Bayesian inference on the number of recurrent events: A joint model of recurrence and survival

Willem van den Boom\textsuperscript{1,2}, Maria De Iorio\textsuperscript{2,3,4}, and Marta Tallarita\textsuperscript{4}

Abstract
The number of recurrent events before a terminating event is often of interest. For instance, death terminates an individual\textquotesingle s process of rehospitalizations and the number of rehospitalizations is an important indicator of economic cost. We propose a model in which the number of recurrences before termination is a random variable of interest, enabling inference and prediction on it. Then, conditionally on this number, we specify a joint distribution for recurrence and survival. This novel conditional approach induces dependence between recurrence and survival, which is often present, for instance, due to frailty that affects both. Additional dependence between recurrence and survival is introduced by the specification of a joint distribution on their respective frailty terms. Moreover, through the introduction of an autoregressive model, our approach is able to capture the temporal dependence in the recurrent events trajectory. A non-parametric random effects distribution for the frailty terms accommodates population heterogeneity and allows for data-driven clustering of the subjects. A tailored Gibbs sampler involving reversible jump and slice sampling steps implements posterior inference. We illustrate our model on colorectal cancer data, compare its performance with existing approaches and provide appropriate inference on the number of recurrent events.

Keywords
Accelerated failure time model, censoring, colorectal cancer, Dirichlet process mixtures, hospital readmission cost burden, number of recurrent events, reversible jump Markov chain Monte Carlo

Introduction
Recurrent events arise in many applications including, amongst others, medicine, science and technology. Typical examples are given by recurrent infections, asthma attacks, hospitalizations, product repairs, and machine failures. Often, the number of recurrent events before termination of the process, such as by death or failure, is of interest. For instance, rehospitalizations are a major financial burden for the health system\textsuperscript{1} and their number is used in policymaking.\textsuperscript{2,3} This work proposes a model which explicitly accounts for the number of recurrent events before termination by building on recent advances in joint modelling of recurrence and termination.

The two main statistical approaches to inference on recurrent events are (1) modelling the intensity or hazard function of the event counts process and (2) modelling the sequence of times between recurrent events, known as gap times or waiting times.\textsuperscript{4} The first approach is most suitable when individuals frequently experience the recurrent event of interest and the...
occurrence does not alter the process itself. Here, we mention some examples that consider the dependence between recurrence and survival time. Liu et al., Rondeau et al., Ye et al., Huang et al., Sinha et al. and Ouyang et al. model the intensity of the recurrent events and the survival time. The latter two approaches propose Bayesian methods with an emphasis on modelling the risk of death and the risk of rejections for heart transplantation patients. Olesen and Parner, Huang and Liu, Yu and Liu, Bao et al. and Liu et al. model the hazard function of the recurrent events and of the survival jointly, with the recurrent events and the survival being independent conditionally on frailty parameters. Yu et al. model the intensity of the recurrent events and the hazard function of survival jointly while assuming independent censoring before death.

The second approach, which focuses on the sequence of gap times, is more appropriate when the recurrent events are relatively infrequent, when individual renewal takes place after an event, or when the goal is prediction of the time to the next event. Therefore, this approach is highly relevant for biomedical applications. For instance, major recurrent cardiac events for one patient are often rather infrequent from a statistical viewpoint. Also, health care planning can benefit from time-to-event predictions, especially if events require hospitalization. Nonetheless, there is less existing work on the second approach than on the first. This work places itself within the second framework, as the events in our application are infrequent but measured on many individuals.

Using the second approach, Li et al. define survival functions via a copula to allow for dependence between recurrence and survival conditionally on frailty parameters. Paulon et al. and Tallarita et al. propose Bayesian non-parametric models for the gap times, with the former also considering dependence with survival time. We take this previous work as a starting point with some important differences. Our autoregressive process for gap times has a constant mean, unlike Tallarita et al. as the process includes regression coefficients which would otherwise be hard to interpret. Unlike Paulon et al., we do not assume independence of gap times conditionally on random effects, and we have separate random effects for gap and survival times, enabling greater flexibility in capturing dependence or lack thereof between recurrence and survival.

The main novel methodological contribution of this work is to explicitly enable inference on the number of recurrent events preceding a termination event. Previous work also considers termination of the observed recurrence process. In that work, a large number of recurrent events are assumed to exist for each individual with the recurrent event process defined also after the terminal event has occurred. Then, the gap times are censored either by the survival censoring time or by the actual survival time. The contribution to the likelihood of the gap times after censoring is set equal to one such that the assumed large number of censored gap times does not affect inference. It is preferable to avoid the often unrealistic assumption of a large arbitrary number of recurrent events or the continuation of the recurrence process beyond the terminal event. More importantly, these approaches prevent inference on the number of events before termination. This can constitute a major limitation in many applications. For example, there is a large literature for hospitalizations on assessing the risk of recurrence for a time window such as within 30 days of hospital discharge as this has serious implications for health care cost. Our model can not only provide such risk estimates but is also able to infer the number of rehospitalizations which can aid health care planning (e.g. budgeting, provisioning).

We overcome the limitations of previous approaches by adopting a conditional approach. More in detail, we explicitly model the number of recurrent events before the terminal event. Then, conditionally on the number of events, we specify a joint distribution for gap times and survival. To the best of our knowledge, this modelling strategy has never been employed in the context of joint modelling of survival and recurrence processes. The resulting posterior inference is computationally more challenging than the models proposed in the available literature. We therefore develop a Markov chain Monte Carlo algorithm that relies on various computational tools such as reversible jump Markov chain Monte Carlo and slice sampling. Our explicit modelling of the number of recurrent events yields a more intuitive model specification and captures the dependence structure between the recurrence and survival processes, which is important in medical applications and beyond. The main motivating result though is intuitive inference and prediction for the number of recurrent events.

An important factor in medical applications related to recurrent events and survival time is the overall frailty. Increased frailty is often associated with both increased disease recurrence and reduced survival. Subject-specific random effects describe the frailty by informing both the survival time and the dependence of subsequent gap times. The random effects are modelled flexibly with a Dirichlet process (DP) prior as in Paulon et al. and Tallarita et al. It is well known that the DP is almost surely discrete. This feature is particularly useful in applications as it allows for data-driven clustering of observations. If $G$ is $DP(M, G_0)$ with concentration parameter $M$ and base measure $G_0$, then it admits a stick-breaking representation and can be represented as

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot)$$
where \( \theta_h \) is a point mass at \( \theta_h \), the weights \( w_h \) follow the stick-breaking process \( w_h = V_h \prod_{j<k} (1 - V_j) \) with \( V_h \sim \text{Beta}(1, M_h) \), and the atoms \( \{\theta_h\}_{h=1}^\infty \) are such that \( \theta_h \sim G_0 \). The sequences \( \{\theta_h\}_{h=1}^\infty \) and \( \{V_h\}_{h=1}^\infty \) are independent. The discreteness of \( G \) induces clustering of the subjects, based on the unique values of the random locations \( \theta_h \), where the number of clusters is learned from the data. This choice allows for extra flexibility, variability between individual trajectories, overdispersion and clustering of observations, and overcomes the often too restrictive assumptions underlying a parametric random effects distribution.

Paulon et al.\(^{19}\) specify a single random effects parameter that influences both the distribution of the gap times and the distribution of survival. Instead, we introduce different random effects parameters, one for the recurrence process and one for the survival. We model these jointly using a DP prior, ensuring dependence between recurrence and survival. Additionally, we specify an autoregressive model for the gap times to capture the dependence between subsequent gap times as some persistence of recurrence across time is expected. Tallarita et al.\(^{20}\) also use an autoregressive model, but on the random effects instead of the gap times themselves.

The paper is structured as follows. Section “Model” introduces the model. Section “Simulation study” checks its performance on simulated data. Section “Application to colorectal cancer data” discusses an application to colorectal cancer data. Section “Comparison with other models” compares our approach with existing ones. Section “Discussion” concludes the paper.

Model

Notation

We consider data on \( L \) individuals. Let \( T_{i0} \) denote the start time of the recurrent event process for individual \( i \). We assume \( T_{i0} = 0 \) for \( i = 1, \ldots, L \). Let \( S_i \) denote the survival time for individual \( i \) since the start of the corresponding event process. Each individual \( i \) experiences \( N_i \) recurrent events over the time interval \((0, S_i] \). Let \( T_{ij} \) denote the \( j \)th event time for individual \( i \). Then, the last event time \( T_{iN_i} \) is less than or equal to \( S_i \).

Some event processes are right censored due to the end of the study, as in the application of Section “Application to colorectal cancer data”, or loss to follow-up. We assume completely independent censoring. This contrasts with the survival time \( S_i \) which our model allows to depend on the event process. Let \( c_i \) denote the minimum of the censoring time and the survival time \( S_i \) for individual \( i \), which is thus observed over the interval \((0, c_i] \). Let \( n_i \leq N_i \) denote the number of events that are observed over the interval \((0, c_i] \). Either \( S_i \) or the censoring time is observed. If \( S_i \) is observed, then \( N_i = n_i \) and \( 0 < T_{i1} < \ldots < T_{iN_i} \leq S_i = c_i \). If \( S_i \) is not observed, then \( N_i \geq n_i \) and \( S_i > c_i \) are unknown and object of inference. In this case, \( 0 < T_{i1} < \ldots < T_{in} \leq c_i < T_{i(n+1)} < \ldots < T_{iN_i} \leq S_i \). We define the log gap times as

\[
Y_{ij} = \log(T_{ij} - T_{(j-1)})
\]

for \( j = 1, \ldots, N_i \). The \( q \)-dimensional vector \( x_i \) contains individual-specific covariates.

Likelihood specification

Firstly, we assume that the number of gap times \( N_i \) follows a negative binomial distribution with shape parameter \( r \) and mean \( \lambda \) independently for \( i = 1, \ldots, L \). Conditionally on \( N_i \), we specify a joint model for the log gap times \( Y_i = (Y_{i1}, \ldots, Y_{in})^T \) and the survival time \( S_i \). We define the joint density \( \mathcal{P}(Y_i, S_i | \ldots) \) where the joint space is constrained by \( T_{iN_i} \leq S_i \). This induces dependence among \( N_i, Y_i \) and \( S_i \) in a principled manner. We build on the existing literature\(^{19,20}\) by assuming that the gap times and survival times follow truncated log-normal distributions where the pairs \((Y_i, S_i)\) are mutually independent for \( i = 1, \ldots, L \), conditionally on the random effects, number of recurrent events \( N_i \) and the other parameters in the model. Conditionally on \( N_i \), the joint distribution for \((Y_i, S_i)\) contains two components:

\[
p(Y_i, S_i | \ldots) \propto f(Y_i | \ldots) f(S_i | \ldots) 1(T_{iN_i} \leq S_i)
\]

one defining the recurrence process and one for the survival, where \( T_{iN_i} = \sum_{j=1}^{N_i} e^{x_i} \) as per (1). The dependence between the two processes is defined through the constraint on the joint space and the specification of a joint random effect distribution on the process-specific parameters.

The random effects parameter for the gap times is the two-dimensional vector \( m_i \), which characterizes an autoregressive model for \( Y_i \) that captures dependence between subsequent gap times:

\[
f(Y_i | \beta, m_i, N_i, \sigma^2, x_i) = \mathcal{N}(Y_{i1} | x_i^T \beta + m_{i1}, \sigma^2)
\]

\[
\times \prod_{j=2}^{N_i} \mathcal{N}(Y_{ij} | x_i^T \beta + m_{ij} + m_{i2}(Y_{(j-1)} - x_i^T \beta - m_{i1}), \sigma^2),
\]

for \( j = 2, \ldots, N_i \).
for \(N_i \geq 1\). When \(N_i = 0\), then we assume that \(f(Y_i | \beta, m_i, N_i, \sigma^2, x_i)\) is equal to a constant. The \(q\)-dimensional vector \(\beta\) consists of covariate effects on the gap times. This resembles the autoregressive model on the random effects in equation (2) of Tallarita et al.\(^{20}\) Two main differences are due to the fact that the mean of \(Y_i\) is the same for all \(j\) conditionally on the remaining parameters in our model and that Tallarita et al.\(^{20}\) do not consider a survival process, which implies the truncation \(T_{N_i} \leq S_i\) in our work. The truncation results from our conditioning on \(N_i\) whereas the existing literature\(^{19,20,28}\) specifies the likelihood as a joint distribution of the number of events \(n_i\) observed over the interval \((0, c_i]\) and their log gap times \(Y_{i1}, \ldots, Y_{in_i}\). The regression coefficient \(\beta\) in (3) has the usual interpretation since the mean of \(Y_i\) equals \(x_i^T \beta + m_i\) for all \(j\).

The survival component of the model is proportional to a log–normal density:

\[
f(S_i | \gamma, \delta_i, x_i) = \mathcal{LN}\{S_i | x_i^T \gamma + \delta_i, \eta^2\}
\]

where the \(q\)-dimensional vector \(\gamma\) consists of covariate effects on the survival time and \(\delta_i\) denotes a random effects parameter. Covariate effects can differ between gap and survival times, for instance, if a therapy delays disease recurrence but does not prolong survival. Therefore, the model on the gap times in (3) and on the survival times in (4) have distinct regression coefficients \(\beta\) and \(\gamma\), respectively. Ultimately, due to the use of a DP prior on \((m_i, \delta_i)\) our sampling model is an infinite mixture with weights and location deriving from the Dirichlet process.\(^{20,30}\) This allows us to overcome the often too restrictive assumptions imposed by a choice of a parametric model.

**Prior specification**

We specify a non-parametric prior for the random effects parameters \(m_i\) and \(\delta_i\) in (3) and (4). In more detail, \((m_i, \delta_i) \sim \mathcal{G}\) independently for \(i = 1, \ldots, L\), where \(G \sim \text{DP}(M, G_0)\), with \(M \sim \text{Gamma}(a_M, b_M)\) and base measure \(G_0 = N_2(0_{2x1}, \sigma_2^2 I_2) \times N(0, \sigma_2^2)\). Finally, the prior distributions on the remaining parameters are \(\beta \sim \mathcal{N}_q(0_{qx1}, \sigma_\beta^2 I_q)\), \(\gamma \sim \mathcal{N}_q(0_{qx1}, \sigma_\gamma^2 I_q)\), \(\sigma^2 \sim \text{Inv-Gamma}(\nu_\sigma^2/2, \nu_\sigma^2 \sigma_\sigma^2/2)\), \(\eta^2 \sim \text{Inv-Gamma}(\nu_\eta^2/2, \nu_\eta^2 \eta_\eta^2/2)\), \(r \sim \text{Gamma}(a_r, b_r)\) and \(\lambda \sim \text{Gamma}(a_\lambda, b_\lambda)\).

**Posterior inference**

The posterior inference is performed through a Gibbs sampler algorithm. This includes imputing \(N_i\), \(Y_{(h_i+1)}\), \(\ldots, Y_{(h_i+1)}\), and \(S_i\) for each censored individual \(i\) by sampling them according to the model in Section “Likelihood specification”. The Gibbs update for \(N_i\) and \(Y_i\) is transdimensional since the number of events \(N_i\) represents the dimensionality of the sequence of log gap times \(Y_i\). This requires devising a reversible jump sampler\(^{24}\) for \(N_i\) and \(Y_i\).

Most full conditional distributions are intractable due to the truncation \(T_{N_i} \leq S_i\), which for instance causes the normalization constant of \(p(Y_i, S_i | \ldots)\) to depend on parameters of interest. We use slice sampling\(^{25}\) to deal with this intractability. The normalization constant of \(p(Y_i, S_i | \ldots)\) is also intractable. We therefore approximate it using the Fenton–Wilkinson\(^{31}\) method. Algorithm 8 from Neal\(^{32}\) is implemented to sample the DP parameters \((m_i, \delta_i)\). Section S1 of the Supplemental material details and derives the Markov chain Monte Carlo algorithms.

**Simulation study**

We investigate the performance of our model and the Markov chain Monte Carlo algorithm via a simulation study. We consider \(L = 150\) individuals spread across three clusters of size 50 each by assigning every first, second and third individual to Cluster 1, 2 and 3, respectively, and \(q = 2\) covariates which are drawn independently from a standard uniform distribution. Then, we sample data according to the likelihood in Section “Likelihood specification” with \(\beta = \gamma = (-1, 1)^T\), \(r = 1\), \(\lambda = 7\), \(\sigma^2 = 1\), \(\eta^2 = 1\), \(m_{i1} = \{h, 0.8(h-2)\}^T\) and \(m_{i2} = \delta_i\) if and only if individual \(i\) belongs to the \(h\)th cluster for \(h = 1, 2, 3\). From these simulated data, we construct four different scenarios, namely where 0%, 50%, 80% and 90% of the individuals, selected uniformly at random, are censored. The censoring times \(c_i\) are sampled uniformly from the time interval \((T_{i1}, S_i)\) between the first event recurrence and death similarly to the simulation in Section 4 of Tallarita et al.\(^{20}\) if \(N_i \geq 1\), and from \((0, S_i)\) otherwise.

We choose hyperparameters yielding uninformative prior distributions as detailed in Section S2 of the Supplemental material. The base measure \(G_0\) of the DP prior has high variance. A priori, \(\sigma^2\) and \(\eta^2\) have an expected value of one and a variance of 100. We run the Gibbs sampler for 200 000 iterations, dis-carding the first 20 000 as burn-in and thinning every 10 iterations, resulting in a final posterior sample size of 18 000.
Figures 1 through 5 summarize the results. The predictions and uncertainty quantification for the number of recurrent events $N_i$ in Figure 1 are sensible with only 16 (5%) of the 330 credible intervals not covering the true $N_i$. Figures 2 and 3 show increased posterior uncertainty for higher levels of censoring. This increase in uncertainty is larger for $\delta_i$, which relates to the survival time $S_i$, than for other parameters which relate to the recurrence process. A reason for this is that for a censored individual $i$, often some event times are observed while $S_i$ is right censored and thus not observed. The posterior mass on the actual number of clusters is marginally higher for the uncensored than for the censored data in Figure 4, although it must be noticed that posterior inference on $K$ is robust across different levels of censoring. Figure 5 shows accurate inference for $\beta$ with the uncertainty in $\gamma$ increasing with the censoring rate as a higher censoring induces more uncertainty about the survival time $S_i$.

Section S3 of the Supplemental material contains additional simulation studies. Our simulations show that posterior inference is robust not only to the choice of hyperparameters of the negative binomial distribution, but also to model misspecification. Nevertheless, it must be noted that posterior mean estimates and their associated bias and mean squared errors are affected by the censoring rate.

## Application to colorectal cancer data

### Data description and analysis

We apply our model to the colorectal cancer data described in Gonzalez et al., which consider $L = 403$ patients diagnosed with colorectal cancer between 1996 and 1998 in Bellvitge University Hospital in Barcelona, Spain. The data consist of hospital readmissions related to colorectal cancer surgery up until 2002 and are available as part of the R package frailtypack. The date of surgery represents the origin of a patient’s recurrence process such that $T_{i0} = 0$ for all $i$. Consequently, $n_i$ represents the number of observed gap times between subsequent hospitalizations. Patients experience between zero and 22 hospitalizations each and $\sum_{i=1}^{L} n_i = 458$ in aggregate. Table 1 shows how they are distributed across patients. Gap times are defined as the difference between successive hospitalizations and, as such, capture both the length of stay in the hospital and the time between discharge and the next hospitalization.

The main clinical outcome of interest is deterioration to death. We therefore define the survival time $S_i$ as the time to death. The survival times of 294 out of the 403 recurrence processes are censored due to the follow-up ending in 2002, resulting in an unobserved total number of gap times $Ni$.

Patient characteristics considered are the binary variables (1) whether the patient received radiotherapy or chemotherapy and (2) gender, and Dukes’ tumour classification, which takes as levels stage A–B, C or D. We dummy code Dukes’ classification with stage A–B as baseline resulting in a subject-specific 4-dimensional covariate vector $x_i$, with $q = 4$. Table 2 and Figures 6 and 7 summarize the patient characteristics, and the gap and survival times.

We use the same priors, from Section S2 of the Supplemental material, and set-up of the Gibbs sampler as the simulation study in Section “Simulation study”. Then, the regression coefficients $\beta$ and $\gamma$ have high prior variance.

### Posterior inference on the number of recurrent events

Figure 8 summarizes the posterior distribution of the total number of gap times $N_i$ for each censored patient. The posterior means for the censored $N_i$ are generally in line with the observed $N_i$ in Table 1 though the lowest posterior mean is with 1.9 higher than the lowest observed $N_i$. This is expected since patients with longer survival times $S_i$ are both more likely to have a higher number of gap times $N_i$ and to be censored due to the end of the study. Our model flexibly captures $N_i$’s uncertainty, which varies notably across censored patients. These findings highlight the importance of modelling and inferring $N_i$ when the number of events is censored and, therefore, unknown.

### Table 1. Frequency table of the number of observed gap times $n_i$ and, for the 109 uncensored patients, the total number of gap times $N_i$ in the colorectal cancer data.

| $n_i$ | 0  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 16 | 22 |
|------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Frequency (all) | 199 | 105 | 45 | 21 | 15 | 8  | 4  | 0  | 1  | 1  | 1  | 1  |     |    |
| Frequency (uncensored) | 36  | 33  | 16 | 10 | 6  | 4  | 0  | 0  | 1  | 1  | 0  | 0  | 1  | 1  |
Figure 1. The number of gap times \(N_i\) (circle) and, if applicable, their posterior means (dot) and 95% posterior credible intervals (lines) for each individual from our model fitted on the simulated data. For censored individuals, the number of observed gap times \(n_i\) is marked by ‘\(\times\)’.
Posterior inference on the regression coefficients

The regression coefficients $\beta$ and $\gamma$ capture the covariate effects on the recurrent event and survival processes, respectively. Figure 9 shows negative coefficients for the more advanced tumour stages C and D in the survival regression but not for the gap times model. This suggests that more severe tumours, while they negatively affect survival, do not have an effect on the rehospitalization rate beyond the link between survival and hospitalization implied by our joint model.

Posterior inference on the cluster allocation

As discussed in the “Introduction” section, the DP prior on $(m_i, \delta_i)$ described in the “Prior specification” section allows for clustering of patients based on their recurrent event and survival profiles. The random effects parameters determine the clustering of patients and capture the dependence between the recurrence and survival processes. Indeed, the posterior predictive distribution of these parameters for a hypothetical new patient is multimodal as shown in Figure 10, indicating the presence of multiple patient subpopulations. The clustering depends on both gap time trajectories and survival outcomes thanks to the joint distribution on $m_i$ and $\delta_i$. In particular, Figure 10 reports the bivariate posterior marginals of $(m_{i1}, \delta_i)$ and $(m_{i2}, \delta_i)$, which are bimodal.

Posterior inference on the clustering structure of the patients is of clinical interest as it might guide more targeted therapies. Our Gibbs sampler provides posterior samples of the cluster allocation. Here, we report the cluster allocation that minimizes the posterior expectation of Binder’s loss function under equal misclassification costs, which is a common choice in the applied Bayesian non-parametric literature. See Appendix B of Argiento et al. for computational details. Briefly, Binder’s loss function measures the difference for all possible pairs of individuals between the true probability of co-clustering and the estimated cluster allocation. In this context, the posterior estimate of the partition of the patients has three clusters, with 99% of the patients allocated to two clusters which are summarized in Table 2. The largest cluster, Cluster 1, has longer gap and imputed survival times than Cluster 2. Moreover, the Kaplan–Meier curves of each cluster in Figure 11 support the conclusion that Cluster 1 includes patients with longer survival times than Cluster 2. As shown in Table 2, Cluster 1 has a higher censoring rate than Cluster 2, as one might expect at longer survival times. The lower prevalence of late-stage tumours in Cluster 1 also confirms that it includes healthier subjects than Cluster 2.

Figures S17 and S18 in the Supplemental material contain additional posterior results. Section S4 of the Supplemental material presents an application to atrial fibrillation data.

Comparison with other models

Cox proportional hazards model

We now compare our results on the colorectal cancer data to those obtained from the Cox proportional hazards model which is one of the most popular semiparametric models in survival analysis with covariates. In this model, the hazard

|                | Full dataset | Cluster 1 | Cluster 2 |
|----------------|--------------|-----------|-----------|
| Number of patients | 403          | 292       | 108       |
| Proportion censored | 73%          | 81%       | 54%       |
| Average uncensored $N_i$ | 1.78 (2.92) | 0.79 (0.93) | 2.68 (3.41) |
| Average posterior mean of $N_i$ (SD) | 2.87 (1.95) | 2.55 (1.08) | 3.66 (2.91) |
| Average uncensored $Y_i$ (SD) | 4.35 (1.84) | 5.55 (1.24) | 3.91 (1.72) |
| Average posterior mean of $Y_i$ (SD) | 6.55 (1.39) | 7.06 (0.82) | 5.03 (1.52) |
| Average uncensored $\log(S_i)$ (SD) | 5.78 (1.03) | 5.66 (1.09) | 5.96 (0.93) |
| Average posterior mean of $\log(S_i)$ (SD) | 8.91 (2.14) | 9.39 (2.03) | 7.72 (1.85) |
| Proportion on chemotherapy | 54%          | 55%       | 50%       |
| Proportion female | 41%          | 42%       | 37%       |
| Proportion stage C | 37%          | 34%       | 45%       |
| Proportion stage D | 19%          | 19%       | 17%       |
function for mortality is

\[ h(t \mid \theta) = h_0(t) \exp(z_i^T \theta), \]

where \( h_0(t) \) is the baseline hazard function, \( z_i \) is a vector of covariates and \( \theta \) is a vector of regression coefficients. Here, a larger value of \( z_i^T \theta \) leads to shorter survival times. This contrasts with (4) from our model where a larger \( x_i^T \gamma \) is associated with longer survival times.

Figure 2. Contour plots of the log of the bivariate posterior predictive densities of \((m_{i1}, \delta_i)\) (left) and \((m_{i2}, \delta_i)\) (right) for a hypothetical new individual from the simulated data.
Figure 3. Posterior density for $\sigma^2$, and posterior predictive density for $\log(S_i)$ and posterior predictive probability mass function for $N_i$ for a hypothetical new patient with covariates equal to their sample medians from our model fitted on the simulated data.
Our model allows for dependence between the gap and survival times. Therefore, for a fairer comparison when fitting the Cox proportional hazard model, we include a patient’s log mean gap time in the covariate vector \( z_i \) in addition to the four covariates included in \( x_i \) described in the “Data description and analysis” section. As a result, the 199 individuals with
no observed gap times are excluded. Table S6 in the Supplemental material shows the covariate effects on survival from the Cox proportional hazards model. The statistically significant effects agree with those from our model in Figure 9.

**Joint frailty model**

We also compare our model with the joint frailty model by Rondeau et al.\(^6\) as implemented in the R package `frailtypack`.\(^{14}\) The model estimates the hazard functions of rehospitalization and mortality jointly using two patient-specific frailty terms, namely \(u_i\) and \(v_i\). The frailty term \(u_i\) captures the association between rehospitalization and mortality while \(v_i\) appears solely in the rehospitalization rate. Specifically, the hazard functions are

\[
\begin{align*}
    r_i(t \mid u_i, v_i, \beta) &= u_i v_i r_0(t) \exp(x_i^T \beta) \\
    \lambda_i(t \mid u_i, \gamma) &= u_i \lambda_0(t) \exp(x_i^T \gamma)
\end{align*}
\]

for rehospitalization and mortality, respectively. Here, \(r_0(t)\) and \(\lambda_0(t)\) are baseline
hazard functions, and $x$, $\beta$ and $\gamma$ are defined as in the “Model” section. The random effects distributions are specified as follows: $v_i \sim \text{Gamma}(1/\rho, 1/\rho)$ and $u_i \sim \text{Gamma}(1/\epsilon, 1/\epsilon)$ independently for $i = 1, \ldots, L$.

The comparison of the joint frailty model results in Table S7 of the Supplemental material with our results in Figure 9 shows that both models find that late-stage tumours are associated with shorter survival. For the other covariates, the comparison is inconclusive with the joint frailty model finding statistically significant effects where our model does not and vice versa.

Finally, the estimate of $\rho$ is 0.007 with a standard error of 0.001. This suggests heterogeneity between patients that is not explained by the covariates. The estimate of $\epsilon$ is 0.073 with a standard error of 0.13. This implies that the rate of rehospitalizations is positively associated with mortality. These results are in line with the posterior clustering results from our model in Table 2, where Cluster 1 is characterized by both the longest gap times and the longest survival times.

**Bayesian semiparametric model from Paulon et al.**

For a more direct comparison, we consider the method proposed by Paulon et al. as it models the gap and survival times jointly using Bayesian non-parametric priors. Paulon et al. assume that conditionally on all parameters and random effects, the gap times are independent of both each other and the survival time. This contrasts with the temporal dependence between gap times in (3). Shared random effects induce dependence between different gap times of the same patient. Specifically, Paulon et al. assume $Y_{ij} \sim \mathcal{N}(x_i^T \beta + \alpha_i, \sigma_i^2)$ independently for $j = 1, \ldots, n_i + 1$ and $i = 1, \ldots, L$, and
\[ \log(S_i) \sim N(\alpha_i^T \gamma + \psi \alpha_i, \eta^2), \]  
(5)

independently for \( i = 1, \ldots, L \), where \( \alpha_i, \beta \) and \( \gamma \) are defined as in the “Model” section, and \( \alpha_i \) and \( \sigma_i \) are random effects. Paulon et al.\(^{19}\) do not model the total number of gap times \( N_i \) but assume that each patient has a censored \((n_i + 1)\)th log gap time \( Y_{(n_i+1)} \). They also assume a priori independence among \( \beta, \gamma, \psi, \eta \) and \( (\alpha, \sigma^2) \). The random effects \((\alpha_i, \sigma_i^2) \sim G\) independently for \( i = 1 \ldots L \) where \( G \sim \text{DP}(M, G_0) \) with \( M \sim \text{Uniform}(a_M, b_M) \) and \( G_0 = \mathcal{N}(0, a_0^2) \times \text{Inv-Gamma}(a_0, b_0) \). The priors on \( \beta, \gamma \) and \( \eta^2 \) are set as in the “Prior specification” section. Finally, \( \psi \sim \mathcal{N}(0, \psi_0^2) \).

In fitting this model to the colorectal cancer data, we specify the same \( \alpha \) and the same hyperparameters for the priors on \( \beta, \gamma \) and \( \eta^2 \) as in the “Data description and analysis” section. Furthermore, we set \( a_M = 0.3, b_M = 5, a_0^2 = 100, a_0 = 2.01, b_\sigma = 1.01 \) and \( \psi_0 = 100 \). This model yields conclusions that are largely consistent with those from our model. In particular, the posterior distributions on the coefficients in Figure S19 of the Supplemental material mimic our results in Figure 9 except for the effect of tumour stage on recurrence. Also, the posterior on \( \psi \) in (5) concentrates between 1.3 and 1.7 as per Figure S20 of the Supplemental material. This parameter captures the strength of the relationship between gap and survival times. Thus, the time between hospitalizations and survival has a positive association. This is consistent with the clustering results obtained from our model in Table 2. Lastly, the posterior on the number of clusters for this model and our model vary slightly, with a mode of three clusters for our model in Figure S18 of the Supplemental material while Figure S21 of the Supplemental material has the mode at five for the model from Paulon et al.\(^{19}\) This is not surprising as our model introduces more structure: a temporal model for the gap time, as well as process-specific frailty terms which are jointly modelled non-parametrically. Moreover, the number of recurrent events is a random quantity and object of inference, which also informs the dependence between the two processes, in addition to the truncation \( T_{N_i} \leq S_i \) of (3). In contrast, Paulon et al.\(^{19}\) can capture such dependence using only \( \psi \) in (5), and the gap times are conditionally independent given \( \psi \) and the remaining parameters, with the total number of gap times per individual assumed arbitrarily large.

Discussion

We introduce a joint model on recurrence and survival that explicitly treats the number of recurrent events \( N_i \) before the terminal event as a random variable and object of inference as \( N_i \) is often of interest in applications. Additionally, the explicit modelling of \( N_i \) as well as the specification of a joint distribution for the random effects of the recurrence and survival processes induces dependence between these processes. Moreover, temporal dependence among recurrent events is introduced through a first-order autoregressive process on the gap times. An extension to a more complex temporal structure is in principle straightforward. The model allows for estimating covariate effects on the recurrence and survival processes by introducing appropriate regression terms. Our model can readily accommodate a different prior on the number of recurrent events than the negative binomial distribution. For instance, an earlier version of this work\(^{19}\) used a Poisson distribution. The use of a non-parametric prior as random effects distribution allows for extra flexibility, patient heterogeneity and data-driven clustering of the patients.

We use log–normal kernels for the non-parametric survival and gap time distributions. The use of different kernels is computationally impracticable as then the evaluation of the normalization constant of \( p(Y_i, S_i | \cdot) \) becomes considerably more expensive. This evaluation is already the computational bottleneck of our method as it happens frequently as part of slice sampling at each iteration of the Gibbs sampler. The reasons why using different kernels is problematic are twofold: firstly, the log–normal kernel for the gap times enables the Fenton–Wilkinson\(^{31}\) approximation for the distribution of \( T_{N_i} \). Other kernels for the gap times might not enable such computationally efficient evaluations. For instance, Proposition 1 of El Bouanani and Ben-Azza\(^{39}\) gives rise to an approximation when using a Weibull kernel, which is computationally more involved than the Fenton–Wilkinson method. Secondly, the current combination of the Fenton–Wilkinson method and a log–normal survival kernel reduce the normalization constant of \( p(Y_i, S_i | \cdot) \) to a tail probability of a univariate Gaussian as detailed in Section S1.1 of the Supplemental material. A different survival kernel, say a Weibull kernel, would require computation which is an order of magnitude more expensive, such as numerical integration to evaluate the normalization constant, or further approximation.

The simulation study shows the effectiveness of posterior inference on the number of recurrent events \( N_i \). In comparison with the Cox proportional hazards model, the joint frailty model\(^6\) and the Bayesian semiparametric model from Paulon et al.\(^{19}\) yields consistent results, with few exceptions in the estimation of covariate effects. This discrepancy might be the result of the fact that these models have fewer parameters and assume a single patient population while our model detects multiple subpopulations. Moreover, in our model specification a further level of dependence is introduced through the distribution of the number of gap times. More flexibility, if required by the application, could be achieved.
by including also the hyperparameters of the distribution of the number of gap times in the non-parametric component of the model.

**Declaration of conflicting interest**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is partially supported by the Singapore Ministry of Health’s National Medical Research Council under its Open Fund – Young Individual Research Grant (OFYIRG19nov-0010).

**ORCID iD**

Willem van den Boom https://orcid.org/0000-0002-1777-3857

**Supplemental material**

Supplemental material for this article is available online. The code that generated the results in this paper is available at https://github.com/willemvandenboom/condi-recur.

**References**

1. Jencks SF, Williams MV and Coleman EA. Rehospitalizations among patients in the medicare fee-for-service program. *N Engl J Med* 2009; 360: 1418–1428.
2. McIlvennan CK, Eapen ZI and Allen LA. Hospital readmissions reduction program. *Circulation* 2015; 131: 1796–1803.
3. Zuckerman RB, Sheingold SH, Orav EJ et al. Readmissions, observation, and the hospital readmissions reduction program. *N Engl J Med* 2016; 374: 1543–1551.
4. Cook RJ and Lawless JF. *The statistical analysis of recurrent events*. New York: Springer, 2007.
5. Liu L, Wolfe RA and Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics* 2004; 60: 747–756.
6. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H et al. Joint frailty models for recurrent events and death using maximum penalized likelihood estimation: Application on cancer events. *Biostatistics* 2007; 8: 708–721.
7. Ye Y, Kalbfleisch JD and Schaubel DE. Semiparametric analysis of correlated recurrent and terminal events. *Biometrics* 2007; 63: 78–87.
8. Huang CY, Qin J and Wang MC. Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* 2010; 66: 39–49.
9. Sinha D, Maiti T, Ibrahim JG et al. Current methods for recurrent events data with dependent termination. *J Am Stat Assoc* 2008; 103: 866–878.
10. Ouyang B, Sinha D, Slate EH et al. Bayesian analysis of recurrent event with dependent termination: An application to a heart transplant study. *Stat Med* 2013; 32: 2629–2642.
11. Olesen AV and Parner ET. Correcting for selection using frailty models. *Stat Med* 2006; 25: 1672–1684.
12. Huang X and Liu L. A joint frailty model for survival and gap times between recurrent events. *Biometrics* 2007; 63: 389–397.
13. Yu Z and Liu L. A joint model of recurrent events and a terminal event with a nonparametric covariate function. *Stat Med* 2011; 30: 2683–2695.
14. Bao Y, Dai H, Wang T et al. A joint modelling approach for clustered recurrent events and death events. *J Appl Stat* 2012; 40: 123–140.
15. Liu L, Huang X, Yaroshinsky A et al. Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics* 2015; 72: 204–214.
16. Yu Z, Liu L, Bravata DM et al. Joint model of recurrent events and a terminal event with time-varying coefficients. *Biom J* 2013; 56: 183–197.
17. Li Z, Chinchilli VM and Wang M. A Bayesian joint model of recurrent events and a terminal event. *Biom J* 2019; 61: 187–202.
18. Li Z, Chinchilli VM and Wang M. A time-varying Bayesian joint hierarchical copula model for analysing recurrent events and a terminal event: an application to the cardiovascular health study. *J R Stat Soc C (Appl Stat)* 2019; 69: 151–166.
19. Paulon G, De Iorio M, Guglielmi A et al. Joint modeling of recurrent events and survival: A Bayesian non-parametric approach. *Biostatistics* 2018; 21: 1–14. DOI: 10.1093/biostatistics/kxy026. kxy026.
20. Tallarita M, De Iorio M, Guglielmi A et al. Bayesian autoregressive frailty models for inference in recurrent events. *Int J Biostat* 2020; 16: 1–18.
21. Kansagara D, Englander H, Salanitro A et al. Risk prediction models for hospital readmission: A systematic review. *JAMA* 2011; 306: 1688.
22. Futoma J, Morris J and Lucas J. A comparison of models for predicting early hospital readmissions. *J Biomed Inform* 2015; 56: 229–238.
23. Mahmoudi E, Kamdar N, Kim N et al. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ* 2020; **369**: m958.

24. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**: 711–732.

25. Neal RM. Slice sampling. *Ann Stat* 2003; **31**: 705–767.

26. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann Stat* 1973; **1**: 209–230.

27. Sethuraman J. A constructive definition of Dirichlet priors. *Stat Sin* 1994; **4**: 639–650.

28. Aalen OO and Husebye E. Statistical analysis of repeated events forming renewal processes. *Stat Med* 1991; **10**: 1227–1240.

29. Lo AY. On a class of Bayesian nonparametric estimates: I. density estimates. *Ann Stat* 1984; **12**: 351–357. DOI: 10.1214/aos/1176346412.

30. De Iorio M, Johnson WO, Müller P et al. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* 2009; **65**: 762–771.

31. Fenton L. The sum of log-normal probability distributions in scatter transmission systems. *IEEE Trans Commun* 1960; **8**: 57–67.

32. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat* 2000; **9**: 249–265.

33. González JR, Fernandez E, Moreno V et al. Sex differences in hospital readmission among colorectal cancer patients. *J Epidemiol Community Health* 2005; **59**: 506–511. DOI:10.1136/jech.2004.028902.

34. Rondeau V, Mazroui Y and Gonzalez JR. frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw* 2012; **47**: 1–28. DOI:10.18637/jss.v047.i04

35. Binder DA. Bayesian cluster analysis. *Biometrika* 1978; **65**: 31–38.

36. Lau JW and Green PJ. Bayesian model-based clustering procedures. *J Comput Graph Stat* 2007; **16**: 526–558.

37. Argiento R, Cremaschi A and Guglielmi A. A “density-based” algorithm for cluster analysis using species sampling Gaussian mixture models. *J Comput Graph Stat* 2014; **23**: 1126–1142.

38. van den Boom W, Tallarita M and De Iorio M. Bayesian joint modelling of recurrence and survival: a conditional approach, 2020. arXiv:2005.06819v1.

39. Bouanani FE and Ben-Azza H. Efficient performance evaluation for EGC, MRC and SC receivers over Weibull multipath fading channel. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Basel: Springer International Publishing, 2015. pp. 346–357. DOI:10.1007/978-3-319-24540-9_28.