Recurrent Events Analysis with Piece-wise exponential Additive Mixed Models

Jordache Ramjith (jordache.ramjith@radboudumc.nl)
Radboud University Medical Center
Andreas Bender
Ludwig-Maximilians-Universität München
Kit C. B. Roes
Radboud University Medical Center
Marianne A. Jonker
Radboud University Medical Center

Research Article

Keywords: Cox proportional hazards, Flexible survival models, Multiple Timescales, Non-linear effects, Non-proportional hazards, Penalized splines, Piece-wise exponential Additive Mixed Model, Recurrent events, Survival Analysis, Time-varying effects

Posted Date: June 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-563303/v1

Version of Record: A version of this preprint was published at Statistical Modelling on September 8th, 2022. See the published version at https://doi.org/10.1177/1471082X221117612.

RESEARCH

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]

*Correspondence:
jordache.ramjith@radboudumc.nl
[1]Department for Health Evidence, Biostatistics Research Group, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, Netherlands
Full list of author information is available at the end of the article

Abstract

Background: Recurrent events analysis plays an important role in many applications, including the study of chronic diseases or recurrence of infections. Historically, most models for the analysis of time-to-event data, including recurrent events, have been based on Cox proportional hazards regression. Recently, however, the Piece-wise exponential Additive Mixed Model (PAMM) has gained popularity as a flexible framework for survival analysis. While many papers and tutorials have been presented in the literature on the application of Cox based models, few papers have provided detailed instructions for the application of PAMMs and to our knowledge, none exist for recurrent events analysis.

Methods: The PAMM is introduced as a framework for recurrent events analysis. We describe the application of the model to unstratified and stratified shared frailty models for recurrent events. We illustrate how penalized splines can be used to estimate non-linear and time-varying covariate effects without a priori assumptions about their functional shape. The model is motivated for both, analysis on the gap timescale ("clock-reset") and calendar timescale ("clock-forward"). The data augmentation necessary for the application to recurrent events is described and explained in detail.

Results: Simulations confirmed that the model provides unbiased estimates of covariate effects and the frailty variance, as well as equivalence to the Cox model when proportional hazards are assumed. Applications to recurrence of staphylococcus aureus and malaria in children illustrates 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` has been extended to facilitate estimation, visualization and interpretation of PAMMs for recurrent events analysis.

Conclusion: PAMMs provide a flexible framework for the analysis of time-to-event and recurrent events data. The estimation of PAMMs is based on Generalized Additive Mixed Models and thus extends the researcher's toolbox for recurrent events analysis.

Keywords: Cox proportional hazards; Flexible survival models; Multiple timescales; Non-linear effects; Non-proportional hazards; Penalized splines; Piece-wise exponential Additive Mixed Model; Recurrent events; Survival analysis; Time-varying effects

Background

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 [1, 2], recurrent diseases and infections like COVID-19 [3], pneumonia [4, 5], pneumococcus[6], staphylococcus aureus (cf. section Example 1: Staphylococcus aureus infection data) [7, 8], asthma [9, 10] and malaria (cf. section Example 2: Childhood malaria data) [11–13].

Typical for this kind of data are the correlations/dependencies of 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 (cf. [14, 15]). 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 Timescale 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 [16, 17]) or by including a subject-specific random effect in the model (the shared frailty model). 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 [18, 19]. Shared frailty models [20] 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 frailty model it is assumed that the hazards are proportional over time (conditional on the random effect) and the baseline hazard function is unspecified, but a parametric form may be chosen as well. For interpretation of the estimated effects, an estimate of the baseline hazard function is important. [21] 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 [22]. They suggest that readers tend to overestimate or overinterpret 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 [23].

Many publications on recurrent events analysis focus on variants of the Cox proportional hazards (CPH) model [24]. This includes the shared frailty model, mentioned earlier, and the variance-adjustment models, specifically the Andersen-Gill (AG) model, the Prentice-Williams-Peterson (PWP) models, and the Wei, Lin & Weissfeld (WLW) model [16, 17, 25]. 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 Timescale) 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 [26, 27] have shown a carry-over effect when using the WLW model and recommend that it be used for modelling multiple event types rather than recurrent events. There are several tutorials available with applications and comparisons of these methods [26–29], and a recent tutorial on frailty models is provided by [30].

The Piece-wise exponential Additive Mixed Model (PAMM) [31], conceptually similar to the Cox model, is 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. While this allows the model to be estimated by any algorithm that can optimize the Poisson likelihood [32], the Generalized Additive Mixed Model (GAMM) framework has proven to be particularly useful, due to its flexibility, availability of robust and established estimation techniques [33] and interpretability of covariate effects. This allows great versatility in modeling choices, as discussed in more detail in section Discussion. 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 monotonicity constraints. Recent examples include spline based stratified baseline hazard estimation in plant biology [34], cumulative effects in critical care [35], treatment effects with non-proportional hazards in patients with coronary stent restenosis [36], estimation of bivariate effect surfaces in the context of randomized controlled trials[37] and progression-free survival in patients with gastroenteropancreatic neuro-endocrine tumors [38], and time-varying hazard ratios in post-transplant outcome assessment [39]. An application to the recurrence of childhood pneumonia was recently presented by Ramjith et al. (2021) [4]. 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 paper 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 paper is accompanied by R package `pammtools` that was extended to facilitate application of PAMMs to recurrent events data. Using `pammtools`, analysis of recurrent events data only requires one additional function call for the required data transformation and additionally facilitates visualization and interpretation of the estimates.

In the remainder of this work, we summarize hazard based model specification and estimation for recurrent events analysis in section Methods and 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 CPH frailty model and the PAMM when proportional hazards are assumed is made through a simulation study in section Results and concludes with examples of PAMMs applied to different recurrent events data sets and illustrates effect estimation of different complexities, i.e., stratified baseline hazards, non-linear, cyclic effects of seasonality and multiple timescales. We conclude with a discussion and an outlook in sections Discussion and Conclusion.

Methods

Models for recurrent events

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) [40] for a thorough discussion). On both timescales, the study start time needs to be clearly defined and meaningful, e.g., 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 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 modelled by stratifying the hazards by event number. Dependence introduced by within-subject correlation can be modelled 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 [41], as any dependence on time is modelled through covariate effects of different representations of time.

Model specification

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

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

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. 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 (1) by relaxing some of the dependencies. For example, with linear effects and under the proportional hazards assumption we get the stratified shared frailty model [14]:

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

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) frailty model [20].

The conditional likelihood for model (1) in calendar time (given the individual frailty terms) is proportional to

$$\mathcal{L} \propto \prod_{i=1}^{n}\prod_{k=1}^{K_i} \lambda_{i,k}(t_{i,k}|z_i)^{\delta_{i,k}} \exp\left(-\int_{t_{i,k-1}}^{t_{i,k}} \lambda_{i,k}(s|z_i)\mathrm{d}s\right), \quad (2)$$

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 [40]). The likelihood in (2) is often given in the simplified form

$$\exp\left(-\sum_{i=1}^{n}\int_0^{t_{i,K_i}} \lambda_{i,k}(s|z_i)\mathrm{d}s\right)\prod_{i=1}^{n}\prod_{k=1}^{K_i-1} \lambda_{i,k}(t_{i,k}|z_i), \quad (3)$$

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

$$\prod_{i=1}^{n}\left(\int_{z_i}\prod_{k=1}^{K_i} \lambda_{i,k}(t_{i,k}|z_i)^{\delta_{i,k}} \exp\left(-\int_{t_{i,k-1}}^{t_{i,k}} \lambda_{i,k}(s|z_i)\mathrm{d}s\right) f_Z(z_i)\mathrm{d}z_i\right), \quad (4)$$

where f_Z is the density function of the frailty variable.

The conditional likelihood for the gap time model is equivalent to (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:

$$\prod_{i=1}^{n}\prod_{k=1}^{K_i} \lambda_{i,k}(d_{i,k}|z_i)^{\delta_{i,k}} \exp\left(-\int_0^{d_{i,k}} \lambda_{i,k}(s|z_i)\mathrm{d}s\right). \tag{5}$$

PAMMs for recurrent events analysis

As described in detail elsewhere (cf. [31]), 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 cutpoints $0 = \tau_0 < \tau_1 < \cdots < \tau_J$. The j^{th} 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 cutpoints 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 An example of recurrent events data in PED format.

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 k^{th} 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.

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

$$\lambda_{i,k}(t|\mathbf{x}_i(t), z_i, k) = \lambda_{j,k} \exp(g(\mathbf{x}_i(t), t, k) + z_i) = \lambda_{ijk} \qquad \forall t \in (\tau_{j-1}, \tau_j] \tag{6}$$

in each interval, the likelihood contribution of subject i (conditional on the frailty variable) in (2) can be rewritten as

$$\lambda_{i,k}(t_{i,k}|z_i)^{\delta_{i,k}} \exp\left(-\int_{t_{i,k-1}}^{t_{i,k}} \lambda_{i,k}(s|z_i)\mathrm{d}s\right) = \left(\prod_{j\in\mathcal{J}_{ik}} \lambda_{ijk}^{\delta_{ijk}}\right) \exp\left(-\sum_{j\in\mathcal{J}_{ik}} \lambda_{ijk}t_{ijk}\right), \tag{7}$$

where $\mathcal{J}_{ik}$ is the set of intervals for which i is at risk for event k. The right-hand side of (7) 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 $\mathcal{L}$ is proportional to the conditional likelihood function that is found under the working assumption just mentioned:

$$\mathcal{L} \propto \mathcal{L}_{Po} = \prod_{k=1}^{K}\prod_{i=1}^{n}\left(\prod_{j\in\mathcal{J}_{ik}} \frac{\mu_{ijk}^{\delta_{ijk}}}{\delta_{ijk}!}\right) \exp\left(-\sum_{j\in\mathcal{J}_{ik}} \lambda_{ijk}t_{ijk}\right). \tag{8}$$

Similar to (4), the conditional likelihood in (8) can be expressed as a marginal likelihood by specifying a frailty distribution:

$$\prod_{i=1}^{n}\int_{z_i}\prod_{k=1}^{K}\left(\prod_{j\in\mathcal{J}_{ik}} \frac{\mu_{ijk}^{\delta_{ijk}}}{\delta_{ijk}!}\right) \exp\left(-\sum_{j\in\mathcal{J}_{ik}} \lambda_{ijk}t_{ijk}\right) f_Z(z_i)\mathrm{d}z_i, \tag{9}$$

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 k^{th} event in interval j and δ_{ijk} indicates whether subject i, experienced an event of type k in interval j. Thus, parameters of model (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 modeling framework, the integral can be approximated using numerical techniques such as adaptive Gauss-Hermite quadrature (c.f. [42]). In the GAMMs framework, the integral is either numerically approximated or the link between penalized splines and random effects models [33, 43–45] is used to treat the frailties as penalized splines where the frailty variance is estimated as part of the penalty (for further reading see [33]).

This reformulation of recurrent events models into Poisson regression models is interesting especially due to flexibility of GAMMs and readily available software.

In PAMMs we then parameterize the interval- and event-specific hazard rates (6) via (non-linear) covariate effects. An exemplary model specification is given below:

$$\lambda\left(t \mid \mathbf{x}_i(t), z_i, k\right) = \exp\left(\beta_{0,k} + f_{0,k}(t_j) + \sum_{p=1}^{P} f_p(x_{i,p}(t_j), t_j) + z_i\right) \qquad \forall\, t \,\epsilon\, (\tau_{j-1}, \tau_j]. \tag{10}$$

In (10), $\beta_{0,k} + f_{0,k}(t_j)$ 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, e.g. $f_{0,k}(s) = \sum_{\ell=1}^{L} \gamma_{\ell,k} b_{\ell,k}(s)$ 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 (cf. [33]). From an estimation perspective, the model is a Poisson regression task with random effects. Therefore, the parameters of the model can be estimated by optimizing the penalized Poisson likelihood using the GAMM framework (e.g., [33]). As discussed elsewhere [31], 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, (10) could be simplified, e.g., by dropping the dependence on event number or by removing the frailty term.

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.

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.

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.

Results

Simulation study

Under the assumption of proportional hazards the proposed PAMM and the CPH frailty model 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 Additional file 1. In summary, it is shown that under the proportional hazards assumption the PAMM and the CPH frailty models 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 CPH frailty model, 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.

Motivating examples

In this section we apply the PAMM for recurrent events to two real-life data sets using R [46, 47]. We use the R packages `pammtools` [48] for the data transformation and the visualization of the results and `mgcv` [33, 45, 49–51] for the estimation of the parametric parameters and the smoothed functions in the generalized additive modelling framework.

The aim of this section is to show the application of several types of recurrent events models using the PAMM framework.

Example 1: Staphylococcus aureus infection data

The staphylococcus aureus (SA) infection data in [8] 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 [8], the gap time PWP model [17] 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 10) in gap time. We consider a different baseline hazard for the first event and for recurrent events (combined). We use the HIV exposure variable, where some children are HIV exposed & 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.

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.
The estimated degrees of freedom, explained in Additional file 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 and survival 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.

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

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

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 ($HR = \exp(\beta_1) = 1.31,\ 95\%\ \text{CI} : 0.98 - 1.74$).
In the paper, 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. i.e.

$$\lambda(t) = \lambda_{0,k}(t) \times \exp(\beta_{1,k} \times \text{HIV})$$

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 : $1.16 - 2.54$)).

Example 2: Childhood malaria data
To illustrate application of PAMMs on calendar timescale, we reanalyze data from a study of childhood malaria undertaken by Kakuru et. al (2020) [13] 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.

The model: We fit the following model

$$\begin{aligned}\log(\lambda(t|\mathbf{x}_i, z_i)) = \beta_0 + f_0(t_j) + f(\text{doy}_i) \\ + \beta_1 \text{DP}_i + \beta_2 \text{sex}_i + \beta_3 \text{DP}_i \cdot \text{sex}_i \\ + \beta_3 \text{preterm}_i + \beta_4 \text{gravidity}_i + z_i\end{aligned} \tag{11}$$

where $\beta_0 + f_0(t_j)$ is the log-baseline hazard as defined in section PAMMs for recurrent events analysis and Equation (10) (here equivalent to the smooth effect of child age). We model the effect of seasonality via variable "day of year" (doy) and set $f(\text{doy}) = \sum_{m=1}^{10} \gamma_m b_m(doy)$, 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 [33, 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 (95% CI): 0.74 $(0.58, 0.95)$), but not for girls (HR (95% CI): 0.98 $(0.78, 1.24)$). The interaction effect showed no statistically significant difference in the treatment effects between boys and girls ($p = 0.0996$). 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.

The estimated non-linear effects of age and seasonality are depicted in Figure 2. The upper row shows the 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.

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 (11) becomes

$$\begin{aligned}\log(\lambda(t|\mathbf{x}_i, z_i)) = \beta_0 + f_0(t_j, \text{doy}_i) + \beta_1 \text{DP}_i + \beta_2 \text{sex}_i + \beta_3 \text{DP}_i \cdot \text{sex}_i \\ + \beta_3 \text{preterm}_i + \beta_4 \text{gravidity}_i + z_i \end{aligned} \tag{12}$$

Note that in (12) 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 (11) 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 [52], as illustrated in Additional file 3.

Discussion

In this paper we introduced PAMMs as an alternative modeling approach for recurrent events data. These models utilize the generalized additive mixed modeling framework for estimation and inference. This facilitates specification and estimation of models for recurrent event data with complex covariate effects, including (multivariate) non-linear effects, time-varying effects and covariates, multiple timescales as well as random effects. In simulation studies we have shown that in the proportional hazards setting, PAMMs give results equivalent to the CPH shared frailty model. Further, we illustrated the application of PAMMs and specification of different covariate effects on real data sets.

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 high-dimensional data efficient estimation methods exist [53, 54]. 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 [32]).

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. Motivating examples, Additional file 2 and Additional file 3).

One advantage of the PAMM framework is that it does not require any particular implementation in order to estimate the model. Rather, any software that can maximize the Poisson likelihood can be used. However, some methods and implementations will be more suitable than others. In this paper, we estimated PAMMs using the R package `mgcv`, which offers great support for complex modeling 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 modeling software might be preferred, e.g., `lme4` [42], `nlme` [55] or `gamm4` [56]. The mixed modeling 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 [57] or fully Bayesian techniques [58].

A possible limitation of these techniques and respective implementations is that they are limited to modeling Gaussian distributed random effects, while Gamma distributed random effects are popular in the context of survival analysis [30]. However, through simulation studies by [59, 60], 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. [59] showed that it is more important to prioritize modeling 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 [61], offers estimation of non-parametric random effects [62]. Both options, however, need to be explored further in the context of survival and recurrent events analysis.

Conclusion

This work highlights the usefulness and possibilities of extending recurrent events models into a well-established framework of flexible models, that offer several advantages, from a modeling perspective, to that of visualization and interpretation of complex models.

Abbreviations
AG - Andersen-Gill; CI - confidence interval; CPH - Cox proportional hazards; doy - day of year; DP - Dihydroartemisininpiperaquine; GAMM - Generalized Additive Mixed Model; HEU - HIV exposed & uninfected; HU - HIV unexposed; SA - staphylococcus aureus; SP - Sulfadoxine-pyrimethamine; PAMM - piece-wise exponential additive mixed model; PED - piece-wise exponential data; PWP - Prentice-Williams-Peterson; WLW - Wei, Lin & Weisfeld (WLW)

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
Data for the Staphylococcus aureus example are contained within the `pammtools` package. The malaria data is not available for public sharing but can be accessed as indicated in the data availability section in the manuscript [13]. 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. Additional file 2 and Additional file 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 additional materials for the simulation study in Additional file 1 can be found in `https://github.com/jordache-ramjith/PAMM` and `doi.org/10.6084/m9.figshare.14638353.v1` respectively.

Competing interests
The authors declare that they have no competing interests.

Funding
AB was funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content. The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions
J.R: Conceptualization, methodology, formal analysis, writing – original draft, review and editing, visualization, simulation study, A.B: Conceptualization, methodology, formal analysis, writing – original draft, review and editing, visualization, software, K.C.B.R: Supervision, writing – review and editing, M.A.J: Conceptualization, supervision, writing – review and editing. All authors have read and approved the manuscript.

Acknowledgements
We thank Shima Mohammed, Mark Nicol and Heather Zar for access and support to use the staphylococcus aureus infection data from the Drakenstein Child Health Study. We also thank Abel Kakuru, Grant Dorsey and the rest of the authors in the Uganda malaria birth cohort study for access and support to use the data for secondary analysis.

Author details
[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.

References

1. Gibson, C.M., Pinto, D.S., Chi, G., Arbetter, D., Yee, M., Mehran, R., Bode, C., Halperin, J., Verheugt, F.W., Wildgoose, P., *et al.*: 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**(4), 323–333 (2017)
2. Varma, N., Bourge, R.C., Stevenson, L.W., Costanzo, M.R., Shavelle, D., Adamson, P.B., Ginn, G., Henderson, J., Abraham, W.T., CHAMPION Investigator Group: Remote hemodynamic-guided therapy of patients with recurrent heart failure following cardiac resynchronization therapy. Journal of the American Heart Association **10**(5), 017619 (2021)
3. Dos Santos, L.A., de Góis Filho, P.G., Silva, A.M.F., Santos, J.V.G., Santos, D.S., Aquino, M.M., de Jesus, R.M., Almeida, M.L.D., da Silva, J.S., Altmann, D.M., *et al.*: Recurrent covid-19 including evidence of reinfection and enhanced severity in thirty brazilian healthcare workers. Journal of Infection **82**(3), 399–406 (2021)
4. Ramjith, J., Roes, K.C., Zar, H.J., Jonker, M.A.: 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**(17) (2021). doi:10.1186/s12874-020-01194-6
5. Barakat, A.N., Hussein, M.M., Fouda, E.M., Zoair, A.M., Abd El-Razek, A.M.: The underlying causes of recurrent pneumonia in children: A two-center study. Journal of Advances in Medicine and Medical Research **33**(6), 62–69 (2021)
6. Hernstadt, H., Cheung, A., Hurem, D., Vasilunas, N., Phuong, L.K., Quinn, P., Agrawal, R., Daley, A.J., Cole, T., Gwee, A.: Changing epidemiology and predisposing factors for invasive pneumococcal disease at two australian tertiary hospitals. The Pediatric infectious disease journal **39**(1), 1–6 (2020)
7. Akinboyo, I.C., Voskertchian, A., Gorfu, G., Betz, J., Ross, T., Carroll, K.C., Milstone, A.M.: Epidemiology and risk factors for recurrent staphylococcus aureus colonization following active surveillance and decolonization in the nicu. Infection control and hospital epidemiology **39**(11), 1334 (2018)

8. Abdulgader, S.M., Robberts, L., Ramjith, J., Nduru, P.M., Dube, F., Gardner-Lubbe, S., Zar, H.J., Nicol, M.P.: Longitudinal population dynamics of staphylococcus aureus in the nasopharynx during the first year of life. Frontiers in genetics **10**, 198 (2019)
9. Chan, K.P., Ko, F.W.S., Ling, K.C., Cheung, P.S., Chan, L.V., Chan, Y.H., Lo, Y.T., Ng, C.K., Lui, M.M.-s., Yee, K.S.W., et al.: A territory-wide study on the factors associated with recurrent asthma exacerbations requiring hospitalization in hong kong. Immunity, Inflammation and Disease, 1–13 (2021)
10. Feng, B., He, C., Liu, X., Chen, Y., He, S.: Effect of congenital heart disease on the recurrence of cough variant asthma in children. BMC Cardiovascular Disorders **21**(1), 1–9 (2021)
11. Lawpoolsri, S., Sattabongkot, J., Sirichaisinthop, J., Cui, L., Kiattibutr, K., Rachaphaew, N., Suk-Uam, K., Khamsiriwatchara, A., Kaewkungwal, J.: Epidemiological profiles of recurrent malaria episodes in an endemic area along the thailand-myanmar border: a prospective cohort study. Malaria journal **18**(1), 1–11 (2019)
12. Ghosh, M., Olaniyi, S., Obabiyi, O.S.: Mathematical analysis of reinfection and relapse in malaria dynamics. Applied Mathematics and Computation **373**, 125044 (2020)
13. Kakuru, A., Jagannathan, P., Kajubi, R., Ochieng, T., Ochokoru, H., Nakalembe, M., Clark, T.D., Ruel, T., Staedke, S.G., Chandramohan, D., *et al.*: 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), 1–11 (2020)
14. Box-Steffensmeier, J.M., De Boef, S.: Repeated events survival models: the conditional frailty model. Statistics in medicine **25**(20), 3518–3533 (2006)
15. Jahn-Eimermacher, A.: Comparison of the andersen–gill model with poisson and negative binomial regression on recurrent event data. Computational Statistics & Data Analysis **52**(11), 4989–4997 (2008)
16. Andersen, P.K., Gill, R.D.: Cox's regression model for counting processes: a large sample study. The annals of statistics **10**(4), 1100–1120 (1982)
17. Prentice, R.L., Williams, B.J., Peterson, A.V.: On the regression analysis of multivariate failure time data. Biometrika **68**(2), 373–379 (1981)
18. Huber, P.J.: The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 221–233 (1967). University of California Press
19. Freedman, D.A.: On the so-called "huber sandwich estimator" and "robust standard errors". The American Statistician **60**(4), 299–302 (2006)
20. Hougaard, P.: Analysis of Multivariate Survival Data. Springer, ??? (2012)
21. Putter, H., Sasako, M., Hartgrink, H., Van de Velde, C., Van Houwelingen, J.: Long-term survival with non-proportional hazards: Results from the dutch gastric cancer trial. Statistics in medicine **24**(18), 2807–2821 (2005)
22. Noordzij, M., van Diepen, M., Caskey, F.C., Jager, K.J.: Relative risk versus absolute risk: one cannot be interpreted without the other. Nephrology Dialysis Transplantation **32**(suppl_2), 13–18 (2017)
23. Ruhe, C.: Quantifying change over time: Interpreting time-varying effects in duration analyses. Political Analysis **26**(1) (2018)
24. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
25. Wei, L.-J., Lin, D.Y., Weissfeld, L.: Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American statistical association **84**(408), 1065–1073 (1989)
26. Kelly, P.J., Lim, L.L.-Y.: Survival analysis for recurrent event data: an application to childhood infectious diseases. Statistics in medicine **19**(1), 13–33 (2000)
27. Therneau, T.M., Hamilton, S.A.: rhdnase as an example of recurrent event analysis. Statistics in medicine **16**(18), 2029–2047 (1997)
28. Amorim, L.D., Cai, J.: Modelling recurrent events: a tutorial for analysis in epidemiology. International journal of epidemiology **44**(1), 324–333 (2015)
29. Ullah, S., Gabbett, T.J., Finch, C.F.: Statistical modelling for recurrent events: an application to sports injuries. British journal of sports medicine **48**(17), 1287–1293 (2014)
30. Balan, T.A., Putter, H.: A tutorial on frailty models. Statistical Methods in Medical Research **29**(11), 3424–3454 (2020)
31. Bender, A., Groll, A., Scheipl, F.: A generalized additive model approach to time-to-event analysis. Statistical Modelling **18**(3-4), 299–321 (2018)
32. Bender, A., Rügamer, D., Scheipl, F., Bischl, B.: A general machine learning framework for survival analysis. In: Machine Learning and Knowledge Discovery in Databases, pp. 158–173. Springer, ??? (2021)
33. Wood, S.N.: Generalized Additive Models: an Introduction with R. CRC press, ??? (2017)
34. Panel, A.D.C., Pen, I., Pannebakker, B.A., Helsen, H.H.M., Wertheim, B.: Seasonal morphotypes of drosophila suzukii differ in key life-history traits during and after a prolonged period of cold exposure. Ecology and Evolution **10**(17), 9085–9099 (2020)
35. Bender, A., Scheipl, F., Hartl, W., Day, A.G., Küchenhoff, H.: Penalized estimation of complex, non-linear exposure-lag-response associations. Biostatistics **20**(2), 315–331 (2019). doi:10.1093/biostatistics/kxy003. Publisher: Oxford Academic
36. Giacoppo, D., Alfonso, F., Xu, B., Claessen, B.E.P.M., Adriaenssens, T., Jensen, C., Pérez-Vizcayno, M.J., Kang, D.-Y., Degenhardt, R., Pleva, L., Baan, J., Cuesta, J., Park, D.-W., Kukla, P., Jiménez-Quevedo, P., Unverdorben, M., Gao, R., Naber, C.K., Park, S.-J., Henriques, J.P.S., Kastrati, A., Byrne, R.A.: Drug-coated balloon angioplasty versus drug-eluting stent implantation in patients with coronary stent restenosis. Journal of the American College of Cardiology **75**(21), 2664–2678 (2020)
37. Argyropoulos, C., Unruh, M.L.: Analysis of time to event outcomes in randomized controlled trials by generalized additive models. PLoS ONE **10**(4), 0123784 (2015). doi:10.1371/journal.pone.0123784
38. Carmona-Bayonas, A., Jiménez-Fonseca, P., Lamarca, Á., Barriuso, J., Castaño, Á., Benavent, M., Alonso, V., Riesco-Martínez, M.d.C., 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, J.Á., Feliciangeli, E., Gallego, J., Llanos, M., Segura, Á., Vilardell, F., Percovich, J.C., Grande, E., Capdevila, J., Valle, J.W., García-Carbonero, R.: 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**(28), 2571–2580 (2019)
39. Wey, A., Hart, A., Salkowski, N., Skeans, M., Kasiske, B.L., Israni, A.K., Snyder, J.J.: Posttransplant outcome assessments at listing: Long-term outcomes are more important than short-term outcomes. American Journal of Transplantation **20**(10), 2813–2821 (2020)
40. Duchateau, L., Janssen, P., Kezic, I., Fortpied, C.: Evolution of recurrent asthma event rate over time in frailty models. Journal of the Royal Statistical Society: Series C (Applied Statistics) **52**(3), 355–363 (2003)
41. Iacobelli, S., Carstensen, B.: Multiple time scales in multi-state models. Statistics in Medicine **32**(30), 5315–5327 (2013). doi:10.1002/sim.5976
42. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. Journal of Statistical Software **67**(1), 1–48 (2015). doi:10.18637/jss.v067.i01
43. Lin, X., Zhang, D.: Inference in generalized additive mixed modelsby using smoothing splines. Journal of the royal statistical society: Series b (statistical methodology) **61**(2), 381–400 (1999)
44. Wand, M.P.: Smoothing and mixed models. Computational statistics **18**(2), 223–249 (2003)
45. Wood, S.N.: Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association **99**(467), 673–686 (2004)
46. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020). R Foundation for Statistical Computing. https://www.R-project.org/
47. RStudio Team: RStudio: Integrated Development Environment for R. RStudio, PBC., Boston, MA (2020). RStudio, PBC. http://www.rstudio.com/
48. Bender, A., Scheipl, F.: Pammtools: Piece-wise Exponential Additive Mixed Modeling tools. arXiv:1806.01042 [stat], ??? (2018)
49. Wood, S.N.: Thin-plate regression splines. Journal of the Royal Statistical Society (B) **65**(1), 95–114 (2003)
50. Wood, S.N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) **73**(1), 3–36 (2011)
51. Wood, S.N., N., Pya, Säfken, B.: Smoothing parameter and model selection for general smooth models (with discussion). Journal of the American Statistical Association **111**, 1548–1575 (2016)
52. Wood, S.N.: Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. Biometrics **62**(4), 1025–1036 (2006)
53. Wood, S.N., Li, Z., Shaddick, G., Augustin, N.H.: Generalized additive models for gigadata: modeling the uk black smoke network daily data. Journal of the American Statistical Association **112**(519), 1199–1210 (2017)
54. Greven, S., Scheipl, F.: Comments on: Inference and computation with generalized additive models and their extensions. TEST **29**(2), 343–350 (2020)
55. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team: nlme: Linear and Nonlinear Mixed Effects Models. (2020). R package version 3.1-149. https://CRAN.R-project.org/package=nlme
56. Wood, S., Scheipl, F.: Gamm4: Generalized Additive Mixed Models Using 'mgcv' and 'lme4'. (2020). R package version 0.2-6. https://CRAN.R-project.org/package=gamm4
57. Pedersen, E.J., Miller, D.L., Simpson, G.L., Ross, N.: Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ **7**, 6876 (2019). doi:10.7717/peerj.6876
58. Bürkner, P.-C.: brms: An r package for bayesian multilevel models using stan. Journal of Statistical Software **80**(1), 1–28 (2017). doi:10.18637/jss.v080.i01
59. Gasparini, A., Clements, M.S., Abrams, K.R., Crowther, M.J.: Impact of model misspecification in shared frailty survival models. Statistics in medicine **38**(23), 4477–4502 (2019)
60. Liu, X.-R., Pawitan, Y., Clements, M.S.: Generalized survival models for correlated time-to-event data. Statistics in medicine **36**(29), 4743–4762 (2017)
61. Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics) **54**(3), 507–554 (2005)
62. Coupé, C.: 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 (2018)

Figure Legends

[width=16cm, height=6cm]staph-base.pdf

Figure 1 The hazard rates for (**A**) first and (**B**) recurrent SA infections respectively. 95% confidence intervals are shaded in gray.

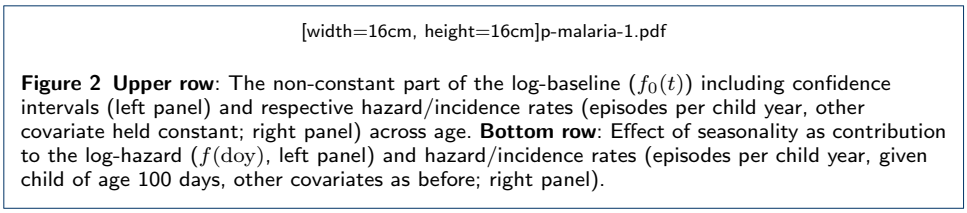

Figure 2 Upper row: The non-constant part of the log-baseline ($f_0(t)$) including confidence intervals (left panel) and respective hazard/incidence rates (episodes per child year, other covariate held constant; right panel) across age. **Bottom row**: Effect of seasonality as contribution to the log-hazard ($f(\text{doy})$, left panel) and hazard/incidence rates (episodes per child year, given child of age 100 days, other covariates as before; right panel).

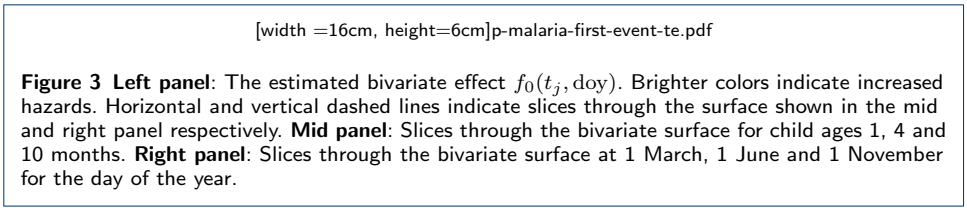

Figure 3 Left panel: The estimated bivariate effect $f_0(t_j, \text{doy})$. Brighter colors indicate increased hazards. Horizontal and vertical dashed lines indicate slices through the surface shown in the mid and right panel respectively. **Mid panel**: Slices through the bivariate surface for child 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.

Tables

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

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

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 2 Recurrent event data in PEM data format on the gap 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	$\log(0.4)$

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

Additional Files

Additional file 1

The simulation study supplement, showing the simulation procedure and the analysis of the simulated data.

Additional file 2

The R code for the staphylococcus aureus infection analysis (Example 1).

Additional file 3
The R code for the childhood malaria analysis (Example 2).

Figures

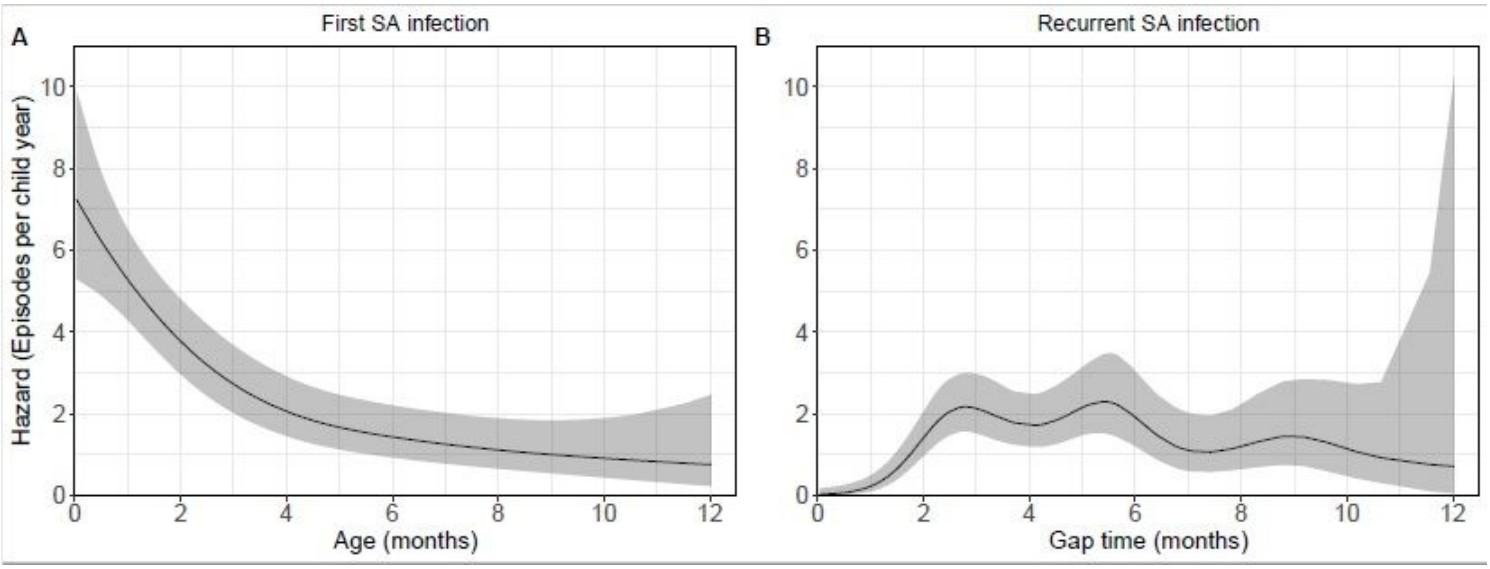

Figure 1

The hazard rates for (A) first and (B) recurrent SA infections respectively. 95% confidence intervals are shaded in gray.

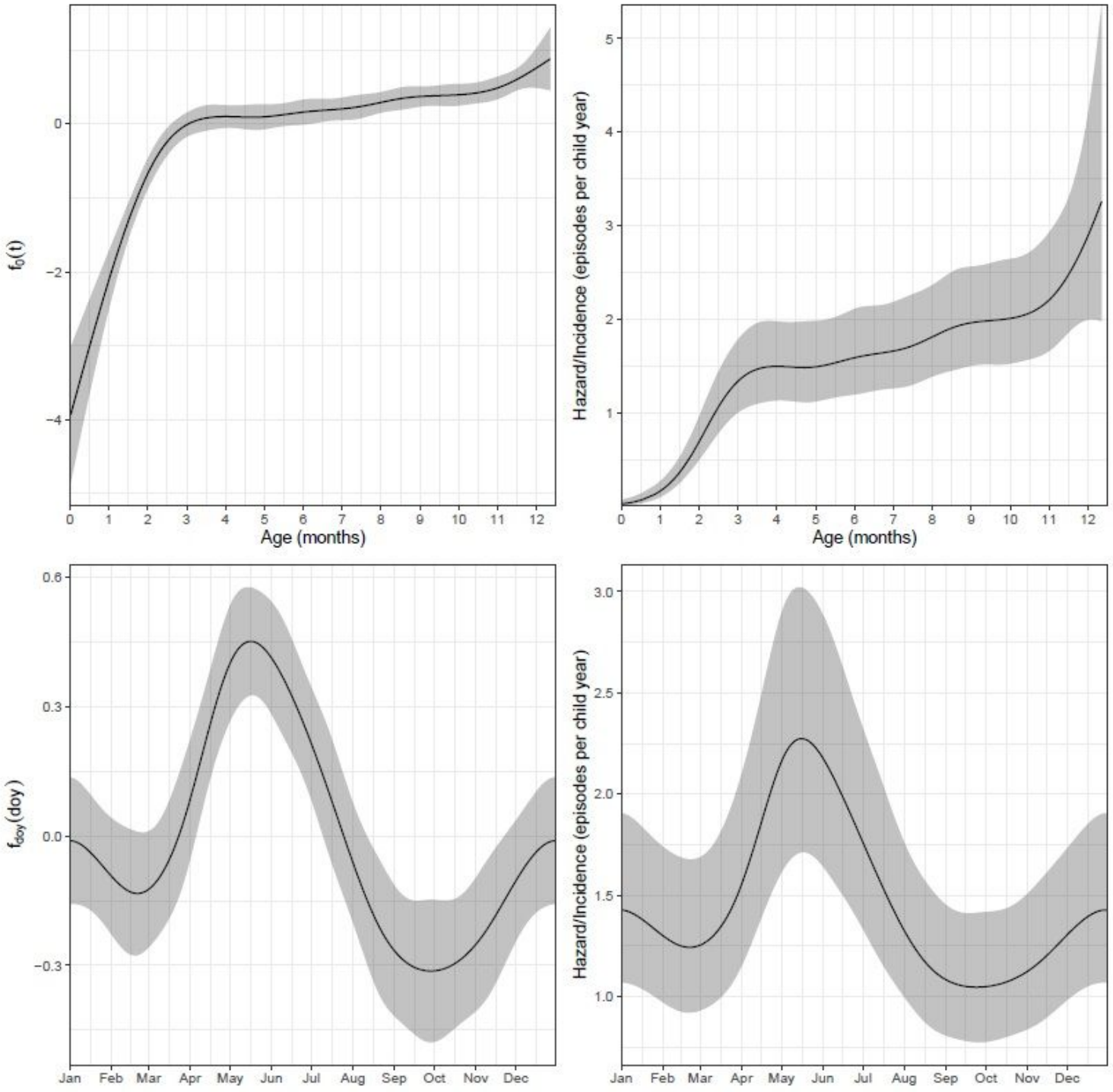

Figure 2

Upper row: The non-constant part of the log-baseline (f0(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).

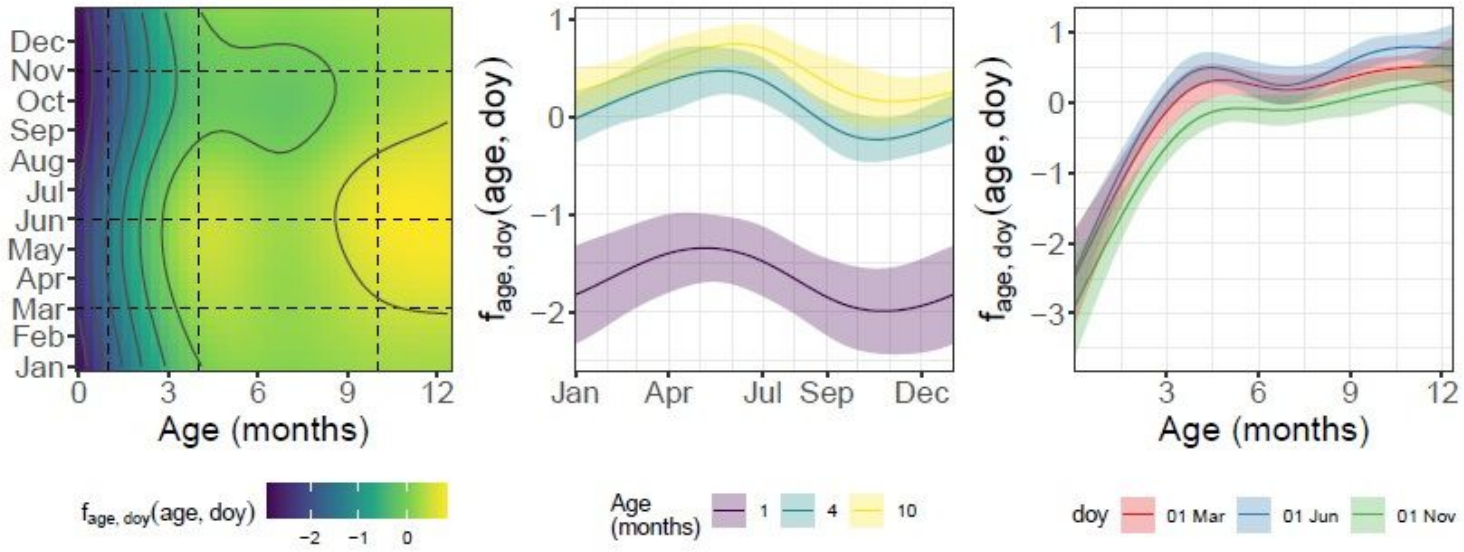

Figure 3

Left panel: The estimated bivariate effect f0(tj ; doy). 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile2.pdf
- Additionalfile3.pdf