frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation.

Virginie Rondeau, Juan R. Gonzalez

FRAILTYPACK: a computer program for the analysis of correlated failure time data using penalized likelihood estimation

Virginie Rondeau[1], Juan R. Gonzalez[2].

[1] *INSERM EMI 0338 (Biostatistic), Université Victor Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE.*

[2] *Genes & Disease Program, Biostatistics Unit. Center for Genomic Regulation (CRG), Barcelona Biomedical Research Park (PRBB). Passeig Maritim, 37-49. 08003 Barcelona, Spain.*

Correspondence to: Virginie Rondeau

INSERM EMI 0338,Université Victor Segalen Bordeaux 2,

146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE.

e-mail: Virginie.Rondeau@isped.u-bordeaux2.fr

Tel.: (33) 557 574 531; Fax.: (33) 556 240 081

Key words: frailty models, correlated survival times, penalized likelihood, recurrent event, colorectal cancer.

Abstract

Correlated survival outcomes occur quite frequently in the biomedical research. Available software is limited, particularly if we wish to obtain smoothed estimate of the baseline hazard function in the context of random effects model for correlated data.

The main objective of this paper is to describe an `R` package called *frailtypack* that can be used for estimating the parameters in a shared gamma frailty model with possibly right-censored, left-truncated stratified survival data using penalized likelihood estimation. Time-dependent structure for the explanatory variables and/or extension of the Cox regression model to recurrent events are also allowed. This program can also be used simply to obtain directly a smooth estimate of the baseline hazard function.

To illustrate the program we used two data sets, one with clustered survival times, the other one with recurrent events, ie the rehospitalizations of patients diagnosed with colorectal cancer. We show how to fit the model with recurrent events and time-dependent covariates using Andersen-Gill approach.

1 Introduction

The frailty models, are useful in a variety of biomedical settings to characterize the risk function of an individual when the observations are clustered

into groups such as geographical areas or families and when the observations are non-independent. This might result from the fact that subjects in the same cluster share similar non-observed environmental factors or genetic factors which affect their survival. The random effects modeling in survival analysis can be also easily applied to deal with repeated measurements on a person, and properly account for dependence within subjects. As a result, analyzes that fail to account for the correlation in survival times are likely to underestimate the variances of the parameters. Research on the statistical analysis of correlated survival data has received considerable attention, shared frailty models are used [1, 2, 3]. A review of the different software packages used for analyzing correlated survival data has been recently written by [4]. Semi-parametric inference for frailty models was introduced by Klein *et al.* [5] and Nielsen *et al.* [6]. Their approaches used an EM algorithm applied to the Cox partial likelihood. Hastie and Tibshirani [7] proposed a general model with time varying coefficients and suggested estimation through penalized partial likelihood. Therneau and Grambsch [3, 8] noted a link between the gamma frailty model and a penalized partial likelihood (where the penalization bears on a regression coefficient). These approaches have some general drawbacks. In particular, the convergence can be slow and a direct estimate of the variance of the frailty term is not provided. Furthermore these methods can not be used to estimate the hazard function, which has often a meaningful interpretation in epidemiology. An alternative method is the non-parametric estimation of the baseline hazard function using a con-

tinuous estimator. This approach is based on the penalized full likelihood, as opposed to the penalized partial likelihood of Therneau and Grambsch [3]. The solution is then approximated using splines. The aim of this paper is to introduce computer program called *frailtypack* which implements this approach using a non-parametric penalized likelihood estimation. *frailtypack* can be used to estimate the parameters in a shared gamma frailty model with possibly right-censored, left-truncated and stratified survival data. Time dependent structure for the explanatory variables and/or extension of the Cox regression model to recurrent events are also allowed. This program can also be used simply to obtain smooth estimates of the baseline hazard function and to plot the hazard's curve, in order to check the proportional hazards assumption, for example. The proposed program *frailtypack* was first written in Fortran and has been ported to R, a very useful free software. In section 2 we present the model, and the estimation procedure. In section 3, *frailtypack* is described and in section 4 two applications on biomedical data are exposed.

2 Computational Methods and theory

This section presents a brief description of the methodology used. A more detailed presentation can be found in a companion statistical paper [9].

2.1 The gamma shared frailty model

We consider models in which the hazard function partly depends on an unobservable random variable thought to act multiplicatively on the hazard, so that a large value of the variable increases the hazard. These models, called frailty models are an extension of the classical Cox proportional hazard model [10].

We treat the case of right-censored and left-truncated data. For the j^{th} $(j = 1, ..., n_i)$ individual of the i^{th} group $(i = 1, ..., G)$, let T_{ij} denote the survival times under study and let C_{ij} be the corresponding right-censoring times. The observations are $Y_{ij} = \min(T_{ij}, C_{ij})$ and the censoring indicators $\delta_{ij} = I_{\{T_{ij} \leq C_{ij}\}}$. The survival times may be left-truncated: only subjects with $T_{ij} > \mathcal{L}_{ij}$ are observed. Our frailty model specifies that the hazard function conditional on the frailty is:

$$\lambda_{ij}(t|Z_i) = Z_i \lambda_0(t) \exp(\beta' X_{ij}) \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function; $X_{ij} = (X_{1ij}, ..., X_{pij})'$ denotes the covariate vector for the j^{th} individual and group i, β is the corresponding vector of regression parameters, and the Z_i's are unobserved random variables (the frailties). It is assumed for mathematical convenience that the Z_i's are independently and identically distributed from a gamma distribution with mean 1 and unknown variance θ at origin time; the density probability function is thus: $g(z) = \frac{z^{(1/\theta - 1)} \exp\{-z/\theta\}}{\Gamma(1/\theta)\theta^{1/\theta}}$.

2.2 Penalized Likelihood

In the shared gamma-frailty models the full log-likelihood for left-truncated and right censored data takes a simple form with an analytical solution for the integrals on the frailty term:

$$
\begin{aligned}
l(\lambda_0(.),\beta,\theta) = \sum_{i=1}^{G}\Bigg\{ & \sum_{j=1}^{n_i}\delta_{ij}\{\beta' X_{ij} + \ln(\lambda_0(Y_{ij}))\} \\
& -(1/\theta + m_i)\ln\left[1+\theta\sum_{j=1}^{n_i}\Lambda_0(Y_{ij})\exp(\beta' X_{ij})\right] \\
& +1/\theta\ln\left(1+\theta\sum_{j=1}^{n_i}\Lambda_0(\mathcal{L}_{ij})\exp(\beta' X_{ij})\right) \\
& +I_{\{m_i\neq 0\}}\sum_{k=1}^{m_i}[\ln(1+\theta(m_i-k))]\Bigg\}
\end{aligned}
\tag{2}
$$

with $\Lambda_0(.)$ the cumulative baseline hazard function and $m_i = \sum_{j=1}^{J_i} I_{\{\delta_{ij}=1\}}$ is the number of observed events in the i^{th} group.

Most often, the baseline hazard function can be expected to be smooth. A possible means for introducing such a priori knowledge is to penalize the likelihood by a term which takes large values for rough functions. The penalized log-likelihood is thus defined as:

$$
pl(\lambda_0(.),\beta,\theta) = l(\lambda_0(.),\theta) - \kappa\int_0^{\infty}\lambda_0''^2(t)dt \tag{3}
$$

where $l(\lambda_0(.),\beta,\theta)$ is the full log-likelihood defined in (2), and $\kappa \geq 0$, is a positive smoothing parameter which controls the trade-off between the data

fit and the smoothness of the functions. Maximization of (3) defines the maximum penalized likelihood estimators (MPnLE) $\hat{\lambda}_0(t)$, $\hat{\beta}$, $\hat{\theta}$. Note that in the approach of the present paper, we penalize the baseline hazard function(s) while Therneau and Grambsch [3] penalize the frailties.

We can use directly $\hat{H}^{-1}$ as a variance estimator of the parameters, where $\hat{H}$ is the converged Hessian of the penalized log-likelihood or a "sandwich estimator" $\hat{H}^{-1}I\hat{H}^{-1}$ where I is the information matrix [11]. A significant test for the variance of the random effects distribution occurs on the boundary of the parameter space; the necessary modification to the usual Wald test simplifies to a comparison of the test statistic with a critical value from a one-sided test [12].

2.3 Approximation using splines of the baseline hazard function

The estimator $\hat{\lambda}(.)$ is approximated by a linear combination of m cubic M-splines $\tilde{\lambda}_0(.) = \sum_{i=1}^{m} \eta_j M_j(.)$ [13]. A spline function is completely defined by a sequence of increasing knots and the coefficients $\eta = (\eta_1, ..., \eta_m)^T$ of the splines. In our approximation we used splines of order 4 (also called cubic splines). In *frailtypack*, a knot is set on the first and last data points and the other knots are put equidistantly between them. With the same vector of coefficients η, we get the cumulative baseline hazard function with I-splines (integrated M-splines).

Increasing the number of knots does not deteriorate the MPnLE: this is because the degree of smoothing in the penalized likelihood method is tuned by the smoothing parameter κ and not by the number of splines. Once a sufficient number of knots is established, there is no advantage in adding more. Moreover, the more knots, the longer the running time will be. Some running problem may arise, particularly for a large number of knots. That is why the maximum number of knots is limited to 20, and the minimum number of knots is limited to 4. So it is recommended to start with a small number of knots (e.g. seven) and increase the number of knots until the graph of the baseline hazard function remains unchanged.

To obtain approximate bayesian pointwise 95% confidence bands for the baseline hazard function we use : $\tilde{\lambda}_0(t) \pm 1.96\sqrt{\mathbf{M}'(\mathbf{t})[\widehat{\text{var}}(\hat{\eta})]\mathbf{M}(\mathbf{t})}$ where $\mathbf{M}'(\mathbf{t}) = (M_1(t), ..., M_m(t))$ is the spline vector in t.

2.4 Smoothing parameter

An empirical estimate of the smoothing parameter can be provided or the smoothing parameter can be chosen by maximizing an approximate cross-validation score as in Joly et al. [14]. The cross-validation procedure has been implemented for a classical Cox proportional hazard model and not for a shared frailty model. We included in the program an option to use a fixed smoothing parameter or to use an automatic selection of the smoothing parameter with the cross-validation criterion (version 2.0-0 of the *frailtypack* library). A relationship linking the model degrees of freedom and the smooth-

ing parameter κ [9], the output of the program gives these two values. The model degrees of freedom decrease in κ from m, the number of parameters (if $\kappa = 0$) to 2 (if $\kappa \to +\infty$) which is the number of degrees of freedom of a straight line. However, in some cases, the search for the smoothing parameter may not be reliable because of local extrema. Thus the estimate of the smoothing parameter is not optimal. This can be examined by taking different starting points. Moreover, it seems that the cross-validation score tends to undersmoothe, especially for small samples, so in this case the smoothing parameter may be fixed a priori in the program.

3 The special case of the extension of the Cox regression model to recurrent events with counting process formulation and/or time-dependent covariates

Frailtypack can also be used to deal with recurrent events. Recurrent events is where for instance the subject experiences repeated occurrences of the same type of event, as repeated asthma attacks, cancer relapses, rehospitalization after surgery. The events from the same subjects may be potentially correlated. Different timescales can be used [15]. The timescale that is most often used is the gap time: after an event, the subject starts again at time 0 and the time to the next event corresponds to the number of days that

it takes to experience the next event. Alternative timescale is the calendar time, also called counting process approach [16] which keeps track of time since randomization. The duration of the time at risk for an event corresponds to the duration of the time at risk in the gap time representation but the start of the at-risk period is not reset to 0. A subject is not considered to be at risk for the kth event until after the (k-1)th event. We will detail this calendar time formulation, because the method of estimation is different from that previously seen. The hazard function for the frailty model with calendar time is the same as the one of the expression 1, but the frailty term Z_i will be specific to each subject and time dependent covariates are allowed: $\lambda_{ij}(t|Z_i) = Z_i\lambda_0(t)\exp(\beta' X_{ij}(t))$. A particular subject has different periods at risk during the total observation time. If there are n_i at-risk periods for patient i, then the complete information for patient i can be represented by n_i triplets $(Y_{i11}, Y_{i12}, \delta_{i1}), ..., (Y_{in_i1}, Y_{in_i2}, \delta_{in_i})$ where, for the jth triplet, Y_{ij1} is the start of the jth at-risk period, Y_{ij2} is the end of the jth at-risk period, δ_{ij} is the censoring indicator and $Y_{i11} = 0$.

The expression of the full log-likelihood is different from that of the ex-

pression 2 and becomes :

$$l(\lambda_0(.),\beta,\theta) = \sum_{i=1}^{G}\left\{\sum_{j=1}^{n_i}\delta_{ij}\{\beta' X_{ij}(Y_{ij2}) + \ln(\lambda_0(Y_{ij2}))\}\right. \quad (4)$$
$$-(1/\theta + m_i)\ln\left[1+\theta\sum_{j=1}^{n_i}(\Lambda_0(Y_{ij2}) - \Lambda_0(Y_{ij1}))\exp(\beta' X_{ij}(Y_{ij2}))\right]$$
$$\left.+I_{\{m_i\neq 0\}}\sum_{k=1}^{m_i}[\ln(1+\theta(m_i-k))]\right\}$$

As before a penalized full log-likelihood is used to estimate the parameters.

4 Computational procedure

The estimated parameters were obtained by the robust Marquardt algorithm [17] which is a combination between a Newton-Raphson algorithm and a steepest descent algorithm. This algorithm has the advantage of being more stable than the Newton-Raphson algorithm while preserving its fast convergence property near the maximum. We stopped the iterations when the difference between two consecutive log-likelihoods was small, the coefficients were stable and the gradient was small enough. To be sure of having a positive function at all stages of the algorithm, we restricted all the spline coefficients η_j to be positive for all j. We imposed a constraint of positivity for the variance parameters, $\theta \geq 0$, so we did not consider a negative dependence in the model, which did not have a frailty interpretation.

When frailty parameter is small, numerical problems may arise. To solve

this problem, an alternative formula of the penalized likelihood is used [9].

The first derivative (the score) and the second derivative (the hessian) of the log-likelihood themselves do not have a simple analytical form, so that we computed numerically these quantities using finite differences.

5 Computer program

5.1 General description

A major problem in the parameter estimation is the likelihood maximization procedure. This procedure is generally very time consuming, making necessary the use of compiled programming languages such as Fortran, C, C++,... The combination of high speed compiled languages and the versatility of `R` (flexible, high-level statistical language) provide us with the ideal framework to deal with these problems. Following this philosophy, *frailtypack* has been implemented as a dynamic link library (dll) in Fortran 77, which is called by *R* functions [18].

It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. This package depends on other `R` packages, such as *survival* and *splines*.

The structure of the data file is ASCII free format. The different columns of the datafile will contain (in any order), the entry time (which is different from zero for left truncated data), the outcome variable, the censoring status, the group identification number, the explanatory variables. The Ap-

pendix 1 describes an extract of a datafile structure, corresponding to the first application (see section (6.1)).

Four functions can be used under the library *frailtypack*: *frailtyPenal, print, summary, plot* with adapted arguments. We indicate the imposed arguments with bold characters, in italic are the names of the variables or the values of the parameters that can be changed. Note that the capitalized characters are different from small letters in R.

- **frailtyPenal(Surv(***time, event* **)** $\sim$ *var1* + *var2* + **cluster(***group***)** + *strata(sex)*, **data =***filename*, **Frailty =** *TRUE* **, n.knots=***13*, **kappa1=** *1000, kappa2=1500, maxit=350, recurrentAG=TRUE, cross.validation=TRUE* **)**

  This function allows arbitrary patterns of time-dependent covariates. It handles not only the clustered failure time data but also the recurrent events data. Furthermore a subject is allowed to be at risk for failure in arbitrary time intervals, which is useful for dealing with non-standard situations like delayed study entries as well as for implementing certain methods on recurrence data.

  The CPU time needed to run this function depends on the number of parameters included (in particular the number of knots), the number of observations, and the speed of the processor and the amount of PC memory.

  - *time*: this is the follow-up time, which can be a censored or a time of

event.

If we have left truncated data the expression becomes: *Surv(entry, time, event)* with *entry* the time for left truncation.

- *event*: the status indicator, normally 0=alive, 1=dead

- *var1, var2 ...*: the names of the explanatory variables which can be time-dependent. In this case the datafile structure need to be modified. For instance if the covariate indicates by the values 1 vs. 0 whether or not the patient has or has not a treatment at a specific time we have to separate the record of this individual into two records corresponding to the two at risk time intervals:

entry	*time*	*event*	*id*	*treatment*
0	96	1	1	1
0	52	0	2	0
52	123	0	2	0
123	165	1	2	1
0	120	0	3	0
120	256	1	3	1
...				

Note that the time dependent structure for the covariates need to be associated with the option $recurrentAG = TRUE$ (see after).

- *cluster*: indicates the name of the level of regrouping for the frailty term.

- *Frailty*: indicates with a logical value TRUE or FALSE if the model

includes a frailty term or not. By default the TRUE value is used and the variance of frailty parameter is estimated. If frailty=FALSE, a cox proportional hazards model is estimated using Penalized likelihood on the hazard function.

- *strata(sex)*: indicates a stratified analysis. In this case two different baseline hazard functions will be estimated, corresponding to the two values of the stratification variable. A maximum of 2 strata is allowed.

- *data=filename*: indicates the name of the data file.

- *n.knots=13*: the integer gives the number of knots to use. It corresponds to the (n.knots+2) splines functions for the approximation of the baseline hazard function or the survival functions. The number of knots must be between 4 and 20.

- *kappa1=1000*: gives a positive value for the smoothing parameter of the penalized likelihood (see equation 3)

- *kappa2=1500*: a second smoothing parameter is required if the analysis is stratified. This parameter will correspond to the smoothing for the second baseline hazard function.

- *maxit=350*: is the maximum number of iterations for the Marquardt algorithm. Default is 350.

- *recurrentAG=FALSE*: if =TRUE, indicates that recurrent event times with the counting process approach of Andersen and Gill, ie, with a calendar timescale, is used. The penalized log-likelihood with the expression 4 is used for the estimation. Default is *FALSE*.

- *cross.validation=FALSE*: if =TRUE, indicates that a cross validation procedure is used for estimating the best smoothing parameter. *kappa1* is used as the seed for the estimation κ. Default is *FALSE*.

- **print**(*model1*)

  This function prints a short summary of the estimates for the model *model1*.

- **summary**(*model1, level=0.95*)

  This function gives the hazard ratios of the regression coefficients and their confidence intervals for the model *model1*. The confidence level is allowed (*level=0.95*).

- **plot**(*model1, type="hazard",conf=TRUE*)

  This function plots the baseline hazard function or the survival function with their confidence bands. If *model1* is a stratified model, then two functions will be plotted.

type="hazard": A character string specifying the type of curve to plot. The two possible values are "hazard" for the hazard function and "survival" for the survival function. The default is "hazard". Only the first letters are required, e.g. "haz" or "su".

conf=TRUE: a logical value TRUE or FALSE indicates whether bayesian confidence bands will be plotted or not. The default, is to do so.

6 Applications

In this section we illustrate the use of *frailtypack* with data taken from two biomedical studies. One on clustered survival data the other one on recurrent events that can be useful to illustrate how to use time-dependent covariates, stratified data and delayed entry.

6.1 Clustered survival data

To illustrate the computations we consider a data set published by Mantel et al. [19] of litter-matched tumorigenesis experiment with one drug treated rat and two placebo treated rats per litter. This dataset on *rats* can be downloaded from the web site http://mayoresearch.mayo.edu/mayo/research/biostat/therneau-book.cfm, they are also present in the `R` library *kinship*. Fifty female litters were considered. Using a shared frailty model, we wanted to evaluate if there existed an intra-litter correlation due to the fact that the risk of tumor formation may depend on the genetic background or the early environmental conditions shared by siblings, but differing between litters.

The `R` instructions are:

```
#for the Cox proportional hazard model estimated with a penalized
#likelihood function:
mod.cox<-frailtyPenal(Surv(time,status)~rx+cluster(litter),
```

```
data=rats,Frailty=FALSE,n.knots=8,kappa1=1500,cross.validation=TRUE)

#for the shared gamma frailty model:
mod.frail<-frailtyPenal(Surv(time,status)~rx+cluster(litter),
data=rats,Frailty=TRUE,n.knots=8,kappa1=1500,cross.validation=TRUE)
```

The output given by the function

```
 print(mod.frail)
```

is:

```
Call: frailtyPenal(formula = Surv(time,status) ~ rx +
cluster(litter),data = rats, Frailty = TRUE,
cross.validation=TRUE, n.knots = 8, kappa1 = 1500)

  Shared Gamma Frailty model parameter estimates
  using a Penalized Likelihood on the hazard function
    coef exp(coef) SE coef (H) SE coef (HIH)    z      p
rx 0.956      2.6       0.325         0.325 2.94 0.0033
    Frailty parameter, Theta: 0.479 (SE (H): 0.465 ) (SE (HIH): 0.463 )
    penalized marginal log-likelihood = -240.09
    n= 150
    n events= 40
    n groups= 50
```

```
number of iterations:  14
Exact number of knots used:  8
Best smoothing parameter estimated by
  an approximated Cross validation: 84586295, DoF: -3.43
```

The second fit *mod.frail*, a gamma frailty fit with each litter defining a group, gives a significant treatment effect ($\beta = 0.956, p - value = 0.0033$). The two estimations for the Standard Errors estimates are given ($\hat{H}^{-1}$ and $\hat{H}^{-1}I\hat{H}^{-1}$) and are in this example, almost identical. When frailty was ignored (*mod.cox*), the standard error for the treatment effect was slightly underestimated as expected ($\beta = 0.954, SE = 0.319$). We can also see that there is no significant random effect because the one-sided Wald statistic $\hat{\theta}/SE(\hat{\theta}) = 0.481/0.465 = 1.03$.

The graph of the baseline hazard function, which represents the risk of tumor can be obtained using the instruction:

```
plot(mod.frail,ylim=c(0,0.02),xlim=c(20,110))
```

The estimation of the hazard with four different numbers of knots (therefore, four different smoothing parameters) is illustrated by Figure 1. For a small number of knots the resulting function space may be not flexible enough to capture the variability of the data. For a large number of knots estimated curves may tend to overfit the data. As a remedy, we recommand a moderately large number of knots (usually between 8 and 12) to ensure enough flexibility, and to guarantee sufficient smoothness of the fitted curves. In this

approach the benefit of using 12 knots compared with 8 knots is small, we then recommand the user to use 8 knots.

FIGURE 1 around here

6.2 Recurrent events

Next biomedical example pertains to the rehospitalization of patients diagnosed with colorectal cancer published by Gonzalez *et al.* [20]. The data provide the calendar time (in days) of the successive hospitalizations after the date of surgery. The first readmission time was considered as the time between the date of the surgical procedure and the first rehospitalization after discharge related to colorectal cancer. Each subsequent readmission time was defined as the difference between the current hospitalization date and the previous discharge date. There were a total of 861 rehospitalization events recorded for the 403 patients included in the analysis. Several readmissions can occur for the same patient, and an individual frailty may influence the ocurrence of subsequent rehospitalizations. Thus, the authors proposed to use a gamma shared frailty model for analyzing the data [20].

Other approaches, based on extensions of the Cox proportional hazards model, have appeared in the literature for dealing with such data [3]. In this paper we will compare the results obtained using the frailty models with a gap timescale with those obtained using the calendar timescale of Andersen and Gill (AG) model [16]. In this last formulation the length

of the time-at-risk period is the same as in the gap-time formulation, but the start of the at-risk period is no reset to 0 but to the actual time since entry to the study. A subject may have a delayed entry or censored period before the subject becomesat risk for the event. This is partly due to the availability of *frailtypack* for accommodating the counting process style of input. Other approaches, such as marginal models, are not yet available using *frailtypack* since it only deals with two strata and both approaches need to fit stratified models depending on the number of reoccurrences. In our example the number of rehospitalizations may be greater than two for many patients.

The aim of the investigators was to investigate social-demographic and clinical inequalities in hospital readmission among patients. The main authors' finding was to determine that women with colorectal cancer are less likely than men to be readmitted to the hospital, after controlling for well-established predictors, such as tumor characteristics and comorbidity. This indicates that a stratified model can be adequate for fitting the data. However, before illustrating how to fit stratified model, we show how to fit a classical Cox proportionnal hazard model, and a shared frailty model with different timescales using penalized procedure. To do so, apart from the gender, we include in the models the most important predicting factor of rehospitalization that was the tumor stage (Dukes classification: A-B, C or D) and a time-dependent variable the Charlson's Index (classification: 0, 1-2, $\geq$ 3). The `R` instructions are:

```
# Cox proportional hazards model
fitCoxPen<-frailtyPenal(Surv(time,event)~as.factor(sex)+
             as.factor(dukes)+cluster(id),data=readmission,
             Frail=FALSE,n.knots=6,kappa1=100000,cross.validation=TRUE)

# Andersen-Gill counting process approach
fitAG<-frailtyPenal(Surv(t.start,t.stop,event)~as.factor(sex)+
           as.factor(dukes)+cluster(id),data=readmission,
           Frail=TRUE,n.knots=6,kappa1=100000,cross.validation=TRUE,
           recurrentAG=TRUE)

# Andersen-Gill counting process approach with a time-dependent covariate
fitAG<-frailtyPenal(Surv(t.start,t.stop,event)~as.factor(sex)+
           as.factor(dukes)+as.factor(charlson)+cluster(id),data=readmission,
           Frail=TRUE,n.knots=6,kappa1=100000,cross.validation=TRUE,
           recurrentAG=TRUE)

# Shared gamma frailty model with gap timescale
fitFraPen<-frailtyPenal(Surv(time,event)~as.factor(sex)+
              as.factor(dukes)+cluster(id),data=readmission,
              n.knots=6,kappa1=100000,cross.validation=TRUE)
```

Table 1 shows the estimates using the approaches mentioned above. First of all, as the authors stated, there is a significant random effect. They used a graphical method to check this assumption [20]. However, using the penalized likelihood approach, standard error of frailty term can be estimated, we can then verify the independence assumption using the one-sided Wald statistic. In the non-adjusted model $\hat{\theta}/SE(\hat{\theta}) = 1.09/0.19 = 5.76$ and in the adjusted model (cf Table 1) the heterogenity decreases but is still significant $\hat{\theta}/SE(\hat{\theta}) = 0.72/0.15 = 4.91$. As we have previously mentioned, heterogeneity among patients may lead to an underestimation of the variance estimates. In our example, confidence intervals are larger for frailty model (second to fourth columns in the table) than for Cox model (first column in the table), as expected.

Regarding to the estimated risk of rehospitalization, we observe that Cox model underestimates the effect of being male, although statistically significant differences due to gender are observed using all models. Differences are also observed in the risk due to Dukes stage between Cox and frailty models. On the other hand, big differences are found in the risk among tumoral stages between the model based on calendar timescale and the model based on gap time. In particular, patients diagnosed with tumoral stage 'D' have more risk of being readmitted if we use calendar timescale approach. The same is observed for the hazard ratio of being male. A larger heterogeneity estimate is obtained also with the model based on a calendar timescale. Note that

depending on the timescale selected the interpretation of the time evolution will be different [15].

After observing that gender is an important prognosis factor of being readmitted, we can fit a stratified model. The R instruction is:

```
# Strafified shared gamma frailty model with gap timescale
fitFraPen<-frailtyPenal(Surv(time,event)~as.factor(dukes)+cluster(id)+
                    strata(sex),Frailty=TRUE,data=readmission,n.knots=8,
                      kappa1=100000,kappa2=100000)
```

In that case, using *summary* function (that is, *summary(fitFraPen)*) we obtain the confidence intervals for the hazard ratios of tumoral stages

```
summary(fitFraPen)
                          hr      95%      C.I.
 as.factor(dukes)2   1.56 (   1.16 -   2.09 )
 as.factor(dukes)3   3.66 (   2.55 -   5.24 )
```

Figure 2 illustrates the probability of hospital readmission depending on gender using a stratified model. By typing *type="surv"* in the *plot* function we print the baseline survivor function, S_0, instead of the baseline hazard function as we have illustrated in the previous example. Sometimes one

wants to estimate the baseline probability distribution function ($F_0 = 1 - S_0$), instead of the baseline survivor function. We can easily obtain this plot using the following R instructions (see Figure 2):

```
fitFra.strat$surv <- 1-fitFra.strat$surv
fitFra.strat$surv2 <- 1-fitFra.strat$surv2

plot(fitFra.strat,type="s",ylim=c(0,0.9))
```

FIGURE 2 around here

We also compared our approach with that of Therneau and Grambsch [3] and of Wei, Lin and Weissfeld [21] using the gap timescale. Therneau and Grambsch developped a penalized partial likelihood estimation for the shared frailty models, whereas a marginal model is used by Wei, Lin and Weissfeld. In the marginal model we chose to analyze a maximum of 6 recurrent events per subject, then a maximum of six strata is used in the analysis, ie, one for each observation per subject. The marginal model or the hazard function for the jth event is: $\lambda_{ij}(t) = \lambda_{0j}(t)\exp(\beta' X_{ij})$. These two models can be implemented with R using the *survival* library and the *coxph* function. A comparison of frailty models and marginal models is given in Therneau et al. [3]. The results obtained in Table 2 match the findings previously obtained (see the second column of Table 1). The major advantage of using the *frailtypack* package to model a frailty model, is first to obtain directlty a smooth curve for the baseline hazard function. Secondly an estimate of the

standard error for the variance of the frailty parameter can also be directly obtained in this approach contrary of Therneau's approach.

7 Availability

The library *frailtypack* (*frailtypack* version 2.0-0) with the function *frailtyPenal* is available to the public at no charge and can be loaded at from *http://cran.r-project.org*. Work is in progress to refine and add new options in the program, as the possibility to model nested frailty models with two levels of regrouping of the data.

Acknowledgements

The authors especially wish to thank T. Groth (the editor) and our referees for helping to improve our paper by their suggestions.

References

[1] D. Clayton, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika 65 (1978) 141-151.

[2] P. Hougaard, Frailty models for survival data, LIDA, 1 (1995) 255-273.

[3] T. Therneau and P. Grambsch, Modeling survival data: extending the Cox model, Springer-Verlag: New York. (2000).

[4] P.J. Kelly, A review of software packages for analyzing correlated survival data, The American Statistician, 58 (2004) 337-342.

[5] J.P. Klein, M.L. Moeschberger, Y.H. Li and S.T. Wang, Estimating random effects in the Framingham heart study, Survival analysis: State of the art, Kluwer Academic, Boston, Massachusetts, (1992) 99-120.

[6] G.G. Nielsen, R.D. Gill, P.K. Andersen and T.I.A. Sorensen, A counting process approach to maximum likelihood estimation in frailty models, Scand J Stat, 19 (1992) 25-43.

[7] T.J. Hastie and R.J. Tibshirani, Varying-coefficient models (with discussion), JRSSB, 55 (1993) 757-796.

[8] T.M. Therneau, P.M. Grambsch, V.S. Pankratz, Penalized survival models and frailty, J Comput Graph Stat, 12 (2003) 156-175.

[9] V. Rondeau, D. Commenges and P. Joly, Maximum penalized likelihood estimation in frailty models, LIDA, 9 (2003) 139-153.

[10] D.R. Cox, Regression models and life tables (with discussion), Journal of the Royal Statistical Society B 34 (1972) 187-220.

[11] R.J. Gray, Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis, JASA, 87 (1992) 942-951.

[12] S.G. Self, K. Liang, Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard Conditions, JASA,82 (1987) 605-610.

[13] J.O. Ramsay, Monotone regression splines in action, Statistical Science, 3 (1988) 425-461.

[14] P. Joly, D. L. Letenneur, A. Alioum, D. Commenges, PHMPL: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data, Computer Methods and Programs in Biomedicine, 60 (1999) 225-231.

[15] L. Duchateau, P. Janssen, I. Kessic and C. Fortpied, Evolution of recurrent asthma event rate over time in frailty models, Applied Statistics, 52 (2003) 355-363.

[16] P. Andersen and R. Gill, Cox's regression model for counting processes: a large sample study, Annals of Statistics 10 (1982) 1100-1120.

[17] D. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, SIAM Journal of Applied Mathematics, 11 (1963) 431-441.

[18] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.(2004)

[19] N. Mantel, N.R. Bohidar, J.L. Cinimera, Mantel-Haenszel analyses of litter-matched time-to-response data, with modificatiosn for recorvery of interlitter information, Cancer Res, 37 (1977) 3863-3868.

[20] J.R. Gonzalez, E. Fernandez, V. Moreno, J. Ribes, M. Peris, M. Navarro, M. Cambray, J.M. Borras, Sex differences in hospital readmission among colorectal cancer patients, J Epidemiol Community Health, 59 (2005) 506-511.

[21] L.J. Wei, D.Y. Lin, L. Weissfeld, Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, J Amer Stat Assoc, 84 (1989) 1065-1073.

Appendix: An extract of the litter-matched tumorigenesis experiment data file.

id	*entry*	*time*	*status*	*litter*	*rx*
1	0	101	0	1	1
2	0	104	0	1	0
3	0	49	1	1	0
4	0	104	0	2	1
5	0	104	0	2	0
6	0	102	0	2	0
7	0	104	0	3	1
8	0	104	0	3	0
9	0	104	0	3	0
	...				
145	0	103	1	49	1
146	0	104	0	49	0
147	0	91	0	49	0
148	0	104	0	50	1
149	0	104	0	50	0
150	0	79	1	50	0

FIGURE LEGEND

Figure 1: Baseline hazard function and confidence bands estimated with a shared gamma-frailty model for the risk of tumor formation on 50 litters of female rats.

Figure 2: Baseline distribution function and confidence bands estimated with a shared gamma-frailty model for the probability of hospital readmission after surgery on 403 patients with colorectal cancer.

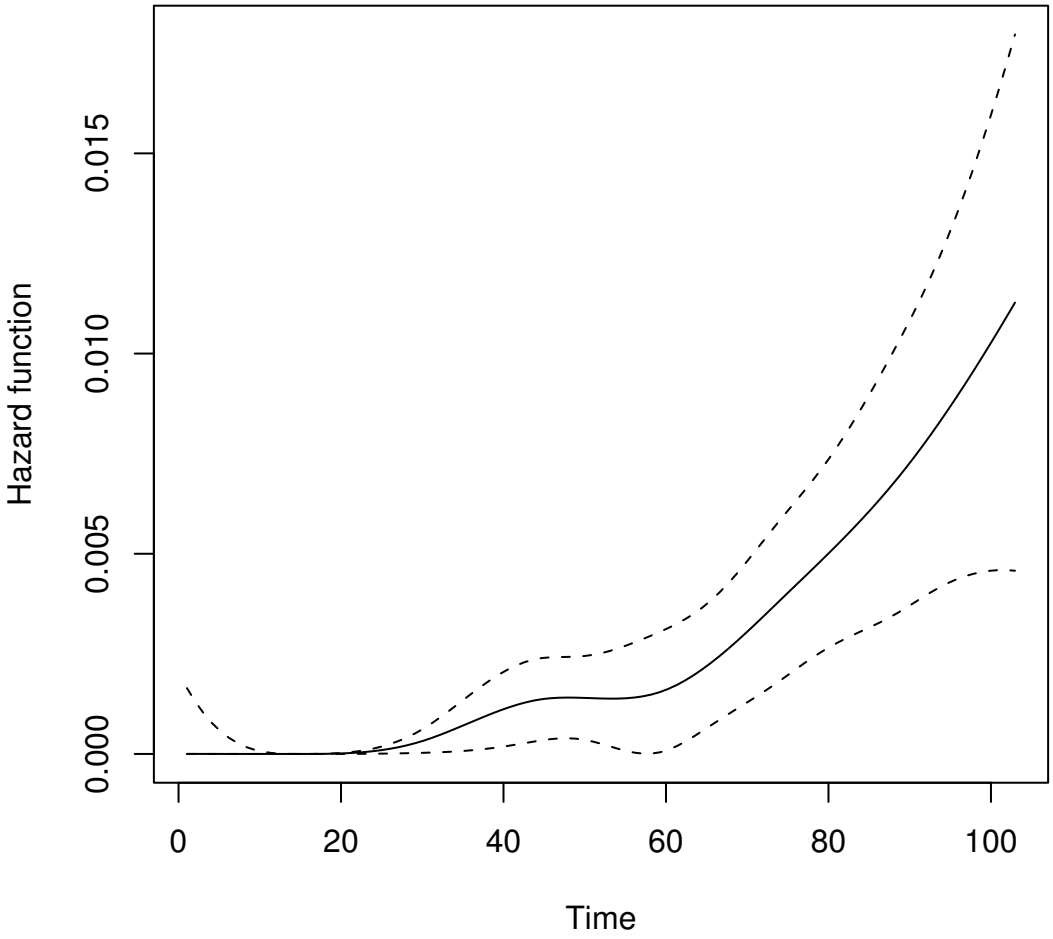

Figure 1: **Baseline hazard function estimated with four different numbers of knots with a shared gamma-frailty model for the risk of tumor formation on 50 litters of female rats.**

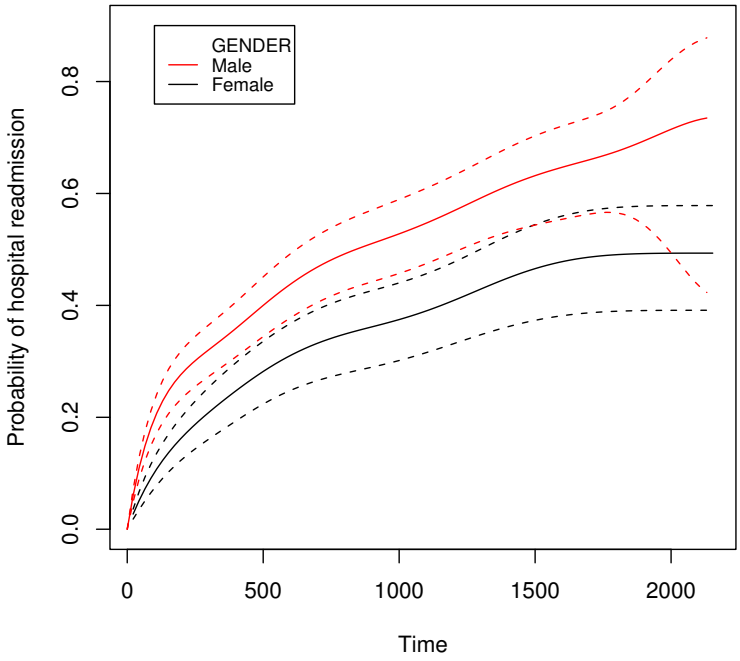

Figure 2: **Baseline distribution function and confidence bands estimated with a shared gamma-frailty model for the probability of hospital readmission after surgery on 403 patients with colorectal cancer.**

TABLE LEGEND

Table 1: Hazard ratios and 95% confidence intervals for the probability of rehospitalization for the readmission data set: estimates using maximum penalized likelihood.

Table 2: Comparison with other approaches using gap timescale: the frailty model of Therneau and the marginal model of Wei, Lin & Weissfeld to study the probability of rehospitalization for the readmission data set.

	Cox model	Shared Gamma Frailty models		
		gap timescale	calendar timescale	
Covariate	HR (CI95%)	HR (CI95%)	HR (CI95%)	HR (CI95%)
Gender				
Female	1	1	1	1
Male	1.52 (1.25-1.85)	1.55 (1.20-2.00)	1.68 (1.25-2.26)	1.71 (1.27-2.31)
Dukes stage				
A-B	1	1	1	1
C	1.56 (1.25-1.94)	1.59 (1.19-2.14)	1.68 (1.19-2.37)	1.50 (1.06-2.12)
D	3.50 (2.74-4.46)	3.72 (2.61-5.32)	5.15 (3.37-7.86)	4.02 (2.58-6.25)
Charlson Index				
0				1
1-2				1.65 (0.92-2.97)
$\geq$3				1.81 (1.35-2.43)
Frailty θ		0.72	1.32	1.28
(SE θ)		(0.15)	(0.19)	(0.19)

Table 1: Hazard ratios and 95% confidence intervals for the probability of rehospitalization for the readmission data set: estimates using maximum penalized likelihood.

	Shared Gamma Frailty model Therneau et al.	Mariginal model Wei-Lin-Weissfeld
Covariate	HR (CI95%)	HR (CI95%)
Gender		
Female	1	1
Male	1.64 (1.26-2.14)	1.62 (1.16-2.26)
Dukes stage		
A-B	1	1
C	1.59 (1.17-2.06)	1.66 (1.14-2.40))
D	3.46 (2.46-4.86)	5.11 (3.32-7.88)
Frailty θ	0.63	

Table 2: Comparison with other approaches using gap timescale: the frailty model of Therneau and the marginal model of Wei, Lin & Weissfeld to study the probability of rehospitalization for the readmission data set.