Recurrent events analysis with piece-wise exponential additive mixed models

Jordache Ramjith,[1] Andreas Bender,[2] Kit C. B. Roes[1] and Marianne A. Jonker[1]

[1]Department for Health Evidence, Biostatistics Research Group, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, Netherlands

[2]Department of Statistics, LMU Munich, Munich, Germany

Abstract: Recurrent events analysis plays an important role in many applications, including the study of chronic diseases or recurrence of infections. Historically, many models for recurrent events have been variants of the Cox model. In this article we introduce and describe the application of the piece-wise exponential Additive Mixed Model (PAMM) for recurrent events analysis and illustrate how PAMMs can be used to flexibly model the dependencies in recurrent events data. Simulations confirm that PAMMs provide unbiased estimates as well as equivalence to the Cox model when proportional hazards are assumed. Applications to recurrence of staphylococcus aureus and malaria in children illustrate the estimation of seasonality, bivariate non-linear effects, multiple timescales and relaxation of the proportional hazards assumption via time-varying effects. The R package `pammtools` is extended to facilitate estimation and visualization of PAMMs for recurrent events data.

Key words: Flexible survival models, non-linear survival analysis, piece-wise exponential models, recurrent events, time-varying effects

1 Introduction

A recurrent events setting in survival analysis is defined by repeated observations of an event of interest over the course of the observation period. In medical, clinical and biological research, such data play an important role, for example, in the context of chronic illnesses and infectious diseases. Some concrete examples include recurrence of hospitalization for cardiovascular events (Gibson et al., 2017; Varma et al., 2021), recurrent diseases and infections like COVID-19 (Dos Santos et al., 2021), pneumonia (Ramjith et al., 2021; Barakat et al., 2021), pneumococcus (Hernstadt et al., 2020), staphylococcus aureus (cf. Section 5.1) (Akinboyo et al., 2018; Abdulgader et al., 2019), asthma (Chan et al., 2021; Feng et al., 2021) and malaria (cf. Section 5.2) (Lawpoolsri et al., 2019; Ghosh et al., 2020; Kakuru et al., 2020).

Address for correspondence: Jordache Ramjith, Department for Health Evidence, Biostatistics Research Group, Radboud Institute for Health Sciences, Radboud University Medical Center, Geert Grooteplein 21, 6525 EZ, Nijmegen, Netherlands.
E-mail: jordache.uct@gmail.com; jordache.ramjith@radboudumc.nl

Typical for this kind of data are the correlations, or more generally dependencies, between event recurrences. Such dependencies may be induced by an individual-specific unmeasured frailty, common across all events of the individual, or by dependence of the occurrence of an event on an individual's history. In the latter case the hazard of an event could depend on the timing (e.g., time since last event) or the frequency (number of previous events). Ignoring either type of dependency, if present, leads to biased estimates or overoptimistic standard errors (Box-Steffensmeier and De Boef, 2006; Jahn-Eimermacher, 2008). To choose an appropriate model for the analysis of recurrent event data, assumptions on this dependency structure must be made as well as a choice on the timescale for the analysis of events: gap time (also referred to as waiting time, clock-reset or renewal approach) or calendar time (clock-forward). See Section 2.1 for more discussion about the timescale.

Several recurrent events models for calendar and gap times have been suggested in the literature. These models attempt to account for the effects of within-subject correlation either by adjusting the variances of the parameter estimators (the variance-adjustment models (Andersen and Gill, 1982; Prentice et al., 1981)) or by including a subject-specific random effect in the model—the shared frailty model (SFM). Variance-adjustment models assume independence among events when estimating the effect of the covariates, while adjusting the standard errors of the parameter estimates for the correlation between recurrent events by calculating robust variance through sandwich estimation (Huber, 1967; Freedman, 2006). SFMs (Hougaard, 2012) account for correlated events within an individual (and thus heterogeneity in the sample) by adding individual specific random effects, which act in a multiplicative way on the hazard function. In this work, we focus on the frailty (or random effects) model. In a standard SFM it is assumed that the hazards are proportional over time (conditional on the random effect) and the baseline hazard function is unspecified (in the Cox SFM), but a parametric form may be chosen as well. For interpretation of the estimated effects, an estimate of the baseline hazard function is important. Putter et al. (2005) mentioned that not knowing the baseline implies that the risk of observing an event at particular points in time remains unknown. The importance of knowing the absolute risk (or baseline hazard) when interpreting relative risks is further highlighted by Noordzij et al. (2017). They suggest that readers tend to overestimate or over-interpret relative risks when information about the absolute risks is missing. This is even strengthened if the hazard ratio is expected to be time-dependent. An understanding of the baseline hazard over time is important for clearer interpretation of time-varying effects (Ruhe, 2018).

Many publications on recurrent events analysis focus on variants of the Cox proportional hazards (CPH) model (Cox, 1972). This includes the SFM, mentioned earlier, and the variance-adjustment models, specifically the Andersen-Gill (AG) model, the Prentice-Williams-Peterson (PWP) models, and the Wei, Lin and Weissfeld (WLW) model (Andersen and Gill, 1982; Prentice et al., 1981; Wei et al., 1989). The AG model estimates a common baseline hazard for all events and covariate effects are usually estimated across events as well. In the PWP and WLW models the baseline hazards are stratified for the different events, and covariate estimates are either estimated across events or event-specific. The PWP models are constructed in either a calendar time approach or a gap time approach (see Section 2.1), while the WLW model is constructed using marginal times (where the time at risk for each event is counted from the study start). In the AG and PWP models, individuals only remain at risk for the event following their last observed event, while for the WLW model, individuals remain at risk for the maximum number of observed events. Studies (Kelly and Lim, 2000; Therneau and Hamilton, 1997) have shown a carry-over effect when using the WLW

model and recommend that it can be used for modelling multiple event types rather than recurrent events. There are several tutorials available with applications and comparisons of these methods (Kelly and Lim, 2000; Therneau and Hamilton, 1997; Amorim and Cai, 2015; Ullah et al., 2014), and a recent tutorial on SFMs is provided by Balan and Putter (2020).

Parametric SFMs can be a useful alternative to the semi-parametric SFM. In the parametric SFM the baseline hazards can be modeled through a choice of many parametric distributions each with its own characteristics (Munda et al., 2012). Misspecification of the baseline hazards can lead to bias in estimated effects and possibly misleading results for the shape of the hazard function (Hutton and Monaghan, 2002; Kwong and Hutton, 2003; Li et al., 1996). In a recurrent events setting where the shape of the baseline hazard differs, for example between first and recurrent events, a mixture of parametric models must be used to appropriately capture the baseline hazards (Duchateau et al., 2003). This can be a complicated task.

More flexible models like the generalized survival models (GSMs) (Liu et al., 2017) are becoming increasingly more popular in modelling the baseline hazards. These models use a smoothing splines approach and allow more complexities such as time-dependent effects, nonlinear effects, and their interactions. The GSM has recently been extended to model correlated survival data: SFMs with time-constant covariates for gap time formulations or an AG variance-adjustment type model for calendar time formulations. However, GSMs still needs to be extended to SFMs in a calendar time framework.

Here we introduce the Piece-wise exponential Additive Mixed Model (PAMM) (Bender et al., 2018; Argyropoulos and Unruh, 2015) for modelling possibly time-varying effects, nonlinear effects and more complex interactions in a recurrent events setting where it is possible to consider multiple time scales, stratified baseline hazards and time-varying covariates. As the Cox model, PAMMs model the hazard rate without parametric assumptions about the distribution of event times and multiplicative covariate effects. Thus, if model specification is equivalent, both models lead to similar results and interpretation.

While conceptually similar to the Cox model, PAMMs offer an alternative method for survival analysis that has recently gained popularity. This is mainly due to the fact that PAMMs reformulate many different survival tasks to regression tasks via a data augmentation step and availability of software that facilitates practical application (Bender and Scheipl, 2018). While this allows the model to be estimated by any algorithm that can optimize the Poisson likelihood (Bender et al., 2021), the Generalized Additive Mixed Model (GAMM) framework has proven to be particularly useful, due to its flexibility, availability of robust and established estimation techniques (e.g., Hastie and Tibshirani (1986); Wood (2017)) and interpretability of covariate effects. This allows great versatility in modelling choices, as discussed in more detail in Section 6. In particular, PAMMs facilitate seamless estimation of multivariate, non-linear, time-varying effects via penalized splines with different basis functions, cyclic splines and splines with shape constraints (Pya and Wood, 2015). Recent examples include spline based stratified baseline hazard estimation in plant biology (Panel et al., 2020), cumulative effects in critical care (Bender et al., 2019), treatment effects with non-proportional hazards in patients with coronary stent restenosis (Giacoppo et al., 2020), estimation of bivariate effect surfaces in the context of randomized controlled trials (Argyropoulos and Unruh, 2015) and progression-free survival in patients with gastroenteropancreatic neuro-endocrine tumors (Carmona-Bayonas et al., 2019), and time-varying hazard ratios in post-transplant outcome assessment (Wey et al., 2020). An application to the recurrence of childhood pneumonia was recently presented in Ramjith et al. (2021). In general, however, detailed instructions on the application of PAMMs to survival tasks,

especially tasks that go beyond the single-event setting with right-censoring as well as software tools that facilitate such analyses, particularly recurrent events analyses, are scarce.

In this article we give an introduction to PAMMs for recurrent events analysis. This includes data transformation specific to recurrent events and discussion on the choice of timescale. We show that the model performs well and generates unique insights by means of simulation studies and real data examples. The article is accompanied by the R package `pammtools` that was extended to facilitate application of PAMMs to recurrent events data. The `pammtools` package takes care of the required data transformation and greatly facilitates model estimation, visualization and interpretation of the estimated effects.

In the remainder of this work, we summarize hazard based model specification and estimation for recurrent events analysis in Section 2. In Section 3, we show how such models can be estimated using PAMMs, including a step by step guide on the necessary data augmentation. A comparison of the SFM and the PAMM when proportional hazards are assumed (conditional on the frailty) is made through a simulation study in Section 4. Section 5 concludes with examples of PAMMs applied to different recurrent events data sets and illustrates effect estimation of different complexities, that is, stratified baseline hazards, non-linear, cyclic effects of seasonality and multiple timescales. We conclude with a discussion and an outlook in Section 6.

2   Models for recurrent events

2.1   Timescale

The choice between the two timescales (gap time or calendar time) is usually driven by the research setting in which the data are collected, and the research question at hand, as the choice of the timescale affects the interpretation of the results (see Duchateau et al., 2003; for a thorough discussion). On both timescales, the study start time needs to be clearly defined and meaningful, for example, time from birth, time from randomization, time from diagnosis. Calendar time is appropriate when the interest is in the full course of the recurrent event process. On this timescale an individual's time starts when entering the study and stops when leaving the study. For the first event, the individual's contribution to time at risk is counted from entering the study. However, the individual's contribution to time at risk for the second event is only counted from the first event, the start time of the risk interval for the second event. A calendar time data set can be viewed as a succession of (internally) left-truncated data sets for each of the events, as a subject's time at risk is left-truncated at the event time of the previous event.

In the gap time approach one is interested in the time between events. Essentially, that means that after every event 'the clock' is reset; after each event the time since the previous event is used to define time intervals that an individual is at risk for an event. Thus, in contrast to the calendar time approach, a subject is at risk for all of his events at time $t = 0$, however, time to event is redefined as time from last event to next event. Historically, the gap time approach facilitates analysis, because the data could be evaluated with standard techniques for survival analysis without the need to take into account left-truncation. Assuming a full renewal process, all observations are considered independent. Dependence induced by number of previous events can be modeled by stratifying the hazards by event number. Dependence introduced by within-subject correlation can be modeled by introducing a frailty term.

In the PAMM framework, the choice of timescale is less important, as we can always add additional time-dependent covariates to describe the past. In calendar time, for example, we could include a variable for the gap time of the previous event or the number of previous events. In the gap time approach, we can include dependence on the total time since entering the study, and so on. More generally, PAMMs support dependence on multiple timescales (Iacobelli and Carstensen, 2013), as any dependence on time is modeled through covariate effects of different representations of time.

2.2 Model specification

In the following we consider models for the (conditional) hazard function for the $k-$ th event for subject i given by

$$\lambda_{i,k}(t|z_i) := \lambda(t|\mathbf{x}_i(t), z_i, k) = \lambda_{0,k}(t) \exp(g(\mathbf{x}_i(t), t, k) + z_i), \qquad i = 1, \ldots, n, \tag{2.1}$$

where, t is the time of interest, $\mathbf{x}_i(t) = (x_{i,1}(t), x_{i,2}(t), \ldots x_{i,p}(t))^\top$ a vector of potentially time-varying covariates, z_i a subject specific unobserved random effect and $g(\mathbf{x}_i(t), t, k)$ is a general function of potentially non-linear, time-varying and event-specific covariate effects, not to be confused with a transformation function of the covariates. Further, $\lambda_{0,k}(t)$ are event-specific baseline hazards. In the CPH model and its extensions, baseline hazards are not specified and estimated non-parametrically. Standard models discussed in the literature follow from 2.1 by relaxing some of the dependencies. For example, with linear effects and under the proportional hazards assumption (conditional on frailty) we get the stratified SFM (Box-Steffensmeier and De Boef, 2006):

$$\lambda(t|\mathbf{x}_i, z_i, k) = \lambda_{0,k}(t) \exp(\mathbf{x}_i^\top \boldsymbol{\beta} + z_i), \qquad i = 1, \ldots, n.$$

Assuming a common baseline hazard for all events, the model could be further simplified by replacing $\lambda_{0,k}$ with λ_0, resulting in the so-called unstratified (or unconditional on episode order) SFM (Hougaard, 2012).

The conditional likelihood for model 2.1 in calendar time (given the individual frailty terms) is proportional to

$$\mathcal{L}(t|z) = \prod_{i=1}^{n} \mathcal{L}_i(t_{i,1}, \ldots, t_{i,K_i}|z_i) \propto \prod_{i=1}^{n} \prod_{k=1}^{K_i} \lambda_{i,k}(t_{i,k}|z_i)^{\delta_{i,k}} \exp\left(- \int_{t_{i,k-1}}^{t_{i,k}} \lambda_{i,k}(s|z_i) \mathrm{d}s \right), \tag{2.2}$$

where K_i is the number of events for which subject i was at risk, $t_{i,k}$ the event time for event k for $k < K_i$, t_{i,K_i} is the time subject i was censored, and $t_{i,0}$ the entering time for subject i which is usually equal to 0. Furthermore, $\delta_{i,k}$ is the event specific status indicator, so $\delta_{i,k} = 1$ for $k < K_i$ and $\delta_{i,K_i} = 0$ (see also Duchateau et al., 2003). The likelihood in (2.2) is often given in the simplified form

$$\mathcal{L}(t|z) \propto \exp\left(- \sum_{i=1}^{n} \int_0^{t_{i,K_i}} \lambda_{i,k}(s|z_i) \mathrm{d}s \right) \prod_{i=1}^{n} \prod_{k=1}^{K_i-1} \lambda_{i,k}(t_{i,k}|z_i), \tag{2.3}$$

The likelihood functions given so far are conditional the frailty variable. The marginal (unconditional) likelihood function is proportional to

$$\mathcal{L}(t) \propto \prod_{i=1}^{n} \left(\int_{z_i} \prod_{k=1}^{K_i} \lambda_{i,k}(t_{i,k}|z_i)^{\delta_{i,k}} \exp\left(-\int_{t_{i,k-1}}^{t_{i,k}} \lambda_{i,k}(s|z_i) \mathrm{d}s \right) f_Z(z_i) \mathrm{d}z_i \right), \qquad (2.4)$$

where f_Z is the density function of the frailty variable.

The likelihood for the gap time model (conditional on frailty) is equivalent to 2.2 with $t_{i,k}$ replaced by $d_{i,k} = t_{i,k} - t_{i,k-1}$ and the lower integration limit set to 0:

$$\mathcal{L}(t|z) = \prod_{i=1}^{n} \mathcal{L}_i(t_{i,1}, ..., t_{i,K_i}|z_i) \propto \prod_{i=1}^{n} \prod_{k=1}^{K_i} \lambda_{i,k}(d_{i,k}|z_i)^{\delta_{i,k}} \exp\left(-\int_0^{d_{i,k}} \lambda_{i,k}(s|z_i) \mathrm{d}s \right). \qquad (2.5)$$

3 PAMMs for recurrent events analysis

3.1 Data augmentation and estimation

As described in detail elsewhere (Bender et al., 2018), application of PAMMs to time-to-event data requires a particular data augmentation step. We refer to the resulting data as piece-wise exponential data (PED). In principle, the data augmentation required for recurrent events analysis is equivalent, however, some specifics need to be taken into account, that also depend on the timescale of the analysis.

As for the single-event PED, the follow-up has first to be discretized into J intervals with cut-points $0 = \tau_0 < \tau_1 < \ldots < \tau_J$. The jth interval is defined as $(\tau_{j-1}, \tau_j]$ and τ_J as the maximum time, which can be set arbitrarily. In practice, the maximum observed event time is a typical choice, since no information about the event process is contained in the data beyond that point. The other cut-points could be set to the unique observed event times. In the gap time analysis, cutpoints could also be set arbitrarily or at a subset of unique event times. In the calendar time approach, cut-points must be set at event times. Note that despite the discretization of the follow up, the exact time-to-event information is used for estimation, thus the PAMM is a method for continuous time-to-event data. Technical details about the data-transformation are provided below. A more intuitive illustration of the data transformation procedure is given in Section 3.2.

Let $\{\tau_j\}_{j=1}^J$ be the set of interval borders that partition the follow-up as before. We define the event-, interval- and subject-specific status and 'time at risk' variables as

$$\delta_{ijk} = \begin{cases} 1 \text{ if } t_{i,k} \in (\tau_{j-1}, \tau_j] \text{ and } \delta_{i,k} = 1 \\ 0 \text{ else} \end{cases} \quad \text{and} \quad t_{ijk} = \begin{cases} t_{i,k} - \tau_{j-1} \text{ if } t_{i,k} \in (\tau_{j-1}, \tau_j) \\ \tau_j - \tau_{j-1} \text{ else} \end{cases},$$

where $i = 1, \ldots, n$ the subject identifier as before, $j = 1, \ldots, J$ denotes the interval and $k = 1, \ldots, K$ the event number. Thus, δ_{ijk} indicates whether subject i, experienced an event of type k in interval j (1=yes, 0=no) and t_{ijk} is the time for which subject i was at risk for the kth event in interval j. Note that for a particular subject i we only calculate these variables for time points (intervals) at which they were at risk for the particular event number. Doing so also naturally

accounts for left-truncation at the beginning of the follow up as well as left-truncation for recurrent events on the calendar timescale (induced by the recurrent events process).

Consider a specific event k. Assuming piece-wise constant hazards

$$\lambda_{i,k}(t|\mathbf{x}_i(t), z_i, k) = \lambda_{j,k} \exp(g(\mathbf{x}_i(t), t, k) + z_i) = \lambda_{ijk} \qquad \forall t \in (\tau_{j-1}, \tau_j] \tag{3.1}$$

in each interval, the likelihood contribution of subject i (conditional on the frailty variable) in 2.2 can be rewritten as

$$\lambda_{i,k}(t_{i,k}|z_i)^{\delta_{i,k}} \exp\left(-\int_{t_{i,k-1}}^{t_{i,k}} \lambda_{i,k}(s|z_i)\mathrm{d}s\right) = \left(\prod_{j \in \mathcal{J}_{ik}} \lambda_{ijk}^{\delta_{ijk}}\right) \exp\left(-\sum_{j \in \mathcal{J}_{ik}} \lambda_{ijk} t_{ijk}\right), \tag{3.2}$$

where $\mathcal{J}_{ik}$ is the set of intervals for which i is at risk for event k. Here, left-truncation can be accounted for by defining $\mathcal{J}_{ik}$ such that it excludes intervals prior to the left-truncation time. The right-hand side of 3.2 can be recognized to be proportional to likelihood contributions to a Poisson likelihood under the working assumption $\delta_{ijk} \overset{iid}{\sim} Po(\mu_{ijk} = \lambda_{ijk} t_{ijk})$ and conditional on the frailties z_i. Making use of this, it can be seen that the conditional likelihood function $\mathcal{L}(t|z)$ is proportional to the conditional likelihood function that is found under the working assumption just mentioned:

$$\mathcal{L}(t|z) \propto \mathcal{L}_{Po}(t|z) = \prod_{k=1}^{K} \prod_{i=1}^{n} \left(\prod_{j \in \mathcal{J}_{ik}} \frac{\mu_{ijk}^{\delta_{ijk}}}{\delta_{ijk}!}\right) \exp\left(-\sum_{j \in \mathcal{J}_{ik}} \lambda_{ijk} t_{ijk}\right). \tag{3.3}$$

Similar to 2.4, the conditional likelihood in 3.3 can be expressed as a marginal likelihood by specifying a frailty distribution:

$$\mathcal{L}_{Po}(t) = \prod_{i=1}^{n} \int_{z_i} \prod_{k=1}^{K} \left(\prod_{j \in \mathcal{J}_{ik}} \frac{\mu_{ijk}^{\delta_{ijk}}}{\delta_{ijk}!}\right) \exp\left(-\sum_{j \in \mathcal{J}_{ik}} \lambda_{ijk} t_{ijk}\right) f_Z(z_i)\mathrm{d}z_i, \tag{3.4}$$

where $\mu_{ijk} = \lambda_{ijk} t_{ijk}$, $\lambda_{ijk} = \lambda_{i,k}(t|\mathbf{x}_i(t), z_i, k) \ \forall t \in (\tau_{j-1}, \tau_j]$, t_{ijk} is the time for which subject i was at risk for the kth event in interval j and δ_{ijk} indicates whether subject i, experienced an event of type k in interval j. Thus, parameters of model 2.1 (i.e., the hazard) can be estimated by fitting a mixed-effects Poisson regression model to outcomes δ_{ijk} and offset $o_{ijk} = \log(t_{ijk})$ with subject-specific random intercepts. In a generalized linear mixed-effects modelling framework, the integral can be approximated using numerical techniques such as adaptive Gauss-Hermite quadrature (Bates et al., 2015). In the GAMM framework, the integral is either numerically approximated or the link between penalized splines and random effects models (Lin and Zhang, 1999; Wand, 2003; Wood, 2004, 2017) is used to treat the frailties as penalized splines where the frailty variance is estimated as part of the penalty.

The reformulation of recurrent events models into Poisson regression models is interesting especially due to flexibility of GAMMs and readily available software. In PAMMs we parameterize the interval- and event-specific hazard rates 3.1 via (non-linear) covariate effects. An exemplary model

specification is given below:

$$\lambda\left(t \mid \mathbf{x}_i(t), z_i, k\right) = \exp\left(\beta_{0,k} + f_{0,k}(t_j) + \sum_{p=1}^{P} f_p(x_{i,p}(t_j), t_j) + z_i\right) \qquad \forall\, t \in (\tau_{j-1}, \tau_j]. \qquad (3.5)$$

In 3.5, $\beta_{0,k} + f_{0,k}(t_j)$ denotes the event-specific baseline log-hazard rate and t_j is simply a covariate that represents time (e.g. $t_j = \tau_j$) that is, used to make the hazard in interval j depend on time, and $f_p(x_{i,p}(t_j), t_j)$ are covariate effects, that might be linear or non-linear and time-dependent. Non-linear functions are represented through basis functions and coefficients, for example, $f_{0,k}(s) = \sum_{\ell=1}^{L} \gamma_{\ell,k} b_{\ell,k}(s)$ where $\gamma_{l,k}$ are unknown regression parameters to be estimated and $b_{\ell,k}(s)$ are known basis functions, with several types of basis functions available (Wood, 2017). Popular choices for basis functions include B-Splines (Eilers and Marx, 1996), cubic regression splines and others. The concrete choice of basis functions, however, is usually not important. The number of basis functions L (or equivalently maximum degree of freedom) are also not important, as long as it is large enough to allow enough flexibility. Overfitting is avoided by enforcing a penalty on the coefficient estimates, that penalizes large deviations between (neighboring) coefficients, leading to a smooth function. Both of these aspects are illustrated in Supplement. As discussed elsewhere (Bender et al., 2018), penalized splines based estimation of the baseline hazard is particularly useful in this context, as it makes the model robust to the choice of interval cutpoints, enforces similarity of neighboring baseline hazards and reduces the number of parameters to estimate ($L \ll J$). Depending on the assumptions about the data generating process, 3.5 could be simplified, for example, by dropping the dependence on event number or by removing the frailty term.

From an estimation and inference perspective, the model is a Poisson regression task with additive and random effects. Therefore, the parameters of the model can be estimated by optimizing the penalized Poisson likelihood 3.6 using the GAMM framework (e.g., Wood, 2017).

$$\mathcal{L}_{Po}(t|z) + \sum_{m=1}^{M} v_m \boldsymbol{\gamma}_m^\top \mathbf{S}_m \boldsymbol{\gamma}_m. \qquad (3.6)$$

In 3.6, M is the total number of smooth and random effects in the model and v_m the respective penalty parameter estimated from the data; $\boldsymbol{\gamma}_m$ is the coefficient vector associated with the m-th smooth or random effect and $\mathbf{S}_m$ the corresponding penalty matrix that will depend on the type of basis functions chosen for the m-th model term. Gaussian random effects can be subsumed in Equation (3.6) because of the equivalence of smooths and random effects (see for example, Wood, 2017, Sec. 5.8 and references therein). For example, if the m-th term is a random effect, $\boldsymbol{\gamma}_m = (\gamma_{1,m}, \ldots, \gamma_{n,m})^\top$ represents the random intercepts, where n is the number of subjects in the data set. The corresponding basis functions in this case are 0/1 coded dummy variables for each subject (i.e. n dummy coded columns in the design matrix) and the penalty matrix $\mathbf{S}_m$ for these terms is an $n \times n$ identity matrix. The tuning parameter v_m controls the shrinkage of the random effects term and the variance of the random effects is the inverse of v_m.

Joint optimization of all parameters in 3.6 can be done in different ways. In the application section, we use the methods implemented in package mgcv, which provides different methods for

Table 1 Recurrent event data in 'standard' format, where each individual i has one row per event k he was at risk.

i	entry ($t_{i,k-1}$)	exit ($t_{i,k}$)	gap ($d_{i,k}$)	status ($\delta_{i,k}$)	event (k)
1	0	0.5	0.5	1	1
1	0.5	3.0	2.5	0	2
2	0	0.8	0.8	1	1
2	0.8	1.2	0.4	1	2
2	1.2	3.0	1.8	0	3

optimization, including generalized cross validation, maximum likelihood and restricted maximum likelihood (see, Wood, 2011 for a comparison).

3.2 An example of recurrent events data in PED format

Exemplary recurrent time-to-event data for two individuals, $i \in \{1, 2\}$ both with a 3 months follow-up is given in Table 1 and will be used to illustrate the data augmentation step discussed in the previous section. Individual $i = 1$ has the first event at 0.5 months since the start of the study and does not experience a second event for the remainder of the follow-up. Individual $i = 2$ experiences the first event at 0.8 months since the start and the second event at 1.2 months since the start (or 0.4 months since the first event) and does not experience the third event in the remainder of the follow-up. As seen in Table 1, the events that each individual is at risk for are captured in different rows. Columns 'entry' and 'exit' define the time span (calendar time) in which the subjects are at risk for event k.

3.2.1 Gap time

In Table 1 the unique event-specific gap times, ordered with respect to length, are given by 0.4, 0.5, 0.8 (values from column $d_{i,k}$ in Table 1 with $\delta_{ik} = 1$). The maximum gap time is 2.5, however, it usually makes sense to set interval border of the last interval J to the maximum event time, which is 0.8 in this example as there is no information about event times beyond that point. Therefore we will use cutpoints (0,0.4,0.5,0.8) and the intervals in this example will be defined as

$$\{(\tau_{j-1}, \tau_j]\}_{j=1}^{J} = \{(0, 0.4], (0.4, 0.5], (0.5, 0.8]\}.$$

The transformation of Table 1 to the PEM data format based on these intervals is shown in Table 2.

3.2.2 Calendar time

On the calendar timescale we consider the ordered time points at which an event took place (or end of study) 0.5, 0.8, 1.2. Once again, we do not consider time points beyond the last observed event time. Next, we split the follow-up into intervals using these time points as cutpoints:

$$\{(\tau_{j-1}, \tau_j]\}_{j=1}^{J} = \{(0, 0.5], (0.5, 0.8], (0.8, 1.2]\}.$$

The resulting data in the piece-wise exponential format is given in Table 3.

Table 2 Recurrent event data in PEM data format on the gap timescale.

i	gap (t_{ik})	k	j	$(\tau_{j-1}, \tau_j]$	δ_{ijk}	t_{ijk}	$o_{ijk} = \log{(t_{ijk})}$
1	0.5	1	1	(0,0.4]	0	0.4	log(0.4)
1	0.5	1	2	(0.4,0.5]	1	0.1	log(0.1)
1	2.5	2	1	(0,0.4]	0	0.4	log(0.4)
1	2.5	2	2	(0.4,0.5]	0	0.1	log(0.1)
1	2.5	2	3	(0.5,0.8]	0	0.3	log(0.3)
2	0.8	1	1	(0,0.4]	0	0.4	log(0.4)
2	0.8	1	2	(0.4,0.5]	0	0.1	log(0.1)
2	0.8	1	3	(0.5,0.8]	1	0.3	log(0.3)
2	0.4	2	1	(0,0.4]	1	0.4	log(0.4)
2	1.8	3	1	(0,0.4]	0	0.4	log(0.4)
2	1.8	3	2	(0.4,0.5]	0	0.1	log(0.1)
2	1.8	3	3	(0.5,0.8]	0	0.3	log(0.3)

Table 3 Recurrent event data in PEM data format on the calendar timescale.

i	t_{ik} (cal)	k	j	$(\tau_{j-1}, \tau_j]$	δ_{ijk}	t_{ijk}	$o_{ijk} = \log{(t_{ijk})}$
1	0.5	1	1	(0,0.5]	1	0.5	log(0.5)
1	3.0	2	2	(0.5,0.8]	0	0.3	log(0.3)
1	3.0	2	3	(0.8,1.2]	0	0.4	log(0.4)
2	0.8	1	1	(0,0.5]	0	0.5	log(0.5)
2	0.8	1	2	(0.5,0.8]	1	0.3	log(0.3)
2	1.2	2	3	(0.8,1.2]	1	0.4	log(0.4)

4   Simulation studies

4.1   Comparison of PAMM and SFM in a proportional hazards setting

Under the assumption of proportional hazards the proposed PAMM and the SFM should give similar estimates of the regression parameters and the frailty variance. In order to check whether this is true an extensive simulation study was performed. The simulation settings as well as the simulation results are given in Supplement 1. In summary, it is shown that under the proportional hazards assumption the PAMM and the SFM give (almost) identical results for the estimated regression parameters in both gap- and calendar timescales, using the stratified and/or unstratified models. Furthermore, both models tend to estimate the regression parameters and the frailty variance more accurately when there is a longer follow-up time or there are more recurrences. It must be noted that when the frailty variances are large both models underestimate the frailty variance, but this underestimation is lower for longer follow-up times. Specifically, in the gap time scenarios, we see that the underestimation is worse in the SFM, and in the calendar time scenarios the underestimation is worse in the PAMM. The consequences of these biases of the estimated frailty variance on the estimates of the regression coefficients are minimal. It should be noted that the varying results for the frailty variances could depend on model implementation, and not necessarily on the model formulation, where `mgcv` uses restricted maximum likelihood to maximize the penalized log likelihood and `coxme` uses a Laplace approximation to calculate an integrated penalized likelihood.

4.2   Comparison of PAMM and SFM in distinguishing frailty and time-varying effects

Balan and Putter (2019) showed in simulation studies that in certain scenarios of clustered survival data – including recurrent events, it is difficult to distinguish the presence of a frailty effect from the presence of a time-varying effect. We replicated their simulation study for recurrent events to compare how the PAMM and the SFM distinguishes between a frailty effect and a time-varying effect. The only difference is that we used log-normal frailties for the Cox-like SFM, since the implementation of PAMM using mgcv is limited to Gaussian distributed random effects only. The results are shown in Supplement 4. In summary, the results indicate that the PAMM distinguishes no worse (and in some cases better) between a frailty effect and a time-varying effect than the Cox-like SFM with Gaussian distributed random effects for at least two events per subject.

5   Motivating examples

In this section we apply the PAMM for recurrent events to two real-life data sets using R (R Core Team, 2020; RStudio Team, 2020). We use the R packages pammtools (Bender and Scheipl, 2018) for the data transformation and the visualization of the results and mgcv (Wood, 2017, 2003, 2004, 2011; Wood et al., 2016) for model estimation. The aim of this section is to show the application of several types of recurrent events models using the PAMM framework.

5.1   Example 1: Staphylococcus aureus infection data

The staphylococcus aureus (SA) infection data in Abdulgader et al. (2019) contains times at which 137 children were colonized with a new staphylococcus aureus infection in the nasopharynx in their first year of life in Paarl, South Africa. In Abdulgader et al. (2019), the gap time PWP model (Prentice et al., 1981) was used to model the survival curves for the different infections (i.e., first, second, third, etc.). A standard CPH model was also used to investigate the effects of risk factors on the time to the first infection. In this section, we aim to repeat their analysis but include recurrent infections using the stratified PAMM (see equation (3.5)) in gap time. We consider a different baseline hazard for the first event and for recurrent events (combined). An advantage of using the PAMM here is that we can easily model, estimate and visualize the baseline hazards stratified for the first and recurrent events. We use the HIV exposure variable, where some children are HIV exposed and uninfected (HEU) and the rest are HIV unexposed (HU). First we look at the baseline hazards and then we look at a proportional hazards model for HIV exposure where the effect of HIV exposure is different for first and recurrent events.

5.1.1   Baseline hazards

We first fit the model without the HIV covariate but including the frailty term (i.e., $\lambda(t) = \lambda_{0,k}(t) \exp(z_i)$) to have an understanding of the functional form of the baseline hazard. The estimated frailty variance is 0.066 ($p = 0.249$). This is a small frailty variance and so we will exclude the frailty term and reduce our model to a simpler model.
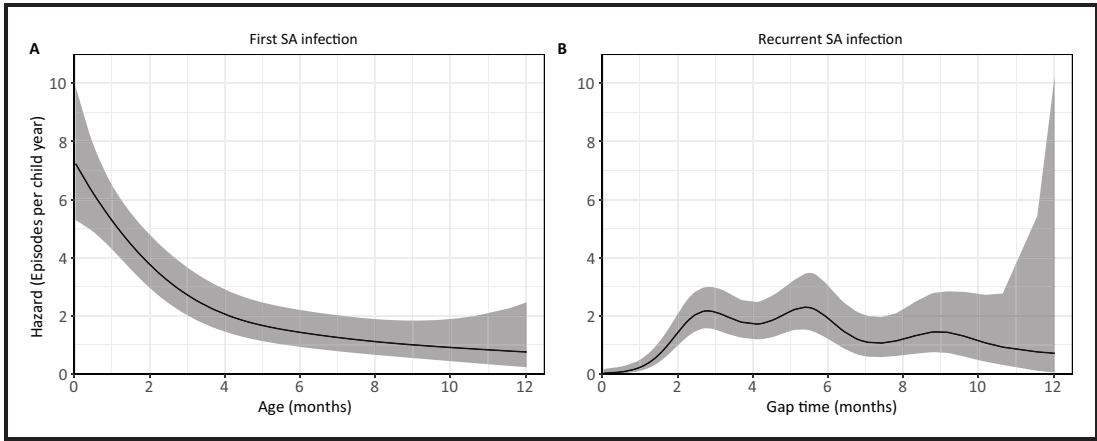
Figure 1 The hazard rates for (**A**) first and (**B**) recurrent SA infections respectively. 95% confidence intervals are shaded in gray.

The estimated degrees of freedom, explained in Supplement 2, for the evolution of the hazard rate over time is different for first and recurrent SA infection, indicating that the evolution of the baseline hazards for first and recurrent events may be different (note: we are not testing for a difference in the baseline hazards). We visualize the estimated baseline hazard in Figure 1, where the hazard rates are re-scaled to correspond with the number of events per child-year. For the first SA event, the hazard rate is highest after birth and decreases quickly until around 5 months, thereafter a very slow decline is seen. For recurrent SA infections, we see an increase in the hazard over time since the previous infection until about three months. Thereafter we see an almost constant hazard for a recurrent infection and a slow decline starting from around 9 months.

5.1.2   PH model

The hazard function of the model we are interested in, is

$$\lambda(t) = \lambda_{0,k}(t) \exp\left(\beta_1 \times \mathrm{HIV}\right)$$

where $\lambda_{0,k}(t) = \exp(\beta_{0,k} + f_{0,k}(t))$ and t is the time in the gap timescale. The case $k = 1$ indicates the first SA infection, and $k = 2$ indicates recurrent SA infection. The results indicate that the HEU children had a higher hazard of SA infection over time than HU children ($HR = \exp(\beta_1) = 1.31$, 95% CI : $0.98 - 1.74$).

In the paper by Abdulgader et al. (2019), the univariable analysis of HIV exposure on the time to first SA infection showed a HR effect close to 1. Next, we model an interaction of HIV exposure and the recurrence indicator to allow a separate HIV exposure effect for first and recurrent events, that is,

$$\lambda(t) = \lambda_{0,k}(t) \times \exp\left(\beta_{1,k} \times \mathrm{HIV}\right) \tag{5.1}$$

where $\lambda_{0,k}(t)$, k and t are as defined above. Here we find that the HR for HIV exposure for the first SA infection is 0.98 (95% CI : $0.64 - 1.50$)) (similar to what was reported in their paper) and for recurrent SA infections is 1.72 (95% CI : $1.16 - 2.54$)).

5.2 Example 2: Childhood malaria data

To illustrate application of PAMMs on calendar timescale, we reanalyze data from a study of child-hood malaria undertaken by (Kakuru et al., 2020) in Uganda to study the effects of two intermittent preventive treatments, with Sulfadoxine-pyrimethamine (SP) and Dihydroartemisininpiperaquine (DP) on the incidence of childhood malaria from birth until children are one year old. Therefore, the timescale for events and recurrences coincides with the age of the children in this example. In the original analysis, treatment effect was investigated with respect to the time to first episode of malaria, and it was found that there may be larger treatment differences between boys but not between girls. Here we analyze time to first event and additionally recurrences and add preterm birth and gravidity (number of previous pregnancies) as additional risk factors in the model. We further illustrate incorporation of cyclic splines and bivariate smooth functions by incorporating a seasonal effect and later its interaction with child age. An advantage using PAMMs, similar to Example 5.1, is that the baseline hazards are flexibly estimated, while simultaneously estimating a non-linear effect of seasonality, all in a SFM setting. Furthermore, it can easily be investigated, through more complex interactions, if there is any indication of the non-linear seasonality effect changing over time). Additionally in Supplement 3 we show the effects of different choices of basis functions for the baseline hazards, as well as degrees of freedom.

5.2.1 Model specification
We fit the following model

$$
\begin{aligned}
\log(\lambda(t|\mathbf{x}_i, z_i)) = \beta_0 \; &+ f_0(t_j) + f(\mathrm{doy}_i) \\
&+ \beta_1 \mathrm{DP}_i + \beta_2 \mathrm{female}_i + \beta_3 \mathrm{DP}_i \cdot \mathrm{female}_i \\
&+ \beta_4 \mathrm{preterm}_i + \beta_5 \mathrm{gravidity}_i + z_i
\end{aligned}
\tag{5.2}
$$

where $\beta_0 + f_0(t_j)$ is the log-baseline hazard as defined in Section 3 and Equation 3.5 (here equivalent to the smooth effect of child age). We model the effect of seasonality via variable 'day of year' (doy) and set $f(\mathrm{doy}) = \sum_{m=1}^{10} \gamma_m b_m(doy)$, where b_m are cyclic cubic regression spline bases that are defined such that the function value is the same at the beginning and end of the year (see Wood, 2017, Section 5.3.2).

The results show a significant frailty variance ($\hat{\sigma}^2 = 0.69^2 = 0.48$, $p < 0.001$), which indicates that the frailty term is needed to account for the correlation between events in a child's longitudinal profile.

The results of the analysis show a significant treatment effect for boys, where DP is associated with a lower hazard of malaria over time compared with SP ($HR = \exp(\hat{\beta}_1)$ (95% CI): 0.74 (0.58, 0.95)), but not for girls ($HR = \exp(\hat{\beta}_1 + \hat{\beta}_3)$ (95% CI): 0.98 (0.78, 1.24)). The model also shows a significant effect of gravidity (HR (95% CI): 1.05 (1.005, 1.09)) indicating that children whose mothers had one more prior pregnancy had an approximately 5% higher hazard over time.
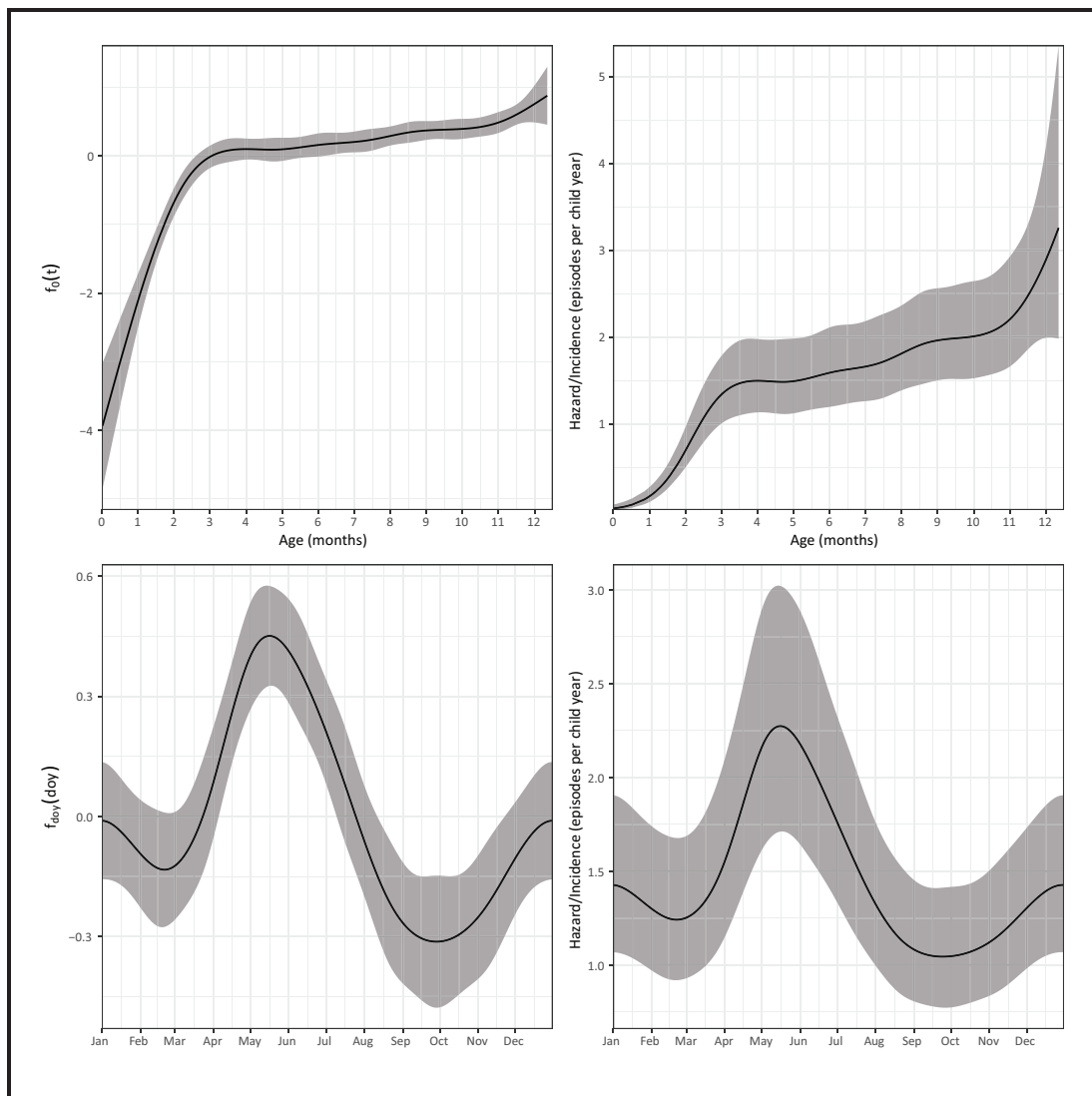
Figure 2 Upper row: The non-constant part of the log-baseline ($f_0(t)$) including confidence intervals (left panel) and respective hazard/incidence rates (episodes per child year, other covariate held constant; right panel) across age. Bottom row: Effect of seasonality as contribution to the log-hazard (f(doy), left panel) and hazard/incidence rates (episodes per child year, given child of age 100 days, other covariates as before; right panel).

The estimated non-linear effects of age and seasonality are depicted in Figure 2. The upper row shows the effect of the child's age as log-hazard contribution of the non-constant baseline term $f_0(t_j)$ (left panel) and as hazard/incidence rates (episodes per child year) (right panel). The hazard rates were calculated by using varying time values while holding all other covariate values constant.

Specifically, doy=1, treatment=DP, sex=Female, preterm=no and gravidity=1. From both graphs, we can see that the hazard of contracting malaria rises quickly as children get older until approximately 3 months of age, where the effects starts to level off. The bottom row of Figure 2 depicts the seasonality of malaria infection, where once again, the left panel depicts the log-hazard contribution of the term $f(\text{doy})$ and the right panel the hazard/incidence rates (episodes per child year, for children of varying levels of seasonality, 100 days of age, other covariates fixed as before). They show that the hazard rates go up from the start of the rainy seasons (January) and start to go down from the start of the dry seasons, with the second rainy season of the year (around June) showing a larger effect on the hazard of malaria. Per construction, the effect is the same at the end of December as in the beginning of January.

5.2.2 Modelling age-dependent effects of seasonality:

As children get older, they might be left unattended for longer periods of time and potentially spent more time outside or otherwise unprotected (e.g., by nets). We therefore additionally investigate a potential interaction between child age and seasonality when it comes to the hazard of malaria infection, for example such that seasonality does not affect the hazard for very young children as much as for older children. Since both variables are continuous, we model the interaction via bivariate penalized splines represented by tensor products, such that Equation 5.2 becomes

$$\log(\lambda(t|\mathbf{x}_i, z_i)) = \beta_0 \; + f_0(t_j, \text{doy}_i) + \beta_1 \text{DP}_i + \beta_2 \text{sex}_i + \beta_3 \text{DP}_i \cdot \text{sex}_i$$
$$+ \beta_4 \text{preterm}_i + \beta_5 \text{gravidity}_i + z_i \qquad (5.3)$$

Note that in Equation 5.3 the non-linear part of the baseline hazard is now a function of both, age (t_j) and day of year (doy). This term can thus be viewed as a time-varying effect of the day of year, as it depends on t_j. In this case, however, it can also be viewed as an example for the application of multiple timescales, as the hazard depends on age (time since origin) and the day of year (calendar time).

The estimated bivariate effect $f_0(t_j, \text{doy})$ is depicted in Figure 3 (as before, similar representations could be obtained for the hazard/incidence rate where the other covariates are fixed at selected values). The left panel shows the surface of the bivariate function, with brighter colours indicating larger hazards. The middle panel shows vertical slices through the surface for fixed values of age, while the right panel shows horizontal slices through the surface for fixed values of day of year (season). These results are in line with the results obtained from model 5.2 and in this case indicate, that there is no substantial interaction between child age and seasonality effect. This conclusion could also be obtained more formally, using a ANOVA type decomposition of the bivariate effect into main and interaction effects (Wood, 2006), as illustrated in Suppplement.

6 Discussion

In this article we introduced PAMMs as an alternative modelling approach for recurrent events data. These models utilize the generalized additive mixed modelling framework for estimation and inference. This facilitates specification and estimation of models for recurrent event data with complex
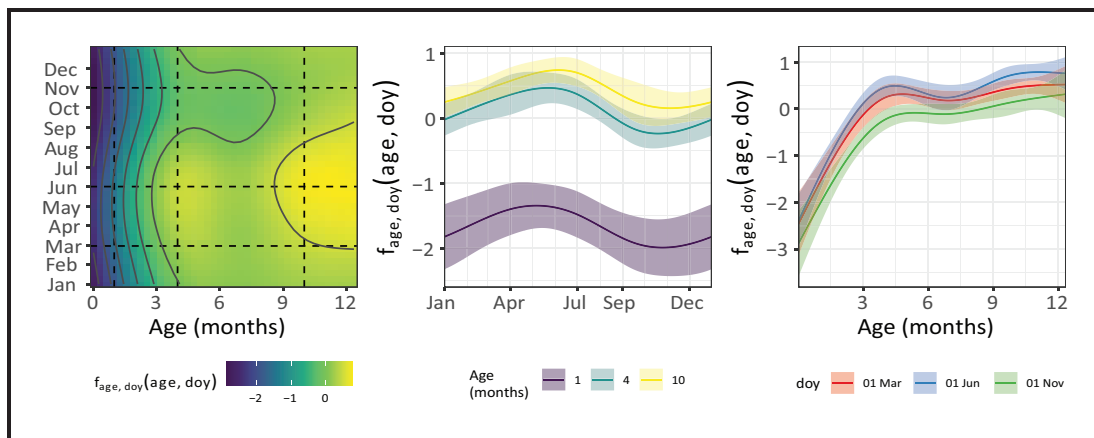
Figure 3   Left panel: The estimated bivariate effect $f_0(t_j, \text{doy})$.
Note: Brighter colors indicate increased hazards. Horizontal and vertical dashed lines indicate slices through the surface shown in the mid and right panel respectively. Mid panel: Slices through the bivariate surface for child ages 1, 4 and 10 months. Right panel: Slices through the bivariate surface at 1 March, 1 June and 1 November for the day of the year.

covariate effects, including (multivariate) non-linear effects, time-varying effects and covariates, multiple timescales as well as random effects. We illustrated the application of PAMMs and specification of different covariate effects on real data sets.

In simulation studies we have shown that in the proportional hazards setting, SFM PAMMs give similar results to the Cox SFM with respect to the estimated effects and frailty variance. In certain scenarios of clustered survival data (including recurrent events) it is difficult to distinguish between unobserved heterogeneity and time-varying effects (non-proportional hazards) in the SFM, especially for smaller cluster sizes (where a cluster size of one implies univariate survival data) (Balan and Putter, 2019). Such identifiability issues result from the data generating process and will persist independently of the model class chosen for the analysis of such data. However, from our simulation study in Supplement 4, we saw that the PAMM has a lower type 1 error rate for detecting a frailty effect in the presence of time-varying effects, even for clusters as small as size two, compared with the Cox SFM. This difference is much more pronounced for larger sample sizes. Also notable in the simulation results, the PAMM may be subject to inflated type 1 errors when used for testing time-varying effects, regardless of whether or not there is unobserved heterogeneity. This simulation study was however limited to a single binary covariate, monotonic shapes of the baseline hazards and without stratification. A considerable amount of work, possibly in terms of a much larger simulation study with a focus on recurrent events, must be carried out to assess these preliminary findings more carefully.

Compared to parametric SFMs, SFMs estimated by PAMMs do not make any assumption about the distribution of event times. Compared to Cox based SFMs, PAMMs directly model the (baseline) hazard via penalized splines. Thus, while for Cox models (and piece-wise exponential models for that matter) the hazard can have implausible jumps between two subsequent time points, PAMMs enforce a penalty on large differences between hazards from neighboring time points (intervals), leading to stable and more plausible hazard estimates.

While PAMMs offer a flexible framework for recurrent events analysis, some readers might be concerned with the expansion of the data set that results from the data augmentation step and, consequently, the computation time. In our experience this is rarely an issue. Especially for small and medium sized data sets, the computation time is barely noticeable (even when using all unique event times as cut-points). For example, on a standard laptop with Intel core-i7 CPU with 1.90GHz, 16GB RAM, fitting the interaction model in Section 5.1 (Equation (5.1)) takes about 1 second; the model estimating incidences of Malaria infection (with bivariate effects, Equation (5.3)), in Section 5.2, takes about 10 seconds. For high-dimensional data efficient estimation methods exist (Wood et al., 2017; Greven and Scheipl, 2020). For models on the gap timescale, computation time could be further improved without sacrificing predictive performance by reducing the number of interval cut points (cf. complexity analysis in Bender et al., 2021).

The R package `pammtools` has been extended to support recurrent events analysis, including the required data augmentation on the gap as well as calendar timescale. The convenience functions in `pammtools` allow the calculation of covariate effects, hazards rates, hazard ratios, cumulative hazards and survival probabilities according to the user's specifications. This means that these quantities, including confidence intervals, can be estimated and visualized in a few simple steps, even in case of complex models (cf. Section 5, Supplement 3).

Another advantage of the PAMM framework is that it does not require any particular implementation in order to estimate the model. While `pammtools` provides many convenience functions, once the data is transformed, any software that can maximize the Poisson likelihood can be used for model estimation. However, some methods and implementations will be more suitable than others. In this article, we estimated PAMMs using the R package `mgcv`, which offers great support for complex modelling of covariate effects. Nevertheless, in some settings, it might have limitations, that can be compensated by using a different software for estimation. For example, while `mgcv` offers support for random effects estimation, it is inefficient when random effects have high cardinality. In such cases, specialized mixed modelling software might be preferred, for example, `lme4` (Bates et al., 2015), `nlme` (Pinheiro et al., 2020) or `gamm4` (Wood and Scheipl, 2020). The mixed modelling framework implemented in these packages can further be exploited to estimate hierarchical models with nested or crossed random effects or more general hierarchical generalized additive models, that can be estimated via empirical Bayes (Pedersen et al., 2019) or fully Bayesian techniques (Bürkner, 2017).

A current limitation of these techniques and respective implementations is that they are limited to modelling Gaussian distributed random effects, while Gamma distributed random effects are popular in the context of survival analysis (Balan and Putter, 2020). However, through simulation studies by Gasparini et al. (2019); Liu et al. (2017), it has been shown that the estimates of the regression coefficients are quite robust to misspecification in the choice of frailty distributions regardless of sample size or the amount of heterogeneity as determined by the frailty variance. Gasparini et al. (2019) showed that it is more important to prioritize modelling the baseline hazard correctly. Alternatively, package `hglm` could be used to estimate Gamma distributed random effects, however, it has no native support for penalized spline estimation. Furthermore, the `gamlss` package (Rigby and Stasinopoulos, 2005), offers estimation of non-parametric random effects (Coupé, 2018). Both options, however, need to be explored further in the context of survival and recurrent events analysis.

This work highlights the usefulness and possibilities of extending PAMMs to recurrent events data.In future research, this model class could be further extended to settings with multivariate recurrent events (i.e., when multiple time-to-event outcomes can (re)occur on the same

observation unit), recurrent events data with a terminal or competing event or for recurrent events data with non-zero time to recovery.

7   Availability of data and materials

Data for the Staphylococcus aureus example are contained within the `pammtools` package and are also available from the publisher's webpage as Supplement 5. The malaria data is not available for public sharing but can be accessed as indicated in the data availability section in the manuscript Kakuru et al. (2020). We did however simulate a malaria dataset with 200 subjects from the original data where we changed the peaks of the seasonality effect. This data is available in Supplement 6. Additionally, the analysis presented here is accompanied by an online supplement that illustrates the data structure and code for data transformation, model fitting and visualization (cf. Supplements 2 and 3) in detail. Further, a vignette for recurrent events PAMMs can be found at https://adibender.github.io/pammtools/articles/recurrent-events.html. The R code and supplementary materials for the simulation study in Supplement 1 can be found in https://github.com/jordache-ramjith/PAMM and doi.org/10.6084/m9.figshare.14638353.v1 respectively. The R code for the simulation data in Supplement 4, can be requested from the authors Balan and Putter (2019) directly.

Acknowledgements

Supplementary material

Supplementary material is available online.

Declaration of conflicting interests

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

Funding

References

Abdulgader SM, Robberts L, Ramjith J, Nduru PM, Dube F, Gardner-Lubbe S, Zar HJ and Nicol MP (2019) Longitudinal population dynamics of staphylococcus aureus in the nasopharynx during the first year of life. *Frontiers in Genetics*, **10**, 198.

Akinboyo IC, Voskertchian A, Gorfu G, Betz J, Ross T, Carroll KC and Milstone AM (2018) Epidemiology and risk factors for recurrent staphylococcus aureus colonization following active surveillance and decolonization in the nicu. *Infection Control and Hospital Epidemiology*, **39**, 13349.

Amorim LD and Cai J (2015) Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology*, **44**, 324–33.

Andersen PK and Gill RD (1982) Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, **10**, 1100–20.

Argyropoulos C and Unruh ML (2015) Analysis of time to event outcomes in randomized controlled trials by generalized additive models. *PLoS ONE*, **10**, e0123784. https://doi.org/10.1371/journal.pone.0123784.

Balan TA and Putter H (2020) A tutorial on frailty models. *Statistical Methods in Medical Research*, **29**, 3424–54.

Balan TA and Putter H (2019) Nonproportional hazards and unobserved heterogeneity in clustered survival data: When can we tell the difference? *Statistics in Medicine*, **38**, 3405–20.

Barakat AN, Hussein MM, Fouda EM, Zoair AM and Abd El-Razek AM (2021) The underlying causes of recurrent pneumonia in children: A two-center study. *Journal of Advances in Medicine and Medical Research*, **33**, 62–9.

Bates D, Mächler M, Bolker B and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. https://doi.org/10.18637/jss.v067.i01.

Bender A and Scheipl F (2018) *pammtools: Piece-wise exponential Additive Mixed Modeling tools*. https://doi.org/10.48550/arXiv.1806.01042.

Bender A, Groll A and Scheipl F (2018) A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, **18**, 299–321.

Bender A, Scheipl F, Hartl W, Day AG and Küchenhoff H (2019) Penalized estimation of complex, non-linear exposure-lag-response associations. *Biostatistics*, **20**, 315–31. ISSN 1465-4644. https://doi.org/10.1093/biostatistics/kxy003.

Bender A, Rügamer D, Scheipl F and Bischl B (2021) A general machine learning framework for survival analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 158–173. Springer International Publishing. https://arxiv.org/abs/2006.15442.

Box-Steffensmeier JM and De Boef S (2006) Repeated events survival models: the conditional frailty model. *Statistics in Medicine*, **25**, 3518–33.

Bürkner P-C (2017) brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, **80**, 1–28. ISSN 1548-7660. https://doi.org/10.18637/jss.v080.i01.

Carmona-Bayonas A, Jiménez-Fonseca P, Lamarca Á, Barriuso J, Castaño Á, Benavent M, Alonso V, Riesco-Martínez MdC, Alonso-Gordoa T, Custodio A, SÁnchez CÁnovas M, Hernando Cubero J, López C, Lacasta A, FernÁndez Montes A, Marazuela M, Crespo G, Escudero P,

Diaz JÁ, Feliciangeli E, Gallego J, Llanos M, Segura Á, Vilardell F, Percovich JC, Grande E, Capdevila J, Valle JW and García-Carbonero R (2019) Prediction of progression-free survival in patients with advanced, well-differentiated, neuroendocrine tumors being treated with a somatostatin analog: The GETNE-TRASGU study. *Journal of Clinical Oncology*, **37**, 2571–80.

Chan KP, Ko FWS, Ling KC, Cheung PS, Chan LV, Chan YH, Lo YT, Ng CK, Lui MM-s, Yee KSW, et al. (2021) A territory-wide study on the factors associated with recurrent asthma exacerbations requiring hospitalization in hong kong. *Immunity, Inflammation and Disease*, **9**, pages 1–13.

Coupé C (2018) Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in psychology*, **9**, 513.

Cox DR (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187–202.

Dos Santos LA, de Góis Filho PG, Silva AMF, Santos JVG, Santos DS, Aquino MM, de Jesus RM, Almeida MLD, da Silva JS, Altmann DM, et al. (2021) Recurrent covid-19 including evidence of reinfection and enhanced severity in thirty brazilian healthcare workers. *Journal of Infection*, **82**, 399–406.

Duchateau L, Janssen P, Kezic I and Fortpied C (2003) Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**, 355–63.

Eilers PHC and Marx BD (1996) Flexible smoothing with B -splines and penalties. *Statistical Science*, **11**, 89–121. https://doi.org/10.1214/ss/1038425655.

Feng B, He C, Liu X, Chen Y and He S (2021) Effect of congenital heart disease on the recurrence of cough variant asthma in children. *BMC Cardiovascular Disorders*, **21**, 1–9.

Freedman DA (2006) On the so-called "huber sandwich estimator" and "robust standard errors". *The American Statistician*, **60**, 299–302.

Gasparini A, Clements MS, Abrams KR and Crowther MJ (2019) Impact of model misspecification in shared frailty survival models. *Statistics in medicine*, **38**, 4477–502.

Ghosh M, Olaniyi S and Obabiyi OS (2020) Mathematical analysis of reinfection and relapse in malaria dynamics. *Applied Mathematics and Computation*, **373**, 125044.

Giacoppo D, Alfonso F, Xu B, Claessen BEPM, Adriaenssens T, Jensen C, Pérez-Vizcayno MJ, Kang D-Y, Degenhardt R, Pleva L, Baan J, Cuesta J, Park D-W, Kukla P, Jiménez-Quevedo P, Unverdorben M, Gao R, Naber CK, Park S-J, Henriques JPS, Kastrati A and Byrne RA (2020) Drug-coated balloon angioplasty versus drug-eluting stent implantation in patients with coronary stent restenosis. *Journal of the American College of Cardiology*, **75**, 2664–78.

Gibson CM, Pinto DS, Chi G, Arbetter D, Yee M, Mehran R, Bode C, Halperin J, Verheugt FW, Wildgoose P, et al. (2017) Recurrent hospitalization among patients with atrial fibrillation undergoing intracoronary stenting treated with 2 treatment strategies of rivaroxaban or a dose-adjusted oral vitamin k antagonist treatment strategy. *Circulation*, **135**, 323–33.

Greven S and Scheipl F (2020) Comments on: Inference and computation with generalized additive models and their extensions. *TEST*, **29**, 343–50.

Hastie T and Tibshirani R (1986) Generalized Additive Models. *Statistical Science*, **1**, 297–310. ISSN 0883-4237. https://doi.org/10.1214/ss/1177013604.

Hernstadt H, Cheung A, Hurem D, Vasilunas N, Phuong LK, Quinn P, Agrawal R, Daley AJ, Cole T and Gwee A (2020) Changing epidemiology and predisposing factors for invasive pneumococcal disease at two australian tertiary hospitals. *The Pediatric infectious disease journal*, **39**, 1–6.

Hougaard P (2012) *Analysis of multivariate survival data*. Springer Science & Business Media.

Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press.

Hutton JL and Monaghan P (2002) Choice of parametric accelerated life and proportional hazards models for survival data: asymptotic results. *Lifetime Data Analysis*, **8**, 375–93.

Iacobelli S and Carstensen B (2013) Multiple time scales in multi-state models. *Statistics in Medicine*, **32**, 5315–27. https://doi.org/10.1002/sim.5976.

Jahn-Eimermacher A (2008) Comparison of the andersen–gill model with poisson and negative binomial regression on recurrent event data. *Computational Statistics & Data Analysis*, **52**, 4989–97.

Kakuru A, Jagannathan P, Kajubi R, Ochieng T, Ochokoru H, Nakalembe M, Clark TD, Ruel T, Staedke SG, Chandramohan D, et al. (2020) Impact of intermittent preventive treatment of malaria in pregnancy with dihydroartemisinin-piperaquine versus sulfadoxine-pyrimethamine on the incidence of malaria in infancy: a randomized controlled trial. *BMC Medicine*, **18**, 1–11.

Kelly PJ and Lim LL-Y (2000) Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, **19**, 13–33.

Kwong G and Hutton J (2003) Choice of parametric models in survival analysis: applications to monotherapy for epilepsy and cerebral palsy. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**, 153–68.

Lawpoolsri S, Sattabongkot J, Sirichaisinthop J, Cui L, Kiattibutr K, Rachaphaew N, Suk-Uam K, Khamsiriwatchara A and Kaewkungwal J (2019) Epidemiological profiles of recurrent malaria episodes in an endemic area along the thailand-myanmar border: a prospective cohort study. *Malaria Journal*, **18**, 1–11.

Li Y-H, Klein JP and Moeschberger M (1996) Effects of model misspecification in estimating covariate effects in survival analysis for small sample sizes. *Computational Statistics & Data Analysis*, **22**, 177–92.

Lin X and Zhang D (1999) Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 381–400.

Liu XR, Pawitan Y and Clements MS (2017) Generalized survival models for correlated time-to-event data. *Statistics in Medicine*, **36**, 4743–62.

Munda M, Rotolo F and Legrand C (2012) parfm: Parametric frailty models in r. *Journal of Statistical Software*, **51**, 1–20.

Noordzij M, van Diepen M, Caskey FC and Jager KJ (2017) Relative risk versus absolute risk: one cannot be interpreted without the other. *Nephrology Dialysis Transplantation*, **32**, ii13–18.

Panel ADC, Pen I, Pannebakker BA, Helsen HHM and Wertheim B (2020) Seasonal morphotypes of drosophila suzukii differ in key life-history traits during and after a prolonged period of cold exposure. *Ecology and Evolution*, **10**, 9085–99.

Pedersen EJ, Miller DL, Simpson GL and Ross N (2019) Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ*, **7**, https://doi.org/10.7717/peerj.6876.

Pinheiro J, Bates D, DebRoy S, Sarkar D and R Core Team (2020) *nlme: Linear and Nonlinear Mixed Effects Models*. URL https://CRAN.R-project.org/package=nlme.

Prentice RL, Williams BJ and Peterson AV (1981) On the regression analysis of multivariate failure time data. *Biometrika*, **68**, 373–9.

Putter H, Sasako M, Hartgrink H, Van de Velde C and Van Houwelingen J (2005) Long-term survival with non-proportional hazards: Results from the dutch gastric cancer trial. *Statistics in Medicine*, **24**, 2807–21.

Pya N and Wood SN (2015) Shape constrained additive models. *Statistics and Computing*, **25**, 543–59. https://doi.org/10.1007/s11222-013-9448-7.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*.

R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramjith J, Roes KC, Zar HJ and Jonker MA (2021) Flexible modelling of risk factors on the incidence of pneumonia in young children in south africa using piece-wise exponential additive mixed modelling. *BMC Medical Research Methodology*, **21**. https://doi.org/10.1186/s12874-020-01194-6.

Rigby RA and Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507–54.

RStudio Team (2020) *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA. URL http://www.rstudio.com/.

Ruhe C (2018) Quantifying change over time: Interpreting time-varying effects in duration analyses. *Political Analysis*, **26**.

Therneau TM and Hamilton SA (1997) rhdnase as an example of recurrent event analysis. *Statistics in medicine*, **16**, 2029–47.

Ullah S, Gabbett TJ and Finch CF (2014) Statistical modelling for recurrent events: an application to sports injuries. *British Journal of Sports Medicine*, **48**, 1287–93.

Varma N, Bourge RC, Stevenson LW, Costanzo MR, Shavelle D, Adamson PB, Ginn G, Henderson J, Abraham WT and CHAMPION Investigator Group (2021) Remote hemodynamic-guided therapy of patients with recurrent heart failure following cardiac resynchronization therapy. *Journal of the American Heart Association*, **10**, e017619.

Wand MP (2003) Smoothing and mixed models. *Computational statistics*, **18**, 223–49.

Wei L-J, Lin DY and Weissfeld L (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84**, 1065–73.

Wey A, Hart A, Salkowski N, Skeans M, Kasiske BL, Israni AK and Snyder JJ (2020) Posttransplant outcome assessments at listing: Long-term outcomes are more important than short-term outcomes. *American Journal of Transplantation*, **20**, 2813–21.

Wood SN (2003) Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, **65**, 95–114.

Wood SN (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**, 673–86.

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, **73**, 3–36.

Wood S and Scheipl F (2020) *gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'*. URL https://CRAN.R-project.org/package=gamm4.

Wood SN (2006) Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics*, **62**, 1025–36.

Wood SN (2017) *Generalized additive models: an introduction with R*. CRC press.

Wood SN, Li Z, Shaddick G and Augustin NH (2017) Generalized additive models for giga-data: modeling the uk black smoke network daily data. *Journal of the American Statistical Association*, **112**, 1199–1210.

Wood S, Pya N. and Säfken B (2016) Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, **111**, 1548–75.