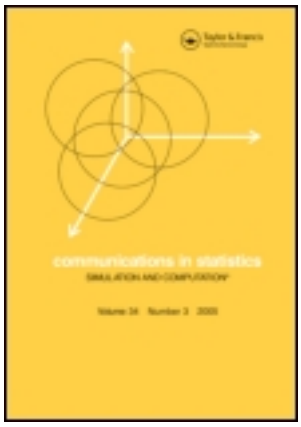

Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Comparison of Selected Methods for Modeling of Multi-State Disease Progression Processes: A Simulation Study

Ella Huszti [a b], Michal Abrahamowicz [a], Ahmadou Alioum [c d] & Catherine Quantin [e]

[a] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

[b] University of Washington -- Harborview Center for Prehospital Emergency Care, Seattle, Washington, USA

[c] Inserm, Bordeaux, France

[d] ISPED-Universitéé Victor Segalen Bordeaux 2, Bordeaux, France

[e] Medical Informatics Department, Dijon University Hospital, Dijon, France

Available online: 01 Jun 2011

To cite this article: Ella Huszti, Michal Abrahamowicz, Ahmadou Alioum & Catherine Quantin (2011): Comparison of Selected Methods for Modeling of Multi-State Disease Progression Processes: A Simulation Study, Communications in Statistics - Simulation and Computation, 40:9, 1402-1421

To link to this article: http://dx.doi.org/10.1080/03610918.2011.575505

Communications in Statistics—Simulation and Computation®, 40: 1402–1421, 2011

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610918.2011.575505

Comparison of Selected Methods for Modeling of Multi-State Disease Progression Processes: A Simulation Study

ELLA HUSZTI[1,2], MICHAL ABRAHAMOWICZ[1], AHMADOU ALIOUM[3,4], AND CATHERINE QUANTIN[5]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada
[2]University of Washington – Harborview Center for Prehospital Emergency Care, Seattle, Washington, USA
[3]Inserm, Bordeaux, France
[4]ISPED-Université Victor Segalen Bordeaux 2, Bordeaux, France
[5]Medical Informatics Department, Dijon University Hospital, Dijon, France

Prognostic studies are essential to understand the role of particular prognostic factors and, thus, improve prognosis. In most studies, disease progression trajectories of individual patients may end up with one of mutually exclusive endpoints or can involve a sequence of different events.

One challenge in such studies concerns separating the effects of putative prognostic factors on these different endpoints and testing the differences between these effects.

In this article, we systematically evaluate and compare, through simulations, the performance of three alternative multivariable regression approaches in analyzing competing risks and multiple-event longitudinal data. The three approaches are: (1) fitting separate event-specific Cox's proportional hazards models; (2) the extension of Cox's model to competing risks proposed by Lunn and McNeil; and (3) Markov multi-state model.

The simulation design is based on a prognostic study of cancer progression, and several simulated scenarios help investigate different methodological issues relevant to the modeling of multiple-event processes of disease progression. The results highlight some practically important issues. Specifically, the decreased precision of the observed timing of intermediary (non fatal) events has a strong negative impact on the accuracy of regression coefficients estimated with either the Cox's or Lunn-McNeil models, while the Markov model appears to be quite robust, under the same circumstances. Furthermore, the tests based on both Markov and Lunn-McNeil models had similar power for detecting a difference between the effects of the same

Received November 23, 2010; Accepted March 23, 2011
Address correspondence to Michal Abrahamowicz, Department of Epidemiology & Biostatistics, McGill University, McGill University Health Centre, Division of Clinical Epidemiology, 687 Pine Avenue West, V Building, Montreal, Que., H3A 1A1, Canada; E-mail: michal.abrahamowicz@mcgill.ca

covariate on the hazards of two mutually exclusive events. The Markov approach yields also accurate Type I error rate and good empirical power for testing the hypothesis that the effect of a prognostic factor on changes after an intermediary event, which cannot be directly tested with the Lunn-McNeil method. Bootstrap-based standard errors improve the coverage rates for Markov model estimates. Overall, the results of our simulations validate Markov multi-state model for a wide range of data structures encountered in prognostic studies of disease progression, and may guide end users regarding the choice of model(s) most appropriate for their specific application.

Keywords Cox's PH model; Lunn-McNeil model; Markov model; Multi-event process; Simulation.

Mathematics Subject Classification 62N01; 62P10.

1. Introduction

Many clinical prognostic studies attempt to model longitudinal processes of disease progression and resulting mortality, in which a patient may experience several clinical events, rather than just a single endpoint (Thein et al., 2009; Wu et al., 2010). These events can be either mutually exclusive, e.g., death due to a disease of interest versus due to other causes; or one event (e.g., a non fatal heart attack) may or may not precede another (e.g., death). Therefore, the appropriate analysis of such studies should account for, respectively, competing risks and/or alternative pathways of disease progression (le Cessie et al., 2009; Meira-Machado et al., 2009). However, the single-endpoint conventional time-to-event methods, such as Cox's proportional hazards model, are still the methods of choice for analyzing many such prognostic studies (Abouassaly et al., 2009; de Voogd et al., 2009; Kumar et al., 2009; Welten et al., 2008).

The conventional applications of Cox's regression to modeling of multi-event data involve fitting a separate Cox's model for each endpoint, while censoring subjects at the time of the competing event(s) (Putter et al., 2007). This approach has been refined by Lunn and McNeil, who proposed extending the Cox's model to the competing risks context, through data augmentation (Lunn and McNeil, 1995). The Lunn and McNeil (LM) method allows a simultaneous estimation of the (separate) effects of covariates on each type of event, as well as direct testing of the differences between the effects of the same prognostic factor on different competing events.

Another class of models particularly relevant for modeling disease progression processes involving multiple events are the Markov multi-state models, which generalize classic (single-event) time-to-event analyses to multiple outcomes (Andersen and Keiding, 2002). Markov models estimate the probability of transitions between different health states and determine how covariates affect the probability of each transition. Markov models are more general than the competing risks models in that the former are able to account not only for competing risks between mutually exclusive endpoints, but also for different multi-event pathways of transitions between consecutive states (Commenges, 1999; Hougaard, 1999). Similar to the competing risks models, they allow for formal testing of statistical significance of the differences between the effects of the same variable on the risks of different events. In an empirical study of colorectal

cancer, Dancourt et al. (2004) investigated several methodological issues related to modeling of disease progression, and compared different analytical approaches. They concluded that the Markov multi-state model provided a better insight into the course of cancer progression, and into the role of recurrence in this process, than the conventional Cox's proportional hazards model. However, to the best of our knowledge, no simulation studies have yet systematically evaluated and compared the accuracy of regression coefficients estimated with the alternative multivariable models for analyzing multi-event prognostic studies.

Another methodological issue, specific to multiple-events analyses, that requires a systematic evaluation, concerns testing of the hypotheses regarding the difference between the effects of the same prognostic factor on the risks of different events. Such testing is important to understand the disease evolution and may help identify subgroups of patients at high risk of particular events, which may ultimately enhance the effectiveness of preventive interventions and optimize allocation of limited resources (Freidlin and Korn, 2005).

In this article, we rely on simulations to investigate the above methodological issues, assuming different, clinically plausible scenarios, which involve multiple events, in the context of both "competing risks" and transitions through consecutive events. We assess and compare the performance of two survival analytical approaches involving: (i) fitting separate event-specifc Cox's models; (ii) the Lunn and McNeil extension of the Cox's model to competing risks analyses (Lunn and McNeil, 1995); as well as (iii) the Markov multi-state model MKVPCI developed by Alioum and Commenges (2001). The next section provides an overview of the above methods. Section 3 describes the simulated scenarios, as well as data generation and data analysis procedures. Section 4 summarizes the simulation results, and discussion in Sec. 5 concludes the article.

2. Methods Compared

2.1. *Cox's Proportional Hazards Model*

The very popular Cox's Proportional Hazards (PH) model estimates how the hazard of a single clinical endpoint depends on a vector of covariates (Cox, 1972):

$$h(t \mid Z) = h_0(t) \exp\left(\sum_{m=1}^{k} \beta_m z_m \right), \tag{1}$$

where $h(t \mid Z)$ is the hazard at time t conditional on covariate vector $\mathbf{Z} = (z_1, \ldots, z_k)$, $h_0(t)$ is the baseline hazard corresponding to $\mathbf{Z} = \mathbf{0}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is the vector of associated regression coefficients, i.e., the logarithm of hazard ratios associated with a unit increase in a given covariate (Cox, 1972; Klein and Moeschberger, 2003).

In the case of prognostic studies involving multiple events, typically a separate Cox's model is estimated for each event of interest (Dancourt et al., 2004). In each model, subjects who reach any of the "competing" events before the event of interest, are censored at that time (Dancourt et al., 2004; Putter et al., 2007). This approach does not allow for "direct" testing of whether the effects of a particular covariate on two different events of interest are the same, as these effects are estimated in *separate* models.

2.2. *Lunn and McNeil Competing Risks Model*

Lunn and McNeil (1995) proposed a nonparametric model that extends the Cox regression model to the competing risks framework under the assumption that the hazard functions for different event types are proportional. The Lunn and McNeil (LM) method involves estimating a single regression model, which incorporates all possible $C \geq 2$ failure types, by restructuring the data. In the competing risks context, a subject may fail of only one of the C distinct causes. As such, the observed time to failure of an individual is defined as the minimum of the potential failure times, i.e., the time to occurrence of the first event of any type. In the LM method, subject's data are duplicated C times, with separate rows corresponding to each failure type, and C-1 indicator variables (f_c) are created for event types $c = 2, 3, \ldots, C$, with Type I event being the "reference." Accordingly, the hazard, conditional on covariates, is modeled as (Lunn and McNeil, 1995):

$$h^*(t_j \mid z_i) = h_0^*(t_j) \exp\left[\sum_{c=2}^{C}\left(\beta_c f_c + \sum_{m=1}^{k} \theta_{cm} f_c z_m\right) + \sum_{m=1}^{k} \alpha_m z_m\right]. \quad (2)$$

In model (2), $h_0^*(t)$ is the baseline hazard, corresponding to covariate vector $\mathbf{Z} = \mathbf{0}$, for the "reference" event "1", and β_c represents the logarithm of the ratio of the baseline hazard for the event of type c ($2 \leq c \leq C$), relative to the "reference" baseline hazard, i.e., accounts for different incidence rates for different competing events. The coefficient α_m represents the effect of prognostic factor z_m on the "reference" hazard of the failure event of Type I. Finally, the coefficients for the "interaction terms" θ_{cm} account for the possibly different effects of prognostic factors on alternative outcomes. Specifically, the log HR for the effect of covariate z_m on the hazard of event c ($c \neq 1$) is estimated as $\alpha_m + \theta_{cm}$ (Lunn and McNeil, 1995). Robust variance estimates are used to account for the inter-dependence of C "observations" per subject (Wilcox, 1997).

An important advantage of the Lunn and McNeil method, in the competing risks context, is that it permits testing the "global" null hypothesis that the effect of a prognostic factor z_m is the same on all C event types. This is achieved through a (C-1) degrees-of-freedom (df) Likelihood Ratio Test (LRT) of all (C-1) interactions involving the prognostic factor z_m (Lunn and McNeil, 1995). The LM method also permits a direct test of a more detailed null hypothesis $\theta_{cm} = 0$, for $c \neq 1$, i.e., testing if the effect of z_m on the hazard of event c differs from its effect on the hazard of the "reference" event 1. Finally, testing the difference of the impact of z_m on two competing "non reference" events ($p \neq 1$ vs. $s \neq 1$) requires (i) restricting $\theta_{pm} = \theta_{sm}$, and (ii) then comparing the deviance of the resulting, restricted model with that of the un-restricted model, through a 1df LRT (Lunn and McNeil, 1995).

2.3. *Multi-State Markov Model: MKVPCI*

Another way of looking at event-time data is to consider an event as a *transition* from one state to another (Andersen and Keiding, 2002; Hougaard, 1999). In this context, multi-state models generalize the conventional survival methods to modeling of several events, which may involve either competing risks, or a sequence

of events, or a combination of these two types. The states analyzed in a multi-state model may include both *absorbing* states (e.g., death), which do not permit further transitions, and *transient* or "intermediary" states (e.g., non fatal heart attack or disease recurrence), which allow further transitions to another state (Andersen and Keiding, 2002; Hougaard, 1999).

A multi-state model is defined as a *stochastic process* $(Y(t), t \in T)$, with a finite *state space* $S = \{1, \ldots, k\}$. The *transition probabilities* are defined as (Andersen and Keiding, 2002):

$$P_{hj}(s, t) = P[Y(t) = j \mid Y(s) = h, \chi_s^-] \tag{3}$$

for $h, j \in S$, $s, t \in T$, $s \leq t$, so that the probability of transition from an (earlier) state h to the (later) state j may depend on the "history", i.e., on the sequence of states through which the subject has transitioned before time s (χ_s^-). *Transition intensities* are defined as the instantaneous risk of the change of the state:

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t, t + \Delta t)}{\Delta t}. \tag{4}$$

From a survival analysis perspective, the *transition intensity* in a multi-state model is equivalent to the hazard function (Andersen and Keiding, 2002; Hougaard, 1999). Markovian processes represent a special class of multi-state models, where the transition intensity depends on the "history" only through the current state (Hougaard, 1999).

In our simulations, we evaluate the time-homogeneous version of the MKVPCI multivariable Markov multi-state model developed by Alioum and Commenges (2001) for estimating the effects of covariates on the intensities of transitions between k states. In the time-homogeneous model, these intensities are assumed to remain constant over time. Therefore, for any given individual, the transition probabilities are assumed to depend only on (i) covariates, and (ii) the length of time interval (w), in which the transition may be observed, but not on the duration of follow-up until time $s : P_{hj}(s, s + w) = P_{hj}(0, w) = P_{hj}(w)$ for all s. The MKVPCI model allows for regression modeling of the transition intensities, under the proportional intensities assumption, that is consistent with the proportional hazards (PH) assumption underlying the Cox's model, which implies constant-over-time covariate effects:

$$\alpha_{hj}(t \mid Z(t)) = \alpha_{hj0} \exp\left(\sum_{m=1}^{k} \beta_{hjm} z_m(t) \right), \tag{5}$$

where $\mathbf{Z(t)} = (z_1(t), \ldots, z_k(t))$ is a matrix of, possibly time-dependent, covariates, α_{hj0} is the "baseline" intensity of the transition from h to j, corresponding to $\mathbf{Z(t)} = \mathbf{0}$, and $\beta_{hj} = (\beta_{hj1}, \ldots, \beta_{hjk})$ is a vector of constant-over-time regression coefficients. These coefficients describe the covariate effects on the intensity of transition from h to j. The MKVPCI estimation process produces full maximum likelihood estimates of both baseline transition intensities and regression coefficients (Alioum and Commenges, 2001).

In practice, the *exact* time of transition to a transient state, such as the onset of a disease, usually cannot be observed. Indeed, typically the occurrence of the

transition to a transient ("intermediary") state can be established only at discrete assessment times, corresponding, for example, to consecutive visits to the clinic. Therefore, the status of subject i is assessed only at m_i discrete times $t_{i,j}$, $j = 0, \ldots, m_i$, and the resulting vector of consecutive states is denoted as $y_{i,j} = Y(t_{i,j})$. Still, the user can set one of the states considered in the analysis as an absorbing state, and assume that transition times to this state are exactly known. In this case, the calculation of the likelihood is modified, and involves the Chapman–Kolmogorov equations (Alioum and Commenges, 2001).

In the MKVPCI Markov model, the null hypothesis of the equality of the effects of the same covariate on two different transitions ($H_0 : \beta_{hj} = \beta_{rv}$) is tested through a 1-df Wald-like test, which compares the model with the constraint: $\beta_{hj} = \beta_{rv}$ vs. the un-constrained model with the same covariates (Alioum and Commenges, 2001).

3. Simulation Scenarios and Data Generation

3.1. *Simulation Design and Data Generation*

We based the general design of our simulations on an empirical prognostic study of colon cancer progression, based on a population-based French colon cancer registry (Dancourt et al., 2004; Quantin et al., 1999). We generated the 10-year follow-up data for a hypothetical cohort of N subjects, diagnosed with colon cancer at time 0. Accordingly, the follow-up started with all patients in state 1 = *cancer diagnosis*. From there, they could have transitioned to either an intermediary state 2 = *cancer recurrence*, or state 3 = *death*, considered an absorbing state. All three transition intensities ($1 \to 2$, $1 \to 3$, and $2 \to 3$) were assumed to depend on three prognostic factors: age at diagnosis, sex, and cancer stage at diagnosis. Most scenarios focused only on competing risks between the two earlier transitions ($1 \to 2$ vs. $1 \to 3$). However, we also developed a scenario that involved testing of the hypothesis that the covariate effect on the hazard of death changed after the intermediary event 2 ($H_0 : \beta_{13} = \beta_{23}$). Thus, the simulated scenarios involved both "competing risks" of recurrence ($1 \to 2$) vs. recurrence-free death ($1 \to 3$), and transitions through consecutive events (death after recurrence: $2 \to 3$).

For simplicity, we focused on, and presented results only for age and sex (one continuous and one binary variable), although event times were always generated conditional on all three covariates. Depending on the simulated scenario, the effects of age and sex on different transitions could be the same or different. In all simulations, when generating times to death and to recurrence, we assumed that all prognostic factors effects conform to the PH assumption, i.e., true hazard ratios remained constant over the entire 10-year follow-up period. Moreover, all baseline transition intensities were assumed constant over time, consistent with exponential survival model and the time-homogeneous Markov model (Marshall et al., 1995).

Covariates. The three covariates were generated to correspond to their empirical distribution in the French colon cancer registry (Le Teuff et al., 2005). First, sex was generated with $P(\text{male}) = P(\text{female}) = 0.5$. Next, age at diagnosis was generated from a log-normal distribution, conditional on sex, with mean age of

71 years for men and 75 years for women. Cancer stage was generated from the marginal multinomial distribution, independent of age and sex, with probability of five consecutive, progressively more severe, stages equal to, respectively, 0.17, 0.35, 0.20, 0.06, and 0.22.

Outcomes. For each of N subjects, $i = 1, \ldots, N$, we first generated, independently of each other, two times, corresponding to the expected times for, respectively, transition $1 \rightarrow 3$, from *cancer diagnosis to death without recurrence* $T_i^{(1\rightarrow 3)}$, and transition $1 \rightarrow 2$ from *cancer diagnosis to cancer recurrence* $T_i^{(1\rightarrow 2)}$. Both times were generated from an exponential distribution, conditional on the covariates. For example, to generate times to recurrence-free death, we use the inverse formula: $T_i^{(1\rightarrow 3)} = (-\ln(u_i))/[\lambda_{13}\exp(\beta'_{13}\mathbf{Z}_i)]$, $i = 1, \ldots, N$, where u_i is an uniform [0; 1] random variable, $\mathbf{Z}_i$ is the generated covariate vector for the ith subject, $\boldsymbol{\beta}_{13}$ represents the vector of "true" covariate effects (log hazard ratios) on death, and λ_{13} the baseline exponential hazard of recurrence-free death, corresponding to $\mathbf{Z} = \mathbf{0}$ (Bender et al., 2005). Times to recurrence were generated in a similar fashion, except for a different baseline hazard λ_{12}, and, in simulations that focused on testing the differences in the covariate effects on alternative transitions, with $\beta_{12} \neq \beta_{13}$.

Finally, we generated N individual times $T_i^{(2\rightarrow 3)}$ *from recurrence to cancer death* (transition $2 \rightarrow 3$). The corresponding time from diagnosis ($t = 0$) to *death after recurrence* was then calculated as the sum of the times for each of the two consecutive transitions: $T_i^{(1\rightarrow 2)} + T_i^{(2\rightarrow 3)}$.

We introduced administrative censoring (AC) of all subjects who remained at risk until the end of the follow-up, i.e., at $AC = 10$ years. We also generated N expected drop-out times C_i, $i = 1, \ldots, N$, from uniform $U[0; AC/w]$ distribution, independent of covariates and outcomes, where $w < 1$ controlled the proportion of subjects expected to drop-out before AC, if there were no deaths (Le Teuff et al., 2005).

For each individual, the sequence of his/her transitions between states, and the observed total follow-up time τ_i, until either death or censoring, were determined by the order of the generated times for (a) different events, (b) drop-out, and (c) administrative censoring. The following sequences could occur. If the shortest time generated for an individual was the time to drop-out or administrative censoring, i.e., $\tau_i = \min\{T_i^{(1\rightarrow 3)}, T_i^{(1\rightarrow 2)}, C_i, AC\}$ corresponded to either C_i or AC, then subject i was censored without any event, i.e., remained in state 1 until the end of his/her follow-up, at $\tau_i = \min\{C_i, AC\}$. If $\tau_i = T_i^{(1\rightarrow 3)}$, then the subject transitioned directly from state 1 to state 3, and at τ_i the event 3 (recurrence-free death) was observed. On the other hand, if $\tau_i = T_i^{(1\rightarrow 2)}$, then subject i experienced transition $1 \rightarrow 2$, i.e., developed cancer recurrence at $T_i^{(1\rightarrow 2)}$. The subject could then further transition to the absorbing state 3 (death after recurrence) if $\min\{T_i^{(1\rightarrow 2)} + T_i^{(2\rightarrow 3)}, C_i, AC\} = T_i^{(1\rightarrow 2)} + T_i^{(2\rightarrow 3)}$. Otherwise, the subject remained in state 2 until the end of his/her follow up at $\tau_i = \min\{C_i, AC\}$.

The generated data were structured so that each subject had multiple entries, since non-absorbing states, such as recurrence (transient state 2), were considered to be assessed only at repeated, discrete observation times (Alioum and Commenges, 2001), the frequency of which was varied throughout the simulations (see Secs. 3.2 and 3.3 for details).

3.2. *Simulated Scenarios*

Several simulated scenarios were considered, in order to investigate different methodological issues relevant to the modeling of multiple-event processes of disease progression.

One practically important aspect of multi-event prognostic studies concerns the accuracy of the information regarding the times of transitions from one state to another. In real-life clinical and epidemiological studies, the occurrence of most transient events, such as cancer recurrence or loss of immunity, can be observed only at discrete "assessment times", corresponding to, for example, visits to the clinic or administration of the relevant laboratory test. Accordingly, the timing of transient events is usually known only as falling within a certain time interval, between the two assessment times. In contrast, the exact date of death is usually known. The first scenario investigated the impact of decreasing frequency of "assessment times", at which the occurrence of the transient event 2 could be determined, on the accuracy of regression coefficient estimates. To this end, we decreased the number of repeated observations (P) from $P = 20$ to $P = 10$ and $P = 5$. Since the total follow-up length was kept at 10 years, this implied increasing the length of the time intervals between consecutive observations from 6 months to, respectively, 1 and 2 years. The "true" generated effects of the prognostic factors for scenario 1, including sensitivity analyses, are shown in Table 1. In a sensitivity analysis, we changed the relative strength of the prognostic factor effects on the recurrence versus death without recurrence in order to assess how these changes will affect the results in the context of imprecise timing of the transient event.

In additional sensitivity analyses we varied the sample size. Specifically, we simulated datasets with $N = 250$, 500, 1,000 and, in limited experiments, we also considered a very large $N = 5{,}000$, while fixing the maximum number of repeated observations to $P = 5$.

The next two scenarios focused on issues specific to hypotheses testing in the context of multi-event analyses. The second scenario involved hypotheses of particular interest for "competing risks" analyses. Specifically, we tested the null hypothesis $H_0 : \beta_{12} = \beta_{13}$ for the binary covariate "sex." Since any test based on two *separate* Cox models ignores the possible covariance of the two estimates, we did not employ Cox's models in our analysis of the second scenario's data, and limited our comparisons to MKVPCI versus LM approaches. In scenario 2a, in order to estimate the Type I error rates, the covariate effects on both transitions were assumed to be equal ($\beta_{12} = \beta_{13} = \ln(2)$), corresponding to HR = 2 for men relative to women). In scenario 2b, in order to estimate the empirical power, we assumed the effects differed with $\beta_{12} = \ln(2)$ vs. $\beta_{13} = \ln(1.3)$.

In the third scenario, we focused on testing whether the effect of a prognostic factor on an absorbing event (state 3) changed after a transient event (state 2) had taken place. Because this situation does *not* involve competing risks, for which the Lunn and McNeil (1995) approach was developed, we compared only the Markov MKVPCI model with a time-dependent Cox's model. Specifically, the null hypothesis tested was $H_0 : \beta_{13} = \beta_{23}$ for the binary covariate "sex." As above, both effects were assumed to be equal for assessing Type I error, while to assess power, we assumed different effects ("true" values are shown in the Results section).

For both the second and third scenarios, the sample size was fixed at $N = 1{,}000$, the maximum number of repeated observations at $P = 20$, and the effects of other covariates were the same as shown in Table 1.

Table 1
Comparison of estimated prognostic factor effects between the three models. Sample size (N) = 500

Transition type	Prognostic factor (True value)	Repeated observations (P)	% Relative bias (95% CI)			% Coverage rate (CR) (% bootstrap CR)			RMSE ratio	
Column 1	2	3	COX 4	LM 5	MKVPCI 6	COX 7	LM 8	MKVPCI 9	COX/MKVPCI 10	LM/MKVPCI 11
1	AGE	$P = 5$	−9.3 (−15.0; −3.6)	−0.3 (−1.3; 0.8)	4.0 (−0.7; 8.6)	98 (97)	91 (93)	85 (91)	0.65	0.74
	ln(1.04)	$P = 10$	2.1 (−0.7; 4.9)	1.5 (−0.9; 3.9)	7.6 (1.0; 14.2)	98 (98)	94 (95)	87 (90)	0.67	0.77
		$P = 20$	5.7 (1.2; 10.3)	1.0 (−0.9; 3.0)	5.2 (0.9; 9.6)	98 (97)	98 (98)	92 (93)	0.79	0.77
2	SEX	$P = 5$	−9.6 (−15.3; −3.8)	2.8 (−0.4; 6.1)	−3.5 (−7.9; 0.8)	92 (90)	98 (98)	83 (88)	0.80	0.83
	ln(2)	$P = 10$	0.1 (−0.4; 0.6)	3.2 (−0.2; 6.7)	−0.2 (−1.4; 0.9)	95 (97)	93 (95)	81 (93)	0.73	0.77
		$P = 20$	4.6 (0.5; 8.6)	1.8 (−0.8; 4.4)	0.6 (−0.9; 2.1)	97 (92)	96 (96)	86 (95)	0.85	0.82
1	AGE	$P = 5$	−39.5 (−49.1; −29.9)	−37.6 (−47.1; −28.1)	−10.5 (−17.8; −3.2)	89 (85)	94 (89)	75 (85)	0.61	0.50
	ln(1.02)	$P = 10$	−15.8 (−22.9; −8.6)	−20.6 (−28.5; −12.7)	−7.2 (−13.6; −0.8)	94 (91)	98 (90)	69 (88)	0.50	0.43
		$P = 20$	−3.2 (−6.7; 0.2)	−9.7 (−15.5; −3.9)	2.8 (−0.4; 6.1)	95 (97)	93 (95)	77 (92)	0.50	0.53
3	SEX	$P = 5$	−51.9 (−61.7; −42.2)	−40.3 (−49.9; −30.7)	7.0 (0.9; 13.0)	82 (81)	94 (83)	78 (87)	0.74	0.54
	ln(1.3)	$P = 10$	−21.6 (−29.7; −13.5)	−22.8 (−31.0; −14.6)	18.4 (8.8; 28.1)	93 (89)	97 (92)	79 (85)	0.61	0.52
		$P = 20$	−6.0 (−10.7; −1.4)	−15.4 (−22.5; −8.4)	3.4 (−0.2; 6.9)	95 (95)	93 (91)	79 (91)	0.57	0.61

3.3. *Analyses of the Simulated Data*

For scenario 1, and in sensitivity analyses for this scenario, we generated 100 random samples for each combination of the relevant parameters considered. In scenarios 2 and 3, to increase precision of the estimated Type I error rates and empirical power of the tests of interest (Burton et al., 2006), we simulated 500 random datasets.

Each simulated dataset was analyzed with the three methods described in Sec. 2: (i) separate Cox's regression models for each event; (ii) LM "competing risks" method (Lunn and McNeil, 1995), except for scenario 3 (see Sec. 3.2); and (iii) time-homogeneous Markov multi-state MKVPCI model (Alioum and Commenges, 2001). Each model included, as independent variables, all three prognostic factors: age, sex, and cancer stage. The effect of age on the logarithm of the hazard for a given transition/event was *a priori* assumed to be linear, which was consistent with the true generated data structure.

Timing of the transient events. For the analyses involving the Cox's regression and the LM method, the transient event ($j = 2$) was considered to happen at the midpoint of the observation interval, in which the exact (assumed to be unknown) time, generated for a given subject, would fall. Therefore, in the analyses focusing on the absorbing event of "death" ($j = 3$), patients for whom event $j = 2$ was generated before $j = 3$, were censored at this interval's midpoint. In sensitivity analysis, we switched the censoring point to (i) either the beginning or (ii) the end of the interval, in order to investigate how this may affect the results.

In the Markov model analyses, for the transient event $j = 2$, the subject's status was assumed to change at the end of the interval, in which the exact generated event time would fall (Alioum and Commenges, 2001). In other words, if subject i experiences the transient event $j = 2$ at t_2^i, $t_m < t_2^i < t_{m+1}$, where $t_1 = 0$, and $t_2, \ldots, t_p$ are the consecutive times of repeated observations, then the status of this subject changes from 1 to 2 at t_{m+1}.

In contrast, for all three models, the exact time of the absorbing event 3 was assumed to be always known. However, for the analyses that employed Cox model with time to cancer recurrence as the outcome, subjects who died without recurrence were censored at the time of the last observation when they were known to be still alive. This analytical strategy corresponds to the FDA regulations (Frydman and Szarek, 2009), which account for the fact that in real life, information about the subject's recurrence-free status would not be available beyond this point.

The third simulated scenario (see Sec. 3.2) required testing whether the effect of a covariate "sex" on the hazard of the absorbing event ($j = 3$) changed after the transient event ($j = 2$) had occurred. For Cox regression analyses of data generated for this scenario, we implemented a Cox model that represented the occurrence of the transient event by a binary time-dependent covariate, which changed its value from 0 to 1 after event 2 was observed. We then tested the interaction between "sex" and this time-dependent covariate.

Criteria to assess the models' performance. The models' performance was assessed using several standard criteria related to, depending on the simulated scenario, accuracy of either regression coefficient estimates and/or hypotheses tests (Burton et al., 2006). Bias in the estimated effect of a prognostic factor was quantified as the difference between the mean of the estimates, from each of the simulated datasets, and the corresponding true log hazard ratio (β). The relative

bias was the ratio of the bias to the true value of β. The root mean square error (RMSE) for each of the three models was calculated as the square root of the sum of the squared bias and the empirical variance of the regression coefficient. Then, to compare the overall relative accuracy of the estimates obtained from different models, we calculated two ratios of the corresponding RMSE's, with the RMSE from either (i) the Cox model or (ii) the Lunn-McNeil model in the numerator, and the RMSE of the MKVPCI model in the denominator.

In addition to conventional "analytical" standard errors (SE) of the regression coefficients, in sensitivity analyses, we also estimated bootstrap SE's. The bootstrap SE was estimated as the standard deviation of the distribution of the corresponding regression coefficient across 300 bootstrap resamples. To assess the accuracy of both analytical and bootstrap-based standard error (SE) estimates, we calculated the ratio of the corresponding mean SE estimate to the empirical standard deviation of the corresponding log hazard ratio estimates, across the simulated samples. The empirical coverage rates of the nominal 95% confidence intervals (95% CI) were estimated, separately for analytical and bootstrap-based SE's, as the proportion of samples, in which the respective 95% CI included the true β. Type I error and empirical power were estimated as the proportion of samples in which a given test yielded a 'significant' result (2-tailed $p < 0.05$) when the null hypothesis was, respectively, true and false.

4. Simulations Results

Table 1 summarizes results of 100 simulation experiments with the sample size fixed at $N = 500$, resulting in about 200 transient events and about 200 absorbing events in each simulated sample. The maximum number of repeated observations varies from $P = 5$ to 10 and 20. Column 2 of Table 1 shows the "true" effect (β) of a prognostic factor on the intensity of each transition.

Results in columns 4 and 5 show that for both survival analytical methods (Cox and Lunn-McNeil (LM)), varying the frequency of repeated observations had a strong impact on the accuracy of the estimated effects of both prognostic factors on transition to recurrence-free death ($1 \rightarrow 3$), but not for transition to recurrence ($1 \rightarrow 2$). For example, for the Cox's model, the relative underestimation bias in the effect of age on the risk of death without recurrence increased from about -3% to about -39%, as the number of repeated observations decreased from $P = 20$ to $P = 5$ (Table 1, column 4). Similar pattern is seen for the binary variable (sex), and for both age and sex estimates obtained with the LM model (column 5).

In sensitivity analyses when, in the dataset used for the analysis, the presumed time to recurrence (transient event 2) was changed from the middle of the interval to the end of the interval, the underestimation bias of the effects of both covariates on the risk of transition $1 \rightarrow 3$ increased further (data not shown). For example, in the Cox's model analyses with $P = 5$, the log hazard ratios (β) for age and sex were underestimated by about 73% and 96%, respectively. In contrast, when the time to recurrence was assumed to correspond to the beginning of the interval, the same effects were *over-estimated* by about 34% and 49%, respectively (data not shown).

Unlike Cox's and Lunn-McNeil's estimates, the Markov model estimates seemed to be much less affected by the decreasing precision of the information on time to transition $1 \rightarrow 2$. For the effect of age on the risk of death without recurrence ($1 \rightarrow 3$), MKVPCI produced a relative biases of about $+3\%$, -7%, and

−10%, as the number of observations decreased from $P = 20$ to $P = 10$ and $P = 5$ (Table 1, column 6). Similarly, no trend toward a systematic under- or over-estimation bias was observed for the Markov estimates of the effect of sex on the same transition (Table 1, column 6). Overall, for a transition to the absorbing state 3 (with known exact event time), the Markov model produced substantially less biased estimates than the two survival analytical models if the timing of the "competing" transient event 2 was measured with low precision, and comparable, small bias if this time was measured with the precision as high as about 5% of the total follow-up time ($P = 20$). Notice that in Markov model, the transition is *always* assumed to occur at the end of the relevant time interval, i.e., at the time when it can be first established in clinical practice (Alioum and Commenges, 2001).

Interestingly, the accuracy of the estimates for the effects of either prognostic factor on transition $1 \rightarrow 2$ were largely unaffected by the decrease in the frequency of repeated observations, i.e., by decreasing precision in the measurement of time to this transition (upper half of Table 1). Indeed, for all three models and both covariates, the relative biases are small, varying between −9% and +8%. Whereas Markov estimates tend to be slightly more biased, the confidence intervals for the relative bias yielded by the three models show considerable overlap.

In all situations represented in Table 1, the Markov model produced estimates with a larger variance than the two survival models. The ratios of the empirical standard deviations (SD) for the estimates obtained with (i) the Cox model and (ii) the LM method, vs. MKVPCI ranged, respectively, from 0.45–0.81 and from 0.31–0.83 (data not shown). As a consequence of the inflated variance, the Root Mean Squared Errors (RMSE) of the Markov model estimates were systematically higher than for both survival analytical models, as shown by the RMSE ratios smaller than 1 in all rows of Table 1 (columns 10 and 11). Thus, with the sample size of $N = 500$, the inflated variance of the Markov estimates seems to outweigh the benefits of reduced bias, resulting in higher RMSEs.

However, in a sensitivity analysis, in which we increased the sample size by a factor of 10, to $N = 5,000$, the bias/variance trade-off between the accuracy of the estimates yielded by the three methods changed. As expected, for all methods, the variance decreased largely, while the relative bias was not systematically affected by the sample size (data not shown). Accordingly, even if the Markov model estimates still had larger variance, in the case of $P = 5$, their smaller bias resulted in RMSE for the Cox and LM being higher than for MKVPCI, with ratios ranging from 1.11–1.33 (data not shown).

The ratios of empirical SD's of the regression coefficients to the corresponding mean value of the conventional analytical SE were close to 1.0 for the Cox's and LM estimates but systematically below 1 for the Markov estimates (data not shown). As a consequence, the vast majority of the coverage rates produced by the two survival models were over 90% (Table 1 columns 7 and 8), while the Markov model produced lower coverage, which ranged between about 70% and 90% (column 9). In sensitivity analysis, we used 300 bootstrap samples in an attempt to better capture the empirical variance of the Markov model-based estimates, and therefore obtain coverage rates. Results presented in columns 7–9 of Table 1 show improved coverage rates for all 3 models in most cases with small bias. However, the most important improvement is seen for the Markov model, and, while still somewhat lower than for the two survival models, the bootstrap-based coverage rates range from 85–95%. Also, the bootstrap-based coverage rates became somewhat smaller for Cox and LM when effect estimates were strongly biased.

Table 2 summarizes results of 100 simulation experiments with the maximum number of repeated observations fixed at $P = 5$, implying subjects were evaluated at 2-year intervals. The table compares the results of the three models across three different sample sizes: 250, 500, and 1,000. The accuracy of the point estimates and confidence intervals for any of the three models is not strongly influenced by the decrease in sample size. For transition $1 \rightarrow 2$, the relative bias for the effects of both sex and age did not exceed $+/-10\%$. Regardless of the sample size, the Markov model produced significantly less biased estimates of the effects of both prognostic factors on the risk of the absorbing event ($1 \rightarrow 3$), compared to either survival analytical model. The consistently strong bias toward the null of the estimates obtained with both Cox and Lunn-McNeil models, ranging between 38% and 52%, did not diminish with increasing sample size. These results are consistent with those from Table 1 when the number of repeated observations is small ($P = 5$). The RMSE ratios (Table 2 columns 10 and 11) show a larger RMSE produced by the Markov model with respect to Cox and LM in all cases, with no clear trend as sample size increases. Coverage rates did not either vary systematically across the sample sizes for any of the models, but they were overall smaller for the Markov model (68–87%) than for both Cox's and Lunn-McNeil models (79–98%).

The coverage rates based on the bootstrap SE obtained in sensitivity analysis, presented in columns 7–9 of Table 2, show similar improvements coverage rates for all 3 models in most cases with small bias as in Table 1. For the Markov model the bootstrap-based coverage rates range from 82–91%, therefore reducing substantially the gap between the coverage rates for the three models.

In the main simulations, we assumed that the effects of age and sex were always stronger on the transition towards recurrence ($1 \rightarrow 2$) than on the transition towards recurrence-free death ($1 \rightarrow 3$). In sensitivity analysis, we inverted this pattern and imposed stronger effects of the prognostic factors on the recurrence-free death ($1 \rightarrow 3$). In the Cox's model with recurrence as the outcome, and recurrence-free death as the censoring event, the relative biases became substantially larger than in the main analysis. These biases further increased as the effects on the censoring event ($1 \rightarrow 3$) became larger relative to the effects on transition $1 \rightarrow 2$, i.e., on the event of interest (data not shown).

Table 3 compares the two models, LM and MKVPCI, with respect to (i) Type I error and (ii) empirical power, for testing the null hypothesis of the same effect of either sex or age on "competing risks" of transition $1 \rightarrow 2$ vs. transition $1 \rightarrow 3$. Sample size was fixed at 1,000, and the number of repeated observations at $P = 10$. Both models performed similarly, falsely rejecting the true null hypothesis in about 6% of the 500 simulated samples, i.e., yielding Type I error rates acceptably close to the nominal 0.05 significance level. When comparing the power of the different models to detect a difference in the effects of the binary variable (sex) on the two risks, the Lunn-McNeil (LM) approach yielded higher power. For a smaller "true" difference ($HR_{1\rightarrow 2} = 2.2$ vs. $HR_{1\rightarrow 3} = 1.5$), the LM-based test had a power of 53% [43%, 63%] vs. a power of 44% [34%, 54%] for the Markov model-based test. When increasing the "true" difference to $HR_{1\rightarrow 2} = 2.5$ vs. $HR_{1\rightarrow 3} = 1.5$, power increased to 92% [87%, 97%] for LM and to 77% [69%, 85%] for Markov.

In contrast, the LM approach could not be directly employed to test whether the effect of cancer stage on the hazard of death changed after the recurrence. Thus, in simulated "scenario 3" we compared only the time-dependent Cox's and Markov models' performance in testing this hypothesis. Both methods yielded

Table 2

Comparison of estimated prognostic factor effects between the three models. Number of repeated observations (P) = 5

Transition type	Prognostic factor (True value)	Repeated observations (P)	% Relative bias (95% CI)			% Coverage rate (CR) (% bootstrap CR)			RMSE ratio	
			COX	LM	MKVPCI	COX	LM	MKVPCI	COX/MKVPCI	LM/MKVPCI
Column 1	2	3	4	5	6	7	8	9	10	11
1 → 2	AGE	$N = 250$	−5.0 (−9.2; −0.7)	1.1 (−0.9; 3.2)	5.1 (0.8; 9.4)	93 (95)	89 (96)	80 (88)	0.73	0.80
	ln(1.04)	$N = 500$	−9.3 (−15.0; −3.6)	−0.3 (−1.3; 0.8)	4.0 (−0.7; 8.6)	98 (97)	91 (93)	85 (91)	0.65	0.74
		$N = 1{,}000$	−6.0 (−10.7; −1.4)	−0.5 (−1.8; 0.9)	0.002 (−0.1; 0.1)	92 (95)	87 (94)	79 (90)	0.72	0.80
	SEX	$N = 250$	−6.5 (−11.4; −1.7)	0.5 (−0.9; 1.9)	−9.5 (−15.3; −3.8)	96 (96)	95 (96)	87 (91)	0.62	0.69
	ln(2)	$N = 500$	−9.6 (−15.3; −3.8)	2.8 (−0.4; 6.1)	−3.5 (−7.9; 0.8)	92 (90)	98 (98)	83 (88)	0.80	0.83
		$N = 1{,}000$	−7.5 (−12.7; −2.4)	−0.8 (−2.6; 0.9)	−9.1 (−14.7; −3.5)	88 (89)	90 (93)	81 (89)	0.80	0.85
1 → 3	AGE	$N = 250$	−34.7 (−44.0; −25.4)	−41.5 (−51.2; −31.9)	−13.8 (−20.6; −7.1)	95 (88)	98 (91)	71 (86)	0.60	0.46
	ln(1.02)	$N = 500$	−39.5 (−49.1; −29.9)	−37.6 (−47.1; −28.1)	−10.5 (−17.8; −3.2)	89 (85)	94 (89)	75 (85)	0.61	0.50
		$N = 1{,}000$	−38.4 (−47.9; −28.8)	−39.3 (−48.9; −29.7)	−12.5 (−19.0; −6.0)	79 (78)	85 (80)	68 (82)	0.61	0.56
	SEX	$N = 250$	−42.6 (−52.2; −32.9)	−42.8 (−52.4; −33.1)	19.9 (12.1; 27.7)	88 (80)	97 (87)	76 (86)	0.55	0.41
	ln(1.3)	$N = 500$	−51.9 (−61.7; −42.2)	−40.3 (−49.9; −30.7)	7.0 (0.9; 13.0)	82 (81)	94 (83)	78 (87)	0.74	0.54
		$N = 1{,}000$	−51.1 (−60.9; −41.3)	−45.4 (−55.1; −35.6)	8.3 (2.9; 13.9)	68 (73)	83 (78)	65 (82)	0.72	0.59

Table 3
Comparison of Type I error and Power results for testing: $H_0 : \mathrm{HR}_{1\to2} = \mathrm{HR}_{1\to3}$

$N = 1{,}000$, $P = 20$	True value	Lunn-McNeil	MKVPCI
Type I error	$H_0 : \mathrm{HR}_{1\to2} = \mathrm{HR}_{1\to3}$	6% (4.5%, 7.5%)	6.2% (4.7%, 7.7%)
Power	$\mathrm{HR}_{1\to2} = 2.2$ vs. $\mathrm{HR}_{1\to3} = 1.5$	53% (43%, 63%)	44% (34%, 54%)
	$\mathrm{HR}_{1\to2} = 2.5$ vs. $\mathrm{HR}_{1\to3} = 1.5$	92% (87%, 97%)	77% (69%, 85%)

correct Type I error rates: the time-dependent Cox's model falsely rejected the "true" null hypothesis of no difference in 5.2% of the 500 simulated samples, while the Markov model in 4.8%. When testing a "true" difference in the effects of cancer stage on death before or after recurrence, with $\mathrm{HR}_{1\to3} = 3$ vs. $\mathrm{HR}_{2\to3} = 2.5$, the test of the time-dependent interaction in Cox's model had an empirical power of 82% [75%, 90%], i.e., very similar to the Markov model-based test, which produced a power of 79% [71%, 87%].

5. Discussion

We have systematically evaluated and compared, through simulations, the performance of alternative multivariable regression methods for modeling prognostic studies, in which a patient may experience more than one type of event. These events, that may be also considered to represent transitions between consecutive "states," may be either mutually exclusive, as in the competing risks analyses, or occur one after another. To model the effects of baseline prognostic factors on the risks of particular transitions, we considered three different statistical methods. Firstly, we adapted the conventional single-endpoint Cox's proportional hazards (PH) model to the analyses of multi-event data. In the "competing risks" framework, this involved using separate Cox's regression analyses to model the conditional hazard of each of mutually exclusive events, with the right censoring on the competing event(s). On the other hand, when using the Cox's model, the sequence of consecutive events was modeled by introducing a binary time-dependent "change-of-state" covariate, which changed its value from 0 to 1 at the time the transient event had occurred. The second method considered relied on the extension of the conventional Cox's PH model, to the competing risks analyses, through a specific data augmentation and manipulation proposed by Lunn and McNeil (1995). Finally, the third method involved the MKVPCI multi-state Markov model (Alioum and Commenges, 2001). To the best of our knowledge, the performance of these methods was not systematically compared and/or evaluated in the context of $k \geq 2$ events, and to date their relative advantages or weaknesses in such multi-state analyses have remained unclear. In order to gain a broad insight into the performance of the three methods, we investigated different clinically plausible simulation scenarios.

Our results highlight some practically important issues related to a frequent limitation of real-life clinical data, where occurrence of many non fatal ("transient") endpoints can be established only at discrete time points, corresponding to clinic

visits or pre-scheduled assessment times. We demonstrated that, in such situations, the decreased precision of the observed timing of these events had a strong impact on the accuracy of regression coefficient estimates obtained with either the Cox's or the Lunn-McNeil models. Indeed, in our simulations, the effects of the prognostic factors on the risk of the absorbing event 3 (with exact transition times known), estimated in either model, were strongly underestimated when we increased the interval between the assessment times for the competing (transient) event 2, with the interval-censored transition times. Furthermore, sensitivity analyses showed that the bias around the estimates of these effects changed the magnitude and/or the direction, depending on where the time to event 2 was placed within the interval, in which the "true" generated time fell.

This pattern of results can be explained by a combination of (i) informative censoring, and (ii) exponential distribution of event times. We assumed that older age and male sex were associated with increasing risks of both event 2 *and* event 3. Therefore, in the analyses focusing on event 3, censoring at the time of event 2 implied an informative censoring. On the other hand, exponential distribution implies that, within each time interval, the censoring events will be more frequent near the beginning than near the end of the interval, so that the median follow-up time for subjects censored within each interval will be shorter than the interval's midpoint. As a result, assuming, as we did, that all the censoring events ($j = 2$) happened at the midpoint of the interval, will overestimate the total person-times "at risk" of those subjects who had the censoring event. Thus, the denominator of the hazard (incidence) rate for the endpoint of primary interest (here event $j = 3$) will be overestimated, and the resulting bias will increase as the length of the inter-assessments intervals increases. In contrast, because the timing of all absorbing events $j = 3$ is assumed to be known exactly, the length of the intervals will not affect the numerator of the estimated hazard rate. Accordingly, the hazard itself will be under-estimated because of the spuriously inflated denominator. As a consequence, if, e.g., males are more prone to both cancer recurrence (2) and death (3), the total "at risk" person-years for male subjects will be more overestimated than for females, and the male/female hazard ratio will be underestimated. In contrast, the true person-time "at risk" will be underestimated if recurrence (the censoring event) is assumed to happen at the beginning of the interval, in which the true time to recurrence was generated. Then, the mechanism described above will induce an under-estimation of the denominators of the hazard rates, and the resulting bias will be stronger for males, therefore, leading to an overestimation of the male/female hazard ratio. In real-life applications such biases can lead to misleading conclusions and sub-optimal treatment decisions. For instance, if the impact of a modifiable prognostic factor is under-estimated, then physicians might decide *not* to allocate necessary resources to the treatment of patients affected by this factor, thus, depriving them of a potential treatment benefit.

In contrast to the Cox's and LM estimates, the MKVPCI model's estimates of the effects of prognostic factors on the risk of recurrence-free death proved to be quite robust with respect to decreasing precision of the timing of the competing transient event (recurrence). Accordingly, especially for a small number of repeated observations, the Markov model produced more accurate estimates of covariate effects on transition $1 \rightarrow 3$, across a range of sample sizes, while estimates obtained with both survival models continued to be strongly biased even when the sample size increased.

The performance of the Cox model, in which the recurrence was the outcome of interest, was affected if subjects who died without recurrence were censored at the last assessment time when they were still "alive," an approach that followed the FDA recommendations (Frydman and Szarek, 2009). In sensitivity analysis, the overestimation bias increased as the true effects of the prognostic factors on the censoring event became stronger. This is most likely explained by the same mechanism described above, in the context of censoring on recurrence at the beginning of the interval. When the effects of the prognostic factors on the censoring event are small in comparison with their effects on the outcome, as in our main scenario, this bias is less marked.

On the other hand, in most simulations we observed an increased variance for the Markov MKVPCI model estimates, as compared to either Cox's or LM estimates. This inflated variance of the Markov model-based estimates may be due to the increased complexity of the model that requires *simultaneous* estimation of the covariate effects on each of the logically possible transitions between consecutive states, as well as of the baseline hazards for each transition (Alioum and Commenges, 2001). Because of the resulting variance inflation, the overall accuracy of the Markov model estimates of the hazard ratios, as measured by the RMSE, was often lower than for the two alternative methods, even if the Markov-based point estimates tended to be less biased. Whereas the bias-variance trade-off offered by the Markov MKVPCI model gradually improved with increasing sample size, it still points out the need for further research to attempt to stabilize the point estimates. It should be noted that this pattern of our simulation results resembles findings reported in many other areas of statistical research, where more complex models may often tend to yield less biased but also less numerical stable estimates than simpler, conventional methods (Le Teuff et al., 2005; Ionescu-Ittu et al., 2009; Abrahamowicz et al., 1996). In a sensitivity analysis, we have evaluated the variance of the Markov model estimates in the case of a very large sample size ($N = 5{,}000$). In that case, the variance of all estimates decreased substantially, while the bias did not change systematically, relative to smaller sample sizes. As a consequence, the bias-variance trade-off tended to favor the (less biased) Markov model estimates, whose RMSE became lower than for the Cox's and LM estimates.

Simulation results confirmed also our expectation that the accuracy of the "analytical" standard errors (SE), derived from asymptotic maximum likelihood estimation theory (Alioum and Commenges, 2001), in the Markov model would improve with increasing sample size. In contrast, with sample sizes below 1,000, the empirical variance of the Markov model estimates was systematically underestimated, resulting in sub-optimal coverage rates of the nominal 95% confidence intervals. This suggests that part of the inflated variability of the Markov estimates is not captured by the "analytical" SE, and that a robust "sandwich" variance estimator (Carroll and Kauermann, 2001), or a bootstrap-based SE (Davison and Hinkley, 1997) may be useful to enhance the accuracy of the SE and confidence intervals. Therefore, in sensitivity analysis we estimated the SE's based on 300 bootstrap resamples. As expected the bootstrap approach improved substantially the coverage rates. Based on these results, we tentatively suggest that coverage rates for Markov model-based regression coefficients should be estimated through bootstrap.

An important advantage of both Markov and Lunn-McNeil models, over fitting separate event-specific Cox's models, is that they allow a "direct" model-based testing of the hypotheses regarding the difference between the effects of the

same prognostic factors on different competing risks (Alioum and Commenges, 2001; Lunn and McNeil, 1995). Indeed, the results of our simulations showed that both models had good power for detecting a "true" difference between the effects of sex on two mutually exclusive transitions, although LM performed somewhat better than the Markov model. In real-life prognostic studies, an efficient test will help identify potentially important differences between the effects of a prognostic factor on competing diseases or events (Dancourt et al., 2004; Freidlin and Korn, 2005). Establishing such differences may help understanding the etiology of the disease, and/or targeting appropriate preventive interventions and treatments to the (correctly identified) subgroups of patients at higher risk of particular outcomes.

On the other hand, whereas the Lunn-McNeil method allows testing hypotheses regarding competing risks, it cannot handle a sequence of different events, one of which precedes the other(s). In contrast, the multi-state Markov model provides a "direct" test of the hypothesis that the effect of a prognostic factor changes after a transient event (e.g., cancer recurrence) has occurred. Simulation results indicated that MKVPCI model-based test performed well in this context. Specifically, both the Type I error rate and the empirical power were similar to the Cox's model-based test of an interaction between a prognostic factor of interest and a time-dependent binary indicator of the transient event. From this perspective, it should be noted that Markov models allow incorporating several transitions and testing, e.g., the compound null hypothesis that the effects of a given factor on all transitions are equal (Alioum and Commenges, 2001). On the other hand, incorporating three or more intermediate events in the Cox's model would require constructing a series of time-dependent covariates, and testing similar compound hypotheses would involve creating complex inter-related interactions, making the entire process very burdensome for the end-users.

Overall, our simulations suggest that each of the three models has its specific strengths and weaknesses, depending on the true structure of the data and some characteristics of the underlying disease progression pathways. Our results indicate that Markov multi-state models may be able to provide acceptably accurate results in a wider range of analyses of multi-event processes than simpler methods based on a series of separate Cox's regression analyses or on the extension of the Cox's model to competing risks proposed by Lunn and McNeil (1995). On the other hand, while yielding less biased point estimates of the effects of prognostic factors on the risks of different transitions, the Markov model tends to inflate their variance. To the best of our knowledge, no published studies provided such insights into the several aspects of the Markov models performance, or allowed a systematic comparison of these models with simpler alternative approaches. From this perspective, the results of our simulations may both encourage a more widespread use of the Markov multi-state models in real-life prognostic studies of disease progression, and alert the future users to some limitations of these models. We also hope that the awareness of these limitations will motivate further developments in multi-state modeling methodology.

As in most simulation studies, the assumptions under which our data were generated were somewhat arbitrary (Bender et al., 2005). Specifically, we considered only the exponential distributions for times to the different competing events, the number of prognostic factors and the number of transitions were restricted to three and, except for a limited sensitivity analysis with $N = 5{,}000$, the sample size did not exceed 1,000. Further studies should consider more variety of the functional forms of the event-specific baseline hazards, larger sets of covariates, larger sample sizes,

and more complex models for transitions between $k > 3$ states. Still, we believe that many empirical findings reported in this article will prove quite robust with respect to the changes in the simulation design and in the values of the relevant design parameters.

In conclusion, our simulation study provided some new insights into the different methodological issues related to analyzing prognostic studies involving multiple events and, therefore, may provide some guidance for end-users of the methods we considered, as well as stimulate further statistical research on refining these methods.

Acknowledgments

We would like to thank Marie-Eve Beauchamp for her careful review of the article. Michal Abrahamowicz is a James McGill Professor at McGill University and this research was partly supported by his grants from the National Sciences and Engineering Research Council of Canada (#228203) and the Canadian Institutes for Health Research (#MOP-8127).

References

Abouassaly, R., Paciorek, A., Ryan, C. J., Carroll, P. R., Klein, E. A. (2009). Predictors of clinical metastasis in prostate cancer patients receiving androgen deprivation therapy. *Cancer* 19:4470–4476.

Abrahamowicz, M., MacKenzie, T., Esdaile, J. M. (1996). Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *JASA* 91:1432–1439.

Alioum, A., Commenges, D. (2001). MKVPCI: A computer program for Markov models with piecewise constant intensities and covariates. *Computer Methods and Programs in Biomedicine* 64:109–119.

Andersen, P. K., Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* 11:91–115.

Bender, R., Augustin, T., Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 11:1713–1723.

Burton, A., Altman, D. G., Royston, P., Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine* 24:4279–4292.

Carroll, R. J., Kauermann, G. (2001). A note on the efficiency of Sandwich covariance matrix estimation. *Journal of the Americal Statistical Association* 96:1387–1396.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis* 5:315–327.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* B34:187–220.

Dancourt, V., Quantin, C., Abrahamowicz, M., Binquet, C., Alioum, A., Faivre, J. (2004). Modeling recurrence in colorectal cancer. *Journal of Clinical Epidemiology* 57:243–251.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

de Voogd, J. N., Wempe, J. B., Koëter, G. H., Postema, K., van Sonderen, E., Ranchor, A. V., Coyne, J. C., Sanderman, R. (2009). Depressive symptoms as predictors of mortality in patients with COPD. *Chest* 3:619–625.

Freidlin, B., Korn, E. L. (2005). Testing treatment effects in the presence of competing risks. *Statistics in Medicine* 11:1703–1712.

Frydman, H., Szarek, M. (2009). Nonparametric estimation in a Markov "illness-death" process from interval censored observations with missing intermediate transition status. *Biometrics* 65:143–151.

Hougaard, P. (1999). Multi-state models: A review. *Lifetime Data Analysis* 5:239–264.

Ionescu-Ittu, R., Delaney, J. A., Abrahamowicz, M. (2009). Bias-variance trade-off in pharmacoepidemiological studies using physician-preference-based instrumental variables: A simulation study. *Pharmacoepidemiology Drug Safety* 19:562–571.

Klein, J. P., Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer-Verlag.

Kumar, S., Shah, J. P., Bryant, C. S., Imudia, A. N., Morris, R. T., Malone Jr., J. M. (2009). A comparison of younger vs older women with vulvar cancer in the United States. *American Journal of Obstetrics and Gynecology* 5:e52–e55.

le Cessie, S., de Vries, E. G., Buijs, C., Post, W. J. (2009). Analyzing longitudinal data with patients in different disease states during follow-up and death as final state. *Statistics in Medicine* 30:3829–3843.

Le Teuff, G., Abrahamowicz, M., Bolard, P., Quantin, C. (2005). Comparison of Cox's and relative survival models when estimating the effects of prognostic factors on disease-specific mortality: a simulation study under proportional excess hazards. *Statistics in Medicine* 24:3887–3909.

Lunn, M., McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics*. 51(2):524–532.

Marshall, G., Guo, W., Jones, R. H. (1995). MARKOV: a computer program for multistate Markov models with covariates. *Computer Methods and Programs in Biomedicine* 47:147–156.

Meira-Machado, L., de Uña-Alvarez, J., Cadarso-Suárez, C., Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research* 2:195–222.

Putter, H., Fiocco, M., Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 11:2389–2430.

Quantin, C., Abrahamowicz, M., Moreau, T., Bartlett-Esquilant, G., MacKenzie, T., Tazi, M. A., Lalonde, L., Faivre, J. (1999). Variation over time of the effects of prognostic factors in a population based study of colon cancer: Comparison of statistical models. *American Journal of Epidemiology* 150:1188–1200.

Thein, H. H., Yi, Q., Heathcote, E. J., Krahn, M. D. (2009). Prognosis of hepatitis C virus-infected Canadian post-transfusion compensation claimant cohort. *Journal of Viral Hepatitis* 11:802–813.

Welten, G. M., Schouten, O., Hoeks, S. E., Chonchol, M., Vidakovic, R., van Domburg, R. T., Bax, J. J., van Sambeek, M. R., Poldermans, D. (2008). Long-term prognosis of patients with peripheral arterial disease: a comparison in patients with coronary artery disease. *Journal of the American College of Cardiology* 16:1588–1596.

Wilcox, R. R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*. New York: Academic Press.

Wu, J. C., Hakama, M., Anttila, A., Yen, A. M., Malila, N., Sarkeala, T., Auvinen, A., Chiu, S. Y., Chen, H. H. (2010). Estimation of natural history parameters of breast cancer based on non-randomized organized screening data: subsidiary analysis of effects of inter-screening interval, sensitivity, and attendance rate on reduction of advanced cancer. *Breast Cancer Research and Treatment* 122(2):553–566.