

# Principled Estimation and Prediction with Competing Risks: a Bayesian Nonparametric Approach

Claudio Del Sole<sup>1</sup>      Antonio Lijoi<sup>2</sup>      Igor Prünster<sup>2</sup>

<sup>1</sup>Department of Economics, Management and Statistics, University of Milano-Bicocca

<sup>2</sup>Bocconi Institute for Data Science and Analytics, Bocconi University

## Abstract

Competing risks occur in survival analysis when multiple causes of death are present. They play a prominent role in several domains extending beyond biostatistics to encompass epidemiology, actuarial sciences, and reliability theory. This paper adopts a multi-state modeling framework to competing risks. We introduce a class of flexible nonparametric priors, defined through hierarchical completely random measures, to model the transition probabilities, and identify the specific (conditionally) conjugate member of this general class. Furthermore, we determine the joint marginal distribution of the data and of a latent random partition, and characterize the posterior distribution of the model. Leveraging these distributional results, we evaluate the predictive probability that a future event is of a specific type (e.g. death from a particular cause), as a function of the time at which the event occurs. The resulting function, derived on sound principles, is termed the *prediction curve*, and represents a major innovation in the literature. In addition, we provide posterior estimates for the survival function, and for the cause-specific incidence and subdistribution functions. Suitable simulation algorithms for posterior inference are also devised. The model's performance, as well as the algorithms' effectiveness, is evaluated through simulation studies. Finally, we illustrate our approach on clinical datasets.

*Keywords:* Bayesian Nonparametrics; Competing risks; Hierarchical Processes; Prediction Curve; Random Partition; Survival Analysis.

# 1 Introduction

Survival analysis must often account for multiple competing sources of risk, where the observed failure (or death) is determined by one among several causes (or types of events). For instance, in clinical studies, a fatal outcome after treatment may be due to a number of different causes, that are either related or unrelated to the treatment, and only one of those causes can be actually observed. *Competing events* or *competing risks* arise when the observation of an event of interest, and of the associated survival time, is potentially prevented by the occurrence of one of a set of distinct alternative events (Kalbfleisch and Prentice, 2002; Geskus, 2024). The natural application of competing risks analysis is in biomedical and clinical studies, in which subjects may die from different causes. However, competing risks play a prominent role in several domains, another notable setting being reliability-testing in industry, where the possible breakdown of a complex system is due to the failure of one of its components.

## 1.1 Approaches to competing risks

Throughout, the observed survival time is denoted by  $T \in \mathbb{R}^+$  and assumed to be a random variable (r.v.) with absolutely continuous distribution (with respect to the Lebesgue measure). Moreover,  $\Delta \in \{1, \dots, D\}$  is the categorical r.v. identifying the observed event type, or equivalently the cause of death/failure. For each event type  $\delta \in \{1, \dots, D\}$ , the *cause-specific hazard rate* at time  $t \geq 0$  is defined as

$$h_\delta(t) := \lim_{s \rightarrow 0} \frac{1}{s} \mathbb{P}(t \leq T < t + s, \Delta = \delta \mid T > t), \quad (1)$$

representing the instantaneous rate of occurrence of an event of type  $\delta$  at time  $t$ , given survival up to that time from *all* possible types of events. Hence, individuals experiencing a competing event are no longer at risk. The cause-specific hazard rates are the building blocks of several important quantities. For instance, the *survival function* of  $T$ , evaluated at time  $t \geq 0$ , is given by

$$S(t) := \mathbb{P}(T > t) = \exp\left(-\sum_{\delta=1}^D \int_0^t h_\delta(s) ds\right). \quad (2)$$

Similarly, the cause-specific *incidence functions*, representing the infinitesimal probability of occurrence of each event type, are defined in terms of the cause-specific hazard rates as

$$f_\delta(t) := \lim_{s \rightarrow 0} \frac{1}{s} \mathbb{P}(t \leq T < t + s, \Delta = \delta) = h_\delta(t) \exp\left(-\sum_{\ell=1}^D \int_0^t h_\ell(s) ds\right), \quad t \geq 0.$$

The probability of occurrence for each event type  $\delta$ , that is, the proportion  $\pi_\delta$  of subjects eventually experiencing  $\delta$ , is the limit of the corresponding *cumulative incidence function*, or *subdistribution*,

$$\pi_\delta = \mathbb{P}(\Delta = \delta) = \lim_{t \rightarrow \infty} \int_0^t f_\delta(s) ds,$$

with the obvious constraint  $\sum_{\delta=1}^D \pi_\delta = 1$ . This setup is consistent with the *multi-state approach* to statistical modeling of competing risks. Indeed, the cause-specific hazards can be seen as transition rates in a multi-state model (Andersen et al., 2002; Putter et al., 2007) for a Markov process  $\{M(t): t \geq 0\}$  with one transient state, ‘alive’, labeled as 0, and  $D$  absorbing states ‘death from cause  $\delta$ ’, with label  $\delta \in \{1, \dots, D\}$ . Hence,  $h_\delta$  is the transition intensity from 0 to  $\delta$ . Specifically, if  $P_{\ell,k}(0, t) = \mathbb{P}(M(t) = k \mid M(0) = \ell)$  denotes the probability of moving from state  $\ell$  to state  $k$  in the time interval  $(0, t)$ , then

$$P_{0,0}(0, t) = S(t), \quad P_{0,\delta}(0, t) = \int_0^t f_\delta(s) ds, \quad (3)$$

are respectively the survival function in (2) and the cause-specific cumulative incidence function.

An alternative approach to competing risks data (Crowder, 2012) considers latent, or *potential*, survival times  $\mathbf{Y}_{1:D} = (Y_1, \dots, Y_D)$ , one for each event type. They are assumed to be distinct almost surely (a.s.) and the observed survival time is  $T = \min\{Y_1, \dots, Y_D\}$ . Whilst intuitive, this approach has been criticized for its lack of plausibility and interpretability in biomedical applications, as it assumes all event types eventually occur for each individual (Geskus, 2024). Additionally, the marginal distributions of the latent survival times cannot be identified without further assumptions on the dependence structure (Kalbfleisch and Prentice, 2002). Even considering a dependent parametric model for  $\mathbf{Y}_{1:D}$ , observed data do not allow to discern this model from one with independent risks (Cox, 1959; Tsiatis, 1975; Crowder, 1991). For this reason, several authors suggest to focus only on the estimation of observable quantities, often assuming independence of competing events for convenience and interpretability. However, this approach is useful to construct simulated datasets; see e.g. Beyersmann et al., 2009 and Section 7.1. Section S1 outlines a Bayesian nonparametric approach to modelling latent failure times.

The statistical analysis of competing risks data typically focuses on: (i) the estimation of the cause-specific cumulative incidence functions, which allows to estimate the probabilities that each competing event occurs within some time interval; (ii) the association of these quantities with different treatments or predictors of interest (Fine and Gray, 1999); (iii) the estimation of the survival function  $S$  in (2). These statistical problems are widely discussed in the frequentist literature, as evidenced by numerous reviews and textbooks; see, e.g.,

Kalbfleisch and Prentice (2002), Lawless (2003), Crowder (2012), Geskus (2015, 2024), Cook and Lawless (2018). In contrast, the literature on Bayesian nonparametric approaches to competing risks is rather limited. Bayesian models are often based on gamma or beta process priors, which were successfully employed to univariate survival data (Dykstra and Laud, 1981; Lo and Weng, 1989; Hjort, 1990). Recently, Arfé et al. (2019) introduced a beta–Stacy process generalization (Walker and Muliere, 1997) for cumulative incidence functions on a discrete timescale. Xu et al. (2016) proposed a model for evaluating dynamic treatment regimes based on dependent Dirichlet processes, while Xu et al. (2020) develop a similar approach for semi–competing risks. Instead, independent gamma processes are considered in Oganisian et al. (2024) for modelling sequential treatment decisions with time-varying confounders. Further pointers and discussions can be found in Müller and Quintana (2004); Müller et al. (2015). An alternative construction based on Bayesian Additive Regression Trees was proposed in Sparapani et al. (2020); we discuss and compare it to our method in Section S7, also through numerical illustrations.

Unlike previous contributions, we take a comprehensive approach by specifying a prior on the transition probabilities of the associated multi–state model, successfully addressing all major inferential problems in the competing risks literature. Additionally, our principled framework leads to a significant breakthrough in prediction: we introduce the concept of *prediction curve*, which corresponds to the conditional probability of a future event being of a specific type given the past observations, as a function of the event’s occurrence time. Prediction curves are not only of foundational and methodological interest, but also offer great potential for practical decision–making due to their clear interpretability.

## 1.2 A Bayesian nonparametric model

Our model for competing risks data employs a dependent prior for the cause–specific hazard rates, which induces a prior on the transition probabilities  $P_{0,\delta}$  and  $P_{0,0}$  in (3). For each  $\delta \in \{1, \dots, D\}$ , assuming  $\tilde{h}_\delta$  is a random cause–specific hazard, the induced prior coincides with the distribution of

$$\tilde{P}_{0,\delta}(0, t) = \int_0^t \tilde{h}_\delta(s) \exp\left(-\sum_{\ell=1}^D \int_0^s \tilde{h}_\ell(u) du\right) ds,$$

while  $\tilde{P}_{0,0}(0, t)$  is expressed as in (2), replacing each  $h_\delta$  with its random counterpart  $\tilde{h}_\delta$ . For modeling the random hazard rates, we adopt a kernel mixture specification, the most popular approach in Bayesian Nonparametrics. In the univariate case, this strategy was pioneered by Dykstra and Laud (1981) and Lo and Weng (1989) using gamma processes, and unraveled in

its full potential by [Ishwaran and James \(2004\)](#) and [James \(2005\)](#) for generic kernels and completely random measures as mixing distributions. In our more complex multivariate setup, we consider dependent cause-specific mixture hazard rates by specifying dependent mixing measures. In particular, among the various possibilities ([Quintana et al., 2022](#)), we adopt a hierarchical specification and resort to hierarchical completely random measures ([Camerlenghi et al., 2021](#)), known for their analytical tractability. This dependent prior is consistent with an assumption of conditional independence of  $\tilde{h}_1, \dots, \tilde{h}_D$ , given a common random measure  $\tilde{\mu}_0$  at the hierarchy’s root, leading to two key advantages: (a) it induces unconditional dependence among the hazard rates; (b) it enables borrowing of information across different causes or types of events  $\delta \in \{1, \dots, D\}$ . This specification allows substantial flexibility in modeling the hazard rates without imposing restrictive dependence assumptions.

We provide a complete picture of the posterior behavior of the proposed model and identify the (conditionally) conjugate prior within our general class. The latter is based on generalized gamma completely random measures, and parallels the role played by the beta process ([Hjort, 1990](#)) in nonparametric models for cumulative hazards and by the beta-Stacy process ([Walker and Muliere, 1997](#)) in neutral-to-the-right models. Moreover, we obtain Bayesian estimates of the cause-specific cumulative incidence functions and of the survival function, along with a characterization of the latent random partition induced by the model. Leveraging the partition distribution, we introduce the concept of *prediction curve*, which corresponds to the probability of a future failure or death due to a specific cause, as a function of the event’s occurrence time,  $t \mapsto \mathbb{P}(\Delta_{n+1} = \delta \mid T_{n+1} = t, \text{ data})$ . Prediction curves exhibit a neat form that reflects the composition structure generated by hierarchical discrete random measures. This principled tool is a major strength of our approach; as a preliminary peek at its practical relevance and immediate interpretability, [Figure 1](#) displays the estimates of the prediction curves, with 0.95 pointwise credible bands, for the “melanoma dataset” analyzed in [Section S6](#). This dataset originates from a clinical study at the Odense University Hospital on stage I melanoma patients: the two competing risks are death from melanoma and from other causes.

### 1.3 Outline

[Section 2](#) introduces kernel mixtures with respect to random measures to construct a prior on cause-specific hazard rates, which induces a prior on the space of transition probabilities of the multi-state model. In [Section 3](#), we define hierarchical completely random measures, introduce the atoms’ marking approach as a key technical tool to derive the main results, and establish consistency results for specific kernel choices. [Section 4](#) investigates the latent

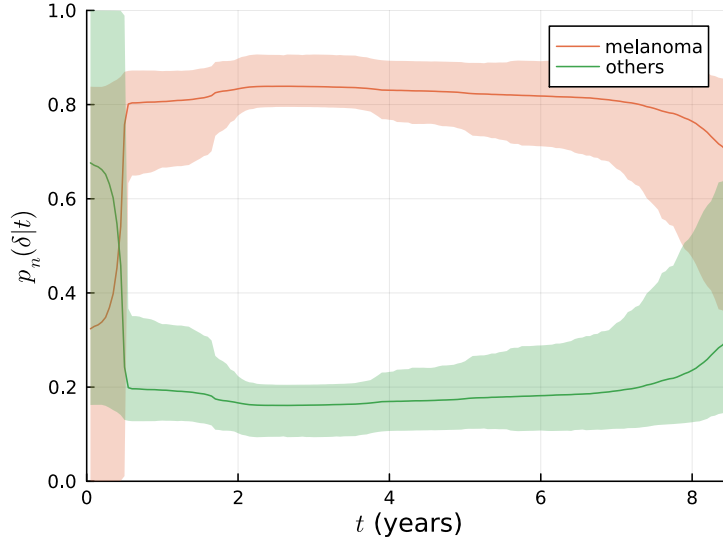


Figure 1: Prediction curves corresponding to melanoma and other causes of death for a future observation  $\Delta_{n+1}$ , as a function of the time  $T_{n+1} = t$  at which the event occurs.

clustering structure, which is pivotal for predictive inference, particularly in determining prediction curves, and for characterizing the posterior distribution, as detailed in Section 5. Simulation algorithms for posterior inference are devised in Section 6. Finally, in Section 7, we present two simulation studies and a clinical application to evaluate the model’s performance and the effectiveness of our samplers. References to the Supplementary Material are identified by the prefix ‘S’, e.g. Section S1. The Julia implementation for the proposed algorithms is available at: [github.com/clauidelsole/CompetingRisks.jl](https://github.com/clauidelsole/CompetingRisks.jl).

## 2 Cause-specific hazards and prediction curves

Kernel mixtures represent the most popular Bayesian nonparametric approach for exchangeable single-cause, and possibly right-censored, survival data (Dykstra and Laud, 1981; James, 2005). In our setting, such a mixture specification applies to the cause-specific random hazard rates

$$\tilde{h}_\delta(t) = \int_{\mathbb{X}} k(t; x) \tilde{\mu}_\delta(dx), \quad \delta \in \{1, \dots, D\}, \quad (4)$$

where  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  are random measures on  $\mathbb{X}$  and  $k: \mathbb{R}^+ \times \mathbb{X} \mapsto \mathbb{R}^+$  is a deterministic non-negative kernel. The distribution of  $\tilde{h}_\delta$  serves as the prior on the space of hazards corresponding to cause  $\delta$ . However, when extending from a single-cause to the multivariate setup with multiple competing risks, we would like to specify dependent cause-specific mixture hazards. This is achieved by acting at the level of the mixing measures: we assume  $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_D)$

to be a vector of dependent random measures with hierarchical dependence structure, which corresponds to a hCRM; see Sections S2 and 3.1. Consequently, the prior on the transition probabilities  $P_{0,\delta}$  and  $P_{0,0}$  is expressed in terms of  $\tilde{\boldsymbol{\mu}}$  as the law of  $\{\tilde{P}_{0,\delta}(0, t) : t \geq 0\}$  and  $\{\tilde{P}_{0,0}(0, t) : t \geq 0\}$ , respectively, where

$$\begin{aligned}\tilde{P}_{0,\delta}(0, t) &= \int_0^t \int_{\mathbb{X}} k(s; x) \tilde{\mu}_\delta(dx) \exp\left(-\sum_{\ell=1}^D \int_0^s \int_{\mathbb{X}} k(u; x) \tilde{\mu}_\ell(dx) du\right) ds, \\ \tilde{P}_{0,0}(0, t) &= \tilde{S}(t) = \exp\left(-\sum_{\ell=1}^D \int_0^t \int_{\mathbb{X}} k(s; x) \tilde{\mu}_\ell(dx) ds\right).\end{aligned}\quad (5)$$

For increasing hazards, the standard choice is the Dykstra–Laud kernel  $k(t; x) = \gamma(x) \mathbb{1}_{\{t \geq x\}}$ , with  $\mathbb{X} = \mathbb{R}^+$  and  $\gamma(\cdot)$  a positive right-continuous function, introduced in Dykstra and Laud (1981). Typically,  $\gamma(x) = \gamma$  for any  $x \in \mathbb{R}^+$  is assumed, a simplifying assumption we also adopt; however, results can be easily extended to any choice of  $\gamma(\cdot)$ . Alternative kernel choices leading to different shapes of the hazard rates are the rectangular kernel  $k(t; x) = \gamma(x) \mathbb{1}_{\{0 \leq t-x \leq \tau\}}$ , with bandwidth  $\tau > 0$ , and the Ornstein-Uhlenbeck kernel  $k(t; x) = \sqrt{2\kappa} \exp(-\kappa(t-x)) \mathbb{1}_{\{t \geq x\}}$ , with rate  $\kappa > 0$  (Ishwaran and James, 2004; Peccati and Prünster, 2008; De Blasi et al., 2009). Further methodological and computational investigations can be found, e.g., in Ishwaran and James (2004); Nieto-Barajas and Walker (2004); Catalano et al. (2020).

We consider the transition probabilities (5) to model competing risks data. Specifically, for exchangeable observations  $(T_i, \Delta_i)$ , where  $T_i \in \mathbb{R}^+$  is the  $i$ -th survival time and  $\Delta_i \in \{1, \dots, D\}$  is the corresponding event type, our nonparametric model takes the form

$$(T_1, \Delta_1), \dots, (T_n, \Delta_n) \mid \tilde{\boldsymbol{\mu}} \stackrel{\text{iid}}{\sim} \tilde{p}, \quad \tilde{\boldsymbol{\mu}} \sim \text{hCRM}, \quad (6)$$

where for any  $i = 1, \dots, n$  and  $\delta \in \{1, \dots, D\}$ ,

$$\tilde{p}(dt, \delta) = \tilde{f}_\delta(t) dt = \int_{\mathbb{X}} k(t; x) \tilde{\mu}_\delta(dx) \exp\left(-\sum_{\ell=1}^D \int_0^t \int_{\mathbb{X}} k(s; x) \tilde{\mu}_\ell(dx) ds\right) dt. \quad (7)$$

The prior distribution on  $\tilde{p}$  is thus induced by the distribution of  $\tilde{\boldsymbol{\mu}}$ , characterized in (10). As a consequence, the prior probability of observing an event of type  $\delta$ , conditionally on  $T$  and  $\tilde{\boldsymbol{\mu}}$ , is

$$\mathbb{P}(\Delta_i = \delta \mid T_i = t, \tilde{\boldsymbol{\mu}}) = \frac{\int_{\mathbb{X}} k(t; x) \tilde{\mu}_\delta(dx)}{\sum_{\ell=1}^D \int_{\mathbb{X}} k(t; x) \tilde{\mu}_\ell(dx)}, \quad (8)$$

for any  $\delta \in \{1, \dots, D\}$  and  $t > 0$ . In the literature on multi-state models, the ratios of

hazards in (8) correspond to the transition probabilities given the transition time and model parameters (Cook and Lawless, 2018, Section 2.5). These are useful for simulation purposes: indeed, the conditional distribution of the event type is multinomial. Importantly, from a Bayesian perspective, the function  $t \mapsto \mathbb{P}(\Delta_i = \delta \mid T_i = t)$  can be interpreted as the prior predictive probability of observing an event of a given type  $\delta$  as the survival time  $T_i$  varies. A primary goal of the paper, achieved in Section 4, is the evaluation of the updated (posterior) function, given the data, which we call the *prediction curve*.

**Definition 1.** The *prediction curve* associated to cause  $\delta \in \{1, \dots, D\}$ , for  $n \geq 1$ , is the function

$$t \mapsto p_n(\delta \mid t) = \mathbb{P}(\Delta_{n+1} = \delta \mid (T_i, \Delta_i)_{i=1}^n, T_{n+1} = t). \quad (9)$$

Hence, conditionally on the observed data, the prediction curve  $p_n(\delta \mid t)$  describes the probability of the next event being of type  $\delta$  as a function of the survival time  $T_{n+1}$ . This serves as a valuable tool for evaluating the relative probabilities of the different event types over time.

The rest of the paper develops a comprehensive distributional theory for this class of models including: (i) a characterization of the marginal distribution of data and latent variables, essential for determining the prediction curve (9); (ii) the posterior distribution of  $\tilde{p}$  in (7), enabling Bayesian estimation and uncertainty quantification of both the survival function and cause-specific incidence functions. The fundamental technical tools to achieve these goals are provided in the next section.

### 3 Dependent prior specification

We construct a tractable and effective dependent prior for  $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_D)$ , inducing a prior on the space of cause-specific hazard rates and transition probabilities, as described in Section 2. Specifically, we adopt hierarchical structures, a popular approach for introducing dependence among *completely random measures* (CRMs). Then, we introduce the marked process, investigate some key prior properties, and provide a consistency result. An overview of CRMs is given in Section S2.

#### 3.1 Hierarchical completely random measures

A prominent research stream in Bayesian Nonparametrics focuses on heterogeneous data or latent structures associated with different yet related experimental conditions. These settings can be framed within partial exchangeability, which assumes within-group homogeneity and

across-group heterogeneity. Success stories of this approach include topic modeling, ANOVA, change-point detection, reliability analysis, species sampling and numerous biostatistics and bioinformatics applications (see [Quintana et al., 2022](#), for a recent review). A popular strategy to model heterogeneity nonparametrically leverages vectors of dependent CRMs, as prior dependence among CRMs enables borrowing of information across different groups. Several effective proposals have appeared in the literature. In survival analysis, [Lijoi and Nipoti \(2014\)](#) model each vector component as a superposition of a common and an independent idiosyncratic CRMs; similar additive structures are employed in [Müller et al. \(2004\)](#) and [Lijoi et al. \(2014\)](#) to model dependent random probabilities. More complex constructions based on multivariate Lévy intensities are found in [Epifani and Lijoi \(2010\)](#) and [Riva-Palacio and Leisen \(2018, 2021\)](#), relying on Lévy copulas, and in [Griffin and Leisen \(2017\)](#), where compound random measures (CoRMs) are introduced. CoRMs represent a general class of dependent random measures encompassing both superpositions of CRMs ([Lijoi and Nipoti, 2014; Lijoi et al., 2014](#)) and weighted CRMs ([Chen et al., 2013](#)). See [Riva-Palacio et al. \(2022\)](#) for an application to survival regression models. Instead, [Nipoti et al. \(2018\)](#) define group-specific baseline hazard rates in frailty models employing independent CRMs. Further Bayesian nonparametric constructions of dependent random measures with biological and medical applications, though not directly based on CRMs, can be found in [Jara et al. \(2010\); Jara and Hanson \(2011\); Soriano and Ma \(2019\); Cassese et al. \(2019\); Christensen and Ma \(2020\)](#).

Unlike contributions above, we employ dependent random measures in an exchangeable framework to jointly model survival times  $T_i$  and competing causes  $\Delta_i$ . Specifically, we focus on hierarchical structures that satisfy a conditional independence assumption, paralleling the standard Bayesian paradigm arising from de Finetti’s representation theorem for exchangeable random elements. A vector of random measures  $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_D)$  has hierarchical structure if  $\tilde{\mu}_1, \dots, \tilde{\mu}_D \mid \tilde{\mu}_0 \stackrel{\text{iid}}{\sim} \tilde{\mathcal{M}}_{\tilde{\mu}_0}$  and  $\tilde{\mu}_0 \sim \mathcal{M}_0$ , where  $\tilde{\mathcal{M}}_{\tilde{\mu}_0}$  is the conditional distribution of each  $\tilde{\mu}_\delta$  given the random measure  $\tilde{\mu}_0$ , which is the hierarchy’s root with distribution  $\mathcal{M}_0$ . The idea of infinite-dimensional hierarchical constructions is due to [Teh et al. \(2006\)](#), who introduced the *hierarchical Dirichlet process*. A general distribution theory for hierarchical random probability measures was developed in [Camerlenghi et al. \(2019\)](#). Instead, we build on the hierarchical construction of CRMs in [Camerlenghi et al. \(2021\)](#), considering a more general formulation.

**Definition 2.** Let  $\tilde{\mu}_0$  be a CRM on  $\mathbb{X}$  with Lévy intensity  $\nu_0(ds, dx) = \rho_0(ds) \Lambda_0(dx)$ , for some  $\sigma$ -finite diffuse  $\Lambda_0$  on  $\mathbb{X}$ , which is termed base measure. A vector of *hierarchical*

completely random measures (hCRM)  $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_D)$  is defined as

$$\tilde{\mu}_1, \dots, \tilde{\mu}_D \mid \tilde{\mu}_0 \stackrel{\text{iid}}{\sim} \text{CRM}(\tilde{\nu}), \quad \tilde{\mu}_0 \sim \text{CRM}(\nu_0), \quad (10)$$

where  $\tilde{\nu}(ds, dx) = \rho(ds) \tilde{\mu}_0(dx)$  and  $\text{CRM}(\nu)$  denotes a CRM with intensity measure  $\nu$ .

Partially exchangeable data are naturally suited for the application of hCRMs, as shown in [Camerlenghi et al. \(2021\)](#) for estimating the marginal survival function of each population. In contrast, we apply hCRMs in a seemingly incompatible setup, namely exchangeable survival data with competing risks, as detailed in [Section 2](#). For illustration, we consider the *generalized gamma* hCRM, a large subclass of hCRM arising, for any  $\beta, \beta_0 > 0$  and  $\sigma, \sigma_0 \in [0, 1)$ , from the specification

$$\rho(ds) = \frac{1}{\Gamma(1-\sigma)} s^{-1-\sigma} e^{-\beta s} ds, \quad \rho_0(ds) = \frac{1}{\Gamma(1-\sigma_0)} s^{-1-\sigma_0} e^{-\beta_0 s} ds. \quad (11)$$

Special cases include the *gamma* hCRM, which is retrieved by setting  $\sigma = \sigma_0 = 0$ , the  $\sigma$ -stable hCRM, characterized by  $\beta = \beta_0 = 0$ , and *inverse Gaussian* hCRM, when  $\sigma = \sigma_0 = 1/2$ .

### 3.2 The marked process

A key feature of hierarchical constructions is that the random measures  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  share the same sequence of locations. Indeed, each measure can be represented as  $\tilde{\mu}_\delta = \sum_{h \geq 1} S_{\delta h} \delta_{X_{\delta h}^*}$  where, conditionally on  $\tilde{\mu}_0$ , the atoms  $(X_{\delta h}^*)_{h \geq 1}$  are i.i.d. from a distribution obtained as a transformation of  $\tilde{\mu}_0$ . Notably, the jumps  $S_{\delta h}$  and  $S_{\delta \ell}$  are indistinguishable if  $X_{\delta h}^* = X_{\delta \ell}^* = X_j^*$ , i.e. if they are associated to the same atom from  $\tilde{\mu}_0$ : indeed, the actual jump of  $\tilde{\mu}_\delta$  at  $X_j^*$  depends on  $S_{\delta h} + S_{\delta \ell}$ , and the individual components cannot be disentangled from their sum. We overcome this issue and considerably simplify the treatment of hCRMs by associating to each jump  $S_{\delta h}$  a diffuse independent mark  $Z_{\delta h}^*$ , which acts as unique label. This is achieved by extending  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  to the enlarged space  $[0, 1] \times \mathbb{X}$  with corresponding augmented intensity  $\tilde{\nu}^e(ds, dz, dx) = \rho(ds) H(dz) \tilde{\mu}_0(dx)$ , where  $H$  is an arbitrary diffuse probability measure on  $[0, 1]$ . Consequently, each extended random measure at the bottom of the hierarchy is

$$\tilde{\mu}_\delta^e = \sum_{h \geq 1} S_{\delta h} \delta_{(Z_{\delta h}^*, X_{\delta h}^*)}, \quad \delta \in \{1, \dots, D\}, \quad (12)$$

where  $(S_{\delta h})_{h \geq 1}$  and  $(Z_{\delta h}^*)_{h \geq 1}$  are independent sequences of random jumps and marks, respectively, while  $(X_{\delta h}^*)_{h \geq 1}$  is a sequence of independent locations whose distribution depends on  $\tilde{\mu}_0$ . The original  $\tilde{\mu}_\delta$  is recovered from  $\tilde{\mu}_\delta^e$  by marginalizing over the marks. From a technical perspective, this augmentation device enables a direct and straightforward proof strategy;

see Section S3. Moreover, the same technique may also simplify arguments in other setups, such as those in e.g. Camerlenghi et al. (2019, 2021).

### 3.3 Distributional properties and weak consistency

Firstly, we investigate the properness of the survival function in (5), namely  $\lim_{t \rightarrow +\infty} \tilde{S}(t) = 0$  a.s., which depends on both the kernel and the underlying random measures. Assuming a deterministic and non-negative kernel satisfying  $\int_0^t k(s; x) ds < \infty$  for any  $t \geq 0$  and  $x \in \mathbb{X}$ , the following result provides sufficient conditions.

**Proposition 3.** *The survival function  $\tilde{S}$  in (5) is proper if either of the following conditions holds:*

- (i) *the  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  and  $\tilde{\mu}_0$  are infinitely active and  $\lim_{t \rightarrow \infty} \int_0^t k(s; x) ds = \infty$  for every  $x \in \mathbb{X}$ ;*
- (ii) *the measure  $\Lambda_0$  is infinite and  $\lim_{t \rightarrow \infty} \int_0^t k(s; x) ds \geq C$  for every  $x \in \mathbb{X}$  and some  $C > 0$ .*

Specifically, the Dykstra-Laud kernel with generalized gamma hCRMs (11) satisfies (i), while the rectangular and Ornstein-Uhlenbeck kernels with  $\Lambda_0$  proportional to the Lebesgue measure meet (ii). A necessary and sufficient condition, although less transparent, is derived within the proof in Section S3.

As anticipated in Section 2, different kernels encode distinct prior properties in the model; e.g. the Dykstra-Laud kernel is suited to distributions with increasing hazards. This naturally motivates an investigation of the prior support under different kernel choices. Let  $\mathcal{P}$  be the prior induced by the distribution of the survival function  $\tilde{S}$  in (5) on the space of probability densities on  $\mathbb{R}^+$ . A density  $f_0$  belongs to the Kullback-Leibler support of  $\mathcal{P}$  if  $\mathcal{P}$  assigns positive probability to any Kullback-Leibler neighbourhood of  $f_0$ . Schwartz’s Theorem (see Ghosal and van der Vaart, 2017, Chapter 6) then guarantees weak consistency at any “true” data-generating density  $f_0$  in the Kullback-Leibler support of  $\mathcal{P}$ . Building on the results for single-cause mixture hazard models in De Blasi et al. (2009), we identify, for different kernel choices, broad classes of hazards  $h_0$  such that the corresponding survival density  $f_0$  belongs to the Kullback-Leibler support of  $\mathcal{P}$ , ensuring weak consistency. For simplicity, consider the generalized gamma hCRM in (11) with  $\Lambda_0$  proportional to the Lebesgue measure. These model assumptions guarantee some technical conditions, but do not compromise the generality of the result; refer to the proof in Section S3 for details.

**Theorem 4.** *Let  $\tilde{\mu}$  be the generalized gamma hCRM (11) and let  $\Lambda_0$  be proportional to the Lebesgue measure on  $\mathbb{R}^+$ . Then, the following hold:*

- (a) Dykstra–Laud kernel  $k(t; x) = \gamma \mathbf{1}_{\{t \geq x\}}$ : the prior  $\mathcal{P}$  is weakly consistent at each  $f_0$  such that  $\int_0^\infty t^2 f_0(t) dt < \infty$ , with  $h_0(0) = 0$  and  $h_0(t)$  strictly positive and non-decreasing for any  $t > 0$ ;
- (b) Rectangular kernel  $k(t; x) = \gamma \mathbf{1}_{\{0 \leq t-x \leq \tau\}}$ : if the bandwidth  $\tau$  is random and its prior assigns positive probability to  $[0, L]$  for some  $L > 0$ , then  $\mathcal{P}$  is weakly consistent at each  $f_0$  such that  $\int_0^\infty t f_0(t) dt < \infty$ , with  $h_0(0) = 0$  and  $h_0(t)$  strictly positive, bounded and Lipschitz for any  $t > 0$ ;
- (c) Ornstein-Uhlenbeck kernel  $k(t; x) = \sqrt{2\kappa} \exp(-\kappa(t-x)) \mathbf{1}_{\{t \geq x\}}$ : the prior  $\mathcal{P}$  is weakly consistent at each  $f_0$  such that  $\int_0^\infty t f_0(t) dt < \infty$ , with  $h_0(0) = 0$  and  $h_0(t)$  strictly positive and differentiable for any  $t > 0$ ; also, for any  $t > 0$  for which  $h_0'(t) < 0$ , one has  $|h_0'(t)|/h_0(t) < \kappa\sqrt{2\kappa}$ .

The class of hazard functions included in the Kullback-Leibler support of  $\mathcal{P}$ , for which weak consistency holds, depends essentially on the choice of the kernel. The rectangular kernel guarantees consistency for basically any bounded Lipschitz hazard, while the Ornstein-Uhlenbeck kernel covers any differentiable hazard subject to a certain upper bound on the local exponential decay. In contrast, the Dykstra-Laud kernel encompasses any non-decreasing hazard satisfying mild conditions. Remarkably, Theorem 4 remains valid in presence of independent right-censoring, provided the censoring distribution has full support (see De Blasi et al., 2009). An empirical confirmation of this consistency result for the Dykstra-Laud kernel is presented in Section S5.2.

Finally, the hierarchical structure in (10), combined with the conditional probability of cause  $\Delta_i$  given the survival time  $T_i$  in (8), implies a discrete uniform prior predictive for  $\Delta_i$ , that is  $\mathbb{P}(\Delta_i = \delta \mid T_i = t) = 1/D$  for  $\delta \in \{1, \dots, D\}$  and any  $t > 0$ . However, this uniformity can be easily removed by assuming either (i)  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  conditionally independent but not identically distributed, or (ii) distinct kernels for each event type. Closed-form expressions for the prior predictives  $\mathbb{P}(\Delta_i = \delta \mid T_i = t)$  are typically unavailable in these cases, but they can be approximated via numerical integration. The results in the following sections extend naturally to this more general setting, with the only drawback being the added notational complexity.

## 4 Latent clustering and prior predictive structures

The nonparametric mixture hazards model (6)–(7) induces a random partition, with a nested structure, of latent random elements. This is somewhat reminiscent of random partitions

induced by dependent discrete random probability measures, often sequentially represented using the popular *Chinese restaurant franchise* metaphor (Teh et al., 2006). Yet, the random partition here is fundamentally different, as it arises from random measures rather than random probability measures: this requires both a distinct analytical treatment and a novel interpretation. To study the distributional properties of the random partition associated to (6), we rely on a novel latent structure that facilitates analytical derivations. These latent variables are also pivotal for the posterior characterization in Section 5. The multiplicative intensity likelihood (Kalbfleisch and Prentice, 2002) for our model is the product of the directing random probability (7) evaluated at the observations, and is expressed as

$$\mathcal{L}(\tilde{\boldsymbol{\mu}}; \mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n}) = \prod_{i=1}^n \left[ \int_{\mathbb{X}} k(T_i; x_i) \tilde{\mu}_{\Delta_i}(dx_i) \exp \left( - \sum_{\ell=1}^D \int_0^{T_i} \int_{\mathbb{X}} k(s; y) \tilde{\mu}_{\ell}(dy) ds \right) \right]. \quad (13)$$

Throughout  $\mathbf{x}_{1:n} = (x_1, \dots, x_n)$  stands for vectors of dimension  $n$ , for any  $n \geq 1$ . We assume the observed survival times to be exact to avoid notational burden, although the inclusion of censored data is straightforward; see Section 7.3 and e.g. James (2005); Lijoi and Nipoti (2014). The likelihood (13) can be equivalently reparameterized in terms of  $\tilde{\boldsymbol{\mu}}^e = (\tilde{\mu}_1^e, \dots, \tilde{\mu}_D^e)$ :

$$\begin{aligned} & \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n}) \\ &= \prod_{i=1}^n \left[ \int_{[0,1] \times \mathbb{X}} k(T_i; x_i) \tilde{\mu}_{\Delta_i}^e(dz_i, dx_i) \exp \left( - \sum_{\ell=1}^D \int_0^{T_i} \int_{[0,1] \times \mathbb{X}} k(s; y) \tilde{\mu}_{\ell}^e(dw, dy) ds \right) \right]. \end{aligned}$$

Standard approaches for the Bayesian analysis of such type of likelihoods rely on suitable sequences of latent random elements  $\mathbf{X}_{1:n} = (X_1, \dots, X_n)$ , which essentially correspond to the (latent) locations sampled from the mixing measures and allow to remove the integration on  $\mathbb{X}$ . See, e.g., James (2005), Lijoi and Nipoti (2014) and Camerlenghi et al. (2021). In contrast, we resort here to a double augmentation: in addition to  $\mathbf{X}_{1:n}$ , we introduce  $[0, 1]$ -valued latent r.v.  $\mathbf{Z}_{1:n} = (Z_1, \dots, Z_n)$ , leading to the following doubly-augmented likelihood

$$\begin{aligned} & \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n}, \mathbf{X}_{1:n}, \mathbf{Z}_{1:n}) \\ &= \prod_{i=1}^n \left[ k(T_i; X_i) \tilde{\mu}_{\Delta_i}^e(dZ_i, dX_i) \exp \left( - \sum_{\ell=1}^D \int_0^{T_i} \int_{[0,1] \times \mathbb{X}} k(s; x) \tilde{\mu}_{\ell}^e(dz, dx) ds \right) \right]. \end{aligned}$$

The a.s. discreteness of  $\tilde{\boldsymbol{\mu}}^e$  naturally induces a random nested partition structure, which is identified by the tied values displayed by the latent sequences  $\mathbf{X}_{1:n}$  and  $\mathbf{Z}_{1:n}$ . Specifically, the coarser partition is induced by elements in  $\mathbf{X}_{1:n}$  and is denoted as

$$\Pi_{\mathbf{X}_{1:n}} = \{ C_j^X : j = 1, \dots, k \}, \quad \text{with } C_j^X = \{ i : X_i = X_j^* \},$$

for some  $k \in \{1, \dots, n\}$ . Hence,  $n_j = \text{card}(C_j^X)$  is the frequency of elements in  $\mathbf{X}_{1:n}$  coinciding with the  $j$ -th distinct value  $X_j^*$ , under the constraint  $n_1 + \dots + n_k = n$ . The series representation (12) of each  $\tilde{\mu}_\delta^e$  entails that these distinct values can be regarded as the latent locations of  $\tilde{\mu}_0$  at the root of the hierarchy. Since locations are shared among the  $\tilde{\mu}_\delta$ 's at the bottom of the hierarchy, the same location can be associated to different competing events. Let  $C_{\delta j}^X = \{i: (\Delta_i, X_i) = (\delta, X_j^*)\}$ , so that  $n_{\delta j} = \text{card}(C_{\delta j}^X)$  stands for the frequency of observations such that  $\Delta_i = \delta$  and  $X_i = X_j^*$ , with  $n_{1j} + \dots + n_{Dj} = n_j$ . For each  $j \in \{1, \dots, k\}$ , one may have  $n_{\delta j} \geq 1$  and  $n_{\delta' j} \geq 1$ , which corresponds to events of type  $\delta$  and  $\delta'$  sharing the same latent location  $X_j^*$ . In view of this,  $X_i = X_\ell$  does not necessarily imply  $\Delta_i = \Delta_\ell$ . Therefore, there is no structural relationship between the random latent partition  $\{C_j^X: j = 1, \dots, k\}$  and the observed partition  $\{C_\delta: \delta = 1, \dots, D\}$  encoded in  $\mathbf{\Delta}_{1:n}$ , where  $C_\delta = \{i: \Delta_i = \delta\}$ . On the contrary, this relationship is accounted for at the finer level of the partition, induced by elements in  $\mathbf{Z}_{1:n}$  and denoted as

$$\Pi_{\mathbf{Z}_{1:n}} = \{C_{\delta j h}^Z: \delta = 1, \dots, D; j = 1, \dots, k; h = 1, \dots, r_{\delta j}\},$$

with  $C_{\delta j h}^Z = \{i \in C_{\delta j}: Z_i = Z_{\delta j h}^*\}$ . This means that, for any  $\delta$  and  $j$  such that  $n_{\delta j} \geq 1$ , the  $n_{\delta j}$  observations for which  $\Delta_i = \delta$  and  $X_i = X_j^*$  are further partitioned into  $r_{\delta j}$  clusters according to the distinct values  $Z_{\delta j 1}^*, \dots, Z_{\delta j r_{\delta j}}^*$  in  $\mathbf{Z}_{1:n}$ ; clearly, if  $q_{\delta j h} = \text{card}(C_{\delta j h}^Z)$ , one has  $q_{\delta j 1} + \dots + q_{\delta j r_{\delta j}} = n_{\delta j}$ . The representation in (12) suggests a straightforward interpretation of these distinct  $Z_{\delta j 1}^*, \dots, Z_{\delta j r_{\delta j}}^*$  as the latent marks from the random measure  $\tilde{\mu}_\delta^e$  that are paired with the same location  $X_j^*$  from  $\tilde{\mu}_0$ . Hence,  $r_j = r_{1j} + \dots + r_{Dj}$  is the total number of different marks across random measures  $\tilde{\mu}_1^e, \dots, \tilde{\mu}_D^e$  associated with the same location  $X_j^*$ , for any latent group  $C_j^X$ . From these definitions, it is apparent that the partitions  $\Pi_{\mathbf{X}_{1:n}}$  and  $\Pi_{\mathbf{Z}_{1:n}}$  are nested. Specifically, if two observations  $T_i$  and  $T_\ell$  are tied in their marks, i.e.  $Z_i = Z_\ell$ , they must also share the same location, i.e.  $X_i = X_\ell$ ; the converse is not true. This description of the latent partition allows to express the augmented likelihood as

$$\begin{aligned} \mathcal{L}(\tilde{\mu}^e; \mathcal{F}_n^*) &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \exp\left(-\sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} K_n(x) \tilde{\mu}_\ell^e(dz, dx)\right) \\ &\quad \times \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \tilde{\mu}_\ell^e(dZ_{\ell j h}^*, dX_j^*)^{q_{\ell j h}}, \end{aligned} \quad (14)$$

where, in order to ease notation, we have set  $\mathcal{F}_n^* = (\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n}, \Pi_{\mathbf{X}_{1:n}}, \Pi_{\mathbf{Z}_{1:n}}, \mathbf{X}^*, \mathbf{Z}^*)$ , with  $\mathbf{X}^* = (X_1^*, \dots, X_k^*)$  and  $\mathbf{Z}^* = (Z_{\delta j h})_{\delta j h}$ , and

$$Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) = \prod_{j=1}^k \prod_{i \in C_j^X} k(T_i; X_j^*), \quad K_n(x) = K_n(x; \mathbf{T}_{1:n}) = \sum_{i=1}^n \int_0^{T_i} k(s; x) ds; \quad (15)$$

for details on the derivation of (14) from (13), see Section S3. The expression (14) neatly singles out the individual contributions to the likelihood due to the observed survival times and event types, and to the sequences of latent variables. Specifically, the survival times  $\mathbf{T}_{1:n}$  impact the likelihood through  $Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n})$  and  $K_n(x)$ . On the other hand, the event types  $\mathbf{\Delta}_{1:n}$  impose constraints on the possible configurations of the nested partition structure, which is encoded into the multiplicities  $(q_{\delta_{jh}})_{\delta_{jh}}$  at the finer  $\mathbf{Z}$ -level, and  $(r_{\delta_j})_{\delta_j}$  at the coarser  $\mathbf{X}$ -level.

The joint distribution of the observations  $(\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n})$  and the latent variables  $(\mathbf{X}_{1:n}, \mathbf{Z}_{1:n})$  is obtained by integrating out  $\tilde{\boldsymbol{\mu}}^e$  from the likelihood. To this end, let  $\psi$  and  $\tau$  be the Laplace exponent and cumulants of the random measures  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  at the bottom of the hierarchy, namely

$$\psi(u) = \int_{\mathbb{R}^+} (1 - e^{-us}) \rho(ds), \quad \tau(m; u) = \int_{\mathbb{R}^+} s^m e^{-us} \rho(ds). \quad (16)$$

The quantities  $\psi_0$  and  $\tau_0$  for  $\tilde{\mu}_0$  at the hierarchy's root are obtained by replacing  $\rho$  with  $\rho_0$ .

**Theorem 5.** *The joint marginal distribution of the observations  $(\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n})$  and the latent variables  $(\mathbf{X}_{1:n}, \mathbf{Z}_{1:n})$ , that is, the marginal distribution of  $\mathcal{F}_n^*$ , is given by*

$$\begin{aligned} \mathbb{P}(\mathcal{F}_n^*) &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \exp\left(-\int_{\mathbb{X}} \psi_0(D\psi(K_n(x))) \Lambda_0(dx)\right) \\ &\quad \times \prod_{j=1}^k \left( \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_n(X_j^*)) H(dZ_{\ell j h}^*) \right) \tau_0(r_j; D\psi(K_n(X_j^*))) \Lambda_0(dX_j^*). \end{aligned}$$

Note that, conditionally on  $\Pi_{\mathbf{Z}_{1:n}}$ , the distinct values  $\mathbf{Z}^*$  can be integrated out, since they are iid from the diffuse probability  $H$ , and hence have no impact on the latent partition structure. In contrast, the random partition  $\Pi_{\mathbf{X}_{1:n}}$  and the distinct values  $\mathbf{X}^*$  are not independent and both impact the marginal distribution. Additionally, the distinct values are tied to the observed survival times through  $Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n})$ . Hence, marginalizing with respect to  $\mathbf{Z}^*$ , Theorem 5 can be restated as

$$\begin{aligned} \mathbb{P}(\mathcal{F}_n) &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \exp\left(-\int_{\mathbb{X}} \psi_0(D\psi(K_n(x))) \Lambda_0(dx)\right) \\ &\quad \times \prod_{j=1}^k \left( \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_n(X_j^*)) \right) \tau_0(r_j; D\psi(K_n(X_j^*))) \Lambda_0(dX_j^*), \quad (17) \end{aligned}$$

where  $\mathcal{F}_n = (\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n}, \Pi_{\mathbf{X}_{1:n}}, \Pi_{\mathbf{Z}_{1:n}}, \mathbf{X}^*)$  is obtained from  $\mathcal{F}_n^*$  by removing the distinct values  $\mathbf{Z}^*$ . This expression serves as the starting point for deriving the computational results in

Section 6: indeed, a marginal Gibbs sampling scheme is obtained from the full conditional distributions derived from (17).

As an important by-product of Theorem 5, we determine the predictive distribution for the event type associated with the  $(n + 1)$ -th observation, that is the predictive distribution of  $\Delta_{n+1}$ , given the survival time  $T_{n+1}$  and  $\mathcal{F}_n$ .

**Theorem 6.** *The predictive distribution of  $\Delta_{n+1}$ , conditional on  $T_{n+1} = t$  and on  $\mathcal{F}_n$ , has probability mass function equal to*

$$\begin{aligned}
p(\delta \mid t, \mathcal{F}_n) \propto & \sum_{j=1}^k k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau(q_{\delta j h} + 1; K_{n+1}(X_j^*))}{\tau(q_{\delta j h}; K_{n+1}(X_j^*))} \\
& + \sum_{j=1}^k k(t; X_j^*) \tau(1; K_{n+1}(X_j^*)) \frac{\tau_0(r_j + 1; D\psi(K_{n+1}(X_j^*)))}{\tau_0(r_j; D\psi(K_{n+1}(X_j^*)))} \\
& + \int_{\mathbb{X}} k(t; x) \tau(1; K_{n+1}(x)) \tau_0(1; D\psi(K_{n+1}(x))) \Lambda_0(dx), \quad (18)
\end{aligned}$$

for any  $\delta \in \{1, \dots, D\}$ , where  $K_{n+1}(x) = K_n(x) + \int_0^t k(s; x) ds$ , since  $T_{n+1} = t$ .

The conditional distribution (18) is essential to derive the prediction curves  $p_n(\delta \mid \cdot)$  in Definition 1, which provide a measure of how likely each future cause of death  $\delta \in \{1, \dots, D\}$  is as a function of the survival time. Although this derivation is not analytically possible,  $p_n(\delta \mid \cdot)$  can be easily obtained from the MCMC algorithm in Section 6.

## 4.1 Predictive structure of generalized gamma mixtures

To conclude, we specialize Theorems 5 and 6 to the noteworthy subclass obtained by specifying  $\tilde{\boldsymbol{\mu}}$  as a *generalized gamma* hCRM (11) and  $k(\cdot; \cdot)$  as the Dykstra-Laud kernel. In this case, the Laplace exponent and cumulants (16) equal  $\psi(u) = ((\beta + u)^\sigma - \beta^\sigma)/\sigma$  and  $\tau(m; u) = (\beta + u)^{\sigma-m} \Gamma(m - \sigma)/\Gamma(1 - \sigma)$ , respectively. Note that  $\lim_{\sigma \downarrow 0} \psi(u) = \log(1 + u/\beta)$ , while  $\psi_0$  and  $\tau_0$  are obtained by replacing  $\beta$  and  $\sigma$  with  $\beta_0$  and  $\sigma_0$ . Moreover, with the Dykstra-Laud kernel  $k(t; x) = \gamma \mathbb{1}_{\{t \geq x\}}$ , the quantities in (15) reduce to

$$\begin{aligned}
Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) &= \prod_{j=1}^k \prod_{i \in C_j^X} \gamma \mathbb{1}_{\{T_i \geq X_j^*\}} = \gamma^n \prod_{j=1}^k \mathbb{1}_{\{\min_{i \in C_j^X} T_i \geq X_j^*\}}, \\
K_n(x) &= \gamma \sum_{i=1}^n \max(T_i - x, 0).
\end{aligned}$$

Therefore, the joint marginal distribution in (17) reduces to the expression

$$\begin{aligned}
\mathbb{P}(\mathcal{F}_n) &= \gamma^n \prod_{j=1}^k \left( \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \frac{\Gamma(q_{\ell j h} - \sigma)}{\Gamma(1 - \sigma)} \right) \frac{\Gamma(r_j - \sigma_0)}{\Gamma(1 - \sigma_0)} \mathbb{1}_{\{\min_{i \in C_j^X} T_i \geq X_j^*\}} \Lambda_0(dX_j^*) \\
&\quad \times \prod_{j=1}^k \left( \beta_0 + \frac{D}{\sigma} \left( (\beta + K_n(X_j^*))^\sigma - \beta^\sigma \right) \right)^{\sigma_0 - r_j} (\beta + K_n(X_j^*))^{r_j \sigma - n_j} \\
&\quad \times \exp \left( -\frac{1}{\sigma_0} \int_{\mathbb{R}^+} \left( \left( \beta_0 + \frac{D}{\sigma} \left( (\beta + K_n(x))^\sigma - \beta^\sigma \right) \right)^{\sigma_0} - \beta_0^{\sigma_0} \right) \Lambda_0(dx) \right).
\end{aligned}$$

Recall that the prior predictive distribution on  $\Delta$  is non-informative, in the sense that  $\Delta \mid (T = t)$  has a discrete uniform distribution on  $\{1, \dots, D\}$ , for every  $t \geq 0$ . The predictive distribution for  $\Delta_{n+1}$  in (18) allows to assess how the data have modified our prior belief; its expression boils down to

$$\begin{aligned}
p(\delta \mid t, \mathcal{F}_n) &\propto \sum_{j=1}^k \mathbb{1}_{\{t \geq X_j^*\}} \left\{ \frac{n_{\delta j} - r_{\delta j} \sigma}{\beta + K_{n+1}(X_j^*)} + \frac{(r_j - \sigma) (\beta + K_{n+1}(X_j^*))^{\sigma-1}}{\beta_0 + \frac{D}{\sigma} \left( (\beta + K_{n+1}(X_j^*))^\sigma - \beta^\sigma \right)} \right\} \\
&\quad + \int_0^t (\beta + K_{n+1}(x))^{\sigma-1} \left( \beta_0 + \frac{D}{\sigma} \left( (\beta + K_{n+1}(x))^\sigma - \beta^\sigma \right) \right)^{\sigma_0-1} \Lambda_0(dx),
\end{aligned}$$

for any  $\delta \in \{1, \dots, D\}$ , where notably  $K_{n+1}(x) = K_n(x) + \gamma \max(t - x, 0)$  for any  $x$  in  $\mathbb{R}^+$ .

## 5 Posterior analysis

### 5.1 General posterior characterization

The predictive system of Theorem 6 provides an indirect glimpse of the posterior of the hCRM model (6), though it was obtained from a probabilistic characterization of the underlying nested random partitions. In this section, we tackle the problem directly and derive a posterior characterization of  $\tilde{\boldsymbol{\mu}}$  and  $\tilde{\mu}_0$ , conditionally on both the observations  $(\mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n})$  and the latent variables  $(\mathbf{X}_{1:n}, \mathbf{Z}_{1:n})$ . This allows to estimate functionals of interest, such as the cause-specific incidence and subdistribution functions, as well as the survival function, and quantify the associated uncertainty. We show that  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  and  $\tilde{\mu}_0$  remain CRMs a posteriori, with significant updates, reflecting the information conveyed by the data. Indeed, each updated CRM consists of two independent components: a *non-homogeneous* CRM without fixed points of discontinuity and a finite number of random jumps at fixed locations. Moreover, the hierarchical structure is preserved a posteriori, as the random measures at the bottom of the hierarchy are still conditionally independent given the random measure at the root. Consequently, our model is structurally conjugate in the sense of Lijoi and Prünster (2010), although the shift from *homogeneous* to *nonhomogeneous* CRMs marks a significant

deviation from the prior structure; this effect does not occur with e.g. normalized random measures, whether in the exchangeable (James et al., 2009) or partially exchangeable (e.g. Camerlenghi et al., 2019) settings. Recall that  $\mathcal{F}_n = (\mathbf{T}_{1:n}, \Delta_{1:n}, \Pi_{\mathbf{X}_{1:n}}, \Pi_{\mathbf{Z}_{1:n}}, \mathbf{X}^*)$  and denote an updated quantity by the superscript  $+$ .

**Theorem 7.** *Given the hCRM model (6), then:*

- (a) *the posterior distribution of  $\tilde{\mu}_0$ , given  $\mathcal{F}_n$ , coincides with the distribution of*

$$\tilde{\mu}_0^+ = \tilde{\mu}_0^* + \sum_{j=1}^k V_j \delta_{X_j^*}, \quad (19)$$

where  $\tilde{\mu}_0^* \sim \text{CRM}(\nu_0^+)$  with  $\nu_0^+(ds, dx) = e^{-D\psi(K_n(x))s} \rho_0(ds) \Lambda_0(dx)$ , and  $V_1, \dots, V_k$  are mutually independent r.v.'s with densities  $f_{V_j}(s) ds \propto s^{r_j} \exp\{-D\psi(K_n(X_j^*))s\} \rho_0(ds)$ ;

- (b) *the posterior distribution of  $\tilde{\boldsymbol{\mu}}$ , given  $\mathcal{F}_n$  and  $\tilde{\mu}_0^+$  in (19), equals the distribution of a vector of random measures, whose  $\delta$ -th component is*

$$\tilde{\mu}_\delta^+ = \tilde{\mu}_\delta^* + \sum_{j=1}^k \sum_{h=1}^{r_{\delta j}} S_{\delta j h} \delta_{X_j^*}, \quad (20)$$

where  $\tilde{\mu}_\delta^* \mid \tilde{\mu}_0^+ \stackrel{\text{iid}}{\sim} \text{CRM}(\tilde{\nu}^+)$  with  $\tilde{\nu}^+(ds, dx) = e^{-K_n(x)s} \rho(ds) \tilde{\mu}_0^+(dx)$  and  $(S_{\delta j h})_{j h}$  are mutually independent r.v.'s with densities  $f_{S_{\delta j h}}(s) ds \propto s^{q_{\delta j h}} e^{-K_n(X_j^*)s} \rho(ds)$ .

Theorem 7 shows that *a posteriori* the hierarchical structure is preserved, though the CRMs at the hierarchy's bottom are no longer conditionally identical in distribution. Indeed,  $\tilde{\mu}_1^+, \dots, \tilde{\mu}_D^+$  are conditionally independent CRMs given  $\tilde{\mu}_0^+$ , but the component with fixed locations on the right-hand side of (20) depends on  $\delta$ , meaning conditional identity in distribution no longer holds. Nevertheless,  $\tilde{\boldsymbol{\mu}}^* = (\tilde{\mu}_1^*, \dots, \tilde{\mu}_D^*)$  remains a hCRM in the sense of Definition 2. Additionally, the result sheds light on the impact of the nested partition structure on the posterior distribution of hCRMs. At the hierarchy's root, the distribution of each jump  $V_j$  depends on the multiplicity  $r_j = r_{1j} + \dots + r_{Dj}$ , i.e. the number of distinct marks across  $\tilde{\mu}_1^e, \dots, \tilde{\mu}_D^e$  associated with the latent location  $X_j^*$ . At the hierarchy's bottom, the distribution of each jump  $S_{\delta j h}$  depends on the number of observations  $q_{\delta j h}$  associated with cause  $\delta$ , location  $X_j^*$  and latent mark  $Z_{\delta j h}^*$ . The interpretation is clear: (i) the posterior distribution of each  $\tilde{\mu}_\delta$  depends on the (finer) partition  $\Pi_{\mathbf{Z}_{1:n}}$  of the observations for which  $\Delta_i = \delta$ , i.e. on the association of the observations with the marks; (ii) the posterior distribution of the root measure  $\tilde{\mu}_0$  depends on how the finer partition  $\Pi_{\mathbf{Z}_{1:n}}$  is nested within the coarser partition  $\Pi_{\mathbf{X}_{1:n}}$ , i.e. on the association of the marks with the locations. Moreover, the components of

the posterior with random locations,  $\tilde{\mu}_1^*, \dots, \tilde{\mu}_D^*$  and  $\tilde{\mu}_0^*$ , are *non-homogeneous*, as evidenced by the structure of  $\tilde{\nu}^+$  and  $\nu_0^+$ . Indeed, the updates of their Lévy intensities: (i) involve non-homogeneous exponential tiltings of the prior homogeneous Lévy intensities; (ii) depend solely on the observed survival times  $\mathbf{T}_{1:n}$  via  $x \mapsto K_n(x) = K_n(x; \mathbf{T}_{1:n})$  in (15), and remain unaffected by the latent partition structure. Notably, at both levels of the hierarchy, the distributions of the jumps depend on the  $\mathbf{Z}$ -level partition and on the fixed latent location  $X_j^*$  through  $K_n(X_j^*)$ , for  $j \in \{1, \dots, k\}$ .

## 5.2 Posterior characterization of generalized gamma mixtures

We specialize the general characterization to the subclass of *generalized gamma* hCRMs (11) discussed in Section 4: the prior to posterior update becomes straightforward. Specifically,  $\tilde{\mu}_1^*, \dots, \tilde{\mu}_D^*$  and  $\tilde{\mu}_0^*$  remain (extended) generalized gamma CRMs, with the same  $\sigma$  and  $\sigma_0$  and updated rate parameters. The latter become functions on the latent space  $\mathbb{X}$ , due to nonhomogeneity *a posteriori*, and depend only on  $\mathbf{T}_{1:n}$  through  $K_n$ . Indeed,  $\beta^+(x) = \beta + K_n(x)$  and  $\beta_0^+(x) = \beta_0 + D\{(\beta + K_n(x))^\sigma - \beta^\sigma\}/\sigma$ . Furthermore, for each fixed latent location  $X_j^*$ , the jumps  $(S_{\delta_{jh}})_{\delta h}$  and  $V_j$  are gamma distributed

$$S_{\delta_{jh}} \sim \text{Gamma}(q_{\delta_{jh}} - \sigma, \beta^+(X_j^*)), \quad V_j \sim \text{Gamma}(r_j - \sigma_0, \beta_0^+(X_j^*)).$$

The generalized gamma hCRM can thus be regarded as the (conditionally) conjugate prior when used as mixing measure for the kernel mixtures defining cause-specific hazards in (6). Moreover, with the Dykstra-Laud kernel,  $K_n(x) = \gamma \sum_{i=1}^n \max(T_i - x, 0)$  is a non-increasing function and  $K_n(x) = 0$  for any  $x \geq \max_i T_i$ . This, combined with  $u \mapsto \psi(u)$  being non-decreasing, implies that the effect of the exponential tilting update in the posterior Lévy intensities  $\tilde{\nu}^+$  and  $\nu_0^+$  is larger for smaller values of  $x$  and becomes negligible for larger values of  $x$ . In other words, for any  $\varepsilon > 0$ , the function  $x \mapsto \rho((\varepsilon, +\infty)) - \int_\varepsilon^{+\infty} e^{-K_n(x)s} \rho(ds)$  is non-increasing and actually equals 0 if  $x \geq \max_i T_i$ . Hence, for large enough  $x$ , the prior and posterior Lévy intensities coincide. The same argument applies to  $\rho_0$  at the hierarchy's root. Moreover, each of the fixed latent locations  $X_1^*, \dots, X_k^*$  is necessarily smaller than the largest observed survival time, due to the kernel product term  $Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n})$ . Therefore, the observations do not provide any information beyond the largest survival time  $\max_i T_i$ , where the prior and posterior distributions of the random measures coincide. Similar considerations apply to other kernels mentioned in Section 2. See also De Blasi et al. (2009) for related discussions.

### 5.3 Conditional functionals estimation given the latent variables

Theorem 7 represents the key ingredient for the estimation of important functionals of  $\tilde{\boldsymbol{\mu}}$  such as the survival function (5) and the cause-specific incidence functions (7). Their posterior estimates, given the observations and the latent variables, are obtained by marginalization with respect to the posterior random measures. To this end, we conveniently single out the jump part of the posterior Lévy intensities  $\tilde{\nu}^+$  and  $\nu_0^+$  as  $\rho^+(ds|x) = \exp(-K_n(x)s)\rho(ds)$  and  $\rho_0^+(ds|x) = \exp(-D\psi(K_n(x))s)\rho_0(ds)$ , with corresponding Laplace exponents and cumulants given by

$$\psi_x^+(u) = \int_{\mathbb{R}^+} (1 - e^{-us}) \rho^+(ds|x), \quad \tau_x^+(m; u) = \int_{\mathbb{R}^+} s^m e^{-us} \rho^+(ds|x); \quad (21)$$

the quantities  $\psi_{0,x}^+$  and  $\tau_{0,x}^+$  are obtained replacing  $\rho^+$  with  $\rho_0^+$  in (21).

First, consider the survival function  $\tilde{S}(t)$  in (5). Since  $\tilde{S}$  is the exponential of a linear functional of  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ , its posterior estimate, under squared error loss, is an immediate consequence of Theorem 7.

**Proposition 8.** *The posterior estimate, given  $\mathcal{F}_n$ , of the survival function  $\tilde{S}(t)$  at  $t \geq 0$  is*

$$\begin{aligned} \mathbb{E} [\tilde{S}(t) | \mathcal{F}_n] &= \exp \left( - \int_{\mathbb{X}} \psi_{0,x}^+ (D\psi_x^+ (K_1(x))) \Lambda_0(dx) \right) \\ &\quad \times \prod_{j=1}^k \left( \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+ (q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+ (q_{\ell j h}; 0)} \right) \frac{\tau_{0, X_j^*}^+ (r_j; D\psi_{X_j^*}^+ (K_1(X_j^*)))}{\tau_{0, X_j^*}^+ (r_j; 0)}, \end{aligned}$$

where  $K_1(x) = K_1(x; t) = \int_0^t k(s; x) ds$ , for any  $t \geq 0$  and  $x$  in  $\mathbb{X}$ .

Moreover, we obtain a posterior estimate of the cause-specific incidence function  $\tilde{f}_\delta(t)$  in (7). Although its dependence on  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  is complex, marginalization remains feasible.

**Proposition 9.** *For any  $t \geq 0$  and  $\delta \in \{1, \dots, D\}$ , the posterior estimate of  $\tilde{p}(dt, \delta) = \tilde{f}_\delta(t) dt$  is*

$$\begin{aligned} \mathbb{E} [\tilde{p}(dt, \delta) | \mathcal{F}_n] &= \mathbb{E} [\tilde{S}(t) | \mathcal{F}_n] \left\{ \sum_{j=1}^k k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau_{X_j^*}^+ (q_{\delta j h} + 1; K_1(X_j^*))}{\tau_{X_j^*}^+ (q_{\delta j h}; K_1(X_j^*))} \right. \\ &\quad + \sum_{j=1}^k k(t; X_j^*) \tau_{X_j^*}^+ (1; K_1(X_j^*)) \frac{\tau_{0, X_j^*}^+ (r_j + 1; D\psi_{X_j^*}^+ (K_1(X_j^*)))}{\tau_{0, X_j^*}^+ (r_j; D\psi_{X_j^*}^+ (K_1(X_j^*)))} \\ &\quad \left. + \int_{\mathbb{X}} k(t; x) \tau_x^+ (1; K_1(x)) \tau_{0,x}^+ (1; D\psi_x^+ (K_1(x))) \Lambda_0(dx) \right\} dt. \end{aligned}$$

**Remark 10.** Proposition 9 also allows to evaluate the predictive distribution for the event type  $\Delta_{n+1}$  associated with the additional survival time  $T_{n+1}$  given  $(T_{n+1}, \mathcal{F}_n)$ , which unsurprisingly coincides with  $p(\delta | t, \mathcal{F}_n)$  in Theorem 6. Details are provided in Section S3.

Propositions 8 and 9 provide explicit posterior estimates, yet conditional on the latent variables  $(\Pi_{\mathbf{X}_{1:n}}, \Pi_{\mathbf{Z}_{1:n}}, \mathbf{X}^*)$ . To obtain posterior estimates depending only on the observations  $(\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n})$ , one needs to integrate these expressions with respect to the posterior distribution of these latent variables. This marginalization is performed via the Gibbs sampling scheme detailed in the next section.

## 6 Simulation algorithms for posterior inference

The latent partition structure identified in Section 4 is the key device to characterize marginal, predictive and posterior quantities, which are expressed conditionally on  $(\Pi_{\mathbf{X}_{1:n}}, \Pi_{\mathbf{Z}_{1:n}}, \mathbf{X}^*)$ . To obtain unconditional quantities, we employ a Gibbs sampling scheme that generates these latent variables conditionally on the observations. Specifically, we derive a marginal algorithm, which leverages the explicit expression of the law of  $\mathcal{F}_n^*$  obtained in Theorem 5 by integrating out the underlying hCRM. We also present a sampling procedure for generating trajectories of  $\tilde{\mu}_0^+$  and  $\tilde{\mu}^+$  from their distributions in Theorem 7, conditionally on the partition structure. These algorithms are tailored to the *generalized gamma* hCRM in Section S4.

### 6.1 Marginal sampling algorithm

The marginal likelihood in Theorem 5 is the starting point for devising a Gibbs sampler that generates the posterior latent variables  $(X_i, Z_i)$ , for  $i \in \{1, \dots, n\}$ . Marginal structures are typically characterized in the nonparametric literature via sequential urn schemes. In the exchangeable case, these are known as generalized Pólya urn schemes or Chinese restaurant processes, derived for the Dirichlet process in Blackwell and MacQueen (1973), the Pitman-Yor process in Pitman (1996) and normalized CRMs in James et al. (2009). In the partially exchangeable case, with a hierarchical dependence structure, they are known as Chinese restaurant franchise processes, and were introduced by Teh et al. (2006) for the hierarchical Dirichlet process and by Camerlenghi et al. (2019) for hierarchical normalized CRMs.

Although we also aim at an urn-type sequential description of the marginal structure, its derivation requires a fundamentally different approach. Indeed, since the underlying model does not build upon normalized (hierarchical) CRMs, we should rely on different tools as well as on the essential augmentation via the latent marks  $\mathbf{Z}_{1:n}$ . The full conditionals correspond to the predictive distributions deduced from Theorem 5 and the marginal of  $\mathcal{F}_n$

in (17). Specifically, the full conditional generating  $(X_i, Z_i)$  can be described in terms of the standard notation where  $q_{\delta jh}^{-i}$ ,  $r_{\delta j}^{-i}$ ,  $r_j^{-i}$  and  $k^{-i}$  have the same frequency interpretation as in Theorem 5 after removing  $(X_i, Z_i)$ . Similarly, we set  $\mathcal{F}_n^{-i} = (\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n}, \mathbf{X}_{1:n}^{-i}, \mathbf{Z}_{1:n}^{-i})$  with the  $i$ -th component removed from  $\mathbf{X}_{1:n}$  and  $\mathbf{Z}_{1:n}$ . Adopting this notation, the predictive distribution of  $(X_i, Z_i)$  is identified by the following cases:

- (1)  $X_i$  and  $Z_i$  coincide with conditioning latent variables, say  $X_i = X_j^*$  and  $Z_i = Z_{\delta jh}^*$ , for some  $j$  and  $h$  and such that  $\Delta_i = \delta$ , with probability

$$\mathbb{P}(X_i = X_j^*, Z_i = Z_{\delta jh}^* \mid \mathcal{F}_n^{-i}) \propto k(T_i; X_j^*) \frac{\tau(q_{\delta jh}^{-i} + 1; K_n(X_j^*))}{\tau(q_{\delta jh}^{-i}; K_n(X_j^*))};$$

- (2)  $X_i$  equals one of the tied values in  $\mathbf{X}_{1:n}^{-i}$ , say  $X_i = X_j^*$ , while  $Z_i$  takes on a new value, not included in  $\mathbf{Z}_{1:n}^{-i}$ , with probability

$$\mathbb{P}(X_i = X_j^*, Z_i = \text{'new'} \mid \mathcal{F}_n^{-i}) \propto k(T_i; X_j^*) \tau(1; K_n(X_j^*)) \frac{\tau_0(r_j^{-i} + 1; D\psi(K_n(X_j^*)))}{\tau_0(r_j^{-i}; D\psi(K_n(X_j^*)))};$$

- (3)  $X_i$  and  $Z_i$  take on new values, i.e. not included in  $\mathbf{X}_{1:n}^{-i}$  and  $\mathbf{Z}_{1:n}^{-i}$ , respectively, with probability

$$\mathbb{P}(X_i = \text{'new'}, Z_i = \text{'new'} \mid \mathcal{F}_n^{-i}) \propto \int_{\mathbb{X}} k(T_i; x) \tau(1; K_n(x)) \tau_0(1; D\psi(K_n(x))) \Lambda_0(dx).$$

Recall from Section 4 that the new value possibly assigned to the latent mark  $Z_i$  does not affect the predictive distributions and, hence, the full conditionals: the only relevant information in  $\mathbf{Z}_{1:n}$  is the induced partition structure. In contrast, the new value possibly assigned to the latent  $X_i$  in case (3) directly impacts the allocation probabilities in the Gibbs sampling scheme, and is sampled from the diffuse probability measure proportional to  $k(T_i; x) \tau(1; K_n(x)) \tau_0(1; D\psi(K_n(x))) \Lambda_0(dx)$ ; with the kernel choices considered in Section 2, this reduces to a univariate density on a bounded interval. Each iteration of the algorithm yields an update of the latent nested partition structure  $(\Pi_{\mathbf{X}_{1:n}}, \Pi_{\mathbf{Z}_{1:n}})$ . A further *acceleration step*, which is known to improve mixing (Escobar and West, 1998; MacEachern, 1998), consists in resampling independently each latent location  $X_j^*$  for  $j \in \{1, \dots, k\}$ , given the observations and latent partition structure, proportionally to

$$\left( \prod_{i \in C_j^X} k(T_i; x) \right) \left( \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell jh}; K_n(x)) \right) \tau_0(r_j; D\psi(K_n(x))) \Lambda_0(dx).$$

Finally, hyperpriors are usually specified for the parameters of the Lévy intensities  $\tilde{\nu}$  and  $\nu_0$  of the underlying hCRM (10) and the kernel  $k(\cdot; \cdot)$ . In particular, if  $\Lambda_0(dx) = \theta dx$  is proportional to the Lebesgue measure, a gamma prior on  $\theta$  is conjugate to the marginal distribution (17). Indeed, if  $\theta \sim \text{Gamma}(a, b)$ , its full conditional depends only on the observed survival times  $\mathbf{T}_{1:n}$  and the number  $k$  of distinct latent locations, being

$$\text{Gamma}\left(a + k, b + \int_{\mathbb{X}} \psi_0(D\psi(K_n(x))) dx\right).$$

Posterior estimates of the survival function, cause-specific incidence functions and prediction curves are then readily obtained. Specifically, at each iteration, (conditional) posterior estimates are computed by plugging the updated latent variables, resampled from their posterior distribution, into the expressions of Propositions 8 and 9 and Theorem 6; thus, unconditioned posterior estimates are obtained as Monte Carlo averages.

## 6.2 Posterior conditional sampling of random measures

Marginal algorithms cannot be used for uncertainty quantification around posterior estimates of functionals involving the random measures, as marginalizing over  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  averages out the associated uncertainty. Indeed, the conditional estimates computed at each iteration as in Section 5.3 are not a sample from the posterior of the corresponding functionals, but rather of their conditional expectations, given the latent variables. To address this issue, one needs to generate samples from the actual posterior of the random measures: this properly accounts for posterior uncertainty and enables the computation of reliable credible bands. Specifically, at each iteration of the marginal algorithm, we sample independently from the posterior hCRMs, given the observations and latent variables. Based on these samples, valid posterior draws of functionals can be computed from their definitions (5) and (7).

The practical implementation of this method requires sampling from the posterior hCRMs in Theorem 7. The Ferguson-Klass algorithm (Ferguson and Klass, 1972) is the most popular approach for generating realizations of CRMs, as it ensures their jumps are a.s. decreasing, allowing to control the associated approximation error. See Barrios et al. (2013) and references therein, and Brück (2025) for a novel approach to sampling based on generative neural networks. For non-homogeneous CRMs, Wolpert and Ickstadt (1998) proposed a simpler alternative, which however does not generate a.s. decreasing jumps. Nevertheless, when non-homogeneity is induced by exponential tilting of a homogeneous prior, as in our case, the sequence of jumps can still be upper-bounded by a decreasing sequence obtained from the prior (Camerlenghi et al., 2021). This theoretical guarantee on the truncation error underpins our sampling method.

Consider the non-homogeneous random measure  $\tilde{\mu}_0^+$  at the root of the posterior hierarchy (Theorem 7), and perform the following steps:

- (i<sub>0</sub>) generate the sequence of random locations  $(X_h^{(0)})_{h \geq 1}$  independently from  $\Lambda_0/\Lambda_0(\mathbb{X})$ ;
- (ii<sub>0</sub>) generate the sequence  $(N_h^{(0)})_{h \geq 1}$  of jump times of a unit-rate Poisson process, namely  $N_0^{(0)} = 0$  and, for  $h \geq 1$ , the increments  $N_h^{(0)} - N_{h-1}^{(0)}$  are i.i.d. exponentially distributed with unit mean;
- (iii<sub>0</sub>) generate the sequence of random jumps  $(V_h^{(0)})_{h \geq 1}$  by solving the equations

$$N_h^{(0)} = \Lambda_0(\mathbb{X}) \int_{V_h^{(0)}}^{\infty} \exp\left(-D\psi\left(K_n\left(X_h^{(0)}\right)\right) s\right) \rho_0(ds), \quad h \geq 1.$$

Since one cannot generate infinite sequences in (i<sub>0</sub>)–(iii<sub>0</sub>), the algorithm is stopped as soon as  $N_{H_0}^{(0)} > \Lambda_0(\mathbb{X}) \int_{\varepsilon}^{\infty} \rho_0(ds)$ , for some  $H_0 \geq 1$ . This guarantees each of the discarded jumps  $V_{H_0+1}^{(0)}, V_{H_0+2}^{(0)}, \dots$  to be smaller than a fixed threshold  $\varepsilon > 0$ . Moreover, if  $\Lambda_0$  is an infinite measure, the sampling procedure must be restricted to a suitable subset  $\mathbb{S} \subseteq \mathbb{X}$  such that  $\Lambda_0(\mathbb{S}) < \infty$ ; for the kernels in Section 2, with  $\mathbb{X} = \mathbb{R}^+$ , we consider  $\mathbb{S} = [0, T_{\max}]$  where  $T_{\max}$  is larger than the largest observed survival time. An approximate realization of the root random measure  $\tilde{\mu}_0$  from its posterior distribution is thus

$$\tilde{\mu}_0^+ \approx \sum_{j=1}^k V_j \delta_{X_j^*} + \sum_{h=1}^{H_0} V_h^{(0)} \delta_{X_h^{(0)}}, \quad (22)$$

where the jumps  $V_1, \dots, V_k$  at the fixed locations  $X_1^*, \dots, X_k^*$  are sampled from the corresponding densities in Theorem 7. Note that, since  $\Lambda_0$  is diffuse, the random locations  $X_1^{(0)}, \dots, X_{H_0}^{(0)}$  are distinct a.s., and also differ from the fixed locations  $X_1^*, \dots, X_k^*$ . Similarly, at the bottom of the posterior hierarchy, for each non-homogeneous  $\tilde{\mu}_\delta^+$  with  $\delta \in \{1, \dots, D\}$ :

- (i<sub>δ</sub>) generate  $(X_h^{(\delta)})_{h \geq 1}$  independently from  $\tilde{\mu}_0^+/\tilde{\mu}_0^+(\mathbb{X})$ ;
- (ii<sub>δ</sub>) generate the sequence  $(N_h^{(\delta)})_{h \geq 1}$  of jump times of a unit-rate Poisson process as in ii<sub>0</sub>);
- (iii<sub>δ</sub>) generate  $(S_h^{(\delta)})_{h \geq 1}$  by solving the equations

$$N_h^{(\delta)} = \tilde{\mu}_0^+(\mathbb{X}) \int_{S_h^{(\delta)}}^{\infty} e^{-K_n(X_h^{(\delta)})s} \rho(ds), \quad h \geq 1,$$

The procedure above is stopped as soon as  $N_{H_\delta}^{(\delta)} > \tilde{\mu}_0^+(\mathbb{X}) \int_{\varepsilon}^{\infty} \rho(ds)$ , for some  $H_\delta \geq 1$  and some fixed  $\varepsilon > 0$ . Clearly, in (i<sub>δ</sub>)–(iii<sub>δ</sub>) we replace  $\tilde{\mu}_0^+$  with its finite approximation on

the right-hand-side of (22). Hence, an approximate realization of  $\tilde{\mu}_\delta$  from its posterior distribution is given by

$$\tilde{\mu}_\delta^+ \approx \sum_{j=1}^k \sum_{h=1}^{r_{\delta j}} S_{\delta j h} \delta_{X_j^*} + \sum_{h=1}^{H_\delta} S_h^{(\delta)} \delta_{X_h^{(\delta)}}, \quad (23)$$

where the jumps at fixed locations are sampled as specified in Theorem 7.

**Remark 11.** The truncation of the infinite sequence of jumps in (23), and the conditioning on a truncated realization of  $\tilde{\mu}_0^+$ , entail two levels of approximation. Moreover, the random locations  $X_1^{(\delta)}, \dots, X_{H_\delta}^{(\delta)}$  display ties with positive probability, and take values among the finite collection of locations  $X_1^{(0)}, \dots, X_{H_0}^{(0)}$  and  $X_1^*, \dots, X_k^*$  characterizing the discrete measure approximating  $\tilde{\mu}_0^+$ . An alternative approach for sampling realizations of each  $\tilde{\mu}_\delta^+$  consists in sampling its  $k + H_0$  cumulative random jumps associated to the locations inherited from the approximation in (22). This exact sampling procedure avoids the approximation at the root of the hierarchy, and is explored for normalized random measures in Lijoi et al. (2020) and more recently in Catalano and Del Sole (2025).

## 7 Applications to synthetic and real datasets

We consider synthetic datasets to assess the performance of our Bayesian nonparametric model, compared to both frequentist and Bayesian alternatives. This evaluation underscores how the borrowing of information across hazard rates, induced by the hierarchical structure, enhances the accuracy of posterior estimates. Moreover, we present an application to a real dataset obtained from the bone marrow transplant registry of the European Blood and Marrow Transplant (EBMT) Group. Section S5 provides further details on the simulated datasets, diagnostic plots, and additional figures. Another application to a dataset of melanoma patients is reported in Section S6. Finally, we compare our method to that of Sparapani et al. (2020) on both synthetic and real datasets in Section S7.

### 7.1 A simulation study with three competing risks

First, we consider a simulation scenario with  $D = 3$  competing sources of risk. The simulated data are generated according to the latent failure times approach (see Section 1). In particular,

$$\begin{aligned} Y_{i,1} &\stackrel{\text{iid}}{\sim} \text{Weibull}(\xi_1 = 1.2, \lambda_1 = 1), & Y_{i,2} &\stackrel{\text{iid}}{\sim} \text{Weibull}(\xi_2 = 1.6, \lambda_2 = 1), \\ Y_{i,3} &\stackrel{\text{iid}}{\sim} \text{Weibull}(\xi_3 = 2.4, \lambda_3 = 1), \end{aligned} \quad (24)$$

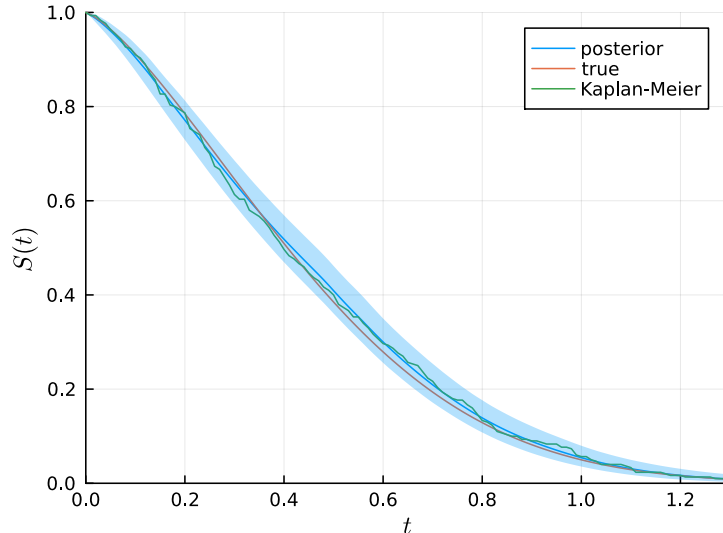


Figure 2: Survival function for a synthetic dataset with  $n = 300$  data points: posterior estimates obtained via the marginal algorithm (blue), true curve (orange), and Kaplan-Meier estimate (green); the pointwise 0.95 posterior credible band is also displayed.

with  $i = 1, \dots, n$ , where  $\xi_\delta$  and  $\lambda_\delta$  are shape and scale parameters, for  $\delta \in \{1, 2, 3\}$ . The observed survival time is  $T_i = \min\{Y_{i,1}, Y_{i,2}, Y_{i,3}\}$  and  $\Delta_i = \delta$  if and only if  $T_i = Y_{i,\delta}$ . The sample size is  $n = 300$ . Note that the number of subjects incurring in each competing event is a r.v. itself, and the parameters in (24) are chosen to yield a fairly balanced allocation of the observations to the different event types.

We rely on the *generalized gamma* hCRM prior with Dykstra-Laud kernel, and implement both algorithms outlined in Section 6. The hyperparameters of our model are  $\beta = \beta_0 = 1.0$  and  $\sigma = \sigma_0 = 0.25$ , while  $\Lambda_0(dx) = \theta dx$  for  $\theta > 0$  is proportional to the Lebesgue measure. Moreover, exponential hyperpriors with large means, reflecting a non-informative specification, are assigned to both  $\theta$  and the kernel parameter  $\gamma$ . The MCMC procedure is run for 25,000 iterations, with a burn-in of 5,000; after thinning, we collect 2,000 samples from the latent partition structure to compute posterior estimates with the marginal algorithm. For the quantification of posterior uncertainty, we generate 10 independent realizations from the posterior of the hCRMs, for each sample from the latent structure. Diagnostic tools suggest convergence of the Markov chains, which show a fairly good mixing; indeed, for the marginal algorithm, the average effective sample size for posterior survival function estimates, evaluated on a grid of points, is about 1,393.

Figure 2 compares the true survival function implied by (24), its Kaplan-Meier estimate and our Bayesian posterior estimate, obtained via the marginal algorithm. The latter successfully recovers the ground truth: its Kolmogorov distance from the true curve (0.0242)

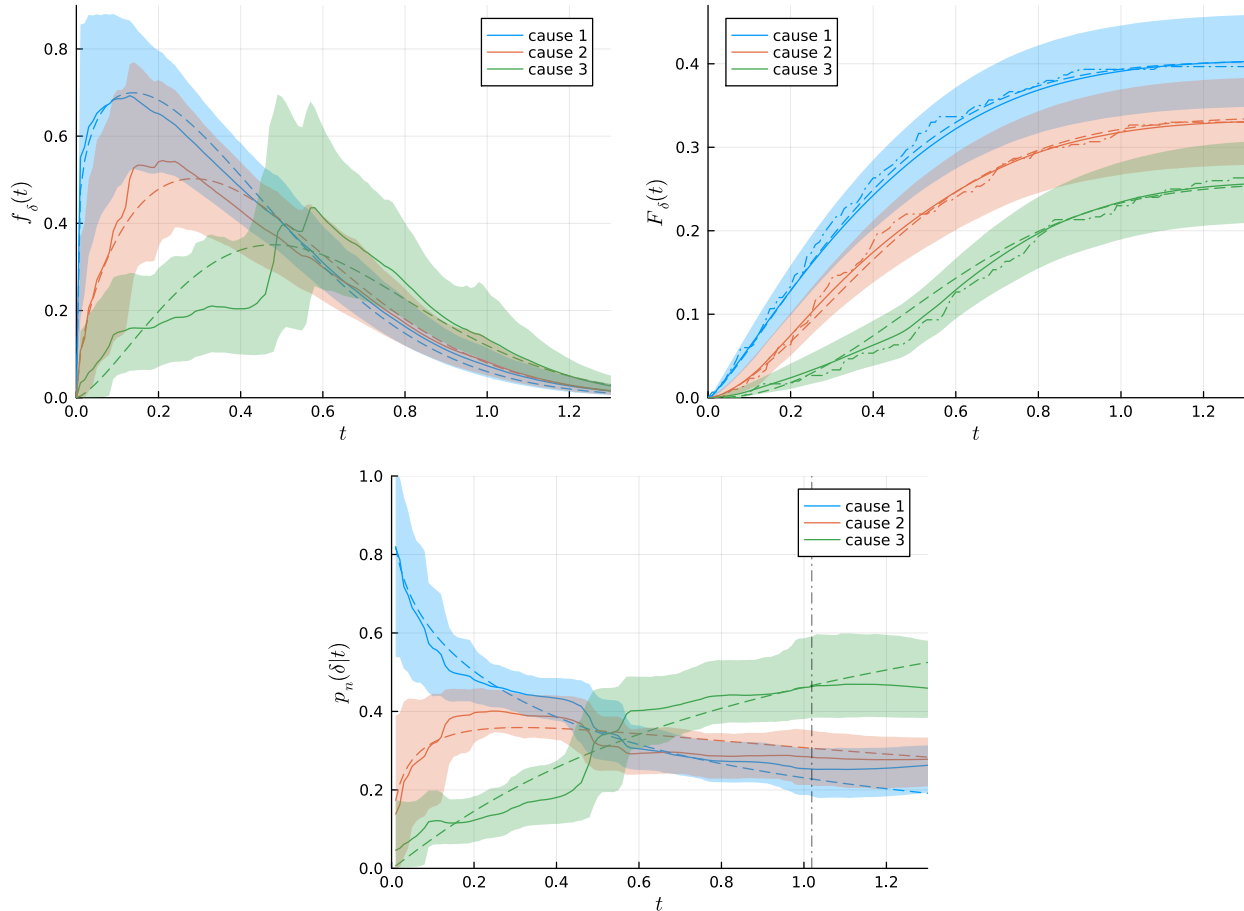


Figure 3: Cause-specific incidence functions (left), subdistributions (middle) and prediction curves (right); true curves (dashed) and posterior estimates (solid), with corresponding pointwise 0.95 credible bands; estimates of subdistributions are also compared with the Aalen-Johansen estimate (dash-dotted).

is substantially smaller than that of the Kaplan-Meier estimate (0.0330). Furthermore, the pointwise 0.95 credible band entirely covers the true function. Figure 3 compares the ground truth with the posterior estimates of the cause-specific incidence functions, subdistributions and prediction curves. The proposed model accurately recovers the true curves. Indeed, the crossing time (around  $t = 0.6$ ) of the cause-specific incidence functions is correctly detected, with pointwise credible bands slightly overlapping for times smaller than the crossing. Moreover, the posterior estimates of the subdistribution functions align with the frequentist Aalen-Johansen nonparametric estimates, and the corresponding pointwise credible bands entirely contain the true subdistribution functions. Finally, the prediction curves slightly depart from the true proportions for large values of  $t$ , as the observations do not inform prediction curves beyond the largest observed survival time; refer to the discussion in Section 5.2.

		incidence function			subdistribution		
		$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 1$	$\delta = 2$	$\delta = 3$
hierarchical	marginal	<b>.0952</b>	<b>.0833</b>	.0963	<b>.1274</b>	<b>.1179</b>	.1280
	conditional	.0964	.0836	<b>.0956</b>	.1293	.1185	<b>.1277</b>
independent	marginal	.1081	.0932	.0974	.1385	.1291	.1339
	conditional	.1085	.0934	.0971	.1392	.1298	.1338
frequentist					.1810	.1976	.2060

Table 1: Comparison of rescaled total variation distances between estimated and true incidence functions, and rescaled Kolmogorov distances between estimated and true subdistribution functions, for the hierarchical and the independent model, using marginal and conditional estimation approaches. Distances are averaged over 100 simulated datasets.

## 7.2 Dependent vs. independent priors

The proposed model assumes dependence among hazard rates associated to different sources of risk. This is one of its main strengths, as it allows for sharing of information between hazards. We illustrate this phenomenon by means of a simulation study, and contrast the performance of our model with a Bayesian nonparametric alternative having independent cause-specific hazard rates. The latter can be regarded as a special case of hierarchical prior distribution over the vector of random measures in (6), namely  $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_D) \mid \tilde{\mu}_0 \stackrel{\text{iid}}{\sim} \text{CRM}(\nu)$ , where  $\nu(ds, dx) = \rho(ds) \Lambda_0(dx)$  is a fixed homogeneous Lévy intensity. Therefore,  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  are also unconditionally i.i.d., since  $\nu$  does not depend on  $\tilde{\mu}_0$ . Such prior specification implies that the posterior estimate of the survival function for  $T$  coincides with the product of the estimates of  $\exp\left(-\int_0^t \int_{\mathbb{X}} k(s, x) \tilde{\mu}_\delta(dx) ds\right)$ , for each  $\delta \in \{1, \dots, D\}$ . These can be interpreted as the survival functions for each event type, in the hypothesis of absence of competing risks, i.e., considering the occurrences of competing events as censored observations.

Consider a data-generating model involving  $D = 3$  competing and independent sources of risk. For each event type  $\delta \in \{1, 2, 3\}$ , the latent survival time  $Y_{i,\delta}$  is sampled from a mixture of a cause-specific and a common distribution, with equal weights; see Section S5 for details. The numerical results are obtained by averaging over 100 simulated datasets, each consisting of  $n = 100$  observations, generated according to the potential survival times approach. The hyperparameters specification coincides with the one adopted in Section 7.1.

Table 1 provides a comparison of the performance of the proposed model with the *independent* model. Bayesian estimates are computed either via the marginal algorithm in Section 6.1 or averaging over the posterior samples generated via the conditional sampling

approach in Section 6.2. In particular, the average errors in estimating the cause-specific incidence function and subdistribution for each event type are reported for both models and both estimation strategies. In addition, the average estimation error for the Aalen-Johansen estimator of the subdistribution functions is also displayed. The comparison is based on total variation and Kolmogorov distances, which are rescaled according to the true probabilities of occurrence of each event type. Specifically, we consider the measures of error

$$e_{\delta}^{\text{TV}} = \frac{1}{2\pi_{\delta}} \int_{\mathbb{R}^+} |\hat{f}_{\delta}(t) - f_{\delta}(t)| dt, \quad e_{\delta}^{\text{K}} = \frac{1}{\pi_{\delta}} \sup_{t \in \mathbb{R}^+} |\hat{F}_{\delta}(t) - F_{\delta}(t)|, \quad \delta = 1, \dots, D,$$

where  $\hat{f}_{\delta}$  and  $f_{\delta}$  are the estimated and true incidence functions,  $\hat{F}_{\delta}$  and  $F_{\delta}$  are the estimated and true subdistribution functions, and  $\pi_{\delta}$  is the true probability of occurrence for event type  $\delta$ , for any  $\delta$ . This rescaling of standard distances is motivated by the need to obtain error measurements that are comparable across different event types. Indeed, the rescaled versions of the incidence and subdistribution functions correspond to proper density and distribution functions, respectively. Our hierarchical model systematically outperforms the independent model in the estimation of both incidence functions and subdistributions. Furthermore, the average number of common latent locations in the hierarchical model is 17.79, whereas the alternative model employs on average 35.46 latent locations among the independent random measures. This significant difference (roughly double) suggests that a substantially larger number of components is required to compensate for the lack of information borrowing in the independent model. Nonetheless, a deeper probabilistic understanding of the role of information borrowing in the competing-risks setting will require further investigation.

### 7.3 Applications to clinical datasets

We apply the proposed Bayesian nonparametric model to a clinical dataset from the bone marrow transplant registry of the European Blood and Marrow Transplant (EBMT) Group, publicly available in the `crrSC` package in R. Data from this registry have previously been used to illustrate methods for competing risks (Zhou et al., 2012), and this specific dataset was recently used in Schmitt et al. (2023) to compare different approaches to clustered data. Refer to Section S6 for an additional application on a dataset of patients affected by stage I melanoma.

Real-world datasets require a slight extension of our model to accommodate both right-censored observations and categorical predictors. For right-censored data, it is enough to modify the likelihood in (13) by removing censored observations from the factor that multiplies the exponential term. The distributional results in Sections 4 and 5 and algorithms

in Section 6 are easily adapted to the presence of right-censored observations once the multiplicative intensity likelihood in (13) is replaced by

$$\mathcal{L}(\tilde{\boldsymbol{\mu}}; \mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n}) = \exp\left(-\sum_{\ell=1}^D \sum_{i=1}^n \int_0^{T_i} \int_{\mathbb{X}} k(s; x) \tilde{\mu}_{\ell}(dx) ds\right) \prod_{i \notin \mathcal{C}} \int_{\mathbb{X}} k(T_i; x) \tilde{\mu}_{\Delta_i}(dx),$$

where  $\mathcal{C} \subseteq \{1, \dots, n\}$  denotes the set of censored observations. This implies that the additional latent r.v.  $\mathbf{X}_{1:n}$  and  $\mathbf{Z}_{1:n}$  are introduced exclusively for non-censored observations, which are those accommodated within the latent partition. In contrast, censored observations contribute to the augmented likelihood only through  $K_n(x) = K_n(x; \mathbf{T}_{1:n}) = \sum_{i=1}^n \int_0^{T_i} k(s; x) ds$ .

The inclusion of predictors is based on a Cox regression, or *proportional hazards*, model, which defines a semiparametric prior on the hazard rates. Specifically, cause-specific hazard rates (4) are extended to

$$\tilde{h}_{\delta}(t; \zeta) = \exp\left(\eta^T \zeta\right) \int_{\mathbb{X}} k(t; x) \tilde{\mu}_{\delta}(dx), \quad \delta = 1, \dots, D,$$

where  $\eta$  are the regression coefficients and  $\zeta$  denotes the categorical predictors. The extension of distributional results and sampling algorithms requires minimal efforts, namely the re-definition of the quantities in (15) as  $Q(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}, \boldsymbol{\zeta}_{1:n}) = \prod_{j=1}^k \prod_{i \in C_j^X} \exp\left(\eta^T \zeta_i\right) k(T_i; X_j^*)$  and

$$K_n(x) = K_n(x; \mathbf{T}_{1:n}, \boldsymbol{\zeta}_{1:n}) = \sum_{i=1}^n \exp\left(\eta^T \zeta_i\right) \int_0^{T_i} k(s; x) ds,$$

where  $\boldsymbol{\zeta}_{1:n} = (\zeta_1, \dots, \zeta_n)$  is the collection of observed predictors. Our application involves a single binary predictor  $\zeta_i \in \{0, 1\}$ , leading to a single regression coefficient  $\eta \in \mathbb{R}$ , for which we specify a non-informative centered Gaussian prior with large variance. Moreover, since the increasing hazard rates assumption from the Dykstra-Laud kernel appears too restrictive, we resort to the Ornstein-Uhlenbeck kernel (see Section 2) with a non-informative hyperprior on the rate parameter.

The EBMT dataset includes data for 400 patients diagnosed with acute myeloid leukemia, who underwent a bone marrow transplantation in 153 different hospitals. The primary event of interest is the occurrence of either acute or chronic *Graft-versus-Host-Disease* (GvHD), with death or relapse without GvHD as competing events; survival times are expressed in years from the graft. Of the patients, 194 (48.5%) observed GvHD, 74 (18.5%) experienced a competing event and the rest (33.0%) were censored. The median follow-up time is 0.50 years, and the maximum observed follow-up time for the primary event of interest is 7.30 years. The main categorical predictor is the stem cells source for transplantation: bone marrow

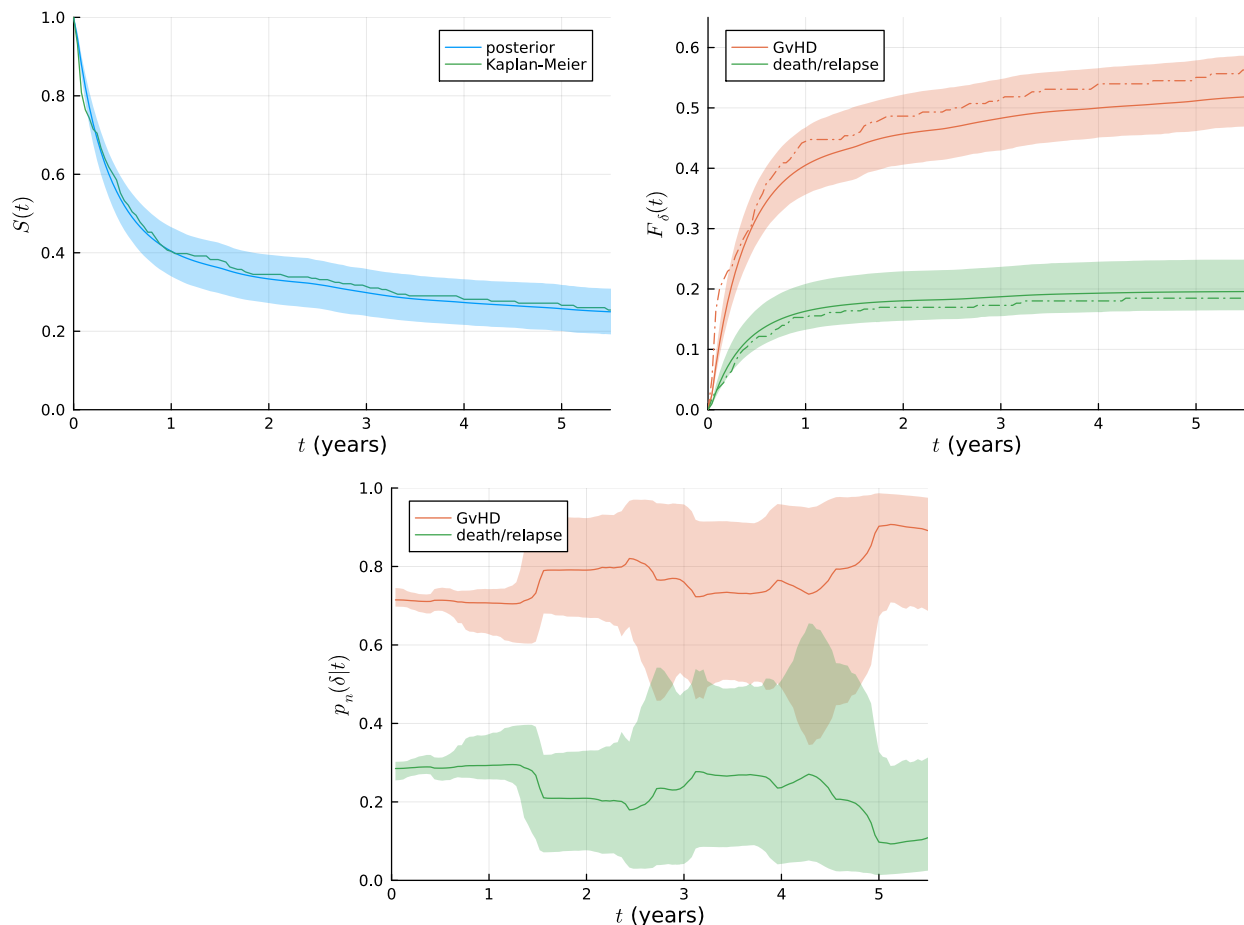


Figure 4: EBMT dataset: posterior estimates of the survival function (left), the subdistributions for the primary (GvHD) and competing (death/relapse) event types (middle), and prediction curves (right); the first two estimates are compared with the frequentist counterparts. Curves are related to patients who experienced graft from bone marrow cells.

grafts were performed on 222 patients (55.5%), while 178 (44.5%) patients received peripheral blood stem cells. The posterior estimates of the survival function, subdistributions for the primary and competing sources of risk, and prediction curves are displayed in Figure 4, along with pointwise 0.95 credible bands, for patients who experienced graft from bone marrow cells; the corresponding Kaplan-Meier and Aalen-Johansen estimators are also shown for comparison. The posterior estimate of the regression parameter  $\eta$  is  $-0.079$ , with 0.95 credible interval  $[-0.325, 0.169]$ , indicating no significant difference between the stem cell sources. These findings align with Zhou et al. (2012), who reported an estimate of  $-0.051$  ( $p$ -value: 0.35) using frequentist techniques based on the Fine-Gray proportional hazards model (Fine and Gray, 1999) on a larger dataset. Finally, the posterior distribution of the kernel rate parameter  $\kappa$  provides strong evidence against the increasing hazard rate assumption.

## Acknowledgements

The authors are grateful to the Associate Editor and two anonymous referees for their constructive and insightful suggestions, which led to major improvements to the paper. The research was largely conducted while Claudio Del Sole was a PhD student at Bocconi University (Milan, Italy).

## Funding

The authors were partially supported by the European Union - Next Generation EU PRIN-PNRR (project 2022CLTYP4).

## References

- Andersen, P. K., S. Z. Abildstrom, and S. Rosthøj (2002). Competing Risks as a Multi-State Model. *Statistical Methods in Medical Research* 11, 203–215.
- Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer.
- Arfé, A., S. Peluso, and P. Muliere (2019). Reinforced Urns and the Subdistribution Beta–Stacy Process for Competing Risks Analysis. *Scandinavian Journal of Statistics* 46, 706–734.
- Barrios, E., A. Lijoi, L. E. Nieto-Barajas, and I. Prünster (2013). Modeling with Normalized Random Measure Mixture Models. *Statistical Science* 28, 313–334.
- Beyersmann, J., A. Latouche, A. Buchholz, and M. Schumacher (2009). Simulating Competing Risks Data in Survival Analysis. *Statistics in Medicine* 28, 956–971.
- Blackwell, D. and J. B. MacQueen (1973). Ferguson Distributions via Pólya Urn Schemes. *Annals of Statistics* 1, 353–355.
- Brück, F. (2025). Generative Neural Networks for Characteristic Functions. *Journal of Computational and Graphical Statistics* 34(4), 1456–1465.
- Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution Theory for Hierarchical Processes. *Annals of Statistics* 47, 67–92.
- Camerlenghi, F., A. Lijoi, and I. Prünster (2021). Survival Analysis via Hierarchically Dependent Mixture Hazards. *Annals of Statistics* 49, 863–884.
- Cassese, A., W. Zhu, M. Guindani, and M. Vannucci (2019). A Bayesian Nonparametric Spiked Process Prior for Dynamic Model Selection. *Bayesian Analysis* 14, 553–572.

- Catalano, M. and C. Del Sole (2025). Hierarchical Random Measures without Tables. arXiv:2505.02653.
- Catalano, M., A. Lijoi, and I. Prünster (2020). Approximation of Bayesian Models for Time-to-Event Data. *Electronic Journal of Statistics* 14, 3366–3395.
- Chen, C., V. Rao, W. Buntine, and Y. Whye Teh (2013). Dependent Normalized Random Measures. In *Proceedings of Machine Learning Research*, Volume 28, pp. 969–977.
- Chipman, H. A., E. I. George, and R. E. McCulloch (2010). BART: Bayesian Additive Regression Trees. *Annals of Applied Statistics* 4, 266–298.
- Christensen, J. and L. Ma (2020). A Bayesian Hierarchical Model for Related Densities by Using Pólya Trees. *Journal of the Royal Statistical Society: Series B* 82, 127–153.
- Cook, R. J. and J. F. Lawless (2018). *Multistate Models for the Analysis of Life History Data*. Boca Raton: Chapman and Hall.
- Cox, D. R. (1959). The Analysis of Exponentially Distributed Life-Times with Two Types of Failures. *Journal of the Royal Statistical Society: Series B* 21, 411–421.
- Crowder, M. J. (1991). On the Identifiability Crisis in Competing Risks Analysis. *Scandinavian Journal of Statistics* 18, 223–233.
- Crowder, M. J. (2012). *Multivariate Survival Analysis and Competing Risks*. Chapman & Hall.
- Daley, D. J. and D. Vere-Jones (2007). *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. New York: Springer.
- De Blasi, P., G. Peccati, and I. Prünster (2009). Asymptotics for Posterior Hazards. *Annals of Statistics* 37, 1906–1945.
- Drzewiecki, K. T., C. Ladefoged, and H. E. Christensen (1980). Biopsy and Prognosis for Cutaneous Malignant Melanomas in Clinical Stage I. *Scandinavian Journal of Plastic and Reconstructive Surgery* 14, 141–144.
- Dykstra, R. L. and P. Laud (1981). A Bayesian Nonparametric Approach to Reliability. *Annals of Statistics* 9, 356–367.
- Epifani, I. and A. Lijoi (2010). Nonparametric Priors for Vectors of Survival Functions. *Statistica Sinica* 20, 1455–1484.
- Escobar, M. D. and M. West (1998). Computing Nonparametric Hierarchical Models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 1–22. New York: Springer.
- Ferguson, T. S. and M. J. Klass (1972). A Representation of Independent Increment Processes without Gaussian Components. *Annals of Mathematical Statistics* 43, 1634–1643.

- Fine, J. P. and R. J. Gray (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 94, 496–509.
- Geskus, R. B. (2015). *Data Analysis with Competing Risks and Intermediate States*. Chapman & Hall.
- Geskus, R. B. (2024). Competing Risks: Concepts, Methods, and Software. *Annual Review of Statistics and Its Application* 11, 227–254.
- Ghosal, S. and A. van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge: Cambridge University Press.
- Griffin, J. E. and F. Leisen (2017). Compound Random Measures and Their Use in Bayesian Non-Parametrics. *Journal of the Royal Statistical Society: Series B* 79, 525–545.
- Hjort, N. L. (1990). Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data. *Annals of Statistics* 18, 1259–1294.
- Ishwaran, H. and L. F. James (2004). Computational Methods for Multiplicative Intensity Models Using Weighted Gamma Processes: Proportional Hazards, Marked Point Processes, and Panel Count Data. *Journal of the American Statistical Association* 99, 175–190.
- James, L. F. (2005). Bayesian Poisson Process Partition Calculus with an Application to Bayesian Lévy Moving Averages. *Annals of Statistics* 33, 1771–1799.
- James, L. F., A. Lijoi, and I. Prünster (2009). Posterior Analysis for Normalized Random Measures with Independent Increments. *Scandinavian Journal of Statistics* 36, 76–97.
- Jara, A. and T. E. Hanson (2011). A Class of Mixtures of Dependent Tail-Free Processes. *Biometrika* 98(3), 553–566.
- Jara, A., E. Lesaffre, M. De Iorio, and F. Quintana (2010). Bayesian Semiparametric Inference for Multivariate Doubly-Interval-Censored Data. *Annals of Applied Statistics* 4, 2126–2149.
- Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data*. Wiley.
- Kingman, J. F. (1967). Completely Random Measures. *Pacific Journal of Mathematics* 21, 59–78.
- Kingman, J. F. (1993). *Poisson Processes*. Oxford: Clarendon Press.
- Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley.
- Lijoi, A. and B. Nipoti (2014). A Class of Hazard Rate Mixtures for Combining Survival Data from Different Experiments. *Journal of the American Statistical Association* 109, 802–814.
- Lijoi, A., B. Nipoti, and I. Prünster (2014). Bayesian Inference with Dependent Normalized Completely Random Measures. *Bernoulli* 20, 1260–1291.

- Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, pp. 80–136. Cambridge University Press.
- Lijoi, A., I. Prünster, and T. Rigon (2020). Sampling Hierarchies of Discrete Random Structures. *Statistics and Computing* 30, 1591–1607.
- Lo, A. and C.-S. Weng (1989). On a Class of Bayesian Nonparametric Estimates: II. Hazard Rate Estimates. *Annals of the Institute of Statistical Mathematics* 41, 227–245.
- MacEachern, S. N. (1998). Computational Methods for Mixture of Dirichlet Process Models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 23–43. New York: Springer.
- Müller, P. and F. Quintana (2004). Nonparametric Bayesian Data Analysis. *Statistical Science* 19, 95–110.
- Müller, P., F. Quintana, A. Jara, and T. E. Hanson (2015). *Bayesian Nonparametric Data Analysis*. Cham: Springer.
- Müller, P., F. Quintana, and G. L. Rosner (2004). A Method for Combining Inference across Related Nonparametric Bayesian Models. *Journal of the Royal Statistical Society: Series B* 66, 735–749.
- Nieto-Barajas, L. E. and S. G. Walker (2004). Bayesian Nonparametric Survival Analysis via Lévy Driven Markov Processes. *Statistica Sinica* 14(4), 1127–1146.
- Nipoti, B., A. Jara, and M. Guindani (2018). A Bayesian Semiparametric Partially PH Model for Clustered Time-to-Event Data. *Scandinavian Journal of Statistics* 45, 1016–1035.
- Oganisian, A., K. D. Getz, T. A. Alonzo, R. Aplenc, and J. A. Roy (2024). Bayesian Semiparametric Model for Sequential Treatment Decisions with Informative Timing. *Biostatistics* 25, 947–961.
- Passeggeri, R. (2023). On Quasi-Infinately Divisible Random Measures. *Bayesian Analysis* 18, 253–286.
- Passeggeri, R. (2025). Random Signed Measures. *Bayesian Analysis* (forthcoming).
- Peccati, G. and I. Prünster (2008). Linear and Quadratic Functionals of Random Hazard Rates: An Asymptotic Analysis. *Annals of Applied Probability* 18, 1910–1943.
- Pitman, J. (1996). Some Developments of the Blackwell–MacQueen Urn Scheme. In *Statistics, Probability and Game Theory*, pp. 245–267. IMS.
- Putter, H., M. Fiocco, and R. B. Geskus (2007). Tutorial in Biostatistics: Competing Risks and Multi-State Models. *Statistics in Medicine* 26, 2389–2430.
- Quintana, F. A., P. Müller, A. Jara, and S. N. MacEachern (2022). The Dependent Dirichlet Process and Related Models. *Statistical Science* 37, 24–41.

- Regazzini, E., A. Lijoi, and I. Prünster (2003). Distributional Results for Means of Normalized Random Measures with Independent Increments. *Annals of Statistics* 31, 560–585.
- Riva-Palacio, A. and F. Leisen (2018). Bayesian Nonparametric Estimation of Survival Functions with Multiple-Samples Information. *Electronic Journal of Statistics* 12, 1330–1357.
- Riva-Palacio, A. and F. Leisen (2021). Compound Vectors of Subordinators and Their Associated Positive Lévy Copulas. *Journal of Multivariate Analysis* 183, 104728.
- Riva-Palacio, A., F. Leisen, and J. E. Griffin (2022). Survival Regression Models with Dependent Bayesian Nonparametric Priors. *Journal of the American Statistical Association* 117, 1530–1539.
- Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge: Cambridge University Press.
- Schmitt, S., A. Buchholz, and A.-K. Ozga (2023). Systematic Comparison of Approaches to Analyze Clustered Competing Risks Data. *BMC Medical Research Methodology* 23, 86.
- Scoggins, C. R., M. I. Ross, D. S. Reintgen, D. Noyes, J. S. Goydos, P. D. Beitsch, M. M. Urist, S. Ariyan, J. J. Sussman, M. J. Edwards, A. B. Chagpar, R. C. G. Martin, A. J. Stromberg, L. Hagendoorn, and K. M. McMasters (2006). Gender-Related Differences in Outcome for Melanoma Patients. *Annals of Surgery* 243, 693–700.
- Soriano, J. and L. Ma (2019). Mixture Modeling on Related Samples by  $\psi$ -Stick Breaking and Kernel Perturbation. *Bayesian Analysis* 14, 161–180.
- Sparapani, R., B. R. Logan, R. E. McCulloch, and P. W. Laud (2020). Nonparametric Competing Risks Analysis Using Bayesian Additive Regression Trees. *Statistical Methods in Medical Research* 29, 57–77.
- Sparapani, R. A., B. R. Logan, R. E. McCulloch, and P. W. Laud (2016). Nonparametric Survival Analysis Using Bayesian Additive Regression Trees (BART). *Statistics in Medicine* 35, 2741–2753.
- Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* 101, 1566–1581.
- Tsiatis, A. (1975). A Nonidentifiability Aspect of the Problem of Competing Risks. *Proceedings of the National Academy of Sciences* 72, 20–22.
- Walker, S. and P. Muliere (1997). Beta-Stacy Processes and a Generalization of the Pólya-Urn Scheme. *Annals of Statistics* 25, 1762–1780.
- Wolpert, R. L. and K. Ickstadt (1998). Simulation of Lévy Random Fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pp. 227–242. Springer New York.

- Xu, Y., P. Müller, A. S. Wahed, and P. F. Thall (2016). Bayesian Nonparametric Estimation for Dynamic Treatment Regimes with Sequential Transition Times. *Journal of the American Statistical Association* 111, 921–950.
- Xu, Y., D. Scharfstein, P. Müller, and M. Daniels (2020). A Bayesian Nonparametric Approach for Evaluating the Causal Effect of Treatment in Randomized Trials with Semi-Competing Risks. *Biostatistics* 23, 34–49.
- Zhou, B., J. Fine, A. Latouche, and M. Labopin (2012). Competing Risks Regression for Clustered Data. *Biostatistics* 13, 371–383.

*Supplementary Material to:*

Principled Estimation and Prediction with  
Competing Risks: a Bayesian Nonparametric Approach

**Abstract**

This supplement contains a discussion of an alternative modeling approach based on latent failure times in Section [S1](#), background material and a key technical result in Section [S2](#), and the proofs of the main statements in Section [S3](#). Section [S4](#) is devoted to additional technical details on the model and algorithms. Complements to the applications presented in the main manuscript, as well as further illustrations with synthetic and real data, including a comparison with an alternative tree-based method, are provided in Sections [S5–S7](#). Numbered elements (i.e., theorems, equations, figures, and so on) specific to this Supplementary Material, and not included in the main manuscript, are labeled with an ‘S’ prefix. References to numbered items without the ‘S’ prefix refer to the main manuscript.

## S1 Latent failure times approach

Section 1.1 of the manuscript mentions an alternative approach to competing risks, which is briefly outlined here. This approach relies on the introduction of latent time-to-event variables  $Y_{i,1}, \dots, Y_{i,D}$ , for each observation  $i = 1, \dots, n$ . The observed survival times are then defined as  $T_i = \min\{Y_{i,1}, \dots, Y_{i,D}\}$ . In a Bayesian nonparametric framework, this setting can be accommodated by assuming partial exchangeability of the array  $\{(Y_{i,\delta})_{i \geq 1} : \delta = 1, \dots, D\}$ . This is a very intuitive distributional invariance property that extends the notion of exchangeability. Specifically, let  $\pi_1, \dots, \pi_D$  denote finite permutations on the integers, i.e. for any  $\delta$  there exists  $I_\delta \in \mathbb{N}$  such that  $\pi_\delta(i) = i$  for any  $i \geq I_\delta$ . Hence, partial exchangeability corresponds to the following distributional equivalence:

$$\{(Y_{i,\delta})_{i \geq 1} : \delta = 1, \dots, D\} \stackrel{d}{=} \{(Y_{\pi_\delta(i),\delta})_{i \geq 1} : \delta = 1, \dots, D\}.$$

This implies that the latent failure times are: (i) exchangeable when they share the same cause  $\delta$ ; (ii) conditionally independent across different causes. In view of a representation theorem for partially exchangeable arrays, there exists a vector of random probability measures  $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \dots, \tilde{p}_D)$  on  $\mathbb{R}^+$  such that

$$(Y_{i_1, \delta_1}, \dots, Y_{i_k, \delta_k}) \mid \tilde{\boldsymbol{p}} \sim \tilde{p}_{\delta_1} \times \dots \times \tilde{p}_{\delta_k},$$

for any  $k \geq 1$ ,  $\delta_1, \dots, \delta_k \in \{1, \dots, D\}$  and  $i_1, \dots, i_k \in \mathbb{N}$ . Remarkably, the dependence among components  $\tilde{p}_1, \dots, \tilde{p}_D$  implies unconditional dependence of the  $Y_{i,\delta}$ 's, both within each cause-specific group and across different causes. This differs from standard approaches that assume independent latent time-to-event for the different sources of risk.

## S2 Completely random measures

The model proposed in (6) for analyzing competing risks data relies on completely random measures, originally introduced in [Kingman \(1967\)](#), as main building blocks. Therefore, before presenting the proofs of the main results, we outline key properties of this class of random measures.

**Definition 12.** A random element  $\tilde{\mu}$ , defined on  $\mathbb{X}$  and taking values in the space of boundedly finite measures, is a *completely random measure* (CRM) if, for any collection of pairwise disjoint sets  $A_1, \dots, A_n$ , the random variables  $\tilde{\mu}(A_1), \dots, \tilde{\mu}(A_n)$  are mutually independent.

We refer the reader to [Daley and Vere-Jones \(2007\)](#) and [Kingman \(1993\)](#) for exhaustive accounts. Throughout, we consider CRMs without deterministic component and fixed points of discontinuity. These are almost surely (a.s.) discrete and can be represented as

$$\tilde{\mu} = \sum_{h \geq 1} S_h \delta_{X_h^*}, \quad (\text{S1})$$

where  $(S_h)_{h \geq 1}$  is a sequence of positive random jumps and  $(X_h^*)_{h \geq 1}$  a sequence of random locations. Remarkably, the pairs  $(S_h, X_h^*)_{h \geq 1}$  are the points of a Poisson random measure  $\tilde{N}$  on  $\mathbb{R}^+ \times \mathbb{X}$  having mean intensity  $\nu$ , which implies that, for any  $A$  such that  $\nu(A) < \infty$ , the integer-valued r.v.  $\tilde{N}(A) = \text{card}(A \cap \{(S_h, X_h^*) : h \geq 1\})$  has Poisson distribution with mean  $\nu(A)$ . In fact, the discreteness of their realizations stems from their characterization as linear functionals of Poisson random measures

$$\tilde{\mu}(dx) = \int_{\mathbb{R}^+} s \tilde{N}(ds, dx),$$

where  $\tilde{N}$  is a Poisson random measure on the product space  $\mathbb{R}^+ \times \mathbb{X}$ . The measure  $\nu$  is also called *Lévy intensity* of  $\tilde{\mu}$  and it uniquely identifies  $\tilde{\mu}$ , which motivates the notation  $\tilde{\mu} \sim \text{CRM}(\nu)$ . This can be also seen from the Laplace functional transform of  $\tilde{\mu}$ , which coincides, for any non-negative bounded  $f: \mathbb{X} \mapsto \mathbb{R}^+$ , with

$$\mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} f(x) \tilde{\mu}(dx) \right) \right] = \exp \left( - \int_{\mathbb{R}^+ \times \mathbb{X}} (1 - e^{-sf(x)}) \nu(ds, dx) \right). \quad (\text{S2})$$

For any bounded set  $A \subset \mathbb{X}$ , the measure  $\nu$  must satisfy the integrability condition

$$\int_{\mathbb{R}^+ \times A} \min(s, 1) \nu(ds, dx) < \infty.$$

Moreover, we assume  $\nu(\mathbb{R}^+ \times A) = \infty$ , which corresponds to  $\tilde{\mu}$  having an infinite number of random jumps on bounded sets and thus guarantees that  $\tilde{\mu}$  is strictly positive a.s. ([Regazzini](#)

et al., 2003). Such an assumption is common in Bayesian nonparametric modeling and is known as *infinite activity* property. The Lévy intensity can be conveniently decomposed as

$$\nu(ds, dx) = \rho(ds | x) \alpha(dx),$$

where  $\rho$  is a transition kernel characterizing the random jumps  $S_h$  and  $\alpha$  is a  $\sigma$ -finite measure characterizing the random locations  $X_h^*$ . If the jumps are independent of the locations, we have a *homogeneous* CRM with  $\nu(ds, dx) = \rho(ds) \alpha(dx)$ . For instance, the popular *gamma* CRM arises when  $\rho(ds) = s^{-1} e^{-\beta s} ds$ , for  $\beta > 0$ ; from (S2), it follows that  $\tilde{\mu}(A)$  has gamma distribution with shape parameter  $\alpha(A)$  and scale parameter  $\beta$ . Some recent extensions of CRMs and of their uses in Bayesian nonparametrics can be found in Passeggeri (2023, 2025).

## S2.1 Preliminary technical lemma

We provide a technical result about completely random measures (CRMs), which constitutes the backbone for the following proofs. The main tool used for its derivation is the Laplace functional of CRMs, in the spirit of the general approach of James et al. (2009). Consider a homogeneous CRM  $\tilde{\mu}$  with Lévy intensity  $\nu(ds, dx) = \rho(ds) \alpha(dx)$ , where  $\alpha$  is a  $\sigma$ -finite diffuse measure on  $\mathbb{X}$ . The Laplace exponent and cumulants of  $\tilde{\mu}$  are, respectively, defined as

$$\psi(u) = \int_{\mathbb{R}^+} (1 - e^{-us}) \rho(ds), \quad \tau(m; u) = \int_{\mathbb{R}^+} s^m e^{-us} \rho(ds).$$

**Lemma 13.** *For any non-negative measurable function  $f: \mathbb{X} \mapsto \mathbb{R}^+$ , and distinct values  $x_1^*, \dots, x_k^* \in \mathbb{X}$ , with multiplicities  $n_1, \dots, n_k \geq 1$ , one has*

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} \frac{\mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} f(x) \tilde{\mu}(dx) \right) \prod_{j=1}^k \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \right]}{\prod_{j=1}^k \alpha(B_\varepsilon(x_j^*))} \\ = \exp \left( - \int_{\mathbb{X}} \psi(f(x)) \alpha(dx) \right) \prod_{j=1}^k \tau(n_j; f(x_j^*)), \end{aligned} \quad (\text{S3})$$

where  $B_\varepsilon(\bar{x}) = \{x \in \mathbb{X} : d(x, \bar{x}) < \varepsilon\}$  denotes the  $\varepsilon$ -ball centered at  $\bar{x} \in \mathbb{X}$ .

Note that the relationship (S3) can be rewritten less formally but more conveniently as

$$\begin{aligned} \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} f(x) \tilde{\mu}(dx) \right) \prod_{j=1}^k \tilde{\mu}(dx_j^*)^{n_j} \right] \\ = \exp \left( - \int_{\mathbb{X}} \psi(f(x)) \alpha(dx) \right) \prod_{j=1}^k \tau(n_j; f(x_j^*)) \alpha(dx_j^*). \end{aligned} \quad (\text{S4})$$

*Proof.* Consider  $\varepsilon > 0$  sufficiently small such that the  $\varepsilon$ -balls  $B_\varepsilon(x_1^*), \dots, B_\varepsilon(x_k^*)$  are pairwise disjoint, and define the complement  $\mathbb{X}^* = \mathbb{X} \setminus (\cup_{j=1}^k B_\varepsilon(x_j^*))$ . By the independence of the evaluations of a CRM on pairwise disjoint sets, one has

$$\begin{aligned} \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} f(x) \tilde{\mu}(dx) \right) \prod_{j=1}^k \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \right] \\ = \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}^*} f(x) \tilde{\mu}(dx) \right) \right] \prod_{j=1}^k \mathbb{E} \left[ \exp \left( - \int_{B_\varepsilon(x_j^*)} f(x) \tilde{\mu}(dx) \right) \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \right]. \end{aligned}$$

The first factor is the Laplace transform of  $\tilde{\mu}$  restricted to  $\mathbb{X}^*$  and evaluated at  $f$ . Therefore

$$\begin{aligned} \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}^*} f(x) \tilde{\mu}(dx) \right) \right] &= \exp \left( - \int_{\mathbb{X}^*} (1 - e^{-f(x)s}) \rho(ds) \alpha(dx) \right) \\ &= \exp \left( - \int_{\mathbb{X}^*} \psi(f(x)) \alpha(dx) \right). \end{aligned}$$

By the monotone convergence theorem, in the limit as  $\varepsilon \rightarrow 0$ , we obtain

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}^*} f(x) \tilde{\mu}(dx) \right) \right] &= \lim_{\varepsilon \rightarrow 0} \exp \left( - \int_{\mathbb{X}^*} \psi(f(x)) \alpha(dx) \right) \\ &= \exp \left( - \int_{\mathbb{X}} \psi(f(x)) \alpha(dx) \right). \end{aligned}$$

On the other hand, for  $j = 1, \dots, k$ , by the monotone convergence theorem, we also get

$$\begin{aligned} M(\varepsilon, x_j^*; n_j) &= \mathbb{E} \left[ \exp \left( - \int_{B_\varepsilon(x_j^*)} f(x) \tilde{\mu}(dx) \right) \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \right] \\ &= \lim_{u \rightarrow 0} \mathbb{E} \left[ \exp \left( - \int_{B_\varepsilon(x_j^*)} f(x) \tilde{\mu}(dx) - u \tilde{\mu}(B_\varepsilon(x_j^*)) \right) \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \right]. \end{aligned}$$

Given that, for  $n_j \geq 0$ ,

$$\frac{d^{n_j}}{du^{n_j}} \exp\left(-u \tilde{\mu}(B_\varepsilon(x_j^*))\right) = (-1)^{n_j} \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \exp\left(-u \tilde{\mu}(B_\varepsilon(x_j^*))\right),$$

the expression above becomes

$$\begin{aligned} M(\varepsilon, x_j^*; n_j) &= (-1)^{n_j} \lim_{u \rightarrow 0} \mathbb{E} \left[ \frac{d^{n_j}}{du^{n_j}} \exp\left(-\int_{B_\varepsilon(x_j^*)} f(x) \tilde{\mu}(dx) - u \tilde{\mu}(B_\varepsilon(x_j^*))\right) \right] \\ &= (-1)^{n_j} \lim_{u \rightarrow 0} \frac{d^{n_j}}{du^{n_j}} \mathbb{E} \left[ \exp\left(-\int_{B_\varepsilon(x_j^*)} (f(x) + u) \tilde{\mu}(dx)\right) \right]. \end{aligned}$$

Again, the expectation coincides with the Laplace transform of  $\tilde{\mu}$  restricted to  $B_\varepsilon(x_j^*)$  and evaluated at  $f + u 1$ , where  $x \mapsto 1(x) = 1$  is the identity function. Hence we have

$$M(\varepsilon, x_j^*; n_j) = (-1)^{n_j} \lim_{u \rightarrow 0} \frac{d^{n_j}}{du^{n_j}} \exp\left(-\int_{B_\varepsilon(x_j^*)} \psi(f(x) + u) \alpha(dx)\right).$$

Since  $\alpha$  is assumed to be finite and diffuse, the measure of the  $\varepsilon$ -ball  $B_\varepsilon(x_j^*)$  vanishes for  $\varepsilon \rightarrow 0$ , and the expression above can be rewritten as

$$\begin{aligned} M(\varepsilon, x_j^*; n_j) &= (-1)^{n_j+1} \lim_{u \rightarrow 0} \int_{B_\varepsilon(x_j^*)} \frac{d^{n_j}}{du^{n_j}} \psi(f(x) + u) \alpha(dx) \\ &\quad \times \exp\left(-\int_{B_\varepsilon(x_j^*)} \psi(f(x) + u) \alpha(dx)\right) + o\left(\alpha(B_\varepsilon(x_j^*))\right). \end{aligned}$$

By definition of Laplace exponent of  $\tilde{\mu}$ , its derivatives coincide with the cumulants of  $\tilde{\mu}$ ,

$$\begin{aligned} \frac{d^m}{du^m} \psi(u) &= \frac{d^m}{du^m} \int_{\mathbb{R}^+} (1 - e^{-us}) \rho(ds) = \int_{\mathbb{R}^+} \frac{d^m}{du^m} (1 - e^{-us}) \rho(ds) \\ &= (-1)^{m+1} \int_{\mathbb{R}^+} s^m e^{-us} \rho(ds) = (-1)^{m+1} \tau(m; u). \end{aligned}$$

Therefore, the quantity above becomes

$$\begin{aligned} M(\varepsilon, x_j^*; n_j) &= \lim_{u \rightarrow 0} \int_{B_\varepsilon(x_j^*)} \tau(n_j; f(x) + u) \alpha(dx) \\ &\quad \times \exp\left(-\int_{B_\varepsilon(x_j^*)} \psi(f(x) + u) \alpha(dx)\right) + o\left(\alpha(B_\varepsilon(x_j^*))\right) \\ &= \int_{B_\varepsilon(x_j^*)} \tau(n_j; f(x)) \alpha(dx) \exp\left(-\int_{B_\varepsilon(x_j^*)} \psi(f(x)) \alpha(dx)\right) + o\left(\alpha(B_\varepsilon(x_j^*))\right). \end{aligned}$$

In conclusion, taking the limit for  $\varepsilon \rightarrow 0$ , by the monotone convergence theorem, we get

$$\begin{aligned}
& \lim_{\varepsilon \rightarrow 0} \frac{1}{\alpha(B_\varepsilon(x_j^*))} \mathbb{E} \left[ \exp \left( - \int_{B_\varepsilon(x_j^*)} f(x) \tilde{\mu}(dx) \right) \tilde{\mu}(B_\varepsilon(x_j^*))^{n_j} \right] \\
&= \lim_{\varepsilon \rightarrow 0} \frac{1}{\alpha(B_\varepsilon(x_j^*))} \int_{B_\varepsilon(x_j^*)} \tau(n_j; f(x)) \alpha(dx) \exp \left( - \int_{B_\varepsilon(x_j^*)} \psi(f(x)) \alpha(dx) \right) \\
&= \lim_{\varepsilon \rightarrow 0} \frac{1}{\alpha(B_\varepsilon(x_j^*))} \int_{B_\varepsilon(x_j^*)} \tau(n_j; f(x)) \alpha(dx) = \tau(n_j; f(x_j^*)).
\end{aligned}$$

□

## S3 Proofs

**Proof of Proposition 3** As a first step, we establish a necessary and sufficient condition that guarantees a proper survival function, a result of independent interest. Let  $\psi$  and  $\psi_0$  denote the Laplace exponents of  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  and  $\tilde{\mu}_0$ , respectively:

$$\psi(u) = \int_{\mathbb{R}^+} (1 - e^{-us}) \rho(ds), \quad \psi_0(u) = \int_{\mathbb{R}^+} (1 - e^{-us}) \rho_0(ds).$$

**Proposition 14.** *The survival function  $\tilde{S}(t)$  in (5) is a proper survival function, that is  $\lim_{t \rightarrow \infty} \tilde{S}(t) = 0$  a.s., if and only if, for any  $\lambda > 0$ ,*

$$\lim_{t \rightarrow \infty} \int_{\mathbb{X}} \psi_0 \left( D\psi \left( \lambda \int_0^t k(s; x) ds \right) \right) \Lambda_0(dx) = \infty. \quad (\text{S5})$$

*Proof.* The survival function is defined as

$$\tilde{S}(t) = \exp \left( - \sum_{\ell=1}^D \int_0^t \int_{\mathbb{X}} k(s; x) \tilde{\mu}_\ell(dx) ds \right).$$

Therefore, the condition  $\lim_{t \rightarrow \infty} \tilde{S}(t) = 0$  a.s. is equivalent to

$$\lim_{t \rightarrow \infty} \sum_{\ell=1}^D \int_0^t \int_{\mathbb{X}} k(s; x) \tilde{\mu}_\ell(dx) ds = \lim_{t \rightarrow \infty} \tilde{H}(t) = \infty, \quad \text{a.s.},$$

where  $\tilde{H}(t)$  denotes the cumulative hazard for all sources of risk. The Laplace transform of  $\tilde{H}(t)$  can be easily computed using (S2) recursively. Indeed, conditionally on the random measure  $\tilde{\mu}_0$  at the root of the hierarchical prior, the random measures  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  are independent CRMs on  $\mathbb{X}$ , characterized by the homogeneous Lévy intensity  $\tilde{\nu}(ds, dx) = \rho(ds) \tilde{\mu}_0(dx)$ . For any  $\lambda \geq 0$ ,

$$\begin{aligned} \mathbb{E} \left[ \exp \left( -\lambda \tilde{H}(t) \right) \mid \tilde{\mu}_0 \right] &= \mathbb{E} \left[ \exp \left( -\lambda \sum_{\ell=1}^D \int_0^t \int_{\mathbb{X}} k(s; x) \tilde{\mu}_\ell(dx) ds \right) \mid \tilde{\mu}_0 \right] \\ &= \prod_{\ell=1}^D \mathbb{E} \left[ \exp \left( -\lambda \int_{\mathbb{X}} \int_0^t k(s; x) ds \tilde{\mu}_\ell(dx) \right) \mid \tilde{\mu}_0 \right] \\ &= \exp \left( -D \int_{\mathbb{X}} \psi \left( \lambda \int_0^t k(s; x) ds \right) \tilde{\mu}_0(dx) \right), \end{aligned}$$

where  $\psi$  is the Laplace exponent of  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ , which are identically distributed. The random measure  $\tilde{\mu}_0$  at the root of the hierarchical prior is itself a CRM with homogeneous Lévy

intensity  $\nu_0(ds, dx) = \rho_0(ds) \Lambda_0(dx)$ . Therefore, the Laplace transform of  $\tilde{H}(t)$  is

$$\begin{aligned} \mathbb{E} \left[ \exp \left( -\lambda \tilde{H}(t) \right) \right] &= \mathbb{E} \left[ \exp \left( -D \int_{\mathbb{X}} \psi \left( \lambda \int_0^t k(s; x) ds \right) \tilde{\mu}_0(dx) \right) \right] \\ &= \exp \left( - \int_{\mathbb{X}} \psi_0 \left( D\psi \left( \lambda \int_0^t k(s; x) ds \right) \right) \Lambda_0(dx) \right), \end{aligned}$$

where  $\psi_0$  is the Laplace exponent of  $\tilde{\mu}_0$ . The Laplace transform of an a.s. infinite random variable is zero for every  $\lambda > 0$ . Moreover, the distribution of the limit of a sequence of non-negative random variables is characterized by the pointwise limit of their Laplace transforms, by the Lévy continuity theorem. Consequently,  $\lim_{t \rightarrow \infty} \tilde{H}(t) = \infty$  if and only if, for any  $\lambda > 0$ ,

$$\lim_{t \rightarrow \infty} \mathbb{E} \left[ \exp \left( -\lambda \tilde{H}(t) \right) \right] = \exp \left( - \lim_{t \rightarrow \infty} \int_{\mathbb{X}} \psi_0 \left( D\psi \left( \lambda \int_0^t k(s; x) ds \right) \right) \Lambda_0(dx) \right) = 0.$$

□

*Proof of Proposition 3.* Remarkably, the Laplace exponents  $\psi$  and  $\psi_0$  are strictly positive and increasing functions on  $(0, \infty)$ . Moreover,

$$\lim_{u \rightarrow \infty} \psi(u) = \lim_{u \rightarrow \infty} \int_{\mathbb{R}^+} (1 - e^{-us}) \rho(ds) = \int_{\mathbb{R}^+} \rho(ds),$$

and similarly  $\lim_{u \rightarrow \infty} \psi_0(u) = \int_{\mathbb{R}^+} \rho_0(ds)$ . Therefore,  $\lim_{u \rightarrow \infty} \psi(u) = \lim_{u \rightarrow \infty} \psi_0(u) = \infty$  if and only if  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  and  $\tilde{\mu}_0$  are infinitely active CRMs; see Section S2 for the definition and implications of infinite activity. To conclude the proof, we show that either one of the conditions contained in the statement imply (S5):

- (i) Assume  $\lim_{t \rightarrow \infty} \int_0^t k(s; x) ds = \infty$  for every  $x \in \mathbb{X}$ . Then, for any  $\lambda > 0$ ,

$$\begin{aligned} \lim_{t \rightarrow \infty} \int_{\mathbb{X}} \psi_0 \left( D\psi \left( \lambda \int_0^t k(s; x) ds \right) \right) \Lambda_0(dx) \\ = \int_{\mathbb{X}} \lim_{u \rightarrow \infty} \psi_0 (D\psi (\lambda u)) \Lambda_0(dx) = \lim_{u \rightarrow \infty} \psi_0 (D\psi (u)) \cdot \Lambda_0(\mathbb{X}). \end{aligned}$$

Since  $\Lambda_0(\mathbb{X}) > 0$ , the quantity above is infinite if and only if  $\psi_0 (D\psi (u)) \rightarrow \infty$ , i.e. if and only if  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  and  $\tilde{\mu}_0$  are infinitely active,

(ii) Assume  $\lim_{t \rightarrow \infty} \int_0^t k(s; x) ds \geq C$  for every  $x \in \mathbb{X}$  and some  $C > 0$ . Then, for any  $\lambda > 0$ ,

$$\begin{aligned} \lim_{t \rightarrow \infty} \int_{\mathbb{X}} \psi_0 \left( D\psi \left( \lambda \int_0^t k(s; x) ds \right) \right) \Lambda_0(dx) \\ \geq \int_{\mathbb{X}} \psi_0 (D\psi(\lambda C)) \Lambda_0(dx) = \psi_0(D\psi(\lambda C)) \cdot \Lambda_0(\mathbb{X}), \end{aligned}$$

which is infinite if and only if  $\Lambda_0(\mathbb{X}) = \infty$ , since  $\psi_0(D\psi(\lambda C)) > 0$ .

□

**Proof of Theorem 4** The proof leverages results from Section 3 in [De Blasi et al. \(2009\)](#), where the authors focus on single-cause mixture hazards

$$\tilde{h}(t) = \int_{\mathbb{X}} k(t; x) \tilde{\mu}(dx), \quad (\text{S6})$$

with mixing measure  $\tilde{\mu}$  a CRM as defined in Section S2 and characterized by the Lévy intensity  $\nu(ds, dx) = \rho(ds | x) \alpha(dx)$ . They also assume that  $\rho$  and  $\alpha$  are diffuse measures and that the support of  $\alpha$  is the entire space  $\mathbb{X}$ . Finally, their Proposition 3 contains a crucial condition on the small-time behaviour of  $\tilde{\mu}$ , namely that there exists  $r > 0$  such that

$$\liminf_{t \rightarrow 0} \frac{\tilde{\mu}((0, t])}{t^r} = \infty \quad \text{a.s.} \quad (\text{S7})$$

The random mixture hazard (S6) induces a prior  $\mathcal{P}$  on the space of densities for the survival time, i.e. densities on  $\mathbb{R}^+$ . [De Blasi et al. \(2009\)](#) provide an explicit characterization of the classes of densities at which  $\mathcal{P}$  is weakly consistent, for several kernel choices. Their proof technique consists in translating the well-known sufficient condition for weak consistency, expressed in terms of Kullback–Leibler neighborhoods in the space of densities, to a condition on uniform neighborhoods in the space of hazards. Hereafter,  $\tilde{H}(t) = \int_0^t \tilde{h}(s) ds$  denotes the cumulative hazard, so that  $\tilde{f}(t) = \tilde{h}(t) \exp(-\tilde{H}(t))$  is the associated density function. Moreover,  $f_0$  is the “true” data-generating density on  $\mathbb{R}^+$ , and  $h_0$  its corresponding hazard function. For convenience, we restate the specific consistency results that are relevant to our purposes.

**Theorem 15** ([De Blasi et al. \(2009\)](#), Theorem 4). *Let  $\tilde{h}$  be a mixture hazard (S6) with Dykstra–Laud kernel  $k(t; x) = \mathbf{1}_{\{t \geq x\}}$  and assume  $\tilde{\mu}$  satisfies (S7). Then,  $\mathcal{P}$  is weakly consistent at each  $f_0$  such that  $\int_0^\infty \mathbb{E}[\tilde{H}(t)] f_0(t) dt < \infty$ , with  $h_0(0) = 0$  and  $h_0(t)$  strictly positive and non-decreasing for any  $t > 0$ .*

An analogous result holds for the slight variation of the Dykstra-Laud kernel  $k(t; x) = \gamma \mathbb{1}_{\{t \geq x\}}$  considered in our paper, since the rescaled measure  $\gamma \tilde{\mu}(dx)$  is again a CRM satisfying (S7).

**Theorem 16** (De Blasi et al. (2009), Theorem 5). *Let  $\tilde{h}$  be a mixture hazard (S6) with rectangular kernel  $k(t; x) = \mathbb{1}_{\{|t-x| \leq \tau\}}$ . Assume that the bandwidth  $\tau$  is random and its prior assigns positive probability to  $[0, L]$  for some  $L > 0$ . Then,  $\mathcal{P}$  is weakly consistent at each  $f_0$  such that  $\int_0^\infty \max(\mathbb{E}[\tilde{H}(t)], t) f_0(t) dt < \infty$ , with  $h_0$  strictly positive, bounded and Lipschitz.*

As with the previous case, an adaptation of the proof of Theorem 16 can be used to establish an analogous result for the rectangular kernel  $k(t; x) = \gamma \mathbb{1}_{\{0 \leq t-x \leq \tau\}}$ , which is the one we consider here. Indeed, the claim rests on the following arguments: (i) the condition on the random bandwidth remains valid if  $\tau$  is replaced by  $2\tau$ ; (ii) the rescaled and shifted measure  $\gamma \tilde{\mu}(dx - \tau)$  is a CRM on  $(\tau, \infty)$ ; (iii) the behavior on  $(0, \tau)$  coincides with that of the Dykstra-Laud kernel. Therefore, Theorem 16 extends to our setting provided  $h_0(0) = 0$  and  $\tilde{\mu}$  satisfies (S7).

**Theorem 17** (De Blasi et al. (2009), Theorem 6). *Let  $\tilde{h}$  be a mixture hazard (S6) with Ornstein-Uhlenbeck kernel  $k(t; x) = \sqrt{2\kappa} \exp(-\kappa(t-x)) \mathbb{1}_{\{t \geq x\}}$  and assume  $\tilde{\mu}$  satisfies (S7). Then,  $\mathcal{P}$  is weakly consistent at each  $f_0$  such that  $\int_0^\infty \max(\mathbb{E}[\tilde{H}(t)], t) f_0(t) dt < \infty$ , with  $h_0(0) = 0$  and  $h_0(t)$  satisfying the following properties for any  $t > 0$ :*

- (a)  $h_0(t)$  is strictly positive and differentiable;
- (b) for any  $t > 0$  for which  $h'_0(t) < 0$ , the local exponential decay rate  $-h'_0(t)/h_0(t)$  is smaller than  $\kappa\sqrt{2\kappa}$ .

Finally, as stated in Remark 1 of De Blasi et al. (2009), if  $\alpha$  is (proportional to) the Lebesgue measure, the integrability conditions on the cumulative hazard  $\tilde{H}(t)$  in the theorems above become, respectively,  $\int_0^\infty t^2 f_0(t) dt < \infty$  for the Dykstra-Laud kernel and  $\int_0^\infty t f_0(t) dt < \infty$  for the rectangular and Ornstein-Uhlenbeck kernels.

To take advantage of these results in our setting, we must show that the hazard rate for the survival time  $T$  from all sources of risk is a mixture hazard as in (S6). Indeed, by construction of our model (see (5)), the hazard rate from all sources of risk is

$$\sum_{\ell=1}^D \int_{\mathbb{X}} k(s; x) \tilde{\mu}_\ell(dx) ds = \int_{\mathbb{X}} k(s; x) \tilde{\mu}_{\text{sum}}(dx) ds,$$

where  $\tilde{\mu}_{\text{sum}} = \sum_{\ell=1}^D \tilde{\mu}_\ell$  is the mixing random measure. Conditionally on  $\tilde{\mu}_0$  at the root of the hierarchy, the measure  $\tilde{\mu}_{\text{sum}}$  is the sum of independent CRMs, and thus is itself a CRM.

Moreover, a hierarchical structure of CRMs is a CRM even unconditionally on the hierarchy's root, as stated in Theorem 1 of [Catalano and Del Sole \(2025\)](#). In particular, its Laplace exponent is  $\psi_{\text{sum}}(u) = \psi_0(D\psi(u))$ , where  $\psi$  and  $\psi_0$  are the Laplace exponents of each  $\tilde{\mu}_\ell$  and  $\tilde{\mu}_0$ , respectively. Note that  $\tilde{\mu}_{\text{sum}}$  is infinitely active if the random measures at both levels of the hierarchy are infinitely active ([Catalano and Del Sole, 2025](#), Lemma 2). Hence, the hazard rate from all sources of risk is indeed of the form (S6).

To conclude the proof, let  $\tilde{\boldsymbol{\mu}}$  be a generalized gamma hCRM and let  $\Lambda_0$  be proportional to the Lebesgue measure. This specification guarantees that the Lévy measure of  $\tilde{\mu}_{\text{sum}}$  is diffuse, and its base measure  $\Lambda_0$  has full support. Moreover, let  $\rho_{\text{sum}}$  be the jump component of its Lévy measure, and note that

$$\int_0^1 s^\varepsilon \rho_{\text{sum}}(ds) = \int_0^1 \int_0^1 s^\varepsilon \rho_{\text{sum}}(ds) \Lambda_0(dx) = D \int_0^1 s^\varepsilon \rho(ds) \int_0^\infty s \rho_0(ds) \int_0^1 \Lambda_0(dx),$$

which is infinite for every  $0 < \varepsilon < \sigma$ . Therefore, by Proposition 47.18 in [Sato \(1999\)](#), the small-time behaviour of  $\tilde{\mu}_{\text{sum}}$  is such that  $\liminf_{t \rightarrow 0} \tilde{\mu}_{\text{sum}}((0, t])/G(t) = C$  a.s. for some constant  $0 < C < \infty$ , where

$$G(t) = \frac{\log \log(t^{-1})}{\psi_{\text{sum}}^{-1}(t^{-1} \log \log(t^{-1}))}.$$

For the generalized gamma hCRM, when  $t \rightarrow 0$ , one obtains

$$G(t) \approx \frac{D^{1/\sigma}}{(\sigma \sigma_0^{1/\sigma_0})^{1/\sigma}} t^{1/\sigma \sigma_0} (\log \log(t^{-1}))^{1-1/\sigma \sigma_0},$$

which implies that condition (S7) is satisfied by taking  $r > 1/\sigma \sigma_0$ . □

**Detailed derivation of (14) from (13)** The multiplicative intensity likelihood for the competing risks model (6) is given in (13),

$$\mathcal{L}(\tilde{\boldsymbol{\mu}}; \mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n}) = \prod_{i=1}^n \left[ \int_{\mathbb{X}} k(T_i; x_i) \tilde{\mu}_{\Delta_i}(dx_i) \exp \left( - \sum_{\ell=1}^D \int_0^{T_i} \int_{\mathbb{X}} k(s; y) \tilde{\mu}_\ell(dy) ds \right) \right].$$

In Section 4, we rewrite this likelihood in terms of the extended random measures  $\tilde{\boldsymbol{\mu}}^e = (\tilde{\mu}_1^e, \dots, \tilde{\mu}_D^e)$ . Moreover, we further disintegrate with respect to such random measures by introducing two sequences of latent variables, namely  $\mathbf{X}_{1:n} = (X_1, \dots, X_n)$  and  $\mathbf{Z}_{1:n} = (Z_1, \dots, Z_n)$ . This leads to the doubly-augmented likelihood

$$\begin{aligned} \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n}, \mathbf{X}_{1:n}, \mathbf{Z}_{1:n}) \\ = \prod_{i=1}^n \left[ k(T_i; X_i) \tilde{\mu}_{\Delta_i}^e(dZ_i, dX_i) \exp \left( - \sum_{\ell=1}^D \int_0^{T_i} \int_{[0,1] \times \mathbb{X}} k(s; x) \tilde{\mu}_\ell^e(dz, dx) ds \right) \right]. \end{aligned}$$

The derivation of (14) from the expression above reflects the nested partition structure induced by the tied values in the sequences  $\mathbf{X}_{1:n}$  and  $\mathbf{Z}_{1:n}$ . Specifically, the coarser level of the partition is denoted by  $\Pi_{\mathbf{X}_{1:n}} = \{C_j^X: j = 1, \dots, k\}$ , for some  $k \in \{1, \dots, n\}$ , where each  $C_j^X = \{i: X_i = X_j^*\}$  contains the observations associated to location  $X_j^*$ . Therefore, the augmented likelihood becomes

$$\begin{aligned} \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n}, \mathbf{X}_{1:n}, \mathbf{Z}_{1:n}) &= \prod_{j=1}^k \prod_{i \in C_j^X} \left[ k(T_i; X_j^*) \tilde{\mu}_{\Delta_i}^e(dZ_i, dX_j^*) \right] \\ &\quad \times \exp \left( - \sum_{i=1}^n \sum_{\ell=1}^D \int_0^{T_i} \int_{[0,1] \times \mathbb{X}} k(s; x) \tilde{\mu}_\ell^e(dz, dx) ds \right), \end{aligned}$$

where the product over  $i$  has been decomposed as: (i) a product over  $j$ , i.e. over the elements of the partition, and (ii) a nested product over the observations in each group  $C_j^X$ . Remarkably, the observations in  $C_j^X$  do not necessarily share the same value of  $\Delta$ : indeed,  $i, \ell \in C_j^X$  does not imply  $\Delta_i = \Delta_\ell$ . The observed partition encoded in  $\boldsymbol{\Delta}_{1:n}$  is taken into account by considering, for each  $j \in \{1, \dots, k\}$ , the nested partition  $\{C_{\delta_j}^X: \delta = 1, \dots, D\}$ , where  $C_{\delta_j}^X = \{i \in C_j^X: \Delta_i = \delta\}$ . The product above is then rewritten as

$$\prod_{j=1}^k \prod_{i \in C_j^X} \left[ k(T_i; X_j^*) \tilde{\mu}_{\Delta_i}^e(dZ_i, dX_j^*) \right] = \left[ \prod_{j=1}^k \prod_{i \in C_j^X} k(T_i; X_j^*) \right] \left[ \prod_{j=1}^k \prod_{\delta=1}^D \prod_{i \in C_{\delta_j}^X} \mu_\delta^e(dZ_i, dX_j^*) \right].$$

Finally, at the finer level of the partition, for each  $j \in \{1, \dots, k\}$  and  $\delta \in \{1, \dots, D\}$ , the observations in  $C_{\delta_j}^X$  are further partitioned as  $\{C_{\delta_j h}^Z: h = 1, \dots, r_{\delta_j}\}$ , for some  $r_{\delta_j} > 0$ , where each  $C_{\delta_j h}^Z = \{i \in C_{\delta_j}^X: Z_i = Z_{\delta_j h}^*\}$  contains the observations associated to latent mark  $Z_{\delta_j h}^*$  from the random measure  $\tilde{\mu}_\delta^e$  paired with location  $X_j^*$ . Indeed,  $i \in C_{\delta_j h}^Z$  implies  $X_i = X_j^*$  and  $\Delta_i = \delta$ . Therefore, the nested product becomes

$$\begin{aligned} \prod_{j=1}^k \prod_{\delta=1}^D \prod_{i \in C_{\delta_j}^X} \mu_\delta^e(dZ_i, dX_j^*) &= \prod_{j=1}^k \prod_{\delta=1}^D \prod_{h=1}^{r_{\delta_j}} \prod_{i \in C_{\delta_j h}^Z} \mu_\delta^e(dZ_{\delta_j h}^*, dX_j^*) \\ &= \prod_{j=1}^k \prod_{\delta=1}^D \prod_{h=1}^{r_{\delta_j}} \mu_\delta^e(dZ_{\delta_j h}^*, dX_j^*)^{q_{\delta_j h}}, \end{aligned}$$

where  $q_{\delta jh} = \text{card}(C_{\delta jh}^Z)$  is the number of observations in  $C_{\delta jh}^Z$ . As a result, the augmented likelihood is expressed as

$$\begin{aligned} \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathbf{T}_{1:n}, \boldsymbol{\Delta}_{1:n}, \mathbf{X}_{1:n}, \mathbf{Z}_{1:n}) &= \left[ \prod_{j=1}^k \prod_{i \in C_j^X} k(T_i; X_j^*) \right] \left[ \prod_{j=1}^k \prod_{\delta=1}^D \prod_{h=1}^{r_{\delta j}} \mu_{\delta}^e(dZ_{\delta jh}^*, dX_j^*)^{q_{\delta jh}} \right] \\ &\quad \times \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} \sum_{i=1}^n \int_0^{T_i} k(s; x) ds \tilde{\mu}_{\ell}^e(dz, dx) \right), \end{aligned}$$

which coincides with (14) by substituting the quantities in (15).  $\square$

**Proof of Theorem 5** Consider the augmented likelihood function in (14),

$$\begin{aligned} \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \\ = Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} K_n(x) \tilde{\mu}_{\ell}^e(dz, dx) \right) \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \tilde{\mu}_{\ell}^e(dZ_{\ell jh}^*, dX_j^*)^{q_{\ell jh}}, \end{aligned}$$

where the quantities  $Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n})$  and  $K_n(x)$  are defined as in (15). Conditionally on  $\tilde{\mu}_0$ , the extended random measures in  $\tilde{\boldsymbol{\mu}}^e$  are independent CRMs on  $[0, 1] \times \mathbb{X}$ , characterized by the homogeneous Lévy intensity  $\tilde{\nu}^e(ds, dz, dx) = \rho(ds) H(dz) \tilde{\mu}_0(dx)$ . The conditional expectation of the augmented likelihood, given  $\tilde{\mu}_0$ , is

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0] &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \\ &\quad \times \mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} K_n(x) \tilde{\mu}_{\ell}^e(dz, dx) \right) \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \tilde{\mu}_{\ell}^e(dZ_{\ell jh}^*, dX_j^*)^{q_{\ell jh}} \mid \tilde{\mu}_0 \right]. \end{aligned}$$

By conditional independence of the extended random measures  $\tilde{\mu}_1^e, \dots, \tilde{\mu}_D^e$ , we have

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0] &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \\ &\quad \times \prod_{\ell=1}^D \mathbb{E} \left[ \exp \left( - \int_{[0,1] \times \mathbb{X}} K_n(x) \tilde{\mu}_{\ell}^e(dz, dx) \right) \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \tilde{\mu}_{\ell}^e(dZ_{\ell jh}^*, dX_j^*)^{q_{\ell jh}} \mid \tilde{\mu}_0 \right], \end{aligned}$$

and exploiting the identity in (S4) with  $\alpha(dz, dx) = H(dz) \tilde{\mu}_0(dx)$  and  $f(z, x) = K_n(x)$ , we obtain

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0] &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \prod_{\ell=1}^D \exp \left( - \int_{\mathbb{X}} \psi(K_n(x)) \tilde{\mu}_0(dx) \right) \\ &\quad \times \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell jh}; K_n(X_j^*)) H(dZ_{\ell jh}^*) \tilde{\mu}_0(dX_j^*) \end{aligned}$$

$$\begin{aligned}
&= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \exp\left(-\int_{\mathbb{X}} D\psi(K_n(x)) \tilde{\mu}_0(dx)\right) \\
&\quad \times \prod_{j=1}^k \tilde{\mu}_0(dX_j^*)^{r_j} \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_n(X_j^*)) H(dZ_{\ell j h}^*).
\end{aligned}$$

The random measure  $\tilde{\mu}_0$  at the root of the hierarchical prior is a CRM with homogeneous Lévy intensity  $\nu_0(ds, dx) = \rho_0(ds) \Lambda_0(dx)$ . Therefore, the joint marginal distribution of the observations  $(\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n})$  and the latent variables  $(\mathbf{X}_{1:n}, \mathbf{Z}_{1:n})$ , i.e. the expectation of the augmented likelihood, is

$$\begin{aligned}
\mathbb{P}(\mathcal{F}_n^*) &= \mathbb{E}[\mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*)] = \mathbb{E}[\mathbb{E}[\mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0]] \\
&= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \mathbb{E}\left[\exp\left(-\int_{\mathbb{X}} D\psi(K_n(x)) \tilde{\mu}_0(dx)\right) \prod_{j=1}^k \tilde{\mu}_0(dX_j^*)^{r_j}\right] \\
&\quad \times \prod_{j=1}^k \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_n(X_j^*)) H(dZ_{\ell j h}^*).
\end{aligned}$$

Leveraging again (S4) with  $\alpha(dx) = \Lambda_0(dx)$  and  $f(x) = D\psi(K_n(x))$ , we conclude that

$$\begin{aligned}
\mathbb{P}(\mathcal{F}_n^*) &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \exp\left(-\int_{\mathbb{X}} \psi_0(D\psi(K_n(x))) \Lambda_0(dx)\right) \\
&\quad \times \prod_{j=1}^k \tau_0(r_j; D\psi(K_n(X_j^*))) \Lambda_0(dX_j^*) \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_n(X_j^*)) H(dZ_{\ell j h}^*).
\end{aligned}$$

□

**Proof of Theorem 6** The predictive distribution for the observation  $(T_{n+1}, \Delta_{n+1})$  and the corresponding latent location  $X_{n+1}$  and latent mark  $Z_{n+1}$  is obtained from the joint marginal distribution in Theorem 5, and given by

$$\mathbb{P}(T_{n+1}, \Delta_{n+1}, X_{n+1}, Z_{n+1} \mid \mathcal{F}_n^*) = \frac{\mathbb{P}(T_{n+1}, \Delta_{n+1}, X_{n+1}, Z_{n+1}, \mathcal{F}_n^*)}{\mathbb{P}(\mathcal{F}_n^*)}. \quad (\text{S8})$$

where the numerator is the joint marginal distribution of the  $n + 1$  observations and corresponding latent variables.

The analytical form of the predictive distribution (S8) depends on the allocation of the observation  $n + 1$  within the latent partition structure, i.e. on the specific values assumed by  $X_{n+1}$  and  $Z_{n+1}$ . Three alternative cases can be identified:

- (1) both  $X_{n+1}$  and  $Z_{n+1}$  display ties with values in  $\mathbf{X}^*$  and  $\mathbf{Z}^*$ , respectively, say  $X_{n+1} = X_j^*$

and  $Z_{n+1} = Z_{\delta_{jh}}^*$ , where necessarily  $\Delta_{n+1} = \delta$ ;

- (2)  $X_{n+1}$  displays a tie with values in  $\mathbf{X}^*$ , say  $X_{n+1} = X_j^*$ , while  $Z_{n+1}$  assumes a value not included in  $\mathbf{Z}^*$ ;
- (3) both  $X_{n+1}$  and  $Z_{n+1}$  assume values not included in  $\mathbf{X}^*$  and  $\mathbf{Z}^*$ , respectively.

These three cases are highlighted in the following expression for the numerator:

$$\begin{aligned}
& \mathbb{P}(T_{n+1} \in dt, \Delta_{n+1} = \delta, X_{n+1} \in dx, Z_{n+1} \in dz, \mathcal{F}_n^*) \\
&= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) k(t; x) \exp\left(-\int_{\mathbb{X}} \psi_0(D\psi(K_{n+1}(y))) \Lambda_0(dy)\right) \\
&\times \prod_{j=1}^k \left( \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_{n+1}(X_j^*)) H(dZ_{\ell j h}^*) \right) \tau_0(r_j; D\psi(K_{n+1}(X_j^*))) \Lambda_0(dX_j^*) \\
&\times \left\{ \sum_{j=1}^k \sum_{h=1}^{r_{\delta j}} \frac{\tau(q_{\delta j h} + 1; K_{n+1}(X_j^*))}{\tau(q_{\delta j h}; K_{n+1}(X_j^*))} \delta_{X_j^*}(dx) \delta_{Z_{\delta j h}^*}(dz) \quad \leftarrow \text{case (1)} \right. \\
&\text{case (2)} \rightarrow \quad \left. + \sum_{j=1}^k \tau(1; K_{n+1}(X_j^*)) \frac{\tau_0(r_j + 1; D\psi(K_{n+1}(X_j^*)))}{\tau_0(r_j; D\psi(K_{n+1}(X_j^*)))} \delta_{X_j^*}(dx) H(dz) \right. \\
&\quad \left. \text{case (3)} \rightarrow \quad + \tau(1; K_{n+1}(x)) \tau_0(1; D\psi(K_{n+1}(x))) \Lambda_0(dx) H(dz) \right\} dt,
\end{aligned}$$

where  $K_{n+1}(x) = K_n(x) + \int_0^t k(s, x) ds$  is the updated kernel term, since  $T_{n+1} = t$ . Remarkably, these three cases coincide with the three cases appearing in the description of the marginal sampling algorithm (Section 6.1). The joint predictive distribution is thus obtained by dividing the expression above by the joint marginal distribution in Theorem 5:

$$\begin{aligned}
& \mathbb{P}(T_{n+1} \in dt, \Delta_{n+1} = \delta, X_{n+1} \in dx, Z_{n+1} \in dz \mid \mathcal{F}_n^*) \\
&= k(t; x) \exp\left(-\int_{\mathbb{X}} (\psi_0(D\psi(K_{n+1}(y))) - \psi_0(D\psi(K_n(y)))) \Lambda_0(dy)\right) \\
&\times \prod_{j=1}^k \left( \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \frac{\tau(q_{\ell j h}; K_{n+1}(X_j^*))}{\tau(q_{\ell j h}; K_n(X_j^*))} \right) \frac{\tau_0(r_j; D\psi(K_{n+1}(X_j^*)))}{\tau_0(r_j; D\psi(K_n(X_j^*)))}
\end{aligned}$$

$$\begin{aligned}
& \times \left\{ \sum_{j=1}^k \sum_{h=1}^{r_{\delta j}} \frac{\tau(q_{\delta j h} + 1; K_{n+1}(X_j^*))}{\tau(q_{\delta j h}; K_{n+1}(X_j^*))} \delta_{X_j^*}(dx) \delta_{Z_{\delta j h}^*}(dz) \right. \\
& \quad + \sum_{j=1}^k \tau(1; K_{n+1}(X_j^*)) \frac{\tau_0(r_j + 1; D\psi(K_{n+1}(X_j^*)))}{\tau_0(r_j; D\psi(K_{n+1}(X_j^*)))} \delta_{X_j^*}(dx) H(dz) \\
& \quad \left. + \tau(1; K_{n+1}(x)) \tau_0(1; D\psi(K_{n+1}(x))) \Lambda_0(dx) H(dz) \right\} dt.
\end{aligned}$$

The predictive distribution for observation  $(T_{n+1}, \Delta_{n+1})$  is obtained by marginalizing the latent variables

$$\begin{aligned}
\mathbb{P}(T_{n+1} \in dt, \Delta_{n+1} = \delta \mid \mathcal{F}_n) &= \exp\left(-\int_{\mathbb{X}} (\psi_0(D\psi(K_{n+1}(y))) - \psi_0(D\psi(K_n(y)))) \Lambda_0(dy)\right) \\
& \times \prod_{j=1}^k \left( \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \frac{\tau(q_{\ell j h}; K_{n+1}(X_j^*))}{\tau(q_{\ell j h}; K_n(X_j^*))} \right) \frac{\tau_0(r_j; D\psi(K_{n+1}(X_j^*)))}{\tau_0(r_j; D\psi(K_n(X_j^*)))} \\
& \times \left\{ \sum_{j=1}^k k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau(q_{\delta j h} + 1; K_{n+1}(X_j^*))}{\tau(q_{\delta j h}; K_{n+1}(X_j^*))} \right. \\
& \quad + \sum_{j=1}^k k(t; X_j^*) \tau(1; K_{n+1}(X_j^*)) \frac{\tau_0(r_j + 1; D\psi(K_{n+1}(X_j^*)))}{\tau_0(r_j; D\psi(K_{n+1}(X_j^*)))} \\
& \quad \left. + \int_{\mathbb{X}} k(t; x) \tau(1; K_{n+1}(x)) \tau_0(1; D\psi(K_{n+1}(x))) \Lambda_0(dx) \right\} dt. \quad (\text{S9})
\end{aligned}$$

Finally, the predictive distribution of  $\Delta_{n+1}$ , given the survival time  $T_{n+1}$  is proportional to the expression above, in which factors not depending on  $\Delta_{n+1} = \delta$  can be ignored:

$$\begin{aligned}
\mathbb{P}(\Delta_{n+1} = \delta \mid T_{n+1} = t, \mathcal{F}_n) &\propto \sum_{j=1}^k k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau(q_{\delta j h} + 1; K_{n+1}(X_j^*))}{\tau(q_{\delta j h}; K_{n+1}(X_j^*))} \\
& + \sum_{j=1}^k k(t; X_j^*) \tau(1; K_{n+1}(X_j^*)) \frac{\tau_0(r_j + 1; D\psi(K_{n+1}(X_j^*)))}{\tau_0(r_j; D\psi(K_{n+1}(X_j^*)))} \\
& + \int_{\mathbb{X}} k(t; x) \tau(1; K_{n+1}(x)) \tau_0(1; D\psi(K_{n+1}(x))) \Lambda_0(dx).
\end{aligned}$$

□

**Proof of Theorem 7** The posterior distribution of the random measure  $\tilde{\mu}_0$  at the root of the hierarchy, given  $\mathcal{F}_n^* = (\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n}, \mathbf{X}_{1:n}, \mathbf{Z}_{1:n})$ , can be characterized through its conditional Laplace transform,

$$\mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0(dx) \right) \mid \mathcal{F}_n^* \right] = \frac{\mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0(dx) \right) \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \right]}{\mathbb{E} [\mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*)]},$$

where  $h_0: \mathbb{X} \mapsto \mathbb{R}^+$  is any non-negative measurable function. The quantity at the denominator is the joint marginal distribution in Theorem 5, while the numerator can be computed similarly. Specifically, the expectation at the numerator, conditionally on  $\tilde{\mu}_0$ , is

$$\begin{aligned} & \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0(dx) \right) \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0 \right] \\ &= \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0(dx) \right) \mathbb{E} [\mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0] \\ &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \exp \left( - \int_{\mathbb{X}} (h_0(x) + D\psi(K_n(x))) \tilde{\mu}_0(dx) \right) \\ & \quad \times \prod_{j=1}^k \tilde{\mu}_0(dX_j^*)^{r_j} \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_n(X_j^*)) H(dZ_{\ell j h}^*). \end{aligned}$$

By marginalization of the conditional expectation with respect to  $\tilde{\mu}_0$ , we get

$$\begin{aligned} & \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0(dx) \right) \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \right] \\ &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} (h_0(x) + D\psi(K_n(x))) \tilde{\mu}_0(dx) \right) \prod_{j=1}^k \tilde{\mu}_0(dX_j^*)^{r_j} \right] \\ & \quad \times \prod_{j=1}^k \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_n(X_j^*)) H(dZ_{\ell j h}^*) \\ &= Q_n(\mathbf{T}_{1:n}, \mathbf{X}_{1:n}) \exp \left( - \int_{\mathbb{X}} \psi_0(h_0(x) + D\psi(K_n(x))) \Lambda_0(dx) \right) \prod_{j=1}^k \Lambda_0(dX_j^*) \\ & \quad \times \prod_{j=1}^k \tau_0(r_j; h_0(X_j^*) + D\psi(K_n(X_j^*))) \prod_{\ell=1}^D \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; K_n(X_j^*)) H(dZ_{\ell j h}^*), \end{aligned}$$

where the last identity is obtained resorting to (S4) with  $f(x) = h_0(x) + D\psi(K_n(x))$ . Therefore, the conditional Laplace transform is

$$\begin{aligned}
& \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0(dx) \right) \mid \mathcal{F}_n^* \right] \\
&= \exp \left( - \int_{\mathbb{X}} \left( \psi_0(h_0(x) + D\psi(K(x))) - \psi_0(D\psi(K(x))) \right) \Lambda_0(dx) \right) \\
& \quad \times \prod_{j=1}^k \frac{\tau_0(r_j; h_0(X_j^*) + D\psi(K(X_j^*)))}{\tau_0(r_j; D\psi(K(X_j^*)))}.
\end{aligned}$$

Thanks to simple algebra with Laplace exponents, the exponential term can be rewritten as

$$\begin{aligned}
& \exp \left( - \int_{\mathbb{X}} \left( \psi_0(h_0(x) + D\psi(K_n(x))) - \psi_0(D\psi(K_n(x))) \right) \Lambda_0(dx) \right) \\
&= \exp \left( - \int_{\mathbb{X}} \psi_{0,x}^+(h_0(x)) \Lambda_0(dx) \right) = \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0^+(dx) \right) \right],
\end{aligned}$$

where  $\psi_{0,x}^+(\cdot)$  is the Laplace exponent corresponding to  $\rho_0^+(ds \mid x) = \exp(-D\psi(K_n(x))s) \rho_0(ds)$ , and  $\tilde{\mu}_0^+$  is a CRM with Lévy intensity  $\nu_0^+(ds, dx) = \rho_0^+(ds \mid x) \Lambda_0(dx)$ . Moreover, for each  $j = 1, \dots, k$ ,

$$\begin{aligned}
\tau_0(r_j; h_0(X_j^*) + D\psi(K_n(X_j^*))) &= \int_{\mathbb{R}^+} s^{r_j} e^{-h_0(X_j^*)s} \exp(-D\psi(K_n(X_j^*))s) \rho_0(ds) \\
&= \int_{\mathbb{R}^+} e^{-h_0(X_j^*)s} s^{r_j} \rho_0^+(ds \mid X_j^*).
\end{aligned}$$

Therefore, the ratios in the expression of the conditional Laplace transform can be regarded as expectations,

$$\frac{\tau_0(r_j; h_0(X_j^*) + D\psi(K_n(X_j^*)))}{\tau_0(r_j; D\psi(K_n(X_j^*)))} = \frac{\int_{\mathbb{R}^+} e^{-h_0(X_j^*)s} s^{r_j} \rho_0^+(ds \mid X_j^*)}{\int_{\mathbb{R}^+} s^{r_j} \rho_0^+(ds \mid X_j^*)} = \mathbb{E} \left[ \exp(-h_0(X_j^*)V_j) \right],$$

where each  $V_j$  is a random variable with density function

$$f_{V_j}(s) ds \propto s^{r_j} \exp(-D\psi(K_n(X_j^*))s) \rho_0(ds) = s^{r_j} \rho_0^+(ds \mid X_j^*).$$

In conclusion, the conditional Laplace transform can be written as

$$\begin{aligned}
\mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0(dx) \right) \mid \mathcal{F}_n^* \right] &= \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0^+(dx) \right) \right] \prod_{j=1}^k \mathbb{E} \left[ \exp(-h_0(X_j^*)V_j) \right] \\
&= \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} h_0(x) \tilde{\mu}_0^+(dx) - \sum_{j=1}^k h_0(X_j^*)V_j \right) \right],
\end{aligned}$$

which highlights the structure of the posterior distribution of  $\tilde{\mu}_0$  as the sum of the non-homogeneous CRM  $\tilde{\mu}_0^*$  and the random jumps  $V_1, \dots, V_k$  at the fixed locations  $X_1^*, \dots, X_k^*$ ; moreover, these components are mutually independent (second line).

A similar characterization holds for the posterior distribution of the extended random measures  $\tilde{\mu}_1^e, \dots, \tilde{\mu}_D^e$  at the bottom of the hierarchy. Indeed, their joint conditional Laplace transform, given  $\mathcal{F}_n^* = (\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n}, \mathbf{X}_{1:n}, \mathbf{Z}_{1:n})$  and the root random measure  $\tilde{\mu}_0$ , is

$$\begin{aligned} & \mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} h_\ell(z, x) \tilde{\mu}_\ell^e(dz, dx) \right) \mid \mathcal{F}_n^*, \tilde{\mu}_0 \right] \\ &= \frac{\mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} h_\ell(z, x) \tilde{\mu}_\ell^e(dz, dx) \right) \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0 \right]}{\mathbb{E} [\mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0]}, \end{aligned}$$

where  $h_\delta: [0, 1] \times \mathbb{X} \mapsto \mathbb{R}^+$ , for  $\delta \in \{1, \dots, D\}$ , are non-negative measurable functions. The quantity at the denominator is the conditional expectation of the augmented likelihood, obtained as an intermediate step in the proof of Theorem 5, while the expectation at the numerator can be computed similarly. In particular, we have

$$\begin{aligned} & \mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} h_\ell(z, x) \tilde{\mu}_\ell^e(dz, dx) \right) \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0 \right] \\ &= Q(\mathbf{T}, \mathbf{X}) \mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} (h_\ell(z, x) + K_n(x)) \tilde{\mu}_\ell^e(dz, dx) \right) \right. \\ & \quad \left. \times \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \tilde{\mu}_\ell^e(dZ_{\ell j h}^*, dX_j^*)^{q_{\ell j h}} \right]. \end{aligned}$$

By conditional independence of the extended random measures  $\tilde{\mu}_1^e, \dots, \tilde{\mu}_D^e$ , and leveraging again (S4) with  $f(z, x) = h_\delta(z, x) + K_n(x)$ , for  $\delta = 1, \dots, D$ , we obtain

$$\begin{aligned} & \mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} h_\ell(z, x) \tilde{\mu}_\ell^e(dz, dx) \right) \mathcal{L}(\tilde{\boldsymbol{\mu}}^e; \mathcal{F}_n^*) \mid \tilde{\mu}_0 \right] \\ &= Q(\mathbf{T}, \mathbf{X}) \prod_{\ell=1}^D \mathbb{E} \left[ \exp \left( - \int_{[0,1] \times \mathbb{X}} (h_\ell(z, x) + K_n(x)) \tilde{\mu}_\ell^e(dz, dx) \right) \right. \\ & \quad \left. \times \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \tilde{\mu}_\ell^e(dZ_{\ell j h}^*, dX_j^*)^{q_{\ell j h}} \right] \end{aligned}$$

$$\begin{aligned}
&= Q(\mathbf{T}, \mathbf{X}) \prod_{\ell=1}^D \exp \left( - \int_{[0,1] \times \mathbb{X}} \psi(h_\ell(z, x) + K_n(x)) H(dz) \tilde{\mu}_0(dx) \right) \\
&\quad \times \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \tau(q_{\ell j h}; h_\ell(Z_{\ell j h}^*, X_j^*) + K_n(X_j^*)) H(dZ_{\ell j h}^*) \tilde{\mu}_0(dX_j^*).
\end{aligned}$$

Therefore, the conditional Laplace transform is

$$\begin{aligned}
&\mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} h_\ell(z, x) \tilde{\mu}_\ell^e(dz, dx) \right) \mid \mathcal{F}_n^*, \tilde{\mu}_0 \right] \\
&= \prod_{\ell=1}^D \exp \left( - \int_{[0,1] \times \mathbb{X}} \left( \psi(h_\ell(z, x) + K_n(x)) - \psi(K_n(x)) \right) H(dz) \tilde{\mu}_0(dx) \right) \\
&\quad \times \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau(q_{\ell j h}; h_\ell(Z_{\ell j h}^*, X_j^*) + K_n(X_j^*))}{\tau(q_{\ell j h}; K_n(X_j^*))}.
\end{aligned}$$

For each  $\delta = 1, \dots, D$ , the exponential term can be rewritten as

$$\begin{aligned}
&\exp \left( - \int_{[0,1] \times \mathbb{X}} \left( \psi(h_\delta(z, x) + K_n(x)) - \psi(K_n(x)) \right) H(dz) \tilde{\mu}_0(dx) \right) \\
&= \mathbb{E} \left[ \exp \left( - \int_{[0,1] \times \mathbb{X}} h_\delta(z, x) \tilde{\mu}_\delta^+(dz, dx) \right) \right],
\end{aligned}$$

where  $\tilde{\mu}_\delta^+$  is a CRM with non-homogeneous Lévy intensity  $\nu^+(ds, dz, dx) = \rho^+(ds \mid x) H(dz) \tilde{\mu}_0(dx)$ , with jump component  $\rho^+(ds \mid x) = e^{-K_n(x)s} \rho(ds)$ .

Moreover, for each  $\delta = 1, \dots, D$ ,  $j = 1, \dots, k$ , and  $h = 1, \dots, r_{\delta j}$ ,

$$\tau(q_{\delta j h}; h_\delta(Z_{\delta j h}^*, X_j^*) + K_n(X_j^*)) = \int_{\mathbb{R}_+} e^{-h_\delta(Z_{\delta j h}^*, X_j^*)s} s^{q_{\delta j h}} \rho^+(ds \mid X_j^*).$$

Therefore, the ratios in the expression of the conditional Laplace transform can be regarded as expectations

$$\frac{\tau(q_{\delta j h}; h_\delta(Z_{\delta j h}^*, X_j^*) + K_n(X_j^*))}{\tau(q_{\delta j h}; K_n(X_j^*))} = \mathbb{E} \left[ \exp \left( -h_\delta(Z_{\delta j h}^*, X_j^*) S_{\delta j h} \right) \right],$$

where each  $S_{\delta j h}$  is a non-negative random variable having density function proportional to

$$s^{q_{\delta j h}} e^{-K_n(X_j^*)s} \rho(ds) = s^{q_{\delta j h}} \rho^+(ds \mid X_j^*).$$

In conclusion, the conditional Laplace transform can be written as

$$\begin{aligned}
\mathbb{E} & \left[ \exp \left( - \sum_{\ell=1}^D \int_{[0,1] \times \mathbb{X}} h_{\ell}(z, x) \tilde{\mu}_{\ell}^{\varepsilon}(dz, dx) \right) \mid \mathcal{F}_n^*, \tilde{\mu}_0 \right] \\
& = \prod_{\ell=1}^D \mathbb{E} \left[ \exp \left( - \int_{[0,1] \times \mathbb{X}} h_{\ell}(z, x) \tilde{\mu}_{\ell}^+(dz, dx) \right) \right] \prod_{j=1}^k \prod_{h=1}^{r_{\ell_j}} \mathbb{E} \left[ \exp \left( -h_{\ell}(Z_{\ell_{jh}}^*, X_j^*) S_{\ell_{jh}} \right) \right] \\
& = \prod_{\ell=1}^D \mathbb{E} \left[ \exp \left( - \int_{[0,1] \times \mathbb{X}} h_{\ell}(z, x) \tilde{\mu}_{\ell}^+(dz, dx) - \sum_{j=1}^k \sum_{h=1}^{r_{\ell_j}} h_{\ell}(Z_{\ell_{jh}}^*, X_j^*) S_{\ell_{jh}} \right) \right],
\end{aligned}$$

which highlights the structure of the posterior distribution of each  $\tilde{\mu}_{\delta}$ , given the root random measure  $\tilde{\mu}_0$ , as the sum of the non-homogeneous CRM  $\tilde{\mu}_{\delta}^*$  and the random jumps  $(S_{\delta_{jh}})_{jh}$  at fixed locations. These components are mutually independent, and the random measures  $\tilde{\mu}_1^{\varepsilon}, \dots, \tilde{\mu}_D^{\varepsilon}$  are conditionally independent, given  $\tilde{\mu}_0$ . The posterior characterization in Theorem 7 is obtained by marginalization with respect to the first component of the random measure, that is  $\tilde{\mu}_{\delta}(dx) = \tilde{\mu}_{\delta}^{\varepsilon}([0, 1], dx)$ , for  $\delta = 1, \dots, D$ .  $\square$

**Proof of Proposition 8** The posterior estimate of the overall survival function  $\tilde{S}(t)$ , given  $\mathcal{F}_n = (\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n}, \Pi_{\mathbf{X}_{1:n}}, \Pi_{\mathbf{Z}_{1:n}}, \mathbf{X}^*)$ , at  $t \geq 0$ , is obtained as

$$\begin{aligned}
\mathbb{E} [\tilde{S}(t) \mid \mathcal{F}_n] & = \mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_0^t \int_{\mathbb{X}} k(s; x) \tilde{\mu}_{\ell}(dx) ds \right) \mid \mathcal{F}_n \right] \\
& = \mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{\mathbb{X}} K_1(x) \tilde{\mu}_{\ell}(dx) \right) \mid \mathcal{F}_n \right],
\end{aligned}$$

where  $K_1(x) = K_1(x; t) = \int_0^t k(s; x) ds$  is the integrated kernel up to time  $t \geq 0$ . Conditionally on the root random measure  $\tilde{\mu}_0$ , the expressions for the posterior distributions of the random measures  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  derived in Theorem 7 are substituted above, leading to

$$\begin{aligned}
\mathbb{E} [\tilde{S}(t) \mid \mathcal{F}_n, \tilde{\mu}_0] & = \mathbb{E} \left[ \exp \left( - \sum_{\ell=1}^D \int_{\mathbb{X}} K_1(x) \tilde{\mu}_{\ell}^+(dx) - \sum_{\ell=1}^D \sum_{j=1}^k \sum_{h=1}^{r_{\ell_j}} K_1(X_j^*) S_{\ell_{jh}} \right) \mid \tilde{\mu}_0 \right] \\
& = \mathbb{E} \left[ \prod_{\ell=1}^D \exp \left( - \int_{\mathbb{X}} K_1(x) \tilde{\mu}_{\ell}^+(dx) \right) \prod_{j=1}^k \prod_{h=1}^{r_{\ell_j}} \exp \left( -K_1(X_j^*) S_{\ell_{jh}} \right) \mid \tilde{\mu}_0 \right];
\end{aligned}$$

By the conditional independence *a posteriori* of the random measures  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ , and of the their components, we have

$$\mathbb{E} \left[ \tilde{S}(t) \mid \mathcal{F}_n, \tilde{\mu}_0 \right] = \prod_{\ell=1}^D \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} K_1(x) \tilde{\mu}_\ell^+(dx) \right) \mid \tilde{\mu}_0 \right] \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \mathbb{E} \left[ \exp \left( -K_1(X_j^*) S_{\ell j h} \right) \right].$$

The posterior characterization of the random measures  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  implies that

$$\begin{aligned} \mathbb{E} \left[ \tilde{S}(t) \mid \mathcal{F}_n, \tilde{\mu}_0 \right] &= \prod_{\ell=1}^D \exp \left( - \int_{\mathbb{X}} \psi_x^+(K_1(x)) \tilde{\mu}_0(dx) \right) \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+(q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\ell j h}; 0)} \\ &= \exp \left( - \int_{\mathbb{X}} D\psi_x^+(K_1(x)) \tilde{\mu}_0(dx) \right) \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+(q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\ell j h}; 0)}, \end{aligned}$$

where  $\psi_x^+$  and  $\tau_x^+$  are the Laplace exponent and cumulants, respectively, corresponding to  $\rho^+$  and defined in (21). The posterior estimate is obtained by substituting the specification of the posterior distribution of the root random measure  $\tilde{\mu}_0$ , and exploiting the mutual independence of its components,

$$\begin{aligned} \mathbb{E} \left[ \tilde{S}(t) \mid \mathcal{F}_n \right] &= \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} D\psi_x^+(K_1(x)) \tilde{\mu}_0^+(dx) \right) \right] \\ &\quad \times \prod_{j=1}^k \mathbb{E} \left[ \exp \left( -D\psi_{X_j^*}^+(K_1(X_j^*)) V_j \right) \right] \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+(q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\ell j h}; 0)}. \end{aligned}$$

In conclusion, the posterior characterization of the random measure  $\tilde{\mu}_0$  implies that

$$\begin{aligned} \mathbb{E} \left[ \tilde{S}(t) \mid \mathcal{F}_n \right] &= \exp \left( - \int_{\mathbb{X}} \psi_{0,x}^+(D\psi_x^+(K_1(x))) \Lambda_0(dx) \right) \\ &\quad \times \prod_{j=1}^k \frac{\tau_{0,X_j^*}^+(r_j; D\psi_{X_j^*}^+(K_1(X_j^*)))}{\tau_{0,X_j^*}^+(r_j; 0)} \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+(q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\ell j h}; 0)}, \end{aligned}$$

where  $\psi_{0,x}^+$  and  $\tau_{0,x}^+$  are the Laplace exponent and cumulants corresponding to  $\rho_0^+$ .  $\square$

**Proof of Proposition 9** The proof follows a similar line of reasoning to that in the proof of Proposition 8. Indeed, the posterior estimate of the cause-specific incidence function  $\tilde{p}(dt, \delta) = \tilde{f}_\delta(t) dt$  for  $\delta \in \{1, \dots, D\}$ , given  $\mathcal{F}_n = (\mathbf{T}_{1:n}, \mathbf{\Delta}_{1:n}, \Pi_{\mathbf{X}_{1:n}}, \Pi_{\mathbf{Z}_{1:n}}, \mathbf{X}^*)$ , at  $t \geq 0$ , is obtained as

$$\mathbb{E} \left[ \tilde{p}(dt, \delta) \mid \mathcal{F}_n \right] = \mathbb{E} \left[ \int_{\mathbb{X}} k(t; x) \tilde{\mu}_\delta(dx) \exp \left( - \sum_{\ell=1}^D \int_{\mathbb{X}} K_1(y) \tilde{\mu}_\ell(dy) \right) dt \mid \mathcal{F}_n \right],$$

where  $K_1(x) = K_1(x; t) = \int_0^t k(s; x) ds$  is the integrated kernel up to time  $t \geq 0$ . Conditionally on the root random measure  $\tilde{\mu}_0$ , the expressions for the posterior distributions of the random

measures  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  are inserted into the previous formula yielding

$$\begin{aligned} \mathbb{E}[\tilde{p}(dt, \delta) \mid \mathcal{F}_n, \tilde{\mu}_0] &= \mathbb{E} \left[ \prod_{\ell \neq \delta} \exp \left( - \int_{\mathbb{X}} K_1(y) \tilde{\mu}_\ell^+(dy) \right) \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \exp \left( -K_1(X_j^*) S_{\ell j h} \right) \right. \\ &\quad \times \left( \int_{\mathbb{X}} k(t; x) \tilde{\mu}_\delta^+(dx) \exp \left( - \int_{\mathbb{X}} K_1(y) \tilde{\mu}_\delta^+(dy) \right) \prod_{j=1}^k \prod_{h=1}^{r_{\delta j}} \exp \left( -K_1(X_j^*) S_{\delta j h} \right) \right. \\ &\quad \left. \left. + \sum_{\ell=1}^k k(t; X_\ell^*) \sum_{\varepsilon=1}^{r_{\delta \ell}} \exp \left( - \int_{\mathbb{X}} K_1(y) \tilde{\mu}_\delta^+(dy) \right) S_{\delta \ell \varepsilon} \prod_{j=1}^k \prod_{h=1}^{r_{\delta j}} \exp \left( -K_1(X_j^*) S_{\delta j h} \right) \right) dt \mid \tilde{\mu}_0 \right]. \end{aligned}$$

By the conditional independence *a posteriori* of  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ , and of the their components,

$$\begin{aligned} \mathbb{E}[\tilde{p}(dt, \delta) \mid \mathcal{F}_n, \tilde{\mu}_0] &= \prod_{\ell \neq \delta} \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} K_1(y) \tilde{\mu}_\ell^+(dy) \right) \mid \tilde{\mu}_0 \right] \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \mathbb{E} \left[ \exp \left( -K_1(X_j^*) S_{\ell j h} \right) \right] \\ &\quad \times \left( \int_{\mathbb{X}} k(t; x) \mathbb{E} \left[ \tilde{\mu}_\delta^+(dx) \exp \left( - \int_{\mathbb{X}} K_1(y) \tilde{\mu}_\delta^+(dy) \right) \mid \tilde{\mu}_0 \right] \prod_{j=1}^k \prod_{h=1}^{r_{\delta j}} \mathbb{E} \left[ \exp \left( -K_1(X_j^*) S_{\delta j h} \right) \right] \right. \\ &\quad \left. + \sum_{\ell=1}^k k(t; X_\ell^*) \sum_{\varepsilon=1}^{r_{\delta \ell}} \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} K_1(y) \tilde{\mu}_\delta^+(dy) \right) \mid \tilde{\mu}_0 \right] \mathbb{E} [S_{\delta \ell \varepsilon} \exp \left( -K_1(X_\ell^*) S_{\delta \ell \varepsilon} \right)] \right. \\ &\quad \left. \times \prod_{j \neq \ell} \prod_{h \neq \varepsilon} \mathbb{E} \left[ \exp \left( -K_1(X_j^*) S_{\delta j h} \right) \right] \right) dt. \end{aligned}$$

The posterior characterization of the random measures  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$  in Theorem 7, implies that

$$\begin{aligned} \mathbb{E}[\tilde{p}(dt, \delta) \mid \mathcal{F}_n, \tilde{\mu}_0] &= \exp \left( -(D-1) \int_{\mathbb{X}} \psi_y^+(K_1(y)) \tilde{\mu}_0(dy) \right) \prod_{\ell \neq \delta} \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+(q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\ell j h}; 0)} \\ &\quad \times \left( \int_{\mathbb{X}} k(t; x) \tau_x^+(1; K_1(x)) \tilde{\mu}_0(dx) \exp \left( - \int_{\mathbb{X}} \psi_y^+(K_1(y)) \tilde{\mu}_0(dy) \right) \prod_{j=1}^k \prod_{h=1}^{r_{\delta j}} \frac{\tau_{X_j^*}^+(q_{\delta j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\delta j h}; 0)} \right. \\ &\quad \left. + \sum_{\ell=1}^k k(t; X_\ell^*) \sum_{\varepsilon=1}^{r_{\delta \ell}} \exp \left( - \int_{\mathbb{X}} \psi_y^+(K_1(y)) \tilde{\mu}_0(dy) \right) \frac{\tau_{X_\ell^*}^+(q_{\delta \ell \varepsilon} + 1; K_1(X_\ell^*))}{\tau_{X_\ell^*}^+(q_{\delta \ell \varepsilon}; 0)} \right. \\ &\quad \left. \times \prod_{j \neq \ell} \prod_{h \neq \varepsilon} \frac{\tau_{X_j^*}^+(q_{\delta j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\delta j h}; 0)} \right) dt. \end{aligned}$$

Rearranging the terms in the expression above, we obtain

$$\begin{aligned} \mathbb{E}[\tilde{p}(dt, \delta) \mid \mathcal{F}_n, \tilde{\mu}_0] &= \exp\left(-\int_{\mathbb{X}} D\psi_y^+(K_1(y)) \tilde{\mu}_0(dy)\right) \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+(q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\ell j h}; 0)} \\ &\quad \times \left( \int_{\mathbb{X}} k(t; x) \tau_x^+(1; K_1(x)) \tilde{\mu}_0(dx) + \sum_{j=1}^k k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau_{X_j^*}^+(q_{\delta j h} + 1; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\delta j h}; K_1(X_j^*))} \right) dt. \end{aligned}$$

The posterior estimator is then derived by inserting the expression of the posterior distribution of the root random measure  $\tilde{\mu}_0$ ,

$$\begin{aligned} \mathbb{E}[\tilde{p}(dt, \delta) \mid \mathcal{F}_n] &= \mathbb{E} \left[ \exp\left(-\int_{\mathbb{X}} D\psi_y^+(K_1(y)) \tilde{\mu}_0^+(dy)\right) \prod_{j=1}^k \exp\left(-D\psi_{X_j^*}^+(K_1(X_j^*)) V_j\right) \right. \\ &\quad \times \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+(q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\ell j h}; 0)} \left( \int_{\mathbb{X}} k(t; x) \tau_x^+(1; K_1(x)) \tilde{\mu}_0^+(dx) \right. \\ &\quad \left. \left. + \sum_{j=1}^k k(t; X_j^*) \tau_{X_j^*}^+(1; K_1(X_j^*)) V_j + \sum_{j=1}^k k(t; X_j^*) \sum_{h=1}^{r_{\delta j}} \frac{\tau_{X_j^*}^+(q_{\delta j h} + 1; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\delta j h}; K_1(X_j^*))} \right) \right] dt. \end{aligned}$$

By rearranging the terms and invoking the mutual independence of  $\tilde{\mu}_0^+$  and  $V_1, \dots, V_k$ , one obtains a sum of three components:

$$\mathbb{E}[\tilde{p}(dt, \delta) \mid \mathcal{F}_n] = \prod_{\ell=1}^D \prod_{j=1}^k \prod_{h=1}^{r_{\ell j}} \frac{\tau_{X_j^*}^+(q_{\ell j h}; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\ell j h}; 0)} \left( I_1(t) + I_2(t) + I_3(t, \delta) \right) dt,$$

where the quantities  $I_1(t)$ ,  $I_2(t)$  and  $I_3(t, \delta)$  are given by

$$\begin{aligned} I_1(t) &= \int_{\mathbb{X}} k(t; x) \tau_x^+(1; K_1(x)) \mathbb{E} \left[ \tilde{\mu}_0^+(dx) \exp\left(-\int_{\mathbb{X}} D\psi_y^+(K_1(y)) \tilde{\mu}_0^+(dy)\right) \right] \\ &\quad \times \prod_{j=1}^k \mathbb{E} \left[ \exp\left(-D\psi_{X_j^*}^+(K_1(X_j^*)) V_j\right) \right], \\ I_2(t) &= \sum_{\ell=1}^k k(t; X_\ell^*) \tau_{X_\ell^*}^+(1; K_1(X_\ell^*)) \mathbb{E} \left[ \exp\left(-\int_{\mathbb{X}} D\psi_y^+(K_1(y)) \tilde{\mu}_0^+(dy)\right) \right] \\ &\quad \times \mathbb{E} \left[ V_\ell \exp\left(-D\psi_{X_\ell^*}^+(K_1(X_\ell^*)) V_\ell\right) \right] \prod_{j \neq \ell} \mathbb{E} \left[ \exp\left(-D\psi_{X_j^*}^+(K_1(X_j^*)) V_j\right) \right], \end{aligned}$$

$$I_3(t, \delta) = \sum_{\ell=1}^k k(t; X_\ell^*) \sum_{h=1}^{r\delta\ell} \frac{\tau_{X_\ell^*}^+(q\delta\ell h + 1; K_1(X_\ell^*))}{\tau_{X_\ell^*}^+(q\delta\ell h; K_1(X_\ell^*))} \mathbb{E} \left[ \exp \left( - \int_{\mathbb{X}} D\psi_y^+(K_1(y)) \tilde{\mu}_0^+(dy) \right) \right] \\ \times \prod_{j=1}^k \mathbb{E} \left[ \exp \left( - D\psi_{X_j^*}^+(K_1(X_j^*)) V_j \right) \right].$$

As for the first component, by the posterior characterization of  $\tilde{\mu}_0$ , and resorting to the identity (S4) with  $f(y) = D\psi_y^+(K_1(y))$ , we get

$$I_1(t) = \int_{\mathbb{X}} k(t; x) \tau_x^+(1; K_1(x)) \tau_{0,x}^+(1; D\psi_x^+(K_1(x))) \Lambda_0(dx) \\ \times \exp \left( - \int_{\mathbb{X}} \psi_{0,y}^+(D\psi_y^+(K_1(y))) \Lambda_0(dy) \right) \prod_{j=1}^k \frac{\tau_{0,X_j^*}^+(r_j; D\psi_{X_j^*}^+(K_1(X_j^*)))}{\tau_{0,X_j^*}^+(r_j; 0)},$$

while for the second and third component, the expectation with respect to  $\tilde{\mu}_0$  and  $V_1, \dots, V_k$  reads

$$I_2(t) = \sum_{\ell=1}^k k(t; X_\ell^*) \tau_{X_\ell^*}^+(1; K_1(X_\ell^*)) \exp \left( - \int_{\mathbb{X}} \psi_{0,y}^+(D\psi_y^+(K_1(y))) \Lambda_0(dy) \right) \\ \times \frac{\tau_{0,X_\ell^*}^+(r_\ell + 1; D\psi_{X_\ell^*}^+(K_1(X_\ell^*)))}{\tau_{0,X_\ell^*}^+(r_\ell; 0)} \prod_{j \neq \ell} \frac{\tau_{0,X_j^*}^+(r_j; D\psi_{X_j^*}^+(K_1(X_j^*)))}{\tau_{0,X_j^*}^+(r_j; 0)},$$

$$I_3(t, \delta) = \sum_{\ell=1}^k k(t; X_\ell^*) \sum_{h=1}^{r\delta\ell} \frac{\tau_{X_\ell^*}^+(q\delta\ell h + 1; K_1(X_\ell^*))}{\tau_{X_\ell^*}^+(q\delta\ell h; K_1(X_\ell^*))} \exp \left( - \int_{\mathbb{X}} \psi_{0,y}^+(D\psi_y^+(K_1(y))) \Lambda_0(dy) \right) \\ \times \prod_{j=1}^k \frac{\tau_{0,X_j^*}^+(r_j; D\psi_{X_j^*}^+(K_1(X_j^*)))}{\tau_{0,X_j^*}^+(r_j; 0)}.$$

In conclusion, by collecting and rearranging the above terms, we finally obtain

$$\mathbb{E}[\tilde{p}(dt, \delta) \mid \mathcal{F}_n] = \exp \left( - \int_{\mathbb{X}} \psi_{0,y}^+(D\psi_y^+(K_1(y))) \Lambda_0(dy) \right) \\ \times \prod_{j=1}^k \frac{\tau_{0,X_j^*}^+(r_j; D\psi_{X_j^*}^+(K_1(X_j^*)))}{\tau_{0,X_j^*}^+(r_j; 0)} \prod_{\ell=1}^D \prod_{h=1}^{r\ell_j} \frac{\tau_{X_j^*}^+(q\ell_j h; K_1(X_j^*))}{\tau_{X_j^*}^+(q\ell_j h; 0)} \\ \times \left( \int_{\mathbb{X}} k(t; x) \tau_x^+(1; K_1(x)) \tau_{0,x}^+(1; D\psi_x^+(K_1(x))) \Lambda_0(dx) \right)$$

$$\begin{aligned}
& + \sum_{j=1}^k k(t; X_j^*) \tau_{X_j^*}^+(1; K_1(X_j^*)) \frac{\tau_{0, X_j^*}^+(r_j + 1; D\psi_{X_j^*}^+(K_1(X_j^*)))}{\tau_{0, X_j^*}^+(r_j; D\psi_{X_j^*}^+(K_1(X_j^*)))} \\
& + \sum_{j=1}^k k(t; X_j^*) \sum_{h=1}^{r_{\delta_j}} \frac{\tau_{X_j^*}^+(q_{\delta_j h} + 1; K_1(X_j^*))}{\tau_{X_j^*}^+(q_{\delta_j h}; K_1(X_j^*))} \Big) dt. \quad (\text{S10})
\end{aligned}$$

Note that the factors in the first two lines of this expression coincide with the posterior estimate of the survival function obtained in Proposition 8.  $\square$

**Alternative proof of Theorem 6 based on Proposition 9 (Remark 10)** The posterior estimate of  $\tilde{p}(dt, \delta)$  detailed in Proposition 9, and more explicitly in (S10), coincides with the predictive distribution for observation  $(T_{n+1}, \Delta_{n+1})$  described in (S9). This fact can be shown by clarifying the relationships of the Laplace exponents and cumulants for the random measures  $\tilde{\mu}_1^+, \dots, \tilde{\mu}_D^+$ , defined in (21) with the corresponding quantities for  $\tilde{\mu}_1, \dots, \tilde{\mu}_D$ , defined in (16). At the bottom of the hierarchy, the Laplace exponents are related as

$$\psi_x^+(u) = \psi(u + K_n(x)) - \psi(K_n(x)),$$

or equivalently,  $\psi_x^+(u) + \psi(K_n(x)) = \psi(u + K_n(x))$ , while at the root of the hierarchy

$$\psi_{0,x}^+(u) = \psi_0(u + D\psi(K_n(x))) - \psi_0(D\psi(K_n(x))).$$

Therefore, exploiting the expressions above and the fact that the updated kernel term is  $K_{n+1}(x) = K_n(x) + \int_0^t k(s; x) ds = K_n(x) + K_1(x)$ , the exponential term in (S10) becomes

$$\begin{aligned}
& \exp\left(-\int_{\mathbb{X}} \psi_{0,y}^+(D\psi_y^+(K_1(y))) \Lambda_0(dy)\right) \\
& = \exp\left(-\int_{\mathbb{X}} (\psi_0(D\psi(K_{n+1}(y))) - \psi_0(D\psi(K_n(y)))) \Lambda_0(dy)\right).
\end{aligned}$$

On the other hand, for the cumulants we have

$$\tau_x^+(m; u) = \tau(m; u + K_n(x)), \quad \tau_{0,x}^+(m; u) = \tau_0(m; u + D\psi(K_n(x))).$$

The quantities appearing in (S10) can be thus rewritten as

$$\tau_{0,x}^+(m; D\psi_x^+(K_1(x))) = \tau_0(m; D\psi_x^+(K_1(x)) + D\psi(K_n(x))) = \tau_0(m; D\psi(K_{n+1}(x))),$$

while  $\tau_{0,x}^+(m; 0) = \tau_0(m; D\psi(K_n(x)))$ , and moreover

$$\tau_x^+(m; K_1(x)) = \tau(m; K_{n+1}(x)), \quad \tau_x^+(m; 0) = \tau(m; K_n(x))$$

In conclusion, the expression in (S9) can be obtained from (S10) using the identities above. The result in Theorem 6 is then retrieved conditioning on the survival time  $T_{n+1} = t$ .  $\square$

## S4 Algorithms for generalized gamma mixtures

In the following, we specialize the sampling algorithms described in Section 6 to the subclass of *generalized gamma* hCRMs (11), with the kernel choices considered in Section 2.

### S4.1 Marginal sampling algorithm (Section 6.1)

The marginal Gibbs sampling scheme for the reallocation of observation  $i \in \{1, \dots, n\}$  to the latent partition structure simplifies to:

- (1) for each latent location  $X_j^*$  such that  $X_j^* \leq T_i$ ,

$$\mathbb{P}(X_i = X_j^*, Z_i = Z_{\delta_{jh}}^* \mid \mathcal{F}_n^{-i}) \propto k(T_i; X_j^*) \frac{q_{\delta_{jh}}^{-i} - \sigma}{\beta + K_n(X_j^*)}, \quad h = 1, \dots, r_{\delta_j};$$

- (2) for each latent location  $X_j^*$  such that  $X_j^* \leq T_i$ ,

$$\begin{aligned} \mathbb{P}(X_i = X_j^*, Z_i = \text{'new'} \mid \mathcal{F}_n^{-i}) \\ \propto k(T_i; X_j^*) (\beta + K_n(X_j^*))^{\sigma-1} \frac{r_j^{-i} - \sigma_0}{\beta_0 + \frac{D}{\sigma} ((\beta + K_n(X_j^*))^\sigma - \beta^\sigma)}; \end{aligned}$$

- (3) both  $X_i$  and  $Z_i$  take on new values with probability

$$\begin{aligned} \mathbb{P}(X_i = \text{'new'}, Z_i = \text{'new'} \mid \mathcal{F}_n^{-i}) \\ \propto \int_0^{T_i} k(T_i; x) (\beta + K_n(x))^{\sigma-1} \left( \beta_0 + \frac{D}{\sigma} ((\beta + K_n(x))^\sigma - \beta^\sigma) \right)^{\sigma_0-1} \Lambda_0(dx), \end{aligned}$$

which depends on  $\mathcal{F}_n^{-i}$  only through the observed survival times  $\mathbf{T}_{1:n}$ ; in this case, the new value assigned to  $X_i$  is sampled between 0 and  $T_i$ , proportionally to

$$k(T_i; x) (\beta + K_n(x))^{\sigma-1} \left( \beta_0 + \frac{D}{\sigma} ((\beta + K_n(x))^\sigma - \beta^\sigma) \right)^{\sigma_0-1} \Lambda_0(dx).$$

Notably, cases (1) and (2) are in closed form, while the integral in case (3) can be evaluated via quadrature formulae over bounded intervals. Moreover, the potential new value of  $X_i$  can be sampled via rejection sampling. As for the acceleration step, each latent location  $X_j^*$  for  $j \in \{1, \dots, k\}$  is independently resampled between 0 and the minimum observed survival time associated with it,  $\min_{i \in C_j^X} T_i$ , proportionally to

$$\left( \prod_{i \in C_j^X} k(T_i; x) \right) (\beta + K_n(x))^{r_j \sigma - n_j} \left( \beta + \frac{D}{\sigma} \left( (\beta + K_n(x))^\sigma - \beta^\sigma \right) \right)^{\sigma_0 - r_j} \Lambda_0(dx);$$

this task is routinely performed by means of Metropolis-Hastings steps. Finally, a straightforward initialization strategy consists in populating an empty partition structure using the same sequential allocation scheme outlined above.

## S4.2 Posterior sampling of random measures (Section 6.2)

For the implementation of this posterior sampling algorithm, a pivotal role in generating the jumps' sequences is played by the upper incomplete gamma function  $\Gamma(a, x) := \int_x^\infty s^{a-1} e^{-s} ds$ , for any  $x \geq 0$  and  $a \in (-1, 0]$ . At the root of the hierarchy, an approximate realization of  $\tilde{\mu}_0^+$  in (22) is obtained as follows:

1<sub>0</sub>) the jumps  $V_1, \dots, V_k$  at fixed locations  $X_1^*, \dots, X_k^*$  are sampled from

$$V_j \stackrel{\text{ind}}{\sim} \text{Gamma} \left( r_j - \sigma_0, \beta_0 + \frac{D}{\sigma} \left( (\beta + K_n(X_j^*))^\sigma - \beta^\sigma \right) \right), \quad j = 1, \dots, k;$$

2<sub>0</sub>) the random jumps  $V_h^{(0)}$ , for  $h \geq 1$ , are sampled by solving the equations

$$\begin{aligned} \Gamma \left( -\sigma_0, V_h^{(0)} \left( \beta_0 + \frac{D}{\sigma} \left( (\beta + K_n(X_h^{(0)}))^\sigma - \beta^\sigma \right) \right) \right) \\ = \frac{N_h^{(0)} \Gamma(1 - \sigma_0)}{\Lambda_0(\mathbb{X})} \left( \beta_0 + \frac{D}{\sigma} \left( (\beta + K_n(X_h^{(0)}))^\sigma - \beta^\sigma \right) \right)^{-\sigma_0}. \end{aligned} \quad (\text{S11})$$

The truncation level  $H_0 \geq 1$  is the first value of  $h$  for which

$$N_h^{(0)} > \Lambda_0(\mathbb{X}) \frac{\Gamma(-\sigma_0, \beta_0 \varepsilon)}{\Gamma(1 - \sigma_0)} \beta_0^{\sigma_0},$$

if  $\beta_0 > 0$ , while if  $\beta_0 = 0$ , the inequality simplifies to  $N_h^{(0)} > \Lambda_0(\mathbb{X}) \varepsilon^{-\sigma_0} / (\sigma_0 \Gamma(1 - \sigma_0))$ . Similarly, at the bottom of the hierarchy, the approximate realization of  $\tilde{\mu}_\delta^+$  in (23) with  $\delta \in \{1, \dots, D\}$  specializes to:

1 <sub>$\delta$</sub> ) the jumps  $(S_{\delta j h})_{j h}$  at fixed locations  $X_1^*, \dots, X_k^*$  are sampled in groups as

$$\sum_{h=1}^{r_{\delta j}} S_{\delta j h} \sim \text{Gamma} \left( n_{\delta j} - r_{\delta j} \sigma, \beta + K_n(X_j^*) \right), \quad j = 1, \dots, k;$$

2<sub>δ</sub>) the random jumps  $S_h^{(\delta)}$ , for  $h \geq 1$ , are sampled by solving the equations

$$\Gamma\left(-\sigma, S_h^{(\delta)}\left(\beta + K_n\left(X_h^{(\delta)}\right)\right)\right) = \frac{N_h^{(\delta)} \Gamma(1 - \sigma)}{\tilde{\mu}_0^+(\mathbb{X}) \left(\beta + K_n\left(X_h^{(\delta)}\right)\right)^\sigma}, \quad (\text{S12})$$

where  $\tilde{\mu}_0^+ = \sum_{j=1}^k V_j + \sum_{h=1}^{H_0} V_h^{(0)}$  is the approximated total mass of  $\tilde{\mu}_0^+$ .

Also in this case, the truncation level  $H_\delta$  is the first value of  $h$  for which

$$N_h^{(\delta)} > \frac{\Gamma(-\sigma, \beta\varepsilon)}{\Gamma(1 - \sigma)} \beta^\sigma \tilde{\mu}_0^+(\mathbb{X}), \quad \text{if } \beta > 0, \quad N_h^{(\delta)} > \frac{\tilde{\mu}_0^+(\mathbb{X}) \varepsilon^{-\sigma}}{\sigma \Gamma(1 - \sigma)}, \quad \text{if } \beta = 0.$$

A potential bottleneck for the concrete implementation of this algorithm is the solution of the equations (S11)-(S12). A simple and effective approach consists in solving these with respect to the logarithm of the jumps, through the Newton-Raphson method. Indeed, the involved functions are strictly convex and their derivatives can be computed in closed form.

## S5 Additional material on the applications in Section 7

We provide further details, diagnostic plots, and additional figures on the illustrations with synthetic and real datasets presented in Section 7, which were omitted from the main manuscript due to space constraints.

### S5.1 A simulation study with three competing risks

The simulation scenario in Section 7.1 considers  $D = 3$  competing sources of risk: synthetic data are generated according to the latent failure times approach, whose distributions are detailed in (24). The sample size is  $n = 300$  and the number of subjects in the simulated dataset experiencing each competing event is 120, 99 and 81, respectively. Figure S1 displays the histogram of the observed survival times, overlapped with the true data-generating density function. The model hyperparameters are set at  $\beta = \beta_0 = 1.0$  and  $\sigma = \sigma_0 = 0.25$ , while  $\Lambda_0 = \theta dx$  for  $\theta > 0$ . Moreover, non-informative exponential hyperpriors are specified for the concentration parameter  $\theta$  and kernel parameter  $\gamma$ , namely  $\theta \sim \text{Exp}(0.1)$  and  $\gamma \sim \text{Exp}(0.1)$ .

The MCMC procedure is run for 25,000 iterations, with a burn-in of 5,000. Posterior samples are summarized for the number  $k$  of shared latent locations in the hierarchical model (Figure S2), the concentration parameter  $\theta$  (Figure S3), and the kernel parameter  $\gamma$  (Figure S4). The burn-in samples are highlighted in gray in traceplots, and discarded when constructing histograms. After thinning, 2,000 samples from the latent partition structure are collected and employed to compute posterior estimates with the marginal method. Figure S5 displays the traceplot (after burn-in and thinning) and corresponding autocorrelation function for the posterior estimates of the survival function, given the latent variables (Proposition 8), at time  $t = 0.4$ . The same plots are reported in Figure S6 for the posterior estimates of incidence functions (Proposition 9), again at time  $t = 0.4$ .

### S5.2 Illustration of posterior consistency

Theorem 4 in Section 3.3 investigates the Kullback-Leibler support and weak consistency of the survival density induced by the mixture hazard model (4) with hierarchical generalized gamma mixing measures, for the different kernel choices. In particular, the Dykstra-Laud kernel is weakly consistent at each density with non-decreasing hazard rate. This section provides an empirical confirmation of this consistency result by means of a simulation study. Consider the simulation setup in Section 7.1, with  $D = 3$  competing sources of risk, and generate synthetic data according to the latent failure times approach. Notably, the distributions of the latent failure times in (24) are Weibull with shape parameters larger than 1, implying a strictly

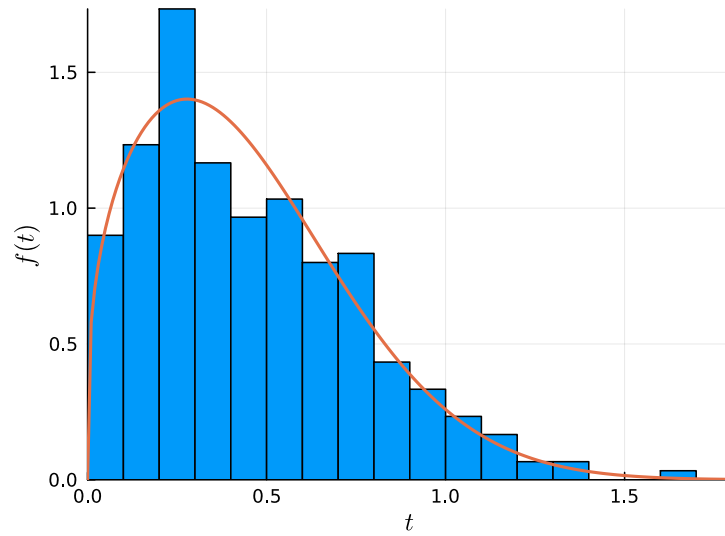


Figure S1: Histogram of the observed survival times for the synthetic dataset with  $n = 300$  data points in Section 7.1, and corresponding data-generating density function; the maximum observed survival time is 1.61.

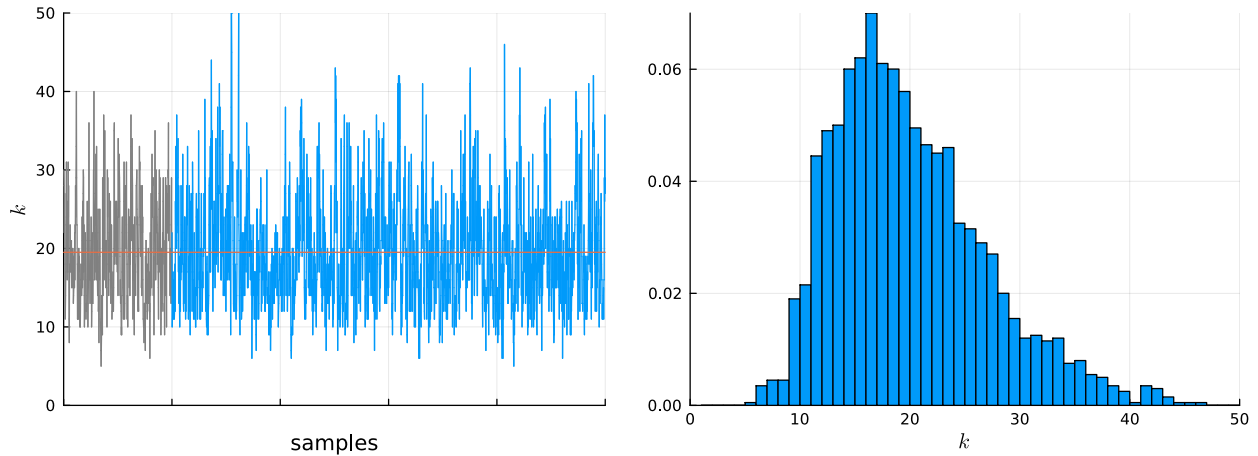


Figure S2: Posterior samples of the number  $k$  of shared latent locations for the synthetic dataset in Section 7.1; the posterior mean is 19.51.

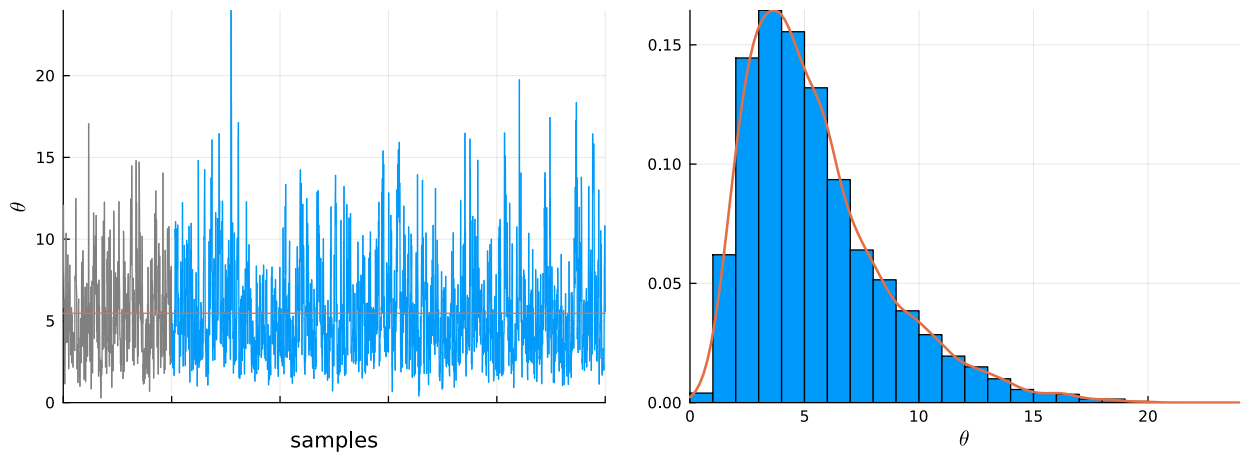


Figure S3: Posterior samples of the concentration parameter  $\theta$  for the synthetic dataset in Section 7.1; the posterior mean is 5.47.

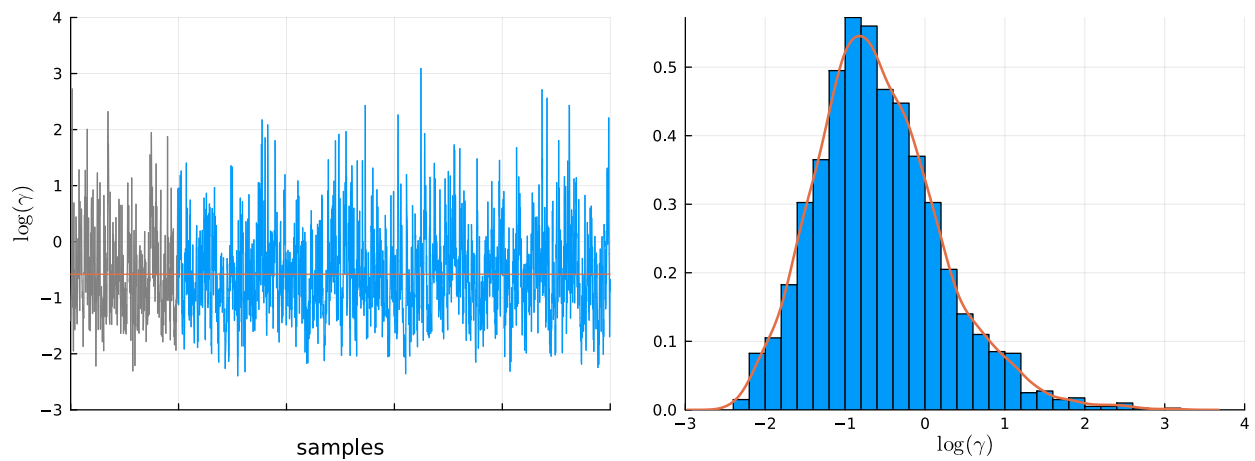


Figure S4: Posterior samples of the Dykstra-Laud kernel parameter  $\gamma$  for the synthetic dataset in Section 7.1, displayed on the log-scale; the posterior mean is 0.815, the acceptance rate of the Metropolis-Hastings step is 0.414.

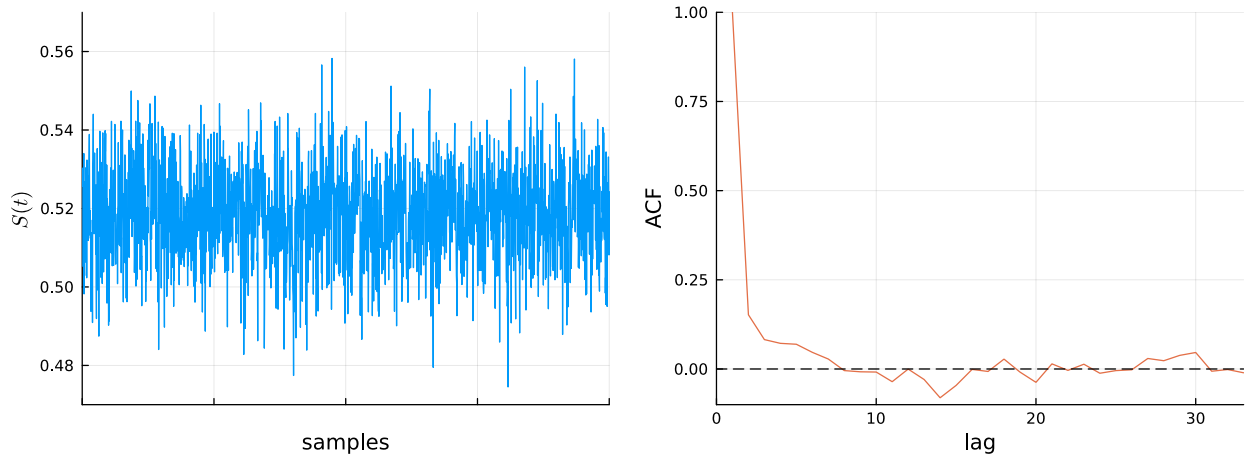


Figure S5: Posterior estimates of the survival function at time  $t = 0.4$ , given the latent variables, for the synthetic dataset in Section 7.1; the effective sample size is 1069.

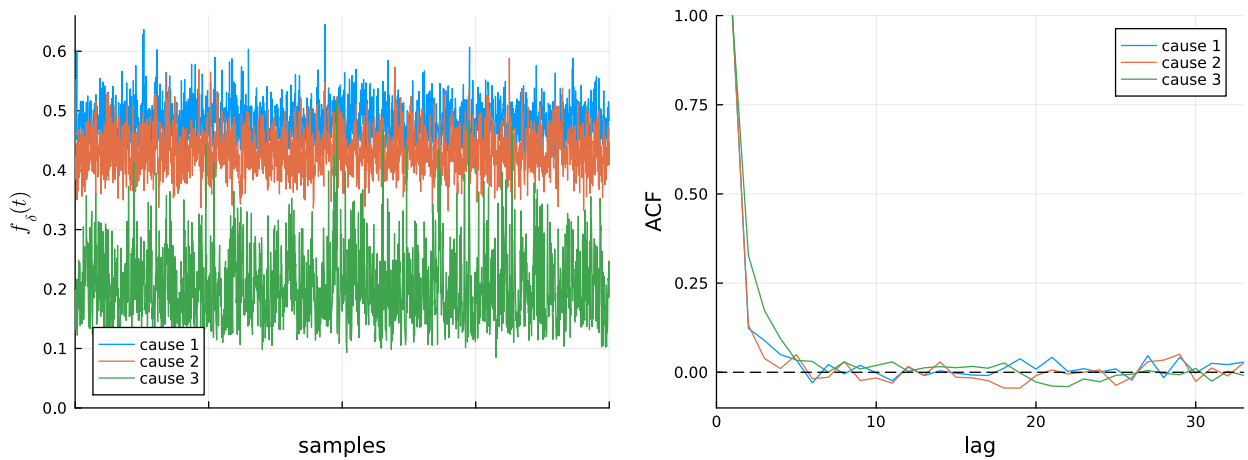


Figure S6: Posterior estimates of the incidence functions at time  $t = 0.4$ , given the latent variables, for the synthetic dataset in Section 7.1; the effective sample sizes are 1267, 1376 and 722, respectively.

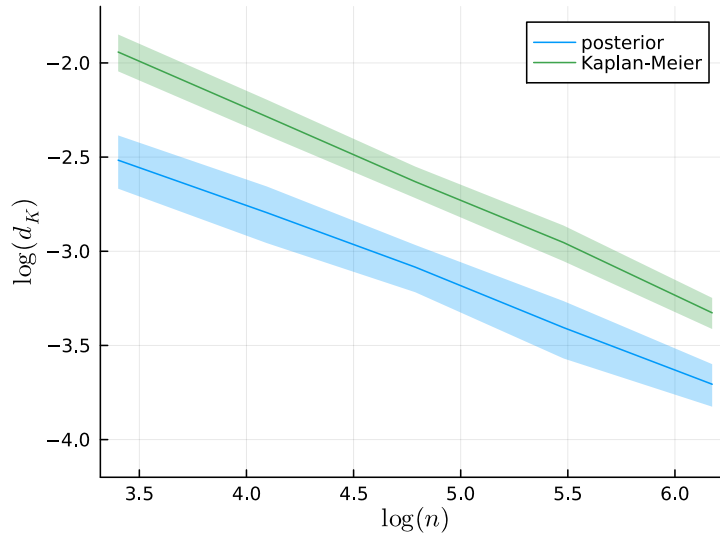


Figure S7: Average errors in estimating the survival function for the hierarchical model (posterior) and Kaplan–Meier estimate (green), in terms of Kolmogorov distances  $d_K$  from the ground truth, for different sample sizes; pointwise asymptotic 0.95 confidence bands for the mean errors are also displayed. Data are generated according to the latent failure times approach detailed in (24); results are averaged over 50 simulated datasets for each sample size.

increasing hazard rate for the survival density. The model hyperparameters and hyperpriors are specified as in Section S5.1 above. The sample sizes are  $n = 30, 60, 120, 240, 480$ , and results are averaged over 50 simulated datasets for each value of  $n$ . The burn-in period and thinning parameter are scaled linearly with  $n$  to ensure proper mixing of the MCMC procedure for each sample size. Figure S7 displays the average errors in estimating the survival function for the proposed hierarchical model and the Kaplan–Meier estimate, in terms of Kolmogorov distances from the ground truth; the posterior estimate is obtained with the marginal approach. For both methods, the estimation errors exhibit a nearly linear relationship with the sample size on the log-log scale. Although far from a comprehensive simulation study, this illustration supports the theoretical results and further confirms the effectiveness of our proposal.

### S5.3 Dependent vs. independent priors

The simulation study in Section 7.2 compares the performances of the proposed model, with hierarchically dependent hazard rates, and of a Bayesian nonparametric model with independent hazard rates. The data-generating model involves  $D = 3$  competing and independent sources of risk. For each of the 100 simulated dataset, observations are generated

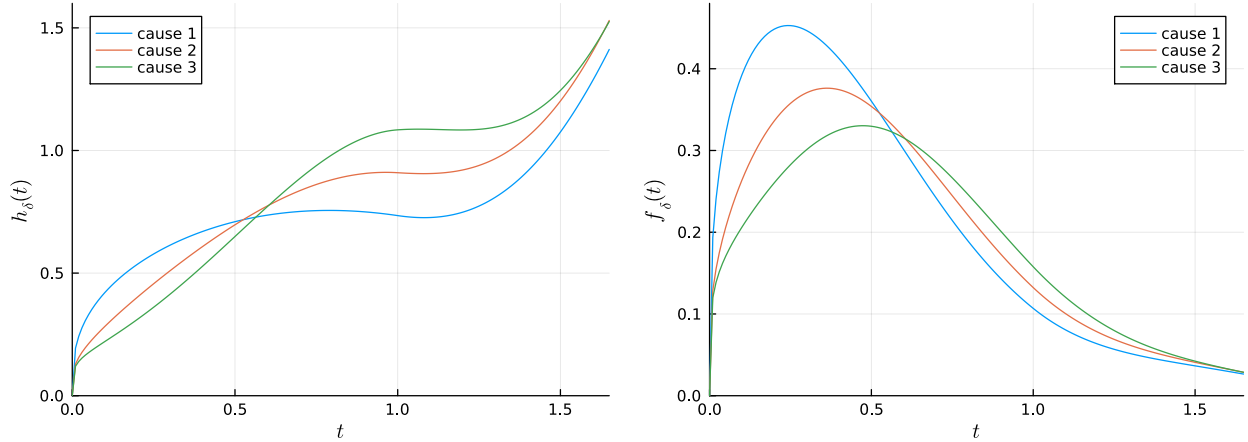


Figure S8: Cause-specific hazard rates (left) and incidence functions (right) for the  $D = 3$  competing independent sources of risks in the data-generating model for the simulation study in Section 7.2.

according to the latent failure time approach. Specifically, for each event type  $\delta \in \{1, 2, 3\}$ , the potential survival time  $Y_{i,\delta}$  is sampled from a mixture of a cause-specific and a common distribution, with equal weights. Denote by  $t \mapsto \text{Weibull}(t; \xi)$  the density of a Weibull distribution with shape  $\xi$  and unit scale parameters. The density of  $Y_{i,\delta}$  is  $g_\delta(t) = 0.5 \phi_\delta(t) + 0.5 \phi_0(t)$ , where  $\phi_0(t) = 0.5 \text{Weibull}(t; 1.2) + 0.5 \text{Weibull}(t-1; 3.0)$ , and  $\phi_\delta(t) = \text{Weibull}(t; \xi_\delta)$ , with shape parameters  $\xi_1 = 1.5$ ,  $\xi_2 = 2.0$  and  $\xi_3 = 2.5$ , respectively. The cause-specific hazard rates and incidence functions resulting from such data-generating model are displayed in Figure S8. The hazard rate functions are essentially increasing, making the Dykstra-Laud kernel an appropriate choice in this scenario. Moreover, the data-generating distributions entails a balanced allocation of the observations to the event types: the average number of subjects experiencing each competing event is 34.11, 33.21 and 32.68, respectively.

#### S5.4 Application to the bone marrow transplantation dataset

The EBMT dataset in Section 7.3 includes data for  $n = 400$  patients diagnosed with acute myeloid leukemia, who underwent a bone marrow transplantation. Our analysis involves a single binary predictor (the stem cells source), and therefore a single regression coefficient  $\eta$ , for which a non-informative centered Gaussian hyperprior  $\eta \sim \mathcal{N}(0, 100)$  is specified. Moreover, the increasing hazard rates assumption implied by the Dykstra-Laud kernel appears too restrictive, and the Ornstein-Uhlenbeck kernel is considered instead,

$$k(t; x) = \sqrt{2\kappa} \exp(-\kappa(t - x)) \mathbf{1}_{\{t \geq x\}},$$

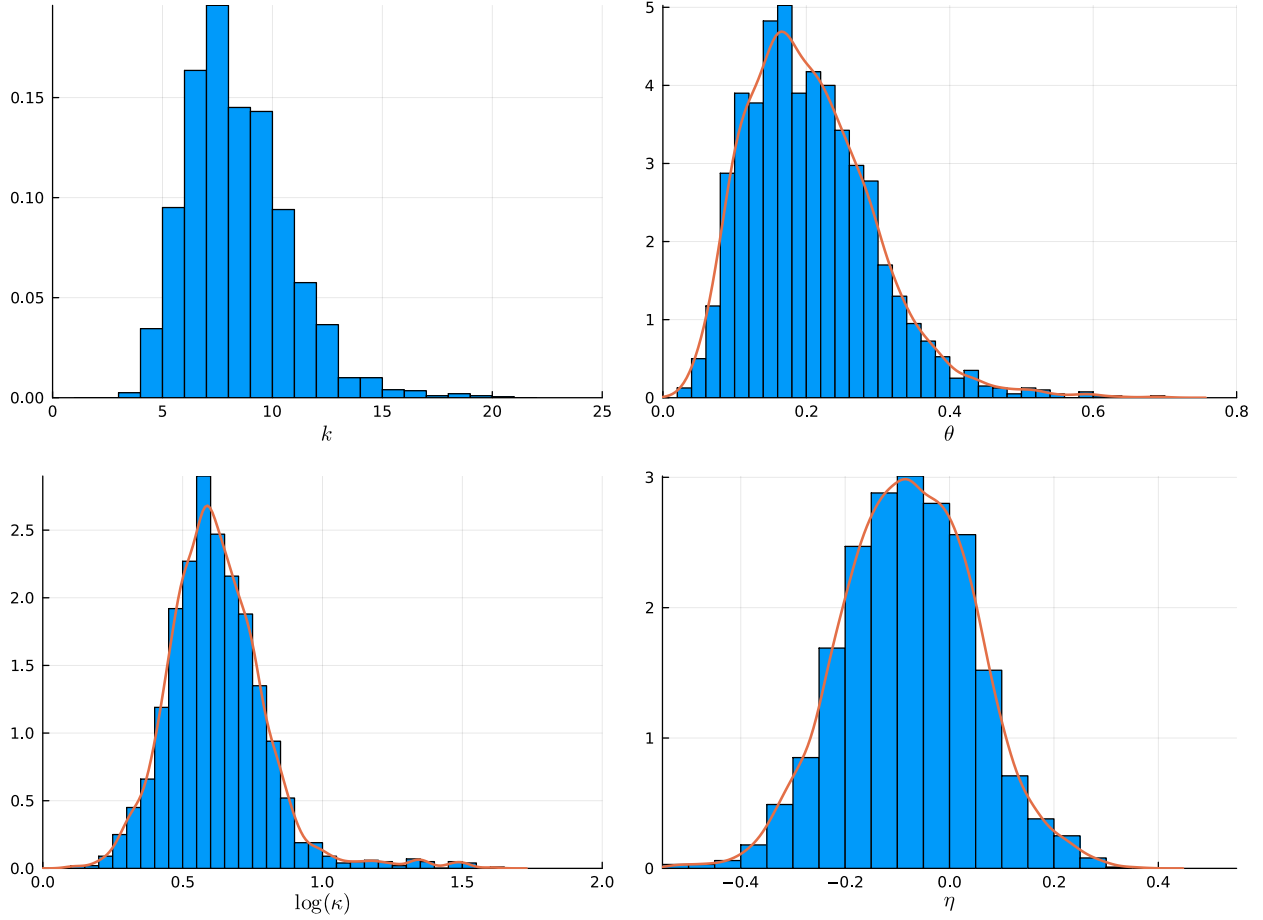


Figure S9: *EBMT* dataset. Histograms of posterior samples of relevant model parameters: number  $k$  of shared latent locations (top left), concentration parameter  $\theta$  (top right), kernel parameter  $\kappa$  on log-scale (bottom left) and regression coefficient  $\eta$  (bottom right).

where a non-informative exponential hyperprior  $\kappa \sim \text{Exp}(0.1)$  is placed on its rate parameter  $\kappa > 0$ ; with this kernel choice, the function  $K_n$  in (15) takes the form

$$K_n(x) = \sqrt{2/\kappa} \sum_{i=1}^n \max(1 - e^{-\kappa(t-x)}, 0).$$

Figure S9 reports histograms of posterior samples of the number  $k$  of shared latent locations, the concentration parameter  $\theta$ , the kernel parameter  $\kappa$  and the regression coefficient  $\eta$ .

## S6 Analysis of the melanoma dataset

The melanoma dataset was collected by [Drzewiecki et al. \(1980\)](#) at the Odense University Hospital, Denmark, and includes  $n = 205$  patients affected by stage I melanoma, who underwent surgery between 1962 and 1977. Data are publicly available as part of the `timereg` package in R. [Andersen et al. \(1993\)](#) illustrate several survival analysis methods on this dataset, while [Arfé et al. \(2019\)](#) analyzed it with their model for competing risks. The primary event is death due to melanoma (57 patients, 27.8%), with death from other causes as competing event (14 patients, 6.8%); the remaining observations are censored (65.4%). The median survival time is 5.89 years, with the maximum observed survival time for the primary event of interest being 9.14 years. The categorical predictor is gender, with 126 women (61.5%) and 79 men (38.5%). As for the EBMT dataset, we consider the Ornstein-Uhlenbeck kernel with exponential hyperprior  $\kappa \sim \text{Exp}(0.1)$  on the rate parameter.

Figure [S10](#) shows the posterior estimates of the survival function and subdistribution functions for the primary and competing event types, for both female and male patients, compared with frequentist estimators. The posterior estimate of the hazard rate ratio  $\exp(\eta)$  is 1.93 with a 0.95 credible interval of [1.19, 2.99], indicating a significant difference between male and female patients. This aligns with clinical evidence showing that male patients with melanoma typically have a higher risk ([Scoggins et al., 2006](#)). The prediction curves for both the primary event (death due to melanoma) and the competing event (death from other causes) are showcased in [Figure 1](#) of the main manuscript. Finally, [Figure S11](#) reports histograms of posterior samples of the number  $k$  of shared latent locations, the concentration parameter  $\theta$ , the kernel parameter  $\kappa$  and the regression coefficient  $\eta$ , for which a centered Gaussian hyperprior  $\eta \sim \mathcal{N}(0, 100)$  is specified.

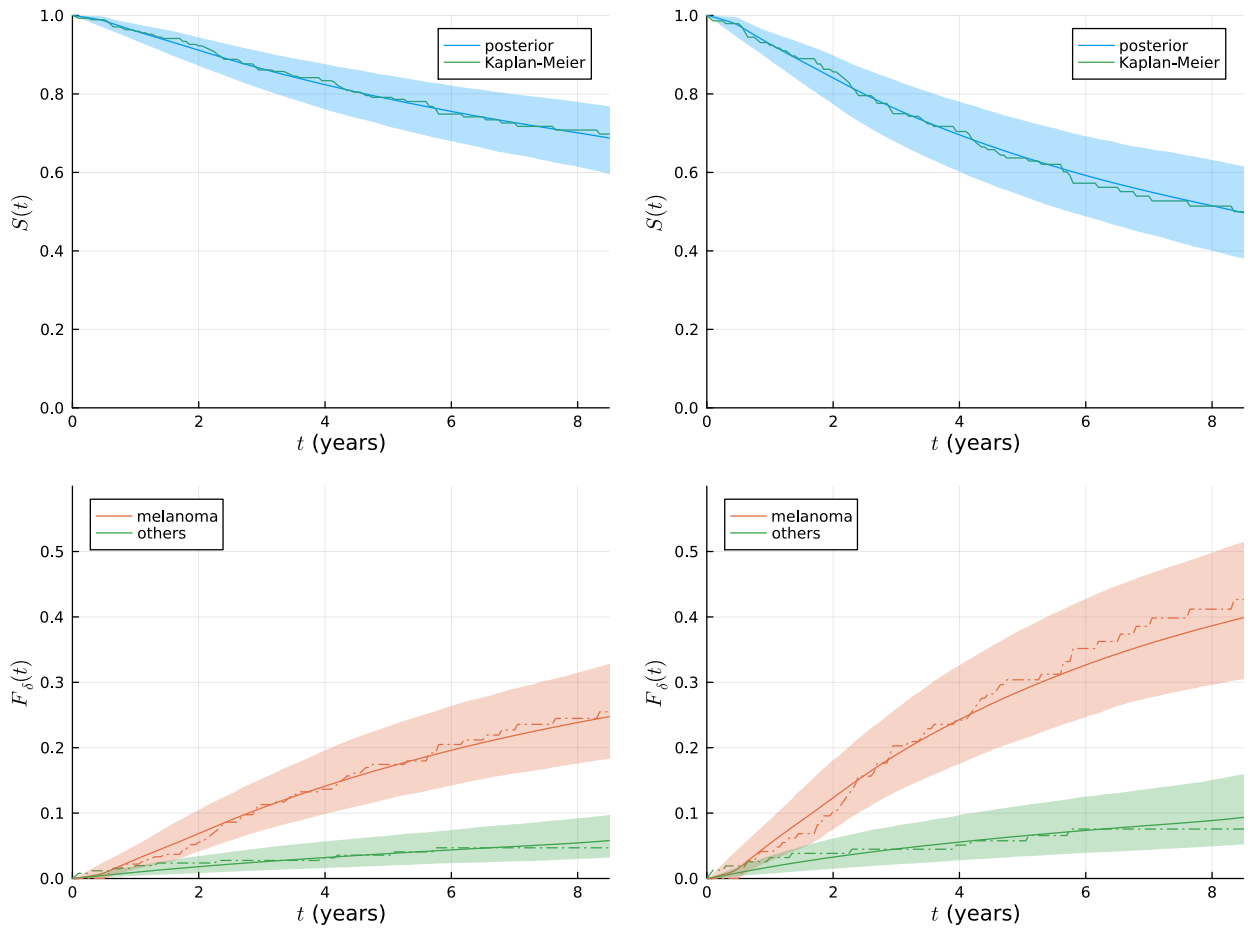


Figure S10: *Melanoma dataset*. Posterior estimates of survival function (top) and subdistribution functions (bottom) for the primary (death due to melanoma) and competing (death from other causes) event types, compared with the frequentist estimators (top: green; bottom: dash-dotted); plots refer to female (left) and male (right) patients; pointwise 0.95 credible bands are also displayed.

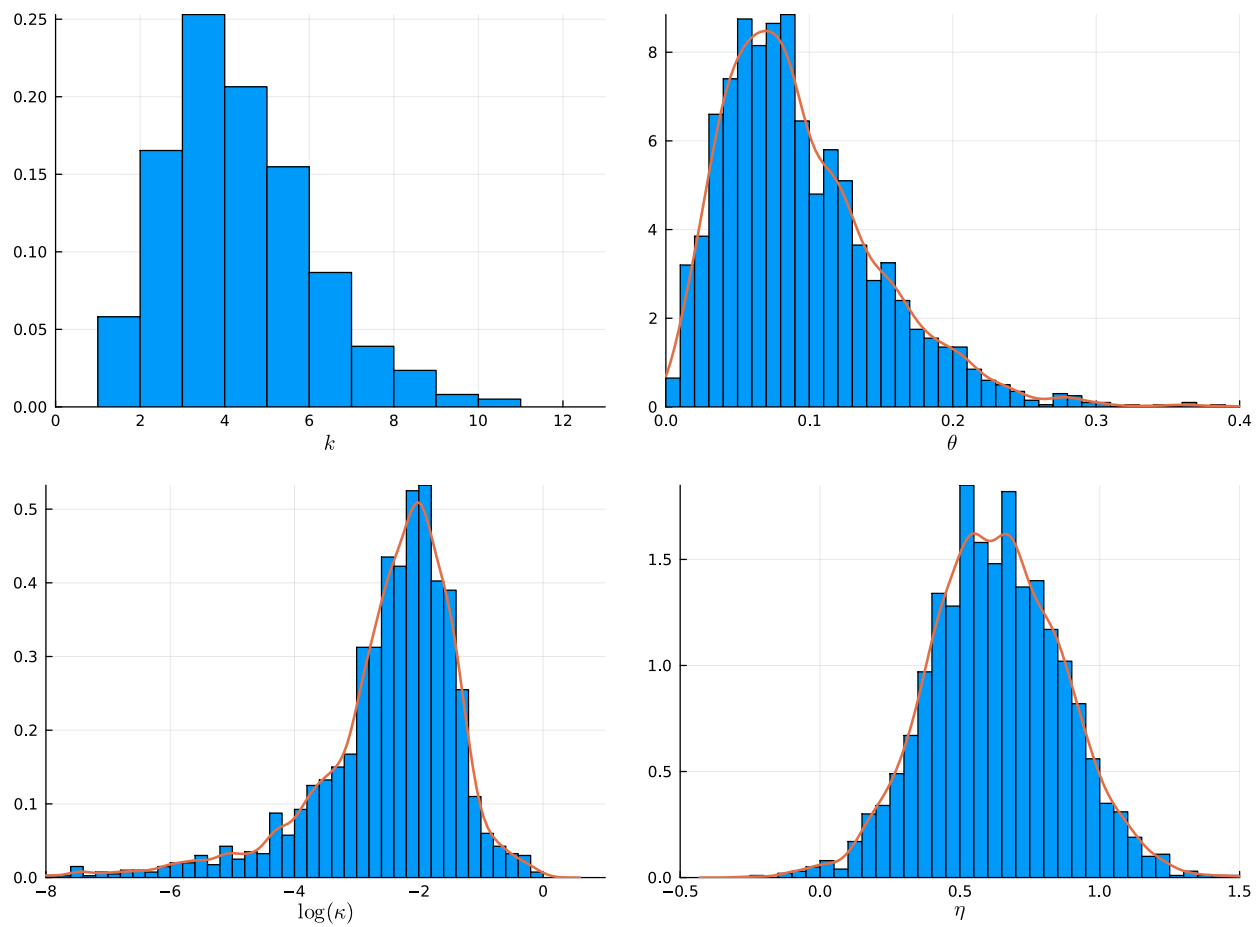


Figure S11: *Melanoma dataset*. Histograms of posterior samples of relevant model parameters: number  $k$  of shared latent locations (top left), concentration parameter  $\theta$  (top right), kernel parameter  $\kappa$  on log-scale (bottom left) and regression coefficient  $\eta$  (bottom right).

## S7 Comparison with a tree-based approach

This section presents a comparison of the method proposed in our paper, based on hCRMs, with the nonparametric approach to competing risks of [Sparapani et al. \(2020\)](#). Their method relies on Bayesian Additive Regression Trees (BART) ([Chipman et al., 2010](#)), a tree-based ensemble technique widely used for both classification and regression. Specifically, the authors adapted the binary probit BART developed in [Sparapani et al. \(2016\)](#) for discrete-time survival analysis to the competing risks scenario. A software implementation of their method is publicly available as part of the `BART` package in R.

Before presenting the main numerical results on synthetic and real data, we highlight several remarkable features that set the two modeling approaches apart. First, our model is based on kernel mixtures, and thus the prior is supported on the space of continuous cause-specific hazards. As a result, our method is naturally suited to model continuous survival data. In contrast, the tree-based approach in [Sparapani et al. \(2020\)](#) assumes discrete-time survival data, since these can be easily accommodated within the BART framework. Additionally, while our approach naturally yields estimates of cause-specific incidence functions and prediction curves, the tree-based method is limited to estimating survival and subdistribution functions. Indeed, these are the functionals included in the output of the `crisk.bart` function. Moreover, introducing the BART method, the authors consider only two competing causes of risks, namely the main cause of interest and all the others together. Accordingly, the comparison focuses on scenarios involving two competing risks. Finally, categorical predictors are incorporated via a Cox proportional hazards model (Section 7.3), which preserves interpretability at the cost of reduced flexibility. In contrast, their proposal allows for a richer dependence structure.

The comparison presented below is necessarily limited to the estimation of survival and subdistribution functions, since, as noted, the tree-based method does not cover cause-specific incidence functions and prediction curves, which are a distinctive capability of our approach. The results should be viewed as preliminary, and we refrain from drawing definitive conclusions, as a more extensive analysis is required and will be the subject of future work.

### S7.1 Illustration with synthetic data

Consider a scenario with  $D = 2$  competing sources of risk. Observations are sampled according to the latent failure times approach (Section S1), and the sample size is  $n = 100$ . Specifically, for  $i = 1, \dots, n$ ,

$$Y_{i,1} \stackrel{\text{iid}}{\sim} \text{Weibull}(\xi_1 = 1.2), \quad Y_{i,2} \stackrel{\text{iid}}{\sim} \text{Weibull}(\xi_2 = 2.4),$$

	survival	subdistribution	
		$\delta = 1$	$\delta = 2$
hCRM	<b>.0339</b>	<b>.0592</b>	<b>.0818</b>
BART	.0686	.1048	.1252
frequentist	.0744	.0938	.1354

Table S1: Comparison of rescaled total variation distances between estimated and true survival and subdistribution functions, for the synthetic dataset in Section S7.1, using the hCRM method, the BART method and the frequentist approach.

where  $\xi_1$  and  $\xi_2$  are shape parameters. Following the illustration in Section 7.1, our method relies on the generalized gamma hCRM prior with Dykstra-Laud kernel; the hyperparameters specification, as well as the MCMC settings, coincide with those adopted in Section S5. For the BART method, we consider the default parameters in the `crisk.bart` function; the survival times are rounded at the second decimal digit. Figure S12 displays the posterior estimates of the survival and subdistribution functions, along with their pointwise 0.95 credible bands, obtained with the two alternative methods; the frequentist estimates are also reported. Notably, the estimated curves produced with the hCRM approach are smoother than those obtained with the BART method, which instead nearly overlap with the frequentist estimates. A smooth estimate is clearly preferable in this scenario, and more generally in any context where the underlying data-generating mechanism is known to be continuous. On the other hand, the uncertainty around the estimates is comparable across the two methods, which yield very similar credible bands. Table S1 reports the estimation errors for the survival and subdistribution functions under the hCRM method (using the marginal algorithm), the BART method, and the frequentist approach. The errors are computed as described in Section 7.2. In this simulation scenario, our method clearly outperforms the tree-based approach due to its smoother estimates.

## S7.2 Application to clinical datasets

The comparison with the tree-based method of Sparapani et al. (2020) is also conducted on the two clinical datasets considered in this paper: the EBMT dataset (Section 7.3) and the melanoma dataset (Section S6). We refer the reader to those sections for details on our modeling choices and the specification of hyperparameters. For the BART model, we use the default parameters with survival times rounded to the nearest week. Figures S13 and S14 display the posterior estimates of the survival and subdistribution functions, along

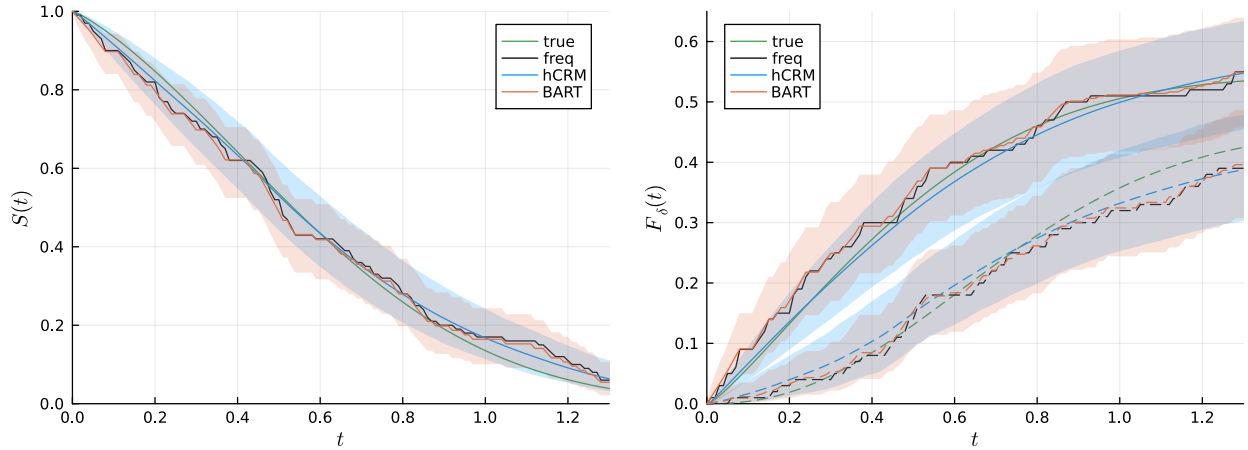


Figure S12: Posterior estimates of the survival function (left) and subdistribution functions (right), for the synthetic dataset in Section S7.1; comparison of the proposed method (hCRM, blue), the tree-based method (BART, orange), and the frequentist estimates (black) with the true curves (green); the pointwise 0.95 credible bands are also displayed.

with their pointwise 0.95 credible bands, obtained using the two alternative methods across different levels of the binary predictors; frequentist estimates are also reported for comparison. The considerations made above for the synthetic dataset continue to hold in the case of the melanoma data: our method yields smoother curves, whereas the BART approach tends to more closely follow the frequentist estimates. Instead, no clear or consistent pattern emerges for the EBMT dataset. The posterior estimates from the BART method appear relatively smooth for larger survival times, but occasionally show noticeable departures from the frequentist estimates. As for the credible bands, they are generally comparable across methods, with a remarkable exception for the subdistribution functions for the EBMT dataset: the credible bands obtained via the BART method are substantially wider.

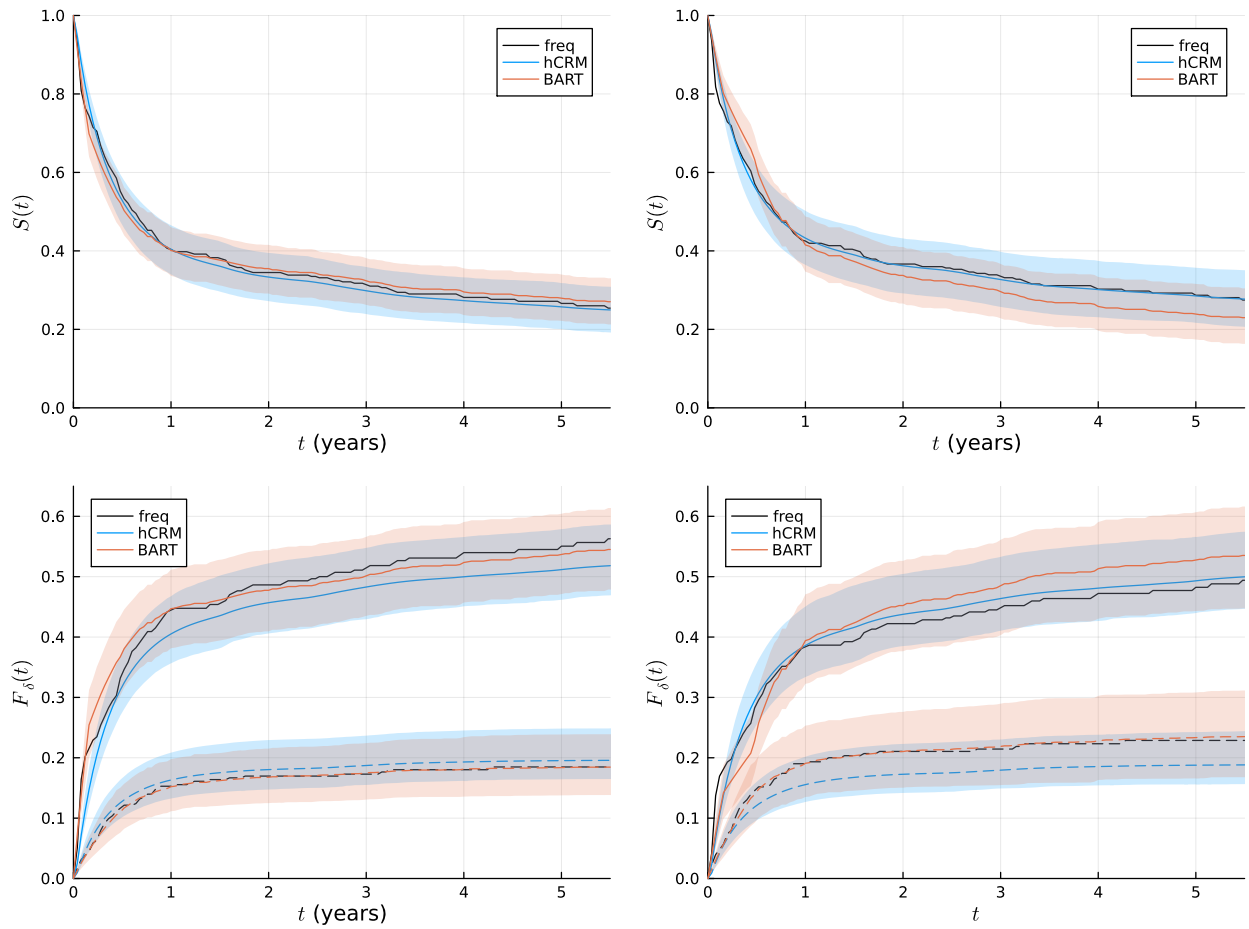


Figure S13: *EBMT dataset*. Posterior estimates of the survival function (top) and subdistribution functions (bottom), for patients who experienced graft from bone marrow (left) and peripheral blood (right) cells; comparison of the proposed method (hCRM, blue), the tree-based method (BART, orange), and the frequentist estimates (black); the pointwise 0.95 credible bands are also displayed.

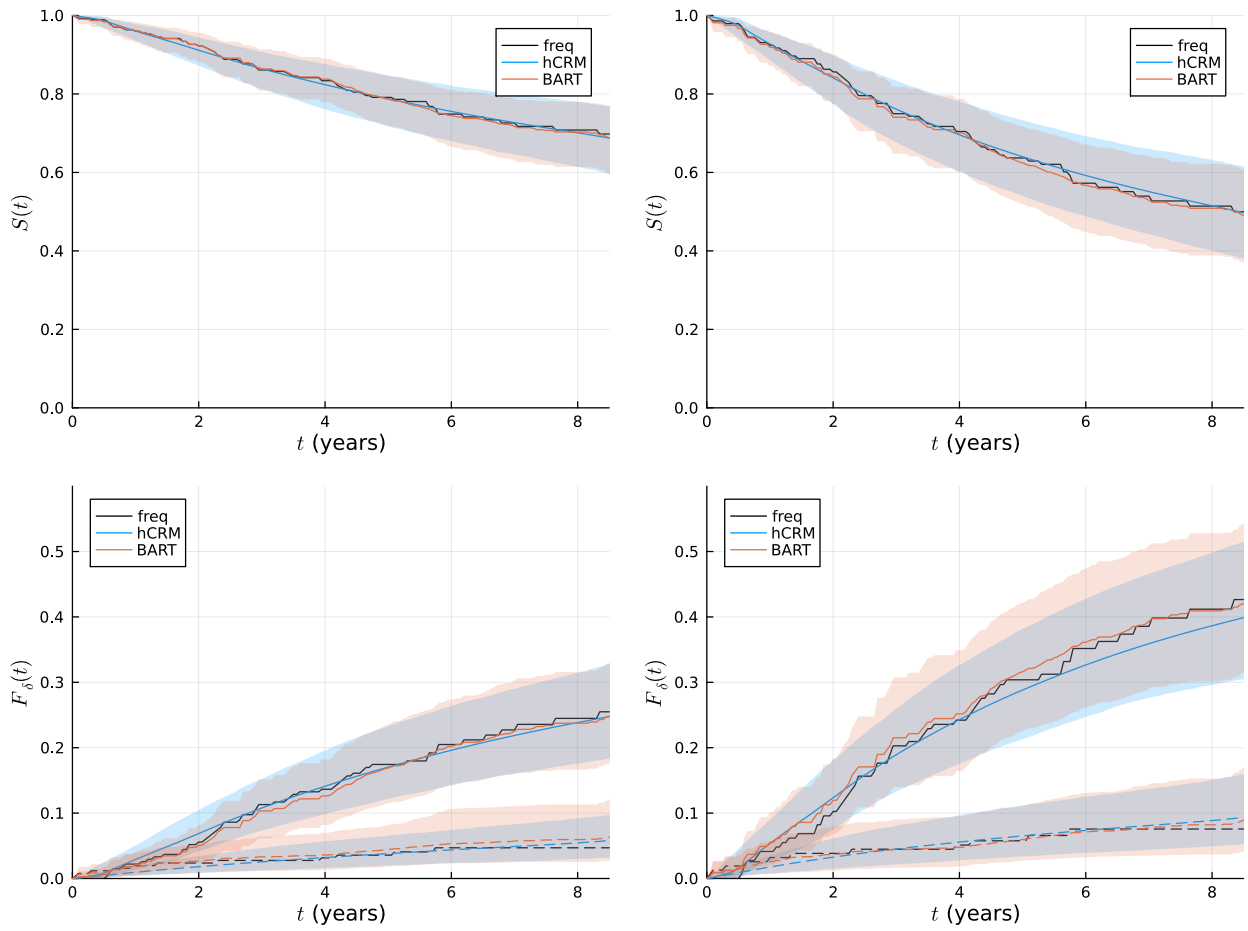


Figure S14: *Melanoma dataset*. Posterior estimates of the survival function (top) and subdistribution functions (bottom), for female (left) and male (right) patients; comparison of the proposed method (hCRM, blue), the tree-based method (BART, orange), and the frequentist estimates (black); the pointwise 0.95 credible bands are also displayed.