Bayesian joint models for multiple longitudinal biomarkers and a time-to-event outcome: software development and a melanoma case study

Sam Brilleman[1,2], Michael J. Crowther[3], Margarita Moreno-Betancur[2,4,5], Serigne Lo[6], Jacqueline Buros Novik[7], Rory Wolfe[1,2]

38th Annual Conference of the International Society for Clinical Biostatistics

Vigo, Spain

9-13th July 2017

[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] Melanoma Institute Australia, Sydney, Australia

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

MONASH University

ViCBiostat

Collection of **biomarker data** from a melanoma patient

Collection of **biomarker data** from a melanoma patient

Analysing **changes** in LDH

Fitting a linear **mixed model** to lymphocyte counts

Fitting a **Cox model** for survival

The **1990's** analysis

Collection of **biomarker data** from a melanoma patient

Analysing **changes** in LDH

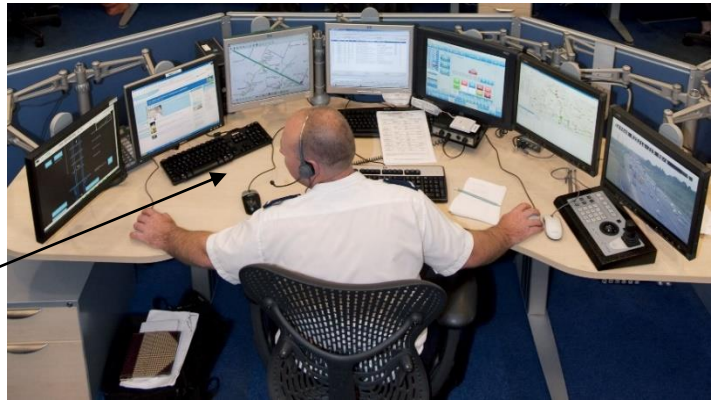
Fitting a linear **mixed model** to lymphocyte counts

Fitting a **Cox model** for survival

The **1990's** analysis

Collection of **biomarker data** from a melanoma patient

The **2017** analysis



Fitting a **joint model** for LDH, lymphocytes and survival

Background

- **What** is joint modelling?
- The joint estimation of regression models which, traditionally, we would have estimated separately:
  - A (multivariate) **longitudinal mixed model** for a longitudinal outcome(s)
  - A **time-to-event model** for the time to an event of interest
  - The "sub"models are linked through **shared (subject-specific) parameters**

- **Why** use it?
  - We want to understand how (some function of) the longitudinal outcome is associated with risk of the event
    - can allow for **measurement error** in the biomarker
    - can allow for **discrete-time** measurement of the biomarker
  - "Dynamic" predictions of the risk of the event
  - Separating out "direct" and "indirect" effects of treatment
  - Adjusting for informative dropout

Joint model formulation

$y_{ijk}(t)$ is the value at time t of the
k^{th} longitudinal marker ($k = 1, \dots, K$)
for the i^{th} individual ($i = 1, \dots, N$)
at the j^{th} time point ($j = 1, \dots, J_{ik}$)
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

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

$$\eta_{ijk}(t) = g_k\left(\mu_{ijk}(t)\right) = \boldsymbol{x'_{ijk}}(t)\boldsymbol{\beta_k} + \boldsymbol{z'_{ijk}}(t)\boldsymbol{b_{ik}}$$

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

Event submodel

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{w'_i}(t)\boldsymbol{\gamma} + \sum_{k=1}^{K} \sum_{q=1}^{Q_k} \alpha_{kq}\, f_{kq}(\boldsymbol{\beta_k}, \boldsymbol{b_{ik}}; t) \right)$$

Joint model formulation

$y_{ijk}(t)$ is the value at time t of the
k^{th} longitudinal marker ($k = 1, \dots, K$)
for the i^{th} individual ($i = 1, \dots, N$)
at the j^{th} time point ($j = 1, \dots, J_{ik}$)
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

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

$$\eta_{ijk}(t) = g_k\left(\mu_{ijk}(t)\right) = \boldsymbol{x'_{ijk}}(t)\boldsymbol{\beta_k} + \boldsymbol{z'_{ijk}}(t)\boldsymbol{b_{ik}}$$

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

Event submodel

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{w'_i}(t)\boldsymbol{\gamma} + \sum_{k=1}^{K}\sum_{q=1}^{Q_k}\alpha_{kq} f_{kq}(\boldsymbol{\beta_k}, \boldsymbol{b_{ik}}; t)\right)$$

association **term**
(some function of parameters
from the longitudinal submodel)

association **parameter**

Association structures

$$h_0(t)\exp\left(\boldsymbol{w_i'}(t)\boldsymbol{\gamma}+\sum_{k=1}^{K}\sum_{q=1}^{Q_k}\alpha_{kq}\boxed{f_{kq}(\boldsymbol{\beta_k},\boldsymbol{b_{ik}};t)}\right)$$

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

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

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

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

$\eta_{ik}(t)\times\eta_{ik'}(t)$ ⟶ Interactions between values of the different biomarkers (for $k\neq k'$)

$\eta_{ik}(t)\times c_i(t)$ ⟶ Interactions with observed data (e.g. for some observed covariate $c_i(t)$)

Joint modelling software

- An abundance of **methodological** developments in joint modelling
- 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***
  - released packages: joineRML (R, available on CRAN)
  - development packages: **survtd**, **rstanarm**, JMbayes (R, available on GitHub); stjm

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

* Hickey et al. (2016) provide a nice "recent" review of multivariate joint model software

Joint modelling software

- An abundance of **methodological** developments in joint modelling
- 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***
  - released packages: joineRML (R, available on CRAN)
  - development packages: **survtd**, **rstanarm**, JMbayes (R, available on GitHub); stjm

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

* Hickey et al. (2016) provide a nice "recent" review of multivariate joint model software

Joint modelling software

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

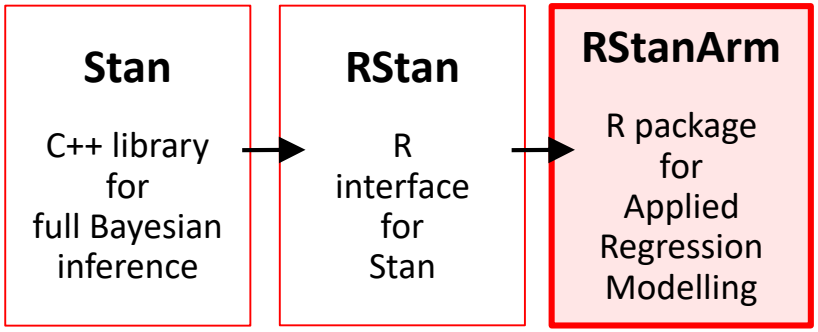
- Well established software for one longitudinal outcome
  - e.g. stjm (Stata); joineR, JM, JMbayes, frailtypack (R); JMFit (SAS)

- Recent software developments for **multiple longitudinal outcomes***
  - released packages: joineRML (R, available on CRAN)
  - development packages: **survtd**, **rstanarm**, JMbayes (R, available on GitHub); stjm

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

* Hickey et al. (2016) provide a nice "recent" review of multivariate joint model software

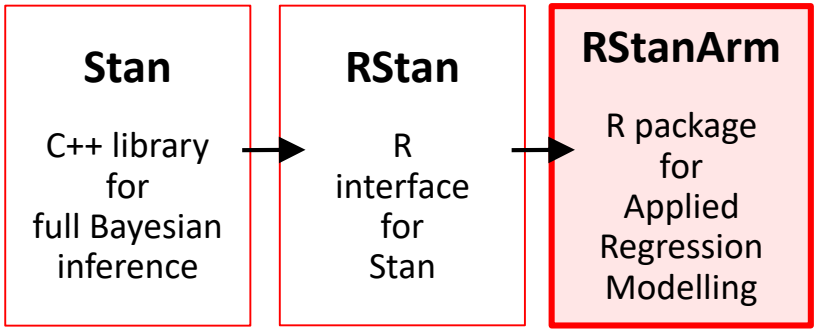
Bayesian joint models via Stan



- Development version available on GitHub
  - https://github.com/sambrilleman/rstanarm

  (soon to be migrated to https://github.com/stan-dev/rstanarm then CRAN)
- 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)
  - Novel decomposition of covariance matrix for the random effects
- **Posterior predictions** – including "dynamic predictions" of event outcome
- Baseline hazard
  - Weibull, piecewise constant, B-splines regression

Bayesian joint models via Stan



- Development version available on GitHub
  - https://github.com/sambrilleman/rstanarm

  (soon to be migrated to https://github.com/stan-dev/rstanarm then CRAN)

- 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)
  - Novel decomposition of covariance matrix for the random effects
- **Posterior predictions** – including "dynamic predictions" of event outcome
- Baseline hazard
  - Weibull, piecewise constant, B-splines regression

Application to Melanoma Institute Australia data

- Background:
  - Approx. **40% of melanoma patients do not respond** to immunotherapy treatment
  - But currently, **no reliable marker** of which patients will (or will not) respond
  - Often clinician must wait until disease progression before altering treatment
  - If risk of progression was known earlier then this could improve patient management
    - e.g. switch drugs, escalation of immunotherapy, initiate more aggressive treatments (e.g. radiotherapy), improve quality of life (e.g. initiate palliative care)

- Data and aims:
  - Phase 1 & 2 clinical trial patients with late-stage melanoma (N = 332)
  - Model the **natural history of several clinical biomarkers (LDH, neutrophils, lymphocytes)**
  - Explore which biomarkers are associated with **progression-free survival**
  - Determine the most appropriate functional form(s) for any associations

Model specification

```
m0 <- stan_jm(
  formulaLong = list(
    ldh ~ t0 + (t0 | id),
    neu ~ t0 + (t0 | id),
    lym ~ t0 + (t0 | id)),
  formulaEvent = Surv(etime, status) ~ agecat + sex,
  dataLong = dat2, dataEvent = dat2.id,
  family = Gamma(log), time_var = "t0")
```

- GLMM (**Gamma, log link**) for each biomarker

$$y_{ijk}(t) \sim Gamma\left(v_k, \lambda_{ijk}(t)\right) \quad \text{with expected value} \quad \mu_{ijk}(t) = \frac{\lambda_{ijk}(t)}{v_k}$$

with linear predictor

$$\eta_{ijk}(t) = \log(\mu_{ijk}(t)) = \beta_{0k} + \beta_{1k}\, t + b_{0ik} + b_{1ik}\, t \quad (i = 1, \ldots, N;\ j = 1, \ldots, J_{ik};\ k = 1, \ldots, 3)$$

- **Weibull** proportional hazards model

$$h_i(t) = h_0(t) \exp\left(\gamma_0 + \gamma_1 G_i + \sum_{a=1}^{3} \gamma_{2a} A_{ia} + \sum_{k=1}^{3} \alpha_k \eta_{ik}(t)\right)$$

where G_i = 1 if individual i is male (or 0 otherwise)

A_{ia} = 1 if individual i is in age category a (or 0 otherwise)

Hazard ratios (event submodel)

	Coefficient (log HR)	Standard error (posterior SD)	Hazard ratio
Age category (years, ref: ≤50)			
51 to 60	-0.166	0.274	0.847
61 to 70	0.113	0.267	1.120
71 and above	0.050	0.279	1.052
Gender (ref: Female)			
Male	-0.110	0.211	0.896
Association parameters			
LDH (log value)	0.461	0.192	1.586
Neutrophils (log value)	2.287	0.454	9.845
Lymphocytes (log value)	-0.472	0.286	0.624

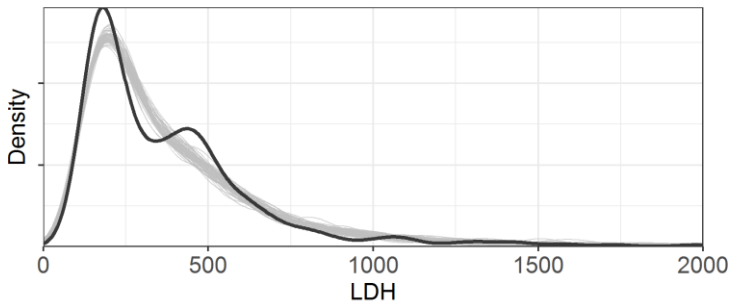
Since a one unit increase in log(LDH) corresponds to an exp(1) = 2.7-fold increase in LDH, we can say that:

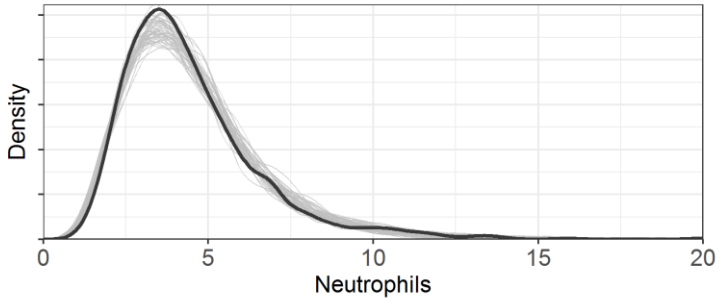
"A 2.7-fold increase in mean LDH is associated with an estimated 1.6-fold increase in the hazard of death or disease progression"

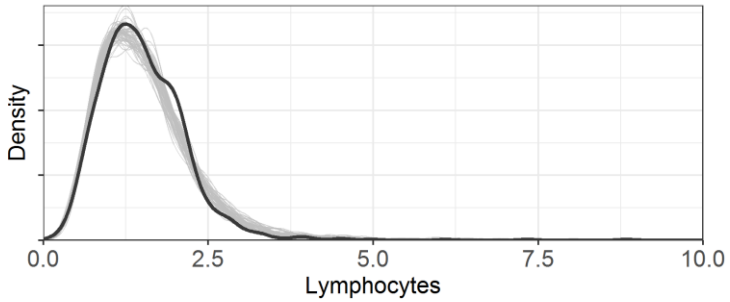
Posterior predictions (longitudinal)

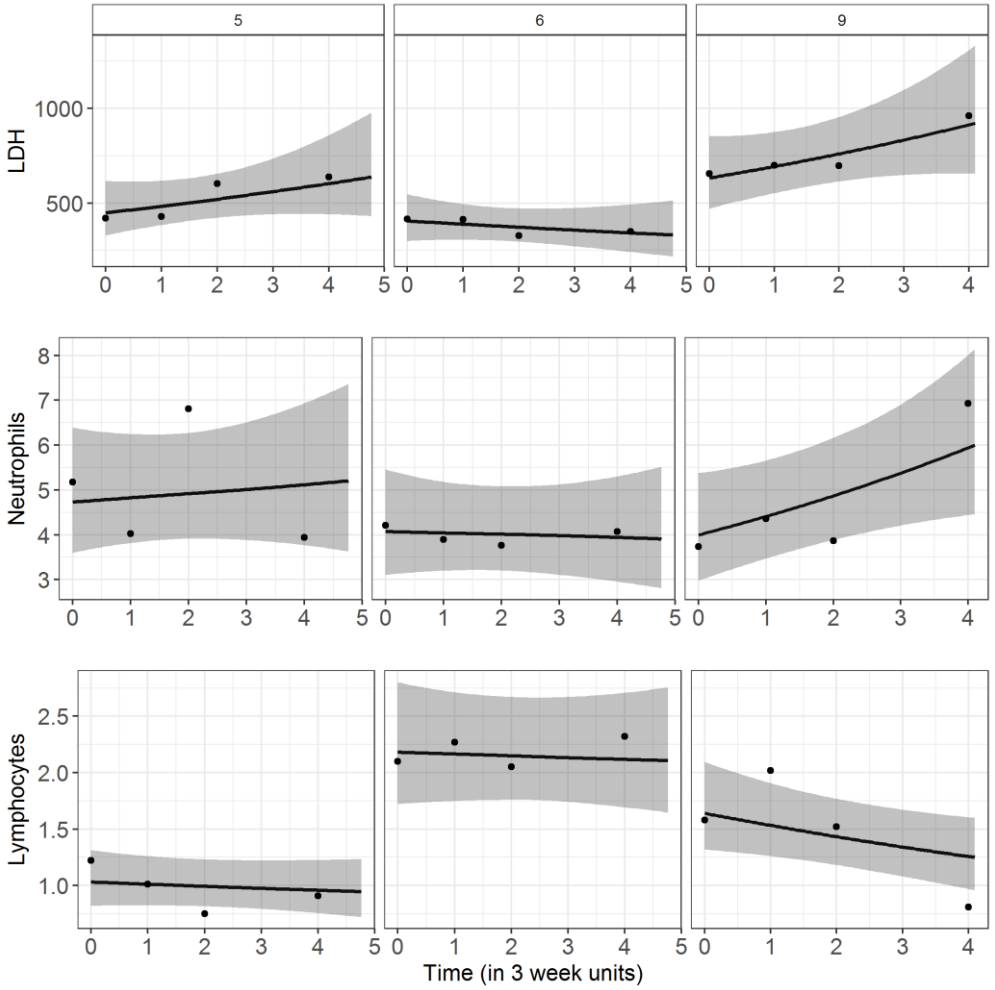
```
pp_check(m0, m = 1)
pp_check(m0, m = 2)
pp_check(m0, m = 3)

plot(posterior_traj(m0, ids = c(5,6,9))
```





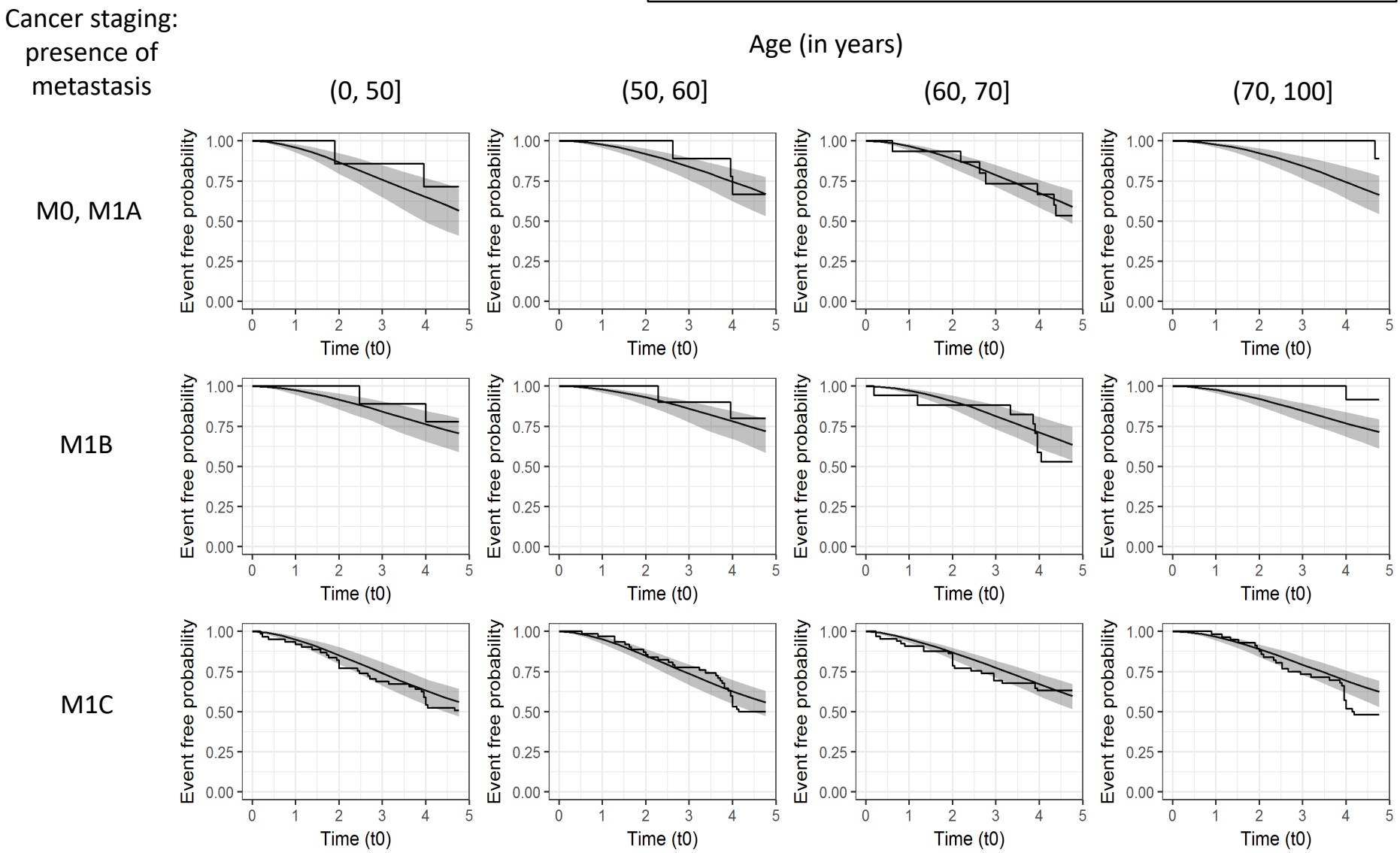




Posterior predictions (event)

```
ps_check # overall standardised survival vs KM

ids <- … # obtain IDs of individuals in risk group
plot(posterior_survfit(m0, ids = ids, time = 0,
                       standardise = TRUE))
```

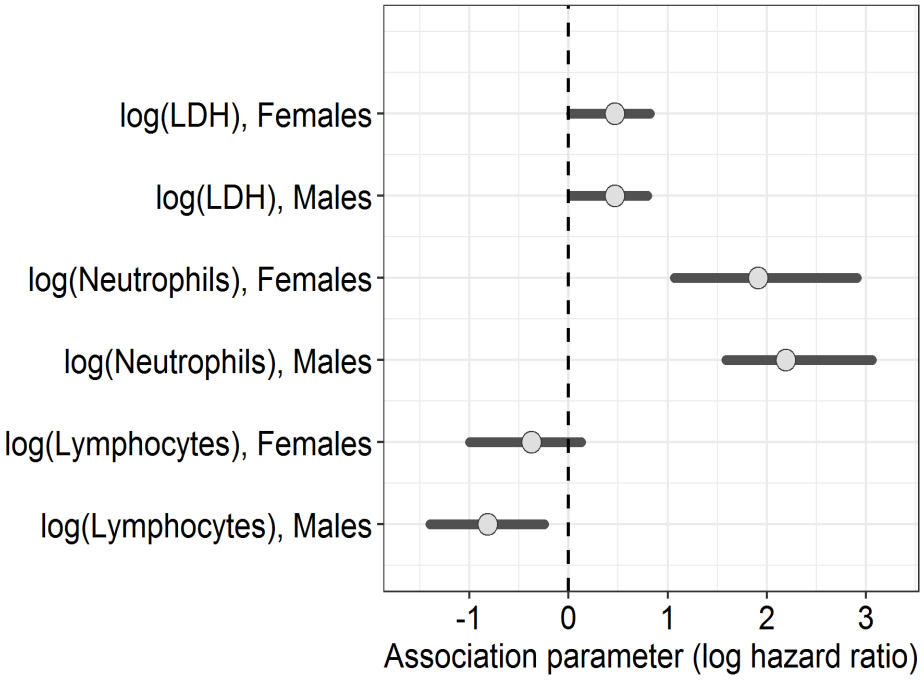


Association structures allowing for effect modification

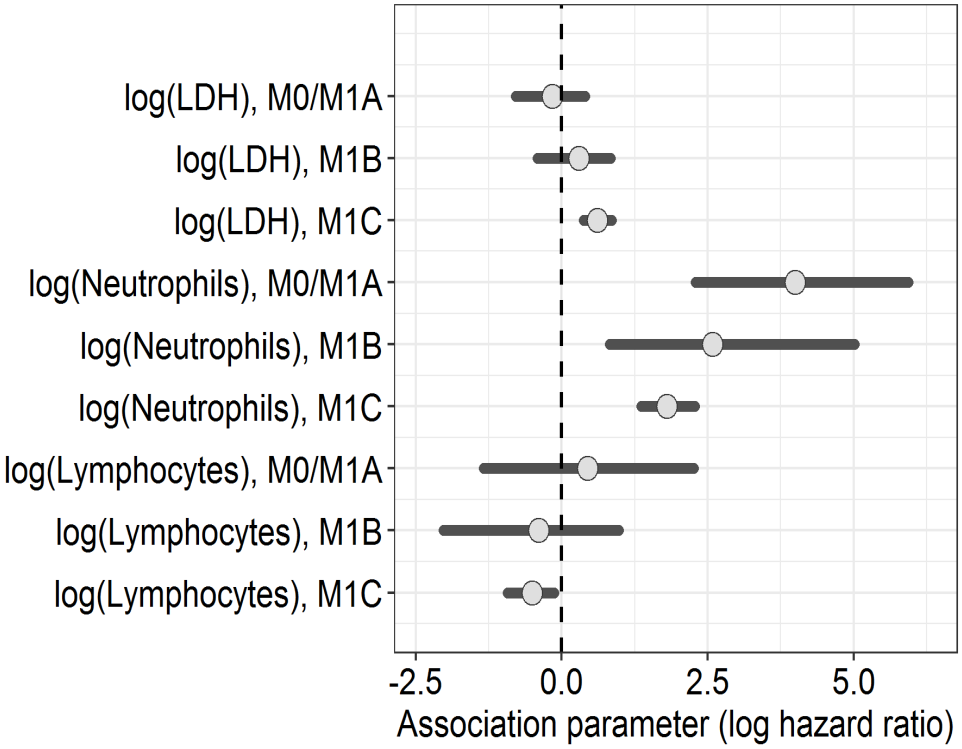
```
m1 <- update(m0, assoc = c("etavalue",
                "etavalue_data(~ gender))

m2 <- update(m0, assoc = c("etavalue",
                "etavalue_data(~ stage))
```

- **Interaction** term between the (log) value of the biomarker and **gender**



- **Interaction** term between the (log) value of the biomarker and **cancer stage/severity**



Acknowledgements

- Ben Goodrich and Jonah Gabry (maintainers of RStanArm)
- **ISCB** (esp. **Nadine Binder**) for support via the Student Conference Award (StCA)
- My PhD supervisors: Rory Wolfe, Margarita Moreno-Betancur, Michael Crowther
- My PhD funders: NHMRC and Victorian Centre for Biostatistics (ViCBiostat)
- Colleagues at Monash University and ViCBiostat

References

- https://github.com/sambrilleman/rstanarm
- http://mc-stan.org/users/interfaces/rstanarm.html
- https://github.com/moreno-betancur/survtd
- Moreno-Betancur M, Carlin JB, Brilleman SL, Tanamas S, Peeters A, Wolfe R. Survival analysis with time-dependent covariates subject to measurement error and missing data: Two-stage joint model using multiple imputation. 2016 (submitted).
- Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Med Res Methodol.* 2016; **16(1)**: 117.

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

Estimation

Likelihood function:

$$p(\boldsymbol{y_{i1}}, \ldots, \boldsymbol{y_{iK}}, T_i, d_i | \boldsymbol{b_i}, \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \left(\prod_{k=1}^{K} \prod_{j=1}^{n_{ik}} p\left(y_{ijk}(t) \middle| \boldsymbol{b_i}, \boldsymbol{\theta_{y_k}}\right) \right) p(T_i, d_i | \boldsymbol{b_i}, \boldsymbol{\theta_T})\, p(\boldsymbol{b_i} | \boldsymbol{\theta_b})\, \mathrm{d}\boldsymbol{b_i}$$

k^{th} longitudinal submodel

event submodel

random effects model

- Assumes **conditional independence**, that is, conditional on $\boldsymbol{b_i}$ the distinct longitudinal and event processes are independent
  - requires we specify the model correctly, including the **association structure**
- rstanarm uses a **full Bayesian specification** (i.e. includes priors)
- **Estimation via MCMC** (Hamiltonian Monte Carlo) or, less preferably, variational Bayes