scientific reports

Check for updates

OPEN

The multivariate shared truncated normal frailty model with application to medical data

Diego I. Gallardo[1], Yolanda M. Gómez[1]✉, John L. Santibañez[2], Osvaldo Venegas[3]✉ & Marcelo Bourguignon[4]

A new multivariate shared frailty model based on the truncated normal distribution is proposed. For the basal distribution of failure times, we assume a parametric approach through the Weibull and piecewise exponential distributions and also a nonparametric approach. Similar to the traditional gamma frailty model, the Laplace transform, the hazard and survival functions of our proposal have a simple and closed form. In addition, the *n*-th derivative of the Laplace transform can be expressed recursively. Parameter estimation is performed by a classical approach through the EM algorithm. A simulation study is presented to demonstrate the consistency of the estimators in finite samples. Finally, two applications to medical data modelling the recurrence of infection in renal patients and patients with fibrosarcoma are presented to demonstrate the effectiveness of the model compared to other classical approaches in the literature. The computational implementation of the model is available in the extrafrail package of R.

Keywords Frailty models, Cox model, Kendall's τ, Extrafrail package, EM algorithm, Kidney Disease

Survival models study the time to event until a certain event of interest occurs. They are characterized by including censored (incomplete) information within the study, either because the individual never presented the event during follow-up or because the follow-up of the individual was truncated during the study. Within this context, due to the fact of not assuming a specific distribution for failure times, one of the most referenced models in the literature is Cox's proportional hazards (PH). For $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ a set of p observed covariates (without intercept term), the hazard risk function for this model is given by

$$h(t \mid \boldsymbol{x}_i) = h_0(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ denotes a vector of p observed covariates and $h_0(\cdot)$ denotes the baseline hazard function. Note that this model provides a proportional hazard structure because the ratio for two individuals with profiles $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$

$$\frac{h(t \mid \boldsymbol{x}_i)}{h(t \mid \boldsymbol{x}_{i'})} = \exp\left((\boldsymbol{x}_i - \boldsymbol{x}_{i'})^\top \boldsymbol{\beta}\right),$$

does not depend on t. A way to break the proportional hazard risk assumption is by using univariate frailty models, although in practice the concept of frailty is more intuitive to explain in the context of data grouped into clusters or data that have some type of association (measurements of the same individual, for example). In the literature, there are many models considered for the frailty distribution in a univariate context. To name a few, gamma[1,2], inverse gaussian (IG)[3,4], Birnbaum-Saunders (BS)[5,6], folded normal[7], weighted Lindley (WL)[8,9], mixture of IG[10], among others, where the restriction that the frailty variable has mean 1 is usually required to avoid identifiability problems.

However, when the observations are grouped in clusters with different sizes, a multivariate frailty model framework is required. In addition to the aforementioned restriction, in this case is required that the derivatives of the Laplace transform have a known form because the joint density function depends on it. Few distributions

[1]Departamento de Estadísticas, Facultad de Ciencias, Univerisidad del Bío-Bío, Concepción 4081112, Chile. [2]Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó 1530000, Chile. [3]Departamento de Ciencias Matemáticas y Físicas, Facultad de Ingeniería, Universidad Católica de Temuco, Temuco 4780000, Chile. [4]Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal 59078-970, Brazil. ✉email: ygomez@ubiobio.cl; ovenegas@uct.cl

in the literature satisfy these conditions. The gamma, IG and the recently proposed WL shared frailty model[9] satisfies those conditions. For this reason, the literature on frailty models in a multivariate context has increased only for the bivariate or trivariate case, in which case all the clusters have 2 or 3 observations, respectively. In addition to the three distributions mentioned above, we found the generalized exponential discussed in[11] and the generalized inverse Gaussian presented in[12]. The truncated normal (TN) model was mentioned as a possible frailty distribution in[7]. However, it was used without imposing any mean restrictions or reparameterization, applying it solely within a copula model and restricting their analysis to the bivariate case. To date, the behavior of the TN model in the context of frailty has not been explored for clusters larger than two, let alone for groups with varying sample sizes.

In this paper, we use the TN distribution as the frailty distribution for clustered survival data. For our model to be identifiable, we employ a TN distribution with mean one and frailty variance as the frailty distribution by using a new parameterization of the TN distribution. The conditional distribution of frailties among the survivors and the frailty of individuals dying at time t can be explicitly determined. Furthermore, we propose a recurrent closed form for the derivatives of the Laplace transform. For parameter estimation, we give a simple EM algorithm, since all conditional expectations involved in the E-step are obtained in explicit form. Finally, the results of this paper have been implemented into R statistical software. The manuscript is organized as follows. Section 2 presents a background of frailty models and introduces the TN frailty model with parameterization such that the mean of the distribution is 1. Section 3 discusses the estimation procedure for the model based on a classical approach. Section 4 presents a simulation study to assess the performance of the proposed estimators in finite samples. In Section 5, we present two real data, the first related to the recurrence times of patients with renal problems and the second fibrosarcoma data. Finally, in Section 6 are presented the main conclusions of this work.

Background of frailty models

In this Section, we introduce the truncated normal distribution and we present a background of frailty models. Then, we introduced the novelty truncated normal frailty model for the univariate and multivariate cases.

The truncated normal distribution

A variable Z has TN distribution defined in the positive axis if its probability density function (PDF) is given by

$$g(z) = \frac{\phi\left(\frac{z-\mu}{\sigma}\right)}{\sigma\,\Phi\left(\frac{\mu}{\sigma}\right)}, \quad z > 0,$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the PDF and cumulative density function (CDF) of the standard normal distribution, $-\infty < \mu < \infty$ represents a location parameter and $\sigma > 0$ a scale parameter. The mean and variance of the TN distribution are given by

$$\mathbb{E}(Z) = \mu + \sigma\frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}, \quad \text{and} \quad \text{Var}\,(Z) = \sigma^2\left\{1 - \frac{\mu\,\phi(\mu/\sigma)}{\sigma\,\Phi(\mu/\sigma)} - \left(\frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}\right)^2\right\}.$$

Considering the reparameterization $\nu = \mu/\sigma$ and the restriction $\sigma = \left(\nu + \frac{\phi(\nu)}{\Phi(\nu)}\right)^{-1}$, we obtain that the pdf of the model is reduced to

$$g(z) = \frac{\gamma\phi\Big(\gamma z - \nu\Big)}{\Phi\big(\nu\big)}, \quad \nu \in \mathbb{R}, z > 0, \tag{2}$$

with $\gamma = \gamma(\nu) = \nu + \phi(\nu)/\Phi(\nu)$, and the mean and variance of the model are given by

$$\mathbb{E}(Z) = 1 \quad \text{and} \quad \theta = \text{Var}\,(Z) = \gamma^{-2} - \frac{\phi(\nu)}{\Phi(\nu)}\gamma^{-1},$$

respectively. From now on we will use the notation TN(ν) to refer to a random variable with PDF given in Equation (2). We note that this parameterization was not proposed in the statistical literature. But, it is not possible to directly reparameterize the frailty variance in terms of θ, however, there is a one-to-one relationship between θ and ν. Thus, this parameterization is very useful because allows us to compare different frailty models also parameterized in the frailty variance directly.

Note that under the restriction $\mathrm{E}(Z) = 1, 0 \leq \theta = \text{Var}\,(Z) \leq 1$. In principle, this can be a disadvantage. However, in practice usually, the frailty variance satisfies this condition (see Section 6).

Figure 1 shows the pdf and variance of the TN(ν) model with different values for ν. The flexibility of the TN distribution is apparent. Furthermore, the variance of the TN distribution decreases as ν increases.

The Laplace transform for the TN(ν) model is given by

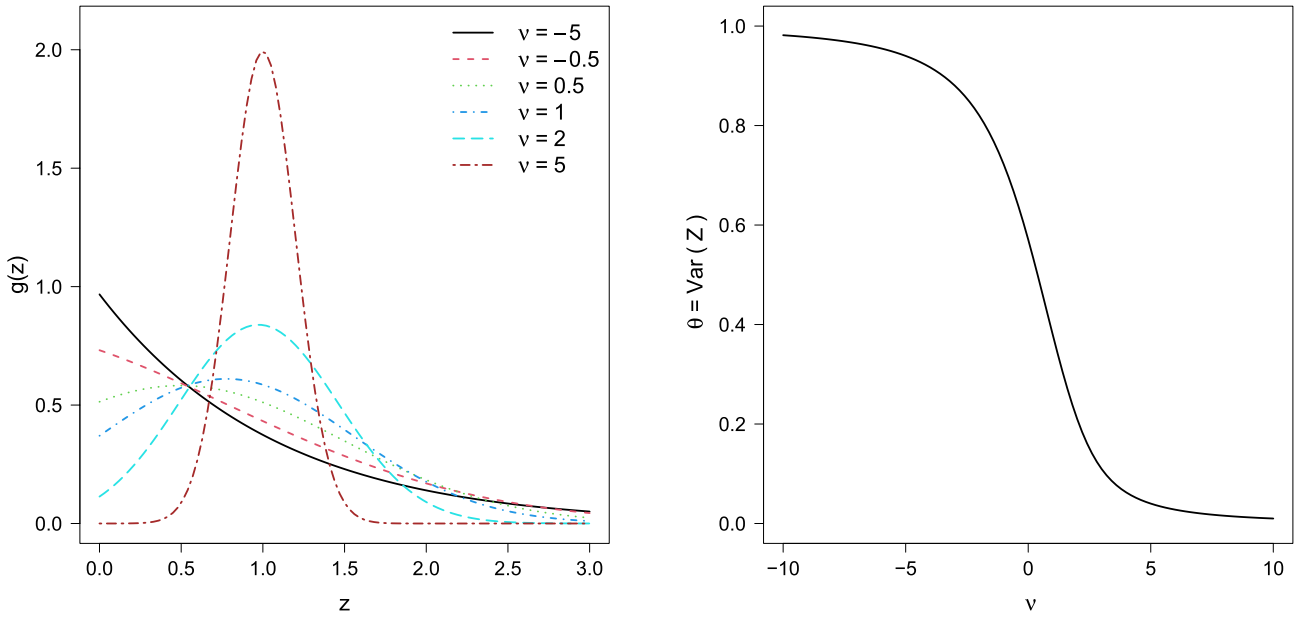

Fig. 1. PDF of the TN model and variance of the distribution in terms of ν. PDF of the TN model (left) and variance of the TN distribution (right).

$$\mathcal{L}_g(s) = \frac{\Phi(\kappa)}{\Phi(\nu)} \exp\left\{\frac{s}{\gamma}\left(\frac{s}{2\gamma} - \nu\right)\right\}, \tag{3}$$

where $\kappa = \kappa(s, \nu) = \nu - s/\gamma$. Let $\mathcal{L}_g^{(d)}(s)$ be the d-th derivative of the Laplace transform. For $d = 1$ and $d = 2$ such term is given by

$$\mathcal{L}_g^{(1)}(s) = -\frac{\mathcal{L}_g(s)}{\gamma}\left(\kappa + \frac{\phi(\kappa)}{\Phi(\kappa)}\right) \quad \text{and} \quad \mathcal{L}_g^{(2)}(s) = \frac{\mathcal{L}_g(s)}{\gamma^2}\left[\kappa\left(\kappa + \frac{\phi(\kappa)}{\Phi(\kappa)}\right) + 1\right]. \tag{4}$$

In[13], Corollary 2.1 presents a recurrence relation for derivatives of order 3 or higher of the generating-moment function (denoted as $M_g(\cdot)$) for the TN model. Using the property $M_g(s) = \mathcal{L}_g(-s)$ we can derive the following relation:

$$\mathcal{L}_g^{(d)}(s) = \frac{(d-1)}{\gamma^2}\mathcal{L}_g^{(d-2)}(s) - \frac{\kappa}{\gamma}\mathcal{L}_g^{(d-1)}(s), \quad d = 3, 4, \ldots, \tag{5}$$

which depends on the two last derivatives, but it is simple to implement computationally. Higher-order Laplace transforms are provided in the table included in the Supplementary Material file. The results in Equations (4) and (5) are very important to the development of our approach for the TN model within the context of frailty models.

Univariate frailty models

In a univariate context, the extended Cox model with the unobserved source of heterogeneity has a conditional hazard function given by

$$h(t \mid z_i, \mathrm{x}_i) = z_i\, h_0(t) \exp(\mathrm{x}_i^\top \boldsymbol{\beta}), \tag{6}$$

where x_i denotes a vector of covariates and z_i is a latent variable representing the unobserved heterogeneity of the i-th observation. For $z_1, z_2, \ldots, z_n$ a positive distribution is assumed (say one with pdf $g(\cdot)$), typically with mean 1 to avoid identifiability problems[14]. Similar to Eq. (1), this implies that the quotient of the conditional hazard function of two individuals does not depend on t, but we remark that in this case it is the conditional (and not marginal) risk function that satisfies this property. Also note that the larger z_i is, the greater the risk associated with that observation. The conditional survival function for the i-th individual obtained from equation (6) is given by

$$S(t \mid z_i, \mathrm{x}_i) = \exp\left\{-\int_0^t h(u \mid z_i, \mathrm{x}_i)du\right\} = \exp\left\{-z_i H_0(t) \exp(\mathrm{x}_i^\top \boldsymbol{\beta})\right\},$$

where $H_0(t) = \int_0^t h_0(u)du$ represents the basal cumulative hazard function. The marginal survival function can be obtained as

$$S(t \mid \mathrm{x}_i) = \int_0^\infty \exp\Big\{ - z_i H_0(t) \exp(\mathrm{x}_i^\top \boldsymbol{\beta})\Big\} g(z_i) dz_i = \mathcal{L}_g\left(H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})\right),$$

where $\mathcal{L}_g(\cdot)$ corresponds to the Laplace transform of the pdf $g(\cdot)$. On the other hand, the marginal hazard function is given by

$$h(t \mid \mathrm{x}_i) = -\frac{\partial S(t \mid \mathrm{x}_i)/\partial t}{S(t \mid \mathrm{x}_i)} = -\frac{h_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})\, \mathcal{L}_g^{(1)}\left(H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})\right)}{\mathcal{L}_g\left(H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})\right)}, \tag{7}$$

where $\mathcal{L}_g^{(d)}(\cdot)$, $d \in \mathbb{Z}$, denotes the d-th derivative of $\mathcal{L}_g(\cdot)$. It is clear from Eq. (7) that the assumption PH is not satisfied in this case. Particularly, when $Z_i \sim$ TN (ν), the marginal survival and hazard functions are reduced to

$$\begin{aligned} S(t \mid \mathrm{x}_i) &= \frac{\Phi\left(\nu - \frac{H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})}{\gamma}\right)}{\Phi\left(\nu\right)} \exp\left\{ \frac{H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})}{\gamma} \left(\frac{H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})}{2\gamma} - \nu \right) \right\}, \quad \text{and} \\ h(t \mid \mathrm{x}_i) &= -\frac{h_0(t)}{\gamma} \exp\left(\mathrm{x}_i^\top \boldsymbol{\beta}\right) \left\{ \nu - \frac{H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})}{\gamma} + \frac{\phi\left(\nu - \frac{H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})}{\gamma}\right)}{\Phi\left(\nu - \frac{H_0(t)\exp(\mathrm{x}_i^\top \boldsymbol{\beta})}{\gamma}\right)} \right\}. \end{aligned} \tag{8}$$

Finally, for the univariate case we present two propositions related to the conditional distribution for the frailty given the events $T > t$ and $T = t$, respectively.

Proposition 1.1 The conditional distribution for the frailty $Z \mid T > t$, follows a TN(ε), where $\varepsilon = \varepsilon(H_0(t), \nu) = \nu - H_0(t)/\gamma$.

Proposition 1.2 The conditional distribution for the frailty $Z \mid T = t$ follows a modified half Normal (MHN)[15], which density function is given by

$$f(z|T=t) = \frac{\gamma^2 \exp\left(-\frac{\kappa^2}{2}\right)}{\sqrt{2\pi}\left(\kappa\Phi(\kappa) + \phi(\kappa)\right)} z \exp\left\{ -\frac{\gamma^2}{2} z^2 + \gamma\kappa z \right\}.$$

Proofs of Propositions 1.1 and 1.2 are provided in the Supplementary Material.

Multivariate shared frailty model

In a more general context, it is possible to consider that the observations are grouped in m clusters and the ith cluster has n_i observations, for $i = 1, \ldots, m$. This scenario is ad hoc when the observations in the same cluster have some kind of dependence. For instance, measurements in the same individual, or members of the same family, among others. The assumption here is that all the observations related to the same cluster are conditionally independent given its corresponding frailty term (z_i). With this assumption, we obtain that the conditional hazard and the joint survival function are given by

$$\begin{aligned} h(t_{i1}, \ldots, t_{in_i} \mid z_i, \mathrm{X}_i) &= \sum_{j=1}^{n_i} h(t_{ij} \mid z_i, \mathrm{x}_{ij}) = z_i \sum_{j=1}^{n_i} \exp\left(\mathrm{x}_{ij}^\top \boldsymbol{\beta}\right) h_0(t_{ij}), \quad \text{and} \\ S(t_{i1}, \ldots, t_{in_i} \mid z_i, \mathrm{X}_i) &= \exp\left(-z_i \sum_{j=1}^{n_i} \exp\left(\mathrm{x}_{ij}^\top \boldsymbol{\beta}\right) H_0(t_{ij}) \right), \end{aligned}$$

respectively, where $\mathrm{x}_{ij}^\top = (x_{ij1}, \ldots, x_{ijp})$ denotes a vector of p covariates related to the j-th individual in the i-th cluster and $\mathrm{X}_i^\top = (\mathrm{x}_{i1}^\top, \ldots, \mathrm{x}_{in_i}^\top)$ denotes the vector with all the information for the p covariates associated with the n_i observations in the i-th cluster, z_i represents the influence of the i-th cluster on its observations. Integrating z_i over its density function is obtained that the marginal survival function for $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})$ is given by

$$S(\boldsymbol{t}_i \mid \mathrm{X}_i) = \mathcal{L}_g\left(\sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathrm{x}_i^\top \boldsymbol{\beta}) \right),$$

and then, the marginal hazard function is given by

$$h(\boldsymbol{t}_i \mid \mathrm{X}_i) = \frac{(-1)^{n_i} \sum_{j=1}^{n_i} h_0(t_{ij}) \exp(\mathrm{x}_i^\top \boldsymbol{\beta}) \, \mathcal{L}_g^{(n_i)} \left(\sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathrm{x}_i^\top \boldsymbol{\beta}) \right)}{\mathcal{L}_g \left(\sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathrm{x}_i^\top \boldsymbol{\beta}) \right)}. \tag{9}$$

For the TN, expressions for Equation (8) can be expressed using the recursive formula in (5). For the bivariate case (i.e., $n_i = 2, \forall i = 1, \ldots, m$), the marginal hazard function is reduced to

$$h(t_{i1}, t_{i2} \mid \mathrm{x}_{i1}, \mathrm{x}_{i2}) = \frac{1}{\gamma^2} \sum_{j=1}^{2} h_0(t_{ij}) \exp(\mathrm{x}_{ij}^\top \boldsymbol{\beta}) \left[\left(\nu - \frac{s_i}{\gamma} \right) \left(\left(\nu - \frac{s_i}{\gamma} \right) + \frac{\phi\left(\nu - \frac{s_i}{\gamma}\right)}{\Phi\left(\nu - \frac{s_i}{\gamma}\right)} \right) + 1 \right],$$

where $s_i = H_0(t_{i1}) \exp(\mathrm{x}_{i1}^\top \boldsymbol{\beta}) + H_0(t_{i2}) \exp(\mathrm{x}_{i2}^\top \boldsymbol{\beta})$.

Kendall's tau

Kendall's τ is a measure that quantifies the dependency between observations in the same cluster. This measure is independent of the unit of measurement of the data, so it works better than the variance and the correlation of the data due to its limitations (non-existence of the second moment, existence of censored observations, different measurement scale, see[16], page 153, for details). Considering the Laplace transform and its second derivative (see Equations (3) and (4), respectively), we can determine the value of τ, for TN distribution, which is defined as

$$\begin{aligned} \tau &= 4 \int_0^\infty s \mathcal{L}^{(2)}(s) \mathcal{L}(s) ds - 1, \\ &= 4 \int_0^\infty s \left[\frac{\Phi(\kappa)}{\gamma \Phi(\nu)} \right]^2 \exp\left\{ \frac{s}{\gamma} \left(\frac{s}{\gamma} - 2\nu \right) \right\} \left[\kappa \left(\kappa + \frac{\phi(\kappa)}{\Phi(\kappa)} \right) + 1 \right] ds - 1. \end{aligned}$$

The integral is solved computationally since it does not have a closed expression. Figure 2 shows the different dependency values (τ) according to the variance value for different frailty models. Note that $\tau \in [0, 0.33]$, for $\theta \in (0, 1)$ in the TN frailty model. We also note that, for a given frailty variance $\theta \in (0, 0.864)$, the TN frailty model produces a higher degree of dependence τ than the GA, IG, and WL frailty models.

On the basal hazard function

The basal hazard function $h_0(t)$ is usually modeled with common distributions with positive support, such as Weibull, gamma, and Gompertz, among others. For the Weibull distribution, we consider the parameterization such as $h_0(t) = \lambda \rho t^{\rho - 1}$ and $H_0(t) = \lambda t^\rho, t, \lambda, \rho > 0$ and we denote $T \sim \mathrm{W}(\lambda, \rho)$ to refer to this particular

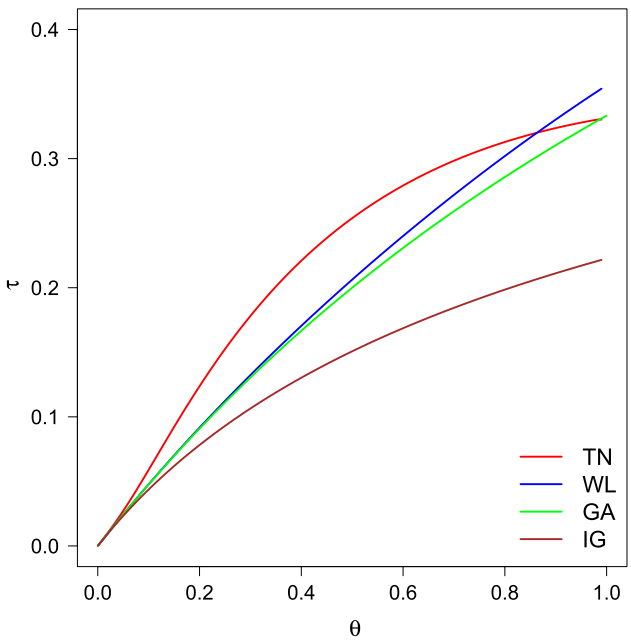

Fig. 2. Comparison among Kendall's τ for TN, weighted Lindley (WL), gamma (GA) and inverse Gaussian (IG).

parameterization. The Weibull model has been widely used in the literature because it adapts well to diverse biological, physical, chemical, and industrial processes, to name a few. Furthermore, its hazard function can assume monotonic forms (increasing, decreasing, or constant), which are controlled only by ρ. On the other hand, the piecewise exponential (PE) distribution introduced in[17] and extended in[18] for the case with covariates. This model considers a constant risk between each predefined interval, say $(a_1, ..., a_L)$ such as $0 = a_0 < a_1 < ... < a_{L-1} < a_L = \infty$. This distribution is extremely useful for adapting critical points where there may be abrupt changes in the baseline risk function and which cannot be captured by non-segmented distributions such as the Weibull distribution. We say that T has PE model with vector of parameters $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_L)$ and known partition time $\boldsymbol{a} = (a_1, ..., a_{L-1})$ (we denote $T \sim PE_a(\boldsymbol{\lambda})$), if its survival function is given by

$$S(t) = \exp\left(-\sum_{l=1}^{L} \lambda_l \nabla_l(t) \right), \ \ t > 0,$$

where

$$\nabla_l(t) = \begin{cases} 0, & \text{if } t < a_{l-1}, \\ t - a_{l-1}, & \text{if } a_{l-1} \le t < a_l, \\ a_l - a_{l-1}, & \text{if } t > a_l. \end{cases}$$

The hazard function is given by

$$h_0(t) = \lambda_\ell, \quad t \in (a_{\ell-1}, a_\ell], \ \ell = 1, ..., L,$$

and the cumulative hazard function is given by

$$H_0(t) = \sum_{l=1}^{L} \lambda_l \Delta_l(t).$$

In the literature, when the PE model is used in the context of frailty models, it is typically referred to as a semi-parametric model[19]. However, in this work, we also consider a non-parametric form for the baseline hazard distribution.

Estimation

In this section, we discuss the parameter estimation for the TN frailty model. Let Y_{ij} and C_{ij} be the failure and censoring times for the *j*-th individual in the *i*-th cluster and x_{ij} be a $p \times 1$ covariate vector (without intercept term), where $1 \le i \le m$ and $1 \le j \le n_i$. Under a right censoring scheme, we observe the random variables $T_{ij} = \min(Y_{ij}, C_{ij})$ and $\delta_{ij} = I(Y_{ij} \le C_{ij})$, where $I(A) = 1$ if the event A occurs (0 otherwise). We assume the frailty terms $Z_1, \ldots, Z_m$ to be a random sample from the TN(θ) distribution. Considering the following assumptions:

i) The pairs $(Y_{i1}, C_{i1}), \ldots, (Y_{in_i}, C_{in_i})$ are conditionally independent given Z_i, and Y_{ij} and C_{ij} are mutually independent for $j = 1, \ldots, n_i$.
ii) $C_{i1}, \ldots, C_{in_i}$ are non-informative about Z_i.

Under this setting, the observed log-likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\beta}, H_0, \nu) &= \prod_{i=1}^{m} \int_0^{+\infty} \prod_{j=1}^{n_i} \left[z_i h_0(t_{ij}) \exp\left(\mathrm{x}_{ij}^\top \boldsymbol{\beta}\right) \right]^{\delta_{ij}} \exp\left(-z_i H_0(t_{ij}) \mathrm{e}^{\mathrm{x}_{ij}^\top \boldsymbol{\beta}} \right) \frac{\gamma \phi(\gamma z_i - \nu)}{\Phi(\nu)} dz_i \\ &= \left(\frac{\gamma \mathrm{e}^{-\nu^2/2}}{\sqrt{2\pi}\Phi(\nu)} \right)^m \exp\left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \mathrm{x}_{ij}^\top \boldsymbol{\beta} \right) \prod_{i=1}^{m} \int_0^{\infty} z_i^{r_i} \exp\left(-b_\nu z_i^2 + c_{\boldsymbol{\psi}}^{(i)} z_i \right) dz_i \prod_{j=1}^{n_i} h_0(t_{ij})^{\delta_{ij}}. \end{aligned}$$

where $r_i = \sum_{j=1}^{n_i} \delta_{ij}$ is the failures in the *i*-th cluster, $b_\nu = \gamma^2/2$ and $c_{\boldsymbol{\psi}}^{(i)} = \gamma\nu - \sum_{j=1}^{n_i} H_0(t_{ij}) e^{\mathrm{x}_{ij}^\top \boldsymbol{\beta}}$. However, the last integral is related to the modified half-normal (MHN) distribution[15] and it can be written as

$$\int_0^{\infty} z_i^{r_i} \exp\left(-b_\nu z_i^2 + c_{\boldsymbol{\psi}}^{(i)} z_i \right) dz_i = \frac{1}{2} b_\nu^{-(r_i+1)/2} \Psi\left(\frac{r_i + 1}{2}, \frac{c_{\boldsymbol{\psi}}^{(i)}}{\sqrt{b_\nu}} \right),$$

where

$$\Psi\left(\frac{\alpha}{2}, x \right) = \sum_{k=0}^{\infty} \frac{\Gamma\left(\frac{\alpha+k}{2}\right)}{k!} x^k,$$

is a specific case of the Fox-Wright function. The supplementary material in[15] discusses different ways to compute this term. Therefore,

$$L(\boldsymbol{\beta}, H_0, \nu) = b_\nu^{-(r+m)/2} \left(\frac{\gamma \mathrm{e}^{-\nu^2/2}}{2\sqrt{2\pi}\Phi(\nu)} \right)^m \exp\left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} \delta_{ij} \mathrm{x}_{ij}^\top \boldsymbol{\beta} \right) \prod_{i=1}^{m} \Psi\left(\frac{r_i+1}{2}, \frac{c_{\boldsymbol{\psi}}^{(i)}}{\sqrt{b_\nu}} \right) \prod_{j=1}^{n_i} h_0(t_{ij})^{\delta_{ij}},$$

with $r = \sum_{i=1}^{m} r_i$ the total failures in the sample. In a parametric approach, $H_0(t)$ or $h_0(t)$ are specified by a set of parameters, say $\boldsymbol{\lambda}$, and then the parameter vector is reduced to $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu)$. For instance, for the Weibull (WEI) distribution, we use the parameterization $H_0(t) = \lambda t^\rho$ and $h_0(t) = \lambda \rho t^{\rho-1}$, where $t > 0$ and $\boldsymbol{\lambda} = (\lambda, \rho) \in \mathbb{R}_+^2$. From a classical approach, the ML estimator can be obtained by maximizing $\log L(\boldsymbol{\beta}, \boldsymbol{\lambda}, \nu)$ relative to $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$ and ν. For the flexibility discussed in previous sections, we also consider the PE model. However, it can be also attractive to discuss a non-parametric approach for the baseline distribution. For this, in the next subsection, we consider an estimation procedure based on the EM algorithm.

EM algorithm

Given the unobservable nature of the frailty terms, the EM algorithm is an ad hoc tool to be applied in this context. Let $\boldsymbol{t}_i^\top = (t_{i1}, ..., t_{in_i})$, $\boldsymbol{\delta}_i^\top = (\delta_{i1}, ..., \delta_{in_i})$ and $\boldsymbol{x}_i^\top = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i})$ the observed times, failure indicators and covariates, related to the n_i observations in the i-th cluster, $i = 1, \ldots, m$. For our particular problem, $\mathcal{D}_c = (\boldsymbol{t}^\top, \boldsymbol{\delta}^\top, \boldsymbol{X}^\top, \boldsymbol{Z}^\top)$ represents the complete data, where $\boldsymbol{t}^\top = (\boldsymbol{t}_1^\top, ..., \boldsymbol{t}_m^\top)$, $\boldsymbol{\delta}^\top = (\boldsymbol{\delta}_1^\top, ..., \boldsymbol{\delta}_m^\top)$, $\boldsymbol{X}^\top = (\boldsymbol{x}_i^\top, ..., \boldsymbol{x}_m^\top)$ and $\boldsymbol{Z}^\top = (z_1, ..., z_m)$, where $\mathcal{D}_o = (\boldsymbol{t}^T, \boldsymbol{\delta}^T, \boldsymbol{X}^T)$ is the observed data and $\boldsymbol{Z}^\top$ represents the vector of latent variables. Note that the complete likelihood function can be written as $L(\boldsymbol{\beta}, H_0, \nu; \mathcal{D}_c) = L_1(\boldsymbol{\beta}, H_0; \mathcal{D}_c) \times L_2(\nu; \boldsymbol{Z})$, where $L_1(\boldsymbol{\beta}, H_0; \mathcal{D}_c) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left[z_i h_0(t_{ij}) \exp(\mathrm{x}_{ij}^\top \boldsymbol{\beta}) \right]^{\delta_{ij}} \exp(-z_i H_0(t_{ij}) \mathrm{e}^{\mathrm{x}_{ij}^\top \boldsymbol{\beta}})$ and $L_2(\nu; \boldsymbol{Z}) = \prod_{i=1}^{m} f(z_i; \nu)$.

The complete log-likelihood function is given by $\ell_c(\boldsymbol{\beta}, H_0, \nu; \mathcal{D}_c) = \ell_{1c}(\boldsymbol{\beta}, H_0; \mathcal{D}_c) + \ell_{2c}(\nu; \boldsymbol{Z})$, where except for a constant that does not depend on $\boldsymbol{\beta}$, H_0 or ν, such functions are given by

$$\ell_{1c}(\boldsymbol{\beta}, H_0; \mathcal{D}_c) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ \delta_{ij} \left[\log h_0(t_{ij}) + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} \right] - z_i H_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) \right\}, \quad \text{and}$$

$$\ell_{2c}(\nu; \boldsymbol{Z}) = \sum_{i=1}^{m} \left\{ \log \gamma - \log \Phi(\nu) - \frac{1}{2} \log(2\pi) - \frac{1}{2} \left(z_i^2 \gamma^2 - 2\gamma\nu z_i + \nu^2 \right) \right\}.$$

Let $\boldsymbol{\psi}^{(k)} = \left(\boldsymbol{\beta}^{(k)}, H_0^{(k)}, \nu^{(k)} \right)$ be the estimated vector of $\boldsymbol{\psi} = (\boldsymbol{\beta}, H_0, \nu)$ at the k-th iteration and

$$Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(k)}) = \mathbb{E}\left(\ell_c(\boldsymbol{\beta}, H_0, \nu; \mathcal{D}_c) \mid \mathcal{D}_o, \boldsymbol{\psi} = \boldsymbol{\psi}^{(k)} \right),$$

i.e., the conditional expectation of $\ell_c(\boldsymbol{\beta}, H_0, \nu; \mathcal{D}_c)$ given the observed data and $\boldsymbol{\psi}^{(k)}$. Note that $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(k)}) = Q_1((\boldsymbol{\beta}, H_0) \mid \boldsymbol{\psi}^{(k)}) + Q_2(\nu \mid \boldsymbol{\psi}^{(k)})$

$$Q_1((\boldsymbol{\beta}, H_0) \mid \boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ \delta_{ij} \left[\log h_0(t_{ij}) + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} \right] - \widehat{z}_i^{(k)} H_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) \right\}, \quad \text{and} \tag{10}$$

$$Q_2(\nu \mid \boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{m} \left\{ \log \gamma - \log \Phi(\nu) - \frac{1}{2} \log(2\pi) - \frac{1}{2} \left(\widehat{z_i^2}^{(k)} \gamma^2 - 2\gamma\nu \widehat{z}_i^{(k)} + \nu^2 \right) \right\}, \tag{11}$$

where $\widehat{z}_i^{(k)} = \mathbb{E}\left[Z_i \mid \mathcal{D}_o, \boldsymbol{\psi} = \boldsymbol{\psi}^{(k)} \right]$ and $\widehat{z_i^2}^{(k)} = \mathbb{E}\left[Z_i^2 \mid \mathcal{D}_o, \boldsymbol{\psi} = \boldsymbol{\psi}^{(k)} \right]$. It is possible to show that

$$Z_i \mid \boldsymbol{t}_i^\top, \boldsymbol{\delta}_i^\top \sim \text{MHN} \left(a_i = 1 + r_i, b_\nu = \frac{\gamma^2}{2}, c_{\boldsymbol{\psi}}^{(i)} = \gamma\nu - \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) \right). \tag{12}$$

Refer to the supplementary material file for a proof of this fact. Using this notation, and applying Lemma 2 from Sun et al.[15], it follows immediately that

$$\widehat{z}_i^{(k)} = \mathbb{E}\left(Z_i \mid \mathcal{D}_o, \boldsymbol{\psi} = \boldsymbol{\psi}^{(k)}\right) = \frac{\Psi\left(\frac{r_i+1}{2}; \frac{c_\psi^{(i)}}{\sqrt{b_\nu}}\right)}{\sqrt{b_\nu}\Psi\left(\frac{r_i}{2}; \frac{c_\psi^{(i)}}{\sqrt{b_\nu}}\right)}, \quad \text{and}$$
$$\widehat{z}_i^{2(k)} = \mathbb{E}\left(Z_i^2 \mid \mathcal{D}_o, \boldsymbol{\psi} = \boldsymbol{\psi}^{(k)}\right) = \frac{\Psi\left(\frac{r_i+1}{2}; \frac{c_\psi^{(i)}}{\sqrt{b_\nu}}\right)}{b_\nu\Psi\left(\frac{r_i}{2}; \frac{c_\psi^{(i)}}{\sqrt{b_\nu}}\right)}. \tag{13}$$

On the other hand, it is possible to construct a discrete version of the cumulative baseline hazard function, considering $H_0^D(t) = \sum_{\ell:t_{(\ell)}\le t} H_0(t_{(\ell)})$, where $t_{(1)}, \ldots, t_{(q)}$ are the ordered distinct failure times and q is the number of different observed failure times. Replacing $H_0(\cdot)$ and $h_0(\cdot)$ in Equation (9) is obtained

$$Q_1((\boldsymbol{\beta}, H_0) \mid \boldsymbol{\psi}^{(k)}) = \sum_{\ell=1}^{q} d_{(\ell)} \log\left[h_0(t_{(\ell)})\right] + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \delta_{ij}\mathrm{x}_{ij}^\top\boldsymbol{\beta} - \sum_{\ell=1}^{q} h_0(t_{(\ell)}) \sum_{i,j\in R(t_{(\ell)})} \widehat{z}_i^{(k)} \mathrm{e}^{\mathrm{x}_{ij}^\top\boldsymbol{\beta}}.$$

Replacing the solution for $h_0(t_{(\ell)})$, i.e., $\widehat{h}_0(t_{(\ell)}) = d_{(\ell)} / \left[\sum_{i,j\in R(t_{(k)})} \exp\left(\mathrm{x}_{ij}^\top\boldsymbol{\beta} + \log\widehat{z}_i^{(\ell)}\right)\right]$, the expression for Q_1 is reduced, up to a constant that does not depend on $\boldsymbol{\beta}$, to

$$Q_1(\boldsymbol{\beta} \mid \boldsymbol{\psi}^{(k)}) = -\sum_{\ell=1}^{q} d_{(\ell)} \log\left(\sum_{i,j\in R(t_{(\ell)})} \exp\left(\mathrm{x}_{ij}^\top\boldsymbol{\beta} + \log\widehat{z}_i^{(k)}\right)\right) + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \delta_{ij}\mathrm{x}_{ij}^\top\boldsymbol{\beta}.$$

Note that $Q_1(\cdot)$ has the same form of the partial log-likelihood function of the Cox model, except for the offset $\log\widehat{z}_i^{(k)}$. For this, to update $\boldsymbol{\beta}$ in the M-step we can use the Cox approach. Finally, the non-parametric estimator for $H_0(\cdot)$ in the k-th step of the algorithm is given by

$$\widehat{H}_0^{(k)}(t) = \sum_{\ell:t_{(\ell)}\le t} \frac{d_{(\ell)}}{\sum_{i,j\in R(t_{(\ell)})} \exp\left(\mathrm{x}_{ij}^\top\boldsymbol{\beta}^{(k)} + \log\widehat{z}_i^{(k)}\right)}, \quad t > 0.$$

In summary, the EM algorithm is given by the following steps.

- **E-step**: For $i = 1, ..., m$, compute $\widehat{z}_i^{(k+1)}$ and $\widehat{z_i^2}^{(k+1)}$ using equations (12) and (13), respectively, with $\boldsymbol{\beta}^{(k)}$, $H_0(\cdot)^{(k)}$ and $\nu^{(k)}$ as the estimated parameters at the k-th iteration.
- **M1-step**: Update $\boldsymbol{\beta}^{(k+1)}$ and $H_0^{(k+1)}(\cdot)$ by fitting a Cox regression model with offset $\log\widehat{z}_i^{(k+1)}$ for the nonparametric case, or maximizing $Q_1(\beta, H_0)$ for the parametric (WEI) and semi-parametric (PE) cases.
- **M2-step**: Update $\nu^{(k+1)}$ by maximizing $Q_2(\nu \mid \boldsymbol{\psi}^{(k)})$ in relation to ν.

Maximization around H_0 refers to optimizing the parameters in $H_0(\cdot)$: ρ and λ for the Weibull baseline distribution, or the vector $\boldsymbol{\lambda}$ for the piecewise exponential case. The unified formulation ensures algorithmic generality. The algorithm iterates until a convergence criterion is satisfied. For instance, we consider $||\widehat{\boldsymbol{\psi}}^{(k-1)} - \widehat{\boldsymbol{\psi}}^{(k)}|| < \epsilon$, where ϵ is a predefined value and $||\cdot||$ denotes the Euclidean norm. Initial values are derived from the ordinary Cox model, taking $\nu^{(0)} = 0.5$. On the other hand, following the suggestion of[20], we estimate the standard error of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\nu}$ via a profile log-likelihood function: $\ell(\boldsymbol{\beta}, \nu) = \log L(\boldsymbol{\beta}, H_0, \nu)$, replacing H_0 with its estimate $\widehat{H}_0$. The variance-covariance matrix of $(\widehat{\boldsymbol{\beta}}, \widehat{\nu})$ is then:

$$I(\widehat{\boldsymbol{\beta}}, \widehat{\nu}) = -\frac{\partial^2\ell(\boldsymbol{\beta}, \nu)}{\partial(\boldsymbol{\beta}, \nu)\partial^\top(\boldsymbol{\beta}, \nu)}\bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \nu=\widehat{\nu}}.$$

Finally, more important than $\widehat{\nu}$ is $\widehat{\theta} := \widehat{\gamma}^{-2} - \frac{\phi(\widehat{\nu})}{\Phi(\widehat{\nu})}\widehat{\gamma}^{-1}$ (the frailty variance) because allows us to compare this term with the variance of other models parameterized directly in the frailty variance. The variance of $\widehat{\theta}$ is estimated as:

$$\widehat{Var(\theta)} = \widehat{Var(\nu)}\left[\frac{\phi(\widehat{\nu})}{\Phi(\widehat{\nu})}\widehat{\gamma}^{-2}\left(1 - \frac{\phi(\widehat{\nu})}{\Phi(\widehat{\nu})}\widehat{\gamma}\right) - 2\widehat{\gamma}^{-3}\left(1 - \frac{\phi(\widehat{\nu})}{\Phi(\widehat{\nu})}\widehat{\gamma}\right) + \frac{\phi(\widehat{\nu})}{\Phi(\widehat{\nu})}\right]^2.$$

Remark 1 Note that the result in Equation (11) is also interesting if a Bayesian approach were applied to the model, because also is valid conditioning on the parameters. This facilitates, among other things, the application of an MCMC type method to simulate from the corresponding conditional distribution related to the frailties.

Computational aspects

The `extrafrail`[21] package of R[22] includes the computational implementation for the TN frailty model considering as the baseline model the Weibull, exponential and PE distributions and the non-parametric specification. For instance, to fit the Weibull case, it can be used

```
frailty.fit(formula, data, dist = "weibull", dist.frail="TN")
```

whereas is usually in survival analysis with random effects in R, the `formula` can be defined as

```
Surv(time, event) ~covariates + cluster(id)
```

A similar syntax can be used to fit the other cases specifying `dist="exponential"`, `dist="pe"` or `dist="np"` in the last sentence. We highlight that the function allows us to perform the estimation even for the case where the clusters have different sizes (i.e., $n_1, n_2, \ldots, n_m$ are not necessarily the same).

Simulation study

In this Section, we present a simulation study to assess the performance of the maximum likelihood estimators obtained via the EM algorithm with samples of different percentages of censoring.

Recovery parameters

We consider the following three different scenarios:

- **Scenario 1**: 19 clusters with 2 observations each and 19 clusters with 4 observations each, totalling 114 observations. ($n_1 = \ldots = n_{19} = 2, n_{20} = \ldots = n_{38} = 4$ and $m = 38$).
- **Scenario 2:** 38 clusters with 2 observations each and 38 clusters with 4 observations each, totalling 228 observations. ($n_1 = \ldots = n_{38} = 2, n_{39} = \ldots = n_{76} = 4$ and $m = 76$).
- **Scenario 3:** 19 clusters with 4 observations each and 19 clusters with 8 observations each, totalizing 228 observations. ($n_1 = \ldots = n_{19} = 4, n_{20} = \ldots = n_{38} = 8$ and $m = 38$).

The idea is to verify if, under a certain amount of data, it is advisable to increase the number of clusters or increase the cluster observations. We consider as baseline model the PE distribution with $L = 3$ and time partition $\boldsymbol{a} = (7/365, 56/365)$. Similar to the real data application, we also consider one dichotomous covariate *x*, which was drawn from the Bernoulli distribution with success probability 20/76. We also consider three values for θ, the variance of the frailty terms: 0.20, 0.50 and 0.75. The percentage of censoring was fixed at 10%, 25% and 50%. In all the cases, the regression coefficient was fixed as $\beta = 1.8$ and the parameters from the PE distribution were fixed as $\boldsymbol{\lambda} = (\lambda_1 = 0.3, \lambda_2 = 2.6, \lambda_3 = 1.9)$. To simulate values from the model, we use the following steps:

i) Draw $z_i \sim \text{TN}\ (\nu)$, $i = 1, \ldots, m$, using the inverse transform method, i.e., do $z_i = \left(\Phi^{-1}\left(u_i\Phi(\nu) + \Phi(-\nu)\right) + \nu\right)\gamma^{-1}$, where $u_i \sim \text{U}\ (0, 1)$ (the standard uniform distribution).
ii) Draw the failure times from the conditional distribution $y_{ij} \mid z_i \sim \text{PE}\ (\boldsymbol{\lambda} z_i \exp(\mathrm{x}_{ij}^{\top}\boldsymbol{\beta}), \boldsymbol{a})$.
iii) Define the censoring times, c_{ij}, as the $100 \times (1 - q)$-th quantile of the corresponding conditional distribution $\text{PE}\ (\boldsymbol{\lambda} z_i \exp(\mathrm{x}_{ij}^{\top}\boldsymbol{\beta}), \boldsymbol{a})$ distribution.
iv) Define the observed failure times and failure indicators as $t_{ij} = \min(y_{ij}, c_{ij})$ and $\delta_{ij} = I(y_{ij} \leq c_{ij})$, respectively, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$.

For each scenario and combination of censoring and θ, we draw 1,000 samples and compute the ML estimates. For each parameter, Tables 1 and 2 summarized the average bias (bias), the root of the estimated mean squared error (RMSE), the mean of the standard errors (SE) and the coverage probabilities (CP) of the asymptotic 95% confidence intervals.

An increase in the sample size improves the precision and accuracy of the estimates. In particular, scenarios 2 and 3, which have larger sample sizes, exhibit better performance than scenario 1. In general, an increase in heterogeneity (θ) and in the censoring percentage tends to raise the bias, standard error, and RMSE, while reducing coverage probability (CP). However, the behavior of the estimator for θ improves under higher censoring, showing reduced bias and increased coverage, possibly due to a better identification of the random effect in the presence of censored events. The most affected estimator is λ_3, since censored information tends to concentrate within its interval. When comparing Scenarios 2 and 3, the former yields better results. This suggests that for a fixed total sample size, increasing the number of clusters is preferable to increasing the number of observations per cluster. This leads to greater diversity in latent effects, which enhances the estimation of frailty terms.

Applications with real data sets

In this Section, we present two applications to illustrate the performance of the TN frailty model in comparison with traditional models. The first application is related to patients with Chronic Kidney Disease (CKD), while the second application is related to patients with fibrosarcoma.

Kidney data set

CKD is the slow and progressive loss of kidney function over time. The main job of these organs is to remove waste and excess water from the body. This disease may be asymptomatic for some time until the kidneys have

Censoring	θ	Parameter	Scenario 1				Scenario 2				Scenario 3			
			Bias	RMSE	SE	CP	Bias	RMSE	SE	CP	Bias	RMSE	SE	CP
10%	0.20	θ	−0.003	0.109	0.106	0.894	0.003	0.074	0.073	0.922	−0.011	0.081	0.076	0.887
		β	0.005	0.299	0.286	0.941	−0.026	0.213	0.201	0.940	−0.020	0.201	0.193	0.945
		λ_1	0.076	0.243	0.261	0.941	0.019	0.171	0.166	0.863	0.028	0.177	0.173	0.875
		λ_2	0.053	0.645	0.599	0.877	0.071	0.460	0.413	0.854	0.049	0.483	0.445	0.874
		λ_3	−0.073	0.462	0.414	0.857	−0.089	0.331	0.279	0.819	−0.059	0.333	0.290	0.848
	0.50	θ	−0.076	0.204	0.210	0.806	−0.044	0.150	0.152	0.871	−0.077	0.173	0.160	0.793
		β	−0.053	0.322	0.287	0.913	−0.046	0.238	0.193	0.877	−0.044	0.206	0.188	0.918
		λ_1	0.086	0.265	0.203	0.708	0.026	0.193	0.096	0.494	0.024	0.191	0.129	0.655
		λ_2	0.058	0.741	0.457	0.615	0.004	0.511	0.224	0.425	0.022	0.558	0.341	0.604
		λ_3	−0.224	0.595	0.377	0.601	−0.237	0.453	0.197	0.460	−0.186	0.448	0.255	0.598
	0.75	θ	−0.226	0.307	0.26	0.695	−0.172	0.237	0.202	0.750	−0.182	0.261	0.220	0.726
		β	−0.082	0.329	0.282	0.888	−0.092	0.246	0.193	0.859	−0.047	0.213	0.186	0.920
		λ_1	0.080	0.265	0.162	0.569	0.039	0.203	0.068	0.345	0.025	0.191	0.105	0.546
		λ_2	−0.081	0.790	0.357	0.464	−0.048	0.538	0.15	0.278	−0.047	0.642	0.285	0.456
		λ_3	−0.419	0.708	0.300	0.443	−0.413	0.578	0.153	0.304	−0.274	0.522	0.226	0.463
25%	0.20	θ	0.038	0.149	0.137	0.937	0.030	0.102	0.089	0.942	0.006	0.095	0.087	0.915
		β	−0.052	0.344	0.307	0.917	−0.053	0.234	0.216	0.922	−0.044	0.225	0.210	0.940
		λ_1	0.099	0.258	0.285	0.977	0.027	0.226	0.179	0.903	0.030	0.199	0.182	0.919
		λ_2	0.086	0.673	0.647	0.924	0.066	0.467	0.455	0.915	0.047	0.491	0.482	0.932
		λ_3	−0.119	0.548	0.478	0.852	−0.180	0.401	0.325	0.817	−0.154	0.400	0.330	0.829
	0.50	θ	−0.048	0.200	0.239	0.856	0.001	0.171	0.187	0.915	−0.045	0.187	0.185	0.830
		β	−0.105	0.353	0.307	0.903	−0.103	0.261	0.217	0.886	−0.070	0.230	0.209	0.919
		λ_1	0.107	0.277	0.248	0.823	0.037	0.184	0.141	0.717	0.044	0.187	0.168	0.831
		λ_2	0.060	0.771	0.565	0.736	0.045	0.519	0.362	0.659	0.042	0.568	0.468	0.792
		λ_3	−0.416	0.687	0.415	0.606	−0.398	0.554	0.274	0.508	−0.295	0.519	0.331	0.658
	0.75	θ	−0.195	0.288	0.298	0.771	−0.119	0.218	0.243	0.838	−0.173	0.254	0.241	0.763
		β	−0.126	0.370	0.309	0.901	−0.117	0.272	0.216	0.876	−0.099	0.242	0.208	0.915
		λ_1	0.104	0.277	0.229	0.762	0.050	0.207	0.122	0.590	0.040	0.201	0.149	0.720
		λ_2	0.003	0.817	0.531	0.653	−0.025	0.541	0.299	0.531	−0.025	0.659	0.426	0.659
		λ_3	−0.600	0.813	0.366	0.459	−0.561	0.672	0.225	0.353	−0.411	0.594	0.298	0.508

Table 1. Estimated bias, RMSE, SE and approximated 95% coverage probabilities for the TN frailty model with basal distribution PE under different scenarios (cases censoring 10% and 25%).

almost stopped working, whereupon kidney disease usually subsides, diagnosed in its final stages. The final stage of CKD is called End-Stage Renal Disease (ESRD). At this stage, the kidneys can no longer sufficiently remove waste and excess fluid from the body, requiring the patient to undergo dialysis (a life-sustaining treatment) or a kidney transplant (US National Library of Medicine). Dialysis is broken down into two main modalities: hemodialysis and peritoneal dialysis. Hemodialysis consists of extracting blood from the body to direct it to a machine that eliminates waste and excess fluid; after filtration, it is reintroduced into the bloodstream. Peritoneal dialysis, for its part, is a simpler process and can be done on an outpatient basis. Liquid is inserted into the peritoneal cavity through a catheter located in the stomach. This solution absorbs waste and excess fluid and is later extracted. The solution is removed through the same channel.

CKD represents one of the most important non-communicable diseases worldwide[23]. For many patients, dialysis is the focal point around which their lives revolve, not only because of the time spent travelling to and from the sessions in specialized centres and the time dedicated to the dialysis treatment itself but also due to the diet that accompanies it, fluid restrictions and medication load[24]. Thus, one of the most advantageous options, considering quality of life, is treatment by ambulatory peritoneal dialysis (with a portable machine). The peritoneal catheter is a foreign body that facilitates the appearance of infections and serves as a reservoir for bacteria. Infection can appear both in the exit orifice and the tunnel (tunnelled path of the catheter) or the peritoneum (peritonitis). Peritonitis continues to be an important complication of PD, as it contributes to technique failure, hospitalization, and even death[25].

We focus on a real dataset named *kidney*, available in the R[22] package *frailtyHL*[26]. For further details, see page 11 of its documentation: https://cran.r-project.org/web/packages/frailtyHL/frailtyHL.pdf. The study collected bivariate times, consisting of the times of first and second recurrence of infection at the catheter insertion point in patients with kidney problems using a portable dialysis machine. The catheter is later removed if infection

Censoring	θ	Parameter	Scenario 1				Scenario 2				Scenario 3			
			Bias	RMSE	SE	CP	Bias	RMSE	SE	CP	Bias	RMSE	SE	CP
50%	0.20	θ	0.054	0.199	0.191	0.904	0.064	0.155	0.131	0.950	0.007	0.107	0.107	0.928
		β	−0.116	0.377	0.352	0.944	−0.113	0.286	0.248	0.925	−0.092	0.267	0.241	0.926
		λ_1	0.121	0.279	0.309	0.994	0.046	0.194	0.196	0.947	0.044	0.198	0.194	0.941
		λ_2	0.086	0.667	0.683	0.964	0.032	0.468	0.480	0.951	−0.010	0.492	0.491	0.930
		λ_3	−0.533	0.841	0.588	0.682	−0.576	0.717	0.405	0.596	−0.486	0.660	0.421	0.674
	0.50	θ	−0.038	0.241	0.302	0.861	0.030	0.203	0.243	0.912	−0.041	0.199	0.214	0.840
		β	−0.191	0.446	0.366	0.893	−0.191	0.342	0.260	0.863	−0.141	0.294	0.249	0.908
		λ_1	0.135	0.304	0.315	0.971	0.060	0.200	0.200	0.933	0.057	0.217	0.198	0.937
		λ_2	−0.053	0.729	0.708	0.894	−0.093	0.534	0.504	0.868	−0.082	0.596	0.552	0.883
		λ_3	−0.917	1.059	0.485	0.443	−0.884	0.971	0.353	0.335	−0.716	0.855	0.407	0.463
	0.75	θ	−0.205	0.316	0.355	0.776	−0.133	0.247	0.290	0.822	−0.182	0.273	0.270	0.754
		β	−0.247	0.483	0.374	0.865	−0.232	0.366	0.263	0.842	−0.162	0.307	0.25	0.896
		λ_1	0.168	0.328	0.325	0.945	0.089	0.233	0.206	0.911	0.051	0.203	0.194	0.911
		λ_2	−0.219	0.794	0.675	0.816	−0.196	0.593	0.499	0.824	−0.135	0.641	0.555	0.852
		λ_3	−1.076	1.178	0.412	0.296	−1.047	1.103	0.307	0.193	−0.798	0.902	0.381	0.408

Table 2. Estimated bias, RMSE, SE and approximated 95% coverage probabilities for the TN frailty model with basal distribution PE under different scenarios (case censoring 50%).

	n	min	Q_1	median	mean	Q_3	max
TR_1	38	0.005	0.042	0.126	0.306	0.403	1.468
TR_2	38	0.011	0.049	0.107	0.251	0.395	1.540

Table 3. Summary of the first and second time of recurrence (TR_1 and TR_2).

occurs and can be removed for other reasons, in which case the observation is censored. Available covariates are sex and type of kidney disease: Glomerulonephritis (GN), acute nephritis (AN), Polycystic kidney disease (PKD) and others. Previous analysis suggests that only sex is significant in this context[27]. The study has 38 patients, 10 men and 28 women, each person has 2 times of recurrence of the infection, so there are a total of 76 observations. A summary of such times is presented in Table 3 and Figure 3 presents the Kaplan-Meier (KM) estimator by both times and by sex.

For comparison purposes, we also consider the GA, WL and IG frailty models with baseline distribution WE and PE. Figure 4 shows the cumulative hazard function for the kidney data. The proposed partition for the PE model was set at 1 and 8 weeks (indicated by the vertical segments in the graph). A change in the slope behavior is evident, as highlighted in the zoomed-in view on the right. This supports the conclusion that the PE model provides a better fit than a non-segmented model for this dataset. Practically, this suggests that the risk of infection at the catheter insertion site is highest during the first week post-insertion and gradually decreases over time. After two months, the risk stabilizes and remains relatively low. This understanding can help healthcare professionals in identifying critical time periods for infection prevention and monitoring patients accordingly.

Table 4 shows the Akaike information criterion (AIC)[28] and the Bayesian information criterion (BIC)[29] for such models. According to the AIC and BIC criteria, it is suggested that the baseline PE model is more appropriate for this data than the WE model, independent of the frailty model used. However, the TN frailty model provides better results. Table 5 presents the estimates for all the models considering the PE baseline distribution, including the ordinary PE model (i.e., without frailty).

Note that the effect of not considering the dependence among the clusters is the underestimation of the effect for sex. On the other hand, the estimated Kendall's τ for the different models is around 0.13. However, the estimated frailty variance for GA, WL and IG is overestimated by at least 50% concerning the frailty TN model. In practical terms, this means that the TN frailty model estimates a greater effect of sex on the recurrence of infection at the catheter insertion point and less variability between the measures associated with the same individual.

Fibrosarcoma data set

Fibrosarcoma is a rare malignant tumor that originates from fibroblasts, the connective tissue cells responsible for the production of collagen and extracellular matrix. This neoplasm exhibits infiltrative growth, a high propensity for local recurrence, and metastatic potential. It can develop in any part of the body, although it is most commonly found in the extremities, trunk, and retroperitoneal region. Clinically, it typically presents as a progressively enlarging mass, initially painless. Diagnosis is based on histopathological findings, where tumor

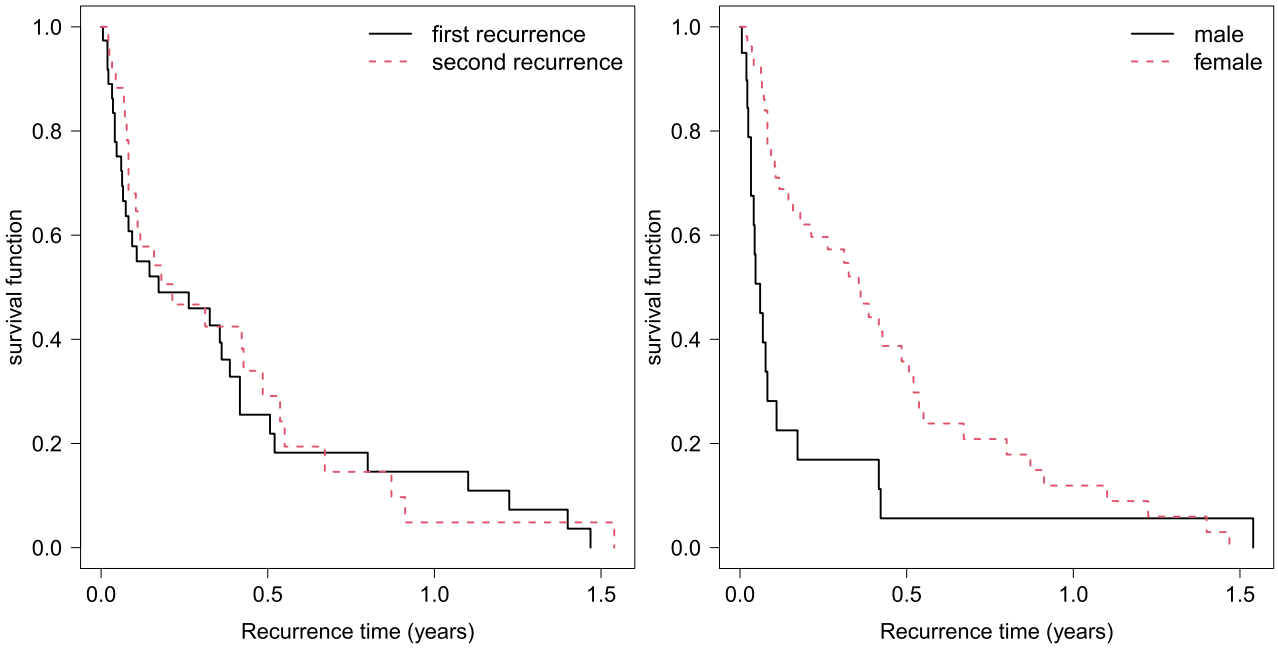

Fig. 3. KM estimator for kidney K-M for kidney dataset considering recurrence time (left panel) and sex (right panel).

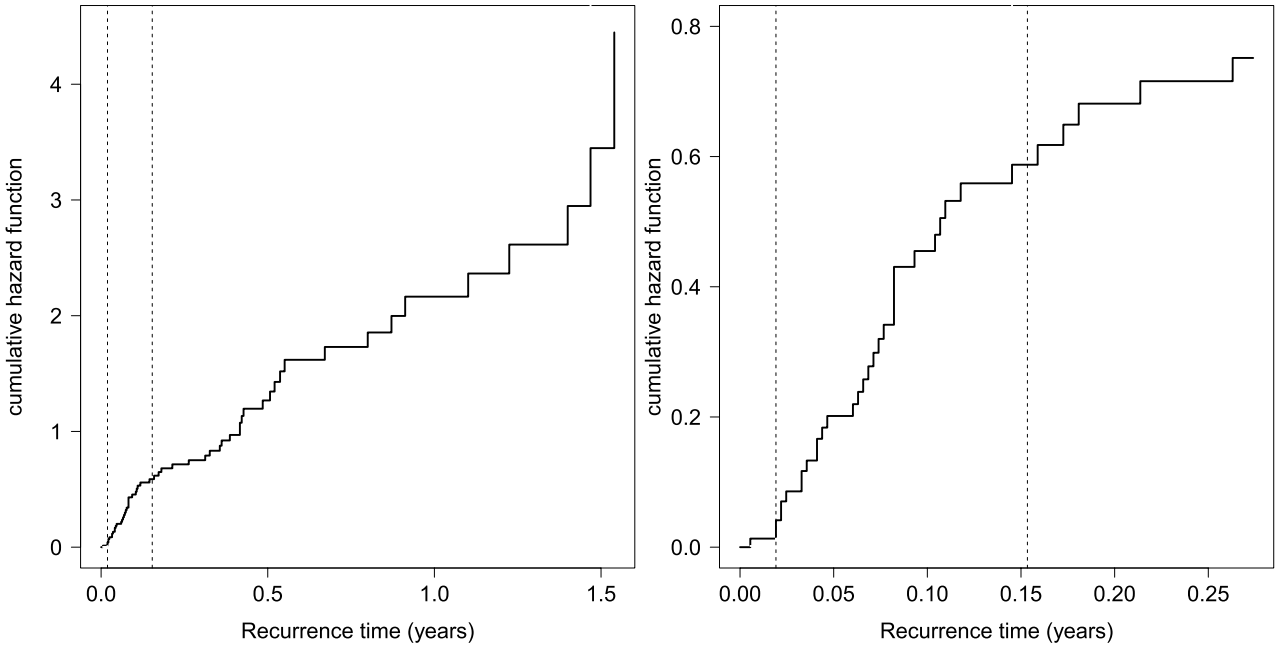

Fig. 4. Cumulative hazard function for kidney dataset considering all the axis time (left panel) and zoom for the 100 first days (right panel).

cells are arranged in a characteristic herringbone pattern, and is often supported by immunohistochemical studies to differentiate it from other soft tissue tumors[30]. The treatment of choice is surgical excision with wide margins, and adjuvant radiotherapy is frequently considered; chemotherapy is generally reserved for advanced or metastatic cases[31].

This dataset includes information from 251 patients diagnosed with fibrosarcoma SOE (from the portugues "sem outra especificacão, meaning "not otherwise specified") with diagnosis dates ranging from 2000 to 2022, and follow-up data extending through December 2022. The dataset was obtained from the Oncocenter Foundation of São Paulo, Brazil (Fundacão Oncocentro de São Paulo, FOSP), which oversees the Hospital Cancer Registry

	TN		GA		WL		IG		-
	WE	PE	WE	PE	WE	PE	WE	PE	PE
log-Like	10.230	**14.786**	9.8384	14.289	9.8914	14.321	8.7783	13.676	11.544
AIC	−12.460	**-19.573**	−11.677	−18.577	−11.783	−18.642	−9.5566	−17.353	−15.088
BIC	−3.1371	**-7.9192**	−2.3539	−6.9236	−2.4599	−6.9885	−0.2337	−5.6991	−5.7649

Table 4. Maximized log-likelihood function (log-Like), AIC and BIC of the TN, GA, WL and IG models for kidney dataset.

Parameter	TN	GA	WL	IG	Without frailty
β_{sex}	1.763 (0.448)	1.644 (0.467)	1.658 (0.470)	1.417 (0.408)	0.935 (0.284)
λ_1	0.328 (0.339)	0.344 (0.357)	0.341 (0.355)	0.384 (0.396)	0.505 (0.509)
λ_2	3.214 (0.808)	3.421 (0.874)	3.406 (0.872)	3.677 (0.921)	3.801 (0.785)
λ_3	2.217 (0.551)	2.377 (0.673)	2.376 (0.667)	2.365 (0.747)	1.689 (0.350)
θ	0.191 (0.111)	0.333 (0.194)	0.328 (0.183)	0.399 (0.341)	-
τ	0.118	0.143	0.143	0.130	-

Table 5. Estimates, standard errors (in parenthesis) and Kendall's τ for the TN, GA, WL and IG frailty model with baseline PE and the ordinary PE model for kidney dataset.

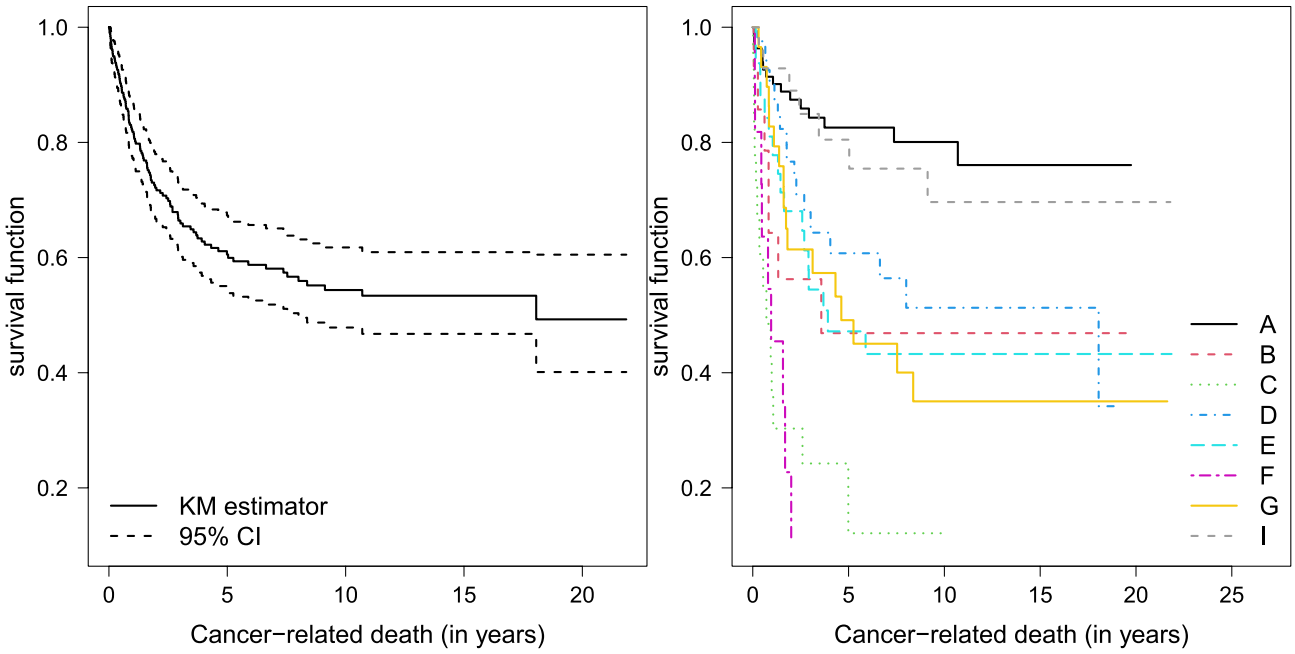

Fig. 5. Kaplan-Meier estimator for the fibrosarcoma data with 95% confidence interval (left panel) and stratified by treatment received (right panel).

of the State of São Paulo (http://fosp.saude.sp.gov.br). This neoplasm is coded as 8810/3 Fibrosarcoma, NOS (not otherwise specified), according to the International Classification of Diseases for Oncology (ICD-O[32]), which is used in cancer registries to classify tumors that lack further histological subtyping at the time of diagnosis.

Cancer-specific death was defined as the event of interest, and time-to-event was measured from the date of diagnosis to the patient's death (in years: mean= 5.72, standard deviation (SD)= 5.78, median= 3.12, range= $0.025 - 21.86$). During the follow-up period, a total of 103 events (39%) occurred. As covariates we use the type of treatment, with eight possible labels: A - surgery (84 patients, 32.2%), B - Radiotherapy (14 patients, 5.4%), C - Chemotherapy (18 patients, 6.9%), D - Surgery + Radiotherapy (42 patients, 16.1%), E - Surgery + Chemotherapy (34 patients, 13.0%), F - Radiotherapy + Chemotherapy (11 patients, 4.2%), G - Surgery + Radiotherapy + Chemotherapy (29 patients, 11.1%) and I - other combination (29 patients, 11.1%). Figure 5 presents the KM estimator by both times and type of treatment. The clusters considered in this analysis correspond to the 26 clinical areas responsible for treating the patients, which are summarized in Table 6. Note

Allergy – immunology (1)	Cardiac surgery (1)	Head and neck surgery (1)	General surgery (1)
Pediatric surgery (1)	Plastic surgery (2)	Thoracic surgery (2)	Vascular surgery (2)
Medical clinic (2)	Dermatology (3)	Endocrinology (3)	Gastrosurgery (3)
Gastroenterology (3)	Geriatrics (4)	Ginecology (5)	Gynecology – Obstetrics (6)
Hematology (8)	Infectology (8)	Nefrology (11)	Neurosurgery (12)
Neurology (12)	Ophthalmology (13)	Surgical Oncology (14)	Clinical Oncology (32)
Pediatric Oncology (34)	Orthopedics (77)		

Table 6. Number of records per medical specialty (cluster size in parenthesis).

Parameter	TN	GA	WL	IG	Without frailty
β_B	1.163 (0.471)	1.143 (0.476)	1.140 (0.476)	1.168 (0.478)	1.284 (0.460)
β_C	2.664 (0.401)	2.638 (0.407)	2.643 (0.408)	2.578 (0.402)	2.450 (0.381)
β_D	0.767 (0.364)	0.749 (0.367)	0.747 (0.367)	0.769 (0.369)	0.939 (0.358)
β_E	1.422 (0.367)	1.408 (0.374)	1.414 (0.374)	1.345 (0.372)	1.169 (0.358)
β_F	2.749 (0.454)	2.772 (0.457)	2.775 (0.457)	2.754 (0.453)	2.651 (0.437)
β_G	1.155 (0.361)	1.150 (0.362)	1.148 (0.362)	1.169 (0.363)	1.281 (0.358)
β_I	0.106 (0.465)	0.114 (0.467)	0.113 (0.467)	0.131 (0.466)	0.174 (0.460)
λ	0.058 (0.018)	0.059 (0.019)	0.059 (0.019)	0.060 (0.019)	0.061 (0.017)
ρ	0.655 (0.055)	0.654 (0.055)	0.655 (0.055)	0.647 (0.055)	0.627 (0.053)
θ	0.226 (0.132)	0.337 (0.250)	0.346 (0.248)	0.271 (0.289)	-
τ	0.139	0.144	0.150	0.099	-
AIC	645.1	647.4	647.3	649.0	650.1
BIC	680.7	683.1	682.9	684.7	682.1

Table 7. Parameter estimates, standard errors (in parentheses), and Kendall's τ for TN, GA, WL, and IG frailty models assuming a Weibull baseline hazard.

that these clusters are highly unbalanced in terms of sample size. In this analysis, we consider the TN, GA, WL, and IG frailty models, using the Weibull distribution for the baseline hazard. The results are summarized in Table 7. Notably, the TN frailty model provides the lowest AIC among the models considered. Once again, the Kendall's τ values provided by the models are similar. However, the estimated intra-cluster variance (0.226) is lower for the TN model compared to the others. Finally, Figure 6 shows the survival functions (SF) for patients treated in neurology and clinical oncology centers, as well as the marginal SF (i.e., the SF for a patient randomly selected from the entire cohort).

Concluding remarks

A new survival model with TN frailty was proposed and studied in detail. This model can lead to a complex structure for the data, because allows to modelling of univariate and multivariate data, being adaptable even for groups of different sizes. For the baseline risk, the Weibull, and PE distributions were adopted as well as a non-parametric approach. For a fixed variance for the frailty, the TN frailty model provides a greater Kendall's than the gamma and IG frailty models. We get a recursive closed-form expression for the derivatives of the Laplace transform for the TN model. Furthermore, the conditional distributions of frailties among the survivors and the frailty of individuals dying at time t were determined explicitly. The simulation studies, based on the EM algorithm, conclude that having more complete information relative to the censored information improves the accuracy and precision of the estimate. Scenarios 2 and 3 did not have a large difference in bias, this suggests that the bias depends on the sample size, not on the data configuration. On the other hand, concerning the RMSE and SE, Scenario 2 showed an improvement in precision for Scenario 3. This suggests that increasing the information in the clusters increases the precision compared to having clusters with little information but more numerous. We fitted the proposed frailty model to a real dataset on times to the first and second recurrence of infection at the catheter insertion point in patients with kidney problems using a portable dialysis machine to show the potential of using the new frailty model. This application demonstrates the practical relevance of the new regression model. In particular, the estimated frailty variance for GA, WL and IG is overestimated in the frailty TN model.

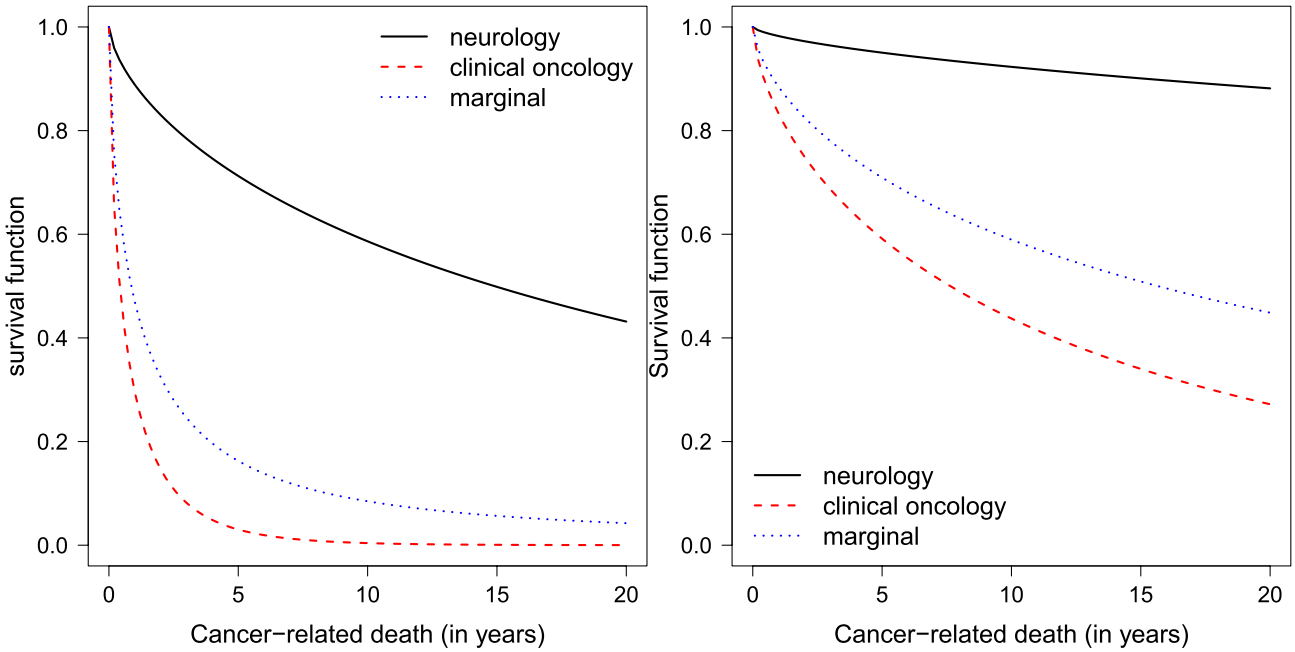

Fig. 6. SF for patients treated with chemotherapy (left panel) and surgery + radiotherapy (right panel), in medical specialty neurology and clinical oncology, as well as the marginal SF.

Data availability

The real dataset used, named kidney, is available in the frailtyHL package in R. For details on its use, refer to page 11 of the manual: https://cran.r-project.org/web/packages/frailtyHL/frailtyHL.pdf.

Received: 27 April 2025; Accepted: 11 August 2025
Published online: 17 August 2025

References

1. Vaupel, J., Manton, K. & Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454 (1979).
2. Congdon, P. Modelling frailty in area mortality. *Statistics in Medicine* **14**, 1859–1874 (1995).
3. Hougaard, P. Life table methods for heterogeneous populations. *Biometrika* **71**, 75–83 (1984).
4. Manton, K., Stallard, E. & Vaupel, J. Alternative models for heterogeneity of mortality risks among the aged. *Journal of the American Statistical Association* **81**, 635–644 (1986).
5. Leão, J., Leiva, V., Saulo, H. & Tomazella, V. Birnbaum-Saunders frailty regression models: Diagnostics and application to medical data. *Biometrical journal* **59**, 291–317 (2017).
6. Gallardo, D. I., Bourguignon, M. & Romeo, J. S. Birnbaum-saunders frailty regression models for clustered survival data. *Statistics and Computing* **34**, 141 (2024).
7. Wang, Y. & Emura, T. Multivariate failure time distributions derived from shared frailty and copulas. *Japanese Journal of Statistics and Data Science* **4**, 1105–1131 (2021).
8. Mota, A. et al. Weighted lindley frailty model: estimation and application to lung cancer data. *Lifetime Data Analysis* **27**, 561–587 (2021).
9. Gallardo, D. I., Bourguignon, M. & Santibáñez, J. L. The shared weighted lindley frailty model for clustered failure time data. *Biometrical Journal* **67**, e70044 (2025).
10. Kiprotich, G., Gallardo, D. I., Ramos, P. L. & Augustin, T. A shared frailty regression model for clustered survival data. *Statistical Methods in Medical Research* **0**, 09622802251338984, https://doi.org/10.1177/09622802251338984 (0).
11. Barreto-Souza, W. & Mayrink, V. Semiparametric generalized exponential frailty model for clustered survival data. *Annals of the Institute of Statistical Mathematics* **71**, 679–701 (2019).
12. Piancastelli, L., Barreto-Souza, W. & Mayrink, V. Generalized inverse-Gaussian frailty models with application to TARGET neuroblastoma data. *Annals of the Institute of Statistical Mathematics* **73**, 979–1010 (2021).
13. Gómez, H. J., Olmos, N. M., Varela, H. & Bolfarine, H. Inference for a truncated positive normal distribution. *Applied Mathematics-A Journal of Chinese Universities* **33**, 163–176 (2018).
14. Elbers, C. & Ridder, G. True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies* **49**, 403–409 (1982).
15. Sun, J., Kong, M. & Pal, S. The modified-half-normal distribution: Properties and an efficient sampling scheme. *Communications in Statistics-Theory and Methods* **52**, 1591–1613 (2023).
16. Wienke, A. *Frailty models in survival analysis* (CRC press, 2010).
17. Feigl, P. & Zelen, M. Estimation of exponential survival probabilities with concomitant information. *Biometrics* 826–838 (1965).
18. Friedman, M. Piecewise exponential models for survival data with covariates. *The Annals of Statistics* **10**, 101–113 (1982).
19. Balakrishnan, N. & Liu, K. Semi-parametric likelihood inference for birnbaum-saunders frailty model. *REVSTAT* 231–255 (2018).
20. Klein, J. P. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics* **48**, 795–806 (1992).
21. Gallardo, D., Bourguignon, M. & Santibáñez, J. *Estimation and Additional Tools for Alternative Shared Frailty Models* (2025). R package version 1.13.

22. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2025).
23. Torres, R. S. L., Ushiña, J., Lloay, A. & Balseca, M. Cuidados de enfermería en pacientes con enfermedad renal crónica en hemodiálisis durante infección por covid-19. *RECIAMUC* **6**, 81–90 (2003).
24. Davenport, A. Portable and wearable dialysis devices for the treatment of patients with end-stage kidney failure: Wishful thinking or just over the horizon?. *Pediatric nephrology* **30**, 2053–2060 (2015).
25. Fariñas, M. C., García-Palomo, J. D. & Gutiérrez-Cuadra, M. Infecciones asociadas a los catéteres utilizados para la hemodiálisis y la diálisis peritoneal. *Enfermedades infecciosas y microbiologia Clinica* **26**, 518–526 (2008).
26. Ha, I. D., Noh, M., Kim, J. & Lee, Y. *frailtyHL: Frailty Models via Hierarchical Likelihood* (2019). R package version 2.3.
27. McGilchrist, C. & Aisbett, C. Regression with frailty in survival analysis. *Biometrics* 461–466 (1991).
28. Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716–723 (1974).
29. Schwarz, G. Estimating the dimension of a model. *The annals of statistics* 461–464 (1978).
30. Fletcher, C. D. M., Bridge, J. A., Hogendoorn, P. C. W. & Mertens, F. *WHO Classification of Tumours of Soft Tissue and Bone* (IARC Press, Lyon, 2020), 5 edn.
31. Pisters, P. W., Leung, D. H., Woodruff, J., Shi, W. & Brennan, M. F. Analysis of prognostic factors in 1,041 patients with localized soft tissue sarcomas of the extremities. *Journal of Clinical Oncology* **25**, 785–790. https://doi.org/10.1200/JCO.2006.08.1363 (2007).
32. World Health Organization. *International Classification of Diseases for Oncology (ICD-O), 3rd Edition, 1st Revision* (WHO Press, 2013).

Acknowledgements

Yolanda M Gómez was partially funded by FONDECYT, project grant number 11230397 from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science, Technology, Knowledge, and Innovation and Marcelo Bourguignon gratefully acknowledges partial financial support of the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq: grant 304140/2021-0).

Author contributions

"D.I.G.: Conceptualization of this study, Methodology, Software. Y.M.G.: Data curation, Methodology, Software, Writing - Original draft preparation. J.L.S.: Data curation, Methodology, Software, Writing - Original draft preparation. O.V.: formal analysis, investigation, writing–review and editing, funding acquisition. M.B.: Data curation, Methodology, Software. All authors reviewed the manuscript."

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-15903-y.

Correspondence and requests for materials should be addressed to Y.M.G. or O.V.