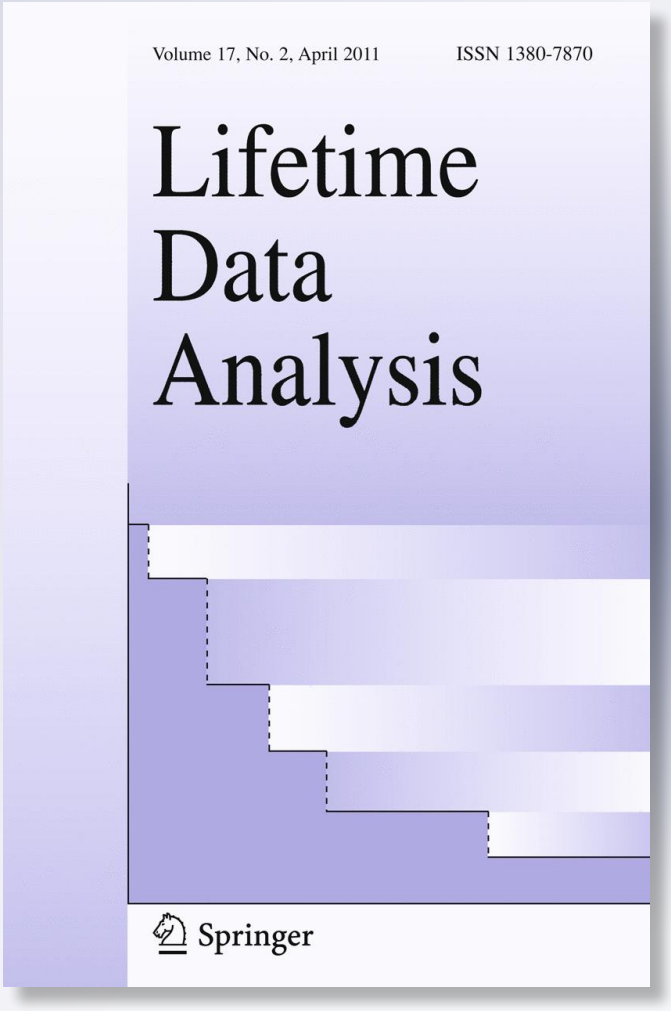
Analysis of recurrent events with non-negligible event duration, with application to assessing hospital utilization

Lifetime Data Analysis
An International Journal
Devoted to Statistical Methods
and Applications for Time-to-
Event Data

ISSN 1380-7870
Volume 17
Number 2

Lifetime Data Anal (2010)
17:215-233
DOI 10.1007/
s10985-010-9183-8



Springer

Lifetime Data Anal (2011) 17:215–233
DOI 10.1007/s10985-010-9183-8

Analysis of recurrent events with non-negligible event duration, with application to assessing hospital utilization

X. Joan Hu · Maria Lorenzi · John J. Spinelli · S. Celes Ying · Mary L. McBride

Received: 24 April 2009 / Accepted: 5 August 2010 / Published online: 22 August 2010

Abstract In an attempt to provide tools for assessing hospital utilization, this paper extends well-known models for recurrent events to address non-negligible event duration and presents a procedure for estimating the model parameters. The model extension is natural and easy to understand. Asymptotic properties of the associated inferences are derived adapting the well-developed methods based on the counting process formulation. Several specifications of the proposed modeling are illustrated with the hospitalization records of childhood cancer survivors from a health care insurance system that motivated this research. The usefulness and robustness of the proposed approach is demonstrated numerically via simulation.

Keywords Proportional intensity · Semiparametric regression · Stratified Cox regression model · Time dependent covariates · Time dependent stratification

1 Introduction

Recent successful improvements in treatment have increased childhood cancer survival rate dramatically. An on-going research program of British Columbia

X. J. Hu (✉) · M. Lorenzi
Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
e-mail: joanh@stat.sfu.ca

M. Lorenzi · J. J. Spinelli · M. L. McBride
Cancer Control Research, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada

S. C. Ying
Abbott Vascular, 3200 Lakeside Drive, Santa Clara, CA, USA

Cancer Agency in Canada, the Childhood/Adolescent/Young Adult Cancer Survivors (CAYACS) program, is to assess the morbidity and long-term resource needs of young cancer survivors using a population health insurance database. A specific aim is to evaluate the hospitalization frequency of the young cancer survivors and to identify the risk factors. When the frequency of hospitalization over time (or the number of hospital admissions, to be more specific) is of primary interest, the problem is formulated into an analysis of recurrent events.

There has been substantial interest in the statistical analysis of recurrent events (i.e., multiple events of the same type per individual). See Therneau and Grambsch (2000) and Cook and Lawless (2007) for various examples and comprehensive reviews on the research in the area. However, most published papers on recurrent event analysis assume negligible event duration, although there are a few notable exceptions (Twisk et al. 2005; Cook and Lawless 2007, Chp 6).

Since a patient who is being hospitalized at a time point is not at risk for a further hospitalization at the same time, one needs to adjust for hospital duration in the analysis of hospitalization records, especially in situations where the hospital stay is considerably long. Otherwise, those patients who have a long hospital stay will likely be categorized as ones with low risk to be admitted to hospital, and the inference on the associated risk factors could be seriously biased. This consideration motivated us to adapt a generalized Cox regression model, which is an extension of the well-known models for recurrent events given in Prentice et al. (1981), Andersen and Gill (1982), and Pepe and Cai (1993). Our modeling method accommodates non-negligible event duration with an adjusted definition for the risk set at a time point, the group of individuals who are at the "risk" of experiencing the event at the time.

Along the lines that Kelly and Lim (2000) use to characterize extensions of the Cox proportional hazards model to handle multiple event data, the four key components of our model are that (i) the elapsed time since study entry is utilized, (ii) either an event-specific or common baseline intensity function is considered, (iii) the risk set of the k-th event stratum is restricted to the individuals who are still in the study and have experienced the whole duration of the events in the previous event stratum, the $(k-1)$-th stratum, and (iv) the within-subject correlation between events is accounted for by introducing additional time-dependent covariates in the model.

People may choose to consider two other methods for analysis of hospitalization records. Cook and Lawless (2007, Chp 6) model the hospitalization process as an alternating two-state process, which parallels the two states, "in hospital" and "out of hospital", and models the transitions between the two states. Our approach focuses on the conditional intensity of the event occurrence, and does not require one to model the event duration. It is suitable for studies such as the young cancer survivor hospitalization project, where the particular interest is in the frequency of the primary event and the associated risk factors. Another advantage of restricting attention to event occurrence is that it allows us to adapt the well-established approaches in recurrent event analysis in order to make inferences and to derive asymptotic properties of the inferences.

Due to their relaxed model assumptions and the ability to achieve greater robustness for inferences, methods that focus on marginal characteristics of the event process, such as the ones presented in Lawless and Nadeau (1995), Lin et al. (2000), and

Cai and Schaubel (2004), are often preferred when inferences on the marginal characteristics are of the primary interest. In many practical situations, the status of a subject at a particular time very likely depends on his history information, such as the duration of his previous event. Specifying this dependence could gain valuable insights into the event process. These consideration lead us to choose an intensity-based approach.

This paper focuses on modeling the risk of the primary event where the event is recurrent and the event duration is non-negligible. We extend the Cox proportional hazards model by introducing time-dependent stratification and using the adjusted risk set to accommodate event duration. The presented analysis procedure is an adaptation of the procedure for estimating the parameters of the Cox model. It can be implemented with available procedures/functions in SAS, Splus or R, similar to, for example, what is described in Therneau and Grambsch (2000, Chp 8) on how to use Splus to make inferences under the models given in Prentice et al. (1981), and Andersen and Gill (1982). Our approach is thus anticipated to be accessible to many practitioners in biomedical research and epidemiology. The rest of the paper is organized as follows. Section 2 introduces the framework and the proposed modeling. Section 3 presents an estimation procedure and the associated asymptotic results. In Sect. 4, the proposed approach is applied to analyze the hospitalization data that motivated this research. Section 5 reports the simulation study designed to examine numerically the proposed approach. Some final remarks are given in Section 6.

2 Statistical modeling

Consider a study with the primary interest in a particular recurrent event. Denote the times to occurrences of the event by $T_1, T_2, \ldots$, and the times when the event durations end by $V_1, V_2, \ldots$, where $0 < T_1 < V_1 < T_2 < V_2 < \ldots$. With the hospitalization process mentioned above, for example, T_j and V_j are the times at admission and discharge of the j-th hospitalization, respectively. Define $N(t) = \sum_{j=1}^{\infty} \mathrm{I}(T_j \le t)$ and $V(t) = \sum_{j=1}^{\infty} \mathrm{I}(V_j < t)$. Here $N(t)$ and $V(t)$ represent the numbers of event occurrences that an individual in the study has started and completed, respectively, up to time $t \ge 0$ since his/her study entry at time 0. Note that $V(t) \le N(t)$, and $N(\cdot)$ and $V(\cdot)$ are right- and left-continuous counting processes, respectively. Let $Z(t)$ be the vector of external covariates for a subject at time t and suppose it is left-continuous. Denote the history information of a subject up to time t by $\mathcal{H}(t) = \mathcal{N}(t) \bigcup \mathcal{V}(t) \bigcup \mathcal{Z}(t)$, where $\mathcal{N}(t) = \{N(s) : 0 \le s < t\}$, $\mathcal{V}(t) = \{V(s) : 0 \le s < t\}$, and $\mathcal{Z}(t) = \{Z(s) : 0 \le s < t\}$. Let $W(\mathcal{N}(t), \mathcal{V}(t))$ be the additional covariates determined by the information of $N(\cdot)$ and $V(\cdot)$ preceding t. We consider a generalized Cox regression model: for $t > 0$,

$$\begin{aligned}\lambda(t\,|\mathcal{H}(t)) &= \lim_{\Delta t \to 0+} \mathrm{P}\left\{N(t) - N(t-\Delta t) = 1 \,|\, \mathcal{H}(t)\right\} \big/ \Delta t \\ &= Y^V(t)\lambda_0\left\{t; \mathcal{H}(t)\right\} \exp\left\{\beta(t; \mathcal{H}(t))' Z(t) + \alpha(t; \mathcal{H}(t))' W(\mathcal{N}(t), \mathcal{V}(t))\right\}, \end{aligned} \tag{1}$$

where $Y^V(t) = \mathrm{I}\{N(t-) = V(t-)\}$ is the indicator for whether the previous (i.e., the $N(t-)$-th) event has completed, $\lambda_0\{t; \mathcal{H}(t)\}$ is an arbitrary baseline

Table 1 Important special cases of the generalized stratified cox regression model in (2)

Model component	Models in the literature			Extended models in this paper		
	PWP[a]	PWP2[b]	AG[c]	GPWP[d]	GPWP2[e]	GAG[f]
$Y^V(t)$	1	1	1	*	*	*
$\beta_s(t)$	β_s	β	β	*	$\beta(t)$	$\beta(t)$
$\alpha_s(t)$	0	0	0	*	$\alpha(t)$	$\alpha(t)$
$\lambda_{0s}(t)$	*	*	$\lambda_0(t)$	*	*	$\lambda_0(t)$

[a] Model (2) in Prentice, Williams and Peterson (1981): $\lambda(t \mid \mathcal{H}(t)) = \lambda_{0s}(t)e^{\beta_s' Z(t)}$

[b] A special case of PWP: $\lambda(t \mid \mathcal{H}(t)) = \lambda_{0s}(t)e^{\beta' Z(t)}$

[c] Model in Andersen and Gill (1982): $\lambda(t \mid \mathcal{H}(t)) = \lambda_0(t)e^{\beta' Z(t)}$

[d] An extension of PWP: $\lambda(t \mid \mathcal{H}(t)) = Y^V(t)\lambda_{0s}(t)e^{\beta_s(t)' Z(t)+\alpha_s(t)' W(t)}$

[e] An extension of PWP2: $\lambda(t \mid \mathcal{H}(t)) = Y^V(t)\lambda_{0s}(t)e^{\beta(t)' Z(t)+\alpha(t)' W(t)}$

[f] An extension of AG: $\lambda(t \mid \mathcal{H}(t)) = Y^V(t)\lambda_0(t)e^{\beta(t)' Z(t)+\alpha(t)' W(t)}$

* Not-specified/arbitrary

function, and both $\beta(t; \mathcal{H}(t))$ and $\alpha(t; \mathcal{H}(t))$ are known functions up to finite dimensional parameters. Here the dimensions of $\beta(t; \mathcal{H}(t))$ and $\alpha(t; \mathcal{H}(t))$ are the same as $Z(t)$ and $W(\mathcal{N}(t), \mathcal{V}(t))$, respectively. We use $A(t-)$ for the left limit of the function $A(\cdot)$ at time t. The indicator $Y^V(\cdot)$ in the model addresses the issue of non-negligible duration. It induces the definition of the risk sets to an adjusted one, which does not include in the risk sets the individuals who are experiencing the event. The conditional intensity function of $N(\cdot)$ in (1) can be viewed as the conditional hazard function of $T_{N(t)}$, the time to the current (i.e., the $N(t)$-th) event occurrence.

Introducing a stratification variable $s\{\mathcal{H}(t)\}$, which is fully determined by the subject's history information, we consider a specification of the model in (1): for $t > 0$ and $s\{\mathcal{H}(t)\} = s$,

$$\lambda(t|\mathcal{H}(t)) = Y^V(t)\lambda_{0s}(t)\exp\left\{\beta_s(t)' Z(t) + \alpha_s(t)' W(\mathcal{N}(t), \mathcal{V}(t))\right\}, \tag{2}$$

where $\beta_s(t)$ and $\alpha_s(t)$ are known functions of t up to parameters β_s and α_s, respectively. The model specification (2) gives an extension of the stratified Cox regression model proposed by Prentice et al. (1981). In the situations where T_j are very close to V_j and $\alpha_s(t) \equiv 0$ and $\beta_s(t) = \beta_s$, it reduces to one of the two semiparametric models in Prentice et al. (1981), and includes the model of Andersen and Gill (1982) as a special case. Table 1 presents important examples of the extended stratified Cox regression model (2). It shows how the models are related to the well-known models by specifying the components in (2) into special forms. We will illustrate the examples further in Sect. 4 with the analysis of the hospitalization data that motivated this research.

The definition of the stratification variable $s\{\mathcal{H}(t)\}$ in (2) may vary from study to study. This paper assumes $s\{\mathcal{H}(t)\}$ to be positive-integered and up-bounded by, say, $S < \infty$. An important special case is that $s\{\mathcal{H}(t)\} = N(t-)$, the total number of events

having been experienced before time t. Section 4 presents another stratification based on the history information of the counting process $N(\cdot)$ using the hospital records that motivated this research, which in general works well in the situations where the total number of events is relatively large and there is relatively little information on the last several events. There can be other types of stratification in practice, such as the ones associated with the covariates $Z(\cdot)$. The model in (2) can capture the differences in event occurrences between strata with different baselines of intensity function and covariate effects. Furthermore, we note that, using an appropriate definition of the stratification variable, the setting above can accommodate the situations where the events are ordered and of different types.

The additional covariates $W(\mathcal{N}(t), \mathcal{V}(t))$, denoted by $W(t)$ in the following, are assumed to bear all the information about the events previous to time t that may affect the occurrence of the next event within each stratum. This addition together with the stratification described above allows the model to accommodate processes without the property of independent increments, that is, non-Poisson processes. In the hospitalization project, for example, $W(t)$ can be specified as $V_{N(t-)} - T_{N(t-)}$, the length of the $N(t-)$-th hospital stay, or $\sum_{j=1}^{N(t-)}(V_j - T_j)$, the total hospitalization duration since the study entry. It is unlikely that in all applications there is such a time-dependent variable $W(t)$. An alternative approach would be to employ frailty models. See, for example, Oakes (1992) for details about frailty models for multiple events.

The stratum-specific regression coefficients $\beta_s(t)$ and $\alpha_s(t)$ in (2) can be time-independent or some particular functions of t to accommodate time-dependent covariate effects. The time-varying coefficients $\beta_s(t)$ and $\alpha_s(t)$ may be specified as linear functions of some known functions, i.e., $\beta_s(t) = \beta_{s0} + \beta_{s1}g(t)$ and $\alpha_s(t) = \alpha_{s0} + \alpha_{s1}h(t)$ with known functions $g(t)$ and $h(t)$. This is a special case of what is considered in Grambsch and Therneau (1994) for time-varying coefficients in proportional hazards testing, which is implemented through the Splus function *cox.zph*. We denote the time-varying coefficients by $\beta(t; \beta_s)$ and $\alpha(t; \alpha_s)$. The relative risk parameters (β_s, α_s) can vary from stratum to stratum.

Another semiparametric model of Prentice et al. (1981), labeled as formula (3) in their paper, is in the same form as the model (2) but using the time scale that measures the time to an event since the last event, rather than from the study entry. Some practical situations may be better formulated by this second specification. In fact, we may use the structure of the model in (2) to incorporate various time scales, such as Markov time-scales. We will explore this topic in a later manuscript.

3 Estimation procedure

Consider a study with n iid realizations of $\{N(\cdot), V(\cdot), Z(\cdot)\}$, denoted by $\{N_i(\cdot), V_i(\cdot), Z_i(\cdot)\}$, $i = 1, \ldots, n$. Suppose the data collection associated with subject i is subject to non-informative censoring conditional on $Z_i(\cdot)$, with the censoring time C_i and the censoring indicator $Y_i^C(t) = \mathrm{I}(t \le C_i)$. Furthermore, let $\mathcal{R}(t; s)$ be the study's risk set of the s-th stratum at time t. That is, $\mathcal{R}(t; s)$ contains all those subjects who are in the stratum s, not experiencing an event, and have not been censored at time t:

$$\mathcal{R}(t; s) = \{i : s\{\mathcal{H}_i(t)\} = s; N_i(t-) = V_i(t-); t \le C_i\}. \quad (3)$$

In this section, we present a procedure with right-censored data for estimating the unknown parameters involved in the model (2): the coefficients for the covariates and the cumulative baseline intensity functions.

3.1 Estimating the coefficient parameters

We first assume there are no ties, and denote the ordered event times in stratum s by $t_{s,1} < \ldots < t_{s,d_s}$, where d_s is the total number of the event times and $s = 1, \ldots, S$. Using the approach of Prentice et al. (1981), a generalized version of the partial likelihood derivation given in Cox (1975), the overall likelihood function based on the data can be factored into two terms, one of which is the partial likelihood of the regression coefficients (β_s, α_s):

$$PL(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{s=1}^{S} \prod_{j=1}^{d_s} \frac{\exp\left\{\beta(t_{s,j}; \beta_s)' Z_{i(t_{s,j})}(t_{s,j}) + \alpha(t_{s,j}; \alpha_s)' W_{i(t_{s,j})}(t_{s,j})\right\}}{\sum_{l \in \mathcal{R}(t_{s,j}; s)} \exp\left\{\beta(t_{s,j}; \beta_s)' Z_l(t_{s,j}) + \alpha(t_{s,j}; \alpha_s)' W_l(t_{s,j})\right\}}, \quad (4)$$

where $i(t_{s,j})$ is the index of the subject who experiences the event occurrence in stratum s at time $t_{s,j}$, and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are, respectively, either $(\beta_1, \ldots, \beta_S)$ and $(\alpha_1, \ldots, \alpha_S)$ or the common values of β_s's and α_s's when the coefficients are the same across the strata. We can estimate the coefficients $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ by maximizing the partial likelihood function (4). Ordinary asymptotic likelihood methods can be applied to (4) to derive asymptotic results of the estimator of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (Cox 1975; Prentice et al. 1981). Some consideration should be given to the size and the number of events of each stratum, according to the specification of the model (2). We will discuss these issues more in the context of the hospitalization project in Sect. 4.

The log-partial likelihood function of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be expressed as $\log\{PL(\boldsymbol{\beta}, \boldsymbol{\alpha})\} = C(\boldsymbol{\beta}, \boldsymbol{\alpha}; \infty)$ with

$$C(\boldsymbol{\beta}, \boldsymbol{\alpha}; t) = \sum_{i=1}^{n} \int_0^t Y_i^C(u) \sum_{s=1}^{S} \Bigg[\beta(u; \beta_s)' Z_i(u) + \alpha(u; \alpha_s)' W_i(u) - \log\left\{ \sum_{j \in \mathcal{R}(u;s)} \exp\{\beta(u; \beta_s)' Z_j(u) + \alpha(u; \alpha_s)' W_j(u)\} \right\} \Bigg] \Bigg|_{s = s\{\mathcal{H}_i(u)\}} dN_i(u). \quad (5)$$

Adapting the well-developed counting process formulation for event history data analysis and the asymptotics derivation using the martingale results presented in, for example, Fleming and Harrington (1991) and Andersen et al. (1992), we can verify

the consistency and asymptotic normality of the maximum partial likelihood estimator (MPLE) of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, the maximum point of (5) at $t = \infty$. The variance of the MPLE, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ is approximately the inverse of the second derivative of $\log\{PL(\boldsymbol{\beta}, \boldsymbol{\alpha})\}$. An outline of the derivation and important technical results are presented in the Appendix. Furthermore, note from expression (5) that the counting process formulation accommodates the situations with tied event times, where are given similar treatments as to the discussion on tied failure times, for example, in Kalbfleisch and Prentice (1980, Chp 4).

A natural testing procedure can be constructed based on the asymptotic normality of the MPLE on the coefficient for a covariate in the model (2) for assessing the covariate effect. We may also provide procedures for model checking on the assumption of the proportional intensity based on the estimation of time-dependent coefficients. The testing procedures are illustrated with the hospitalization data in Sect. 4.

3.2 Estimating the baseline intensity functions

With fixed $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, the following estimating equations are unbiased:

$$\sum_{i \in \mathcal{R}(t;s)} Y_i^C(t) \left[dN_i(t) - Y_i^V(t) \lambda_{0s}(t) \exp\{\beta(t;\beta_s)' Z_i(t) + \alpha(t;\alpha_s)' W_i(t)\} dt \right] = 0,$$

for $t > 0$ and $s = 1, \ldots, S$. This motivates the following estimation procedure. When the baseline intensity function in model (2) varies from stratum to stratum, a consistent estimator of the cumulative baseline intensity function of stratum s is

$$\hat{\Lambda}_{0s}(t;\beta_s,\alpha_s) = \int_0^t \frac{\sum_{i \in \mathcal{R}(u;s)} dN_i(u)}{\sum_{j \in \mathcal{R}(u;s)} \exp\{\beta(u;\beta_s)' Z_j(u) + \alpha(u;\alpha_s)' W_j(u)\}}, \quad t > 0, \tag{6}$$

for $s = 1, \ldots, S$. Here we take the convention $0/0 = 0$. In the situations with a single cumulative baseline intensity function, that is, $\lambda_{0s}(t) = \lambda_0(t)$ for $s = 1, \ldots, S$, we estimate the baseline function using

$$\hat{\Lambda}_0(t;\boldsymbol{\beta},\boldsymbol{\alpha}) = \int_0^t \frac{\sum_{i=1}^n Y_i^C(u) dN_i(u)}{\sum_{s=1}^S \sum_{j \in \mathcal{R}(u;s)} \exp\{\beta(u;\beta_s)' Z_j(u) + \alpha(u;\alpha_s)' W_j(u)\}}, \quad t > 0. \tag{7}$$

With some regularity conditions and by the martingale central limit theorem, we can show that $\hat{\Lambda}_{0s}(t;\boldsymbol{\beta},\boldsymbol{\alpha})$ and $\hat{\Lambda}_0(t;\boldsymbol{\beta},\boldsymbol{\alpha})$, after standardization, converge weakly to Gaussian processes with mean zero.

The above estimators are generalizations of the Breslow estimator for the baseline functions under model (2). They can be used to estimate the cumulative intensity function of a particular group with the covariates fixed at the corresponding levels, without

Table 2 Characteristics of CAYACS subjects diagnosed with leukemia (Total $n = 349$)

Diagnosis period		Age at diagnosis (in years)		Gender		Previous hospitalization (in days)		dropout study	
Category	Size	Category	Size	Category	Size	Category	Size	Category	Size
1980–1985	84	$0 \sim 5$	193	Male	198	None	39	No	286
1985–1990	119	$5 \sim 13$	107	Female	151	(0,30]	144	Yes (death)	19
1990–1995	146	$13 \sim 20$	49			(30,60]	107	Yes (migration)	44
						(60,90]	42		
						>90	56		

any further model assumption. This indicates an approach for model checking. We will illustrate it with the hospitalization data in Sect. 4. The nonparametric versions of (6) and (7), taking $\beta(t; \beta_s) = 0$ and $\alpha(t; \alpha_s) = 0$, may be viewed as generalizations of the Nelson-Aalen estimator for the cumulative intensity functions with adjustment for event duration.

We obtain estimators for the baseline intensity functions $\Lambda_{0s}(\cdot)$ and $\Lambda_0(\cdot)$ conventionally by plugging in the corresponding MPLE of the unknown coefficients in (6) and (7), respectively. Given the continuity of $\hat{\Lambda}_{0s}(t; \boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\hat{\Lambda}_0(t; \boldsymbol{\beta}, \boldsymbol{\alpha})$ as functions of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, the resulting estimators $\hat{\Lambda}_{0s}(\cdot; \hat{\beta}_s, \hat{\alpha}_s)$ and $\hat{\Lambda}_0(\cdot; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ are consistent and, after standardization, weakly converge to mean zero Gaussian processes.

4 Practical example

The CAYACS program that motivated this research is concerned with people who were diagnosed with cancer between 1981–1995, before the age of 20 in British Columbia, Canada, and had survived five years or longer after their diagnosis. See MaCarthur et al. (2007a,b) and Ying (2006) for more details. One of its study objectives is to evaluate the hospital utilization during 1986–2000 of the cancer patients and to identify risk factors. We analyzed the study data with the approach discussed in the previous sections. For simplicity, the analysis presented below focuses on the study subjects who were diagnosed with leukemia.

Among the potential risk factors identified by the study team, three external covariates were considered: diagnosis period (in 1981–1985, 1985–1990 or 1990–1995), age at diagnosis, and gender. To evaluate the potential association of the hospitalization process during the study with previous hospitalizations, we took the total length of all hospitalizations after the diagnosis till the study beginning (i.e., up to five year survival) as an additional time-independent covariate. Table 2 gives a summary of the patient characteristics. The group includes 349 subjects with an average age at diagnosis of 6.2 years old, and the average length in the study of 6.25 years. Among the 497 recorded hospitalizations from the patient group, there were 144, 95, 102, 92, 47 and 17 cases with the length of hospital stay of $1, 2, 3-4, 5-9, 10-30$ and longer than

30 days, respectively. A total of 63 patients ended the study earlier than anticipated due to either out-migration or death. Because out-migration can be safely assumed response-noninformative and the mortality rate of the study subjects was low (19 out of 349 subjects died during the study period), we assumed in the analysis that the censoring is noninformative.

We first grouped the subjects according to the diagnosis period, age at diagnosis (0–5, 5–13 and 13–20 years), and gender. For each of the different subgroups, we obtained the Nelson-Aalen estimate of the cumulative intensity function and evaluated the generalized Nelson-Aalen estimator in Sect. 3.2 to adjust for hospital duration. The nonparametric estimates indicate that the cumulative intensity function of the study population is likely proportional to the three risk factors. In addition, we stratified the hospitalizations into five strata: the first, second, third, fourth to sixth, and seventh or later hospital utilizations during the study. That is, we used the stratification variable $s\{\mathcal{H}(t)\} = 1, 2, 3, 4,$ or 5 for $N(t-) = 0, 1, 2, N(t-) \in \{3, 4, 5\}$, or $N(t-) \geq 6$, respectively. The Nelson-Aalen estimates for the cumulative conditional intensity functions of the five strata, without or with adjustment for hospital duration, were obtained. The estimates differ from stratum to stratum. They show that the larger the number of previous hospitalizations is, the larger the risk for having the next hospitalization becomes. The nonparametric estimates without adjustment for hospital duration are very close to the ones with adjustment for the duration. This is probably due to the relatively long study period of each individual, compared to the length of his/her hospital stay.

We evaluated the maximum partial likelihood estimates (MPLE) of the regression parameters β in the models given in Table 1 with the four time-independent covariates mentioned above, assuming all the risk effects were time-independent. Table 3 presents the MPLE of β and the values of the log-partial likelihood functions at the parameter estimates with the models. The associated estimated standard errors are shown in parentheses. The coefficient estimates of the covariates with significant effects are bolded in the table.

The results indicate that the covariate effects vary from stratum to stratum: the p-values of the partial likelihood ratio tests for comparing the fit of the two models of Prentice et al. (1981), denoted by PWP and PWP2 in Table 1, and the two generalized PWP models, denoted by GPWP and GPWP2 in Table 1, are $p < 0.001$ and $p = 0.003$ respectively. The average covariate effects across the strata obtained under models PWP2 and GPWP2 are close except for gender. It appears that the difference between females and males became non-statistically significant when the risk sets were adjusted for hospital duration. This is probably due to the relatively lower frequency of hospitalizations and longer hospital stays in the male group, compared to the female group.

The overall covariate effects without stratification, i.e., with the model of Andersen and Gill (1982) and the generalized version, denoted by AG and GAG respectively in Table 1, are in agreement with the average covariate effects under models PWP2 and GPWP2 except for the indicator of diagnosis period II. Results from all models indicate that the patients with diagnosis period III have a statistically significantly lower risk, and the patients diagnosed at an older age have a higher hospitalization rate than those diagnosed at a younger age. There was no statistically significant difference detected

Table 3 Maximum partial likelihood estimates of regression parameters β_s with four time-independent covariates[a]

Risk factor	with risk set not-adjusted for hospital duration ($Y^V(t) = 1$)						
Model	PWP					PWP2	AG
	β_1	β_2	β_3	β_4	β_5	$\beta(=\beta_s)$	$\beta(=\beta_s,)$ $\lambda_{0s} = \lambda_0)$
log PL			−1296.26			−1335.63	−2616.81
Diagnosis period II	−0.262	**0.499**	0.388	**0.590**	**0.647**	**0.464**	0.037
	(0.201)	(0.300)	(0.370)	(0.244)	(0.259)	(0.109)	(0.100)
Diagnosis period III	**−1.029**	**−1.439**	**1.500**	−0.275	**−3.548**	**−0.797**	**−1.150**
	(0.292)	(0.515)	(0.849)	(0.425)	(1.079)	(0.196)	(0.173)
Age at diagnosis (in years)	0.010	**0.052**	**0.066**	0.042	−0.001	**0.034**	**0.030**
	(0.018)	(0.030)	(0.034)	(0.026)	(0.024)	(0.010)	(0.009)
Male (vs female)	−0.044	−0.093	**0.691**	0.040	**−1.069**	**−0.201**	−0.074
	(0.180)	(0.303)	(0.393)	(0.255)	(0.247)	(0.103)	(0.091)
Hospitalization length before study (in months)	**0.184**	**0.061**	−0.033	0.004	**−0.086**	**0.037**	**0.123**
	(0.031)	(0.028)	(0.034)	(0.034)	(0.026)	(0.008)	(0.029)
	with risk set adjusted for hospital duration						
Model	GPWP					GPWP2	GAG
	β_1	β_2	β_3	β_4	β_5	$\beta(=\beta_s)$	$\beta(=\beta_s,$ $\lambda_{0s} = \lambda_0)$
log PL			−1061.00			−1077.20	−2612.15
Diagnosis period II	−0.265	**0.657**	**0.863**	**0.936**	**1.203**	**0.658**	0.032
	(0.201)	(0.311)	(0.472)	(0.253)	(0.326)	(0.113)	(0.100)
Diagnosis period III	**−1.034**	**−1.593**	**2.417**	0.284	**−3.813**	**−0.711**	**−1.161**
	(0.292)	(0.519)	(0.901)	(0.420)	(1.095)	(0.197)	(0.173)
Age at diagnosis (in years)	0.010	**0.068**	**0.199**	**0.073**	0.019	**0.048**	**0.030**
	(0.018)	(0.031)	(0.056)	(0.028)	(0.029)	(0.011)	(0.009)
Male (vs female)	−0.046	−0.063	**2.096**	**0.440**	**−1.177**	−0.104	−0.070
	(0.180)	(0.325)	(0.617)	(0.265)	(0.275)	(0.105)	(0.091)
Hospitalization length before study (in months)	**0.182**	**0.057**	0.061	0.033	**−0.094**	**0.028**	**0.124**
	(0.027)	(0.025)	(0.069)	(0.031)	(0.029)	(0.007)	(0.003)

[a] Estimated standard error in parentheses; significant effect with P-value ≤ 0.05 in **boldface**

for gender under either AG or GAG. Moreover, the results indicate that the additional covariate, the total length of hospital stay before the study, was statistically significant under all models, PWP2, AG, GPWP2 and GAG. A subject who had longer hospital stay in the first 5 years post-diagnosis tends to have higher risk of hospitalization after the five-year-survival.

We estimated the associated baseline intensity functions using (6) or (7) for models AG/GAG or PWP2/GPWP2. Figure 1 (a) presents two sets of estimates of the cumulative intensity functions with the subjects stratified according to their diagnosis periods, which are the generalized Nelson-Aalen estimates and the semiparametric estimates based on model GAG with hypothetical individuals with covariates at the average levels. It indicates a good fit of the semiparametric model, model GAG. Figure 1 (b) gives two sets of estimates of the cumulative intensity functions with stratification according to previous number of hospitalizations. Again it indicates the GPWP2 fits reasonably well.

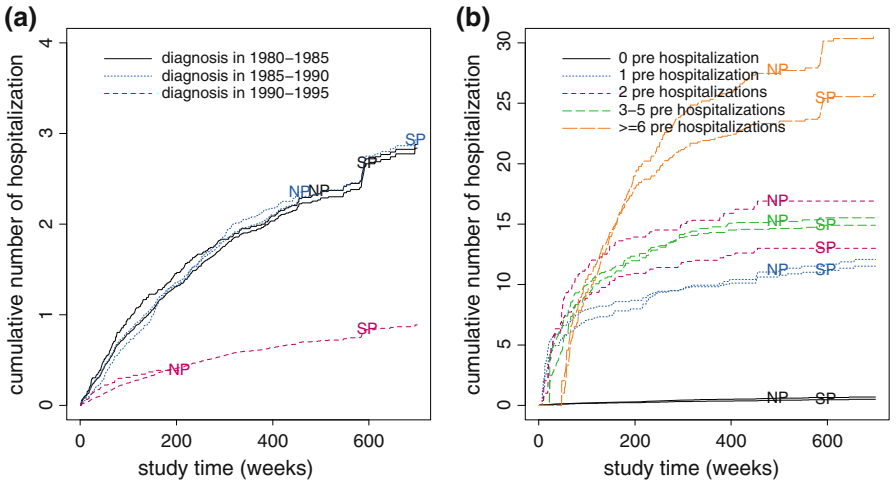


Fig. 1 Model fitting: "NP" for generalized Nelson-Aalen estimates; "SP" for semiparametric estimates under GAG/GPWP2. (**a**) GAG fitting: diagnosis period. (**b**) GPWP2 fitting: previous hospitalization

To understand the discrepancy in assessing the effect of diagnosis period II between PWP2 and AG, and GPWP2 and GAG, we considered models with a time-dependent effect of the indicator of diagnosis period II. In addition, we considered a time-dependent covariate, the length (in months) of the previous hospitalization. Table 4 summarizes the further analysis results with models AG and GAG. The obtained time-varying differences between the diagnosis period II group and the diagnosis period I group by AG and GAG are in a close agreement with the difference between the two sets of corresponding curves shown in Figure 1 (a). The results indicate a statistically significant time trend for the effect of diagnosis period II. The estimates for the slope of the time-varying effect together with the estimates for the intercept term explain the discrepancy of results from PWP2 and AG, and from GPWP2 and GAG. Although both PWP2 (or GPWP2) and AG (or GAG) summarize the covariate effect with a single parameter, they summarize it differently: while PWP2 (or GPWP2) has stratified baselines, AG (or GAG) has a common baseline. In addition, no statistically significant effect of the new time-dependent covariate, the length of the previous hospital stay, was detected.

We remark that the models with stratification can identify variation across strata, which is rather key when the variation is large. When the primary interest is in identifying risk factors, however, the unstratified models (AG/GAG) may be preferred, since the associated analyses are easier to conduct and the outcomes are easier to interpret and to convey to the medical investigators. The PWP2/GPWP2 models can be used to support the unstratified models when analyzing similar data. While having different baselines to address the potential variation across the strata, the PWP2/GPWP2 models provide an averaged risk effect for each of the risk factors with the corresponding coefficient. Finally, an adjustment for the event duration needs to be made when it is considerably long; otherwise, there could be misleading results such as the statistically significant difference between the female and male groups "detected" by the conventional approaches.

Table 4 Maximum partial likelihood estimates of regression parameters β_s in AG/GAG models with time-dependent effect/time-dependent factor[a]

Risk Factor	Model	$\beta_s(t)=\beta$ $\alpha_s(t)=0$	$\beta_s(t)=\beta+\beta_1\log t$ $\alpha_s(t)=0$	$\beta_s(t)=\beta$ $\alpha_s(t)=\alpha$	$\beta_s(t)=\beta+\beta_1\log t$ $\alpha_s(t)=\alpha$
	AG approaches	with risk set not-adjusted for hospital duration ($Y^V(t)=1$)			
	$\log PL(\hat{\beta},\hat{\alpha})$	−2616.81	−2613.92	−2615.39	−2613.91
Diagnosis period II		0.037	**−1.740**	0.038	**−1.738**
		(0.100)	(0.550)	(0.100)	(0.332)
Diagnosis period III		**−1.150**	**−1.263**	**−1.150**	**−1.263**
		(0.173)	(0.175)	(0.173)	(0.116)
Age at diagnosis (in years)		**0.030**	**0.031**	**0.030**	**0.031**
		(0.009)	(0.009)	(0.009)	(0.005)
Male (vs female)		−0.074	−0.074	−0.072	−0.073
		(0.091)	(0.091)	(0.091)	(0.058)
Length of hospitalization before study (in months)		**0.123**	**0.117**	**0.120**	**0.123**
		(0.034)	(0.032)	(0.033)	(0.034)
Time-dept effect of diag period II (β_1, coefficient to $\log t$)		–	**0.269**	–	**0.269**
		–	(0.082)	–	(0.049)
Length of last hospital stay (in months)		–	–	−0.027	−0.012
		–	–	(0.156)	(0.153)
	GAG approaches	with risk set adjusted for hospital duration			
	$\log PL(\hat{\beta},\hat{\alpha})$	−2612.15	−2609.19	−2610.69	−2609.19
Diagnosis period II		0.032	**−1.758**	0.033	**−1.757**
		(0.100)	(0.550)	(0.100)	(0.551)
Diagnosis period III		**−1.161**	**−1.274**	**−1.161**	**−1.274**
		(0.173)	(0.175)	(0.173)	(0.175)
Age at diagnosis (in years)		**0.030**	**0.031**	**0.030**	**0.031**
		(0.009)	(0.009)	(0.009)	(0.009)
Male (vs female)		−0.070	−0.070	−0.069	−0.070
		(0.091)	(0.091)	(0.091)	(0.091)
Length of hospitalization before study (in months)		**0.124**	**0.123**	**0.123**	**0.121**
		(0.003)	(0.003)	(0.003)	(0.003)
Time-dept effect of diag period II (β_1, coefficient to $\log t$)		–	**0.271**	–	**0.271**
		–	(0.082)	–	(0.082)
Length of last hospital stay (in months)		–	–	−0.012	−0.006
		–	–	(0.156)	(0.157)

[a] Estimated standard error in parenthese; significant effect with P-value ≤ 0.05 in **boldface**

5 Simulation study

We examined the proposed approach numerically via simulation. A hospitalization study was simulated with a total of $n = 300$ independent subjects, who were allocated equally likely to Groups A or B. The event occurrences associated with each subject were simulated following a GPWP model (an extended stratified Cox regression model) with three strata in total. Specifically, having introduced the indicator of Group A, $z_i = 1$ or 0 if subject i was from Group A or not, the hospitalization process of subject i from the study was generated as follows. For $k \geq 1$, given his $(k-1)$-th discharge time $V_{i,k-1}$ with $V_{i,0} = 0$,

Step 1. Subject i's time to the k-th hospital admission, denoted by $T_{i,k}$, was generated from the distribution with the hazard function $\lambda_{0s}\exp\{(\beta_{0s}+\beta_{1s}\log t)z_i\}$ with $s=\min(k, S)$, where S is a pre-determined integer.
Step 2. The duration time of the k-th hospitalization, denoted by $W_{i,k}$, was generated from the uniform distribution $U(0, D_A)$ or $U(0, D_B)$ according to subject i from Group A or B (i.e., $z_i=1$ or not), where D_A and D_B are the longest hospital stays associated with Groups A and B, respectively.
Step 3. The k-th discharge time was then attained as $V_{i,k}=T_{i,k}+W_{i,k}$.

We fixed $S=3$ and chose the baseline parameters λ_{0s} to be $1/40$, $1/30$ and $1/20$ for $s=1,2$ and 3, respectively. We used $D_A=8$ or 30 (months) and $D_B=5$ (months) to simulate situations with relatively short or long hospital stay in Group A, and with similar length or not in the two groups. The data collection was simulated as it was subject to a noninformative censoring, where the censoring times (C_i) were independent to the event processes and from the truncated uniform distribution $\min\{U(c_{\min}, c_{\max}), 180\}$ (months), with $c_{\min}=90$ and $c_{\max}=300$. The following procedures were taken to attain estimates of the regression parameters and variance estimates of the regression estimators for $s=1,2,3$ as well as the baseline functions with different model specifications.

(a). *PWP/GPWP Procedures:* the simulated event processes were analyzed with the PWP/GPWP models, the hazard functions of the times to hospital admissions being $\lambda_{0s}(t)\exp\{(\beta_{0s}+\beta_{1s}g(t))z_i\}$ with $t>t^*$ for PWP or $t>v^*$ for GPWP with t^* and v^* the admission and discharge times of the previous hospitalization, respectively. We first simulated situations where the analyst considered the correct stratification, with s in the hazard functions to be $1,2,3$ according to the total number of previous admissions as 0, 1 or larger than 1. Three scenarios were considered to simulate situations when the covariate effect is taken as it is (i) time-independent ($\beta_{1s}=0$), (ii) time-dependent with correct functional form, $\beta_{0s}+\beta_{1s}\log(t)$, and (iii) time-dependent with an incorrect form, $\beta_{0s}+\beta_{1s}\sqrt{t}$. Note that, the simulated data were truly from the GPWP model in scenario (ii). In each of the scenarios, we evaluated the baseline estimators and the coefficient estimators and their variance estimators without/with adjustment for hospital duration. The corresponding estimates are referred to as (i) "PWP/GPWP", (ii) "PWP$_T$/GPWP$_T$", and (iii) "PWP$_{T*}$/GPWP$_{T*}$", respectively. We repeated the above procedures with an "incorrect" stratification, having s in the intensity functions to be $1,2$ according to the total number of previous admissions as 0, or larger than 0. The corresponding estimates are denoted by "PWP$_B$/GPWP$_B$", "PWP$_{TB}$/GPWP$_{TB}$", and "PWP$_{T*B}$/GPWP$_{T*B}$" in the following.

(b). *PWP2/GPWP2 Procedures:* all the PWP/GPWP procedures were repeated with the intensity functions replaced with $\lambda_{0s}(t)\exp\{(\beta_0+\beta_1 g(t))z_i\}$. The corresponding estimators were evaluated, which are referred to as the same as for the PWP/GPWP estimates given above but with "PWP/GPWP" substituted by "PWP2/GPWP2" in the notation.

(c). *AG/GAG Procedures*: all the PWP/GPWP procedures were repeated with the intensity functions replaced with $\lambda_0(t)\exp\{(\beta_0+\beta_1 g(t))z_i\}$, which totally ignored the stratification. The corresponding estimators of the model parame-

Table 5 Estimates of the regression parameters and their standard error estimates based on 200 simulation replications ($(\beta_{0s}, \beta_{1s}) = (0.8, 0.00), (0.1, 0.01), (0.2, 0.02)$ for $s = 1, 2, 3$; $D_A = 30$ and $D_B = 5$

Estimates	with risk set not-adjusted for event duration ($Y^V(t) = 1$)											
	AG	PWP			PWP2	PWP_{T*}			$PWP2_{T*}$	PWP_{T*B}		$PWP2_{T*B}$
		s=1	s=2	s=3		s=1	s=2	s=3		s=1	s=2	
$\tilde{\beta}_0$[a]	−.123	.669	−.118	−.284	−.083	.808	−1.244	−.969	−.366	.808	−.684	−.351
$\tilde{se}(\hat{\beta}_0)$[b]	.044	.075	.093	.027	.025	.308	.432	.271	.184	.308	.220	.188
$\bar{se}(\hat{\beta}_0)$[c]	.051	.123	.121	.063	.051	.301	.441	.356	.160	.301	.241	.159
$\tilde{\beta}_1$[d]	0	0	0	0	0	.000	.148	.059	.088	.000	.039	.085
$\tilde{se}(\hat{\beta}_1)$	–	–	–	–	–	.065	.057	.026	.018	.065	.021	.019
$\bar{se}(\hat{\beta}_1)$	–	–	–	–	–	.063	.057	.034	.017	.063	.024	.017
	with risk set adjusted for event duration											
	GAG	GPWP			GPWP2	$GPWP_{T*}$			$GPWP2_{T*}$	$GPWP_{T*B}$		$GPWP2_{T*B}$
		s=1	s=2	s=3		s=1	s=2	s=3		s=1	s=2	
$\tilde{\beta}_0$	.349	.669	.177	.361	.381	.808	.123	.275	.687	.808	.079	.601
$\tilde{se}(\hat{\beta}_0)$	.060	.075	.089	.041	.034	.308	.428	.358	.154	.308	.270	.163
$\bar{se}(\hat{\beta}_0)$	.051	.123	.121	.063	.051	.301	.452	.365	.159	.301	.243	.157
$\tilde{\beta}_1$	0	0	0	0	0	.000	.008	.011	.028	.000	.021	.032
$\tilde{se}(\hat{\beta}_1)$	–	–	–	–	–	.065	.056	.034	.017	.065	.026	.017
$\bar{se}(\hat{\beta}_1)$	–	–	–	–	–	.063	.059	.034	.017	.063	.024	.017

[a] Sample means of β_0 estimators: $\tilde{\beta}_0 = \sum_{j=1}^{200} \hat{\beta}_{0,j} \Big/ 200$ with $\hat{\beta}_{0,j}$ the j-th evaluation of $\hat{\beta}_0$

[b] Sample standard deviations of β_0/β_1 estimators: $\tilde{se}(\hat{\beta}_0) = \left\{\sum_{j=1}^{200} (\hat{\beta}_{0,j} - \tilde{\beta}_0)^2 \Big/ 200\right\}^{1/2}$

[c] Sample means of standard error estimators for $\hat{\beta}_0$ or $\hat{\beta}_1$: $\bar{se}(\hat{\beta}_0) = \sum_{j=1}^{200} \hat{se}(\hat{\beta}_0)_j \Big/ 200$ with $\hat{se}(\hat{\beta}_0)_j$ the j-th evaluation of $\hat{se}(\hat{\beta}_0)$, standard error estimators of $\hat{\beta}_0$

[d] Sample means of β_1 estimators: $\tilde{\beta}_1 = \sum_{j=1}^{200} \hat{\beta}_{1,j} \Big/ 200$ with $\hat{\beta}_{1,j}$ the j-th evaluation of $\hat{\beta}_1$

ters and their variance estimators were evaluated, which are referred to in the following as the same as the PWP/GPWP estimates but with "PWP/GPWP" in the notation replaced by "AG/GAG".

Our primary program was written in C. The *runif* function in Splus was used to generate uniform-distributed random variables in the simulation. Iterations were terminated when the change of the current estimate from the previous estimate was below 10^{-5}. Sets of values for the parameters (β_{0s}, β_{1s}) for $s = 1, 2, 3$ were chosen to generate the event processes, to examine performance of the above procedures in various situations. The following comments and discussions are based on 200 repetitions in each of the simulation settings, with a focus on estimation of the regression parameters. All the simulation outcomes are available upon request.

Table 5 illustrates the simulation outcomes with the simulated scenario of a relative large event duration in Group A, $D_A = 30$, and a large first event risk and small

risk rates of the following events. In this scenario, the approaches with stratification gave very close estimates for the first stratum (about the first event) regardless of risk set definition, for event duration influences only the inferences about the part of the event process with event duration. The table shows large differences in results for all the models, with or without stratification, between approaches with and without adjustment for event duration. Furthermore, the outcomes of all the approaches with adjustment for the risk sets show the higher event rate in Group A than in Group B for all strata. This indicates robustness of the proposed approach against misspecifying covariate effects and/or model stratification. The results associated with the approaches without the adjustment and without stratification suggested Group A had a significantly lower event rate. This illustrates clearly the bias caused by ignoring the event duration in the analysis: subjects with a long event duration were categorized as ones with a low event rate in the analyses without the adjustment.

The following summarizes our observations in general from the simulation results.

(a). In all the simulated scenarios, the sample variances of the estimators for the regression parameters and the sample means of the variance estimators of the regression estimators were close sufficiently for practical use, regardless of the implemented procedures.

(b). When used to identify risk factors, the outcomes of the approaches without adjustment for the event duration were similar to those of the approaches with the adjustment when the event duration was short on average. One may think the outcomes are also similar in the situations where the first event rate is much larger than the rates of the following events, even when the event duration is long. However, with long event duration, the approaches without adjustment for the duration led to obviously biased inferences in the simulation. Table 5 exemplifies this.

(c). Regardless of the sets of values chosen for (β_{0s}, β_{1s}), the sample means of the regression estimates with GPWP$_T$ (i.e., under the correct model specification) were very close to the true values. This confirms the consistency of the MPLE for the regression parameters. The GPWP$_{T*}$ estimates, with the misspecified $g(t)$ in the time-dependent effect to the covariate, were similar to the GPWP$_T$ estimates. When used to identify risk factors and to classify events, all the other GPWP-type estimates, the GPWP estimates with time-independent effects and the GPWP$_{TB}$ and GPWP$_{T*B}$ estimates with misspecified stratification, led to conclusions close to the ones based on the GPWP$_T$ estimates. This indicates robustness of the GPWP-type approach against misspecification of event stratification and misspecification of time-dependent covariate effect.

(d). When aiming at identifying risk factors only, the GPWP2 and GPWP2$_{TB}$ estimates were in close agreement with the GPWP estimates. So were the GAG estimates.

6 Discussion

This paper presents a generalized Cox regression model for the conditional intensity function of a counting process, which accommodates non-negligible event duration using adjusted risk sets. Various specifications of the model are illustrated by analyses

of the hospitalization data that motivated this research. The usefulness of the proposed approach is demonstrated via a simulation study. The proposed approach provides a tool for assessing hospitalization, with particular attention to identifying risk factors.

We assume that the response process is observed subject to a noninformative right-censoring. The data collection of the hospitalization project was subject to the subjects' provincial medical insurance coverage, and some subjects were unregistered during a portion of the study time. This leads to some extra missingness in hospitalization records, and suggests another data structure to consider. Methods for analyses of event history data with interval censoring, such as those reviewed in Sun (2006), may be adapted to analyze the data. On the other hand, in many practical situations, the observation of the response process is terminated by informative dropouts or failure events such as death. Approaches presented in Wang et al. (2001), Ghosh and Lin (2003), and Miloslavsky et al. (2004), for example, may be extended to analyze recurrent events with non-negligible duration in the presence of dependent censoring.

Several other further investigations would also be worthwhile. One of theoretical and practical interest is another formulation of hospital utilization, using the length of hospitalization over time as the primary measure of interest. In some situations, the total length of hospitalization may be more informative than the count of hospital admissions. Another is, as mentioned in Sect. 2, to investigate the modeling in situations with other time scales, such as time to an event since the end of the previous event. A third is to consider non-Cox type of regression models for the recurrent event data, such as the threshold regression models discussed in Lee and Whitmore (2006).

Acknowledgments The research was partially supported by the grants of the authors from the Canadian Institutes of Health Research (CIHR), the National Institutes of Health of USA (NIH) and the Natural Sciences and Engineering Research Council of Canada (NSERC). The CAYACS Research Program is funded by the Canadian Cancer Society. Cooperation in allowing access, use and linkage of the data to support this program is gratefully acknowledged from the following: BC Cancer Registry, BC Cancer Agency, BC Children's Hospital, the BC Ministry of Health Services, BC Vital Statistics, and the Centre for Health Services and Policy Research at the University of British Columbia. The authors thank an AE and an anonymous reviewer for their helpful comments and suggestions, which led to a much improved version of the paper.

Appendix: Asymptotics of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$

In the following, we assume all the necessary conditions are satisfied. Define for $t \geq 0$

$$M(t) = \int_0^t Y^C(u)dN(u) - \int_0^t Y^C(u)\lambda(u|\,\mathcal{H}(u))du$$

and

$$\Lambda(t) = \int_0^t Y^C(u)Y^V(u)\sum_{s=1}^{S}\lambda_{0s}(u)\exp\{\beta(u;\beta_s)'Z(u) + \alpha(u;\alpha_s)'W(u)\}\Big|_{s=s\{\mathcal{H}(u)\}}du.$$

Note that, under the noninformative censoring assumption, $\Lambda(\cdot)$ is the compensator of the counting process $\left\{\int_0^t Y^C(u)dN(u) : t \geq 0\right\}$, and thus $\{M(t) : t \geq 0\}$ is a compensated counting process, a martingale, with its predictable variation the same as the compensator $\Lambda(\cdot)$. Let $\{M_i(t) : t \geq 0\}$ for $i = 1, \ldots, n$ be n iid realizations of $\{M(t) : t \geq 0\}$. Denote, for $s = 1, \ldots, S$,

$$B_s^{(r)}(t; \beta_s, \alpha_s) = \sum_{j \in \mathcal{R}(t;s)} \exp\{\beta(t; \beta_s)' Z_j(t) + \alpha(t; \alpha_s)' W_j(t)\} \left\{ \frac{\partial \beta(t; \beta_s)' Z_j(t)}{\partial \boldsymbol{\beta}} \right\}^{\otimes r}$$

and

$$A_s^{(r)}(t; \beta_s, \alpha_s) = \sum_{j \in \mathcal{R}(t;s)} \exp\{\beta(t; \beta_s)' Z_j(t) + \alpha(t; \alpha_s)' W_j(t)\} \left\{ \frac{\partial \alpha(t; \alpha_s)' W_j(t)}{\partial \boldsymbol{\alpha}} \right\}^{\otimes r}$$

with $r = 0, 1, 2$, where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa'$, and

$$BA_s(t; \beta_s, \alpha_s) = \sum_{j \in \mathcal{R}(t;s)} \exp\{\beta(t; \beta_s)' Z_j(t) + \alpha(t; \alpha_s)' W_j(t)\} \left\{ \frac{\partial \beta(t; \beta_s)' Z_j(t)}{\partial \boldsymbol{\beta}} \right\} \left\{ \frac{\partial \alpha(t; \alpha_s)' W_j(t)}{\partial \boldsymbol{\alpha}} \right\}'.$$

The partial score function of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $\partial \log\{PL(\boldsymbol{\beta}, \boldsymbol{\alpha})\} \big/ \partial(\boldsymbol{\beta}, \boldsymbol{\alpha})$, is

$$U(\boldsymbol{\beta}, \boldsymbol{\alpha}; \infty) = \left(U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}; \infty)', U_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}; \infty)' \right)'$$

with

$$\begin{aligned} U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}; t) &= \frac{\partial C(\boldsymbol{\beta}, \boldsymbol{\alpha}; t)}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \int_0^t Y_i^C(u) \sum_{s=1}^{S} \left[\frac{\partial \beta(u; \beta_s)' Z_i(u)}{\partial \boldsymbol{\beta}} - \frac{B_s^{(1)}(u; \beta_s, \alpha_s)}{B_s^{(0)}(u; \beta_s, \alpha_s)} \right] \Bigg|_{s = s\{\mathcal{H}_i(u)\}} dN_i(u) \end{aligned} \tag{8}$$

and

$$\begin{aligned} U_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}; t) &= \frac{\partial C(\boldsymbol{\beta}, \boldsymbol{\alpha}; t)}{\partial \boldsymbol{\alpha}} \\ &= \sum_{i=1}^{n} \int_0^t Y_i^C(u) \sum_{s=1}^{S} \left[\frac{\partial \alpha(u; \alpha_s)' W_i(u)}{\partial \boldsymbol{\alpha}} - \frac{A_s^{(1)}(u; \beta_s, \alpha_s)}{A_s^{(0)}(u; \beta_s, \alpha_s)} \right] \Bigg|_{s = s\{\mathcal{H}_i(u)\}} dN_i(u) \end{aligned} \tag{9}$$

for $C(\boldsymbol{\beta}, \boldsymbol{\alpha}; t)$ defined in (5).

With some regularity conditions and arguments similar to those in Andersen and Gill (1982), we can establish the consistency of the PMLE $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ by noting $n^{-1}\left[C(\mathbf{b}, \mathbf{a}; \infty) - C(\boldsymbol{\beta}, \boldsymbol{\alpha}; \infty)\right]$ with fixed $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is a concave function with large n with a unique maximum at $(\mathbf{b}, \mathbf{a}) = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$. The following facts along with the central limit theorems for local martingales lead to the asymptotic normality of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$, that is $n^{1/2}\left\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) - (\boldsymbol{\beta}, \boldsymbol{\alpha})\right\} \to N(0, \Sigma(\boldsymbol{\beta}, \boldsymbol{\alpha}))$, where $\Sigma(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is defined below. Firstly,

$$\frac{1}{\sqrt{n}} U(\boldsymbol{\beta}, \boldsymbol{\alpha}; \infty) = \frac{1}{n} \mathcal{I}(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*) \sqrt{n} \left\{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})' - (\boldsymbol{\beta}, \boldsymbol{\alpha})'\right\}$$

by the definition of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$, where $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*)$ is on the line segment between $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ and $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, and $\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is the matrix

$$-\frac{\partial^2 \log\{PL(\boldsymbol{\beta}, \boldsymbol{\alpha})\}}{\partial(\boldsymbol{\beta}, \boldsymbol{\alpha})^2} = \begin{pmatrix} \mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) & \mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha})' & \mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \end{pmatrix}$$

with $\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\mathcal{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ being

$$\sum_{i=1}^{n} \int_0^{\infty} Y_i^C(u) \sum_{s=1}^{S} \left[\frac{B_s^{(2)}(u; \beta_s, \alpha_s)}{B_s^{(0)}(u; \beta_s, \alpha_s)} - \left\{ \frac{B_s^{(1)}(u; \beta_s, \alpha_s)}{B_s^{(0)}(u; \beta_s, \alpha_s)} \right\}^{\otimes 2} \right] \Bigg|_{s=s\{\mathcal{H}_i(u)\}} dN_i(u),$$

$$\sum_{i=1}^{n} \int_0^{\infty} Y_i^C(u) \sum_{s=1}^{S} \left[\frac{BA_s(u; \beta_s, \alpha_s)}{B_s^{(0)}(u; \beta_s, \alpha_s)} - \left\{ \frac{B_s^{(1)}(u; \beta_s, \alpha_s)}{B_s^{(0)}(u; \beta_s, \alpha_s)} \right\} \left\{ \frac{A_s^{(1)}(u; \beta_s, \alpha_s)}{A_s^{(0)}(u; \beta_s, \alpha_s)} \right\}' \right] \Bigg|_{s=s\{\mathcal{H}_i(u)\}} dN_i(u)$$

and

$$\sum_{i=1}^{n} \int_0^{\infty} Y_i^C(u) \sum_{s=1}^{S} \left[\frac{A_s^{(2)}(u; \beta_s, \alpha_s)}{A_s^{(0)}(u; \beta_s, \alpha_s)} - \left\{ \frac{A_s^{(2)}(u; \beta_s, \alpha_s)}{A_s^{(0)}(u; \beta_s, \alpha_s)} \right\}^{\otimes 2} \right] \Bigg|_{s=s\{\mathcal{H}_i(u)\}} dN_i(u),$$

respectively. At the true value $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}; t)$ and $U_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}; t)$ are the same as the ones having $N_i(\cdot)$ in (8) and (9) replaced with the martingales $M_i(\cdot)$; $n^{-1/2}U(\boldsymbol{\beta}, \boldsymbol{\alpha}; t)$ converges weakly to a Gaussian process with mean zero. Plus $n^{-1}\mathcal{I}(\mathbf{b}, \mathbf{a})$, in the neighborhood of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ converges in probability to the covariance of $n^{-1/2}U(\boldsymbol{\beta}, \boldsymbol{\alpha}; \infty)$, denoted by $\Sigma(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1}$. It is thus easy to see that the variance of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ can be estimated consistently by the inverse of $\mathcal{I}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}; \infty)$.

References

Andersen PK, Borgan O, Gill RD, Keiding N (1992) Statistical models based on counting processes. Springer, New York

Andersen PK, Gill RD (1982) Cox's regression model for counting processes: a large sample study. Anna Stat 10:1100–1120

Cai J, Schaubel DE (2004) Marginal means/rates models for multiple type recurrent event data. Lifetime Data Anal 10:121–138

Cook RJ, Lawless JF (2007) The statistical analysis of recurrent events. Springer, New York

Cox DR (1975) Partial likelihood. Biometrika 62:269–276

Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley, New York

Ghosh D, Lin DY (2003) Semiparametric analysis of recurrent events data in the presence of dependent censoring. Biometrics 59:877–885

Grambsch P, Therneau T (1994) Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81:515–526

Kalbfleisch JD, Prentice RL (1980) The statistical analysis of failure time data. Wiley, New York

Kelly PJ, Lim LL-Y (2000) Survival analysis for recurrent event data: an application to childhood infectious diseases. Stat Med 19:12–33

Lawless JF, Nadeau C (1995) Some simple robust methods for the analysis of recurrent events. Technometrics 37:158–168

Lee MLT, Whitmore GA (2006) Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Stat Sci 21:501–513

Lin DY, Wei LJ, Yang I, Ying Z (2000) Semiparametric regression for the mean and rate functions of recurrent events. J R Stats Soc, Series B 62:711–730

MaCarthur AC, Spinelli JJ, Rogers P, Goddard KJ, Abanto Z, McBride ML (2007a) Mortality among 5-year survivors of cancer diagnosed during childhood or adolescence in British Columbia, Canada. Pediatr Blood Cancer 48(4):460–467

MaCarthur AC, Spinelli JJ, Rogers P, Goddard KJ, Abanto Z, McBride ML (2007b) Risk of a second malignant neoplasm among 5-year survivors of cancer in childhood and adolescence in British Columbia, Canada. Pediatr Blood Cancer 48(4):453–459

Miloslavsky S, Keles S, Laan MJvan der , Butler S (2004) Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. J R Stats Soc, Series B 66:239–257

Oakes D (1992) Frailty models for multiple event times. In: Klein J, Goel P (eds) Survival analysis: state of the art. Kluwer, Dordrecht pp 371–379

Pepe MS, Cai J (1993) Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. J Am Stats Assoc 88:811–820

Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. Biometrika 68:373–379

Sun J (2006) The statistical analysis of interval-censored failure time data. Springer, New York

Therneau T, Grambsch P (2000) Modeling survival data: extending the cox model. Springer, New York

Twisk JWR, Smidt N, Vente Wde (2005) Applied analysis of recurrent events: a practical overview. J Epidemiol Community Health 59:706–710

Wang MC, Qin J, Chiang CT (2001) Analyzing recurrent event data with informative censoring. J Am Stats Assoc 96:1057–1065

Ying SC (2006) Generalized longitudinal data analysis, with application to evaluating hospital utilization based on administrative database. Master's Thesis, Department of Statistics and Actuarial Science, Simon Fraser University