Recurrence of staphylococcus aureaus infection in children

Jordache Ramjith, Andreas Bender, Kit C.B. Roes & Marianne A. Jonker

Introduction

We illustrate application of PAMMs in the analysis of the effect of HIV exposure on the time to staphylococcus aureaus infection in children, with possible recurrences. This guide is setup to make the code used for the manuscript results available. The children are anonymized and random `id`s were made.

Packages

The PAMM application requires the installation and loading of two specific user-written packages. These are the **`pammtools`** and **`mgcv`** packages. The pammtools package includes the data augmentation function that can restructure your data into the required structure for piece-wise exponential models, and some utility functions that can easily provide estimates of the hazards, cumulative hazards and survival probabilities, which can be used for visualization. See the package page for more details. The package **`mgcv`** is used for building the actual PAMM. We also use additional packages for data wrangling and visualization.

```
library(mgcv)
library(pammtools)
library(dplyr) #data wrangling
library(data.table) #data wrangling
library(ggplot2) #visualization
library(ggpubr) #visualization
library(readxl)
```

Example data

The data set (`staph`), is read in below and has 374 observations from 137 children (from the Drakenstein child health study) with a maximum of 6 recurrences. The `staph` data is in longitudinal format reflecting the recurrences for children over different rows. Here

- `t.start` and `t.stop` indicate the entry and exit time into the risk set for the respective recurrences
- `event` indicates whether the k-th recurrence was observed (1 = yes, 0 = censored for the k-th recurrence)
- `enum` is the event number k
- `HIVexposure` indicates whether the mother of the child was HIV positive (1 = yes, 0 = no)

```
staph <- read_excel("staph.xlsx")
```

The data for the first two children are

```
staph %>% filter(id %in% c("1", "2"))
```

```
## # A tibble: 7 x 6
##      id t.start t.stop event  enum   hiv
##   <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1     1       0     16     1     1     0
## 2     1      16    324     1     2     0
## 3     1     324    365     0     3     0
```

```
## 4      2        0      59      1      1      1
## 5      2       59     157      1      2      1
## 6      2      157     227      1      3      1
## 7      2      227     368      0      4      1
```

Piece-wise exponential data

In order to apply PAMMs, we first transform the data to the *piece-wise exponential data* (PED) format (see here for details). We use the `as_ped()` function in **`pammtools`**. To read the help with this function, use the command `?as_ped()`. For the analyses of these data, the timescale we use is gap-time.

The individual inputs are given as follows:

- `formula`: specifies the `Surv` object on the left hand side which contains information about the risk set entry and exit times as well as the event indicator; and the variables that should be retained in the data set after data transformation. Note that the variables `id` and `enum` will be retained in the data without specification.
- `id`: specifies the variable in the data set the indicates indivual subjects
- `data`: the data to be transformed
- `transition`: the variable that indicates transitions from one state to another (here state transitions are transtions from event number $k-1$ to k)
- `timescale`: the time scale of the ouput data (defaults to gap time)
- `max_time`: The maximum time considered. All observations with $t > max\_time$ will be set to `max_time` and their event indicator set to 0. Here we restrict the follow-up to `366` days, as it marks one year under observation and few children were under observation beyond that time.
- `cut`: This argument is unspecified her, but could be used to control the time points at which the follow-up is partitioned. If unspecified all unique event times are used.

```
ped <- as_ped(
  formula     = Surv(t.start,t.stop,event) ~ hiv,
  id          = "id",
  data        = staph,
  transition  = "enum",
  timescale   = "gap",
  max_time    = 366)
```

The resulting data for the first two infants is indicated below (we show the first an last observation of each infant for each event number they were at risk):

```
ped %>%
  filter(id %in% c("1", "2")) %>%
  group_by(id, enum) %>%
  mutate(offset=round(offset,2)) %>%
  slice(1, n()) %>%
  knitr::kable()
```

id	tstart	tend	interval	offset	ped_status	hiv	enum
1	0	1	(0,1]	0.00	0	0	1
1	15	16	(15,16]	0.00	1	0	1
1	0	1	(0,1]	0.00	0	0	2
1	295	308	(295,308]	2.56	1	0	2
1	0	1	(0,1]	0.00	0	0	3
1	40	41	(40,41]	0.00	0	0	3
2	0	1	(0,1]	0.00	0	1	1
2	57	59	(57,59]	0.69	1	1	1
2	0	1	(0,1]	0.00	0	1	2

id	tstart	tend	interval	offset	ped_status	hiv	enum
2	97	98	(97,98]	0.00	1	1	2
2	0	1	(0,1]	0.00	0	1	3
2	69	70	(69,70]	0.00	1	1	3
2	0	1	(0,1]	0.00	0	1	4
2	140	141	(140,141]	0.00	0	1	4

Baseline model

Stratification by event number

We first model the baseline hazards over time. Biologically, the infection incidence in gap time may be different for the first event compared with the recurrences. Statistically, estimation of the baseline hazard for each of the event numbers is not useful/feasible since only a few subjects experienced more than 3 events. So we will create a new variable to indicate whether the event a child is at risk for is the first event or a recurrent event. Note, however, that we do this after PED data transformation and use the full information to create the PED data. We only use `enum2` for stratification when estimating the baseline hazard.

```
ped <- ped %>%
  mutate(enum2 =as.factor(ifelse(enum>1,"recurrent","first")),
         id=as.factor(id))
```

Including random effects/frailties

Additionally to model subject specific random effects using `mgcv`, we must ensure that `id` is a factor variable and not numeric (we have done this in the previous chunk of code).

Fitting the model

In the code chunk below we use the `pamm` function to fit a Piecewise exponential Additive Mixed Model (PAMM), which is a wrapper around `mgcv::gam` or `mgcv::bam`, depending of the specification of the `engine` argument. The other arguments of the functions are directly passed to these functions (with `family = poisson()` and `offset` set to the `offset` variable in the `ped` data set for convenience).

In the formula specification of the model

- `s(tend,by=enum2)` indicates the smooth effect that estimates the deviation of the log baseline hazard over time (`tend`) from the estimated intercept, by using `by=enum2` inside `s()` and including `enum2` in the model, we are modelling stratified smooth functions for first and recurrent events respectively, and
- `s(id, bs = "re")` indicates a random effect (frailty) for each child where random effects basis are specified for id `bs="re"`, to allow child-specific Gaussian distributed random effects.

The `s()` functions are smooth functions with thin-plate splines as the default basis functions and 10 as the default degrees of freedom. We recommend reading more about the different possibilities for basis splines in the `mgcv` package documentation.

```
pam0 <- pamm(
  ped_status~ enum2+s(tend,by=enum2)+s(id,bs="re"),
  data = ped,
  engine   = "bam",
  method   = "fREML",
  discrete = TRUE)

summary(pam0)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## ped_status ~ enum2 + s(tend, by = enum2) + s(id, bs = "re")
##
## Parametric coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.9287     0.1213  -40.62  < 2e-16 ***
## enum2recurrent  -1.0991     0.1857   -5.92 3.23e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf  Ref.df Chi.sq p-value
## s(tend):enum2first       1.995   2.487  35.69 < 2e-16 ***
## s(tend):enum2recurrent   6.101   7.240  33.49 3.7e-05 ***
## s(id)                   14.052 136.000  15.17   0.249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  -0.0265   Deviance explained = -0.97%
## fREML =  18453  Scale est. = 1         n = 18889
```

The resulting output is separated into two parts, one for the "parametric coefficients" and one for the "smooth terms". The intercept is the average log baseline hazard rate. The parametric coefficients for covariates are the estimated average log hazard ratios over time. The $s(tend)$ term in the smooth terms part of the output corresponds with the $f_0(t_j)$ function in equation (9) in the manuscript, which tells us how the log baseline hazard deviates from the estimated average (the intercept) over time. The estimated degrees of freedom (edf) gives us an idea of how "wiggly" the respective smooth functions are, and not the "strength" of these effects, and the p-values test whether these are different from a flat line (see package documentation). For the random effects terms, we report their estimated variances and p-values. From the output, we see that the random effects are not statistically significant ($p = 0.249$) and not necessary in the model. The standard deviation of the random effects can be found using the command:

```
gam.vcomp(pam0)
```

```
##
## Standard deviations and 0.95 confidence intervals:
##
##                             std.dev        lower       upper
## s(tend):enum2first     0.0005741588 8.308116e-05 0.003967907
## s(tend):enum2recurrent 0.0062687807 2.563565e-03 0.015329284
## s(id)                  0.2565797067 4.015591e-02 1.639438617
##
## Rank: 3/3
```

Excluding random effects/frailties

Because we showed that the frailty variance was small and not statistically significant, it may be better for interpretation to fit a simpler model without the random effects.

Fitting the model

```
pam0 <- pamm(
  ped_status~ enum2+s(tend,by=enum2),
  data = ped,
  engine   = "bam",
  method   = "fREML",
  discrete = TRUE)

summary(pam0)
```

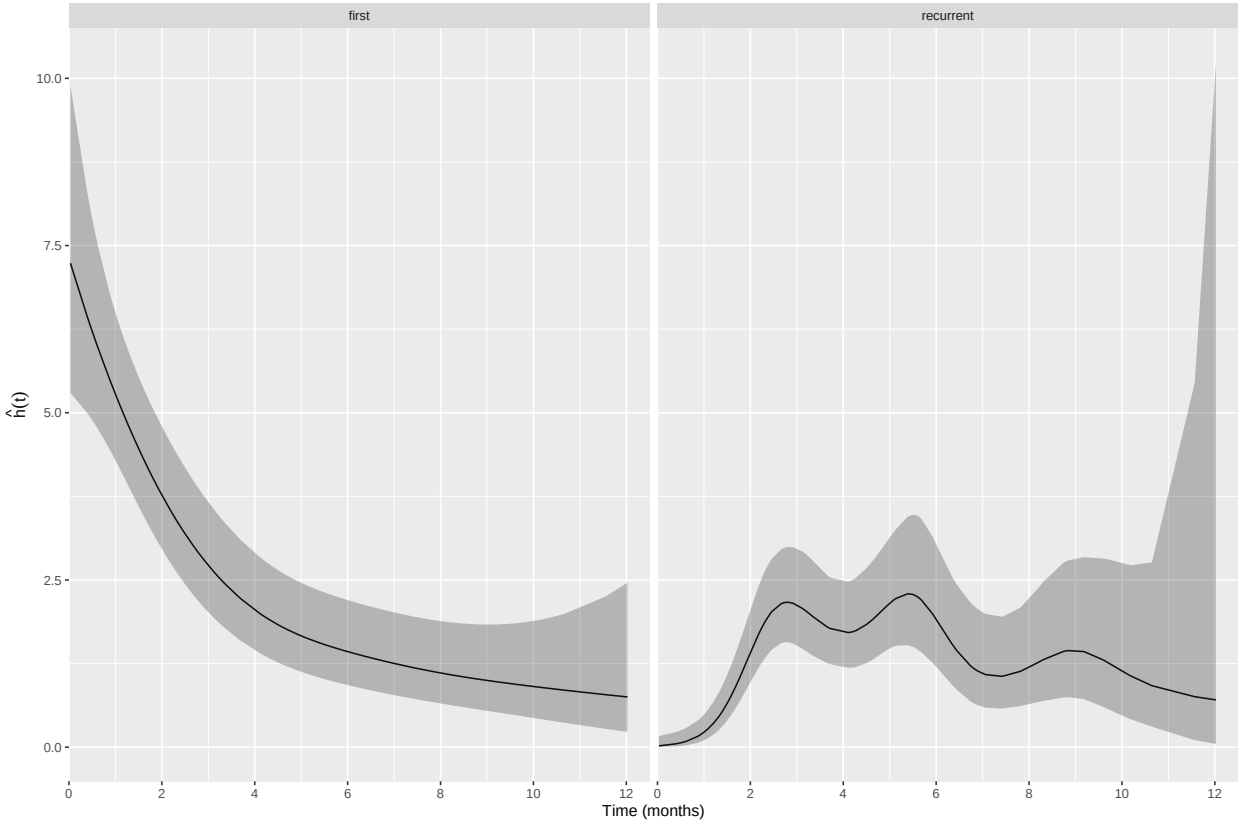
```
##
## Family: poisson
## Link function: log
##
## Formula:
## ped_status ~ enum2 + s(tend, by = enum2)
##
## Parametric coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -5.0624     0.1231 -41.124  < 2e-16 ***
## enum2recurrent  -0.8860     0.1801  -4.919 8.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq  p-value
## s(tend):enum2first      2.099  2.616  41.71  < 2e-16 ***
## s(tend):enum2recurrent 6.099  7.238  33.07 3.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  -0.0276   Deviance explained = -2.34%
## fREML =  18453  Scale est. = 1         n = 18889
```

The edf for the baseline log-hazard rate over time is ≈ 2 and ≈ 6 for the first and recurrent events respectively, both with statistically significant p-values (< 0.001) indicating sufficient evidence of a log hazard rate that is not constant over time for both first and recurrent events. The intercept is estimated as -5.1 which means that the geometric mean baseline hazard is $\approx \exp(-5.1) = 0.0061$ new first infections per child day. The coefficient for recurrences is -0.9, which means that the geometric mean baseline hazard is $\approx \exp(-5.1 - 0.9) = \exp(-6) = 0.0025$ new recurrent infections per child day. Multiplying by 365.25, this is ≈ 2.23 new first infections per child year over the first year of life and ≈ 0.91 new recurrent infections per child year over the first year of life. For visualization, we also wanted to visualize the hazards in terms of episodes per child-year, so we simply multiplied by 365.25.

Estimates over time and visualization

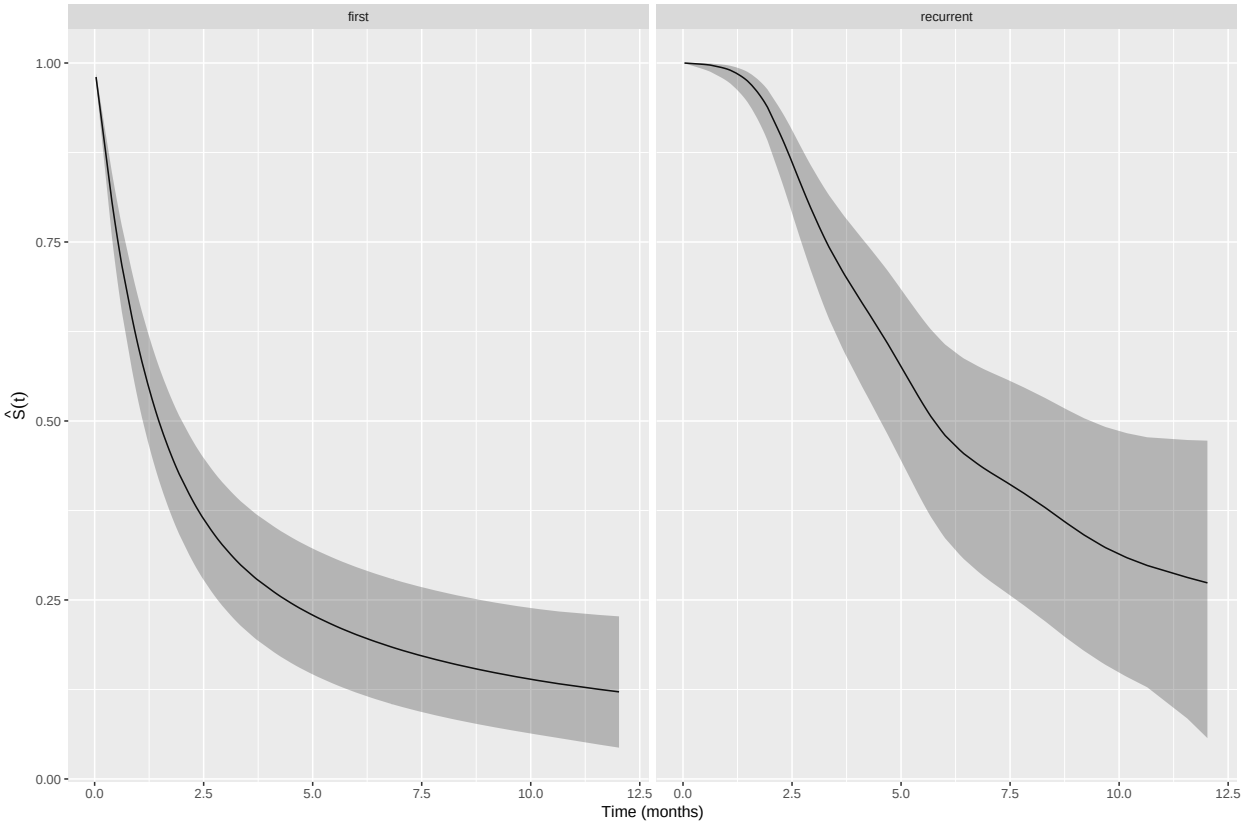
To visualize these results - especially the way in which the hazard/incidence evolves over time, we could create a new dataset (using the `make_newdata()` convenience function in **`pammtools`**) to find the estimates we are interested in. This new dataset must include a value for all variables used in the model and the correct variable names. In the `pam0` model above, the variables used were the time `tend` and the created event number `enum2` variable. We can use the `add_hazard()` convenience function from **`pammtools`** to calculate the hazard for the data provided.

```
newdata<- ped %>% make_newdata(tend = unique(tend),
                        enum2 = unique(enum2)) %>%
          group_by(enum2) %>%
          add_hazard(pam0, type = "response")

ggplot(newdata, aes(x = tend/(365.25/12), y = hazard*365.25)) +
  geom_line() +
  geom_ribbon(aes(ymin  = ci_lower*365.25, ymax = ci_upper*365.25), alpha = .3) +
  ylab(expression(hat(h)(t))) + xlab("Time (months)") +
  scale_x_continuous(limits = c(0, 12.5),breaks=seq(0,12,2),expand=c(0,0)) +
  facet_wrap(~enum2)
```



Similarly, we can also calculate the survival probabilities and visualize.

```
newdata <-newdata %>% add_surv_prob(pam0)

ggplot(newdata, aes(x = tend/(365.25/12), y = surv_prob)) +
  geom_line() +
  geom_ribbon(aes(ymin  = surv_lower, ymax = surv_upper), alpha = .3) +
  ylab(expression(hat(S)(t))) + xlab("Time (months)") +
  scale_x_continuous(limits = c(0, 12.1)) +
  facet_wrap(~enum2)
```



Modelling the effects of HIV assuming proportional hazards

The HIV exposure variable indicates whether children were HIV exposed and uninfected (HEU) by being born to HIV positive mothers or HIV uninfected (HU). We will fit this model in the PAMM framework to evaluate the effect of HIV exposure while assuming proportional hazards. This means that the effects of HIV act to shift the log-hazard by some constant over time. We fit two models. In the first model, we assume that the hazard ratio is the same for first and recurrent infections. In the second model, we allow different hazard ratios for first and recurrent infections, but we still assume both these hazard ratios are proportional over time.

Model 1

```
pam1 <- pamm(ped_status ~ enum2 + s(tend, by=enum2)+hiv,
             data = ped,
             engine   = "bam",
             method   = "fREML",
             discrete = TRUE)

summary(pam1)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## ped_status ~ enum2 + s(tend, by = enum2) + hiv
```

```
##
## Parametric coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -5.1351     0.1303 -39.416  < 2e-16 ***
## enum2recurrent  -0.8779     0.1801  -4.874 1.09e-06 ***
## hiv              0.2681     0.1434   1.870   0.0615 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq  p-value
## s(tend):enum2first      2.121  2.643  41.78  < 2e-16 ***
## s(tend):enum2recurrent 6.084  7.224  33.22 3.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  -0.0279   Deviance explained = -2.18%
## fREML =  18453  Scale est. = 1         n = 18889
```

We can see that estimated effect of HIV exposure in terms of the hazard ratio is $HR = \exp(0.2681) = 1.31$ ($p = 0.062$). A simple way to calculate the HR and 95% confidence intervals is shown in the code chunk below.

```
ped %>%
  make_newdata(hiv=c(1)) %>%
  add_hazard(
    pam1,
    reference = list(hiv = c(0))) %>%
  select(hazard,ci_lower,ci_upper)
```

```
##     hazard  ci_lower ci_upper
## 1 1.307541 0.9814742 1.741934
```

Model 2

```
pam2 <- pamm(ped_status ~ enum2+s(tend, by=enum2)+hiv:enum2,
             data = ped,
             engine   = "bam",
             method   = "fREML",
             discrete = TRUE)

summary(pam2)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## ped_status ~ enum2 + s(tend, by = enum2) + hiv:enum2
##
## Parametric coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.05753    0.13330 -37.940  < 2e-16 ***
## enum2recurrent     -1.03583    0.19753  -5.244 1.57e-07 ***
## enum2first:hiv     -0.02007    0.21300  -0.094  0.92493
```

```
## enum2recurrent:hiv  0.54272    0.19553   2.776  0.00551 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq  p-value
## s(tend):enum2first     2.098  2.615  41.71  < 2e-16 ***
## s(tend):enum2recurrent 6.067  7.208  33.42 3.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  -0.0274   Deviance explained = -2.01%
## fREML =  18451  Scale est. = 1         n = 18889
```

We can see that estimated effect of HIV exposure in terms of the hazard ratio is $HR = \exp(-0.0201) = 0.98$ ($p = 0.925$) for the first infection and $HR = \exp(0.5427) = 1.72$ ($p = 0.006$) for recurrent infections.

```
ped %>%
  make_newdata(hiv=c(1),enum2=unique(enum2)) %>%
  add_hazard(
    pam2,
    reference = list(hiv = c(0))) %>%
  select(enum2,hazard,ci_lower,ci_upper)
```

```
##       enum2    hazard ci_lower ci_upper
## 1     first 0.9801305 0.640137 1.500704
## 2 recurrent 1.7206799 1.163762 2.544110
```

These models can also easily be fitted using the popular survival functions in `R` like `coxph()`. Here we showed how they can be fitted through the PAMM. In Example 2 - the childhood malaria incidence example, we show how to fit more complex models including non-linear effects of seasonality, time-varying effects and non-linear effects that possibly vary over time.