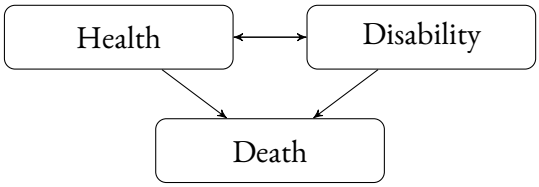

Figure 3.1: The reversible health-disability-death multi-state model.

disability debate is shown in Figure 3.1. It is an example of an illness-death model, with disability as the 'illness' state. The illness-death model of Figure 3.1 is reversible, since recovery from disability is possible. In general, a multi-state model is a stochastic process with outcomes in a finite space that represents the different stages in a subject's life course or disease/recovery process. Such multi-state models enable the estimation of the effect of explanatory factors on the transition intensities, but they do not give a direct quantification of the effect of these factors on the expected length of stay (ELOS) in a given state. Furthermore, these models typically rely on the assumption that the process is Markov.

We propose to use pseudo-observations to fit regression models that directly quantify the effect of explanatory variables on ELOS. Pseudo-observations has previously been proposed for regression on different multi-state objects such as the state occupation probabilities[9] and the restricted mean survival time[6]. The restricted ELOS is a generalization of the restricted mean survival to the multi-state setting. Here we will consider the restricted residual ELOS, which provides ELOS for a subject who has already survived up to a certain time-point, e.g. the expected remaining life in health and disability for a person of age 75. To incorporate this aspect we combine the concept of pseudo-observations with landmarking. Landmarking was introduced by[11], as a way to deal with time-dependent covariates in survival analysis, while avoiding immortal time bias. Pseudo-observations have previously been combined with landmarking for regression on the cumulative incidence function in a competing risks setting[66].

In the case of ordinary survival data Oakes and Dasu[67] proposed a proportional

mean residual life model, for the (unrestricted) residual mean survival time. Previous work on the restricted mean survival time has been focused on comparison of two groups adjusted for covariates[51,106,17,104]. Our method provides a straightforward way of fitting the proportional mean residual life model. At the same time, we allow for the more general setting with multiple states and the possibility of nonproportionality.

Section 3.2 describes the method; a formal definition of ELOS is given in section 3.2.1, section 3.2.2 defines the pseudo-observations in a general setting, and section 3.2.3 describes how to construct dynamic pseudo-observations for ELOS. These pseudo-observations are then used to construct direct regression models for ELOS in section 3.2.4. To illustrate the method we apply it to data from the US Health and Retirement Study described in section 3.3. In section 3.4 a simulation study is conducted to study the performance of the method under different degrees of incompleteness and non-Markovianity. Section 3.5 contains a discussion of the implications of the model assumptions, the performance of the method, as well as possible extensions and applications.

3.2 Method

3.2.1 Expected length of stay

A multi-state model is defined as a stochastic process $X(t)$ which has outcomes in a finite state space $\mathcal{K} = \{1, \ldots, K\}$. We are interested in how long time the process spends in a given state $h \in \mathcal{K}$, not necessarily consecutively, until a threshold τ, which will typically be taken to be large. The restricted length of stay in state h is defined by $\int_0^\tau I(X(t) = h)dt$. Hence, the restricted ELOS in state h is given by

$$e_h = E\left(\int_0^\tau I(X(t) = h)dt\right) = \int_0^\tau P\left(X(t) = h\right) dt \ ,$$

and it can be reformulated as the integral of the state occupation probability

$$P_h(t) = P(X(t) = h),$$

i.e. the probability that the process is in state h at time t. This detail is important for the construction of the pseudo-observations later on.

We can generalize the restricted ELOS to start at an arbitrary fixed time-point $s \geq 0$. This *residual restricted ELOS* is the expected length of stay in the time period from s to τ, conditional on the subjects being *alive* at time s, i.e. being in a non-absorbing state. Formally it is defined as

$$\begin{aligned} e_h(s) &= E\left(\int_s^\tau I(X(t)=h)dt \,\middle|\, X(s) \in \mathcal{A}\right) \\ &= \int_s^\tau P\left(X(t)=h \mid X(s) \in \mathcal{A}\right) dt \ , \end{aligned} \tag{3.1}$$

where $\mathcal{A}$ is the set of non-absorbing states in the model. Conceptually, conditioning on being alive is similar to *partly conditioning* as defined in [105,56]. The state h in $e_h(s)$ that indicates which state the process spends time in, will be referred to as the *target state*. In the remainder, the restricted residual ELOS will also be referred to as ELOS, unless confusion can arise.

3.2.2 Pseudo-observations

Assume that the data $(X_i, Z_i)_{i=1}^n$ consists of i.i.d. observations of an outcome X and covariates Z. The outcome may be the trajectory of a multi-state process $X(t)$. We are interested in constructing regression models based on a (possibly complex) function f of our outcome, i.e. our aim is to fit regression models of the form

$$\theta_i = E\left(f(X_i)|Z_i\right) = g^{-1}\left(\beta^\top Z_i\right) \ ,$$

for some known link function g, where $f(X_i)$ may be one-dimensional or a vector. As with most time-to-event data, some X_i are incompletely observed and hence so is

$f(X_i)$. Consider now the unconditional expectation, which is the parameter

$$\theta = E\left(f(X)\right) \ . \tag{3.2}$$

Assuming there exists a consistent estimator $\hat{\theta}$ of θ, the pseudo-observation for subject i is defined as

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{(-i)} \ , \tag{3.3}$$

where $\hat{\theta}$ is the estimate based on the entire data set and $\hat{\theta}^{(-i)}$ is the estimate where subject i has been removed. The pseudo-observation $\hat{\theta}_i$ can be seen as the contribution of subject i to the estimate of θ. The idea is to use the pseudo-observations as outcome, instead of $f(X_i)$, to fit a generalized linear regression model using generalized estimating equations (GEE)[58]. GEE are employed, since we want to avoid making assumptions about the full distribution of the outcome, because we are using pseudo-observations, and at the same time we want to account for possible dependence between observations on the same subject. The assumptions underlying the GEE are

1. Observations between subjects are independent.

2. The conditional mean depends linearly on the covariates through a known link-function g
$$E\left(f(X_i)|Z_i\right) = \theta_i \ , \ g(\theta_i) = \beta^\top Z_i \ .$$

Furthermore a structure for the working covariance matrix V_i of the pseudo-observations should be specified. The first assumption is satisfied as the pseudo-observations are approximately independent[88,10].

A consistent estimate of β can be obtained as the solution to the estimating equations

$$U(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \theta_i}{\partial \beta}\right)^\top V_i^{-1} \left(\hat{\theta}_i - \theta_i\right) = 0 \ .$$

Notice that pseudo-observations are used also for those individuals where the outcome was completely observed. The covariance matrix can be estimated by the sandwich estimator

$$\widehat{\text{Cov}}(\hat{\beta}) = \hat{\Sigma}^{-1}\widehat{\text{Cov}}(U(\beta))\hat{\Sigma}^{-1} \ ,$$

where

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\theta_i}{\partial\beta}\right)^{\top} V_i^{-1}\left(\frac{\partial\theta_i}{\partial\beta}\right) \quad \text{and} \quad \widehat{\text{Cov}}(U(\beta)) = \frac{1}{n}\sum_{i=1}^{n} U_i(\hat{\beta})U_i(\hat{\beta})^{\top} \ .$$

The choice of working covariance matrix influences the efficiency of the estimator $\hat{\beta}$. Nevertheless, if assumption 1 and 2 are satisfied and $\hat{\theta}$ is a consistent estimator of θ, then $\hat{\beta}$ is consistent for any suitable choice of working covariance matrix. So far only Graw et al[39] has provided proofs regarding the asymptotic properties of the pseudo-observation based regression in the setting of cumulative incidences for a competing risks.

3.2.3 Dynamic pseudo-observations

This section describes how to create dynamic pseudo-observations for ELOS, by combining the concept of pseudo-observations with landmarking. In our setting X is the multi-state process $X(t)$ and the parameter θ is formed by $e_h(s)$ from equation (3.1). Note that $e_h(s)$ is indeed the expectation of a complex function of the data, as in (3.2). The analysis can be limited to specific states of interest. Let therefore $\mathcal{H} \subseteq \mathcal{K}$ denote this set and let H be the cardinality of $\mathcal{H}$. To construct the pseudo-observations we need consistent estimators of $e_h(s)$, $h \in \mathcal{H}$. The first step is to find a consistent estimator for the state occupation probabilities. Let

$$P_h^s(t) = P(X(t) = h \mid X(s) \in \mathcal{A})$$

denote the state h occupation probability at time t conditional on being at alive at time s. It can be estimated by the non-parametric Aalen-Johansen estimator $\hat{P}_h^s(t)$[7] using

landmarking[11,92]. The estimate is based on the sub-sample of subjects alive at time s. Let $n(s)$ denote the number of subjects alive and at risk at time s and let $Y_g(s+)$ be the number of those subjects that occupied state $g \in \mathcal{A}$. $\hat{P}_h^s(t)$ is then the weighted sum over the estimated transition probabilities $\hat{P}_{gh}(s,t)$, where the weights are equal to the corresponding empirical initial occupation probabilities $\frac{Y_g(s+)}{n(s)}$. The non-parametric Aalen-Johansen estimator of the state occupation probabilities is consistent under independent right-censoring, even when the process is non-Markovian and in the presence of left-truncation[23,62]. The second step is then to find the area under the Aalen-Johansen estimator. Let $s = t_0 < t_1 < \cdots < t_J \leq t_{J+1} = \tau$ be the ordered transition times pooled over all transitions. A consistent estimator for $e_h(s)$ in (3.1) is then given by

$$\hat{e}_h(s) = \sum_{j=0}^{J} \hat{P}_h^s(t_j)(t_{j+1} - t_j) \ . \tag{3.4}$$

Inserting the estimator of (3.4) into the equation (3.3) gives rise to the pseudo-observations

$$\hat{e}_{ih}(s) = n(s)\hat{e}_h(s) - (n(s) - 1)\, \hat{e}_h^{(-i)}(s) \ ,$$

one for every subject i at risk at time s and for every state h of interest. Pseudo-observations are only created for subjects at risk, but left-truncated individuals still contribute to the estimate of the pseudo-observations through $\hat{e}_h(s)$. In principle it is possible to create pseudo-observations for left-truncated individuals, but typically the value of the (time-dependent) covariates at times will be unknown for such subjects. An alternative way of estimating the state occupation probability is to base $\hat{P}_h^s(t)$ only on the sub-sample of people alive and at risk at time s. In this way left-truncated individuals would be completely discarded in the construction of the pseudo-observations. We call this latter approach the *strict* approach and the former approach (where subjects not yet at risk are included in the calculation of the state occupation probabilities) the *non-strict* approach.

An interesting feature is that had the data been completely observed, the pseudo-observations would be the actual observed length of stay of the subjects.

3.2.4 Regression models

In this section we describe how the dynamic pseudo-observations may be used to construct direct regression models for ELOS. Section 3.2.4 describes the situation for one fixed landmark. In section 3.2.4 several landmarks are selected, for the purpose of modeling the development of ELOS over the landmark time. To this end a so-called super model is employed to construct one regression model.

Models for a fixed landmark

Let s be a fixed landmark and recall H to be the cardinality of $\mathcal{H}$ the states of interest. For every individual i at risk at time s and every state h of interest, a pseudo-observation $\hat{e}_{ih}(s)$ is created as described in section 3.2.3. Hence, each individual at risk has H pseudo-observations which may be dependent.

For a time-dependent covariate $Z(t)$ the value fixed at the landmark $Z(s)$ is used as a time-fixed covariate[11,92,95]. A covariate of special interest is $X(s) = g$, which is the state that the process occupies at time s and it will be referred to as the *current state*.

It is natural to assume that the effect of some covariates will differ according to the target state h, e.g. the effect of BMI or smoking is different for ELOS in health and ELOS in disability. We therefore introduce *target-specific* covariates. The idea is similar to that of transition-specific covariates in regression models for the transition intensities in multi-states models[4]. Let $Z_{ih}(s)$ denote the p dimensional target-specific covariate vector for subject i fixed at time s. Define the conditional mean

$$e_{ih}(s) = E\left(\int_s^\tau I(X(t) = h)dt \,\middle|\, X(s) \in \mathcal{A}, Z_{ih}(s)\right) \ .$$

We assume that the conditional mean has the structure

$$g(e_{ih}(s)) = \beta(s)^\top Z_{ih}(s) \ , \tag{3.5}$$

where $\beta(s)$ is a vector of p parameters. The covariate vector may include 1 to allow for target-specific intercepts. The current state $X(s)$ may also be included as a covariate. In some situations covariates may also interact with the current states, e.g. being disabled at time s could modify the effect of BMI on time spent in disability. These interactions will be a part of the p covariates contained in $Z_{ih}(s)$.

The model in (3.5) can be fitted by GEE using a suitable working covariance matrix. The following section shows how the concept can be extended from one to several landmarks. The GEE for a fixed landmark therefore follows from the more general GEE case with several landmarks.

Supermodels using several landmarks

Let $\mathcal{S} = \{s_1, \cdots, s_D\}$ be a set of fixed landmark time points. To study the development of ELOS over time we could repeat the fixed landmark method to make D separate regression models. It is, however, appealing to think that covariate effects change smoothly over s, and the (pseudo-) data could instead be considered as longitudinal data[95,66]. Let $\mathcal{S}_i \subseteq \mathcal{S}$ denote the set of the D_i landmarks where subject i was at risk. For every subject i, every $s_d \in \mathcal{S}_i$ and every $h \in \mathcal{H}$ we create a pseudo-observation $\hat{e}_{ih}(s_d)$. Subject i therefore has $H \cdot D_i$ pseudo-observations, which are stacked into the vector $\hat{e}_i$. As before, with one fixed landmark, we make use of target-specific covariates to handle interactions between covariates and target state. In addition to this there is also the new possibility of covariates interacting with landmark time, i.e. effects may be time-varying.

The conditional mean is assumed to follow 3.5, where $\beta(s)$ is no longer a vector of parameters, but a q vector of suitable smooth function of $s \in [s_1, s_D]$ that we have to

specify. The lth element of $\beta(s)$ is

$$\beta_l(s) = \beta_l^\top b_l(s) \ ,$$

where b_l is a vector of fixed basis functions, and β_l a vector of parameters. Let $B(s)$ denote the $p \times q$ matrix of basis functions and let β denote the stacked vector of b_l's. It then follows that $\beta(s) = B(s)\beta$. The conditional mean in equation (3.5) can be rewritten in terms of the covariate $Z_{ih}^*(s) = B(s)^\top Z_{ih}(s)$,

$$g(e_{ih}(s)) = \beta(s)^\top Z_{ih}(s) = (B(s)\beta))^\top Z_{ih}(s) = \beta^\top Z_{ih}^*(s) \ .$$

The estimating equations of the super model can be formulated as

$$U(\beta) = \sum_{i=1}^{n} \left(\frac{\partial e_i}{\partial \beta} \right)^\top V_i^{-1} (\hat{e}_i - e_i) = 0 \ , \tag{3.6}$$

where $\hat{e}_i = [\hat{e}_{ih}(s_d)]_{h,d}$ is the stacked vector of all pseudo-observations for subject i. The solution to the estimating equations $\hat{\beta}(s)$ is a consistent estimator of β, provided that

1. The estimator of $e_h(s)$ is consistent.

2. The regression model is correctly specified.

Furthermore, it is necessary to assume working independence between observations at different landmarks. Kurland and Heagerty[56] point out that for partly conditional models, i.e. models such as ours where we condition on being alive ($X(t) \in \mathcal{A}$), the number of observations on an individual is stochastic. If the working correlation matrix would be anything else than a diagonal matrix, the inverse variance matrix V_i^{-1} would depend on the cluster size. Since V_i^{-1} no longer is a known quantity conditional on the covariates, this may destroy the unbiasedness of the estimating equations in (3.6).

Covariates effect may be tested by a Wald test, in the same fashion as with standard GEE.

3.3 Application

To illustrate the method and to show how it can be used to contribute to the health-disability debate, data from the Asset and Health Dynamics Among the Oldest Old (AHEAD), now part of the wider US Health and Retirement Study (HRS), will be used[49]. The HRS has been collecting data since 1992, including health and socio-economic status on a population of elderly. Of these we selected a subpopulation of people of age 75 and older. The time scale is age. Table 3.1 shows the frequency in the HRS data of the time-fixed covariates considered in the illustration (body-mass index (BMI) and smoking status are assessed at entry into the study). Disability status is defined according to the Basic Activities of Daily Living (ADL) scale[52], which includes items for walking, bathing, dressing, toileting and feeding. A subject is defined to be ADL disabled here if he/she responds "with difficulty" for at least one of the ADL items.

In the following we will study the dynamics of disability and recovery in the health-disability-death multi-state model of Figure 3.1. In this data, for a total of 4026 subjects, 1929 transitions from healthy to ADL disabled occurred and 679 recoveries (transitions from ADL disability to healthy). A total of 1982 deaths were observed, 916 from the healthy state and 1066 from ADL disability. More details about the results and the code used for the analysis can be found in the Supplementary Material.

3.3.1 Fixed landmark model

We begin with considering a fixed landmark model for the age of 75, to investigate the effect of the covariates on the ELOS in health ($h = 1$) and disability ($h = 2$). Pseudo-observations were created for these two states, with $\tau = 110$, using the `mstate` package[26] in `R` to estimate ELOS.

Table 3.1: Baseline covariates in the HRS study.

Covariate	n	(%)
Gender		
Male	1561	(39%)
Female	2465	(61%)
Education		
Less than high school	1732	(43%)
High school	1211	(30%)
Some college	1083	(27%)
BMI (kg/m^2)		
≤ 25	2241	(56%)
$25 - 30$	1386	(34%)
> 30	389	(10%)
Missing	10	
Smoking		
Never	1996	(50%)
Past	1680	(42%)
Current	322	(8%)
Missing	28	(1%)

It is natural to assume that the effect of the covariates on expected healthy life will differ from the effect on expected life in disability, in other words that the covariates will interact with the target state. We therefore make use of target-specific covariates. Furthermore, the effect of covariates may not only differ by target-state, but also by current state. We therefore fit a model where all the target-specific covariates also interact with the current state. This amounts to estimating separate covariate effects for each of the four combinations of target state and current state. The link function is assumed to be the identity function and a working independence covariance matrix is applied. The model was fitted using the `geepack` package [45] in R.

Table 3.2 shows the estimated regression parameters of the model, with robust standard errors and 95% confidence intervals. It is presented in terms of the target-specific co-

variates conditional on the current state. We see that females who are healthy at age 75, with a high school education, a BMI<25 and who never smoked, are expected to have 10.057 more years in health, and 3.587 more in disability. Corresponding males spent less time in health than the females, and even less time in disability. Interestingly, both high BMI and current smoking are associated with less time spent in health, but the effect on time spent in disability is quite different: negative for current smoking, positive for high BMI. This supports the claim of "smoking kills, obesity disables"[75]. More parsimonious models could have been found, e.g. by removing the non-significant covariates by current state interactions, but this was not pursued at this stage.

The procedure was repeated for a whole set of landmarks from the age of 75 to 95 at every 2.5 years. Figure 3.2 illustrates the change of ELOS with age for the baseline characteristics, i.e. the intercepts of all the landmark models. Not surprisingly, the ELOS is declining in all four groups as people become older. The drop seems to be particularly fast for time spent in health.

Figure 3.3 shows the covariate effects on time spent in health given healthy at age s. It is interesting to see that the effect of current smoking seems to decline over time. Naturally this is also forced by the fact that there is less time to spend, but it may also be that individuals who live to an old age are especially robust and therefore less susceptible to die from smoking.

These plots motivate the idea that the changes over age could be reasonably modeled with linear functions for the covariates and quadratic functions for the intercepts. This can be achieved by employing a landmark super model.

3.3.2 Super model

In this section we applied a super model to the stacked pseudo-data for all the landmarks from age 75 to 95 at every 2.5 years. Landmark time was rescaled as $\tilde{s} = (s - 75)/20$, thus taking values between 0 and 1. The only time-varying covariate included in the model was the current state, which was fixed at its current value at time s. Quadratic interactions between landmark time and the target- and current state were included.

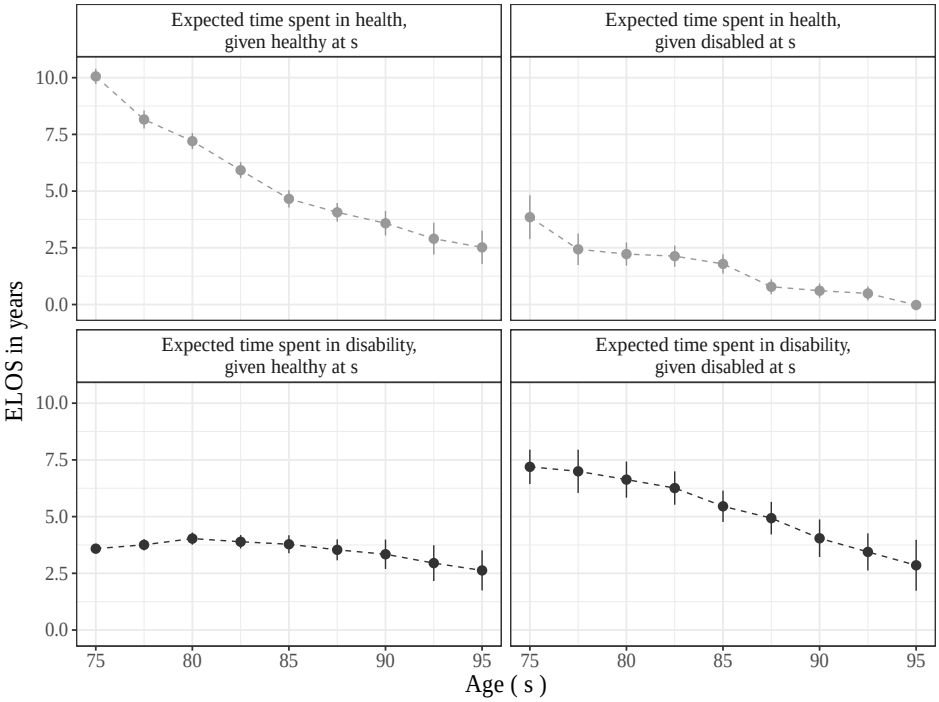

Figure 3.2: Estimated ELOS for the baseline characteristics (Intercepts) according to the fixed landmark models with 95% point-wise confidence intervals. The dashed line only serves as an visual aid.

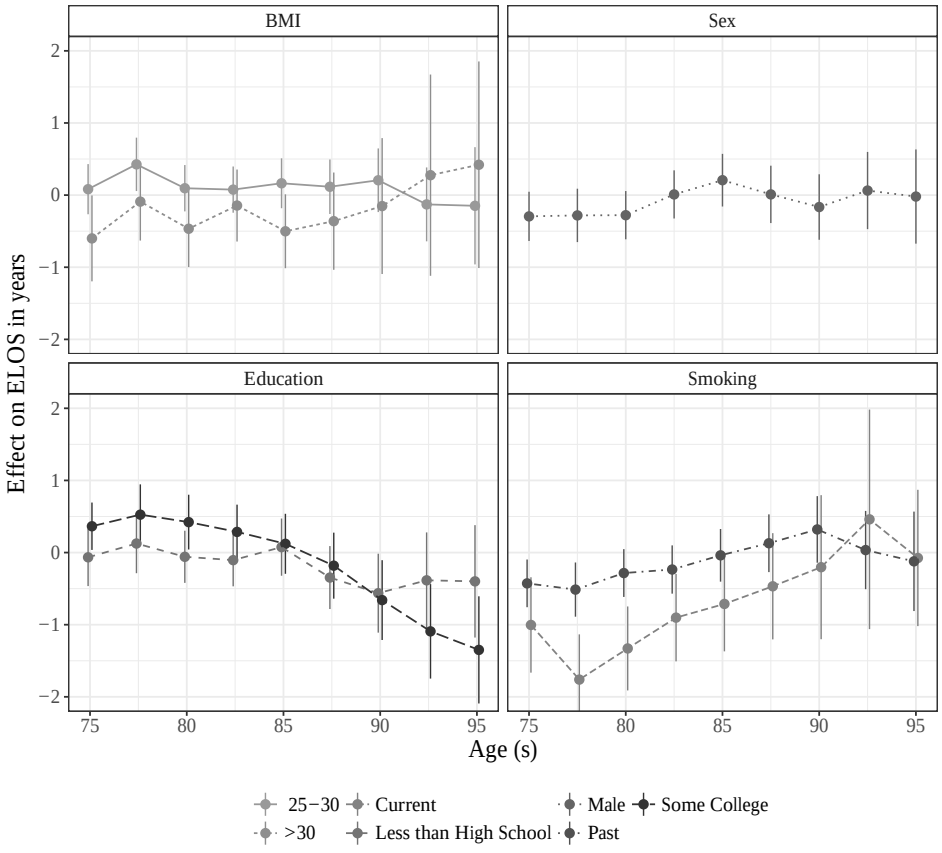

Figure 3.3: Estimated covariate effect on ELOS in health, given healthy at age s, according to the fixed landmark models with 95% point-wise confidence intervals.

Table 3.2: Estimates of the fixed landmark model for the age of 75.

Covariate	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)
Target state : Health		Current state : Health			Current state : Disability	
Intercept	10.057	0.172	(9.775, 10.339)	3.851	0.493	(3.039, 4.663)
Sex						
Male	−0.295	0.174	(−0.581, −0.008)	−0.442	0.431	(−1.150, 0.266)
BMI						
$25 - 30$	0.082	0.177	(−0.210, 0.374)	−0.329	0.467	(−1.097, 0.439)
> 30	−0.600	0.304	(−1.100, −0.101)	0.455	0.500	(−0.367, 1.277)
Education						
Less than high school	−0.066	0.204	(−0.401, 0.269)	−0.711	0.430	(−1.419, −0.003)
Some college	0.365	0.168	(0.089, 0.641)	0.305	0.617	(−0.710, 1.320)
Smoking						
Past	−0.427	0.169	(−0.704, −0.149)	−0.321	0.420	(−0.371, −1.012)
Current	−1.004	0.337	(−1.558, −0.449)	−0.353	0.520	(−0.502, 1.209)
Target state : Disability		Current state : Health			Current state : Disability	
Intercept	3.587	0.098	(3.426, 3.749)	7.194	0.386	(6.559, 7.829)
Sex						
Male	−0.171	0.085	(−0.310, −0.031)	−0.453	0.428	(−1.157, 0.250)
BMI						
$25 - 30$	−0.114	0.089	(−0.033, 0.260)	0.583	0.475	(−0.199, 1.364)
> 30	0.055	0.165	(−0.216, 0.326)	0.434	0.406	(−0.233, 1.102)
Education						
Less than high school	−0.021	0.103	(−0.148, 0.190)	−0.436	0.460	(−1.192, 0.320)
Some college	0.059	0.091	(−0.090, −0.208)	0.198	0.476	(−0.585, 0.981)
Smoking						
Past	−0.072	0.091	(−0.221, −0.077)	−0.520	0.388	(−1.158, 0.118)
Current	−0.293	0.159	(−0.554, −0.031)	−0.631	0.395	(−1.281, 0.031)

Covariate effects were assumed to vary linearly over landmark time and differ according to both target- and current state. The link function was assumed to be the identity function and an independence working covariance matrix was employed.

The results of the analysis are shown in Table 3.3 and 3.4 with the estimates of the regression parameters, robust standard errors and corresponding 95% confidence intervals. The table shows the target-specific effects conditional on the current state. The *constant* part of the super model corresponds to the effect on ELOS at age 75, and it is therefore comparable to the fixed landmark model in Table 3.2. In the super model a female, with low BMI, high school education, who never smoked is expected to spend 9.693 years in health from the age of 75. This is comparable to the fixed landmark model at age 75 in the previous section. The *landmark* part of the super model shows the estimated change of the effects over s. Since the intercept is assumed to change as a quadratic function over time, a similar person of age 85 ($\tilde{s} = 0.5$) is expected to spend $(9.693 - 11.590 \cdot 0.5 + 4.129 \cdot 0.5^2 =) \, 4.930$ years in health. If this person instead had been disabled at age 85 she would be expected to spend $(3.745 - 5.827 \cdot 0.5 + 2.354 \cdot 0.5^2 =) \, 1.42$ years in health.

Figure 3.4 illustrates the overall impact of gender on ELOS and how it interacts with the target- and current state over time. It shows males and females with low BMI, a high school education, who never smoked. The upper left graph show ELOS in health conditional on being healthy at the current age s. At age 75, both males and females are expected to live around 10 more years in health. On the other hand, the lower left graph shows that the females are expected to live longer in disability than the males. This difference is even larger for subjects that were disabled at time s (the lower right graph).

3.4 Simulations

The performance of the method under different degrees of right-censoring, left-truncation and non-Markovianity was investigated through simulations by comparing the true ef-

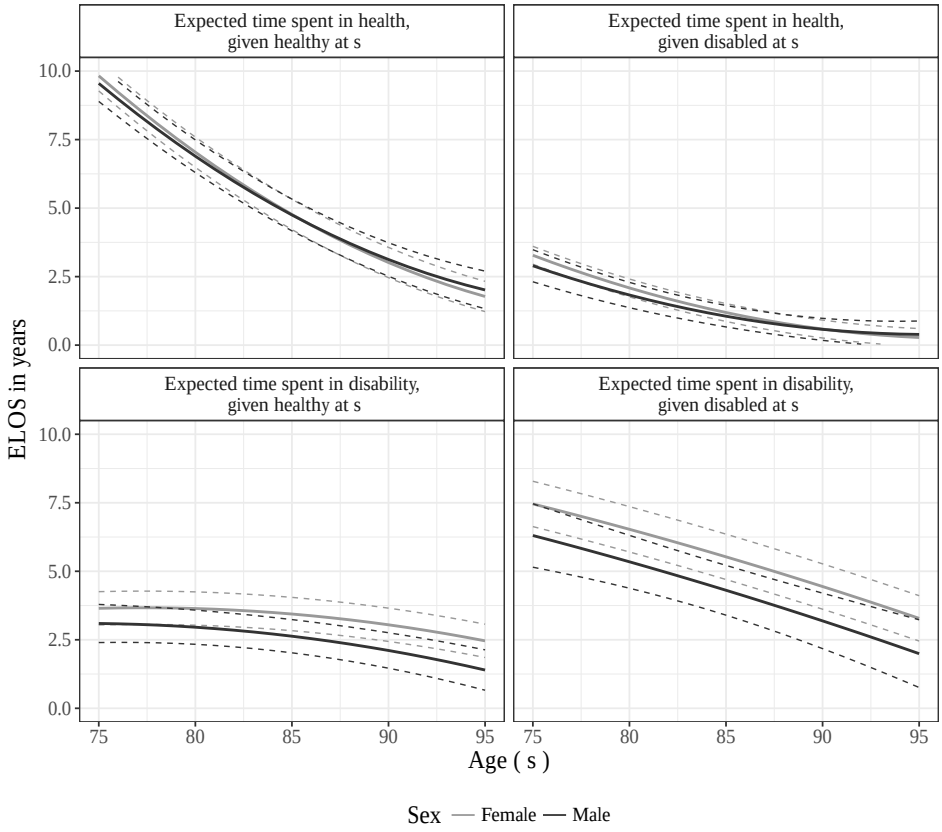

Figure 3.4: Estimated ELOS over time in health and disability for males and females, given healthy or disabled, with point-wise 95% confidence intervals.

Table 3.3: Estimates of the landmark super model for the target state health.

Covariate	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)
Target state : Health		Current state : Health			Current state : Disability	
Constant						
Intercept	9.693	0.195	(9.312, 10.075)	3.745	0.281	(3.195, 4.296)
Sex						
Male	−0.276	0.171	(−0.611, 0.06)	−0.382	0.242	(−0.857, 0.092)
BMI						
$25 - 30$	0.240	0.166	(−0.085, 0.565)	0.550	0.242	(0.076, 1.023)
> 30	−0.355	0.274	(−0.891, 0.181)	0.212	0.289	(−0.355, 0.779)
Education						
Less than high school	0.133	0.191	(−0.241, 0.508)	−0.467	0.244	(−0.945, 0.012)
Some college	0.795	0.194	(0.415, 1.175)	0.173	0.297	(−0.409, 0.754)
Smoking						
Past	−0.537	0.174	(−0.878, −0.196)	−0.163	0.249	(−0.651, 0.325)
Current	−1.710	0.294	(−2.286, −1.133)	−1.023	0.338	(−1.685, −0.361)
Landmark						
Intercept						
$\tilde{s}$	−11.590	0.589	(−12.745, −10.435)	−5.827	0.717	(−7.232, −4.422)
$\tilde{s}^2$	4.129	0.486	(3.177, 5.08)	2.354	0.517	(1.341, 3.367)
Sex						
Male $\cdot \tilde{s}$	0.513	0.349	(−0.171, 1.197)	0.495	0.379	(−0.248, 1.238)
BMI						
$25 - 30 \cdot \tilde{s}$	−0.219	0.335	(−0.876, 0.437)	−0.663	0.351	(−1.351, 0.025)
$> 30 \cdot \tilde{s}$	0.146	0.621	(−1.07, 1.363)	−0.504	0.479	(−1.444, 0.435)
Education						
Less than high school $\cdot \tilde{s}$	−0.590	0.400	(−1.374, 0.194)	0.471	0.364	(−0.241, 1.184)
Some college $\cdot \tilde{s}$	−1.651	0.397	(−2.429, −0.872)	−0.145	0.438	(−1.003, 0.714)
Smoking						
Past $\cdot \tilde{s}$	0.928	0.360	(0.223, 1.633)	0.265	0.372	(−0.463, 0.993)
Current $\cdot \tilde{s}$	1.931	0.641	(0.675, 3.187)	1.579	0.643	(0.319, 2.839)

Table 3.4: Estimates of the landmark super model for the target state disability.

Covariate	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)	$\hat{\beta}$	SE($\hat{\beta}$)	(CI)
Target state : Disability	Current state : Health			Current state : Disability		
Constant						
Intercept	3.848	0.160	(3.535, 4.161)	7.694	0.461	(6.791, 8.598)
Sex						
Male	−0.554	0.143	(−0.835, −0.273)	−1.152	0.406	(−1.948, −0.356)
BMI						
$25 - 30$	0.000	0.139	(−0.273, 0.273)	0.227	0.391	(−0.539, 0.993)
> 30	0.504	0.214	(0.085, 0.923)	0.442	0.509	(−0.555, 1.44)
Education						
Less than high school	−0.200	0.153	(−0.499, 0.099)	−0.237	0.409	(−1.037, 0.564)
Some college	−0.298	0.161	(−0.614, 0.018)	−0.560	0.462	(−1.465, 0.344)
Smoking						
Past	0.164	0.146	(−0.122, 0.45)	−0.799	0.409	(−1.6, 0.002)
Current	0.001	0.222	(−0.435, 0.437)	−0.843	0.648	(−2.114, 0.428)
Landmark						
Intercept						
$\tilde{s}$	0.542	0.623	(−0.679, 1.762)	−4.146	1.283	(−6.661, −1.631)
$\tilde{s}^2$	−1.541	0.555	(−2.629, −0.453)	−0.630	1.023	(−2.634, 1.375)
Sex						
Male $\cdot \tilde{s}$	−0.026	0.019	(−0.064, 0.013)	−0.007	0.040	(−0.085, 0.071)
BMI						
$25 - 30 \cdot \tilde{s}$	1.137	0.388	(0.376, 1.899)	1.042	0.720	(−0.37, 2.454)
$> 30 \cdot \tilde{s}$	−0.735	0.627	(−1.965, 0.494)	0.415	1.008	(−1.561, 2.391)
Education						
Less than high school $\cdot \tilde{s}$	−0.188	0.394	(−0.961, 0.585)	0.599	0.754	(−0.88, 2.078)
Some college $\cdot \tilde{s}$	0.795	0.466	(−0.118, 1.709)	1.482	0.862	(−0.207, 3.172)
Smoking						
Past $\cdot \tilde{s}$	−0.559	0.390	(−1.323, 0.205)	0.675	0.776	(−0.847, 2.196)
Current $\cdot \tilde{s}$	−0.679	0.649	(−1.951, 0.592)	0.823	1.430	(−1.981, 3.626)

fect of one covariate with the estimates. In addition, the approach was compared to estimates based on regression models for the transition intensities, which will be referred to as the multi-state model. In general, it is not possible to compare to alternative methods, as no other methods are available for direct regression on ELOS. It was however possible to make a direct comparison in this simulations study as the model only includes one categorical covariate. The setup of the simulation study is inspired by the HRS data in Section 3.3 and the generated data follows the multi-state model of Figure 3.1. In the following section the setup of the simulation study is described in brief and the last section describes the results. Additional results can be found in the Supplementary Material.

3.4.1 Setup

Simulating from a multi-state model

The data for the simulation study was generated by assuming constant transition intensities. One categorial covariate Z with two levels $\{0, 1\}$ was considered and subjects with $Z = 1$ would have lower transition intensities into death than those with $Z = 0$, which may illustrate the situation of non-smokers versus smokers. Non-Markovian data was generated by including individual frailties on the transition to disability. The intuition is that an individual who is currently healthy, but has a history of being disabled, would more likely be a frail individual, which by construction has a higher transition intensity to disability. The probability of making a transition therefore depends on the process history, which is a violation of the Markov assumption.

For both the Markov and non-Markov setup a total of 1000 data sets, each with 2000 individuals, were simulated. In each data set one half of the subjects had $Z = 0$ and the other half had $Z = 1$. To mimic the setup of the HRS data, where subjects were followed from approximately age 75 onwards, we simply added 75 (years) to all simulated time values of the complete data. Subjects were followed for 35 years until $\tau = 110$. Six different scenarios of random right-censoring and left-truncation were

subsequently imposed on the complete data.

Estimated- and true parameters

This section describes the models that were fitted to the simulated data using either the pseudo-observations or the multi-state approach.

For the pseudo-observation approach landmarks from 75 to 105 at every 2.5 were selected and the corresponding pseudo-data created. The pseudo-data was then analyzed using either fixed landmark models or a super model. Let $Z_{gh}(s)$ denote the target-specific covariates of Z, which also include interactions with current state g. In the fixed landmark models the mean was assumed to be

$$E\left(\int_s^\tau I(X(t)=h)dt \,\middle|\, X(s)=g\in\{1,2\}, Z_{gh}(s)\right) = \alpha(s)+\beta(s)Z_{gh}(s)\ , \tag{3.7}$$

where $\alpha(s)$ and $\beta(s)$ are parameters. In the super model quadratic functions were assumed for both the intercept and the effect of $Z_{gh}(s)$ over s,

$$\begin{aligned} E\left(\int_s^\tau I(X(t)=h)dt \,\middle|\, X(s)=g\in\{1,2\}, Z_{gh}(s)\right) = \\ \alpha_1+\alpha_2 s+\alpha_3 s^2+(\beta_1+\beta_2 s+\beta_3 s^2)Z_{gh}(s)\ , \end{aligned} \tag{3.8}$$

where the α's and β's are parameters and $\beta(s) = \beta_1+\beta_2 s+\beta_3 s^2$ is the effect of the covariate at time s. All models were fitted with a working independence covariance matrix.

In the multi-state approach the transition intensity from state g to h was assumed to be

$$\lambda_{gh}(t|Z) = \lambda_{0,gh}(t)\exp(\beta Z)\ , \tag{3.9}$$

where $\lambda_{0,gh}(t)$ is the unspecified baseline intensity and β is the transition specific covariate effect. Estimates of the transition intensities was then used for obtaining estimates of the transition probabilities $P_{gh}(s,t)$. Finally, the area under the estimated transition

probabilities was used as estimates of the conditional mean of interest.

The true effect of the covariate was approximated, since the simulation setup does not allow explicit analytical expressions, unless in the Markov case. The true effect $\bar{\beta}(s)$ of $Z_{gh}(s)$ may depend on both time s, the current state g and the target state h. The true value was therefore approximated by averaging the length of stay $e_h(s)$, over the 1000 complete data sets (before censoring or truncation was applied). This was done separately for each landmark s and each current state $g = 1, 2$.

Comparison

The estimated effects were compared with the true effects $\bar{\beta}(s)$ by calculating the bias, root mean square error (RMSE) and coverage as measures of performance. Let $\hat{\beta}(s)$ denote the estimated effect of $Z_{gh}(s)$ in either the fixed landmark models or the super model. For a given s and covariate $Z_{gh}(s)$ the bias and RMSE are defined as

$$\text{bias} = E\left(\hat{\beta}(s) - \bar{\beta}(s)\right) \;\; \& \;\; \text{RMSE} = \sqrt{E\left((\hat{\beta}(s) - \bar{\beta}(s))^2\right)} \; .$$

The coverage was estimated by the proportion of simulated data sets from which the estimated confidence interval contained the true value. All three measures may depend on the target- and the current state, but this is suppressed in the notation.

3.4.2 Results

This section describes the results of the simulation study, where the performance of the method is evaluated. Figure 3.5 shows the results of the fixed landmark models using non-strict and strict pseudo-observations, as well as the multi-state model approach. The models was fitted on non-Markov data with 20% truncation and 10% censoring. The estimated effect of the covariate $Z_{gh}(s)$ for each of the 1000 data sets are depicted with boxplots. The top left graph shows the results for $Z_{11}(s)$, i.e the effect of the covariate on ELOS in health, given healthy at time s. The true value $\bar{\beta}(s)$ is denoted

with a white diamond. The estimates of the strict approach are virtually unbiased, and the variability is acceptable, although occasionally negative estimates for ELOS are obtained. The non-strict approach show some bias, especially at earlier landmarks, for which the degree of left-truncation is substantial. The bias disappears for later landmark time points. Interestingly (details not shown), RMSE for the strict and non-strict approaches are comparable; coverage is very good for the strict approach and increasing from 94% for earlier landmarks to 95% for later landmarks. The multi-state model is biased, which is due to the non-Markovian nature of the data.

It is clear that the strict approach is favorable in this situation. However, for a small data set with some truncation the non-strict version will be more stable, which is supported by the simulation study. This is due to the fact that the non-strict version borrows information from more individuals, without truncation the two will coincide. As expected, lower levels of censoring and truncation lead to smaller bias for the non-strict approach. When applied to the Markov data there was no change in performance for the pseudo-observations and the multi-state model performed reasonably. The super model showed very similar results as the fixed landmark models and is therefore not shown.

3.5 Discussion

In this paper we explored the use of pseudo-observations in combination with landmarking to construct direct regression models for the restricted residual expected length of stay (ELOS) in multi-state models. The traditional approach to model ELOS is to fit regression models for all the transition hazards. The estimated covariate effects in these model, however, do not translate directly into the effect of the covariates on ELOS. Our method conveniently avoids the need to fit this kind of, possibly complicated, multi-state models and the estimated effect of the covariates have a direct interpretation. In combination with landmarking it furthermore allows for time-dependent covariates and time-varying effects.

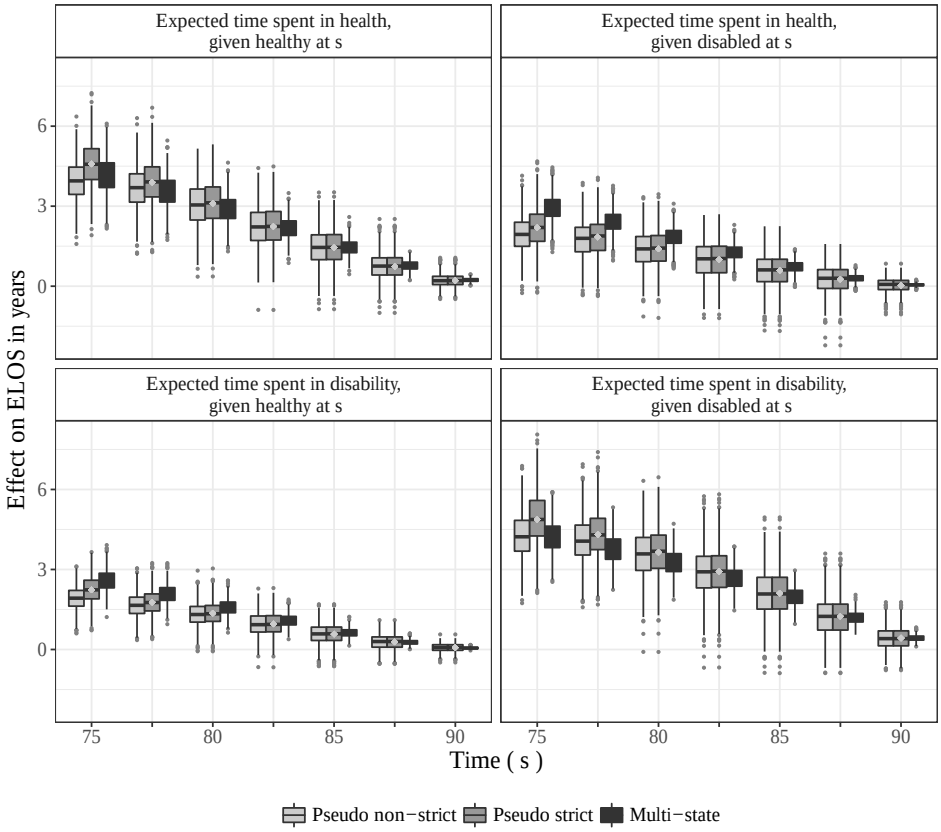

Figure 3.5: Estimated effect of $Z_{gh}(s)$ in the 1000 non-Markov data sets with 10% censoring and 20% truncation, using the fixed landmark models with non-strict and strict pseudo-observations and the multi-state model respectively. The true value is denoted with a white diamond.

The proportional mean residual life model[67] is a special case of the super models we consider in this paper, namely for the alive-death multi-state model, with log-link function and covariate effects which do not vary over landmark time. The pseudo-observation approach thus provides a straightforward way of fitting the proportional mean residual life model, and extends it, both by allowing for multiple states and other choices of link function. Our choice of identity link function in the application has the disadvantage of not guaranteeing positive ELOS, but in our view the advantage of directly interpreting the regression coefficients as adding/subtracting life years outweighs this disadvantage. As in any situation, it would be appropriate to make a goodness-of-fit assessment. It is however an open question how this could be achieved when pseudo-observations are used and we did not pursue this.

We showed how the method can be applied in a reversible illness-death model to estimate the direct effects of socio-economic factors on ELOS in health and disability for a population of elderly. The fixed landmark models had comparable standard errors with the super model, but this model may be too rich. Although we did not pursue it here, the method allows for model selection, and a more parsimonious model may have been found for the super model. In general better efficiency, in terms of improved standard errors, may be obtained by using a super model, at the possible expense of bias introduced by incorrect specification of such a super model. Further improvements in terms of efficiency may be achieved by selecting an appropriate working covariance matrix, as long as observations between landmarks are taken as independent in the working covariance matrix.

We conjecture that the approach yields consistent estimates provided that the estimator for the state-occupation probabilities is consistent and the regression model is correctly specified. We have chosen to use the Aalen-Johansen estimator, which in the presence of independent right-censoring is consistent even under non-Markovianity[23] and left-truncation[62]. Depending on the setting alternatives to the Aalen-Johansen estimator could be considered in order to obtain consistency and to improve efficiency. E.g. in the situation with state dependent censoring, Datta and Satten[24] proposed an estim-

ator for the state occupation probabilities under non-Markovianity, and others[14,25] have also considered different alternatives and settings. At present, as far as we know, there has been no work related to pseudo-observations under left-truncation. In our motivating example with the HRS data both non-Markovianity and left-truncation was present. We therefore relied on a simulation study to evaluate bias and root mean square error of our approach in this context. Under right-censoring and even non-Markovianity, but in the absence of left-truncation, the performance was good, which is in line with the theory[23]. The non-strict approach did seem to be sensitive to left-truncation, however. The strict approach, which for landmark time s uses only the subjects alive and at risk at time s in the calculation of the state occupation probabilities, performed quite well. For a small to moderate degree of left-truncation, bias and root mean square error of the non-strict approach are acceptable, but it is not completely clear how our approach performs when there is a considerable degree of left-truncation. Thus, the non-strict approach needs to be used with caution. The crucial issue might be in the correct choice of $n(s)$ in the definition of the pseudo-observations. We used the number of subjects alive and at risk at time s, even though additional subjects were used for calculation of the transition intensities and the state occupation probabilities. We also evaluated the non-strict approach, taking $n(s)$ to be the number of subjects used in the calculation of the state occupation probabilities, but this also resulted in a moderate bias. Perhaps an intermediate "effective sample size" governing the asymptotics of the state occupation probability estimates should be used, but it is unclear as yet how to define this. Further theoretical research and practical experience is needed in this case.

There are a number of directions for future research. First, the method is applicable for general multi-state models and is not restricted to the illness-death model. Depending on the objective of the data analysis it may also be of interest to select a prediction window, instead of a time-horizon τ, i.e. to investigate the ELOS in health over the next 10 years. Another possible extension of the pseudo-observation approach is to consider other outcomes, where one important possibility is regression models for quality-adjusted (remaining) life years. A utility q_h (per time unit spent in state) is then

assigned to each state h, and one is interested in $\sum_h q_h e_h(s)$. In another application, q_h could be (medical) costs associated with being in state h. Another outcome of interest may be the proportion of remaining life spent in health; in our setting that would be $e_{h=1}(s)/(e_{h=1}(s) + e_{h=2}(s))$.