Lifetime Data Anal (2008) 14:447–463
DOI 10.1007/s10985-008-9099-8

# Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data

**Hans C. van Houwelingen · Hein Putter**

Received: 6 March 2008 / Accepted: 10 September 2008 / Published online: 3 October 2008

**Abstract** This paper considers the problem of obtaining a dynamic prediction for 5-year failure free survival after bone marrow transplantation in ALL patients using data from the EBMT, the European Group for Blood and Marrow Transplantation. The paper compares the new landmark methodology as developed by the first author and the established multi-state modeling as described in a recent Tutorial in Biostatistics in Statistics in Medicine by the second author and colleagues. As expected the two approaches give similar results. The landmark methodology does not need complex modeling and leads to easy prediction rules. On the other hand, it does not give the insight in the biological processes as obtained for the multi-state model.

**Keywords** Dynamic prediction · Landmark model · Multi-state model · ALL

## 1 Introduction

This paper discusses the problem of obtaining dynamic x-year survival predictions during the follow-up of patients using all current information. We compare two approaches to obtain such predictions. One approach is the recently developed landmark methodology of van Houwelingen (2007). The other approach is the more traditional multi-state modeling (Putter et al. 2007).

The data used in this paper are obtained from the European Group for Blood and Marrow Transplantation (EBMT, http://www.ebmt.org/) registry. We consider all 2297 acute lymphoid leukemia (ALL) patients who had an allogeneic bone marrow transplant from an HLA-identical sibling donor between 1985 and 1998. The data were

H. C. van Houwelingen · H. Putter (✉)
Department of Medical Statistics and Bioinformatics, Leiden University Medical Center,
Post Zone S5-P, P.O. Box 9600, Leiden 2300 RC, The Netherlands
e-mail: h.putter@lumc.nl

extracted from the EBMT database in 2004. All patients were transplanted in first complete remission. Events recorded during the follow-up of these patients were: acute graft versus host disease (AGvHD), platelet recovery (PR, the recovery of platelet counts to a level of $>20\times10^9$/l), relapse and death. AGvHD has been defined as a GvHD of grade 2 or higher, appearing before 100 days post-transplant. Prognostic information at time of transplant are: donor recipient gender mismatch, T-cell depletion (TCD), year of transplant and age at transplant. The same data have been studied in a multi-state model by Fiocco et al. (2008). This system of engraftment and acute GvHD has been previously modeled and analyzed using quite similar multi-state models, see Klein et al. (1993) and Klein and Shu (2002).

For the sake of this paper we combine the events relapse and death into a single event "Failure". The clinical purpose of our modeling is to obtain a dynamic prognostic model for 5-year failure free survival given the history on AGvHD and PR and the prognostic covariates.

The outline of the paper is as follows. In Sect. 2 we will describe a traditional data analysis of failure free survival with AGvHD and PR as time-dependent covariates and the prognostic information as fixed covariates. In Sect. 3 we will describe the landmark approach and develop a dynamic prediction model for 5-year failure free survival given the patient history during the first year. In Sect. 4 we will develop a multi-state model for the data, use that to obtain dynamic 5-year failure free survival predictions and compare those predictions with the landmark predictions. In Sect. 5 we will discuss the pros and contras of the two approaches and competitors.

## 2 A closer look at the data

To simplify the presentation and the calculations tied event times are broken randomly. We first show the Kaplan–Meier estimates of the failure free survival distribution and the censoring distribution in Fig. 1a and b.

It is clear from the survival graph that the situation of the patients is quite stable after 5 years. The graph of the censoring distribution shows that the follow-up is quite complete in the first 5 years. Attempting to predict long term survival, e.g. 10-year survival, using this data set would be tricky because of the lack of follow-up in the recent cohorts. (At 10-years of follow-up there are 298 patients at risk, 237 from the first cohort (1985–1989), 61 from the second cohort (1990–1994) and 0 from the last cohort (1995–1998).

Table 1 shows the distribution of the risk factors and the (univariate) hazard ratios for treatment failure.

We categorized age and year of transplantation for didactical reasons. Both factors show a clear effect. Younger patients have a better prognosis and transplantations before 1990 had a worse prognosis. Donor recipient gender mismatch seems to be of minor importance, while TCD shows a clear negative effect on failure free survival.

To get an impression when AGvHD and PR occur, we show estimates of the cumulative incidence functions of time to AGvHD and time to PR in Fig. 1c and d. By definition, AGvHD appears before 100 days post-transplant. Since PR after 1 year is

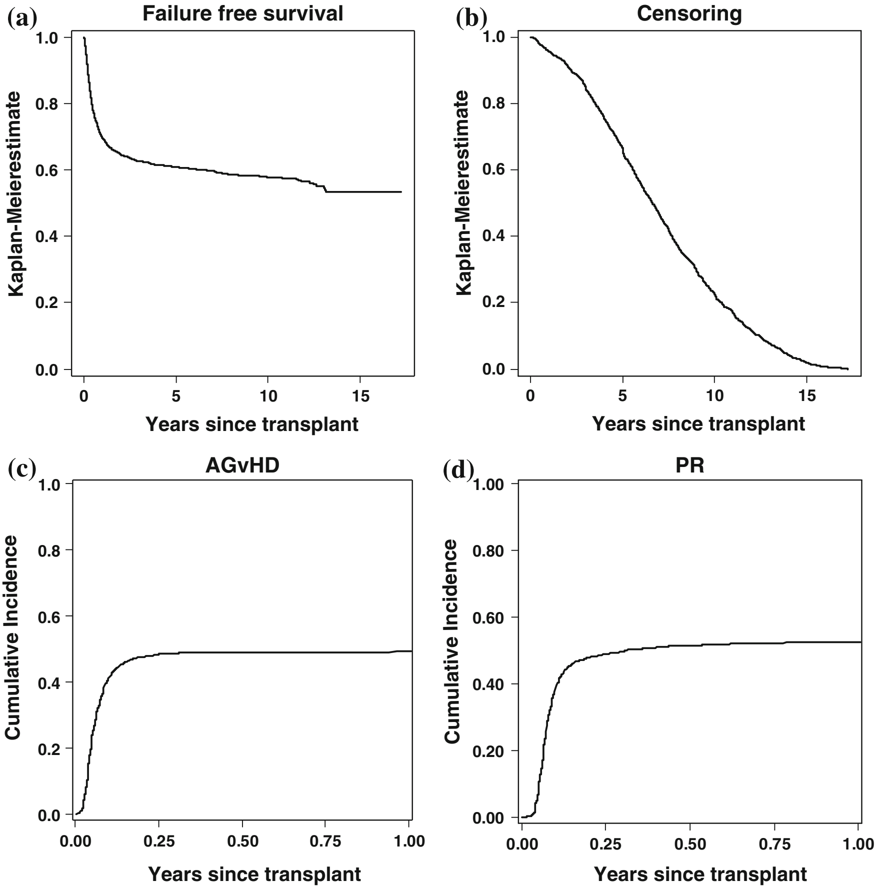

**Fig. 1** Kaplan–Meier estimates of the failure-free survival distribution (**a**), the censoring distributing (**b**), and estimates of the cumulative incidence functions for time until AGvHD (**c**) and time until PR (**d**)

very rare, we truncated the time axis after 1 year. It is clear that both AGvHD and PR mostly occur within the first 3 months.

In order to define sensible landmark models in the next section, we explore the potentially time varying effect of the time dependent factors AGvHD and PR. In order to do so, we fit a Cox model with four binary time-dependent covariates: AGvHD($t$), recent-AGvHD($t$), PR($t$) and recent-PR($t$). Here AGvHD($t$) stands for having experienced AGvHD before time $t$, while recent-AGvHD is defined as having experienced AGvHD within the last month, that is between $t - 1/12$ and $t$. The definition of "recent" as in the past month is based on some exploratory analysis. We have no biological rationale other than the general observation in the analysis of this type of data that the effects of intermediate events have a tendency to " fade out". The results in Table 2 show that recent-AGvHD is not significant, while recent-PR is highly significant. There remains a significant, but much smaller effect of PR after 1 month of its occurrence.

**Table 1** Overview of the prognostic factors and the corresponding hazard ratios for treatment failure (relapse or death)

| Prognostic factor | Category | N (%) | Hazard ratio (95% CI) |
|---|---|---|---|
| Donor recipient gender mismatch | No gender mismatch | 1,734 (76) | 1 |
| | Gender mismatch (F donor, M patient) | 545 (24) | 1.14 (0.98–1.32) |
| GvHD prevention | No TCD | 1,730 (76) | 1 |
| | TCD | 549 (24) | 1.29 (1.11–1.49) |
| Year of transplant | 1985–1989 | 634 (28) | 1 |
| | 1990–1994 | 896 (39) | 0.74 (0.63–0.86) |
| | 1995–1998 | 749 (33) | 0.73 (0.62–0.87) |
| Age at transplant (years) | ≤20 | 551 (24) | 1 |
| | 20–40 | 1,213 (53) | 1.37 (1.16–1.63) |
| | >40 | 515 (23) | 1.64 (1.35–1.99) |

**Table 2** Estimated parameters for time-dependent effects of AGvHD and PR

| | $\hat{\beta}$ | SE($\hat{\beta}$) | *P*-value | exp($\hat{\beta}$) |
|---|---|---|---|---|
| AGvHD($t$) | 0.405 | 0.072 | 0.000 | 1.500 |
| Recent-AGvHD($t$) | −0.220 | 0.179 | 0.218 | 0.803 |
| PR($t$) | −0.263 | 0.073 | 0.000 | 0.768 |
| Recent-PR($t$) | −1.070 | 0.257 | 0.000 | 0.343 |

The effects of these time-dependent covariates are assumed to be additive; so for instance the effect on failure of recent-PR is the sum of the coefficients of PR($t$) and recent-PR($t$)

In further model building we will consider AGvHD($t$), PR($t$) and recent-PR($t$) and will denote them simply by AGvHD, PR and recent-PR.

We also checked in the same Cox model that there is no significant interaction between AGvHD and PR. Therefore, we will not consider this interaction in our future modeling. For simplicity in further analyses we will consider year of transplantation prior to 1990 versus 1990 and later.

## 3 Dynamic prediction based on the landmark model

As described in the introduction we want to develop a dynamic prediction model for 5-year failure free survival based on the time-dependent covariates AGvHD, PR (and recent-PR) and the fixed covariates: donor recipient gender mismatch, TCD, year of transplantation and age at transplantation. Traditionally, this is done by making a model for time to failure with time-dependent and fixed covariates, plus models for time to AGvHD and time to PR depending on the fixed covariates and history and deriving a predictive model from that by conditioning on being failure free and the AGvHD and PR-status at the moment of prediction. As argued by Zheng and Heagerty (2005) and

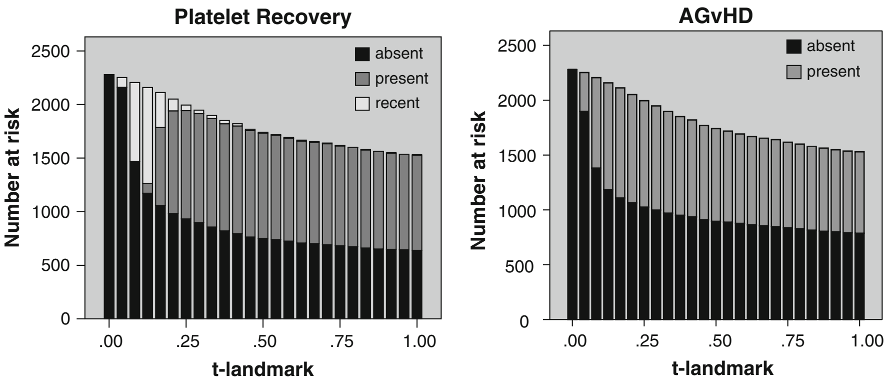

**Fig. 2** The distribution of the current value of AGvHD and PR at each landmark point

van Houwelingen (2007) predicting models can well be constructed without making comprehensive models by constructing landmark data sets with all relevant information needed for the prediction. From the original data set we constructed such data sets for 25 landmark (prediction) time points $t_{LM} = 0, 1/24, 2/24, \ldots, 1$ year. We denote $t_{LM}$ by $s$ to simplify the notation whenever convenient. In all data sets we take all patients still at risk for failure, compute the current value of AGvHD, PR and recent-PR and set the horizon for the failure time at $t_{hor} = 5$ to indicate that we want to obtain dynamic models for 5-year failure free survival. The variation of AGvHD and PR over time is shown in Fig. 2.

It is interesting to observe that the potentially important predictor recent-PR is only present in a very limited time window between the first and the third month. We will come back to that at the end of the section.

At each landmark point we can fit a simple Cox model on the interval $(t_{LM}, t_{hor})$ and use that to obtain a prediction of failure free survival at $t_{hor} = 5$. As argued in van Houwelingen (2007) using arguments similar to those of Xu and O'Quigley (2000), this will give a reasonably accurate estimate even if the predictors available at $t_{LM}$ have time-varying effects in the interval $(t_{LM}, t_{hor})$.

In general terms we take a grid of landmark time points $s_0 \leq s \leq s_1$ (in our application $s_0 = 0, s_1 = 1$ and we have 25 equally space landmark points). Let $X(s)$ stand for the current vector of predictors that might depend on the landmarking time-point.

For each landmark point we postulate the prediction model

$$P(T > t_{hor}|T > t_{LM} = s, X(s)) = \exp(-\exp(X(s)^T \beta(s)) H_0(s, t_{hor})) \qquad (1)$$

and estimate the parameters of this model by fitting the simple Cox model

$$h(t|X(s), s) = h_0(t|s) \exp(X(s)^T \beta(s)) \quad \text{for} \quad s = t_{LM} \leq t \leq t_{hor} \qquad (2)$$

enforcing administrative censoring at $t_{hor}$.

Fitting this model for each landmark point separately would ignore the "overlap" between landmark data sets. We can expect that the coefficients $\beta(s)$ depend on s in rather a smooth way. We can bring more structure into the analysis by modeling the regression parameters $\beta(s)$ as a function of $s$. Generally, we can take any parametric model. In this application we take the simple model

$$\beta(s) = \beta_0 + \beta_1 s \tag{3}$$

and gather the components of $\beta_0$ and $\beta_1$ into a single parameter vector $\beta_{LM}$. Fitting the model (1)–(2) is equivalent to maximizing the pseudo partial log-likelihood

$$ipl(\beta_{LM}) = \sum_i d_i \left( \sum_{s<t_i} \left[ X_i(s)^T \beta(s) - \ln(\sum_{t_j \geq t_i} e^{X_j(s)^T \beta(s)}) \right] \right) \tag{4}$$

It can be fitted to the data in standard software by using a stacked data set, containing all the 25 landmark data sets with stratification on the landmark. This can be used to inspect whether the coefficients depend on the landmark. However, such a fit from standard software cannot be used to test the statistical significance of the components of $\hat{\beta}_0$ and $\hat{\beta}_1$ since the data of the same patient are used repeatedly in the different landmark strata. The correct standard errors can by obtained by taking into account the "clustering" of the data and using the sandwich estimators of Lin and Wei (1989). This approach is incorporated in the landmarking software we developed (van Houwelingen 2007), but it can also be performed in software packages like R and Stata. For a further description how this can be done, see www.msbi.nl/multistate.

After fitting the model the baseline hazard at the event time $t_i$ can be estimated by a Breslow-type estimator

$$\hat{h}_0(t_i|s) = \frac{1}{\sum_{t_i \leq t_j} \exp(X_j(s)^T \hat{\beta}(s))} \tag{5}$$

It is interesting to observe that $\hat{h}_0(t|s)$ does not depend on $s$ if $X(s)$ and $\hat{\beta}(s)$ are constant. See Van Houwelingen (2007) for a more elaborate discussion. In our application some of the components of $X(s)$ vary with $s$ and not all $\hat{\beta}(s)$ turn out to be constant either. That implies that $\hat{h}_0(t|s)$ will depend on $s$. To add more structure and to make it easier to interpret the models we assume a model

$$h_0(t|s) = h_0(t) \exp(\gamma(s)) \tag{6}$$

with the restriction $\gamma(s_0) = 0$ to warrant identifiably. In our application we take $\gamma(s)$ to be a third degree polynomial

$$\gamma(s) = \gamma_1 s + \gamma_2 s^2 + \gamma_3 s^3 \tag{7}$$

The model (2), (3), (6), (7) can be fitted directly by applying a simple Cox model to the stacked data set, provided the software allows for delayed entry at $s$. Repeated observations on the same subject automatically leads to the presence of many ties.

**Table 3** Estimated parameters in the ipl*-model

(a) *The $\beta$-parameters and their standard errors*

| Factor | $\hat{\beta}_0$ (SE) | $\hat{\beta}_1$ (SE) |
|---|---|---|
| AGvHD | 0.317 (0.077) | |
| Recent-PR | −0.179 (0.042) | |
| Age 20–40 | 0.285 (0.099) | −0.193 (0.097) |
| Age 40+ | 0.502 (0.113) | −0.298 (0.152) |
| Tx < 1990 | 0.259 (0.092) | |
| TCD | 0.254 (0.094) | |

(b) *The $\gamma$-parameters with standard errors*

| $\hat{\gamma}_1$ (SE) | $\hat{\gamma}_2$ (SE) | $\hat{\gamma}_3$ (SE) |
|---|---|---|
| −0.712 (0.094) | 1.544 (0.270) | −0.828 (0.155) |

Fitting the model with the Breslow partial likelihood for those tied observations is equivalent to maximizing a different pseudo partial log-likelihood, namely

$$\text{ipl}^*(\beta_{LM}, \gamma) = \textstyle\sum_{t_i} d_i \left[ \sum_{s<t_i} (X_i^T \beta(s) + \gamma(s)) - \ln(\sum\limits_{t_j \geq t_i > s} \exp(X_j^T \beta(s) + \gamma(s)) \right] \tag{8}$$

The estimator of the corresponding baseline hazard is given by

$$\hat{h}_0^*(t_i) = \frac{\#(s < t_i)}{\sum\limits_{t_j \geq t_i > s} \exp(X_j^T \hat{\beta}(s) + \hat{\gamma}(s)))} \tag{9}$$

Again, standard errors for the regression parameters can be obtained by sandwiching and this is implemented in our software. Alternatively, R and Stata can be used for this purpose as well, among others. Standard errors for the baseline hazard and estimated survival probabilities are not included yet. The convenience of the ipl*-model is that landmark survival probabilities can easily be estimated as

$$\hat{P}(T > t | T \geq s, X(s)) = \exp(-e^{X_i^T \hat{\beta}(s) + \hat{\gamma}(s)} (\hat{H}_0^*(t) - \hat{H}_0^*(s-))) \tag{10}$$

where $\hat{H}_0^*(t)$ is the cumulative baseline hazard $\hat{H}_0^*(t) = \sum_{t_i \leq t, d_i = 1} \hat{h}_0^*(t_i)$.

We fitted the ipl*-model to the data of our application. The parameters of the final model are given in Table 3.

Table 3a can be interpreted losely as giving the relative risks for dying before 5 years at different landmark points. For AGvHD the relative risk $RR = \exp(0.317) = 1.37$, for recent-PR $RR = 0.84$, for Tx $< 1990$ $RR = 1.30$ and for TCD $RR = 1.29$. For the age groups, the relative risks compared to the baseline groups with age $< 20$, varies with the time of landmarking. For the 20–40 age group it varies from $RR = 1.33$ at the start of the follow-up to $RR = 1.09$ after 1 year. Similarly, the relative risk for the 40+ age group varies from $RR = 1.65$ to $RR = 1.23$. The graph of the $\gamma(s)$- function (not shown) has a "dip" about $s = 0.3$ related to the changes in the distribution of the dynamic predictors shown in Fig. 2. This curve is used in the computation of the 5-year survival probability by means of formula (10). It is hard to give it a simple interpretation. The bending of the curve near $s = 1$ might be an artifact of the polynomial model. Using B-splines might give a more "natural" curve, but we stick to the polynomial model for the sake of simplicity. The shape of $\hat{\gamma}(s)$ is partly caused by the transient behavior of recent-PR as shown in Fig. 2b.

The curves for $\hat{H}_0(t)$ and its quadratic B-spline fit with knots at $t = 1$ and $t = 2$

$$\hat{H}_0^*(t) = 0.392 \cdot t - 0.157 \cdot t^2 + 0.129 \cdot ((t-1)^+)^2 + 0.024 \cdot ((t-2)^+)^2 \quad (11)$$

(not shown) virtually coincide. (Here, $a^+ = \max(a, 0)$) The function rapidly increases in the beginning, slows down later and reaches a value of about 0.32 at 5 years after transplant.

Table 3 only gives the results as obtained after some data-driven model building. The main findings are:

(1) Recent-PR suffices to describe the PR-effect. In contrast with Table 2, the PR-effect itself was not significant in a model with recent-PR.
(2) The effect of AGvHD and PR does not depend significantly on the landmark $s$.
(3) Donor recipient gender mismatch has no significant effect.
(4) TCD has a significant constant effect.
(5) The cohort "before 1990" has a significantly worse prognosis. There is no significant difference between the other two cohorts.
(6) Age at transplantation has a major effect that depends significantly on $s$.

It might seem a bit unnatural that only recent-PR is used in the predictive model. The statistical explanation is that (i) there is very little effect of PR for those who are still alive at the landmark time-point after 6 months; (ii) for landmarking in the very beginning of the follow-up recent-PR has a stronger effect that PR-present. The reader should bear in mind that the effects in Table 3 apply to 5-year survival while those in Table 2 apply to the hazard.

The predicted 5-year failure free survival probabilities for TCD = 0 and Tx after 1989 are shown for the three age categories in Fig. 3.

The curves look quite smooth showing that the model allows predictions at any time in the first year, not only in the 25 landmark points. However, the smooth appearance does not imply that the dynamic prediction itself is smooth. Everybody starts in the upper left corner curve with No AGvHD, no recent-PR. The other curves are actually meaningless for $s$ close to zero. At the occurrence of either PR or AGvHD a patient shifts to one of the other curves. One month after PR, a patient with recent-PR moves back to the No recent-PR status. This also implies that the curves for recent-PR are only relevant in the first quarter of the follow-up. Another way of presenting would be to show how the prognosis changes over time for a patient who, for example, experiences PR after 6 weeks and AGvHD after 8 weeks. We leave that to the reader. It

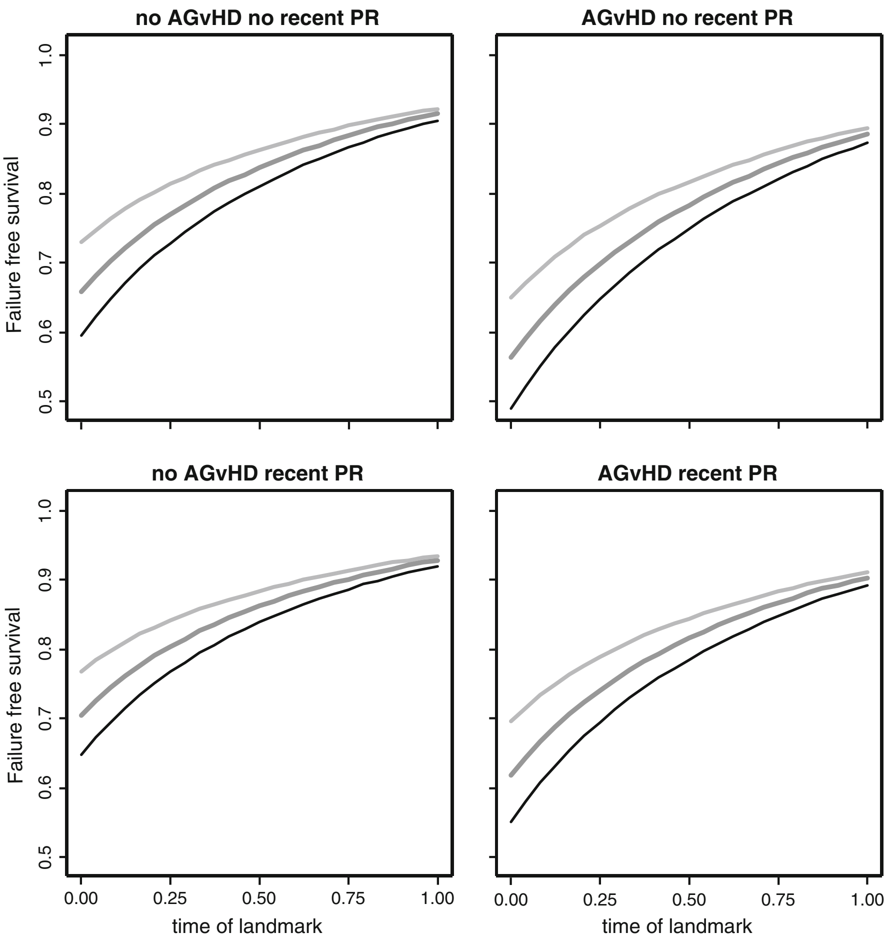

**Fig. 3** Estimated 5-year failure-free survival probabilities for subject with no TCD and transplanted after 1989 based on the landmark model. In light grey age <20, in grey age 20–40, in black age ≥40

is not hard to write a little program that implements the predictions based upon the ipl*-model given by Table 3 and approximation (11).

## 4 Dynamic prediction based on the multi-state model

The multi-state model we use for the comparison follows the methods outlined in Putter et al. (2007) and the *mstate* software developed for the prediction based upon the multi-state model. In this approach we use a single time-scale, years since transplantation. The transition intensities and their covariate effects are modeled by imposing Cox proportional hazards models for the transition intensities. Some restrictions on the baseline hazards and the covariate effects are made in order to follow the landmark model and some states have to be introduced to handle the time-varying effect of PR.

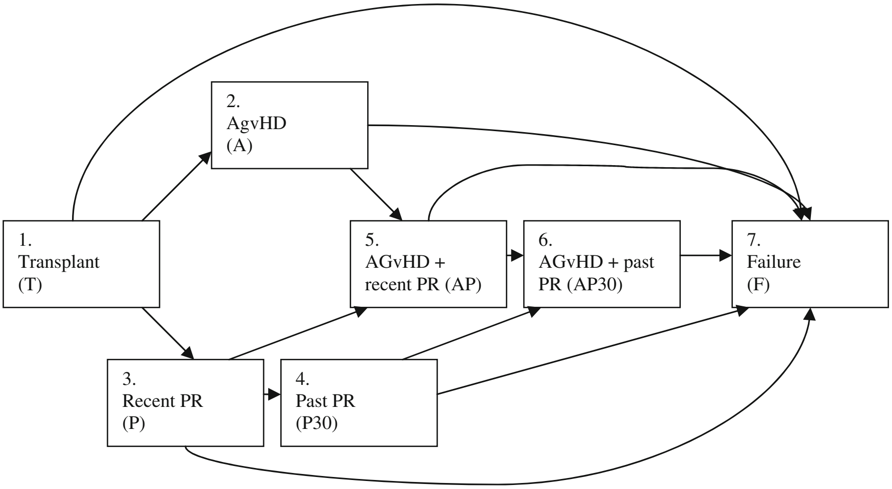

**Fig. 4** The multi-state model

After estimating the baseline hazards and the covariate effects, the 5 years relapse-free survival probabilities for a patient with a given set of covariate values are obtained by first "deducing" the patient-specific transition intensities for all the transitions in the model and by subsequently applying the Aalen and Johansen formula to these patient-specific transition intensities. Each of these steps is detailed below.

The multi-state model that allows an effect of recent-PR as described in Sect. 3 is shown in Fig. 4.

The multi-state model has seven states, indicated by boxes, and thirteen direct transitions between states, indicated by arrows. In Fig. 4 we have also indicated the abbrevations that will be used to denote these states. We distinguish between PR within the last 30 days and PR more than 30 days ago by assigning two distinct states to PR, states P (recent-PR) and P30 (past-PR). Since PR may occur after AGvHD, also two states are used to indicate AGvHD plus PR. A patient in state P will automatically move to state P30 after 30 days, unless another event or a censoring occurs before 30 days after PR. A similar transition after 30 days is defined from state AP to AP30. We further want to create a model in which AGvHD($t$) and PR($t$) act as time-dependent covariates in a Cox model for the transition to Failure with a time-varying effect for PR($t$) (recent-PR versus past-PR). Moreover, we want to exclude such a time-varying effect of PR($t$) on the transition to AGvHD for the patients who have experienced PR. We create such a model by imposing the following restrictions:

- No covariates are incorporated into the transitions P $\rightarrow$ P30 and AP $\rightarrow$ AP30
- Transitions P $\rightarrow$ AP and P30 $\rightarrow$ AP30 have identical baseline transition intensities and covariate effects
- Transitions P $\rightarrow$ F and P30 $\rightarrow$ F differ only with respect to a proportionality coefficient indicating the effect of recent-PR (the effect of prognostic factors is assumed to be identical)

- Transitions AP → F and AP30 → F differ only with respect to a proportionality coefficient indicating the effect of recent-PR (the effect of prognostic factors is assumed to be identical)
- All transitions into state F (failure) are assumed to be proportional, with transition T → F being the baseline, and two proportionality coefficients being estimated, indicating the effect of recent-PR and of AGvHD.

For the remainder the model has been chosen as free as possible; all transition intensities are freely estimated and the effects of all fixed covariates are allowed to differ between transitions. The fixed covariates we consider are the same as in Sect. 3: TCD, transplant before 1990 and age at transplant (in three groups). In terms of parameters that need to be estimated, the basis is a Cox proportional hazards model for the transition intensity of transition $i \to j$ of the form

$$h_{i,j}(t) = h^0_{i,j}(t)\exp(\beta^T_{i,j} Z),$$

where $h^0_{i,j}(t)$ is the baseline transition intensity of transition $i \to j$, and $\beta_{i,j}$ is a regression vector of transition-specific covariate effects. In terms of these baseline hazards and regression vectors, the above restrictions read as follows:

$$\beta_{P,P30} = \beta_{AP,AP30} = 0;\ h^0_{P,AP}(t) = h^0_{P30,AP30}(t);\ \beta_{P,AP} = \beta_{P30,AP30}$$
$$\beta_{P,F} = \beta_{P30,F};\ \beta_{AP,F} = \beta_{AP30,F};\ h^0_{i,F}(t) = h^0_{1,F}(t)\exp(\gamma_i),$$

where by definition $\gamma_T = 0$, and also $\gamma_{P30} = 0$. Two unknown parameters $\gamma_A$ and $\gamma_P$ indicate the effects of AGvHD and of recent-PR on failure free survival, and we have $\gamma_{AP} = \gamma_A + \gamma_P$ and $\gamma_{AP30} = \gamma_A$.

All this implies that seven different baseline intensities are estimated, namely $h^0_{T,A}(t)$, $h^0_{T,P}(t)$, $h^0_{A,AP}(t)$, $h^0_{P,AP}(t)$, $h^0_{P,P30}(t)$, $h^0_{AP,AP30}(t)$ and $h^0_{T,F}(t)$ and that four distinct covariate effects are estimated on 8 distinct transitions. After some data preparation, using transition-specific covariates [see e.g. Andersen et al. (1993) and Putter et al. (2007)] and defining appropriate strata, all regression parameters and baseline transition hazards can be fitted within a single stratified Cox regression model. For more details the reader is referred to our website mentioned under Software.

The estimated baseline cumulative transition hazard estimates of the transitions (not shown) resemble the form of the plots of Fig. 1. The cumulative transition intensities of P → P30 and AP → AP30 were estimated by adding a time to P30 (AP30) in the data used for estimation of the parameters of the multi-state model for patients reaching P (AP). These times were given by $t + 30$ for patients reaching P (AP) at time $t$. The transition intensities were then estimated from the data without using covariates for those transitions (see restrictions stated above). The estimates of the cumulative transition intensity result in very high cumulative hazards, reflecting the fact that almost all patients reaching the P state will move on to the P30 state after 30 days. The estimated cumulative hazards also have large jumps, and we shall see later that this results in slight irregularities in some of the predictions, but it is a consequence of our wish to stay in the Markov model framework in order to apply the Aalen–Johansen formulas for obtaining prediction probabilities. In this case it would have been more logical to

**Table 4** Estimated hazard ratios in the multi-state model with 95% confidence intervals

| | TCD | Year > 1990 | Age 20–40 | Age > 40 |
|---|---|---|---|---|
| (a) *Estimates of the prognostic factors (exogeneous covariates)* | | | | |
| T → A | **0.79 (0.67–0.92)** | 0.89 (0.77–1.03) | 1.12 (0.95–1.32) | 1.07 (0.88–1.31) |
| P → AP | 0.76 (0.53–1.10) | 0.90 (0.70–1.18) | 1.31 (0.91–1.87) | **1.57 (1.07–2.32)** |
| T → P | **0.65 (0.54–0.78)** | **1.29 (1.11–1.50)** | 1.06 (0.89–1.26) | **1.28 (1.05–1.57)** |
| A → AP | 1.07 (0.84–1.37) | **1.85 (1.51–2.27)** | **0.67 (0.53–0.84)** | **0.75 (0.56–0.99)** |
| T → F | 1.16 (0.90–1.51) | 0.85 (0.63–1.14) | **1.51 (1.15–1.98)** | **1.68 (1.19–2.38)** |
| A → F | **1.53 (1.16–2.02)** | 1.09 (0.80–1.49) | **1.40 (1.05–1.86)** | **1.72 (1.21–2.45)** |
| P → F | 1.26 (0.88–1.82) | 0.99 (0.72–1.37) | 0.81 (0.58–1.14) | 1.26 (0.86–1.84) |
| AP → F | 1.01 (0.74–1.38) | **0.77 (0.59–1.00)** | **1.50 (1.13–2.00)** | **1.81 (1.29–2.53)** |
| (b) *Estimates of AGvHD and recent PR on failure (endogeneous covariates)* | | | | |
| Acute GvHD | 1.23 (0.90–1.69) | | | |
| Recent platelet recovery | **0.39 (0.24–0.62)** | | | |

Covariate effects significant at the 5% level are shown in bold

have used the semi-Markov or clock-reset approach [see e.g. Lagakos et al. (1978), Dabrowska et al. (1994), Putter et al. (2006)], but that would have made prediction far more difficult.

The effects of the covariates are shown in Table 4. Again a number of transitions are not shown. The transitions P → P30 and AP → AP30 are assumed not to depend on covariates. The effects of covariates are assumed to be identical for the transitions P → F and P30 → F, and also for P → AP and P30 → AP30, and for AP → F and AP30 → F. The most consistent findings are the effects of age; higher age generally implies a higher transition rate to failure, and a lower transition rate to PR after AGvHD. TCD, given as a treatment to prevent AGvHD is effective as such, but also has negative side effects, such as a lower PR rate. Moreover, due to the well known reverse biological mechanisms of AGvHD and relapse, TCD has a higher relapse rate and hence generally a higher intensity of the transitions into failure. A more detailed analysis distinguishing between relapse and death as endpoints (Fiocco et al. 2008) revealed a higher relapse rate for TCD but no direct effect of TCD on death. The adverse effect of TCD on failure was not seen as clearly here because in this analysis no distinction was made between relapse and death.

Figure 5 shows similar pictures as Fig. 3, that is estimated 5-year failure free survival for patients transplanted after 1989 and with TCD = 0. They were obtained by first deriving the patient-specific transition intensities for all transitions in the multi-state model from the estimated regression coefficients of Table 4 and the estimated baseline cumulative transition intensities. Subsequently, the formula of Aalen and Johansen (1978) (see also Andersen et al. (1993), Sect. VII.2.3) was used to obtain predictions of 5-year failure free survival.

Each picture is based on predictions starting from a different state in the multi-state model (states T, A, P, AP, P30, AP30, respectively). The jumps in the prediction No AGvHD, recent-PR (state 3), and to a lesser extent those in the prediction AGvHD,

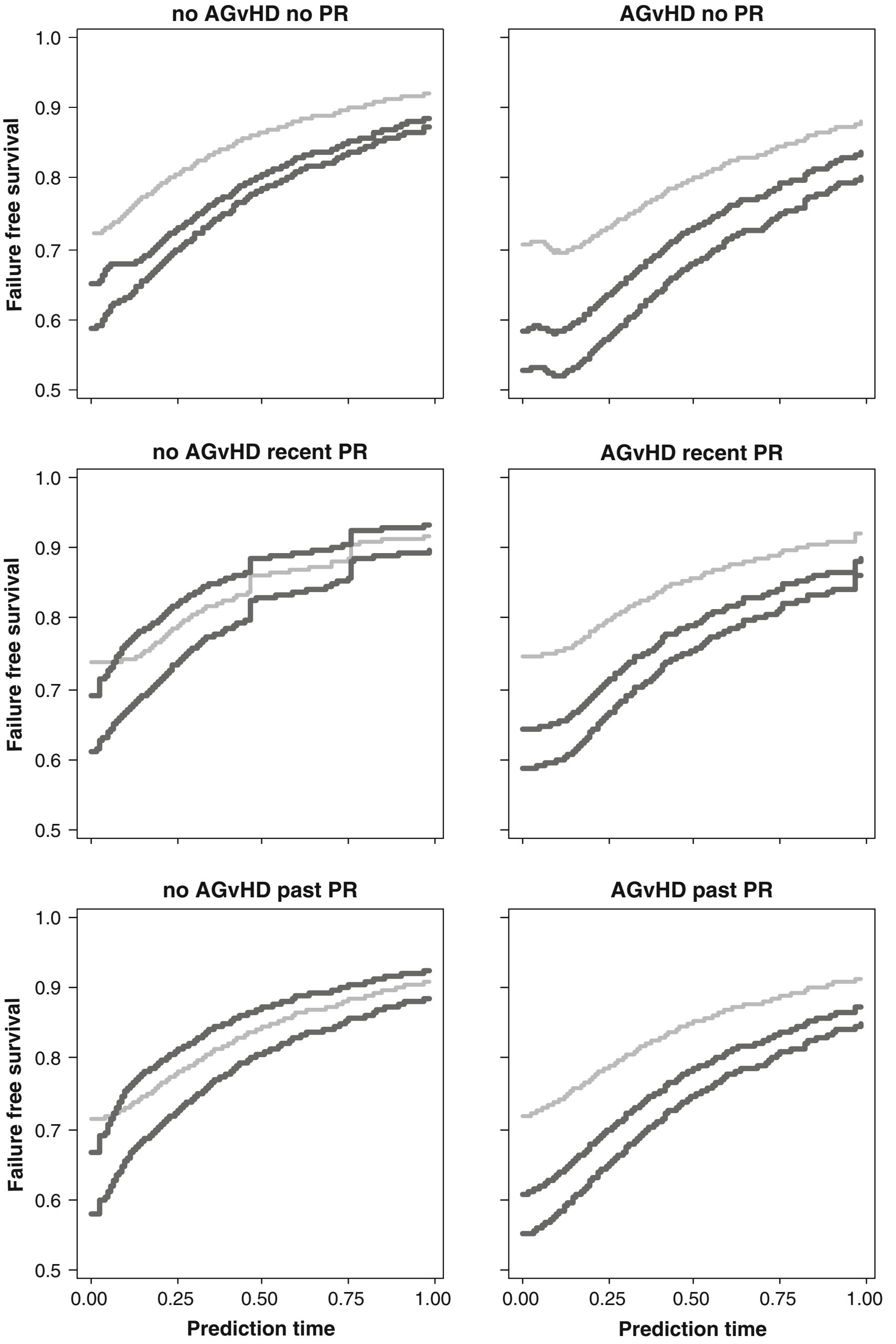

**Fig. 5** Estimated 5-year failure-free survival probabilities for subject with no TCD and transplanted after 1989 based on the multi-state model. In light grey age $< 20$, in grey age 20–40, in black age $\geq 40$

recent-PR (state 4) are caused by the jumps in the cumulative hazards of the P → P30 and AP → AP30 transitions. Although we attempted to stay close to the model of Sect. 3, there are two striking differences: in Sect. 3 only recent-PR matters, merging patients with no PR and patients with past-PR. Such a merging is not possible in our implementation of the multi-state modeling. The caveats given in Sect. 3 for the interpretation of such curves apply here as well. Another striking difference between Figs. 3 and 5 is the reversed age effect in the group of patients with No AGvHD and recent or past-PR.

It might be a bit surprising that some curves for no PR are non-monotone. For instance, the second plot of Fig. 5 shows $P$(failure free at $t|$ state AGvHD at $s$) for fixed $t$ and varying $s$. A plot of $P$(failure free at $t|$ state AGvHD at $s$) for fixed $s$ and varying $t$ is indeed monotone since Failure is an absorbing state in the multi-state model, but there is no reason why this should be the case for fixed $t$ and varying $s$. The fact that the curves are non-monotone in this instance is a consequence of the dynamic nature of PR status. The curve can be interpreted as: the longer you have waited for PR, the more likely that it will never happen and the worse the prognosis. Actually, the non-monotone behavior in Fig. 5 might be an indication that the multi-state modeling stays closer to the clinical data.

## 5 Discussion

We have chosen to build a landmark prediction model that is valid for the whole first year of follow-up. On retrospect, this is open for debate because the intermediate events mostly occur in the first half year. Although we did not find a statistically signifcant interaction between the (dynamic) predictors and the time-point of landmarking, the picture might slightly change if we restrict the landmarking to the first half year. However, we think that our choice for the first year gives a good insight in the potential of the landmarking methodology and its pros and cons when compared with multi-state modeling.

The big advantage of using a Markov multi-state model is the availability of the formula of Aalen and Johansen (1978) that converts the transition hazards into transition or state probabilities through repeated multiplication of matrices containing the transition hazard increments. The multiplication is over the event time points; when these multiplications are performed in increasing order of event time points this yields predictions forward in time. In order to obtain our 5-years predicted probabilities of failure, we have instead performed these matrix multiplications backward in time. This makes the prediction of 5-years failure free survival from different points in time quite straightforward once the required transition hazards are obtained. But in the presence of covariates Aalen–Johansen's formula is only valid for Markov models (see Datta and Satten (2001) however, who show that the Markov property is not needed in the absence of covariates). The wish to stay within the framework of Markov models is in fact not very well compatible with the nature of the intermediate events found in Sect. 2. In particular the distinction between recent and past-PR has forced us to make some arguably unnatural steps. It has led us to introduce two extra states in the multi-state model with some restrictions on the transition intensities. In order to compensate for the large number of additional parameters, we have modeled the transitions P → P30, AP → AP30, and P → AP and P30 → AP30, with additional restrictions on equality of the baseline transition intensities and covariate effects. These restrictions appeared to make sense a priori (certainly in clock reset time-scale, perhaps less so in clock forward time-scale) and limited checks have not shown violations of the underlying assumptions. Of course it would have been most natural to use the time-scale of time since entrance in state P (recent-PR) for the transition to P30 (past-PR), and for the similar transitions after AGvHD (the transition AP → AP30), but this destroys the Markov property. Models that use time since entrance of the present state as time-scale are called clock reset or semi-Markov or Markov renewal models (under additional Markov-like assumptions). Prediction in this type of multi-state models is far more difficult to do exactly. Instead simulation could be used to approximate the required transition probabilities.

The multi-state model could be characterized as an indirect way of obtaining a prediction through a complete model for the follow-up of a patient. Such a modeling can be quite useful for a biological understanding of the underlying process, but the predictions derived from these models can be off the mark if the assumptions underlying the model are violated or the fit of the model is not perfect. Actually, the validity of the predictions can be checked in the same data set, but that is hardly ever done in practice. In principle, such a goodness-of-fit check could be combined with the landmark analysis by comparing the observed survival in each landmark data set with the predictions derived from the multi-state, but in practice it would take quite some energy to carry out such an analysis.

The landmark approach can be seen as a way of direct modeling. One useful approach is the pseudo-value approach developed by Klein and Andersen (see Andersen et al. (2003) and Andersen and Klein (2007)) which is inspired by the wish to have simple regression models for multi-state transition probabilities $P$(in state $S$ at time $t|$ covariate $X$). An alternative is the direct modeling of Scheike and Zhang (2007) and Scheike et al. (2008) which directly estimates similar probabilities from the data using logistic regression type models and an ingenious way of handling the censoring during the follow-up. The typical graphical presentation of the results of such an analysis shows how $P$(in state $S$ at time $t|$ covariate $X$) changes over time.

The main difference with our landmark approach is that we are interested in the dynamic prediction $P$(in state $S$ at time $t_{hor}|$ covariate $X$, history at time $s$). A generalization that includes both perspectives is to let both $t_{hor}$ and the landmark time-point $s$ vary.

A technical difference is that we derive the prediction by fitting a simple Cox model on the interval $(s, t_{hor})$ to obtain an estimate of the survival up to $t_{hor}$. This circumvents the censoring problem that inspired the pseudo-values of Andersen and Klein, but could be biased if there is much censoring between $s$ and $t_{hor}$.

The main limitation of the approach is that it can only handle survival type data, that is data with a single absorbing state. A next step would be to develop a similar approach for competing risk data. Klein and Andersen (2005) apply the pseudo-value approach to competing risk data to obtain simple estimates of the cumulative incidence functions, comparable to the estimates coming from the models of Fine and Gray (1989) based on the modeling the sub-distribution hazard. Scheike et al. (2008) give methodology for directly estimating the cumulative incidence functions by binomial models. However it is not quite clear how to define and estimate *dynamic* cumulative incidence functions. Of course it is always possible to define landmark sets and to estimate landmark specific cumulative incidence functions, but it is not yet clear how different landmark models could be combined into one (pseudo) model as used in this paper.

The big advantage of the landmark approach is that it can easily incorporate any type of information about the patients history. Multi-state models as used in Sect. 4 can only be used to obtain predictions given the current state in the model. In the ALL example, it could not pool patients with no-PR with patients with past-PR while this was easily done in the landmark model, which can handle any type of time-dependent covariate without modeling the dynamics of the process itself.

Another advantage of the landmark approach is the sparseness of the model. It only has to estimate a few parameters and a single baseline hazard, while the multi-state model has much more parameters, which could lead to serious over-fitting. The implication is that it has a similar simplicity as the direct models of Klein and Andersen (2005) and Scheike and Zhang (2007).

A final caveat is that one might be tempted to use prediction formula (10) for all $s < t \leq t_{hor}$. Such a prediction might be biased if in the landmark data set the effect of the last observed covariate $X(s)$ is time-varying, that is $\beta(t|s) \neq \beta(s)$ for some $s < t \leq t_{hor}$. It would be better to create a new data set with a different horizon. Ultimately, this would lead to data-sets truncated at $t_{LM} = s$ and administratively censored at $t_{hor}$ that are used to "predict from $t_{LM} = s$ to $t_{hor}$". For each prediction problem we can fit simple PH models with coefficients $\beta(t_{LM}, t_{hor})$ and baseline hazard $h_0(t|t_{LM}, t_{hor})$. The challenge would be to develop methodology that allows a smooth dependence of both the coefficients $\beta(t_{LM}, t_{hor})$ and the baseline hazard $h_0(t|t_{LM}, t_{hor})$ on the pair $(t_{LM}, t_{hor})$.

The caveat above is especially relevant in case of time-dependent covariates. Since the landmark approach as applied here can only use the current value of the time-dependent covariate $X(s)$ it can be expected that its effect will decrease over time. Potential biases in the predictive probabilities can be avoided by using joint models for the time-dependent covariate and the survival hazard. However, any lack of fit of such models could lead to similar biases.

## 6 Software

The mstate software is available as an R package on www.msbi.nl/multistate. Macro's for dynamic predicting using landmarking as well as the data used in this paper can also be found on this website.

**Acknowledgements** The EBMT is gratefully acknowledged for providing the data for this study. Research leading to this paper was supported by the Netherlands Organization for Scientific Research Grants ZONMW-912-02-015 "Survival analysis for complicated data" and ZONMW-912-07-018 "Prognostic modelling and dynamic prediction for competing risks and multi-state models". Moreover, much of the research by the first author was carried out during a stay at the Centre for Advanced Study at the Norwegian Academy of Science and Letters in Oslo during the academic year 2005/2006, taking part in the joint work of an international research group studying Statistical Analysis of Complex Event History Data.

## References

Aalen OO, Johansen S (1978) Empirical transition matrix for nonhomogeneous Markov-chains based on censored observations. Scand J Stat 5:141–150

Andersen PK, Klein JP (2007) Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. Scand J Stat 34:3–16. doi:10.1111/j.1467-9469.2006.00526.x

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Andersen PK, Klein JP, Rosthoj S (2003) Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 90:15–27. doi:10.1093/biomet/90.1.15

Dabrowska DM, Sun GW, Horowitz MM (1994) Cox regression in a Markov renewal model: an application to the analysis of bone-marrow transplant data. J Am Stat Assoc 89:867–877. doi:10.2307/2290911

Datta S, Satten GA (2001) Validity of the Aalen–Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non-Markov models. Stat Probab Lett 55:403–411. doi:10.1016/S0167-7152(01)00155-9

Fiocco M, Putter H, van Houwelingen JC (2008) Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. Stat Med 27:4340–4358. doi:10.1002/sim.3305

Fine JP, Gray RJ (1989) A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 94:496–509. doi:10.2307/2670170

Klein JP, Andersen PK (2005) Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics 61:223–229. doi:10.1111/j.0006-341X.2005.031209.x

Klein JP, Shu Y (2002) Multi-state models for bone marrow transplantation studies. Stat Methods Med Res 11:117–139. doi:10.1191/0962280202sm277ra

Klein JP, Keiding N, Copeland (1993) Plotting summary predictions in multistate survival models: probabilities of relapse and death in remission for bone-marrow transplantation patients. Stat Med 12: 2315–2332. doi:10.1002/sim.4780122408

Lagakos SW, Sommer CJ, Zelen M (1978) Semi-markov mdoels for partially censored data. Biometrika 65:311–317. doi:10.1093/biomet/65.2.311

Lin DY, Wei LJ (1989) The robust inference for the Cox proportional hazards model. J Am Stat Assoc 84:1074–1078. doi:10.2307/2290085

Putter H, van der Hage J, de Bock GH, Elgalta R, van de Velde CJH (2006) Estimation and prediction in a multi-state model for breast cancer. Biom J 48:366–380. doi:10.1002/bimj.200510218

Putter H, Fiocco M, Geskus RB (2007) Tutorial in biostatistics: competing risks and multi-state models. Stat Med 26:2389–2430. doi:10.1002/sim.2712

Scheike TH, Zhang MJ (2007) Direct modelling of regression effects for transition probabilities in multistate models. Scand J Stat 34:17–32. doi:10.1111/j.1467-9469.2006.00544.x

Scheike TH, Zhang MJ, Gerds TA (2008) Predicting cumulative incidence probability by direct binomial regression. Biometrika 95:205–220. doi:10.1093/biomet/asm096

van Houwelingen HC (2007) Dynamic prediction by landmarking in event history analysis. Scand J Stat 34:70–85. doi:10.1111/j.1467-9469.2006.00529.x

Xu R, O'Quigley J (2000) Estimating average regression effect under non-proportional hazards. Biostatistics 1:423–439. doi:10.1093/biostatistics/1.4.423

Zheng Y, Heagerty PJ (2005) Partly conditional survival models for longitudinal data. Biometrics 61: 379–391. doi:10.1111/j.1541-0420.2005.00323.x