Biostatistics (2018), **0**, 0, *pp.* 1–26
doi:10.1093/biostatistics/revision˙20180515

Joint modelling of recurrent events and survival: a Bayesian nonparametric approach

GIORGIO PAULON*,

University of Texas at Austin (USA)

giorgio.paulon@utexas.edu

MARIA DE IORIO,

Department of Statistical Science, University College London, UK

ALESSANDRA GUGLIELMI, FRANCESCA IEVA

Dipartimento di Matematica, Politecnico di Milano, Italy

Summary

Heart failure (HF) is one of the main causes of morbidity, hospitalization and death in the western world and the economic burden associated with HF management is relevant and expected to increase in the future. We consider hospitalization data for heart failure in the most populated Italian Region, Lombardia. Data were extracted from the administrative data warehouse of the regional healthcare system. The main clinical outcome of interest is time to death and research focus is on investigating how recurrent hospitalizations affect the time to event. The main contribution of the paper is to develop a joint model for gap times between consecutive re-hospitalizations and survival time. The probability models for the gap times and for the survival outcome share a common patient specific frailty term. Using a flexible Dirichlet process

*To whom correspondence should be addressed.

model for the random effects distribution accounts for patient heterogeneity in recurrent event trajectories. Moreover, the joint model allows for dependent censoring of gap times by death or administrative reasons and for the correlations between different gap times for the same individual. It is straightforward to include covariates in the survival and/or recurrence process through the specification of appropriate regression terms. The main advantages of the proposed methodology are wide applicability, ease of interpretation and efficient computations. Posterior inference is implemented through Markov chain Monte Carlo methods.

Key words: AFT model, Dirichlet process mixtures, frailty, heart failure, re-hospitalizations, waiting times.

1. Introduction

Congestive Heart Failure (HF) is a chronic disease caused by many conditions that damage the heart muscle, including coronary artery disease, heart attack, cardiomyopathy and conditions that overwork the heart (high blood pressure, valve disease, thyroid disease, kidney disease, diabetes or heart defects present at birth). HF prevalence significantly increases with age (see Desai and Stevenson, 2012, for instance), and the number of people living with chronic health conditions for a long time is growing fast. These individuals are often admitted to hospitals and outpatient care services (Gasperoni *and others*, 2017). Multiple re-admissions are largely burdensome to the patient and the healthcare system; for instance, it has been estimated that the average cost of a HF-related event in the most populated region in Italy, Lombardia, is around 6,000 euros (see Mazzali *and others*, 2016).

Despite the efforts to improve the efficiency and the efficacy of treatments and management, re-hospitalization rates remain persistently high. Moreover, the ageing of the population and improved survival of cardiac patients due to modern therapeutic innovations have led to an increasing impact of HF on healthcare systems all over the western countries. This makes the management and planning of healthcare services a crucial issue (Naylor *and others*, 2004). As for many other chronic diseases, clinical interest lies in both the final outcome (death or survival time) and the dynamics of the process itself, since it determines the subsequent quality of patients' life.

For example, the termination time may be dependent on the recurrent event history, and modelling such dependency is one of the principal aims of this work. It is therefore paramount to develop a comprehensive model for disease management, mortality and associated clinical event histories, which is also able to account for the significant inter-individual variability in disease course which is typical of HF. This will aid also in the identification of the drivers of good policies as well as the main factors resulting in the considerable economic burden of HF.

However, most of the previous analyses attempting to identify those patients who are at risk of readmission focused their attention only on the single re-hospitalization and failed to capture the entire pattern of health risks associated to HF. Hence there is a clear need for methods able to consider all possible clinical pathways and assess their dependence on patient characteristics and clinical variables. To address this issue, we consider data on episodes of hospitalization for HF related events and survival, obtained from an administrative healthcare database of Lombardia. The database was originally intended to help policy makers to better manage HF burden of the healthcare costs of Lombardia (Mazzali *and others*, 2016). Administrative databases are usually characterised by high numbers, universal coverage and systematic collection of data over time and offer an invaluable source to study prevalence and incidence of major diseases. They usually provide useful information about the patient's status and pattern of care, especially when the data are fully integrated with clinical data generated as part of routine patient care. As such, there is a growing interest in exploiting administrative data to address epidemiological and healthcare questions.

In our application detailed measurements on clinical biomarkers are not available and, therefore, the main research trust of this work is to develop a joint model for waiting times between re-hospitalizations and survival outcome in HF within a Bayesian nonparametric framework. This choice is motivated by the fact that there is a known relationship between several re-hospitalizations and risk of death due to HF (Postmus *and others*, 2012; Jhund *and others*, 2009). Individuals are clustered according to their history of recurrent events and termination and the ability of assessing the relationship between event occurrence and survival. We also include patient characteristics as potential explanatory factors and asses their ability to predict risk of death and rehospitalizations. The strength of the association between recurrences and terminal event may then be interpreted in terms of patients' risk profile, and a better understanding of how re-hospitalizations affect survival may lead to a more effective planning of healthcare resources.

There have been different approaches to modeling recurrent events and survival, though some of them refer more properly to longitudinal data than recurrent events. Here we mention some of those considering recurrent events and dependent termination. A first approach consists of modeling the intensity of the recurrent events and the survival time. See, for example, Liu *and others* (2004), Ye *and others* (2007), Rondeau *and others* (2007), Huang *and others* (2010); Ouyang *and others* (2013) and Sinha *and others* (2008). The latter two papers propose a Bayesian framework, where the emphasis is on modelling the risk of death and the risks of rejections for heart transplantation patients. A second strategy models the hazard rates of the recurrent gap times and of the survival jointly (e.g. Huang and Liu, 2007; Yu and Liu, 2011).

Similar to Huang and Liu (2007), we model the time dependency between recurrent events assuming that, conditional on subject-specific random effects parameters, the gap (or waiting) times between such events are independent. We then assume that the conditional distribution of the survival time for each individual depends on the same random effect parameters. In other words both conditional distributions, i.e., the one of the j-th gap times and the one of the survival time, share a common subject-specific frailty, that is a subject-specific random effect on the log scale. The joint model takes into account the dependent "censoring" of gap times by death, and the dependency between different gap times for the same patient. However, unlike Huang and Liu (2007), we model each event time distribution as a regression model, and our approach is Bayesian.

The shared frailty parameters are modelled flexibly with a Dirichlet process (DP) (Ferguson, 1973). It is well known that the DP is almost surely discrete, and that if G is a DP(M, G_0) with total mass parameter M and baseline distribution G_0, then G can be represented as (Sethuraman, 1994),

$$G(\cdot) = \sum_{h \geqslant 1} w_h \delta_{\theta_h}(\cdot) \tag{1.1}$$

where δ_θ is a point-mass at θ, the weights follow a stick-breaking process, $w_h = V_h \prod_{j<h}(1 - V_j)$, with $V_h \overset{\text{iid}}{\sim} \text{Beta}(1, M)$, and the atoms $\{\theta_h\}_{h \geqslant 1}$ are such that $\theta_h \overset{\text{iid}}{\sim} G_0$.

The discreteness of the DP induces clustering of the subjects in the sample based on the unique values of the random effects parameters, where the number K of clusters is unknown and learned from the data. This choice allows for extra flexibility, variability between individual trajectories, over-dispersion and clustering of the observations and overcomes the often too restrictive assumptions underlying a parametric distribution. Brown and Ibrahim (2003) use a similar strategy for specifying a joint model for survival and longitudinal outcome to allow for extra flexibility and robustness in the model.

In Section 2 we introduce the model, while in Section 3 we describe the application in detail. In Section 4 posterior inference results are presented, while in 5 we compare our approach to other competing methods. *We conclude the paper in Section 6.*

2. A joint model for gap times of recurrent events with termination

We consider data on N individuals. We assume that $0 := T_{i0}$ corresponds to the start of the event process for individual i and that subject i is observed over the time interval $[0, \zeta_i]$. If n_i events are observed at times $0 < T_{i1} < \cdots < T_{in_i} < \zeta_i$, let $W_{ij} = T_{ij} - T_{ij-1}$ for $j = 1, \ldots, n_i$ denote the waiting times (gap times) between events of subject i. Let S_i denote the survival time of patient i since the start of the corresponding event process: either the time S_i or the censoring time ζ_i is observed. If S_i is observed, then $\zeta_i = S_i$ and $T_{in_i} < S_i$, otherwise $T_{in_i} < \zeta_i$ and $S_i > \zeta_i$. In our application the censoring time is administrative and therefore fixed. In what follows we set $T_{in_i+1} > \zeta_i$ with gap-time $T_{in_i+1} - T_{in_i}$ always censored. Let J be the maximum number of observed repeated events, i.e., $J := \max_{i=1,\ldots,N} n_i$.

As mentioned in the Introduction, our goal is to jointly model the gap times and survival time of each subject in the sample. We assume that, conditional on all parameters and random effects, gap times are independent of each other and are also independent of the survival time. The censoring is assumed to be independent of the survival time and the sequence of gap-times. Shared random effects are responsible for the dependent "censoring" of gap times by termination and the correlation between different gap times for the same subject.

Distinct from Huang and Liu (2007), we assume an accelerated failure time survival model with random effects linking the two processes. This choice is different from non-proportional hazards, and implies, instead, proportional quantiles (see Meeker and Escobar, 2014, Chapter 17). In particular, we specify the following hierarchical structure for the log-transformation of waiting times and survival times, i.e., $Y_{ij} = \log(W_{ij})$, $j = 1, \ldots, n_i+1$, $U_i := \log(S_i)$, $i = 1, \ldots, N$:

$$Y_{ij} | \boldsymbol{x}_{ij}, \boldsymbol{\beta}_j^*, \alpha_i, \sigma_i^2 \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j^* + \alpha_i, \sigma_i^2) \quad j = 1, \ldots, n_i + 1, \tag{2.2}$$

$$U_i | \boldsymbol{z}_i, \boldsymbol{\gamma}, \alpha_i, \psi, \eta^2 \sim \mathcal{N}(\boldsymbol{z}_i^T \boldsymbol{\gamma} + \psi \alpha_i, \eta^2). \tag{2.3}$$

for $i = 1, \ldots, N$. Therefore, $(Y_{i1}, \ldots, Y_{in_i+1})$ and U_i are conditionally independent for each i, given parameters, and trajectories for different patients are conditionally independent as well. Here $\boldsymbol{\beta}_j^* := (\boldsymbol{\beta}_0, \boldsymbol{\beta}_j)^T = (\beta_{01}, \ldots, \beta_{0p}, \beta_{j1}, \ldots, \beta_{jq})^T$ is the vector of regression coefficients, $\boldsymbol{x}_{ij}$ is a set of p fixed and q time-varying covariates influencing the gap times, while $\boldsymbol{\gamma} := (\gamma_1, \ldots, \gamma_r)$ and $\boldsymbol{z}_i$ denote the vector of regression coefficients and fixed covariates, respectively, which are potential predictors of the time-to-event. Note that the effects on disease recurrence and survival are not necessarily the same, since, in general, some therapies may delay disease recurrence but not prolong survival. For this reason, the covariates $\boldsymbol{z}_i$ and the components of $\boldsymbol{x}_{ij}$ may be distinct.

Since the terminal event censors event recurrence, but not vice versa, we need to assume a semi-competing risks model, i.e., a model taking into account that, when subjects are at risk of another recurrent event, they are also at risk of the terminal event; see, for instance Cook and Lawless (2007, Sect. 6.6). The likelihood for subject i, under independent censoring, is then proportional to:

$$\left(\prod_{j=1}^{n_i} f_Y(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_j, \alpha_i, \sigma_i^2)\right) S_Y(\log(\tau_i - (\mathrm{e}^{y_{i1}} + \ldots + \mathrm{e}^{y_{in_i}})))$$
$$\times\, f_U^{1-\nu_i}(\log\tau_i|\boldsymbol{z}_i, \boldsymbol{\gamma}, \alpha_{ij}, \eta^2) S_U^{\nu_i}(\log\tau_i|\boldsymbol{z}_i, \boldsymbol{\gamma}, \alpha_{ij}, \eta^2),$$

where f_Y, f_U are the densities of the gap and survival times (both Gaussian), respectively, S_Y, S_U denote the corresponding survival functions, $\tau_i = \min(S_i, \zeta_i)$ and ν_i is the censoring indicator, which is equal to 1 if the survival time is censored and 0 otherwise.

Note that for each patient the likelihood contribution from the waiting time process includes always a last censored waiting time, independently of censoring by death or administrative reasons. We are making the assumptions (which is common in this type of problems) that a patient cannot experience a recurrent and a terminal event at the same time and that, conditional on the parameters and the frailty terms, the intensity for the recurrent process is of renewal type.

Moreover, as in Olesen and Parner (2006) and Huang and Liu (2007), we assume the existence of L, with L large, gap times for each subject. As each individual is observed up to time τ_i, we observe n_i gap times; the n_i+1-th is censored by τ_i (which explains the term S_Y in the likelihood) and the non-initiated waiting times (after the censored one) are set to infinity. These last gap times have a likelihood contribution equal to 1. Refer to Olesen and Parner (2006) and Huang and Liu (2007) for a detailed discussion.

We assume a priori independence among parameters $\boldsymbol{\beta}_0$, $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$, $\boldsymbol{\gamma}$, ψ, η^2 and $\{(\alpha_i, \sigma_i^2)$, $i = 1 \ldots, N\}$. As random-effect distribution we specify

$$(\alpha_i, \sigma_i^2) | G \overset{\text{iid}}{\sim} G, \quad i = 1, \ldots, N, \qquad G \sim \text{DP}(M, G_0), \tag{2.4}$$

i.e., the random effects distribution is a Dirichlet Process. In summary, our modelling assumptions imply that (i) waiting times are independent of each other, conditional on the other parameters; (ii) the subject-specific random effect α_i links the distribution of the waiting times and survival times; (iii) the shared parameter α_i allows the clustering to depend on both gap times trajectories and survival outcome. We complete the model by setting the following prior distribution on the remaining parameters:

$$\begin{aligned}
&\boldsymbol{\beta}_0 \sim \mathcal{N}_p(\mathbf{0}, \beta_0^2 I_p) \\
&\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J \mid \boldsymbol{\mu} := (\mu_1 \ldots, \mu_q)^T, \Sigma := \text{diag}(\sigma^2_{\beta_1}, \ldots, \sigma^2_{\beta_q}) \overset{\text{iid}}{\sim} \mathcal{N}_q(\boldsymbol{\mu}, \Sigma) \\
&\mu_1, \ldots, \mu_q \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\mu); \quad \sigma^2_{\beta_1}, \ldots, \sigma^2_{\beta_q} \overset{\text{iid}}{\sim} \text{Inv-Gamma}(a_\beta, b_\beta) \\
&\boldsymbol{\gamma} \sim \mathcal{N}_r(\mathbf{0}, \gamma_0^2 I_r); \quad \psi \sim \mathcal{N}(0, \psi_0^2) \\
&\eta^2 \sim \text{inv-Gamma}(a_\eta, b_\eta) \\
&G_0 = \mathcal{N}(0, \alpha_0^2) \times \text{inv-Gamma}(a_\sigma, b_\sigma); \quad M \sim \mathcal{U}(a_M, b_M).
\end{aligned} \tag{2.5}$$

Note the use of a further level of hierarchy in the marginal prior distribution of the time-varying regression coefficients to ensure exchangeability. This allows the coefficients to exchange information over time and leads to better estimates, in particular for the last gap times as often fewer observations are available. Note that identifiability is not an issue here, since the mapping relating the parameter $(\{\boldsymbol{\beta}_j^*\}_j, \alpha_i, \sigma_i^2, \boldsymbol{\gamma}, \psi, \eta)$ to the joint likelihood (2.2)-(2.3) is clearly injective. Moreover, we specify proper priors on ψ and the α_i's, centred in zero, to improve MCMC convergence issues and, when fitting the model to the HF data, we obtain a very low correlation (≈ -0.05) between the chains of ψ and α_i, which confirms that both parameters are identifiable.

Finally, in our analysis, we do not include an intercept term in (2.2)-(2.3) to improve mixing of the MCMC chain.

3. Congestive heart failure dataset

We apply the model described in Section 2 to an administrative dataset extracted from the healthcare data warehouse of Regione Lombardia, which contains information on patient healthcare usage and the relative economic impact on the national health system (e.g. hospitalizations, drugs, visits); see Mazzali *and others* (2015) and Mazzali *and others* (2016) for details. We consider data on a subsample of $n = 1000$ patients, which is representative of the entire population in terms of age, gender, comorbidity burden, number of procedures and groups. We focus on hospitalizations due to Congestive Heart Failure (HF) in the time window January 1st, 2006 - December 31th, 2012. Therefore, for administrative reasons, the censoring time for all the patients in the sample is December 31th, 2012. In the analysis the gap times refer to times in days between successive hospitalizations. The first recorded hospitalization for each patient represents here the origin of the recurrent process ($T_{i0} := 0$ for all i); consequently, n_i represents the number of completely observed gap times between subsequent hospitalizations, given the initial one.

We only consider patients with at least two recurrences (including the first one), i.e. at least one observed gap time but no more than $J = 10$. The resulting dataset for the analysis consists of $N = 810$ patients for a total of 2920 gap times (this subset covers 74.64% of all the events). Table 1 in Supplementary Materials reports the distribution of the number patients N_j for which j, $j = 1, \ldots, 10$, waiting times are observed, where $\sum_j N_j = N$. Figure 1 in Supplementary Materials displays the histogram of the observed gap times in log-scale: 356 out of 810 patients are right-censored (in terms of event time), which corresponds to a high censoring rate, approximately 44%. In Figure 2 in Supplementary Materials we show the empirical distribution of event times (in days) for censored (blue) and non-censored (red) observations on a log scale.

Information has also been collected on several covariates, fixed or time-varying. We report below a list of the covariates included in the model:

- *gender* of the patient. In Table 2 in Supplementary Materials, we report the percentage p_j of male patients among the observations available at each gap time j.

- *age* [years] of the patient at each hospitalization. Empirical mean of *age* at entrance in the study is 75.77 (sd 10.78). The sample means of *age* stratified by gender are 78.44 (sd 10.02) for women and 73.14 (sd 10.86) for men, respectively. In the model we include only *age* at the entrance in the study, as we find that there is no gain including *age* at the end of each gap time.

- *group*: indicator variable which identifies the clinical classification of the patient according to criteria detailed in Mazzali *and others* (2016). As a result, four different groups were defined (see Figure 2 and Table 2 in Mazzali *and others*, 2016): G1 denotes the group of patients having HF as the cause of admission or complicating another cardiac disease. G2 includes patients with Myocardial or cardiopulmonary diseases. G3 refers to patients with Acute HF as a complication of other diseases or for whom HF is reported as comorbidity. Finally, G4 identifies the remaining subjects (only three). The "group" variable is defined at the time of the first heart failure event, independently of subsequent events. As G1 represents the most frequent classification, as well as the most traditional characterization of hearth failure, we reduce the variable *group* to a binary covariate, which is set equal to 0 if the label of the patient is G1 (560 patients), and 1 otherwise (250 patients). Hence, *group* denotes the indicator of non-standard characterization of hearth failure.

- *rehab*: binary variable indicating if any time during the hospitalization is spent in a rehabilitation unit: 11.78% of the hospitalizations are spent at least partially in a rehabilitation unit, corresponding to 29.01% of the patients.

- *ic*: binary variable indicating if at least part of the hospitalization is spent in a intensive care unit. This occurs in 11.95% of the hospitalizations (i.e. for 31.48% of patients).

- *n_com*: number of comorbidities for each hospitalization. Table 3 in Supplementary Materials reports the average number of comorbidities for all patients at the j-th gap times.

- *n_pro*: number of surgical procedures for each hospitalization. Table 4 in Supplementary Materials reports the average number of procedures for all patients at the j-th gap times.

Note that 426 (52.59%) patients spent some time in either a rehabilitation or intensive care unit; none of the patients was admitted in both rehabilitation and intensive care unit at the same hospitalization; 129 (15.93%) patients entered a rehabilitation unit in at least one hospitalization, but never an intensive care one, while for 149 (18.40%) patients the opposite occurs.

Moreover, it is important to highlight that each gap time is calculated as the difference between two successive hospitalizations and, as such, it captures both the length of stay in hospital and the time between the discharge and the next hospitalization of the patient. In the analysis we include the value of the time dependent covariates measured at the end of each gap time.

When fitting model, variables *gender*, *age* and *group* are treated as fixed covariates ($p = 3$), whereas *rehab*, *ic*, *n_com* and *n_pro* are time varying ($q = 4$). In the analysis, *age*, *n_com* and *n_pro* have been standardized to have mean zero and variance one. Therefore, the linear regression term for patient i at time j is given by

$$\beta_{01}x_{i1} + \beta_{02}x_{i2} + \beta_{03}x_{i3} + \beta_{j1}x_{ij1} + \beta_{j2}x_{ij2} + \beta_{j3}x_{ij3} + \beta_{j4}x_{ij4},$$

where $\boldsymbol{x}_{ij} := (x_{i1}, x_{i2}, x_{i3}, x_{ij1}, x_{ij2}, x_{ij3}, x_{ij4})$; x_{i1} and x_{i2} correspond to indicators for *gender* ($= 0$ if the patient is a female) and *group* ($= 0$ for standard characterization of hearth failure, i.e. G1), respectively, x_{i3} is the standardized age at the beginning of the study; x_{ij1} and x_{ij2} denote the binary variables *rehab* and *ic* for patient i at the j-th time, respectively, while x_{ij3} and x_{ij4}

are the standardized number of comorbidities and number of surgical procedures of patient i during the j-th gap time.

The time-varying covariates are likely to have an effect also on survival. To capture such a relationship we include as predictors in the regression term in (2.3) functions of the time-varying covariates. In particular, we define two binary variables, x_{i4} and x_{i5} representing if a patient was admitted in a rehabilitation or an intensive care unit at least once during the observation period, respectively. We also include as covariates in (2.3) the average number of comorbidities (x_{i6}) and of surgical procedures (x_{i7}) over time. Therefore, we assume that $\boldsymbol{z}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7})$, i.e., the time-homogenous covariates included in the survival regression are *gender*, *group* and *standardized age* at the beginning of the study and the summary statistics of the time varying covariates. This approach leads to satisfactory results for our application. It is in principle straightforward to specify a joint model for survival, recurrent event and longitudinal measurements, for example following the approach in van Dijkhuizen *and others* (2017) and Li *and others* (2016), but at an increased computational cost.

4. Posterior inference

In this Section, we report posterior inference results. Details on the Gibbs sampler algorithm, on the prior choice and robustness with respect to different choices of the prior distribution are presented in Section 2 in Supplementary Materials.

First we show inference results for the regression parameters in order to understand how covariates influence the recurrent events distribution and survival, regardless of the underlying structure of the trajectories (which is captured by the subject-specific parameters). We report in Table 1 the probability that each regression parameter on the gap times ($\beta_{01}, \beta_{02}, \beta_{03}$) and on the survival ($\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7$) is larger than 0. Recall that $\gamma_4, \gamma_5, \gamma_6$ and γ_7 refer to time-varying covariates, which are included in the survival regression via summary statistics. Moreover, Figure 3 in Supplementary Materials shows the 95% credible intervals for the posterior marginals of the same regression parameters. To summarize the results in Table 1, by "effect" we mean that the marginal posterior probabilities are large or small (e.g. larger than 0.9 or smaller that 0.1). There is an effect of the group and age variables on both gap times and survival, as well as an effect on survival of gender, average number of comorbidities and whether a patient has been admitted to a rehabilitation unit at least once.

In order to better interpret the effect of the covariates, Table 2 reports posterior predictive survival probabilities for hypothetical new patients characterised by different levels of the covariates. Survival probabilities are evaluated at three different time thresholds *(s_i, $i = 1, 2, 3$),* set to be equal to the quartiles of the empirical distribution of the observed terminal times, *i.e. s_1 is 1.1, s_2 is 2.5 and s_3 is around 4 years.* We first consider a baseline covariate combination and then we investigate the effect on survival of each covariate individually by varying its level with respect to the baseline. Since the covariates included in the study correspond to intrinsic characteristics of the patients or to HF clinical classification that cannot be changed, our analysis highlights important risk factors for survival; *for instance, Table 2 confirms that age considerably decreases both survival probabilities and median survival time.* Hence, in general, it remains difficult to suggest to the healthcare provider ways of reducing the risk of death based on our results. From Table 2, it is clear that patients spending longer times in a rehabilitation unit seem to benefit from it in terms of predicted survival probabilities. Consequently, a potential practical importance of this study is to suggest to healthcare providers that investing in rehabilitation units might lead to a better level of care for patients affected by congestive HF. Table 2 also shows the median survival time for each combination of covariates, again in agreement with the previous findings.

In addition, Figure 1 shows predictive survival curves for four hypothetical patients, corresponding to four different combinations of covariates: we consider two 70-years-old patients (top) versus two 83-years-old patients (bottom), men (left) and women (right); the selected ages correspond to the 25% and 75% empirical quantiles of the age distribution in the dataset.

Time-varying covariates are reported at the end of each gap time. We need to take this into account when analysing the results for the regression coefficients of such covariates. In general they do not appear to have an effect on the recurrence process, except for few early gap times. As expected, the uncertainty on the effect estimates increases over time, due to the smaller number of available observations. Figure 4 in Supplementary Materials displays the 95% CIs of the marginal posterior densities of the regression coefficients $(\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_{j4})$ of the time-varying covariates. It is evident that *ic* has an effect on the distribution of waiting times, as patients with *ic* equal 1 show shorter gap times. Moreover, it appears that a large number of surgical procedures yield frequent hospitalizations at the beginning of the study, followed by an opposite effect (delayed hospitalizations) for later gap times.

Figure 2 displays the marginal posterior distribution of the parameter ψ, which links the survival outcome to the recurrent event process and reflects the relationship between the two processes. In our application, this distribution is centred away from zero, on the positive axis (posterior mean equal to 1.41 and 95% CI $[1.37; 1.44]$), indicating that as the time between hospitalizations increases, the median survival time tends to increase as well. This result is consistent with the fact that rehospitalization is common among patients affected by HF and that the course of this disease is often characterized by repeated hospital admissions at relatively short intervals and a limited prognosis for survival (Neumann *and others*, 2009); this conclusion is also confirmed by the clustering output where clusters with highest risk of death (Clusters 4 and 6 in Table 5 in Supplementary Materials) are characterised by shortest gap times between subsequent rehospitalizations.

The model described by (2.2)-(2.3) and (2.5)-(2.4) induces a prior on the partition of the subjects in the sample (see, for instance, Barcella *and others*, 2015). We summarise the MCMC output by reporting the clustering that minimizes the posterior expectation of Binder's loss function (Binder, 1978) under equal misclassification costs. Briefly, the Binder loss function measures the distance for all possible pairs of subjects between the true probability of co-clustering and the estimated cluster allocation. See Section 1 in Supplementary Materials for more details. The estimated partition contains 6 clusters as confirmed also by the posterior distribution of K, the number of clusters (see Figure 5 in Supplementary Materials). This is also evident from the posterior predictive density of the random effect parameter $\alpha^{\star}$, corresponding to a hypothetical new patient from our model (see Figure 6 in the Supplementary Materials), which is multimodal. In Figure 3 we report the cluster specific posterior estimates (conditional on cluster assignment) of the random effect parameters α_i's (panel (a)) and the Kaplan-Meier survival estimates within each of the six estimated clusters, respectively. Note that the parameters (α_i, σ_i^2) determine the clustering of patients. Since the parameter α_i links the failure and recurrence processes, our modelling strategy allows the clustering to depend on both gap times trajectories and survival outcome.

In Table 5 in Supplementary Materials we report cluster-specific summary statistics. Trajectories of the gap times, $Y_{ij},\ j = 1, ..., n_i$, for all the patients in each cluster are displayed in Figure 7 in Supplementary Materials. The largest cluster of patients (56.85% of the patients), denoted as Cluster 2 in Table 5 in Supplementary Materials, is characterized by large survival times. Clusters denoted as 3 and 5 include mostly censored observations; however, they are different as Cluster 3 shows large survival times and shorter gap times with a large average number of hospitalizations, whereas Cluster 5 includes younger patients with the largest waiting times but with only few recurrent events. We note that the percentage of patients with standard pathology (i.e. group=0) is similar to the overall rate ($\sim 70\%$) in each cluster except for Cluster 5, where it is 47%. This is an interesting result as it points towards the existence of clinical unobserved factors affecting intra patients variability.

5. Comparison with other models

As a comparison, we fit to our data the generalized mixed Poisson process (GMPP) model by Ouyang *and others* (2013), jointly modelling recurrent event counts and survival time in a Bayesian framework. The model is described in terms of the intensity λ_{N_i} of the recurrent process and the hazard h_{S_i} for the survival time:

$$\lambda_{N_i}(t|w_i, \boldsymbol{\beta}_0, \boldsymbol{z}_i, \eta, \delta) = \lambda_0(t)\ w_i e^{\boldsymbol{z}_i^T \boldsymbol{\beta}_0}; \quad h_{S_i}(t|w_i, \psi, \boldsymbol{\gamma}, \boldsymbol{z}_i, \eta^\star, \delta^\star) = \phi_0(t)\ w_i^{\psi} e^{\boldsymbol{z}_i^T \boldsymbol{\gamma}}. \tag{5.6}$$

Here ϕ_0 and λ_0 are modelled as Weibull hazards. This model introduces dependence via the individual frailty terms $w_1, \ldots, w_n$, which account for the heterogeneity in the patient population. As random effect distribution for the w_i, the authors opt for a Gamma distribution. Our implementation of the GMPP follows the suggestions in Ouyang *and others* (2013) This model allows only for time-homogeneous covariates, which, in our case, implies to set $\boldsymbol{z}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7})$. The prior specification for the regression coefficients $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}$ and for ψ is the same as in our model. Note that ψ plays the same role as in (2.3), linking the recurrent process with survival in the likelihood.

From Figure 9 in Supplementary Materials, it is evident that we obtain the same conclusions for covariate effects as in our model. In Ouyang *and others* (2013), covariates have an impact on the hazard functions, therefore a positive effect implies a larger risk of event occurrence or death time. In our model, instead, positive effects would increase the median time of death or recurrence. The most influential covariates are the same in the two models, as well as the direction of their effects. The estimate of ψ is 5.76 (95% CI: $[4.45, 7.29]$); this means that the incidence of recurrences is positively associated with death, again in agreement with our model.

We also compare the in-sample predictive performances of the two models in Figure 10 in Supplementary Materials, which reports the posterior distribution of the Brier score (Brier, 1950), that measures the accuracy of in-sample prediction. As the Brier score is usually evaluated for binary classification problems, we need to dichotomise our prediction to adapt the Brier score to continuous data, as proposed by Barcella *and others* (2016). Therefore, we are interested in predicting whether the survival time of an individual is above or below a specific threshold. We use the quartiles of the observed survival times as thresholds. We discretise the observed data so that $\tilde{u}_i^{(k)} = 0$ if $\tilde{u}_i \leqslant Q_k$ and $\tilde{u}_i^{(k)} = 1$ if $\tilde{u}_i > Q_k$, where Q_k is the k-th quartile of the data, $k = 1, 2, 3$. Details on how to evaluate the Brier score from an MCMC output are given in Section 4 of Supplementary Materials. Small values of the Brier statistic indicates good classification performance. From Figure 10 in Supplementary Materials, it is evident that our approach outperforms the one in Ouyang *and others* (2013) for smaller survival times, and it has similar performances on the other quartiles of the data.

In Figure 1, we show predictive survival curves, obtained from both models, for four hypothetical patients. The two models generally agree in terms of inferred survival profiles. Our models shows wider credible intervals due to a higher number of parameters and the extra flexibility. In Supplementary Materials, we also compare our model with the Cox Proportional Hazards model, as well as with the model described in Rondeau *and others* (2007) and implemented the R package `frailtypack`. We asses the out-of-sample performance of the proposed approach in Section 5 in Supplementary Materials.

From these experiments we conclude that advantages of the Bayesian nonparametric model described here include flexibility of the random effect distribution, automatic clustering of the individuals according to their risk profile, ability to include time-homogeneous and inhomogeneous covariates in both the survival and recurrence process, predictive performance and ease of interpretation, still allowing efficient computations.

6. Discussion

Treating and appropriately managing Heart Failure (HF) patients is a major public health issue. Indeed, HF is one of the major causes of hospitalization and death in adult population and the main reason for hospital admission in patients aged over 65 years in western countries. As the population ages and the prevalence of heart failure increases, expenditures related to the care of these patients are climbing dramatically. As a result, the health care industry must develop strategies to contain the economic burden without compromising the effectiveness of the care. To this aim, it is essential to gain a deeper understanding of the most influential factors contributing to lengthen/decrease short-, middle- and long-term survival jointly with the length and recurrence of hospitalizations.

In this paper we have proposed a joint semiparametric model for recurrent hospitalizations due to HF and time to death. Our approach jointly models survival and the hospitalizations times, specifying a DP as random effect distribution of the frailty parameter that links the survival and gap time trajectories. This strategy allows us to introduce extra flexibility in the model, able to account for patients heterogeneity. The data available to us provide information on a broad population of HF patients and our analysis has highlighted important determinants of repeated hospitalizations as well as survival. In particular, advanced age and higher morbidity load increases the risk of death and of being rehospitalized (Figure 3 and Figure 4 in Supplementary Materials). This is not surprising as older age is usually associated with worse health conditions and comorbidity load. The effect of the variable *group* on gap times reflects different protocols of HF treatments. Furthermore, we have investigated the effect over time of the time-varying covariates, highlighting possible temporal patterns. Moreover, the model is able to account for patient-specific heterogeneity through the data-driven clustering of patients based on their re-hospitalizations trajectory and survival outcome, showing that frequent hospitalizations are usually associated to a higher risk of death. Note that the nonparametric approach and the implied clustering of subjects captures possible subgroup structure in the administrative database. These results highlight different ways to manage specific patients' patterns of care according to their risk profiles or subgroup structure and may help public authorities to achieve a more efficient planning of healthcare resources, tailoring healthcare paths on specific patients' needs in a more efficient way (see Driscoll *and others*, 2016; Gasperoni *and others*, 2017). In our particular application, the main practical recommendation for the healthcare provider is to invest in rehabilitation units as they seem to improve survival of certain categories of patients.

Instead of modelling the hazards of gap and survival times, e.g. through proportional hazards models, we adopted an accelerated failure time model for both processes, thus giving an alternative to the PH model when supported by the data. Moreover, the model is simply a linear regression on the log scale, leading to an easy interpretability of the relationship between covariates and event processes as well as of the dependence between the two processes. See Tseng *and others* (2005) for some of the advantages of accelerated failure time models over the Cox model.

The proposed approach can be generalized by making different distributional assumptions for either/both gap and survival times (e.g. Weibull distributions), by allowing the variance of the gap times to depend on the time index and/or including more complex temporal dependence structure between gap times (e.g. an autoregressive model as in Tallarita *and others* (2016)). Further developments may include the extension of the methodology to a much richer dataset, including a wider patient population and new potential explanatory variables, such as treatment. In addition, a hospital effect and spatial information can be easily incorporated in the model. Moreover, it is easy to perform variable selection in this context by assuming, for example, a spike and slab prior on the regression coefficients and performing Stochastic Search Variable Selection. See Rockova *and others* (2012) for a review of Bayesian variable selection strategies. These future directions of research will most likely require generalising the methodology to combine aggregated and individual level information. The approach could also be extended by specifying a joint model for survival, recurrent events and longitudinal biomarkers, by linking the different processes through random effects, or alternatively modelling the error distribution as a mixtures of Polya Trees model following the approach in Hanson *and others* (2011). Finally, the model assumes a time-homogeneous effect of the gap times on survival, which can be a strong modelling assumption when many subsequent hospitalizations are associated with a higher risk of death. This limitation can be overcome by, for example, by assuming time varying effect parameters α_{it}. In conclusion, the flexibility of the proposed approach makes it an ideal building block of more complex hierarchies.

7. Supplementary Materials

Supplementary material is available online at `http://biostatistics.oxfordjournals.org`. It includes further analysis results and comparison with other methods.

Acknowledgments

Maria De Iorio and Alessandra Guglielmi acknowledge The Royal Society for providing funding under award [IE140575]. The authors thank all the referees, the Associate Editor and the Editor for useful comments and suggestions which greatly improved the manuscript.

References

Barcella, William, De Iorio, Maria and Baio, Gianluca. (2015). A comparative review of variable selection techniques for covariate dependent Dirichet process mixture models. *arXiv preprint arXiv:1508.00129*.

Barcella, William, Iorio, Maria De, Baio, Gianluca and Malone-Lee, James. (2016). Variable selection in covariate dependent random partition models: an application to urinary tract infection. *Statistics in Medicine* **35**, 1373–1389.

Binder, David A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

Brier, Glenn W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

Brown, Elizabeth and Ibrahim, Joseph. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221–228.

Cook, Richard J and Lawless, Jerald F. (2007). *The statistical analysis of recurrent events*. Springer, New York.

Desai, Akshay S and Stevenson, Lynne W. (2012). Rehospitalization for heart failure. *Circulation* **126**, 501–506.

Driscoll, Andrea, Meagher, Sharon, Kennedy, Rhoda, Hay, Melanie, Banerji, Jayant, Campbell, Donald, Cox, Nicholas, Gascard, Debra, Hare, David, Page, Karen *and others*. (2016). What is the impact of systems of care for heart failure on patients diagnosed with heart failure: a systematic review. *BMC cardiovascular disorders* **16**(1), 195.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Gasperoni, Francesca, Ieva, Francesca, Barbati, Giulia, Scagnetto, Arjuna, Iorio, Annamaria, Sinagra, Gianfranco and Di Lenarda, Andrea. (2017). Multi-state modelling of heart failure care path: A population-based investigation from italy. *PloS one* (6), e0179176.

Hanson, Timothy E, Branscum, Adam J and Johnson, Wesley O. (2011). Predictive comparison of joint longitudinal-survival modeling: a case study illustrating competing approaches. *Lifetime data analysis* **17**, 3–28.

Huang, C-Y, Qin, J and Wang, M-C. (2010). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* **66**(1), 39–49.

Huang, Xuelin and Liu, Lei. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics* **63**, 389–397.

Jhund, Pardeep S, MacIntyre, Kate, Simpson, Colin R, Lewsey, James D, Stewart, Simon, Redpath, Adam, Chalmers, James WT, Capewell, Simon and McMurray, John JV. (2009). Long-term trends in first hospitalization for heart failure and subsequent survival between 1986 and 2003. *Circulation* **119**(4), 515–523.

Li, Qiuju, Pan, Jianxin and Belcher, John. (2016). Bayesian inference for joint modelling of longitudinal continuous, binary and ordinal events. *Statistical methods in medical research* **25**(6), 2521–2540.

Liu, Lei, Wolfe, Robert A. and Huang, Xuelin. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60**, 747–756.

Mazzali, Cristina, Maistriello, Mauro, Ieva, Francesca and Barbieri, Pietro. (2015). Methodological issues in the use of administrative databases to study heart failure. In: *Advances in Complex Data Modeling and Computational Methods in Statistics*. Springer, pp. 149–160.

Mazzali, Cristina, Paganoni, Anna Maria, Ieva, Francesca, Masella, Cristina, Maistrello, Mauro, Agostoni, Ornella, Scalvini, Simonetta and Frigerio, Maria. (2016). Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012. *BMC Health Services Research* **16**, 234.

Meeker, William Q and Escobar, Luis A. (2014). *Statistical methods for reliability data*. John Wiley & Sons.

Naylor, Mary D, Brooten, Dorothy A, Campbell, Roberta L, Maislin, Greg, McCauley, Kathleen M and Schwartz, J Sanford. (2004). Transitional care of older adults hospitalized with heart failure: a randomized, controlled trial. *Journal of the American Geriatrics Society* **52**(5), 675–684.

Neumann, Till, Biermann, Janine, Erbel, Raimund, Neumann, Anja, Wasem, Jürgen, Ertl, Georg and Dietz, Rainer. (2009). Heart failure: the commonest reason for hospital admission in germany: medical and economic perspectives. *Deutsches Ärzteblatt International* **106**(16), 269.

Olesen, Anne Vingaard and Parner, Erik Thorlund. (2006). Correcting for selection using frailty models. *Statistics in medicine* **25**(10), 1672–1684.

Ouyang, Bichun, Sinha, Debajyoti, Slate, Elizabeth H and Van Bakel, Adrian B. (2013). Bayesian analysis of recurrent event with dependent termination: an application to a heart transplant study. *Statistics in Medicine* **32**, 2629–2642.

Postmus, Douwe, Veldhuisen, Dirk J, Jaarsma, Tiny, Luttik, Marie Louise, Lassus, Johan, Mebazaa, Alexandre, Nieminen, Markku S, Harjola, Veli-Pekka, Lewsey, James, Buskens, Erik *and others*. (2012). The coach risk engine: a multistate model for predicting survival and hospitalization in patients with heart failure. *European journal of heart failure* **14**(2), 168–175.

Rockova, Veronika, Lesaffre, Emmanuel, Luime, Jolanda and Löwenberg, Bob. (2012). Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in Medicine* **31**, 1221–1237.

Rondeau, Virginie, Mathoulin-Pelissier, Simone, Jacqmin-Gadda, Hélène, Brouste, Véronique and Soubeyran, Pierre. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics* **8**, 708–721.

Sethuraman, Jayaram. (1994). A constructive definition of Dirichet priors. *Statistica Sinica* **4**, 639–650.

Sinha, Debajyoti, Maiti, Tapabrata, Ibrahim, Joseph G and Ouyang, Bichun. (2008). Current methods for recurrent events data with dependent termination. *Journal of the American Statistical Association* **103**, 866–878.

Tallarita, Marta, De Iorio, Maria, Guglielmi, Alessandra and Malone-Lee, James. (2016). Bayesian nonparametric modelling of joint gap time distributions for recurrent event data. *arXiv preprint arXiv:1607.08141*.

Tseng, Yi-Kuan, Hsieh, Fushing and Wang, Jane-Ling. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92**(3), 587–603.

van Dijkhuizen, EH Pieter, Deakin, Claire T, Wedderburn, Lucy R and De Iorio, Maria. (2017). Modelling disease activity in juvenile dermatomyositis: a Bayesian approach. *Statistical Methods in Medical Research*, 0962280217713233.

Ye, Yining, Kalbfleisch, John D and Schaubel, Douglas E. (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics* **63**(1), 78–87.

Yu, Zhangsheng and Liu, Lei. (2011). A joint model of recurrent events and a terminal event with a nonparametric covariate function. *Statistics in Medicine* **30**, 2683–2695.

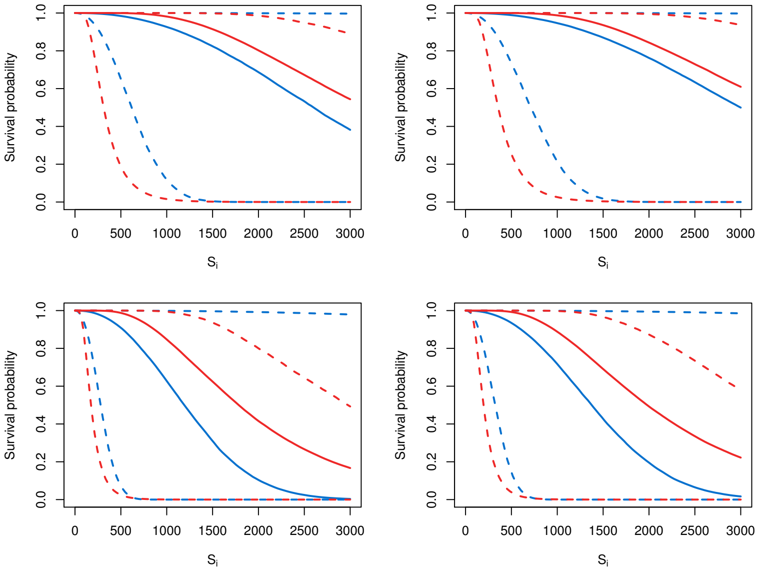


Fig. 1: Predictive survival curves for four hypothetical new patients obtained from the proposed model (in red) and to Ouyang et al. (in blue), along with the 95% CIs. Top left, 70-year-old man. Top right, 70-year-old woman. Bottom left, 83-year-old man. Bottom right, 83-year-old woman. The other covariates are fixed at the sample mode for binary and at the sample mean for continuous covariates.

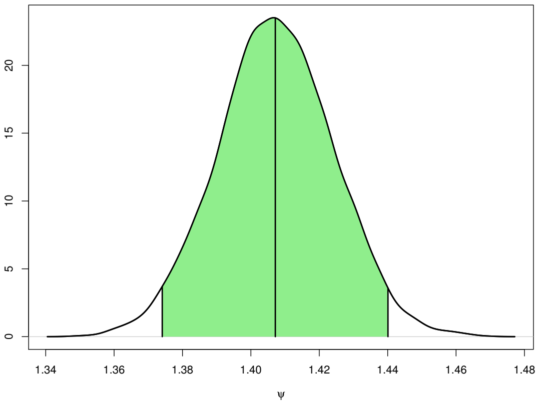


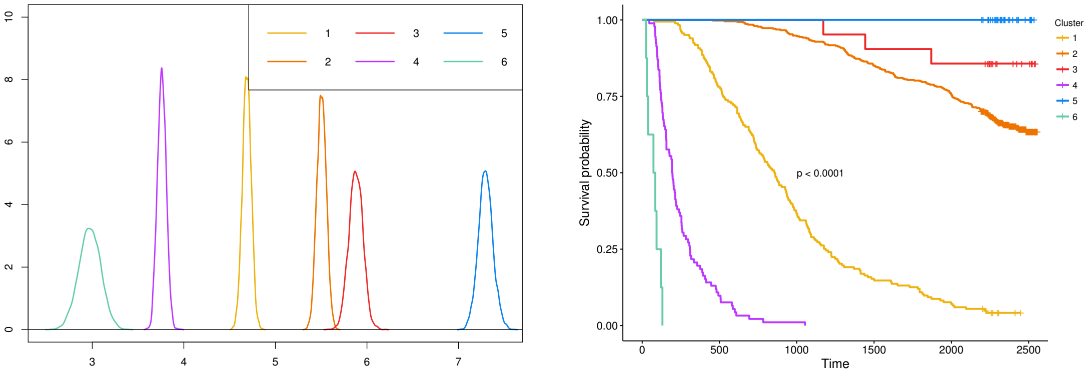
Fig. 2: Marginal posterior density of ψ.

Table 1: Marginal posterior probability of each fixed coefficient to be greater than 0.

	β_{01} *gender*	β_{02} *group*	β_{03} *age*	γ_1 *gender*	γ_2 *group*	γ_3 *age*	γ_4 *rehab*	γ_5 *ic*	γ_6 *n_com*	γ_7 *n_pro*
$\mathbb{P}(\,\cdot > 0 \vert \text{data})$	0.480	0.959	0.051	0.094	0.067	0	1	0.784	0.004	0.278

Table 2: Posterior predictive survival probabilities evaluated at times $s_1 = 404$ days, $s_2 = 928.5$ days, $s_3 = 1496.25$ for different combination of covariates. The baseline patient is a 60 years old woman, group G1, who never transitioned through rehabilitation nor intensive care, and with number of procedures and comorbidities fixed to the sample means. Computation of these probabilities was obtained by varying the value of one covariate at a time, fixing the remaining ones at the baseline levels. The last column represents the estimate of the median survival time for each covariate group.

	$\mathbb{P}(S > s_1)$	$\mathbb{P}(S > s_2)$	$\mathbb{P}(S > s_3)$	Median Survival
Baseline:	0.945	0.860	0.776	4420.6
gender: male	0.935	0.845	0.751	3964.9
group: other	0.933	0.841	0.745	3877.6
age: 70	0.903	0.783	0.660	2819.2
83	0.831	0.629	0.465	1576.4
rehab: Yes	0.975	0.909	0.855	6957.7
ic: Yes	0.950	0.868	0.790	4727.3



(a) Estimates of the α_i's within the clusters

(b) Kaplan-Meier estimate of the survival within each cluster

Fig. 3: Estimates of cluster specific parameters α_i's and Kaplan-Meier estimates of the survival time within each clusters; colours indicate clusters and cluster names are consistent with Table 5 in Supplementary Materials.