Lifetime Data Anal (2009) 15:216–240
DOI 10.1007/s10985-008-9106-0

A semi-Markov multistate model for estimation of the mean quality-adjusted survival for non-progressive processes

Gisela Tunes-da-Silva · Antonio C. Pedroso-de-Lima · Pranab K. Sen

Received: 23 January 2008 / Accepted: 3 November 2008 / Published online: 10 December 2008

Abstract We discuss the estimation of the expected value of the quality-adjusted survival, based on multistate models. We generalize an earlier work, considering the sojourn times in health states are not identically distributed, for a given vector of covariates. Approaches based on semiparametric and parametric (exponential and Weibull distributions) methodologies are considered. A simulation study is conducted to evaluate the performance of the proposed estimator and the jackknife resampling method is used to estimate the variance of such estimator. An application to a real data set is also included.

Keywords Quality of life · Survival analysis · Multistate models · Exponential distribution · Weibull distribution · Semiparametric proportional hazards

1 Introduction

Over the last decade, the concept of health has been broadened, comprising physical, functional, mental and social well being in addition to survival and disease-specific responses. In this new setting, the quality of life of patients become an important aspect to be considered in statistical analyses. As pointed out by Fairclough (1997), the main

G. Tunes-da-Silva · A. C. Pedroso-de-Lima (✉)
Departamento de Estatística, Universidade de São Paulo, Rua do Matão, 1010,
CEP 05508-090 São Paulo, SP, Brazil
e-mail: acarlos@ime.usp.br

G. Tunes-da-Silva
e-mail: tunes@ime.usp.br

P. K. Sen
The University of North Carolina at Chapel Hill, Biostatistics, CB #7420, Chapel Hill, NC 27599, USA
e-mail: pksen@bios.unc.edu

objective of studies concerned with quality of life assessment is to compare the overall quality of life related to different medical treatments. Most of these studies, however, consider only the quality of life and do not include survival times or disease-specific responses. On the other hand, only the survival time is considered when comparing treatments based on the usual survival methodologies. A better criteria of comparison should include both survival times and quality of life. This idea becomes even more appealing when treatments provide the same survival experience for the patients.

In this context, new methodologies have been developed in order to take into account both, survival and quality of life. As an example, we may refer to the joint estimation of quality of life, assessed longitudinally, and time to event, as presented by Pocock et al. (1987), or the application of multistate models, as suggested by Andersen and Keiding (2002).

The quality-adjusted survival time is an alternative approach that has the advantage of incorporating both survival and quality of life information in a single response variable. The idea of quality-adjusted survival was introduced by Gelber et al. (1989), with the Q-TWiST (Quality-adjusted Time Without Symptoms of disease and Toxicity of treatment) methodology.

Several authors have discussed the quality-adjusted survival. For right censored data, the classical Kaplan–Meier estimator based on the quality-adjusted survival time may not perform well, and Zhao and Tsiatis (1997) proposed an alternative estimator based on weighted estimating equations, assuming independent censoring. The estimator was further modified by Zhao and Tsiatis (1999) in order to increase its efficiency. Another approach to estimate the survival distribution of the quality-adjusted survival time, allowing for covariates, was proposed by van der Laan and Hubbard (1998). The comparison of survival functions of quality-adjusted lifetime for two different treatments when the data is right-censored was considered by Zhao and Tsiatis (2001). The estimation of the mean quality-adjusted lifetime was also discussed by Zhao and Tsiatis (2000) and an estimator based on influence functions was derived by Robins et al. (1994). The estimation of the mean quality-adjusted survival when clinical evaluations are made on a periodic basis (e.g., monthly) was considered by Chen and Sen (2001). They also considered that patients could experience more than one type of health status between two visits.

Tunes-da-Silva et al. (2008) proposed an estimator for the mean quality-adjusted survival time using a multistate model for the sojourn times, considering parametric as well as semiparametric approaches. The estimator was developed based on the assumption that the sojourn times in each health state are independent and identically distributed random variables for a given vector of covariates. In this paper, we will generalize that estimator allowing the sojourn times to be non-identically distributed. In Sect. 2 we present the definition for the quality-adjusted survival time and obtain an expression for its expected value in this new setting. In Sect. 3 parametric and semiparametric models for the sojourn times are described and in Sect. 4 we deal with the three state model. Section 5 contains a real data example. Finally, simulation studies are presented and discussed in Sect. 6. Further discussions and concluding remarks are presented in Sect. 7.

2 The mean quality-adjusted survival time

In order to formally define the quality-adjusted survival time, assume that n individuals are being followed up and consider that the health history of the i-th patient can be described by a process $\{V_i(t), t \geq 0\}$, where $V_i(t)$ may assume any of the $K+1$ states belonging to the state space $\Gamma = \{0, 1, \ldots, K\}$. Suppose that the states $1, 2, \ldots, K$ are transient and the state 0 is absorbing so that if $V_i(t) = 0$, then $V_i(s) = 0$, $\forall s \geq t$. The usual survival time for the i-th individual is given by $T_i(t) = \inf\{t : V_i(t) = 0\}$. Define also the function Q that maps the state space into a pre-specified set of real numbers (the utility coefficients). By this notation, the quality-adjusted survival time is given by

$$U_i = \int_0^{T_i} Q\{V_i(t)\}dt = \int_0^{\infty} Q\{V_i(t)\}dt, \quad i = 1, \ldots, n. \tag{1}$$

The function Q may also be defined such that it depends on both health state and time, i.e., $Q\{V_i(t), t\}$, meaning that the quality of life associated to each state may change over time. However, this situation will not be considered in this paper.

It is assumed each patient is observed as long as the absorbing state is not reached nor the patient is censored, and that the occurrence of any change in health state during the period the patient is being followed up is known. We also consider that a transient state can be visited more than once and denote by $T_j^{(k)}$ the sojourn time in the j-th visit to a state k. Note that it is not assumed a progressive structure in this case, but progressive processes are special cases in this general setting. Using the above introduced notation and considering that the quality of life remains constant when a patient is in a given health state, the quality-adjusted survival time can be simplified as

$$U = \int_0^{\infty} Q\{V(t)\}dt = q_1 \sum_{j=1}^{N_1} T_j^{(1)} + q_2 \sum_{j=1}^{N_2} T_j^{(2)} + \cdots + q_K \sum_{j=1}^{N_K} T_j^{(K)},$$

where q_k is the coefficient related to the k-th health state and N_k is the number of visits to state k, $k = 1, 2, \ldots, K$. When there are only two states, one corresponding to perfect health and the other corresponding to death, with utility coefficients equal to 1 and 0, respectively, the quality-adjusted survival time reduces to the usual survival time.

It is important to note that censoring has an informative pattern in the quality-adjusted time scale. Glasziou et al. (1990) argue that patients with poor quality of life tend to *accumulate* the quality-adjusted survival time slowly and, because of that, small censored quality-adjusted survival times are associated with poor quality of life.

The main goal of this paper is related to the estimation of the mean quality-adjusted lifetime for a given vector $\mathbf{Z}$ of covariates, given by

$$\mu_Q = \mathrm{E}(U|\mathbf{Z}) = \mathrm{E}\left(\int_0^\infty Q\{V(t)\}dt\middle|\mathbf{Z}\right)$$
$$= q_1\mathrm{E}\left[\sum_{j=1}^{N_1} T_j^{(1)}\middle|\mathbf{Z}\right] + q_2\mathrm{E}\left[\sum_{j=1}^{N_2} T_j^{(2)}\middle|\mathbf{Z}\right] + \cdots + q_K\mathrm{E}\left[\sum_{j=1}^{N_K} T_j^{(K)}\middle|\mathbf{Z}\right]. \quad (2)$$

Expression (2) is general and can be applied in different situations. In this paper, we consider that the mean time spent in some health states may decrease as the number of previous visits to that state increases. More specifically, it is assumed that the expectation of the sojourn times in state k may decrease for $k = 1, \ldots, r$ and, for $k = r + 1, \ldots, K$, the sojourn times $T_j^{(k)}$ are random variables with the same distribution, for a given vector of covariates. The number r of states whose mean sojourn times may decrease depends on the process considered. In a illness-death process, for example, it may be reasonable to assume the frequency of illness episodes increases with time, i.e., the mean sojourn time in the healthy state decreases and it is evident that $r = 1$.

It is also assumed a competitive risk structure for the sojourn times in the states, i.e., the observed sojourn time in state k is

$$T_j^{(k)} = \min_{l \in B_{(k)}} \{T_j^{k\to l}\},$$

where $T_j^{k\to l}$ is the time spent in state k up to a transition to state l (which may not be observable) and $B_{(k)}$ is the set of all states that can be reached from k.

In order to work with expression (2), it is needed to compute the quantities $\mathrm{E}\left[\sum_{j=1}^{N_k} T_j^{(k)}\middle|\mathbf{Z}\right]$, $k = 1, \ldots, K$. To simplify the notation, from now on we will write this expectation as $\mathrm{E}\left[\sum_{j=1}^{N_k} T_j^{(k)}\right]$, i.e., the vector of covariates $\mathbf{Z}$ will be omitted, but all results will be obtained for a given vector of covariates. We will compute this expectation separately for the states in which the mean sojourn time decreases and for the states in which the mean sojourn time remains constant.

Consider initially the states for which the mean sojourn time decreases as the number of previous visits increases. It is assumed the distribution functions of the sojourn times $T_j^{(k)}$, $k = 1, \ldots, r$ belong to the Lehmann family (see Hajek et al. 1999), i.e., the specific hazard function for the transition from state k to state l in the j-th visit to k is given by

$$\lambda_j^{k\to l}(t) = (1 + d_j^{(k)})\lambda_1^{k\to l}(t),$$

$l \in B_{(k)}$, where $j = 1, \ldots, N_k$ accounts for the number of visits to state k, $B_{(k)}$ is the set of all states that can be reached immediately after state k and $d_j^{(k)}$ are constants, to be discussed later on. Assuming that

$$\lambda_1^{k\to l}(t) = \lambda_\circ^{k\to l}(t)e^{\boldsymbol{\beta}_{kl}^T\mathbf{Z}},$$

we have

$$\lambda_j^{k\to l}(t) = \lambda_\circ^{k\to l}(t)(1 + d_j^{(k)})e^{\boldsymbol{\beta}_{kl}^T \mathbf{Z}},$$

$l \in B_{(k)}$. Since we are working on a competing risk framework, the hazard and survival functions of the minimum among $T_j^{k\to l}$, for $l \in B_{(k)}$, are respectively given by

$$\begin{aligned}\lambda_{k_j}(t) &= (1 + d_j^{(k)}) \left(\sum_{l\in B_{(k)}} \lambda_\circ^{k\to l}(t) e^{\boldsymbol{\beta}_{kl}^T \mathbf{Z}} \right) \\ &= (1 + d_j^{(k)})\lambda_{k_1}(t)\end{aligned}$$

and

$$S_{k_j}(t) = \left[\prod_{l\in B_{(k)}} (S_\circ^{k\to l}(t))^{e^{\boldsymbol{\beta}_{kl}^T \mathbf{Z}}} \right]^{(1+d_j^{(k)})} = (S_{k_1}(t))^{(1+d_j^{(k)})},$$

where $\lambda_{k_1}(t)$ and $S_{k_1}(t)$ are the hazard and survival functions associated with the first visit to state k.

Since $\mathrm{E}(T_j^{(k)}) = \int_0^\infty S_{k_j}(t)dt$, using the first order Taylor approximation of $S_{k_j}(t)$, we have

$$\mathrm{E}(T_j^{(k)}) \approx \eta_k + d_j^{(k)} \int_0^\infty S_{k_1}(t) \log S_{k_1}(t) dt$$

and

$$\mathrm{E}\left[\sum_{j=1}^{N_k} T_j^{(k)} \right] \approx \mathrm{E}(N_k)\eta_k + \mathrm{E}\left[\sum_{j=1}^{N_k} d_j^{(k)} \right] \int_0^\infty S_{k_1}(t) \log S_{k_1}(t) dt, \qquad (3)$$

$k = 1, \ldots, r$, where $\eta_k = \mathrm{E}(T_1^{(k)})$ and $d_1^{(k)} = 0$.

If we make further assumptions on the coefficients $d_j^{(k)}$, the expression for the expected value of the total time spent in each state can be simplified. Two different assumptions are considered. The first one is

$$d_j^{(k)} = \begin{cases} (e^{(j-1)\gamma^{(k)}} - 1), & j = 1, \ldots, s^{(k)}, \\ (e^{s^{(k)}\gamma^{(k)}} - 1), & j > s^{(k)}, \end{cases} \qquad (4)$$

where $s^{(k)}$ are known quantities and $\gamma^{(k)}$ are unknown parameters. The expectation $\mathrm{E}\left[\sum_{j=1}^{N_k} d_j^{(k)}\right]$ is given by

$$\mathrm{E}\left[\sum_{j=1}^{N_k} d_j^{(k)}\right] = \sum_{j=1}^{\infty} P(N_k \geq j) d_j^{(k)} = \zeta_k,$$

and must be computed for each process. In Sect. 5 we discuss how to compute this quantity for a process with three states.

The second assumption is to assume that there is a constant $\bar{d}_k$ such that $\sum_{j=1}^{N} d_j^{(k)}/N \to \bar{d}_k$ as $N \to \infty$, so that

$$\mathrm{E}\left[\sum_{j=1}^{N_k} d_j^{(k)}\right] \approx \mathrm{E}\left[N_k \bar{d}_k\right] = \bar{d}_k \mathrm{E}(N_k). \tag{5}$$

In this case, expression (3) may be written as

$$\mathrm{E}\left[\sum_{j=1}^{N_k} T_j^{(k)}\right] \approx \mathrm{E}(N_k)\eta_k + \mathrm{E}(N_k)\bar{d}_k \int_0^{\infty} S_{k_1}(t) \log S_{k_1}(t) dt. \tag{6}$$

The assumptions made for the coefficients d_j are not unrealistic. We must ensure that the hazards associated to transitions from a given state are bounded, i.e., they do not become extremely high so that the sojourn times in that state becomes too small. If that happens, it will imply that the patient would stay in the good health state for a very short period of time, eventually tending to zero. In practice, a patient in that situation would not be considered as having moved back to the good health state. Therefore, the hazards of transitions may increase as the number of previous visits to that state increases, but it must be bounded.

For states $r+1, \ldots, K$, the expression for the expected total time spent in the state is derived by Tunes-da-Silva et al. (2008) and we discuss it briefly here. For the states from which the absorbing state can be reached, for $j = 1, \ldots, N_k - 1$, it is known that, given N_k, the next state visited after $T_j^{(k)}$ must not be the absorbing one. Therefore, the distribution of $T_j^{(k)}$ given N_k is the distribution of the minimum of all latent times $T_j^{k\to l}$ given that the time $T_j^{k\to 0}$ is greater than the minimum of all others, i.e., we have that $T_j^{(k)}$ given N_k has the distribution of $\min_{l\in B_{(k)}}\{T_j^{k\to l}\}$ given that $T_j^{k\to 0} > \min_{l\in B_{(k)}^*}\{T_j^{k\to l}\}$, where $B_{(k)}^* = B_{(k)}\backslash\{0\}$. Denote by $T_j^{*(k)}$ the random variable with the same distribution of $\min_{l\in B_{(k)}}\{T_j^{(k)\to(l)}\}$ given $T_j^{(k)\to(0)} > \min_{l\in B_{(k)}^*}\{T_j^{(k)\to(l)}\}$.

As for the last time state k is visited, the distribution of $T_{N_k}^{*(k)}$ given N_k is the distribution of $\min_{l\in B_{(k)}}\{T_j^{(k)\to(l)}\}$ given that $T_j^{(k)\to(m)} > \min_{l\in B_{(k)}^{**}}\{T_j^{(k)\to(l)}\}$ for

$m \in B_{(k)} \backslash B^{**}_{(k)}$, where $B^{**}_{(k)}$ is the set of all states that can be reached when it is known that state k was visited for the last time. This random variable will be denoted by $T^{\dagger(k)}_{N_k}$.

It follows that with the assumption of identically distributed sojourn times for states $r+1, \ldots, K$, we have that

$$\mathrm{E}\left[\sum_{j=1}^{N_k} T_j^{(k)}\right] = [\mathrm{E}(N_k) - 1]\,\mathrm{E}\left(T_1^{*(k)}\right) + \mathrm{E}\left(T^{\dagger k}_{N_k}\right).$$

Assume now that the latent random variables $T_j^{(k)\to(l)}$ for $l \in B_{(k)}$ have proportional hazards, i.e.,

$$\lambda_{k,l}(t) = \lambda_k^{\circ}(t) e^{\boldsymbol{\beta}_{kl}^T \mathbf{Z}},$$

$l \in B_{(k)}$, where $\lambda_k^{\circ}(t)$ is the baseline hazard function. Under this assumption, Tunes-da-Silva et al. (2008) show that $T_j^{*(k)}$ and $T^{\dagger(k)}_{N_k}$ are identically distributed and that

$$\mathrm{E}\left[\sum_{j=1}^{N_k} T_j^{(k)}\right] = \mathrm{E}(N_k)\mathrm{E}(T_1^{(k)}). \tag{7}$$

Expression (7) is also valid for states from which the absorbing state cannot be reached. Applying results (4) and (7) in (2), we have

$$\begin{aligned}\mu_Q = \sum_{k=1}^{r} q_k \left[\mathrm{E}(N_k)\mathrm{E}(T_1^{(k)}) + \zeta_k \int_0^{\infty} S_{k_1}(t) \log S_{k_1}(t) dt\right] \\ + \sum_{k=r+1}^{K} q_k \mathrm{E}(N_k)\mathrm{E}(T_1^{(k)}),\end{aligned} \tag{8}$$

and by (6) it follows that

$$\begin{aligned}\mu_Q = \sum_{k=1}^{r} q_k \mathrm{E}(N_k) \left[\mathrm{E}(T_1^{(k)}) + \bar{d}_k \int_0^{\infty} S_{k_1}(t) \log S_{k_1}(t) dt\right] \\ + \sum_{k=r+1}^{K} q_k \mathrm{E}(N_k)\mathrm{E}(T_1^{(k)}).\end{aligned} \tag{9}$$

It is important to note that if $\zeta_k = 0$ and $\bar{d}_k = 0$ for $k = 1, \ldots, K$, we have that all sojourn times are identically distributed and expressions (8) and (9) simplifies to the ones obtained by Tunes-da-Silva et al. (2008).

Further simplifications of these expressions can be obtained by imposing certain models for the sojourn times. Two different approaches are considered in this respect:

parametric (exponential) and semiparametric models. We note that expressions (8) and (6) can be applied to any other model for the sojourn times that allows the estimation of the mean time, such as the ones presented by Huzurbazar (2004).

3 Modeling the sojourn times

Usual survival models can be applied with the inclusion of time independent covariates. We consider a semiparametric approach, based on Dabrowska et al. (1994) and Cox (1972), and also a parametric approach, based on the exponential distribution. For both approaches, we assume the data is subject to right censoring, non-informative in the chronological time scale.

3.1 Semiparametric model

The approach considered by Dabrowska et al. (1994) can be adapted in our case with some minor modifications. Intuitively, if we assume that each visit to a state defines a new state, the results presented their work remain valid.

Denote by $0 = \tau_0 < \tau_1 < \tau_2 < \tau_3 < \cdots$ the instants of entrance in states $V_0, V_1, V_2, V_3, \ldots$, respectively. Following Dabrowska et al. (1994), it is necessary to define stopping times given by

- $U_n = \tau_{n+1}$ if τ_n, τ_{n+1}, V_n and V_{n+1} are known;
- $\tau_n < U_n < \tau_{n+1}$ if V_n is known, V_{n+1} is unknown and $\tau_{n+1} - \tau_n > U_n - \tau_n$;
- $U_n = \tau_n$ if no information is available on $\tau_{n+1} - \tau_n$, V_n and V_{n+1}.

Also, for $k = 1, \ldots, r$, define

$$M_{k,h} = \sum_{h'=1}^{h} I(V_{h'} = k),$$

that counts the number of times state k was visited up to the h-th transition is observed. Consider now the counting processes

$$\tilde{N}_{k_j l}(t) = \sum_{h \geq 1} I(\tau_h \leq t, V_{h-1} = k, V_h = l, M_{k,h} = j), \quad k = 1, \ldots, r, l \in B_{(k)},$$
$$\tilde{N}_{kl}(t) = \sum_{h \geq 1} I(\tau_h \leq t, V_{h-1} = k, V_h = l), \quad k = r+1, \ldots, K, l \in B_{(k)},$$

and

$$\tilde{N}(t) = \sum_{k=1}^{r} \sum_{j} \sum_{l} \tilde{N}_{k_j l}(t) + \sum_{k=r+1}^{K} \sum_{l} \tilde{N}_{kl}(t).$$

We assume that the processes $\{\tilde{N}_{k_jl}(t) : t \in [0,\tau]\}$, $k = 1,\ldots,r$, and $\{\tilde{N}_{kl}(t) : t \in [0,\tau]\}$, $k = r+1,\ldots,K$, have intensities given, respectively, by

$$\begin{aligned}\Lambda_{k_jl}(dt) &= I(V(t^-) = k, M_k(t^-) = j)\lambda_{k_jl}\,(L(t);\mathbf{Z})\,dt\\ &= I(V(t^-) = k, M_k(t^-) = j)\lambda_k^\circ\,(L(t))\,(1 + d_j^{(k)})e^{\boldsymbol{\beta}_{kl}^T\mathbf{Z}}dt,\end{aligned}$$

and

$$\begin{aligned}\Lambda_{kl}(dt) &= I(V(t^-) = k)\lambda_{kl}\,(L(t);\mathbf{Z})\,dt\\ &= I(V(t^-) = k)\lambda_k^\circ\,(L(t))\,e^{\boldsymbol{\beta}_{kl}^T\mathbf{Z}}dt,\end{aligned}$$

where $L(t) = t - \tau_{\tilde{N}(t^-)}$ is the backward recurrence time. It is assumed that the intensities are computed with respect to the history or filtration $\mathcal{F}_t$ generated by the counting processes $\tilde{N}_{k_jl}(t)$ and $\tilde{N}_{kl}(t)$ and the information available at $t = 0$.

Note that in this case it is assumed that the hazards functions associated to transitions from the same state to others are proportional, so that the baseline hazard function $\lambda_k^\circ\,(L(t))$ depends only on the actual state k.

We can define the processes associated to the sojourn times:

$$\begin{aligned}N_{k_jl}(x) &= \sum_{h\geq 0} I\left(\tau_{h+1} - \tau_h \leq x, V_h = k, V_{h+1} = l, U_h = \tau_{h+1}, M_{kh} = j\right),\\ N_k(x) &= \sum_j \sum_{l\in B_{(k)}} N_{kl}(x),\\ Y_{k_j}(x) &= \sum_{h\geq 0} I\left(\tau_{h+1} - \tau_h \geq x, V_h = k, M_{kh} = j, U_h - \tau_h \geq x\right),\end{aligned}$$

$j = 1, 2, \ldots$, for $k = 1,\ldots,r$ and $l \in B_{(k)}$ and

$$\begin{aligned}N_{kl}(x) &= \sum_{h\geq 0} I\left(\tau_{h+1} - \tau_h \leq x, V_h = k, V_{h+1} = l, U_h = \tau_{h+1}\right),\\ N_k(x) &= \sum_{l\in B_{(k)}} N_{kl}(x),\\ Y_k(x) &= \sum_{h\geq 0} I\left(\tau_{h+1} - \tau_h \geq x, V_h = k, U_h - \tau_h \geq x\right),\end{aligned}$$

for $k = r+1,\ldots,K$ and $l \in B_{(k)}$, where $B_{(k)}$ is the set of states that can be visited from k.

Let $(1 + d_j^{(k)}) = e^{\alpha_j^{(k)}}$ and denote by $\boldsymbol{\beta}$ the vector collecting all unknown parameters. We also define the vectors $\mathbf{Z}_{i,k_jl}$ for $k = 1,\ldots,r$ and $\mathbf{Z}_{i,kl}$ for $k = r+1,\ldots,K$ such that $\mathbf{Z}_{i,k_jl}^T\boldsymbol{\beta} = \mathbf{Z}_i^T\boldsymbol{\beta}_{kl} + \alpha_j^{(k)}$ for $k = 1,\ldots,r$ and $\mathbf{Z}_{i,kl}^T\boldsymbol{\beta} = \mathbf{Z}_i^T\boldsymbol{\beta}_{kl}$ for $k = r+1,\ldots,K$. Assuming that there are n replications of these processes, the log-likelihood is given by

$$l(\boldsymbol{\beta},\tau)=\sum_{i=1}^{n}\left\{\sum_{k=1}^{r}\left[\sum_{l\in B_{(k)}}\sum_{j=1}^{J_k}\int_0^{\tau}\left[\boldsymbol{\beta}^T\mathbf{Z}_{i,k_jl}+\alpha_j-\log\left(nS_k^{(0)}(t,\boldsymbol{\beta})\right)\right]dN_{i,k_jl}(t)\right]\right.$$
$$\left.+\sum_{k=r+1}^{K}\left[\sum_{l\in B_{(k)}}\int_0^{\tau}\left[\boldsymbol{\beta}^T\mathbf{Z}_{i,kl}-\log\left(nS_k^{(0)}(t,\boldsymbol{\beta})\right)\right]dN_{i,kl}(t)\right]\right\},$$

where

$$S_{k_jl}^{(0)}(t,\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}Y_{i,k_j}(t)(1+d_j^{(k)})e^{\boldsymbol{\beta}_{kl}^T\mathbf{Z}_i},\quad k=1,\ldots,r,l\in B_{(k)},$$
$$S_{kl}^{(0)}(t,\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}Y_{i,k}(t)e^{\boldsymbol{\beta}_{kl}^T\mathbf{Z}_i},\quad k=r+1,\ldots,K,l\in B_{(k)},$$

and

$$S_k^{(0)}(t,\boldsymbol{\beta})=\sum_{l\in B_{(k)}}S_{kl}^{(0)}(t,\boldsymbol{\beta}).$$

Properties for estimators obtained under this model are presented by Dabrowska et al. (1994) using arguments that do not make use of the usual martingale properties, since there is not an appropriate filtration for the processes associated to the sojourn times (e.g. Andersen et al. 1993, p. 680) leading to suitable martingales. Asymptotic theory of a simpler model have been worked out by Shu et al. (2007).

After obtaining the estimator $\widehat{\boldsymbol{\beta}}$ from $l(\boldsymbol{\beta},\tau)$, the cumulative hazard functions can be estimated by (Andersen et al. 1993)

$$\widehat{\mathcal{A}}_k^{\circ}(t,\widehat{\boldsymbol{\beta}})=\sum_{i=1}^{n}\int_0^{t}\frac{dN_{i,k}(u)}{nS_k(u,\widehat{\boldsymbol{\beta}})},$$

for $k=1\ldots,K$, and the baseline survival function by

$$\widehat{S}_k^{\circ}(t)=\exp\{-\hat{\mathcal{A}}_k^{\circ}(t,\widehat{\boldsymbol{\beta}})\}$$

or

$$\tilde{S}_k^{\circ}(t)=\prod_{u\le t}\left(1-d\widehat{\mathcal{A}}_k^{\circ}(u,\widehat{\boldsymbol{\beta}})\right).$$

The survival functions related to each transition are estimated by either

$$\widehat{S}_{k_jl}(t)=\widehat{S}_k^{\circ}(t)^{\exp(\mathbf{Z}_{k_jl}^T\widehat{\boldsymbol{\beta}})},$$
$$\tilde{S}_{k_jl}(t)=\tilde{S}_k^{\circ}(t)^{\exp(\mathbf{Z}_{k_jl}^T\widehat{\boldsymbol{\beta}})}$$

or

$$\breve{S}_{k_j l}(t) = \mathop{\boldsymbol{\pi}}_{u \le t} \left(1 - \exp\left(\mathbf{Z}_{k_j l}^T \widehat{\boldsymbol{\beta}}\right) d\widehat{\mathcal{A}}_k^{\circ}(u, \widehat{\boldsymbol{\beta}})\right)$$

for $k = 1, \ldots, r$, and analogous expressions are valid for $k = r + 1, \ldots, K$. The survival function for the sojourn time in state k at the j-th visit is given by

$$\widehat{S}_{k_j}(t) = \prod_{l \in B_{(k)}} \widehat{S}_{k_j l}(t),$$

$$\tilde{S}_{k_j}(t) = \prod_{l \in B_{(k)}} \tilde{S}_{k_j l}(t)$$

or

$$\breve{S}_{k_j}(t) = \mathop{\boldsymbol{\pi}}_{u \le t} \left(1 - \left[\sum_{l \in B_{(k)}} \exp\left(\mathbf{Z}_{k_j l}^T \widehat{\boldsymbol{\beta}}\right)\right] d\widehat{\mathcal{A}}_k^{\circ}(u, \widehat{\boldsymbol{\beta}})\right)$$

for $k = 1, \ldots, r$. Finally, the mean sojourn time of the j-th visit to state k can be estimated by

$$\widehat{\mathrm{E}(T^{(k_j)})} = \int_0^{\infty} \hat{S}_{k_j}(x) dx.$$

In order to obtain an estimator for the mean quality-adjusted survival time, it is also needed to estimate ζ_k in expression (8) or $\bar{d}_k$ in expression (9) as well as $\int_0^{\infty} S_{k_1}(x) \log S_{k_1}(x) dx$. Estimators of ζ_k or $\bar{d}_k$ are easily obtained given the partial likelihood estimators of the parameters d_j. The expression $\int_0^{\infty} S_{k_1}(x) \log S_{k_1}(x) dx$ can be estimated as the area under the estimated function.

3.2 Parametric model

In this section we consider that the latent time transitions are well described by an exponential model, i.e., the hazard functions associated to the variables $T_j^{k \to l}$, $k = 1, \ldots, r$ and $l \in B_{(k)}$, are given by

$$\lambda_j^{k \to l}(t) = (1 + d_j^{(k)}) e^{-\boldsymbol{\beta}_{kl}^T \mathbf{Z}},$$

where $\boldsymbol{\beta}_{kl}$ is a vector of unknown parameters and $d_j^{(k)}$ defined as before. Define now

$$\alpha_{k_j} = \log(1 + d_j^{(k)}),$$

so that $d_j^{(k)} = e^{\alpha_{k_j}} - 1$, $j = 1, 2, \ldots$. Denoting by $\boldsymbol{\beta}$ the vector of all unknown parameters, including α_{k_j}, $\forall k, j$, we define the vector $\mathbf{Z}_{kl_j}$ such that $\boldsymbol{\beta}_{kl}^T \mathbf{Z} - \alpha_{k_j} = \boldsymbol{\beta}^T \mathbf{Z}_{kl_j}$. With this notation, the likelihood function is given by

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \left\{ \prod_{k=1}^{r} \left[\prod_{j=1}^{N_{i,k}} \left(\prod_{l \in B_{(k)}} \lambda_j^{k \to l} (t_i^{(k_j)})^{\delta_{i,k_j l}} \right) \right. \right.$$
$$\left. \times \exp\left(- \sum_{l \in B_{(k)}} \int_0^{t_i^{(k_j)}} \lambda_j^{k \to l}(u) du \right) \right]$$
$$\left. \times \prod_{k=r+1}^{K} \left[\prod_{j=1}^{N_{i,k}} \left(\prod_{l \in B_{(k)}} \lambda^{k \to l} (t_i^{(k_j)})^{\delta_{i,k_j l}} \right) \exp\left(- \sum_{l \in B_{(k)}} \int_0^{t_i^{(k_j)}} \lambda^{k \to l}(u) du \right) \right] \right\},$$

where

$$\delta_{i,k_j l} = \begin{cases} 1, & \text{if state } l \text{ is visited after the } j\text{-th visit to } k; \\ 0, & \text{otherwise,} \end{cases}$$

and $t_i^{(k_j)}$ is the observed sojourn time in the j-th visit to state k (possibly censored).

Maximum likelihood estimators are obtained by solving $U(\boldsymbol{\beta}) = \mathbf{0}$, where $U(\boldsymbol{\beta})$ is the score function based on $L(\boldsymbol{\beta})$. It is important to note that the maximization procedure must take into account the restriction $\alpha_j^{(k)} \geq 0$.

It is easy to show that, for this parametric model,

$$\mathrm{E}(T_1^{(k)}) = \frac{1}{\sum_{l \in B_{(k)}} \lambda_1^{k \to l}}$$

and

$$\int\limits_0^{\infty} S_{k_1}(t) \log S_{k_1}(t) dt = -\frac{1}{\sum_{l \in B_{(k)}} \lambda_1^{k \to l}},$$

so that expressions (8) and (9) can be written as

$$\mu_Q = \sum_{k=1}^{r} q_k \frac{1}{\sum_{l \in B_{(k)}} \lambda_1^{k \to l}} [\mathrm{E}(N_k) - \zeta_k] + \sum_{k=r+1}^{K} q_k \mathrm{E}(N_k) \frac{1}{\sum_{l \in B_{(k)}} \lambda_1^{k \to l}}$$

and

$$\mu_Q = \sum_{k=1}^{r} q_k \frac{1}{\sum_{l \in B_{(k)}} \lambda_1^{k \to l}} \mathrm{E}(N_k)(1 - \bar{d}_k) + \sum_{k=r+1}^{K} q_k \frac{1}{\sum_{l \in B_{(k)}} \lambda_1^{k \to l}} \mathrm{E}(N_k).$$

Estimators for μ_Q are obtained plugging the estimators for the corresponding unknown quantities into these expressions.

3.3 Expected number of visits to each state

For each state k, define the variables

$$M_{k,m} = \begin{cases} 1, & \text{if state } k \text{ was visited at least } m \text{ times;} \\ 0, & \text{if the absorbing state has been reached earlier.} \end{cases}$$

Note that $N_k = \sum_m M_{k,m}$ and $\mathrm{E}(N_k) = \sum_m \mathrm{E}(M_{k,m}) = \sum_m P(M_{km} = 1)$. For states $k = r+1, \ldots, K$, the probability that a transition from k to k' takes place is

$$p_{kk'} = \int_0^\infty \lambda^{k\to k'}(u) \left(\prod_{l \in B_{(k)}} S^{k\to l}(u) \right) du.$$

For states 1 to r, we consider the general situation in which the hazard function associated to a transition from k to k' in the j-th visit to k is given by $\lambda_j^{k\to k'}(t) = (1+d_j^{(k)})\lambda_1^{k\to k'}(t)$ and the hazard function of the transition $k \to k'$ in the first visit to k is $\lambda_1^{k\to k'}(t) = \lambda_k^\circ(t) g(\boldsymbol{\beta}_{kk'}, \mathbf{Z})$, where $g(\cdot)$ is any positive function. Considering this notation, the probability of a transition to k' after the j-th visit to k is

$$\begin{aligned} p_{k(j)k'} &= \int_0^\infty (1+d_j^{(k)})\lambda_1^{k\to k'}(u) \left(\prod_{l\in B_{(k)}} \left(S_1^{k\to k'}(u)\right)^{(1+d_j^{(k)})} \right) du \\ &= (1+d_j^{(k)}) g(\boldsymbol{\beta}_{kk'}, \mathbf{Z}) \int_0^\infty \lambda_k^\circ(u) \left(S_k^\circ(u)\right)^{\sum_{l\in B_{(k)}} (1+d_j^{(k)}) g(\boldsymbol{\beta}_{kl}, \mathbf{Z})} du \\ &= \frac{(1+d_j^{(k)}) g(\boldsymbol{\beta}_{kk'})}{(1+d_j^{(k)}) \sum_{l\in B_{(k)}} g(\boldsymbol{\beta}_{kl}, \mathbf{Z})} = \frac{g(\boldsymbol{\beta}_{kk'})}{\sum_{l\in B_{(k)}} g(\boldsymbol{\beta}_{kl}, \mathbf{Z})}. \end{aligned}$$

This result shows us that the transition probabilities are the same for all j. Therefore, it is possible to construct a matrix $\mathbf{P}$ of transition probabilities that does not depend on the number of previous visits to each state, and the results presented by Tunes-da-Silva et al. (2008) are directly applicable here. Assuming that all patients enter in the study in the same health state, it is possible to compute the probability f_{1k} of reaching k for the first time, given by the sum of the probabilities of reaching k in one, two, three, … steps. Denote by $f_{kk}^{(n)}$ the probability of going from k to k in n steps, which can also be computed based on the transition matrix $\mathbf{P}$, and let

$$f_{kk} = \sum_{n=0}^{\infty} f_{kk}^{(n)}$$

be the probability of returning to k from k. We have that $P(M_{k1} = 1) = f_{1k}$ and $P(M_{km} = 1) = f_{1k}\,(f_{kk})^{m-1}$, $m = 1, 2, \ldots$. Therefore,

$$\mathrm{E}(N_k) = f_{1k} \sum_{m=0}^{\infty} (f_{kk})^m = \frac{f_{1k}}{1 - f_{kk}}.$$

It is evident that the expected number of visits in each state depends on the process considered. Also, the way it is computed may be different for each particular process. In the next section, we present the derivation for a three state process in details.

4 Three state process

The estimator of the mean quality-adjusted survival time in the situation with three states (two transient and one absorbing—see Fig. 1) is derived in this section. We assume sojourn times in state B independent with the same distribution for a given vector of covariates, whereas the mean sojourn times in state A decreases as the number of previous visits increases. For this process, applying the results presented in Sect. 2, the mean quality-adjusted survival is given by

$$\mu_Q = q_A\,[\mathrm{E}(N_A)]\,\mathrm{E}(T_1^{(A)}) + q_A \left(\int_0^{\infty} S_{A_1}(t) \log S_{A_1}(t) dt \right) \mathrm{E}\left[\sum_{j=1}^{N_A} d_j \right] + q_B\,[\mathrm{E}(N_B)]\,\mathrm{E}\left(T_1^{(B)}\right).$$

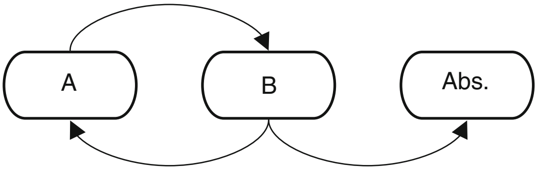
Assuming that all patients enter the study in the same health state, we have $\mathrm{E}(N_A) = \mathrm{E}(N_B)$. Defining the variables

$$M_{A,m} = \begin{cases} 1, & \text{if state } A \text{ was visited at least } m \text{ times;} \\ 0, & \text{otherwise,} \end{cases}$$

for, $m \geq 1$, we have that $N_A = \sum_m M_{A,m}$, $P(M_{A,1} = 1)$ and

$$P(M_{A,2} = 1) = P(A \to B \to A) = \int_0^{\infty} \lambda_{BA|\mathbf{Z}}(u) \left(S_{BA|\mathbf{Z}}(u) S_{B0|\mathbf{Z}}(u)\right) du = p_{\mathbf{Z}},$$

Fig. 1 Three state process



where $\lambda_{BA|\mathbf{Z}}(u)$ is the hazard function for the transition $B \to A$ for a given $\mathbf{Z}$ and $S_{BA|\mathbf{Z}}(u)$ and $S_{B0|\mathbf{Z}}(u)$ are the survival functions for the respective transitions. It follows that

$$\mathrm{E}(N_A) = \sum_m \mathrm{E}(M_{Am}) = 1 + \sum_{m=2}^{\infty} p_{\mathbf{Z}}^{m-1} = \frac{1}{1-p_{\mathbf{Z}}},$$

with $p_{\mathbf{Z}} = \int_0^\infty \lambda_{BA|\mathbf{Z}}(u) S_{BA|\mathbf{Z}}(u) S_{B0|\mathbf{Z}}(u) du$. In addition, $P(N_A \geq j) = P(M_{A,j} = 1) = p_{\mathbf{Z}}^{j-1}$ and, under assumption (4),

$$\begin{aligned}\mathrm{E}\left[\sum_{j=1}^{N_A} d_j\right] = \zeta_{A|\mathbf{Z}} = \sum_{j=1}^{\infty} P(N_A \geq j) d_j &= \sum_{j=1}^{s} p_{\mathbf{Z}}^{j-1}(e^{j\gamma}-1) + \sum_{j=s+1}^{\infty} p_{\mathbf{Z}}^{j-1}(e^{s\gamma}-1) \\ &= \left[\left(e^{\gamma} p_{\mathbf{Z}}\right)^s - 1\right]\left[\frac{1}{1-p_{\mathbf{Z}}} - \frac{e^{\gamma}}{1-p_{\mathbf{Z}}e^{\gamma}}\right]. \end{aligned} \tag{10}$$

The mean quality-adjusted survival time under assumption (4) is given by

$$\begin{aligned}\mu_Q \approx\, & q_A \frac{1}{1-p_{\mathbf{Z}}} \mathrm{E}(T_1^{(A)}) + q_A \zeta_{A|\mathbf{Z}} \left(\int_0^\infty S_{A_1}(t) \log S_{A_1}(t) dt\right) \\ & + q_B \frac{1}{1-p_{\mathbf{Z}}} \mathrm{E}\left(T_1^{(B)}\right), \end{aligned} \tag{11}$$

and, under (6),

$$\mu_Q \approx q_A \frac{1}{1-p_{\mathbf{Z}}} \left(\mathrm{E}(T_1^{(A)}) + \bar{d} \int_0^\infty S_{A_1}(t) \log S_{A_1}(t) dt\right) + q_B \frac{1}{1-p_{\mathbf{Z}}} \mathrm{E}\left(T_1^{(B)}\right). \tag{12}$$

Assuming the semiparametric model, under (4), we have

$$\mu_Q = q_A \frac{e^{\mathbf{Z}^T \boldsymbol{\beta}_{BA}} + e^{\mathbf{Z}^T \boldsymbol{\beta}_{BO}}}{e^{\mathbf{Z}^T \boldsymbol{\beta}_{BO}}} \mathrm{E}(T_1^{(A)}) + q_A \Phi_A \zeta_{A|\mathbf{Z}} + q_B \frac{e^{\mathbf{Z}^T \boldsymbol{\beta}_{BA}} + e_{BO}^{\mathbf{Z}^T \boldsymbol{\beta}}}{e^{\mathbf{Z}^T \boldsymbol{\beta}_{BO}}} \mathrm{E}\left(T_1^{(B)}\right) \tag{13}$$

and under (5) we have

$$\mu_Q = \frac{e^{\mathbf{Z}^T \boldsymbol{\beta}_{BA}} + e^{\mathbf{Z}^T \boldsymbol{\beta}_{BO}}}{e^{\mathbf{Z}^T \boldsymbol{\beta}_{BO}}} \left[q_A \left(\mathrm{E}(T^{(A)}) + \bar{d}\Phi_A\right) + q_B \mathrm{E}(T^{(B)})\right], \tag{14}$$

where $\Phi_A = e^{\mathbf{Z}^T \boldsymbol{\beta}_A} \int_0^\infty (S_0(t))^{e^{\mathbf{Z}^T \boldsymbol{\beta}_A}} \log(S_0(t))\, dt$ and $\zeta_{A|\mathbf{Z}}$ is defined in (10).

Assuming an exponential model, the expressions can be simplified to

$$\mu_Q = q_A \frac{1}{\lambda_{A|\mathbf{Z}}} \left(\frac{\lambda_{BA|\mathbf{Z}} + \lambda_{BO|\mathbf{Z}}}{\lambda_{BO|\mathbf{Z}}} - \zeta_{A|\mathbf{Z}} \right) + q_B \frac{1}{\lambda_{BO|\mathbf{Z}}}, \tag{15}$$

where $\lambda_{A|\mathbf{Z}} = e^{-\mathbf{Z}^T \boldsymbol{\beta}_A}$,

$$\begin{aligned} \zeta_{A|\mathbf{Z}} = &\left[\left(e^{\gamma} \frac{\lambda_{BA|\mathbf{Z}}}{\lambda_{BA|\mathbf{Z}} + \lambda_{BO|\mathbf{Z}}} \right)^{s} - 1 \right] \left(\lambda_{BA|\mathbf{Z}} + \lambda_{BO|\mathbf{Z}} \right) \\ &\times \left(\frac{1}{\lambda_{BO|\mathbf{Z}}} - \frac{e^{\gamma}}{\lambda_{BA|\mathbf{Z}}(1 - e^{\gamma}) + \lambda_{BO|\mathbf{Z}}} \right), \end{aligned}$$

and

$$\mu_Q = q_A \frac{(1 - \bar{d})}{\lambda_{A|\mathbf{Z}}} \frac{\lambda_{BA|\mathbf{Z}} + \lambda_{BO|\mathbf{Z}}}{\lambda_{BO|\mathbf{Z}}} + q_B \frac{1}{\lambda_{BO|\mathbf{Z}}}. \tag{16}$$

The results for the parametric approach remain valid for any proportional hazards model. Therefore, a Weibull distribution can also be considered. The derivation of the estimator for the Weibull distribution is exactly the same as done for the exponential distribution. Assuming that the sojourn times follow a Weibull distribution with hazard rates given by

$$\lambda_j^{k \to l}(t) = (1 + d_j^{(k)}) e^{-\boldsymbol{\beta}_{kl}^T \mathbf{Z}} \nu t^{\nu - 1},$$

for $k = A, B, l \in B_{(k)}$, $B_{(A)} = \{B\}$ and $B_{(B)} = \{A, O\}$, expressions (11) and (12) can be simplified to, respectively,

$$\begin{aligned} \mu_Q = q_A &\frac{\Gamma(1 + 1/\nu)}{\lambda_{A|\mathbf{Z}}^{1/\nu}} \left(\frac{\lambda_{BA|\mathbf{Z}} + \lambda_{BO|\mathbf{Z}}}{\lambda_{BO|\mathbf{Z}}} - \frac{\zeta_A}{\nu} \right) \\ &+ q_B \frac{\left(\lambda_{BA|\mathbf{Z}} + \lambda_{BO|\mathbf{Z}} \right)^{1 - 1/\nu}}{\lambda_{BO|\mathbf{Z}}} \Gamma \left(1 + \frac{1}{\nu} \right) \end{aligned}$$

and

$$\begin{aligned} \mu_Q = q_A &\frac{\Gamma(1 + 1/\nu)}{\lambda_{A|\mathbf{Z}}^{1/\nu}} \frac{\lambda_{BA|\mathbf{Z}} + \lambda_{BO|\mathbf{Z}}}{\lambda_{BO|\mathbf{Z}}} \left(1 - \frac{\bar{d}}{\nu} \right) \\ &+ q_B \frac{\left(\lambda_{BA|\mathbf{Z}} + \lambda_{BO|\mathbf{Z}} \right)^{1 - 1/\nu}}{\lambda_{BO|\mathbf{Z}}} \Gamma \left(1 + \frac{1}{\nu} \right). \end{aligned}$$

For both, parametric and semiparametric approaches, the model for the sojourn times and estimation procedures are exactly as described before.

Table 1 Frequencies for hospitalization data

Sex	Disease		Total
	Respiratory	Digestive	
Female	47	24	71
Male	43	26	69
Total	90	50	140

Table 2 Results for the hospitalization data

Estimator		Respiratory		Digestive	
		Female	Male	Female	Male
Exponential	$\hat{\mu}_Q$	12.55	6.27	9.82	5.04
Not ident. dist. (13)	std	0.75	0.38	0.50	0.23
Exponential	$\hat{\mu}_Q$	10.90	5.68	8.87	4.75
Identic. dist. assumption	std	0.32	0.18	0.35	0.15
Semiparametric	$\hat{\mu}_Q$	5.29	2.73	4.49	2.48
Not ident. dist. (15)	std	0.17	0.10	0.17	0.10
Semiparametric	$\hat{\mu}_Q$	7.58	4.34	6.04	3.56
Identic. dist. assumption	std	0.17	0.11	0.17	0.09

5 An application to hospitalization data

In this section we apply the methodology discussed in the previous sections to a data retrieved from a larger data set, based on the information of all patients that were hospitalized (including readmissions) in a medical facility in Brazil, from July 2006 to June 2007. This data was first analyzed by Castro and Carvalho (2005). Only a sample of the data is available and we selected from this sample patients aged 55 years or older, hospitalized due to problems in the respiratory or digestive systems. Since the censoring rate is extremely high and our main objective is to illustrate the methodology, we selected a sample with a censoring rate of 60% and ended up with a total of 140 patients, as shown in Table 1. We consider a three state process for this data: state A corresponding to the periods (in years) in which the patient is out of the hospital, state B if the patient was admitted to the hospital, and the third state is the absorbing state, corresponding to death. The maximum number of readmissions of a patient observed in this sample was six, but most patients had one or two hospitalizations. We considered both, parametric and semiparametric approaches, with two binary covariates: gender (female or male) and disease (respiratory system or digestive system). Age was not included in the model because the sample was somewhat homogeneous with respect to that characteristic (all patients selected in the sample were aged more than 55 years). The utility coefficients were set as 1/2 for state B and 1 for state A. Jackknife resampling was used to compute standard errors. For each approach, we fit the model considering two scenarios with respect to the sojourn times: being and not being identically distributed. The results are shown in Table 2.

The figures for the exponential model are very different than the ones observed for the semiparametric model. This is not surprising, since the assumption of exponential distribution for the sojourn times may be not be tenable. In both situations considered for the semiparametric model, the estimator of the mean quality adjusted survival time with the identically distributed assumption for the sojourn times in state A (for a given set of covariates) showed larger estimates, for all combinations of gender and disease. The opposite behavior was observed for the parametric model. Comparing the mean quality adjusted times for the semiparametric approach, we may conclude that patients with digestive problems have smaller quality adjusted survival time and that males have smaller quality adjusted survival times when compared to females.

6 Simulation study

In this section, we present simulation studies of the proposed estimator. We consider two different scenarios for the three state process shown in Fig. 1.

For the first simulation study, only one binary covariate, denoted x, was included in the model assuming values 0 or 1. The data was generated in such way that the proportion of covariates equal to 1 was the same as the proportion equal to 0. It was also assumed that all observations were in the good health state at the beginning of the study.

The sojourn time of the j-th visit to state A was considered exponentially distributed, with hazard rate given by

$$\lambda_{A_j|x} = \exp(-2 - \beta_A x + \gamma j), \quad j = 1, \ldots, n_d$$

and

$$\lambda_{A_j|x} = \exp(-2 - \beta_A x + \gamma n_d), \quad j > n_d.$$

Note that if $\gamma > 0$, the mean sojourn time in state A decreases as the number of previous visits increases for $j = 1, \ldots, n_d$.

The sojourn times in state B is the minimum between the time until a transition to A, $T_j^{(B)\to(A)}$, and the time up to a transition to the absorbing state, $T_j^{(B)\to(O)}$. $T_j^{(B)\to(A)}$ and $T_j^{(B)\to(O)}$ are exponentially distributed with hazard rates given by

$$\lambda_{BA|x} = \exp(-1 - \beta_B x)$$

and

$$\lambda_{BO|x} = \exp(-2 - \beta_B x).$$

In the simulations considered we generated data with $\beta_A = \beta_B = 1/2$. Under these assumptions, the expected number of visits to state A is

$$\mathrm{E}(N_A) = \frac{\lambda_{BA|x} + \lambda_{BO|x}}{\lambda_{BA|x}} \approx 3.72.$$

This result motivated us to choose $n_d = 5$ in our simulated data.

Censoring was included in the data with proportions fixed at 0%, 30% and 50%. We assumed random censoring and the hazard for the censoring variable was computed so that the probability of an observation being censored was equal to the established proportion of censoring. In order to do so, first note that from (15), if we set $q_A = q_B = 1$, it is possible to obtain an expression for the expected survival time for the two possible values of the covariate:

$$\mu_x = \frac{1}{\lambda_{A|x}} \frac{\lambda_{BA|x} + \lambda_{BO|x}}{\lambda_{BO|x}} + \frac{1}{\lambda_{BO|x}},$$

for $x = 0, 1$, and

$$\mu = (1/2)\mu_1 + (1/2)\mu_0,$$

since the probability of $x = 1$ is $1/2$.

For T and C independent and exponentially distributed random variables, it is known that $P(C > T) = \frac{\lambda_T}{\lambda_C + \lambda_T}$ and $\lambda_T = \frac{1}{\mu_T}$. Therefore, the hazard for the censoring variables was considered as

$$\lambda_C = \frac{p_c}{(1 - p_c)\mu},$$

where p_c is the desired proportion of censored observations. The simulated data showed that the observed proportions of censored observations were very close to the desired ones.

We fit a parametric (exponential) model and a semiparametric model. The jackknife resampling method was employed to compute the estimated standard error as well as the estimated bias for the estimators. If no transition between two states was observed, the corresponding sample was disregarded since, in that case, it is not possible to estimate the parameters associated to that transition. Therefore, we considered samples with at least two observed transitions from state A to B, B to A and B to 0, so that the jackknife could be used. The jackknife was applied removing all sojourn times observed in a patient, i.e., removing one patient at each jackknife sample. Table 3 shows the results for both, parametric and semiparametric models, based on 1,000 simulations. For the parametric approach, the estimator was obtained using the second order approximation based on a Taylor expansion applied to (3). We assumed that the QOL scores for states A and B were 1 and 0.3, respectively.

The second study was performed in order to compare different estimators for the mean quality adjusted survival. We generated data analogously to the first study, but we also included a continuous covariate in addition to the binary one, following a normal distribution with zero mean and variance equals to 0.2. This variable was included only for the sojourn times in state A. Therefore, the sojourn time of the j-th visit to state A is exponentially distributed with hazard rate given by

$$\lambda_{A_j|x} = \exp\left(-2 - \beta_{A_1}x + \beta_{A_2}y + \gamma j\right), \quad j = 1, \ldots, n_d$$

and

$$\lambda_{A_j|x} = \exp\left(-2 - \beta_{A_1}x + \beta_{A_2}y + \gamma n_d\right), \quad j > n_d,$$

Table 3 Simulation results for parametric and semiparametric models: sample size (n), sample average of $\hat{\mu}_Q$, sample average of the jackknife estimator $\hat{\mu}_{Q,J}$, sample average of the estimated bias using jackknife, sample variance of $\hat{\mu}_Q$ (SV), sample average of the estimated variance using jackknife (EVJ) and mean squared error of $\hat{\mu}_Q$ (MSE)

n	Censoring rate (%)	μ_Q	Exponential						Semiparametric					
			$\hat{\mu}_Q$	$\hat{\mu}_{Q,J}$	Bias	SV	EVJ	MSE	$\hat{\mu}_Q$	$\hat{\mu}_{Q,J}$	Bias	SV	EVJ	MSE
200	0	21.25	21.44	21.35	0.19	4.28	4.56	4.31	19.46	19.55	−1.79	2.99	2.74	6.18
	0	35.04	35.47	35.09	0.43	11.99	12.59	12.18	32.33	32.32	−2.71	7.37	7.52	14.71
	30	21.25	23.21	22.52	1.96	10.72	12.34	14.57	19.39	19.45	−1.86	5.01	5.59	8.48
	30	35.04	39.16	37.83	4.12	36.32	39.77	53.30	33.24	33.02	−1.80	17.61	18.85	20.85
	50	21.25	28.11	24.67	6.86	63.85	86.17	110.93	18.53	18.44	−2.72	16.55	19.74	23.94
	50	35.04	49.80	42.35	14.76	317.80	308.71	535.76	34.83	34.62	−0.21	47.94	60.87	47.98
100	0	21.25	21.66	21.39	0.41	9.93	10.76	10.10	19.42	19.59	−1.83	5.40	5.84	8.76
	0	35.04	35.63	34.90	0.59	24.90	28.93	25.25	32.35	32.42	−2.68	15.14	15.72	22.35
	30	21.25	23.80	22.08	2.55	27.69	33.68	34.20	19.44	19.55	−1.81	10.81	13.10	14.10
	30	35.04	40.69	37.09	5.65	101.45	133.87	133.37	33.61	33.41	−1.43	35.33	44.44	37.38
	50	21.25	28.88	21.90	7.63	87.41	144.77	145.55	18.15	18.31	−3.10	51.11	64.29	60.72
	50	35.04	51.47	37.37	16.43	367.10	577.22	637.10	34.63	34.48	−0.41	132.42	172.50	132.58

where x is the binary covariate and y is the continuous covariate. For state B, sojourn times and censoring times were generated exactly as described before. For this configuration, we used $\beta_{A_1} = \beta_B = 0.2$, $\beta_{A_2} = 0.4$, $n_d = 5$, $q_a = 0.8$ and $q_b = 0.3$. There were $n = 100$ and $n = 200$ subjects per simulated data set and 2,000 replicates per data configuration. We computed the estimated mean quality adjusted survival using nine different estimators: the parametric estimators proposed in this paper, using the first and second order Taylor expansion (denoted parametric 1 and parametric 2, respectively); the parametric estimator proposed by Tunes-da-Silva et al. (2008) that assumes identically distributed sojourn times in state A, for a given vector of covariates; the semiparametric estimator proposed in this work; the semiparametric estimator proposed by Tunes-da-Silva et al. (2008), that also assumes sojourn times in state A have the same distribution for a given vector of covariates; two estimators proposed by Zhao and Tsiatis (2000) and two estimators proposed by Huang and Louis (1999). The first estimator proposed by Zhao and Tsiatis (2000) is a simple weighted estimator, given by

$$\hat{\mu}_{WT} = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i U_i}{\hat{K}(Z_i)}, \tag{17}$$

where $\hat{K}(Z_i)$ is the Kaplan–Meier estimator for the survival of the censoring variable, Z_i is the minimum between the time a patient reaches the absorbing state and the censoring time and Δ_i is the failure indicator variable. The second estimator is an improved one, obtained by adding a new term to the right-hand side of (17). The estimators proposed by Huang and Louis (1999) are functions of Nelson-Aalen estimators and also take into account the distribution of the censoring variable.

The proposed estimators as well as the estimators in Tunes-da-Silva et al. (2008) allow for covariates while Zhao-Tsiatis and Huang-Louis estimators do not. Therefore, for the estimators allowing covariates we computed the estimated mean quality adjusted survival time for both values of the binary covariate with $y = 0$. For estimators that do not allow for covariates, we computed the estimated mean quality-adjusted survival for observations with $x = 0$ and with $x = 1$ separately. The estimators allowing for covariates were computed including the covariate x in the model (i.e., we did not compute separately for different values of x). Results of the simulation for $x = 0$ and $x = 1$ are listed in Tables 4 and 5, respectively.

The resulting figures from the simulations suggest that, in general, the proposed estimators provide accurate estimates for low and moderate censoring rates. The second order Taylor approximation estimator is the best one in most scenarios. For high censoring rates, the variance of the estimator may be extremely high; the estimator overestimates the mean quality-adjusted survival time in the parametric approach and underestimates in the semiparametric situation. In Table 3, the difference between the parametric and semiparametric approaches is due to the approximations used in each situation: it was used the second order Taylor approximation for the parametric estimator and the first order for the semiparametric estimator. In these cases, it can be verified that the jackknife estimator considerably reduces the bias of the proposed

Table 4 Simulation results for nine different estimators of the mean quality-adjusted survival time for $\mu_Q = 21.10$: sample size (n), sample average of $\hat{\mu}_Q$, sample average of the bias, sample variance of $\hat{\mu}_Q$ (SV) and mean squared error of $\hat{\mu}_Q$ (MSE)

n	Estimator	0 % censoring				30 % censoring				50 % censoring			
		$\hat{\mu}_Q$	Bias	SV	MSE	$\hat{\mu}_Q$	Bias	SV	MSE	$\hat{\mu}_Q$	Bias	SV	MSE
200	Parametric 1	18.54	−2.56	2.07	8.63	18.11	−3.00	3.91	12.89	17.02	−4.08	9.32	26.00
	Parametric 2	21.69	0.59	2.40	2.75	22.13	1.03	3.51	4.57	22.67	1.57	4.89	7.35
	Parametric Identic.	25.88	4.78	5.18	28.04	27.02	5.92	8.07	43.06	28.62	7.52	11.94	68.44
	Semiparametric	18.60	−2.50	2.11	8.36	18.27	−2.84	4.04	12.10	17.56	−3.54	9.87	22.40
	Semiparam. Identic.	26.00	4.90	5.26	29.23	27.25	6.15	8.39	46.19	29.40	8.30	13.65	82.54
	Zhao-Tsiatis 1	20.99	−0.12	3.30	3.32	20.47	−0.63	6.26	6.66	18.90	−2.21	14.01	18.88
	Zhao-Tsiatis 2	20.99	−0.12	3.30	3.32	20.00	−1.11	6.21	7.43	17.74	−3.36	14.56	25.85
	Huang-Louis 1	20.99	−0.12	3.30	3.32	21.39	0.29	5.76	5.84	21.76	0.66	11.37	11.81
	Huang-Louis 2	20.99	−0.12	3.30	3.32	21.21	0.11	5.44	5.45	21.26	0.15	8.69	8.72
100	Parametric 1	18.44	−2.66	4.49	11.57	17.91	−3.19	8.07	18.24	16.64	−4.46	20.76	40.65
	Parametric 2	21.75	0.65	4.78	5.21	22.12	1.02	6.75	7.79	22.77	1.66	10.30	13.06
	Parametric Identic.	26.00	4.90	10.24	34.25	27.01	5.91	15.33	50.22	28.75	7.64	26.25	84.68
	Semiparametric	18.57	−2.53	4.64	11.05	18.20	−2.91	8.37	16.83	17.54	−3.56	20.28	32.98
	Semiparam. Identic.	26.19	5.09	10.39	36.31	27.31	6.21	15.89	54.42	29.48	8.38	29.47	99.66
	Zhao-Tsiatis 1	21.06	−0.05	6.84	6.84	20.13	−0.98	11.55	12.50	18.16	−2.95	24.32	33.02
	Zhao-Tsiatis 2	21.06	−0.05	6.84	6.84	19.30	−1.80	11.38	14.63	16.30	−4.80	22.79	45.83
	Huang-Louis 1	21.06	−0.05	6.84	6.84	21.39	0.28	11.14	11.22	21.67	0.57	20.38	20.70
	Huang-Louis 2	21.06	−0.05	6.84	6.84	21.13	0.03	9.89	9.89	21.00	−0.10	15.46	15.47

Table 5 Simulation results for nine different estimators of the mean quality-adjusted survival time for $\mu_Q = 25.78$: sample size (n), sample average of $\hat{\mu}_Q$, sample average of the bias, sample variance of $\hat{\mu}_Q$ (SV) and mean squared error of $\hat{\mu}_Q$ (MSE)

n	Estimator	0 % censoring				30 % censoring				50 % censoring			
		$\hat{\mu}_Q$	Bias	SV	MSE	$\hat{\mu}_Q$	Bias	SV	MSE	$\hat{\mu}_Q$	Bias	SV	MSE
200	Parametric 1	22.63	−3.15	3.08	12.99	22.24	−3.54	6.33	18.86	21.01	−4.76	14.43	37.10
	Parametric 2	26.47	0.70	3.58	4.07	27.17	1.40	5.45	7.40	27.99	2.22	7.61	12.52
	Parametric Identic.	31.59	5.81	7.74	41.52	33.17	7.39	11.98	66.65	35.34	9.56	18.28	109.70
	Semiparametric	22.78	−2.99	3.14	12.10	22.60	−3.18	6.37	16.48	22.14	−3.64	14.27	27.51
	Semiparam. Identic.	31.73	5.96	7.82	43.29	33.45	7.68	12.61	71.55	36.30	10.52	22.16	132.91
	Zhao-Tsiatis 1	25.59	−0.18	4.94	4.97	24.98	−0.79	10.35	10.98	22.16	−3.61	26.83	39.87
	Zhao-Tsiatis 2	25.59	−0.18	4.94	4.97	24.29	−1.48	11.00	13.20	20.43	−5.34	27.16	55.73
	Huang-Louis 1	25.59	−0.18	4.94	4.97	26.27	0.50	9.65	9.90	26.70	0.92	18.33	19.18
	Huang-Louis 2	25.59	−0.18	4.94	4.97	25.96	0.19	8.56	8.60	25.95	0.17	13.87	13.89
100	Parametric 1	22.44	−3.33	6.60	17.72	21.97	−3.81	11.98	26.46	20.52	−5.26	32.77	60.43
	Parametric 2	26.47	0.69	7.00	7.48	27.14	1.36	10.23	12.08	28.05	2.27	16.22	21.39
	Parametric Identic.	31.63	5.86	14.96	49.25	33.14	7.37	23.47	77.75	35.41	9.63	39.34	132.14
	Semiparametric	22.73	−3.05	6.63	15.93	22.55	−3.22	11.91	22.29	22.16	−3.62	28.65	41.74
	Semiparam. Identic.	31.85	6.07	15.11	52.01	33.45	7.67	24.39	83.20	36.09	10.31	44.79	151.18
	Zhao-Tsiatis 1	25.61	−0.17	9.45	9.48	24.29	−1.49	20.56	22.76	20.52	−5.25	44.10	71.68
	Zhao-Tsiatis 2	25.61	−0.17	9.45	9.48	23.13	−2.65	20.39	27.38	18.32	−7.46	39.21	94.81
	Huang-Louis 1	25.61	−0.17	9.45	9.48	26.10	0.32	19.05	19.15	26.07	0.29	37.99	38.08
	Huang-Louis 2	25.61	−0.17	9.45	9.48	25.73	−0.05	16.19	16.20	25.18	−0.59	25.99	26.35

estimators. Also, the jackknife estimator for the variance of $\hat{\mu}_Q$ provides, in general, very good approximations.

The results in Tables 4 and 5 show that the estimator denoted by Huang-Louis 2 has the smallest bias in most situations. However, the variance as well as the mean squared error of the estimators that do not allow for covariates are greater than both parametric and semiparametric estimators. This results shows the importance of including covariates in the model. The parametric estimator with second order Taylor approximation usually has the smallest variance and mean squared error. The results also show that the estimators proposed by Tunes-da-Silva et al. (2008), where it is assumed that the distribution of the sojourn times in state A does not change as the number of previous visits to A increases, performs very badly when this assumption is violated.

7 Discussion

In this paper, we generalize the estimator proposed by Tunes-da-Silva et al. (2008) for the mean quality-adjusted survival time allowing the sojourn times to be non-identically distributed. This makes possible to apply the methodology to a broader class of applied problems, since that, in practice, the mean sojourn times in each health state usually changes over time. Although only right censoring was considered, the extension for interval and left censored observations can be derived.

We assumed that the mean sojourn time in some states may decrease as the number of previous visits increases, however the situation in which the mean sojourn times increase may also be considered without further complications. Another assumption made in this paper is the independence among sojourn times for a given patient, but further work is under development to allow the incorporation of a possible correlation among sojourn times for a patient. The use of frailty terms in gap times models, however, is not straightforward due to the fact that when the overall follow-up time is subject to right independent censoring, gap times (except the first one) are subject to dependent censoring (see Lin et al. 1999). The inclusion of time dependent covariates is also an issue of interest in applications; nevertheless, some of the results presented in this paper do not remain valid in this case (in particular, results concerning the expected number of visits to states), and the resulting expressions may be extremely complex, compromising the simplicity of the estimator.

Simulation results show that the estimator behaves properly and may be a helpful tool for treatment comparisons. The proposed estimator, in addition to be the only one to include covariates, seems to have smaller mean squared error, although the bias may be greater. The simulations also show us the importance of considering the decrease of mean sojourn times in the model and how covariates can improve the estimation when they are available. Finally, model validation is an important aspect that must be carefully considered. The proposed estimator is based on multistate models for sojourn times and it will be appropriate whenever the multistate model is correctly specified. Model validation techniques for multistate models is a topic that needs to be urgently developed.

Acknowledgements This work was partially supported by CAPES, Brazil.

References

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

Andersen PK, Keiding N (2002) Multistate models for event history analysis. Stat Meth Med Res 11:91–115

Castro MMS, Carvalho MS (2005) Grouping of the international classification of diseases for analysis of hospital readmissions. Cadernos de Saúde Pública 21(1):317–323

Chen P, Sen PK (2001) Quality-adjusted survival estimation with periodic observations. Biometrics 57:868–874

Cox DR (1972) Regression models and life tables. J Roy Stat Soc Ser B 34(2):187–220

Dabrowska DM, Sun G, 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(427):867–877

Fairclough DL (1997) Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy. Stat Med 16:1197–1209

Gelber RD, Gelman RS, Goldhrisch A (1989) A quality-of-life-oriented endpoint for comparing therapies. Biometrics 45(3):781–795

Glasziou PP, Simes RJ, Gelber RD (1990) Quality adjusted survival analysis. Stat Med 9:1259–1276

Hajek J, Sidak Z, Sen PK (1999) Theory of rank tests. Academic Press, New York

Huang YJ, Louis TA (1999) Expressing estimators of expected quality-adjusted survival as functions of Nelson-Aalen estimators. Lifetime Data Anal 5(3):199–212

Huzurbazar A (2004) Multistate models, flowgraph models and semi-markov processes. Commun Stat Theor Meth 33(3):457–474

Lin DY, Sun W, Ying Z (1999) Nonparametric estimation of the gap times distributions for serial events with censored data. Biometrika 86(1):59–70

Pocock SJ, Geller NL, Tsiatis AA (1987) The analysis of multiple endpoints in clinical trials. Biometrics 43:487–498

Robins MJ, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89(427):846–866

Shu Y, Klein JP, Zhang M-J (2007) Asymptotic theory for the Cox semi-Markov illness-death model. Lifetime Data Anal 13:91–117

Tunes-da-Silva G, Sen PK, Pedroso-de-Lima AC (2008) Estimation of the mean quality-adjusted survival using a multistate model for the sojourn times. J Stat Plann Infer 138(8):2267–2282

van der Laan MJ, Hubbard A (1998) Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed. Biometrika 85(4):771–783

Zhao H, Tsiatis AA (1997) A consistent estimator for the distribution of quality adjusted survival time. Biometrika 84(2):339–348

Zhao H, Tsiatis AA (1999) Efficient estimation of the distribution of quality adjusted survival time. Biometrics 55:1101–1107

Zhao H, Tsiatis AA (2000) Estimating mean quality adjusted lifetime with censored data. Sankhya: Indian J Stat Ser B 62:175–188

Zhao H, Tsiatis AA (2001) Testing equality of survival functions of quality adjusted lifetime. Biometrics 57:861–867