Joint longitudinal and time-to-event models via Stan

Sam Brilleman[1,2], Michael J. Crowther[3], Margarita Moreno-Betancur[2,4,5], Jacqueline Buros Novik[6], Rory Wolfe[1,2]

StanCon 2018

Pacific Grove, California, USA

10-12th January 2018

[1] Monash University, Melbourne, Australia

[2] Victorian Centre for Biostatistics (ViCBiostat)

[3] University of Leicester, Leicester, UK

[4] Murdoch Childrens Research Institute, Melbourne, Australia

[5] University of Melbourne, Melbourne, Australia

[6] Icahn School of Medicine at Mount Sinai, New York, US

MONASH University

ViCBiostat

Outline

- Context and background
- Joint model formulation
- Association structures
- Software implementation via Stan / rstanarm
- Example application

Context

- Suppose we observe **repeated measurements** of a **clinical biomarker** on a group of individuals
- May be clinical trial patients or some observational cohort

Collection of **serum bilirubin** and **serum albumin** from patients with liver disease

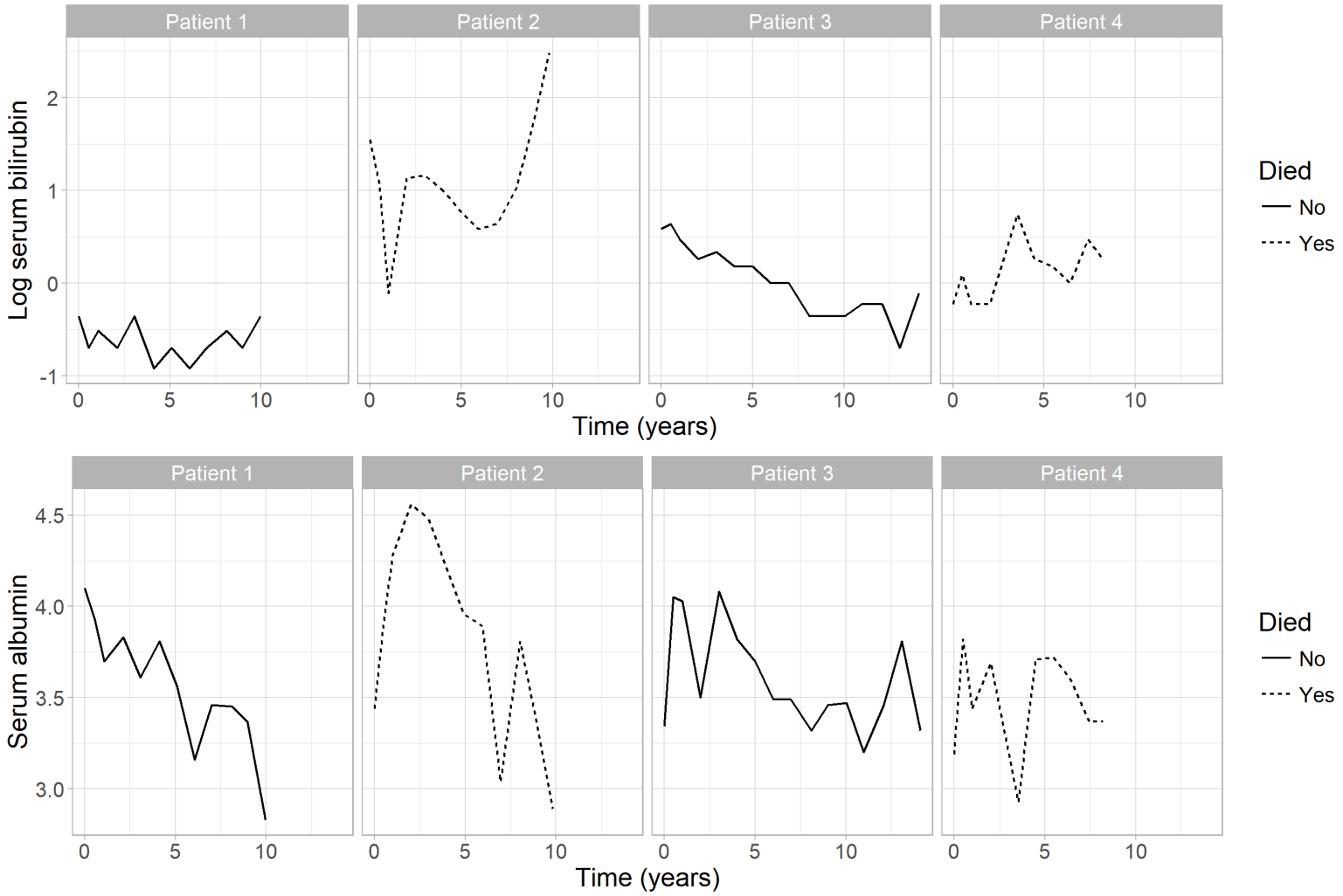
Context

- Suppose we observe **repeated measurements** of a **clinical biomarker** on a group of individuals
- May be clinical trial patients or some observational cohort

Collection of **serum bilirubin** and **serum albumin** from patients with liver disease

- In addition we observe the **time to some event** endpoint, e.g. death

Longitudinal and time-to-event data



What is “joint modelling” of longitudinal and time-to-event data?

- Treats both the longitudinal biomarker(s) and the event as outcome data
- Each outcome is modelled using a distinct regression submodel:
  - A (multivariate) **mixed effects model** for the longitudinal outcome(s)
  - A **proportional hazards model** for the time-to-event outcome
- The regression submodels are linked through **shared individual-specific parameters** and **estimated simultaneously** under a joint likelihood function

Why use “joint modelling”?

- Want to understand **whether (some function of) the longitudinal outcome is associated with the risk of the event** (i.e. epidemiological questions)
  - Joint models offer advantages over just using the biomarker as a time-varying covariate (described in the next slide!)
- Want to develop a **dynamic prognostic model**, where predictions of event risk can be updated as new longitudinal biomarker measurements become available (i.e. clinical risk prediction)
- Possibly other reasons:
  - e.g. adjusting for informative dropout, separating out “direct” and “indirect” effects of treatment

Joint model formulation

$y_{ijm}(t)$ is the value at time t of the
m^{th} longitudinal marker ($m = 1, \dots, M$)
for the i^{th} individual ($i = 1, \dots, N$)
at the j^{th} time point ($j = 1, \dots, n_{im}$)
T_i^* is "true" event time, C_i is the censoring time
$T_i = \min(T_i^*, C_i)$ and $d_i = I(T_i^* \leq C_i)$

- Longitudinal submodel

$y_{ijm}(t)$ follows a distribution in the exponential family with expected value $\mu_{ijm}(t)$ and

$$\eta_{ijm}(t) = g_m\left(\mu_{ijm}(t)\right) = \boldsymbol{x_{ijm}^T}(t)\boldsymbol{\beta_m} + \boldsymbol{z_{ijm}^T}(t)\boldsymbol{b_{im}}$$

$$\begin{bmatrix} \boldsymbol{b_{i1}} \\ \vdots \\ \boldsymbol{b_{iM}} \end{bmatrix} = \boldsymbol{b_i} \sim N(0, \boldsymbol{\Sigma})$$

- Event submodel

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w_i^T}(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \alpha_m \, \mu_{im}(t) \right)$$

Joint model formulation

$y_{ijm}(t)$ is the value at time t of the
m^{th} longitudinal marker ($m = 1, \dots, M$)
for the i^{th} individual ($i = 1, \dots, N$)
at the j^{th} time point ($j = 1, \dots, n_{im}$)
T_i^* is "true" event time, C_i is the censoring time
$T_i = \min(T_i^*, C_i)$ and $d_i = I(T_i^* \leq C_i)$

- Longitudinal submodel

$y_{ijm}(t)$ follows a distribution in the exponential family with expected value $\mu_{ijm}(t)$ and

$$\eta_{ijm}(t) = g_m\left(\mu_{ijm}(t)\right) = \boldsymbol{x_{ijm}^T}(t)\boldsymbol{\beta_m} + \boldsymbol{z_{ijm}^T}(t)\boldsymbol{b_{im}}$$

$$\begin{bmatrix} \boldsymbol{b_{i1}} \\ \vdots \\ \boldsymbol{b_{iM}} \end{bmatrix} = \boldsymbol{b_i} \sim N(0, \boldsymbol{\Sigma})$$

- Event submodel

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w_i^T}(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \alpha_m \, \mu_{im}(t) \right)$$

- Known as a **current value "association structure"**

Joint model formulation

$y_{ijm}(t)$ is the value at time t of the
m^{th} longitudinal marker ($m = 1, \dots, M$)
for the i^{th} individual ($i = 1, \dots, N$)
at the j^{th} time point ($j = 1, \dots, n_{im}$)
T_i^* is "true" event time, C_i is the censoring time
$T_i = \min(T_i^*, C_i)$ and $d_i = I(T_i^* \leq C_i)$

- Longitudinal submodel

$y_{ijm}(t)$ follows a distribution in the exponential family with expected value $\mu_{ijm}(t)$ and

$$\eta_{ijm}(t) = g_m\left(\mu_{ijm}(t)\right) = \boldsymbol{x_{ijm}^T}(t)\boldsymbol{\beta_m} + \boldsymbol{z_{ijm}^T}(t)\boldsymbol{b_{im}}$$

$$\begin{bmatrix} \boldsymbol{b_{i1}} \\ \vdots \\ \boldsymbol{b_{iM}} \end{bmatrix} = \boldsymbol{b_i} \sim N(0, \boldsymbol{\Sigma})$$

- Event submodel

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w_i^T}(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \alpha_m \, \mu_{im}(t) \right)$$

$y_{ijm}(t)$ is both:
- error-prone
- measured at discrete times

Whereas $\mu_{im}(t)$ is both:
- error-free
- modelled in continuous time

Therefore less bias in α_m compared with a time-dependent Cox model.

- Known as a **current value "association structure"**

Association structures

- A more **general form** for the event submodel is

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w_i^T}(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \sum_{q=1}^{Q_m} \alpha_{mq} f_{mq}(\boldsymbol{\beta_m}, \boldsymbol{b_{im}}; t) \right)$$

Association structures

- A more **general form** for the event submodel is

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w_i^T}(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \sum_{q=1}^{Q_m} \alpha_{mq} f_{mq}(\boldsymbol{\beta_m}, \boldsymbol{b_{im}}; t) \right)$$

- This posits an **association** between the **log hazard of the event** and **any function of the longitudinal submodel parameters**

Association structures

- A more **general form** for the event submodel is

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w_i^T}(t)\boldsymbol{\gamma} + \sum_{m=1}^{M} \sum_{q=1}^{Q_m} \alpha_{mq} f_{mq}(\boldsymbol{\beta_m}, \boldsymbol{b_{im}}; t) \right)$$

- This posits an **association** between the **log hazard of the event** and **any function of the longitudinal submodel parameters**; for example, defining $f_{mq}(.)$ as:

$\eta_{im}(t)$ ⟶ Linear predictor (or expected value of the biomarker) at time t

$\frac{d\eta_{im}(t)}{dt}$ ⟶ Rate of change in the linear predictor (or biomarker) at time t

$\int_0^t \eta_{im}(s)\, ds$ ⟶ Area under linear predictor (or biomarker trajectory), up to time t

$\eta_{im}(t-u)$ ⟶ Lagged value (for some lag time u)

Joint modelling software

- An abundance of **methodological** developments in joint modelling
- But not all methods have been translated into **"user-friendly" software**

- Well established software for one longitudinal outcome
  - e.g. stjm (Stata); joineR, JM, JMbayes, frailtypack (R); JMFit (SAS)

- Recent software developments for **multiple longitudinal outcomes**
  - R packages: **rstanarm**, joineRML, JMbayes, survtd

- Each package has its strengths and limitations
  - e.g. (non-)normally distributed longitudinal outcomes, selected association structures, speed, etc.

Joint modelling software

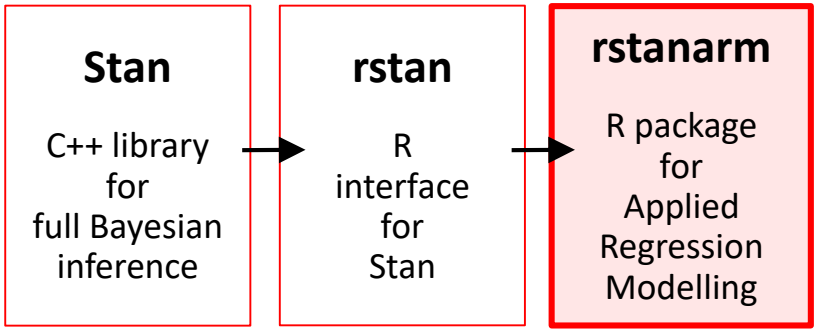
- An abundance of **methodological** developments in joint modelling
- But not all methods have been translated into **"user-friendly" software**

- Well established software for one longitudinal outcome
  - e.g. stjm (Stata); joineR, JM, JMbayes, frailtypack (R); JMFit (SAS)

- Recent software developments for **multiple longitudinal outcomes**
  - R packages: **rstanarm**, joineRML, JMbayes, survtd

- Each package has its strengths and limitations
  - e.g. (non-)normally distributed longitudinal outcomes, selected association structures, speed, etc.

Bayesian joint models via Stan



- Included in **rstanarm** version ≥ 2.17.2
  - https://cran.r-project.org/package=rstanarm
  - https://github.com/stan-dev/rstanarm
- Can specify **multiple longitudinal outcomes**
- Allows for **multilevel** clustering in longitudinal submodels (e.g. time < patients < clinics)
- Variety of **families** (and link functions) for the longitudinal outcomes
  - e.g. normal, binomial, Poisson, negative binomial, Gamma, inverse Gaussian
- Variety of **association structures**
- Variety of **prior distributions**
  - Regression coefficients: normal, student t, Cauchy, shrinkage priors (horseshoe, lasso)
- **Posterior predictions** – including "dynamic predictions" of event outcome
- Baseline hazard
  - B-splines regression, Weibull, piecewise constant

Application to the PBC dataset

- Data contains 312 **liver disease patients** who participated in a clinical trial at the Mayo Clinic between 1974 and 1984
- Secondary analysis to explore whether **log serum bilirubin** and **serum albumin** are associated with risk of **mortality**
- Longitudinal submodel:
  - **Linear mixed model for each biomarker**
  - w/ patient-specific intercept and linear slope (i.e. random effects)
- Event submodel:
  - Gender included as a baseline covariate
  - **Current value** association structure (i.e. expected value of each biomarker)
  - B-splines baseline hazard

```
> fit1 <- stan_jm(
>   formulaLong = list(
>     logBili ~ year + (year | id),
>     albumin ~ year + (year | id)),
>   formulaEvent = Surv(futimeYears, death) ~ sex,
>   dataLong = pbcLong, dataEvent = pbcSurv,
>   time_var = "year", assoc = "etavalue", basehaz = "bs")
```

```
> fit1 <- stan_jm(
>   formulaLong = list(
>     logBili ~ year + (year |
>     albumin ~ year + (year |
>   formulaEvent = Surv(futime
>   dataLong = pbcLong, dataEv
>   time_var = "year", assoc =
```

```
> print(fit1)
```

```
# stan_jm
#  formula (Long1): logBili ~ year + (year | id)
#  family  (Long1): gaussian [identity]
#  formula (Long2): albumin ~ year + (year | id)
#  family  (Long2): gaussian [identity]
#  formula (Event): Surv(futimeYears, death) ~ sex
#  baseline hazard: bs
#  assoc:           etavalue (Long1), etavalue (Long2)
# ------
#
# Longitudinal submodel 1: logBili
#             Median MAD_SD
# (Intercept) 0.678  0.192
# year        0.227  0.042
# sigma       0.354  0.017
#
# Longitudinal submodel 2: albumin
#             Median MAD_SD
# (Intercept)  3.520  0.082
# year        -0.161  0.025
# sigma        0.290  0.014
#
# Event submodel:
#                 Median   MAD_SD   exp(Median)
# (Intercept)        7.054    2.870 1157.757
# sexf              -0.182    0.674    0.834
# Long1|etavalue     0.745    0.281    2.105
# Long2|etavalue    -3.141    0.857    0.043
...
# Group-level error terms:
#  Groups Name               Std.Dev. Corr
#  id     Long1|(Intercept) 1.2425
#         Long1|year        0.1937    0.50
#         Long2|(Intercept) 0.5029   -0.64 -0.51
#         Long2|year        0.1022   -0.59 -0.81  0.47
```

```
> fit1 <- stan_jm(
>   formulaLong = list(
>     logBili ~ year + (year |
>     albumin ~ year + (year |
>   formulaEvent = Surv(futime
>   dataLong = pbcLong, dataEv
>   time_var = "year", assoc =
```

```
> print(fit1)
```

```
# stan_jm
#  formula (Long1): logBili ~ year + (year | id)
#  family  (Long1): gaussian [identity]
#  formula (Long2): albumin ~ year + (year | id)
#  family  (Long2): gaussian [identity]
#  formula (Event): Surv(futimeYears, death) ~ sex
#  baseline hazard: bs
#  assoc:           etavalue (Long1), etavalue (Long2)
# ------
#
# Longitudinal submodel 1: logBili
#             Median MAD_SD
# (Intercept) 0.678  0.192
# year        0.227  0.042
# sigma       0.354  0.017
#
# Longitudinal submodel 2: albumin
#             Median MAD_SD
# (Intercept)  3.520  0.082
# year        -0.161  0.025
# sigma        0.290  0.014
#
# Event submodel:
#                   Median   MAD_SD   exp(Median)
# (Intercept)        7.054    2.870 1157.757
# sexf              -0.182    0.674    0.834
# Long1|etavalue     0.745    0.281    2.105
# Long2|etavalue    -3.141    0.857    0.043
...
# Group-level error terms:
#  Groups Name               Std.Dev. Corr
#  id     Long1|(Intercept) 1.2425
#         Long1|year         0.1937    0.50
#         Long2|(Intercept) 0.5029   -0.64 -0.51
#         Long2|year         0.1022   -0.59 -0.81  0.47
```

A one **unit increase in log serum bilirubin** is associated with an estimated **2.1-fold increase in the hazard of death**

```
> fit1 <- stan_jm(
>   formulaLong = list(
>     logBili ~ year + (year | id)
>     albumin ~ year + (year | id)
>   formulaEvent = Surv(futimeYear
>   dataLong = pbcLong, dataEvent
>   time_var = "year", assoc = "et
```

```
> print(fit1)
```

```
> summary(fit1, pars = "assoc")
```

```
# Model Info:
#
#  function:         stan_jm
#  formula (Long1): logBili ~ year + (year | id)
#  family  (Long1): gaussian [identity]
#  formula (Long2): albumin ~ year + (year | id)
#  family  (Long2): gaussian [identity]
#  formula (Event): Surv(futimeYears, death) ~ sex
#  baseline hazard: bs
#  assoc:           etavalue (Long1), etavalue (Lo
#  algorithm:       sampling
#  priors:          see help('prior_summary')
#  sample:          4000 (posterior sample size)
#  num obs:         304 (Long1), 304 (Long2)
#  num subjects:    40
#  num events:      29 (72.5%)
#  groups:          id (40)
#  runtime:         2.9 mins
#
# Estimates:
#                           mean   sd     2.5%  97.5%
# Assoc|Long1|etavalue  0.748  0.281  0.204  1.302
# Assoc|Long2|etavalue -3.204  0.903 -5.121 -1.566
#
# Diagnostics:
#                          mcse  Rhat  n_eff
# Assoc|Long1|etavalue 0.004 1.000 4000
# Assoc|Long2|etavalue 0.018 1.001 2452
```

```
> fit1 <- stan_jm(
>   formulaLong = list(
>     logBili ~ year + (year | id),
>     albumin ~ year + (year | id)),
>   formulaEvent = Surv(futimeYears, death) ~ sex,
>   dataLong = pbcLong, dataEvent = pbcSurv,
>   time_var = "year", assoc = "etavalue", basehaz = "bs")
```

```
> print(fit1)
```

```
> summary(fit1, pars = "assoc")
```

```
> p1 <- posterior_traj(fit1, m = 1, ids = 7:8, extrapolate = TRUE)
> p2 <- posterior_traj(fit1, m = 2, ids = 7:8, extrapolate = TRUE)
> p3 <- posterior_survfit(fit1, ids = 7:8)

> pp1 <- plot(p1, vline = TRUE, plot_observed = TRUE)
> pp2 <- plot(p2, vline = TRUE, plot_observed = TRUE)

> plot_stack_jm(yplot = list(pp1, pp2), survplot = plot(p3))
```

```
> fit1 <- stan_jm(
>   formulaLong = list(
>     logBili ~ year + (yea
>     albumin ~ year + (yea
>   formulaEvent = Surv(fut
>   dataLong = pbcLong, dat
>   time_var = "year", asso
```

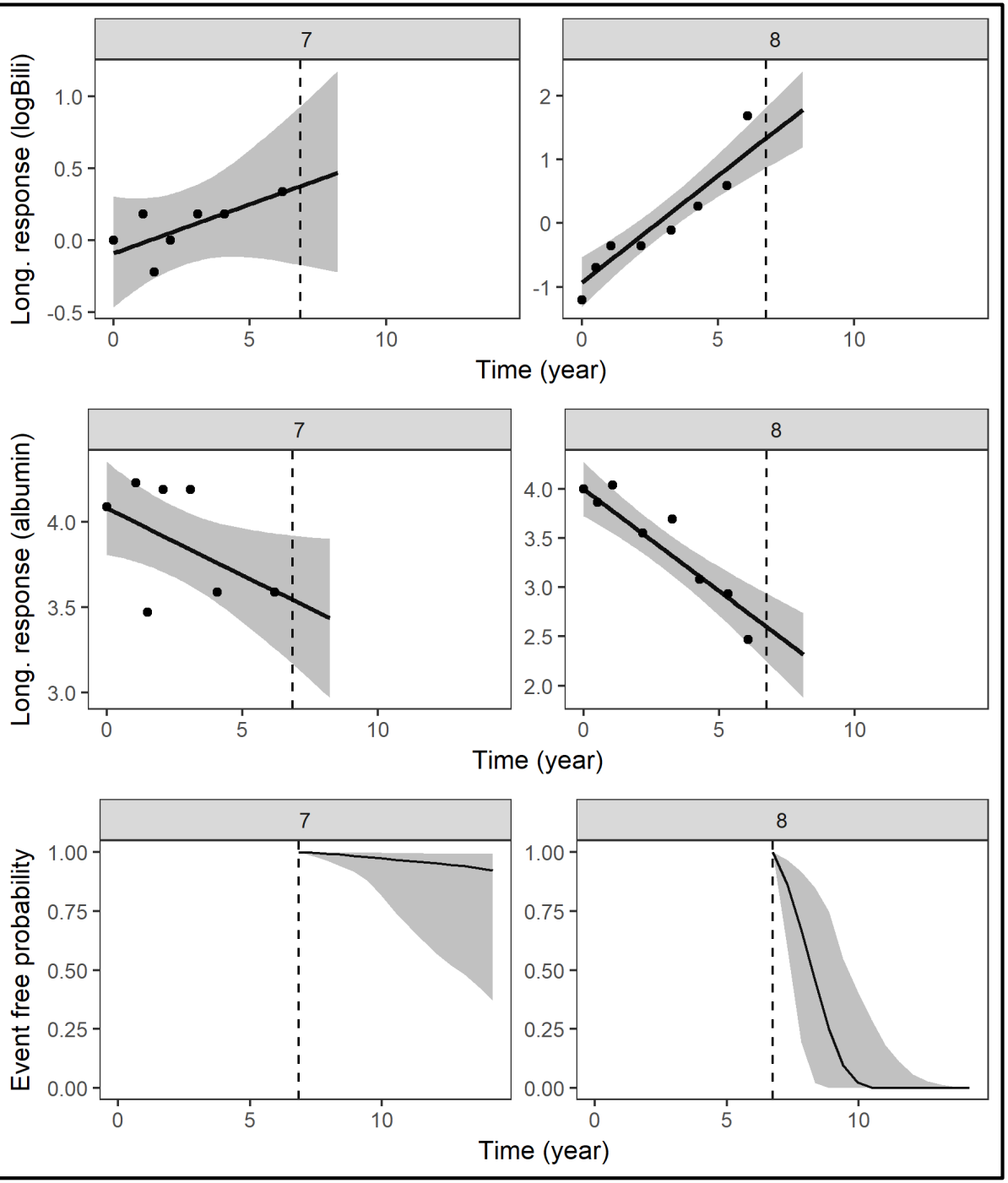
```
> print(fit1)
```

```
> summary(fit1, pars = "ass
```

```
> p1 <- posterior_traj(fit1,
> p2 <- posterior_traj(fit1,
> p3 <- posterior_survfit(fit

> pp1 <- plot(p1, vline = TRU
> pp2 <- plot(p2, vline = TRU

> plot_stack_jm(yplot = list
```



Acknowledgements

- StanCon committee and sponsor for support via the Student Scholarship
- Eric Novik and Daniel Lee at Generable, for both academic support and financial support to get me here! ☺
- Ben Goodrich and Jonah Gabry (maintainers of **rstanarm**)
- Collaborators on motivating projects: Nidal Al-Huniti, James Dunyak and Robert Fox at AstraZeneca; Serigne Lo at Melanoma Institute Australia
- My PhD supervisors: Rory Wolfe, Margarita Moreno-Betancur, Michael Crowther
- My PhD funders: Australian National Health and Medical Research Council (NHMRC) and Victorian Centre for Biostatistics (ViCBiostat)

References

- http://mc-stan.org/users/interfaces/rstanarm.html
- https://github.com/stan-dev/rstanarm

MONASH University

ViCBiostat

ISCB ASC18

26-30 AUGUST 2018
MELBOURNE, AUSTRALIA

Key Dates

Registration opens **August 2017**

Abstract submission opens **October 2017**

Abstract submission closes **March 2018**

Early bird registration deadline **May 2018**

Joint International Society for Clinical Biostatistics and Australian Statistical Conference 2018

HOSTED BY: ISCB International Society for Clinical Biostatistics; Statistical Society of Australia

MANAGED BY: arinex

ISCB ASC 2018 Conference Managers: Arinex Pty Ltd
91–97 Islington St, Collingwood, VIC 3066, Australia
Ph: +61 3 8888 9500

Joint International Society for Clinical Biostatistics and Australian Statistical Conference 2018

Confirmed Keynote Speakers:

Chris Holmes
University of Oxford

Louise Ryan
University of Technology, Sydney

Susan Murphy
University of Michigan

Thomas Lumley
University of Auckland

web: www.iscbasc2018.com

email: iscbasc2018@arinex.com.au