



Additive Frailty Model for Recurrent Events Data with Application to Cancer Data

Authors: FELIPE RODRIGUES DA SILVA 
– Federal University of Piauí
Brazil
feliperodrigues@ufpi.edu.br

VERA TOMAZELLA 
– Federal University of São Carlos
Brazil
vera@ufscar.br

CLEIDE MAYRA MENEZES LIMA 
– Federal University of Piauí
Brazil
cleide.mayra@ufpi.edu.br

SARALEES NADARAJAH  
– Department of Mathematics, University of Manchester
UK
mbbssn2@manchester.ac.uk

Received: Month 0000

Revised: Month 0000

Accepted: Month 0000

Abstract:


- Both the additive and proportional intensity models provide two principal frameworks for studying the association between risk factors and disease recurrences. When the events of interest are not terminal and can occur more than once for the same individual, we have the so-called recurrent events, these types of data appear in areas such as biomedicine, criminology and industrial reliability. In this paper, we study an additive intensity model with gamma frailty and propose an estimator for the individual frailties of patients. An advantage of the studied model is the possibility to jointly consider the heterogeneity among patients and to evaluate the dependence within recurrent events captured by the frailty variable. Such a distribution has theoretical arguments to model medical data and has been shown empirically to be a good option. We consider likelihood-based methods to estimate the model parameters, and also investigate large-sample properties of the estimators. In order to illustrate our methodology, we consider two data sets including one from an experimental animal carcinogenesis study.

Keywords:

- *Additive hazard rate models; Cancer data; Frailty models; Recurrent events.*

AMS Subject Classification:

- Primary 62E15.

 Corresponding author

1. INTRODUCTION

The methodology for survival data is designed to determine variables affecting the hazard rate function and to obtain estimates of these functions for each individual. Studies involve following units (individuals) until the occurrence of some event of interest, for example, the fault (death) of the unit. The model proposed by [Cox \(1972\)](#) is one of the best known and used in analysis of survival data, however this model assumes that risks are proportional, an assumption that is often unreasonable. To try to solve this limitation, the additive hazard rate model was initially proposed by [Aalen \(1980\)](#). In the additive model the effect of covariates is additive in the hazard rate function and not multiplicative as in the Cox model.

Another characteristic of survival data is that some events of interest are not terminal. Existing events sometimes occur more than once for the same individual, producing recurrent events. Lifetime data where more than one event is observed on each subject arises in areas such as biomedical studies, criminology, demography, manufacturing and industrial reliability. For example, an offender may be convicted several times; several tumors may be observed for an individual; recurrent pneumonia episodes arise in patients with human immunodeficiency syndrome; a piece of equipment may experience repeated failures or warranty claims.

A reasonable assumption for recurrent event data is that times are dependent and that risks are not proportional. An alternative to model these data types is the additive hazard rate model. To measure the dependence between the recurrence times, we will use the frailty term that can be inserted in an additive or multiplicative way. In this paper, we will use the frailty term inserted additively. A parametric approach to additive models with frailty was presented by [Tomazella \(2003\)](#). A Bayesian inference procedure for additive models with frailty was considered in [Tomazella et al. \(2006\)](#).

Several methodologies have been proposed to analyze the problem of recurrent events. [Lawless and Nadeau \(1995\)](#) applied the Poisson process to develop models that focus on the expected number of events occurring in a determined time interval. The development of statistical models based on counting process data was originally introduced by [Aalen \(1978\)](#). There is an extensive literature about point process models (see, for example, [Cox and Isham \(1980\)](#)).

Frailty models are characterized by the inclusion of a random effect representing information that has not been observed or cannot be measured such as environmental or genetic factors. Also, there may be information that, for some reason, was not considered in the planning process of the study. A way to incorporate this random effect, called frailty variable, is to include it in the baseline hazard rate for controlling the unobservable heterogeneity of the units under analysis. The frailty can be included into the model additively or multiplicatively to assess the heterogeneity among units by means of the hazard rate [Tomazella \(2003\)](#). In survival analysis, these units can be patients possessing different frailties, patients who are “frail” or “prone” may have the disease earlier than those who are less frail.

Although early techniques developed for handling recurrent event data suppose independence among the recurrent event times (“lifetimes”) related to the same subject, an

assumption of dependence among these lifetimes is reasonable. The dependence can be taken into account by incorporating a random effect (so-called frailty) in modelling (Clayton, 1978). The frailty term generates dependence among the lifetimes of each subject, which are assumed to be conditionally independent given the frailty.

Advances in medical treatments increase researchers' interest in considering survival models for cancer data. The occurrence of an event of interest (for example, the patient's death) can be due to one or several competing causes. There are also some unobserved external factors that can affect the onset of a tumor. The interest may be in understanding and characterizing the event, illustrating the process for the individual subject, or the factors may focus on time-based treatment comparisons for each distinct event, the number of events, the type of events, and the interdependence between events. The idea is to explain the nature of variation between subjects in terms of fixed covariates of treatments or other factors such as unobservable factors.

This paper is motivated by two real medical data sets. The first is a medical data set corresponding to animal carcinogenicity described by (Gail et al., 1980). The second data set refers to readmission times after surgery in patients diagnosed with colorectal cancer (González et al., 2005).

In this paper, we consider a class of parametric regression models which are extensions of Aalen's model with structure of recurrent event data. The research goals of analyzing such data often include characterizing the rate of different event types, estimating the treatment effects on each event process, and understanding the correlation structure among different event types. The proposed model assumes that the intensity given the covariates and a random frailty has an additive hazard rate form. The frailty in the proposed model is assumed to follow a gamma distribution. We also include a baseline hazard rate assumed to follow the exponential distribution or a Weibull distribution. We employ Laplace transform for finding the survival function unconditional on the individual frailty. We use the maximum likelihood (ML) method for estimating the corresponding parameters. We evaluate the performance of the ML estimators via a Monte Carlo simulation method.

This paper is organized as follows. In Section 2, we present the additive frailty model with a gamma frailty distribution. Inference methods based on the likelihood function are also presented. In Section 3, we consider a simulation study in different scenarios. We evaluate numerically the asymptotic properties of the estimators. In Section 4, we apply these procedures to two real medical data sets. Some final remarks are made in Section 5.

2. METHODOLOGY

2.1. ADDITIVE FRAILTY MODEL FOR RECURRENT EVENT DATA

Rocha (1995) proposed a linear frailty model assuming the intensity function at the instant $t_{i,j}$, where individual i 's recurrent failure times are of the form $0 \leq t_{i,1} < t_{i,2} < \dots < t_{i,m_i} \leq T_i$, where T_i is the total lifetime and $t_{i,j}$ is the observed lifetime of individual

i in the j th failure, where $i = 1, 2, 3, \dots, n$ the index that identifies the individual and $j = 1, 2, 3, \dots, m_i$ the index representing the recurrent event of individual i ($j = 0$ is the initial event).

From Aalen's linear regression model (Aalen, 1980), we may define the additive intensity model by

$$(2.1) \quad \lambda_i(t) = \lambda_0(t) + g(\mathbf{x}_i, \boldsymbol{\beta}),$$

where $\lambda_0(\cdot)$ is the baseline intensity function, $\boldsymbol{\beta}$ is the vector of regression coefficients, and \mathbf{x}_i is covariate vector of the i th individual, $i = 1, 2, \dots, n$. Note that (2.1) is an alternative to the established Cox regression model (Cox, 1972) that is defined by an intensity function given by $\lambda_i(t) = \lambda_0(t) g(\mathbf{x}_i, \boldsymbol{\beta})$.

A simple way of composing the intensity model with a frailty term is to introduce an additive random effect in (2.1). Hence, following Silva (2001), the additive homogeneous Poisson process with a frailty term is

$$(2.2) \quad \lambda_i(t|v_i, \mathbf{x}_i) = \lambda_0(t) + \mathbf{x}_i' \boldsymbol{\beta} + v_i,$$

where $\lambda_0(t)$, $\boldsymbol{\beta}$ and \mathbf{x}_i are defined in (2.1) and v_i is the frailty variable with a known distribution function. The frailty term in (2.2) represents the information that may not be observed such as environment and genetics factors or information that, by some reason, was not considered at the planning. These models are doubly additive, since both the observed covariates \mathbf{x}_i and the frailty v_i are introduced in (2.2) additively.

The survival function, conditioned to the frailty variable v_i and to the effects of the observed factors, obtained by the relation with the accumulated hazard rate function is

$$\begin{aligned} S(t|v_i, \mathbf{x}_i) &= \exp \left\{ - \int_0^{T_i} \lambda(u|v_i, \mathbf{x}_i) du \right\} \\ &= \exp \left[- \int_0^{T_i} \{ \lambda_0(u) + \mathbf{x}_i' \boldsymbol{\beta} + v_i \} du \right] \\ &= \exp \{ - \mathbf{x}_i' \boldsymbol{\beta} T_i - v_i T_i - \Lambda(T_i) \}. \end{aligned}$$

The hazard rate function and the unconditional survival function can be obtained through the Laplace transform; see Hougaard (1984). When looking for distributions for the frailty variable, distributions having explicit Laplace transformations are a natural choice. This facilitates the use of traditional ML methods for parameter estimation.

Since survival times are absolutely continuous random variables here, the unconditional survival function for the i th individual associated with (2.2) is

$$S(t) = \int_0^\infty S(t|v_i, \mathbf{x}_i) f(v_i) dv_i.$$

Hence,

$$\begin{aligned} S(t) &= \int_0^\infty \exp \{ - \Lambda_0(T_i) - \mathbf{x}_i' \boldsymbol{\beta} T_i - v T_i \} f(v) dv \\ &= \int_0^\infty \exp \{ - \Lambda_0(T_i) - \mathbf{x}_i' \boldsymbol{\beta} T_i \} \exp(-v T_i) f(v) dv \\ (2.3) \quad &= \exp \{ - \Lambda_0(T_i) - \mathbf{x}_i' \boldsymbol{\beta} T_i \} L(T_i), \end{aligned}$$

where $\Lambda_0(t)$ is the cumulative baseline hazard rate function and $L(t)$ is the Laplace transform of v . The unconditional intensity function of (2.2) for the i th individual is

$$(2.4) \quad \lambda_i(t) = \lambda_0(t) + \mathbf{x}'_i \boldsymbol{\beta} - \frac{L'(t)}{L(t)},$$

where $L'(t)$ is the derivative of the Laplace transform $L(t)$ of the frailty distribution.

For a fully parametric analysis, we assume a gamma distribution for the frailty variable. Our choice was essentially made by mathematical convenience. Interested readers can refer to Hougaard (2000) for a comprehensive discussion about the choice of the frailty term distribution.

2.1.1. ADDITIVE MODEL WITH GAMMA FRAILITY DISTRIBUTION

Due to the way the frailty term acts in the hazard rate function, the candidates for the frailty distribution are supposed to be non-negative, usually continuous and not time-dependent, such as the gamma, inverse Gaussian or log-normal distributions (Hougaard, 2000). The gamma distribution has been widely applied as a frailty distribution. From a computational and analytical point of view, the gamma distribution fits very well as a frailty distribution to failure data. Closed-form expressions for the unconditional survival, cumulative density, and hazard rate functions are easy to derive for the gamma distribution, due to the simplicity of its Laplace transform. This is also the reason why the gamma distribution has been used in most applications published to date.

Thus, we consider the frailty variable v to have the gamma $(\frac{1}{\alpha}, \frac{1}{\alpha})$ distribution with the probability density function given by

$$f(v) = \frac{\left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha}}}{\Gamma\left(\frac{1}{\alpha}\right)} v^{\frac{1}{\alpha}-1} \exp\left(-\frac{v}{\alpha}\right),$$

where $v > 0$ and α quantifies the amount of heterogeneity among subjects. The Laplace transform $L(t)$ for the distribution is

$$(2.5) \quad L(t) = \int_0^{\infty} e^{-tv} f(v) dv = (1 + \alpha t)^{-1/\alpha}.$$

The first derivative of the Laplace transform is

$$(2.6) \quad L'(t) = -(1 + \alpha t)^{-\frac{1}{\alpha}-1}.$$

Thus, replacing (2.5) with (2.6), we can rewrite (2.4) as

$$(2.7) \quad \lambda_i(t) = \lambda_0(t) + \mathbf{x}'_i \boldsymbol{\beta} + (1 + \alpha t)^{-1}.$$

The unconditional survival function (2.3) can be expressed by

$$(2.8) \quad S(t) = \exp\{-\mathbf{x}'_i \boldsymbol{\beta} T_i - \Lambda_0(T_i)\} (1 + \alpha T_i)^{-1/\alpha}.$$

Different parametric forms can be taken for the base hazard rate function $\lambda_0(t)$. In this paper, we consider the Weibull(μ, γ) and exponential(μ) distributions. The probability density function of the Weibull(μ, γ) distribution is

$$f(t) = \gamma \mu t^{\gamma-1} \exp(-\mu t^\gamma),$$

where $\lambda(t) = \gamma \mu t^{\gamma-1}$ and $\Lambda(t) = \mu t^\gamma$. The exponential(μ) distribution is the particular case for $\gamma = 1$ with $\lambda(t) = \mu$ constant and $\Lambda(t) = \mu t$. The unconditional hazard rate and survival functions for the Weibull gamma additive frailty model are

$$(2.9) \quad \lambda_i(t) = (1 + \alpha t)^{-1} + \gamma \mu t^{\gamma-1} + \mathbf{x}'_i \boldsymbol{\beta}$$

and

$$(2.10) \quad S(t) = (1 + \alpha T_i)^{-1/\alpha} \exp(-\mu T_i^\gamma - \mathbf{x}'_i \boldsymbol{\beta} T_i),$$

respectively.

2.2. NONPARAMETRIC ESTIMATOR FOR MARGINAL SURVIVAL FUNCTION

The nonparametric estimator Wang and Chang (1999) for recurrent events is used to estimate the marginal survival function in the presence of correlation between the times of occurrence. Consider the censored recurrence times $(t_{i,1}, t_{i,2}, \dots, t_{i,m_i}^+)$ and define the observed recurrence times by

$$y_{i,j} = \begin{cases} t_{i,j}, & \text{for } j = 1, \dots, m_i - 1, \\ t_{i,m_i}^+, & \text{for } j = m_i. \end{cases}$$

Let $(y_1^*, y_2^*, \dots, y_k^*)$ be ordered and distinct uncensored times. The estimator for censored data using risk estimation techniques assumes the expression of limit product defined by

$$\widehat{S}(t) = \prod_{\{y_i^* \leq t\}} \left\{ 1 - \frac{d^*(y_i^*)}{R^*(y_i^*)} \right\},$$

where $R^*(t)$ and $d^*(t)$ are given by

$$R^*(t) = \sum_{i=1}^n \left\{ \frac{a_i}{m_i^*} \sum_{j=1}^{m_i^*} I(y_{i,j} \geq t) \right\}$$

and

$$d^*(t) = \sum_{i=1}^n \left\{ \frac{a_i I(m_i \geq 2)}{m_i^*} \sum_{j=1}^{m_i^*} I(y_{i,j} = t) \right\},$$

respectively, where $a_i = a(C_i)$ with $a(\cdot)$ representing a positive valued function subject to the constraint $E(a_i) < \infty$ and times of censorship C_i .

2.2.1. INFERENCE

We present here the estimation procedures related to the gamma additive frailty model. The data on the i th individual consists of the total number, m_i , of the events observed over the time period $(0, T_i]$ and the ordered epoch of the m_i events, $0 \leq t_{i,1} < t_{i,2} < \dots < t_{i,m_i} \leq T_i$. Additionally, we assume that $\delta_{i,j} = 1$ if $t_{i,j} \leq T_i$ and $\delta_{i,j} = 0$ if $t_{i,j} > T_i$.

Note that the conditional cumulative intensity function $\Lambda_i(t|v_i, \mathbf{x}_i) = \int_0^t \lambda_i(u|v_i, \mathbf{x}_i) du$ for the i th individual is obtained by integrating Equation 2.7 from the initial time to the total time of the study (T_i).

The unconditional likelihood function for the Weibull gamma additive frailty model, considering the unconditional intensity (2.9) and the survival function (2.10) for a sample of n independent individuals with m_i events observed by time $t_{i,j}$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m_i$, is given by

$$(2.11) \quad L(\alpha, \mu, \gamma, \boldsymbol{\beta} | T_i, \mathbf{x}_i; \tau_i) = \prod_{i=1}^n \prod_{j=1}^{m_i} \left\{ \gamma \mu t_{i,j}^{\gamma-1} + \mathbf{x}'_i \boldsymbol{\beta} + (\alpha t_{i,j} + 1)^{-1} \right\}^{\delta_{i,j}} \cdot \prod_{i=1}^n \exp(-\mu T_i^\gamma - \mathbf{x}'_i \boldsymbol{\beta} T_i) (1 + \alpha T_i)^{-1/\alpha}.$$

The corresponding log-likelihood function is given by

$$(2.12) \quad l(\alpha, \mu, \gamma, \boldsymbol{\beta} | T_i, \mathbf{x}_i; \tau_i) = \sum_{i=1}^n \sum_{j=1}^{m_i} \delta_{i,j} \log \left\{ \gamma \mu t_{i,j}^{\gamma-1} + \mathbf{x}'_i \boldsymbol{\beta} + (1 + \alpha t_{i,j})^{-1} \right\} + \sum_{i=1}^n \left(-\mu T_i^\gamma - \mathbf{x}'_i \boldsymbol{\beta} T_i \right) - \frac{1}{\alpha} \log(1 + \alpha T_i).$$

The log-likelihood function (2.12) can be maximized numerically to obtain the ML estimates. There are various routines available for numerical maximization. Here, we use the function `optim` of the R software (R Core Team, 2022) for the numerical maximization. We used the ‘BFGS’ method for maximization, for details see Fletcher and Reeves (1964).

In many cases, construction of confidence intervals is necessary to indicate the precision or accuracy of point estimates of the parameters. The confidence intervals of model parameters can be based on the asymptotic normality properties of the ML estimators. If $\widehat{\boldsymbol{\theta}}$ denotes the ML estimators of the parameter vector $\boldsymbol{\theta}$, then the distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ can be approximated by a multivariate normal distribution with mean zero and covariance matrix $I^{-1}(\widehat{\boldsymbol{\theta}})$, where $I(\widehat{\boldsymbol{\theta}})$ is referred to as the observed information matrix. Thus, a $100(1 - \alpha)\%$ asymptotic confidence interval for each parameter θ_i is

$$CI(\boldsymbol{\theta}, 100(1 - \alpha)\%) = \left(\widehat{\theta}_i - z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\theta_i)}, \widehat{\theta}_i + z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\theta_i)} \right),$$

where $\widehat{\text{Var}}(\theta_i)$ is the i th main diagonal element of $I^{-1}(\widehat{\boldsymbol{\theta}})$ and $z_{\alpha/2}$ is the $(1 - \alpha)\%$ quantile of the standard normal distribution.

In addition to parameter estimates, estimates of individual frailties are necessary. Using the idea developed by Munda et al. (2012), we propose the following estimates for individual

frailties

$$\widehat{v}_i = \frac{\int_0^\infty v_i^{d_i+1} \exp(-v_i T_i) f(v_i) dv_i}{\int_0^\infty v_i^{d_i} \exp(-v_i T_i) f(v_i) dv_i} = \frac{E \{V^{d_i+1} \exp(-VT_i)\}}{E \{V^{d_i} \exp(-VT_i)\}},$$

where

$$d_i = \sum_{j=1}^{m_i} \delta_{i,j}$$

is the number of recurrent events for individual i and T_i is the total lifetime of the study. By writing the expected values given in (2.13) in terms of derivatives of the Laplace transform, we have

$$E \{V^{d_i+1} \exp(-VT_i)\} = (-1)^{d_i+1} L^{(d_i+1)}(T_i)$$

and

$$E \{V^{d_i} \exp(-VT_i)\} = (-1)^{d_i} L^{(d_i)}(T_i),$$

that is,

$$\widehat{v}_i = \frac{-L^{(d_i+1)}(T_i)}{L^{(d_i)}(T_i)}.$$

Using the result in (Munda et al., 2012), we still have that for $V \sim \text{gamma}(1/\alpha, 1/\alpha)$,

$$L^{(d_i)}(T_i) = (-1)^{d_i} (1 + \alpha T_i)^{-d_i} \left\{ \prod_{k=0}^{d_i-1} (1 + k\alpha) \right\} (1 + \alpha T_i)^{-1/\alpha}$$

and

$$L^{(d_i+1)}(T_i) = (-1)^{d_i+1} (1 + \alpha T_i)^{-(d_i+1)} \left\{ \prod_{k=0}^{d_i} (1 + k\alpha) \right\} (1 + \alpha T_i)^{-1/\alpha}.$$

In this way, we can estimate individual frailty for recurrent event data from the gamma $(1/\alpha, 1/\alpha)$ additive frailty model as

$$(2.13) \quad \widehat{v}_i = (1 + \widehat{\alpha} T_i)^{-1} (1 + d_i \widehat{\alpha}).$$

In this paper, we consider that the truncation time T_i is the same for all individuals. If the individual i does not have recurrent times then $d_i = 0$, that is, we can estimate its individual frailty for gamma $(1/\alpha, 1/\alpha)$ frailty. It is given by $\widehat{v}_i = (1 + \widehat{\alpha} T_i)^{-1}$.

3. SIMULATION STUDY

In this study, times were generated from the additive model with gamma frailty distribution $(1/\alpha, 1/\alpha)$ and an exponential baseline hazard rate. The inversion method was used with $\mu = 0.03$, $\beta = -0.1$, $\alpha = 10, 50$, sample sizes $n = 100, 250, 400$ and censoring percentages 0%, 10% and 20%. Each sample was replicated 300 times. The Mean Squared Error

(MSE), the standard deviation (SD) of the estimates, the mean of the asymptotic standard errors (SE) and bias were computed for α , μ and β of (2.7). To assess the covariate effects on the hazard function and time effect, we divided the sample into two groups. Subjects in the control and treatment groups were assigned covariate values of 0 and 1, respectively. Censors were obtained as follows: we randomly generated the recurrence number of each individual in the sample in such a way to obtain 0%, 10% and 20% censorships in each generated sample. Figure 1 shows the MSE of the estimates for α considering the generated scenario.

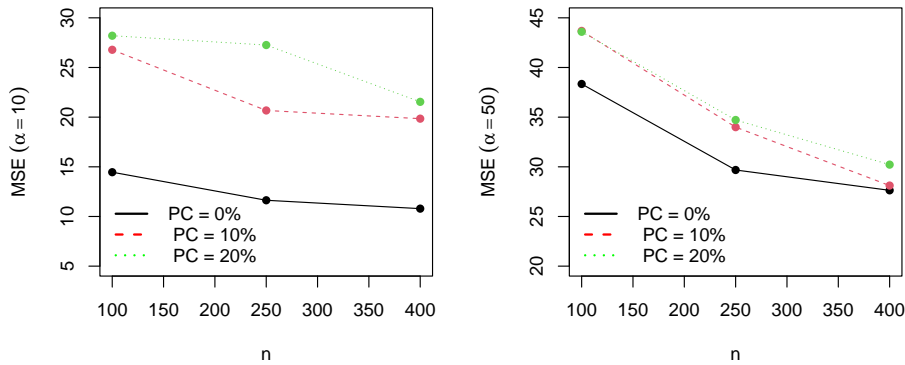


Figure 1: MSE of the estimates of α .

According to Table 2, we observe in general the MSE and the bias of the estimates for α decrease when the sample size increases. When the percentage of censorship increases, the MSE and the bias of the estimates for α and β tend to increase. For the estimates of β the MSE and the bias do not change much when the value of n is higher. However, the MSE values and the bias of the estimates for μ grow for most censored scenarios.

Table 1: MSE, SD, SE and bias of the estimators of the gamma ($1/\alpha, 1/\alpha$) additive model and an exponential baseline hazard rate with $\mu = 0.03$, $\beta = -0.1$ for different values of α and n with percentage of censorship (PC) ranging from 0% to 20%.

PC	n	α	$\hat{\alpha}$				$\hat{\mu}$				$\hat{\beta}$			
			MSE	SD	SE	bias	MSE	SD	SE	bias	MSE	SD	SE	bias
0%	100	10	14.4572	3.0686	0.6302	3.7497	0.0000	0.0049	0.0028	0.0007	0.0171	0.0131	0.0036	0.0589
		50	38.3403	2.7262	4.0151	4.7138	0.0001	0.0034	0.0029	0.0087	0.0192	0.0050	0.0042	0.0929
	250	10	11.6319	3.1830	0.3859	3.3887	0.0000	0.0046	0.0017	-0.0001	0.0169	0.0159	0.0022	0.0573
		50	29.6807	2.5224	2.5478	4.8155	0.0001	0.0025	0.0019	0.0087	0.0192	0.0033	0.0026	0.0930
	400	10	10.7890	2.9957	0.2995	3.2710	0.0000	0.0043	0.0014	0.0004	0.0170	0.0146	0.0018	0.0567
		50	27.6321	2.6287	2.0169	4.8543	0.0001	0.0019	0.0015	0.0089	0.0193	0.0029	0.0021	0.0925
10%	100	10	26.7829	2.7468	0.7001	4.3890	0.0004	0.0028	0.0017	-0.0196	0.0122	0.0108	0.0016	0.0573
		50	43.6808	2.4574	4.4230	4.9109	0.0002	0.0022	0.0019	-0.0148	0.0133	0.0034	0.0026	0.0908
	250	10	20.6736	2.5618	0.4482	4.5247	0.0004	0.0023	0.0011	-0.0194	0.0122	0.0092	0.0010	0.0579
		50	33.9930	2.6538	2.8054	5.1110	0.0002	0.0017	0.0012	-0.0148	0.0133	0.0027	0.0016	0.0906
	400	10	19.8484	2.9093	0.3511	4.4413	0.0004	0.0024	0.0009	-0.0192	0.0123	0.0212	0.0008	0.0559
		50	28.1127	2.4725	2.2008	4.8238	0.0002	0.0016	0.0009	-0.0149	0.0133	0.0022	0.0013	0.0906
20%	100	10	28.1979	2.8348	0.7300	4.4932	0.0006	0.0021	0.0014	-0.0236	0.0113	0.0106	0.0010	0.0558
		50	43.5955	2.4197	4.5589	4.7762	0.0004	0.0020	0.0015	-0.0204	0.0120	0.0029	0.0020	0.0899
	250	10	27.2608	2.5052	0.4605	4.5832	0.0006	0.0020	0.0009	-0.0238	0.0113	0.0088	0.0006	0.0566
		50	34.7141	2.5276	2.9088	5.1237	0.0004	0.0015	0.0010	-0.0204	0.0120	0.0022	0.0013	0.0900
	400	10	21.5413	2.5324	0.3632	4.6270	0.0006	0.0018	0.0007	-0.0238	0.0113	0.0088	0.0004	0.0565
		50	30.2201	2.5337	2.2967	4.9945	0.0004	0.0014	0.0008	-0.0203	0.0120	0.0020	0.0010	0.0899

In general, the estimates for β coincided with the sign of the parameter, even with values

close to zero. Thus, there are indications that the classical method is robust to estimate the model (2.7).

3.1. MISSPECIFICATION

Model misspecification simulation studies are often conducted to evaluate the performance of statistical models under misspecification. In the context of model misspecification simulation studies, AIC can be used to assess the impact of misspecification on model selection and to evaluate the reliability of these criteria in identifying the true underlying model. The motivation behind these model misspecification simulation studies is to understand how proposed models behave when they are applied to data that violate their underlying assumptions. The AIC criterion provides quantitative measures of the tradeoff between model fit and complexity. It penalizes models with excessive complexity, discouraging overfitting, while favoring models that provide a good fit to the data.

For the misspecification study, data were generated from the additive model with gamma frailty and base hazard of an exponential distribution(μ), in a similar way to that described in section 3. The same scenarios as the simulation study were used. We compared the data generating model for different values of α and n , with the additive model assuming the inverse Gaussian distribution for the frailty variable with density given by

$$(3.1) \quad f(v) = \left(\frac{\alpha}{\pi}\right)^{1/2} \exp\{2\alpha\} v^{-3/2} \exp\{-\alpha v - \alpha/v\}, \quad \alpha > 0,$$

where $E(v) = 1$ and $Var(v) = 1/2\alpha$. The moment generating function is $M_v(t) = \exp\{2\alpha - 2\sqrt{\alpha}\sqrt{\alpha-t}\}$, the Laplace transform is $L(t) = M_v(-t) = \exp\{2\alpha - 2\sqrt{\alpha}\sqrt{\alpha+t}\}$. So, the unconditional survival function, considering the effect of observed factors, is

$$(3.2) \quad S(t|\mathbf{x}_i) = \exp\{-\Lambda_0(t)\} \exp\{-\mathbf{x}'_i \beta t\} \exp\{2\alpha - 2\sqrt{\alpha}\sqrt{\alpha+t}\}$$

and as the derivative of $L(t)$ with respect to t is $L'(t) = \exp\{2\alpha - 2\sqrt{\alpha}\sqrt{\alpha+t}\} \left(-\frac{\sqrt{\alpha}}{\sqrt{\alpha+t}}\right)$, by 2.4, the unconditional hazard function is

$$(3.3) \quad \lambda(t|\mathbf{x}_i) = \lambda_0(t) + \mathbf{x}'_i \beta + \frac{\sqrt{\alpha}}{\sqrt{\alpha+t}}.$$

The results are presented in Table 2 in which different scenarios were generated and the percentages of cases in which the model with the lowest AIC was correctly selected were analyzed.

Table 2: Proportion of cases with the lowest AIC for the well-specified model for different values of α , n , and censoring percentages (CP), ranging from 0% to 20%.

CP	α	β	μ	n		
				100	250	400
0%	10	-0.1	0.03	0.9967	1.0000	1.0000
	50	-0.1	0.03	0.9633	0.9833	0.9800
10%	10	-0.1	0.03	0.9870	0.9810	0.9850
	50	-0.1	0.03	0.9970	1.0000	1.0000
20%	10	-0.1	0.03	0.9867	0.9567	0.9767
	50	-0.1	0.03	0.9967	0.9900	1.0000

The results consistently show that the model selection method with the smallest AIC is effective in correctly selecting the appropriate model when compared to the models. Furthermore, as the sample size increases, the ability to correctly select the model with the smallest AIC also increases. This suggests that a larger sample size provides better discrimination between models and increases the probability of selecting the correct model based on AIC. In general, in all tested scenarios the model with gamma frailty distribution was better than the model with inverse Gaussian frailty distribution in terms of AIC.

4. APPLICATION

In this section, the methodology is illustrated using two sets of data. The first is a medical data set corresponding to animal carcinogenicity described by [Gail et al. \(1980\)](#). The experiment used 48 mammary tumors from rats. There were 23 rats in group 1 (treatment) and 25 rats in group 2 (control), and the data are the days when new tumors occurred for each animal; a given animal may have multiple tumors. The main objective of the analysis is to assess the difference between treatment groups 1 and 2 with regard to tumor development. The second data set refers to readmission times after surgery in patients diagnosed with colorectal cancer ([González et al., 2005](#)), available in the package `fragilitypack` ([Rondeau et al., 2012](#)) of the R software ([R Core Team, 2022](#)). The study took place in Hospital de Bellvitge, a 960 bed public university hospital in the metropolitan area of Barcelona, Spain. Between January of 1996 and December of 1998, a total of 523 patients with incident colorectal cancer were identified. This study was based on 403 patients who had an operation. The outcome variable was readmission. It can be considered as a potential recurrent event (colorectal cancer patients may have several readmissions after discharge). The date of surgery was taken at the beginning of the study period. Patients were actively followed up until June 2002. Consequently, the length of follow up can differ for each patient, depending on the surgery date. Survival curves, by groups, were estimated by the nonparametric estimator for recurrent event data proposed by [Wang and Chang \(1999\)](#) and using the package `newTestSurvRec`.

4.1. DATA FROM ANIMALS WITH CARCINOGENESIS

The rats received a carcinogenic compound and after 60 days they were randomly divided into two groups (control group and treatment group). A follow-up period of 122 days began after randomization, during which they were examined twice weekly to assess for new tumor development (failure). The rats were divided into a group of 23 that received the treatment and another of 25 rats that formed the control group. In this study, the only covariate was treatment (yes = 1, no = -1). [Figure 2](#) shows recurrent cancerous tumor detection times considering both treatment and control groups.

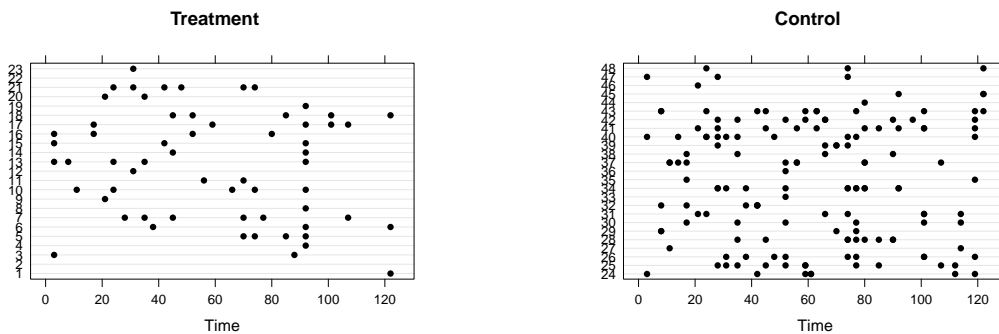


Figure 2: Recurrent cancer detection times.

[Figure 2](#) shows that the rats in the treatment group have few tumor recurrences, while the rats in the control group have many recurrences of the event of interest during the study period, suggesting that the treatment contributed to the decrease in the number of tumors in the study.

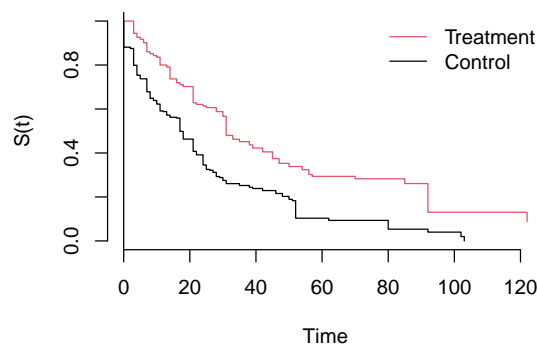


Figure 3: Survival curves estimated for treatment and control groups.

[Figure 3](#) represents the estimate of the survival function by group. We see that the treated group seems to have a longer survival, which seems to be related to the amount of tumors detected in each group.

Suppose that the i th individual rat has tumors occurring according to an additive frailty model with intensity (2.2), where x_i is a covariate indicating whether an individual is in treatment group ($x_i = 1$) or control group ($x_i = -1$). Time $t_{i,j}$ is defined to be number of days from the start, so that the observation intervals $(0, T_i)$ are $(0, 122)$ for all animals. The interest is to find characteristics for the parameters of the frailty model.

Table 3: ML estimates, standard errors (SEs) and 95 percent confidence intervals (CIs) for the parameters of the gamma additive frailty model with exponential and Weibull base risk functions.

Parameters	Exponential			Weibull		
	MLE	SE	CI(95%)	MLE	SE	CI(95%)
β	-0.014	0.0024	(-0.0180; -0.0099)	-0.014	0.0025	(-0.0179; -0.0098)
α	68.79	31.91	(16.14; 121.4)	82.19	43.32	(10.71; 153.67)
μ	0.035	0.0025	(0.0311; 0.0392)	0.028	0.0091	(0.0125; 0.0427)
γ	-	-	-	1.051	0.0673	(0.9397; 1.1617)

Applying the Likelihood Ratio (TRV) test to compare gamma additive frailty models with Weibull(μ, γ) (larger model) and exponential(μ) (smaller model) base risks, we obtained the value $LRT = 1.5134$ and $p - value = 0.2186$. Therefore, we do not reject the null hypothesis that the smaller model is more suitable at a 5% level of significance.

For both models, according to Table 3, we can conclude that the treatment covariate is significant because the confidence intervals for $\hat{\beta}$ do not contain zero, that is, the risk of developing new tumors decreases for mice that are in the treated group. In addition, the estimated variance for the frailty variable of both models with exponential baseline hazard rate ($\hat{\alpha} = 68.79$) and Weibull baseline hazard rate ($\hat{\alpha} = 82.19$) indicate the existence of heterogeneity among mice and a dependence between tumor recurrences in each mouse.

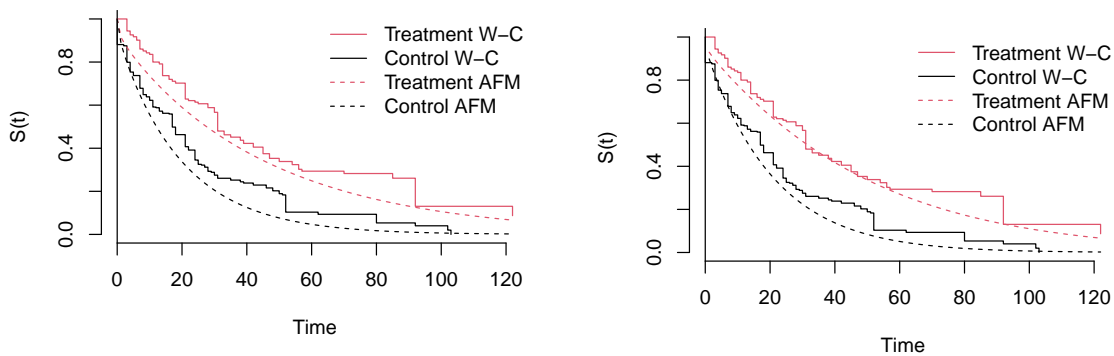


Figure 4: Wang and Chang (WC) and Additive Frailty Model (AFM) estimated curves with exponential and Weibull base risks.

In Figure 4, we can see that the survival curves estimated by the gamma additive frailty model with an exponential base risk are very close to the curves estimated by the WC estimator.

When fitting the Aalen additive model without frailty, the AIC and BIC values were, respectively, 1804.316 and 1808.059, which compared to the values for the gamma exponential additive frailty model were lower. However, the Aalen additive model with gamma frailty and exponential base risk is preferable due to the possibility of finding individual frailties. It is observed that the estimated variance for the gamma exponential additive frailty model was $\hat{\alpha} = 68.79$, providing a high heterogeneity in the data due to factors not considered in the model.

Considering the gamma-exponential additive frailty model, Table 4 shows the individual frailties estimated from Equation 2.13. Comparing the estimated frailties in Table 4 with the number of recurrences per rats, we can see that the greater the number of recurrences (presence of tumors) the greater the frailty of the rat. This is expected due to the nature of the event under study.

Also through Table 4, we notice that untreated individuals (control group) are generally more fragile than individuals in the group that received treatment, which was also expected due to the negative value of the estimate of the parameter β . The rat that showed more tumor recurrences (rat 34) was the one with the greatest frailty among all the rats in the study. The least fragile rats were rats 2 and 22.

Table 4: Individual frailties of the 48 rats divided by group.

i	Times (treatment group)	\hat{v}_i	i	Times (control group)	\hat{v}_i
1	122	0.0083	24	3, 42, 59, 61,61,112,119,122+	0.0575
2	122+	0.0001	25	28,31,35,45,52,59,59,77,85,107,112,122+	0.0903
3	3,88,122+	0.0165	26	31,38,48,52,74,77,101,101,119,122+	0.0739
4	92,122+	0.0083	27	11,114,122+	0.0165
5	70,74,85,92,122+	0.0329	28	35,45,74,74,77,80,85,90,90,122+	0.0739
6	38,92,122	0.0247	29	8,8,70,77,122+	0.0329
7	28,35,45,70,77,107,122+	0.0493	30	17,35,52,77,101,114,122+	0.0493
8	92,122+	0.0083	31	61,24,66,74,101,101,114,122+	0.0575
9	21,122+	0.0083	32	8,17,38,42,42,42,122+	0.0493
10	11,24,66,74,92,122+	0.0411	33	52,122+	0.0083
11	56,70,122+	0.0165	34	28,28, 31,38,52,74,74,77,77,80,80,92,92,122+	0.1067
12	31,122+	0.0083	35	17,119,122+	0.0165
13	3,8,24,35,92,122+	0.0411	36	52,122+	0.0083
14	45,92,122+	0.0165	37	11,11,14,17,52,56,56,80,80,107,122+	0.0821
15	3,42,92,122+	0.0247	38	17,35,66,90,122+	0.0329
16	3,17,52,80,122+	0.0329	39	28,66,70,70,74,122+	0.0411
17	17,59,92,101,107,122+	0.0411	40	3,14,24,24,28,31,35,48,74,77,119,122+	0.0903
18	45,52,85,101,122	0.0411	41	21,28,45,56,63,80,85,92,101,101,119,122+	0.0903
19	92,122+	0.0083	42	28,35,52,59,66,66,90,97,119,122+	0.0739
20	21,35,122+	0.0165	43	8,8,24,42,45,59,63,63,77,101,119,122	0.0985
21	24,31,42,48,70,74,122+	0.0493	44	80,122+	0.0083
22	122+	0.0001	45	92,122,122	0.0247
23	31,122+	0.0083	46	21, 122+	0.0083
-	-	-	47	3,28,74,122+	0.0247
-	-	-	48	24,74,122	0.0247

A second modelling of the data from the carcinogenic animals was performed using the additive model with inverse Gaussian frailty and with a base risk of exponential distribution (μ). Table 5 shows the results of the parameter estimates (MLE), standard errors (SE) and

95% confidence intervals obtained by fitting the inverse Gaussian additive frailty model with an exponential base hazard function.

Table 5: Maximum likelihood estimates (MLE), standard error (SE), confidence interval - CI (95%) for the inverse Gaussian additive frailty model with exponential base hazard function.

Parameters	MLE	SE	CI(95%)
β	-0.020	0.0029	(-0.026; -0.015)
α	0.0057	0.0027	(0.0003; 0.0111)
μ	0.0309	0.0013	(0.0283; 0.0336)

From Table 5 we can see that the additive frailty model with gamma frailty and exponential base risk is better in terms of AIC and BIC, with AIC and BIC values of 1811.188 and 1816.801, respectively, for the gamma exponential MFA and 1835.768 and 1841.381 for the GI exponential MFA. Furthermore, we can see that for the data of this application, the choice of the inverse Gaussian frailty did not provide a good fit.

4.2. REHOSPITALIZATION DATA

This data set includes 403 patients who have been followed for approximately 6 years. Observed times are the days of hospitalization after surgery. The first readmission time was considered as the time between the date of the surgical procedure and the first readmission after hospital discharge. The other readmission times were defined as the difference between the last admission date and the previous discharge date. For this application, we only use the variable that indicates whether the patient has received chemotherapy ($x = 1$: No chemotherapy, $x = 2$: Chemotherapy treatment received).

Of the 403 patients in the study, 217 (53.85%) were treated with chemotherapy and the other 186 (46.15%) did not receive chemotherapy treatment. Patients treated with chemotherapy had an average of 0.81 readmissions, while an individual in the group that did not receive this treatment was readmitted an average of 1.52 times, that is, almost twice as much as the first group. Patient 350, who belongs to the group that did not receive chemotherapy, had the maximum number of readmissions for a study subject reaching 22 readmissions during the follow-up period.

Figure 5 shows the survival curves obtained by the non-parametric estimator for data on recurrent events for each group of patients. The treated group seems to have a longer survival, which may be related to the treatment used in each group.

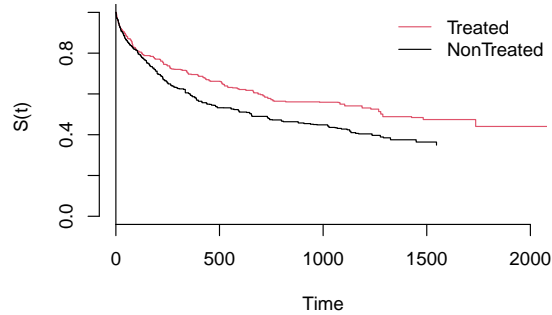


Figure 5: Survival curves estimated for treated and untreated patient groups.

Table 6 shows the ML estimates, standard errors and confidence intervals for the parameters of the gamma-exponential additive frailty model.

Table 6: ML estimates, SEs and 95 percent CIs for the parameters of the gamma additive frailty model with exponential base risk function.

Parameters	MLE	SE	CI(95%)
β	-0.0003	4.8×10^{-5}	(-0.0004; -0.0002)
α	30.16	0.3406	(29.49; 30.83)
μ	0.0009	8.5×10^{-5}	(0.0007; 0.0010)

From the results in Table 6, we see that the covariate indicates the chemotherapy treatment is significant because the confidence interval $[-0.0004; -0.0002]$ for the parameter β does not contain zero, that is, the risk of readmission decreases for patients who are in the treated group ($\hat{\beta} = -0.0003$). Through the value of $\hat{\alpha} = 30.16$, we observe that there is heterogeneity among patients, that is, there are unobserved factors (for example, genetic or environmental), which can influence the lifespan of patients and there is still dependence between the rehospitalization times of patients, that is, as new hospitalizations happen, the patient becomes more fragile.

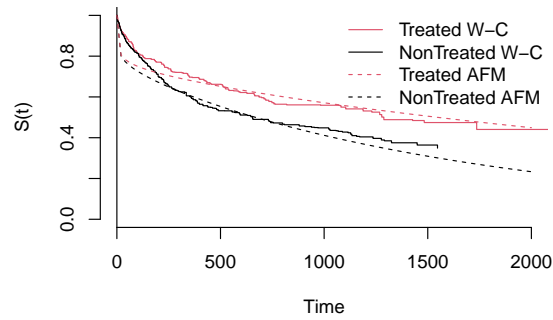


Figure 6: Comparison of the survival curves estimated by the $\text{gamma}(1/\alpha, 1/\alpha)$ additive frailty model and $\text{exponential}(\mu)$ base risk with the curves estimated by the nonparametric estimator.

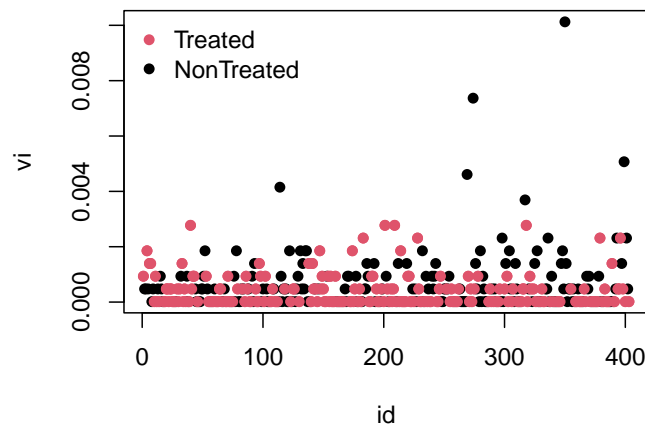


Figure 7: Individual frailties estimated by the gamma-exponential additive model for the 403 patients divided by group (treated and untreated).

In [Figure 6](#), we see that the survival curves estimated by the gamma-exponential additive frailty model are very close to the curves estimated by the nonparametric estimator, which indicates a good fit of the model. Individual frailty can be estimated from [Equation 2.13](#) with estimates in [Table 6](#).

For the readmission data, we notice that the greater the number of recurrences (re-hospitalizations), the greater the patient's frailty. This is expected due to the nature of the event under study. From [Figure 7](#), we note that individuals not treated with chemotherapy are generally more fragile than individuals in the group that received the treatment, which was also expected due to the negative value of the estimate of β . The patient with the most readmissions (patient 350) was the one with the greatest frailty among all the patients in the

study. The least frail were the patients who did not have readmissions.

5. CONCLUDING REMARKS

The additive models initially proposed by [Aalen \(1980\)](#) are interesting alternatives when there is no guarantee of proportionality of risks, as in the case of recurrent events when the assumption of proportional risks is not reasonable. However, this additive modeling has the disadvantage of providing negative values for the risk function.

For the recurrent event data studied in this paper, we adjusted a gamma frailty model with exponential base risk, that is, we used a simple model and achieved a relatively close adjustment to the survival curves estimated by the nonparametric estimator. By additive modeling with additive frailty, we conclude that the treatment applied to mice decreases the risk of them presenting cancerous tumors, that is, decreases the risk of death of the treated mice. Still through the model used, we conclude that there are factors not observed in the study that influence rats' lifetime and that recurrent times are dependent. In the second application, we notice that in a larger data set and for a not too big α the estimates are more accurate in the sense that they have smaller standard errors, which is expected due to asymptotic properties of the ML estimators.

We observed in the simulated study that as the percentage of censorship increases, the ML estimates for the additive model parameters with additive brittleness worsen, in terms of both MSE and bias. We also noted that the classical estimation method has a limitation to estimate the parameters of the additive model studied in scenarios of great variability among individuals.

As in every parametric approach, several models can be fitted to the same data. Depending on the choice of distribution for frailty and for the base risk of the additive model, other adjustments can be found, including better ones, however we still do not know of a more objective method for a better choice, other than adjusting several models and comparing them through some selection criteria. The model presented, even with some limitations, is a useful model for recurrent event data and can be used with some reservations.

For the model studied in this paper, the estimator proposed to calculate the individual frailties of patients proved to be very intuitive, in the sense that it attributes greater frailty to individuals who presented greater recurrence of the event of interest.

ACKNOWLEDGMENTS

The authors would like to thank the Editor and the referee for careful reading and comments which greatly improved the paper.

REFERENCES

- Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726.
- Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical Statistics and Probability Theory*, pages 1–25. Springer Verlag, New York.
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.
- Cox, D. R. and Isham, V. (1980). *Point Processes*, volume 12. CRC Press, London.
- Fletcher, R. and Reeves, C. (1964). Function minimization by conjugate gradients. *Computer Journal*, 7:148–154.
- Gail, M., Santner, T., and Brown, C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, 36:255–266.
- González, J. R., Fernandez, E., Moreno, V., Ribes, J., Peris, M., Navarro, M., Cambray, M., and Borràs, J. M. (2005). Sex differences in hospital readmission among colorectal cancer patients. *Journal of Epidemiology and Community Health*, 59(6):506–511.
- Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71(1):75–83.
- Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer Verlag, New York.
- Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37(2):158–168.
- Munda, M., Rotolo, F., and Legrand, C. (2012). parfm: Parametric frailty models in R. *Journal of Statistical Software*, 51(11):1–20.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rocha, C. M. T. S. (1995). *Modelos com fragilidade em Análise de Sobrevivência*. PhD thesis, Universidade de Lisboa.
- Rondeau, V., Mazroui, Y., and Gonzalez, J. R. (2012). frailtypack: An r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47(4):1–28.
- Silva, G. L. (2001). *Análise Bayesiana de modelos de sobrevivência com fragilidade*. PhD thesis, Universidade Técnica de Lisboa - Instituto Superior Técnico.
- Tomazella, V., Louzada-Neto, F., and Silva, G. (2006). Bayesian modeling of recurrent events data with an additive gamma frailty distribution and a homogeneous Poisson process. *Journal of Statistical Theory and Applications*, 5(4):417–429.
- Tomazella, V. L. D. (2003). *Modelagem de dados de eventos recorrentes via processo de Poisson com termo de fragilidade*. PhD thesis, Universidade de São Paulo.
- Wang, M.-C. and Chang, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, 94(445):146–153.