simsurv: A Package for Simulating Simple or Complex Survival Data

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

useR! Conference 2018

Brisbane, Australia

10-13th July 2018

[1] Monash University, Melbourne, Australia
[2] Victorian Centre for Biostatistics (ViCBiostat)
[3] Murdoch Children's Research Institute, Melbourne, Australia
[4] University of Melbourne, Melbourne, Australia
[5] University of Leicester, Leicester, UK

MONASH University

ViCBiostat

Outline

- Background to survival analysis
- A general method for simulating event times
- Examples of using the ‘simsurv’ package
- Summary

What is survival analysis?

- The analysis of a variable that corresponds to the **time from a defined baseline** (e.g. diagnosis of a disease) until **occurrence of an event** of interest (e.g. heart failure).

What is survival analysis?

- The analysis of a variable that corresponds to the **time from a defined baseline** (e.g. diagnosis of a disease) until **occurrence of an event** of interest (e.g. heart failure).
- Also known as:
  - Time-to-event analysis
  - Duration analysis (economics)
  - Reliability analysis (engineering)
  - Event history analysis (sociology)

What is survival analysis?

- The analysis of a variable that corresponds to the **time from a defined baseline** (e.g. diagnosis of a disease) until **occurrence of an event** of interest (e.g. heart failure).
- Also known as:
  - Time-to-event analysis
  - Duration analysis (economics)
  - Reliability analysis (engineering)
  - Event history analysis (sociology)
- The context for this talk will be **health research**
  - Each observational unit will be an “individual” (e.g. a patient)

Why simulate survival data?

- To evaluate the performance of new or existing statistical methods for survival analysis

Why simulate survival data?

- To evaluate the performance of new or existing statistical methods for survival analysis

- To calculate statistical power, e.g. in planning clinical trials

Why simulate survival data?

- To evaluate the performance of new or existing statistical methods for survival analysis
- To calculate statistical power, e.g. in planning clinical trials
- To calculate uncertainty in model predictions, e.g. transition probabilities in multistate models

Why simulate survival data?

- To evaluate the performance of new or existing statistical methods for survival analysis
- To calculate statistical power, e.g. in planning clinical trials
- To calculate uncertainty in model predictions, e.g. transition probabilities in multistate models
- …others?

Modelling survival data

- Let T_i^* denote the "true" event time for individual i
- In practice, T_i^* may not be observed due to right censoring, e.g. the study ending before an individual experiences the event

Modelling survival data

- Let T_i^* denote the "true" event time for individual i
- In practice, T_i^* may not be observed due to right censoring, e.g. the study ending before an individual experiences the event
- **Possible to model T_i^* directly**, e.g. "accelerated failure time (AFT)" models

Modelling survival data

- Let T_i^* denote the "true" event time for individual i
- In practice, T_i^* may not be observed due to right censoring, e.g. the study ending before an individual experiences the event
- **Possible to model T_i^* directly**, e.g. "accelerated failure time (AFT)" models
- But **more common to model the *rate*** of occurrence of the event (e.g. the "Cox" model)
- The *hazard* at time *t* is defined as the *instantaneous rate* of occurrence for the event at time *t*

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_i^* < t + \Delta t \mid T_i^* > t)}{\Delta t}$$

The hazard, cumulative hazard & survival

- Hazard (for individual i): $h_i(t)$
- Cumulative hazard: $H_i(t) = \int_{s=0}^{t} h_i(s)ds$
- Survival probability: $S_i(t) = P(T_i^* > t) = \exp\big(-H_i(t)\big)$

The hazard, cumulative hazard & survival

- Hazard (for individual i): $h_i(t)$
- Cumulative hazard: $H_i(t) = \int_{s=0}^{t} h_i(s)ds$
- Survival probability: $S_i(t) = P(T_i^* > t) = \exp\left(-H_i(t)\right)$

This is the complement of the CDF for the distribution of event times

The hazard, cumulative hazard & survival

- Hazard (for individual i): $h_i(t)$
- Cumulative hazard: $H_i(t) = \int_{s=0}^{t} h_i(s)ds$
- Survival probability: $S_i(t) = P(T_i^* > t) = \exp\left(-H_i(t)\right)$

This is the complement of the CDF for the distribution of event times

- The "probability integral transformation" tells us $1 - F_X(X) = U$, where $F_X(.)$ is the CDF of a continuous random variable X, and U is a uniform random variable on the range 0 to 1

Cumulative hazard inversion

- The result from the previous slide tells us

$$\exp\left(-H_i(T_i^S)\right) = U_i \quad \Longrightarrow \quad T_i^S = H_i^{-1}(-\log(U_i))$$

where

- T_i^S is a randomly drawn (i.e. simulated) event time for individual i
- U_i is a random uniform variable on the range 0 to 1
- $H_i(t) = \int_{s=0}^{t} h_i(s)\, ds$ is the cumulative hazard evaluated at time t

Cumulative hazard inversion

- The result from the previous slide tells us

$$\exp\left(-H_i(T_i^S)\right) = U_i \quad \Longrightarrow \quad T_i^S = H_i^{-1}(-\log(U_i))$$

where

  - T_i^S is a randomly drawn (i.e. simulated) event time for individual i
  - U_i is a random uniform variable on the range 0 to 1
  - $H_i(t) = \int_{s=0}^{t} h_i(s)\, ds$ is the cumulative hazard evaluated at time t

- Commonly known as the 'cumulative hazard inversion method' [1,2]
- Easy and efficient when $H_i(t)$ has a closed form and is invertible

[1] Leemis LM. Variate Generation for Accelerated Life and Proportional Hazards Models. *Operations Research*, 1987: 35(6); 892–894.

[2] Bender R et al. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005: 24(11); 1713–1723.

Cumulative hazard inversion

- The result from the previous slide tells us

$$\exp\left(-H_i(T_i^S)\right) = U_i \quad \Longrightarrow \quad T_i^S = H_i^{-1}(-\log(U_i))$$

where

  - T_i^S is a randomly drawn (i.e. simulated) event time for individual i
  - U_i is a random uniform variable on the range 0 to 1
  - $H_i(t) = \int_{s=0}^{t} h_i(s)\, ds$ is the cumulative hazard evaluated at time t

- Commonly known as the 'cumulative hazard inversion method' [1,2]
- Easy and efficient when $H_i(t)$ has a closed form and is invertible
- But for complex specifications of $h_i(t)$:

[1] Leemis LM. Variate Generation for Accelerated Life and Proportional Hazards Models. *Operations Research*, 1987: 35(6); 892–894.

[2] Bender R et al. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005: 24(11); 1713–1723.

Cumulative hazard inversion

- The result from the previous slide tells us

$$\exp\left(-H_i(T_i^S)\right) = U_i \quad \Longrightarrow \quad T_i^S = H_i^{-1}(-\log(U_i))$$

where

  - T_i^S is a randomly drawn (i.e. simulated) event time for individual i
  - U_i is a random uniform variable on the range 0 to 1
  - $H_i(t) = \int_{s=0}^{t} h_i(s)\, ds$ is the cumulative hazard evaluated at time t

- Commonly known as the 'cumulative hazard inversion method' [1,2]
- Easy and efficient when $H_i(t)$ has a closed form and is invertible
- But for complex specifications of $h_i(t)$:
  - $H_i(t)$ may not have a closed form

[1] Leemis LM. Variate Generation for Accelerated Life and Proportional Hazards Models. *Operations Research*, 1987: 35(6); 892–894.

[2] Bender R et al. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005: 24(11); 1713–1723.

Cumulative hazard inversion

- The result from the previous slide tells us

$$\exp\left(-H_i(T_i^S)\right) = U_i \quad \Longrightarrow \quad T_i^S = H_i^{-1}(-\log(U_i))$$

where

  - T_i^S is a randomly drawn (i.e. simulated) event time for individual i
  - U_i is a random uniform variable on the range 0 to 1
  - $H_i(t) = \int_{s=0}^{t} h_i(s)\, ds$ is the cumulative hazard evaluated at time t

- Commonly known as the 'cumulative hazard inversion method' [1,2]
- Easy and efficient when $H_i(t)$ has a closed form and is invertible
- But for complex specifications of $h_i(t)$:
  - $H_i(t)$ may not have a closed form
  - $H_i(t)$ may not be invertible

[1] Leemis LM. Variate Generation for Accelerated Life and Proportional Hazards Models. *Operations Research*, 1987: 35(6); 892–894.

[2] Bender R et al. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005: 24(11); 1713–1723.

Cumulative hazard inversion

- The result from the previous slide tells us

$$\exp\left(-H_i(T_i^S)\right) = U_i \quad \Longrightarrow \quad T_i^S = H_i^{-1}(-\log(U_i))$$

where

  - T_i^S is a randomly drawn (i.e. simulated) event time for individual i
  - U_i is a random uniform variable on the range 0 to 1
  - $H_i(t) = \int_{s=0}^{t} h_i(s)\, ds$ is the cumulative hazard evaluated at time t

- Commonly known as the 'cumulative hazard inversion method' [1,2]
- Easy and efficient when $H_i(t)$ has a closed form and is invertible
- But for complex specifications of $h_i(t)$:
  - $H_i(t)$ may not have a closed form → numerical integration (quadrature)
  - $H_i(t)$ may not be invertible

[1] Leemis LM. Variate Generation for Accelerated Life and Proportional Hazards Models. *Operations Research*, 1987: 35(6); 892–894.

[2] Bender R et al. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005: 24(11); 1713–1723.

Cumulative hazard inversion

- The result from the previous slide tells us

$$\exp\left(-H_i(T_i^S)\right) = U_i \quad \Longrightarrow \quad T_i^S = H_i^{-1}(-\log(U_i))$$

where

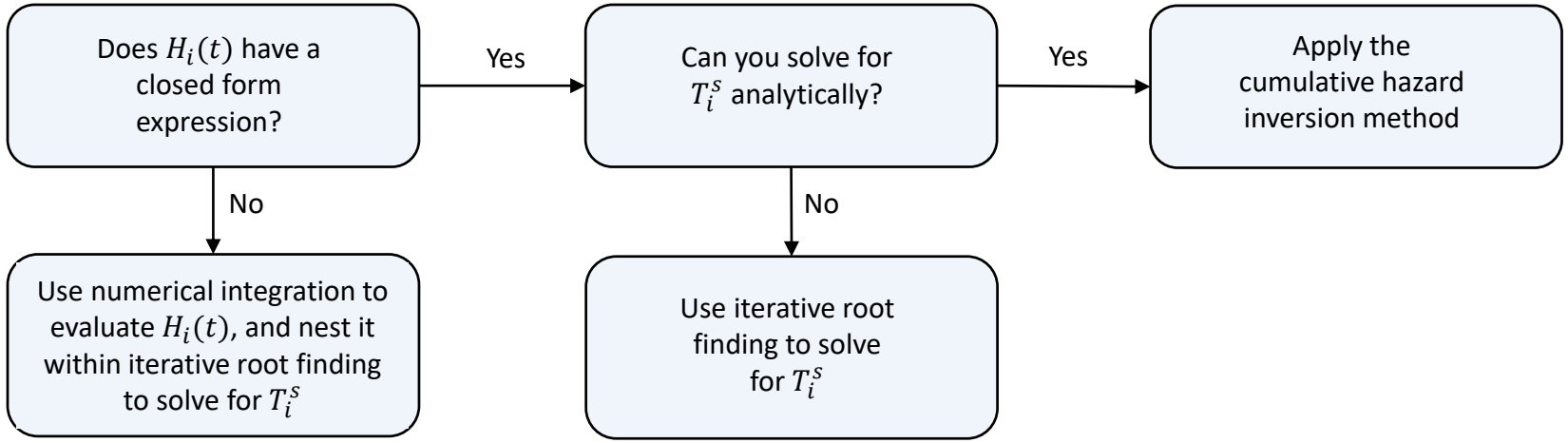
  - T_i^S is a randomly drawn (i.e. simulated) event time for individual i
  - U_i is a random uniform variable on the range 0 to 1
  - $H_i(t) = \int_{s=0}^{t} h_i(s)\, ds$ is the cumulative hazard evaluated at time t

- Commonly known as the 'cumulative hazard inversion method' [1,2]
- Easy and efficient when $H_i(t)$ has a closed form and is invertible
- But for complex specifications of $h_i(t)$:
  - $H_i(t)$ may not have a closed form → numerical integration (quadrature)
  - $H_i(t)$ may not be invertible → iterative univariate root finding

[1] Leemis LM. Variate Generation for Accelerated Life and Proportional Hazards Models. *Operations Research*, 1987: 35(6); 892–894.
[2] Bender R et al. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005: 24(11); 1713–1723.

A general algorithm for simulating event times

- Crowther and Lambert [3] describe an algorithm as follows



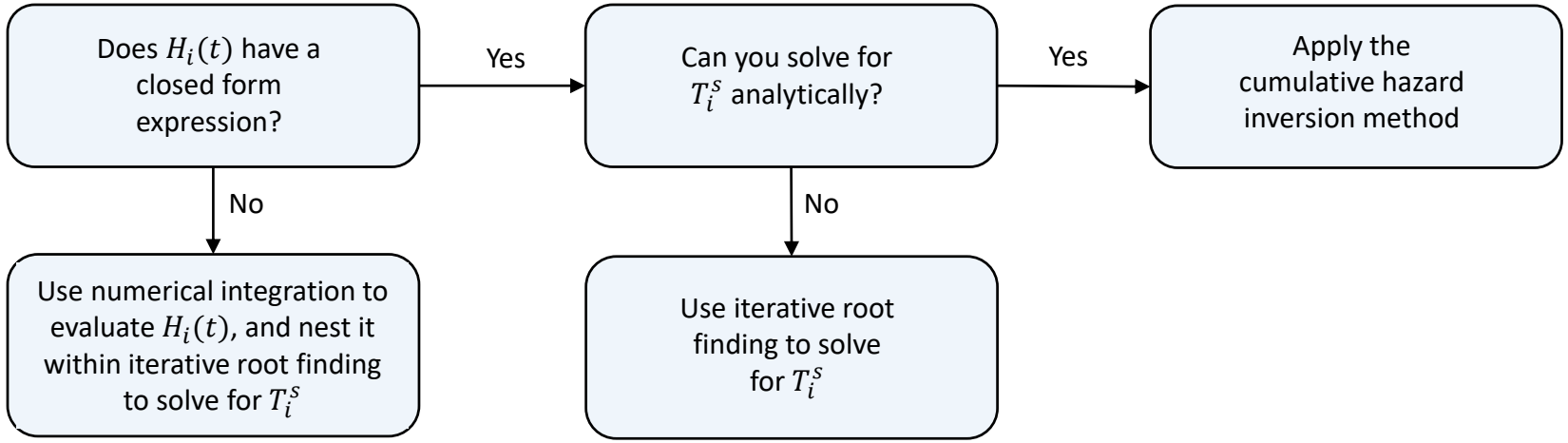
[3] Crowther MJ, Lambert PC. Simulating Biologically Plausible Complex Survival Data. *Statistics in Medicine*, 2013: 32(23); 4118–4134.

[4] Crowther MJ, Lambert PC. Simulating Complex Survival Data. *The Stata Journal*, 2012: 12(4); 674–687.

[5] Brilleman S. (2018) simsurv: Simulate Survival Data. R package version 0.2.2. https://CRAN.R-project.org/package=simsurv

A general algorithm for simulating event times

- Crowther and Lambert [3] describe an algorithm as follows



- This method was implemented in a Stata package [4]
- Now also implemented in R as part of the 'simsurv' package [5]

[3] Crowther MJ, Lambert PC. Simulating Biologically Plausible Complex Survival Data. *Statistics in Medicine*, 2013: 32(23); 4118–4134.
[4] Crowther MJ, Lambert PC. Simulating Complex Survival Data. *The Stata Journal*, 2012: 12(4); 674–687.
[5] Brilleman S. (2018) simsurv: Simulate Survival Data. R package version 0.2.2. https://CRAN.R-project.org/package=simsurv

The ‘simsurv’ package

- Built around one function: `simsurv()`

The ‘simsurv’ package

- Built around one function: `simsurv()`
- Can simulate survival times from:
  - Standard parametric survival distributions (exponential, Weibull, Gompertz)
  - Two-component mixture survival distributions
  - Covariate effects under proportional hazards
  - Covariate effects under non-proportional hazards (i.e. time-dependent effects)
  - Clustered survival times (e.g. shared frailty, meta-analytic models)
  - Time-varying covariates
  - Any user-defined hazard, log hazard, or cumulative hazard function

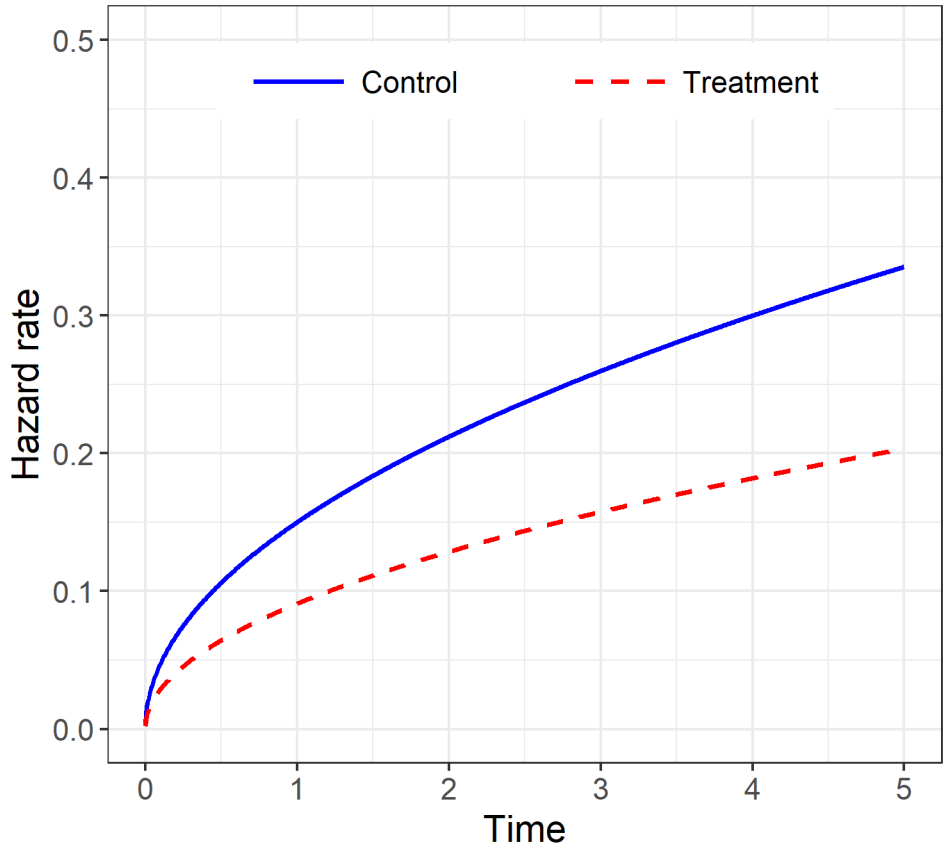
The ‘simsurv’ package

- Built around one function: `simsurv()`
- Can simulate survival times from:
  - Standard parametric survival distributions (exponential, Weibull, Gompertz)
  - Two-component mixture survival distributions
  - Covariate effects under proportional hazards
  - Covariate effects under non-proportional hazards (i.e. time-dependent effects)
  - Clustered survival times (e.g. shared frailty, meta-analytic models)
  - Time-varying covariates
  - Any user-defined hazard, log hazard, or cumulative hazard function
- Uses analytical forms where possible, otherwise
  - Gauss-Kronrod quadrature to evaluate $H_i(t)$
  - Brent’s univariate root finder to invert $H_i(t)$ (via the uniroot function in R)

The ‘simsurv’ package

- Built around one function: `simsurv()`
- Can simulate survival times from:
  - Standard parametric survival distributions (exponential, Weibull, Gompertz)
  - Two-component mixture survival distributions
  - Covariate effects under proportional hazards
  - Covariate effects under non-proportional hazards (i.e. time-dependent effects)
  - Clustered survival times (e.g. shared frailty, meta-analytic models)
  - Time-varying covariates
  - Any user-defined hazard, log hazard, or cumulative hazard function
- Uses analytical forms where possible, otherwise
  - Gauss-Kronrod quadrature to evaluate $H_i(t)$
  - Brent’s univariate root finder to invert $H_i(t)$ (via the uniroot function in R)

Example 1: Standard parametric proportional hazards model



General model:

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{X_i^T\beta}\right)$$

Example model: Weibull model with proportional hazards

$$h_i(t) = \lambda\,\gamma\,t^{\gamma-1}\exp(X_i\beta)$$

Covariates:

$X_i \sim \text{Bern}(0.5)$ (e.g. a binary treatment indicator)

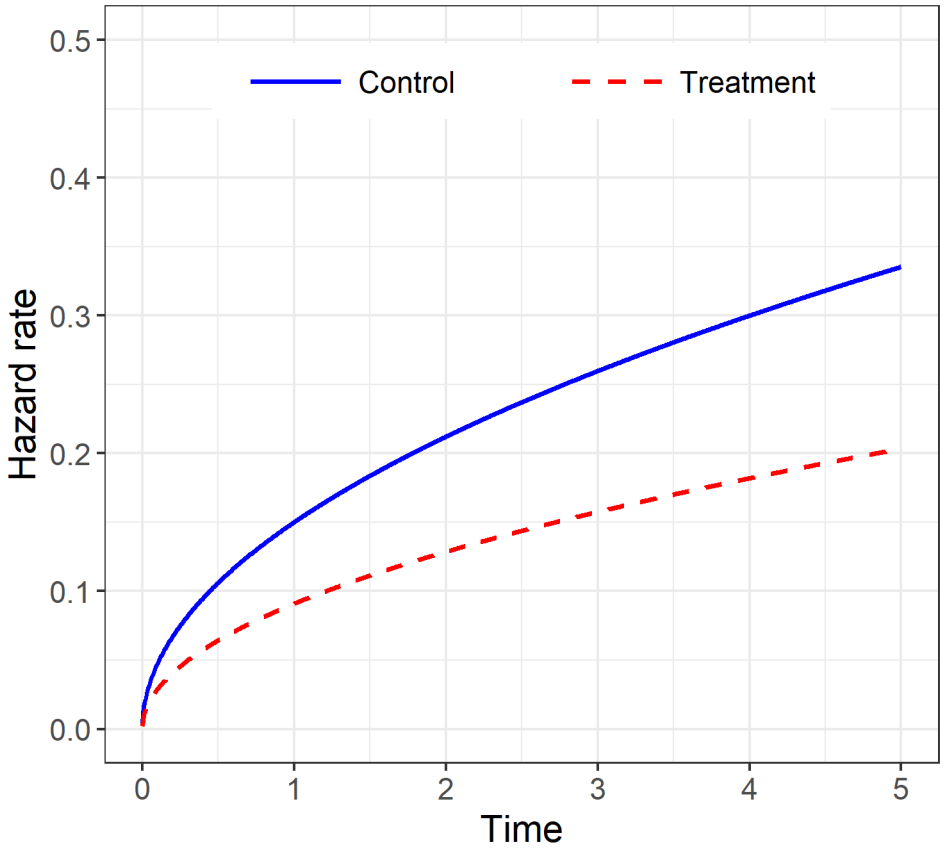
Parameters:

$\lambda = 0.1$ (scale parameter)

$\gamma = 1.5$ (shape parameter)

$\beta = -0.5$ (log hazard ratio)

Example 1: Standard parametric proportional hazards model



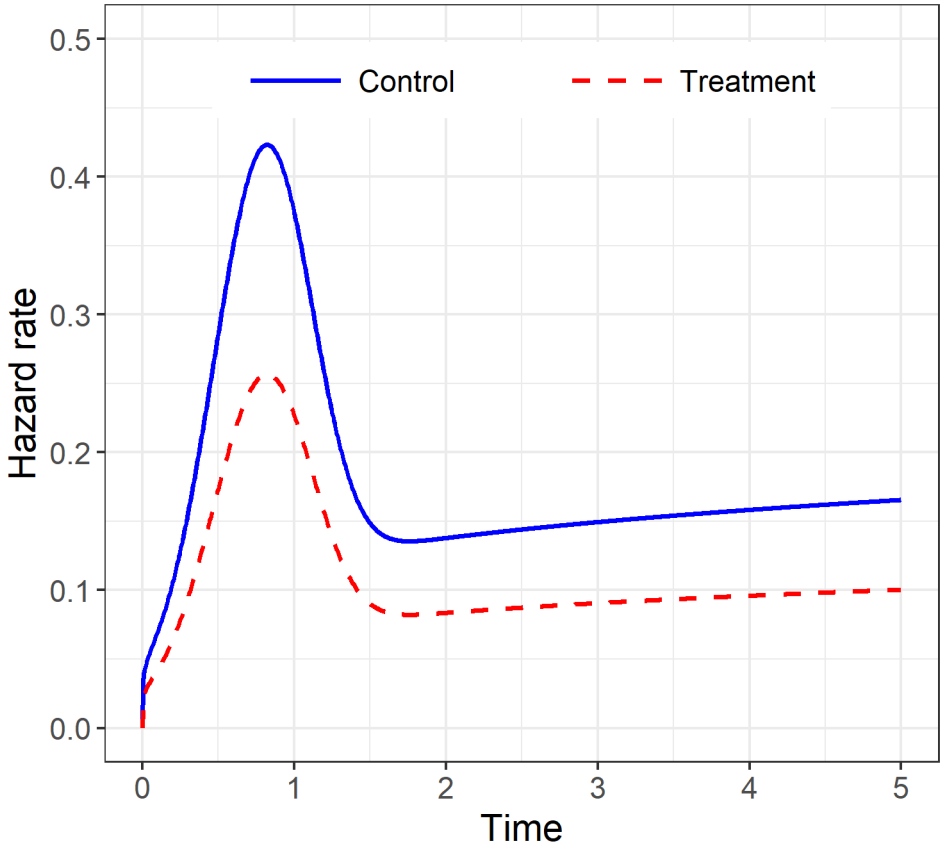

```
# Dimensions
N <- 1000   # total number of patients

# Define covariate data
covs <- data.frame(id = 1:N,
                   trt = rbinom(N, 1, 0.5))

# Define true coefficient (log hazard ratio)
pars <- c(trt = -0.5)

# Simulate the event times
times <- simsurv(dist    = 'weibull',
                 lambdas = 0.1,
                 gammas  = 1.5,
                 x       = covs,
                 betas   = pars)
```

Example 2: Two-component mixture survival distribution



General model:

$$S_i(t) = \left(p\, S_1(t) + (1-p)\, S_2(t)\right)^{\exp(\boldsymbol{X}_i^T \boldsymbol{\beta})} \quad \text{where } 0 < p < 1$$

Example model: Weibull mixture model with proportional hazards

$$S_i(t) = \left(p \exp(-\lambda_1 t^{\gamma_1}) + (1-p) \exp(-\lambda_2 t^{\gamma_2})\right)^{\exp(X_i \beta)}$$

Covariates:

$X_i \sim \text{Bern}(0.5)$ (e.g. a binary treatment indicator)

Parameters:

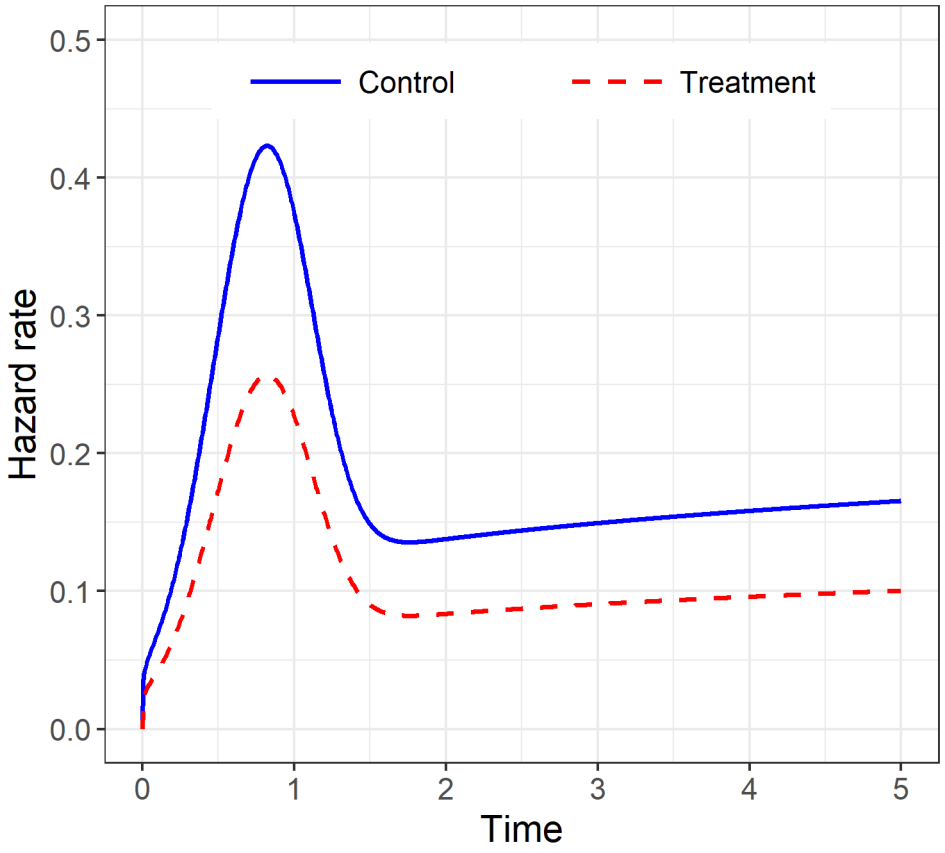
$\lambda_1 = 1.5,\ \lambda_2 = 0.1$ (scale parameters)

$\gamma_1 = 3.0,\ \gamma_2 = 1.2$ (shape parameters)

$p = 0.2$ (mixing parameter)

$\beta = -0.5$ (log hazard ratio)

Example 2: Two-component mixture survival distribution



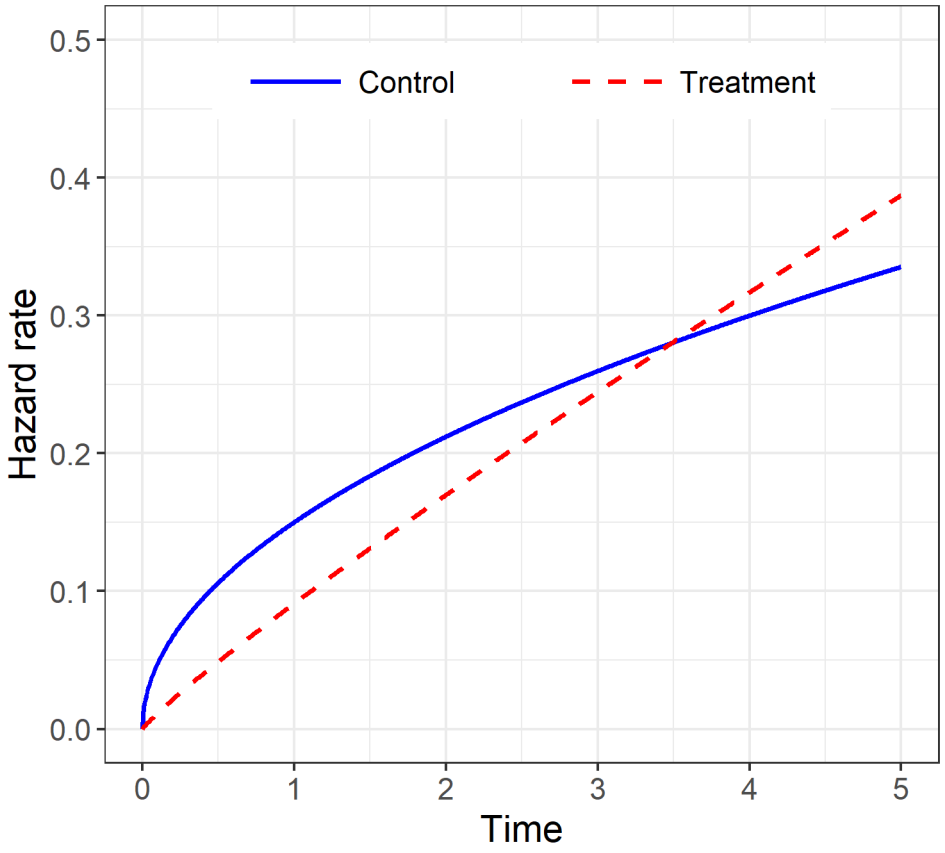

```
# Dimensions
N <- 1000   # total number of patients

# Define covariate data
covs <- data.frame(id = 1:N,
                   trt = rbinom(N, 1, 0.5))

# Define true coefficient (log hazard ratio)
pars <- c(trt = -0.5)

# Simulate the event times
times <- simsurv(dist    = 'weibull',
                 lambdas = c(1.5, 0.1),
                 gammas  = c(3.0, 1.2),
                 mixture = TRUE,
                 pmix    = 0.2,
                 x       = covs,
                 betas   = pars)
```

Example 3: Non-proportional hazards



General model:

$$h_i(t) = h_0(t)\exp\left(\boldsymbol{X_{i1}^T\beta_1} + \boldsymbol{X_{i2}^T\beta_2} f(t)\right)$$

Example model: Weibull model with non-proportional hazards

$$h_i(t) = \lambda\,\gamma\,t^{\gamma-1}\exp(\beta_0 X_i + \beta_1 X_i \log(t))$$

Covariates:

$X_i \sim \text{Bern}(0.5)$ (e.g. a binary treatment indicator)

Parameters:

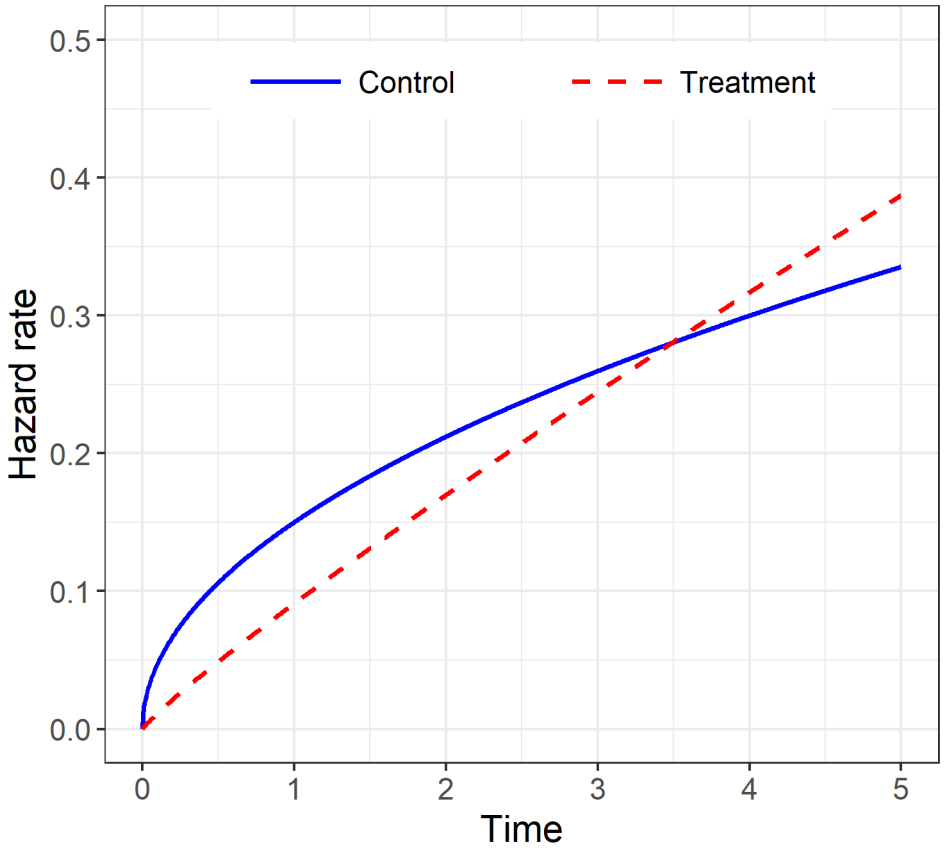
$\lambda = 0.1$ (scale parameter)

$\gamma = 1.5$ (shape parameter)

$\beta_0 = -0.5$ (log hazard ratio when $\log(t) = 0$)

$\beta_1 = 0.4$ (change in log hazard ratio per unit change in $\log(t)$)

Example 3: Non-proportional hazards



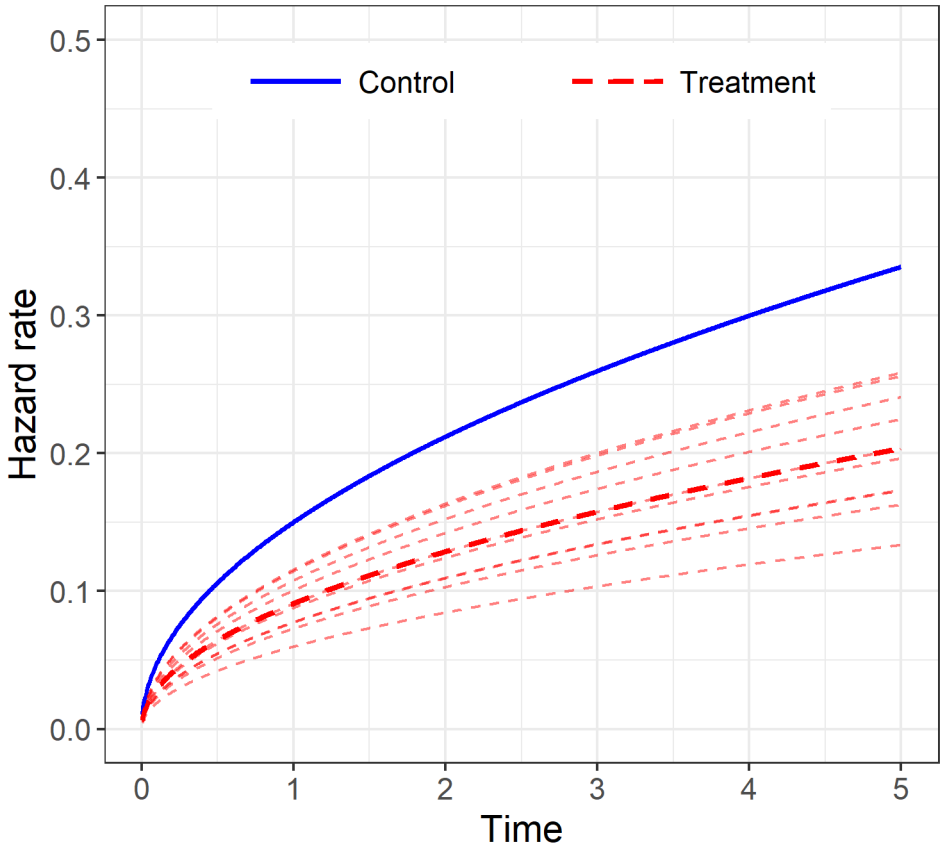

```
# Dimensions
N <- 1000   # total number of patients

# Define covariate data
covs <- data.frame(id  = 1:N,
                   trt = rbinom(N, 1, 0.5))

# Define true coefficients
pars     <- c(trt = -0.5) # time-fixed coefficient
pars_tde <- c(trt =  0.4) # time-varying coefficient

# Simulate the event times
times <- simsurv(dist    = 'weibull',
                 lambdas = 0.1,
                 gammas  = 1.5,
                 x       = covs,
                 betas   = pars,
                 tde     = pars_tde,
                 tdefun  = 'log')
```

Example 4: Clustered survival times



General model:

$$h_{ij}(t) = h_0(t)\exp\left(\boldsymbol{X_{ij}^T\beta} + \boldsymbol{Z_{ij}^T b_j}\right)$$

Example model: Weibull meta-analytic model for RCTs

$$h_{ij}(t) = \lambda\,\gamma\,t^{\gamma-1}\exp\left(X_{ij}\left(\beta + b_j\right)\right)$$

Covariates:

$X_{ij} \sim \text{Bern}(0.5)$ (e.g. a binary treatment indicator)

Parameters:

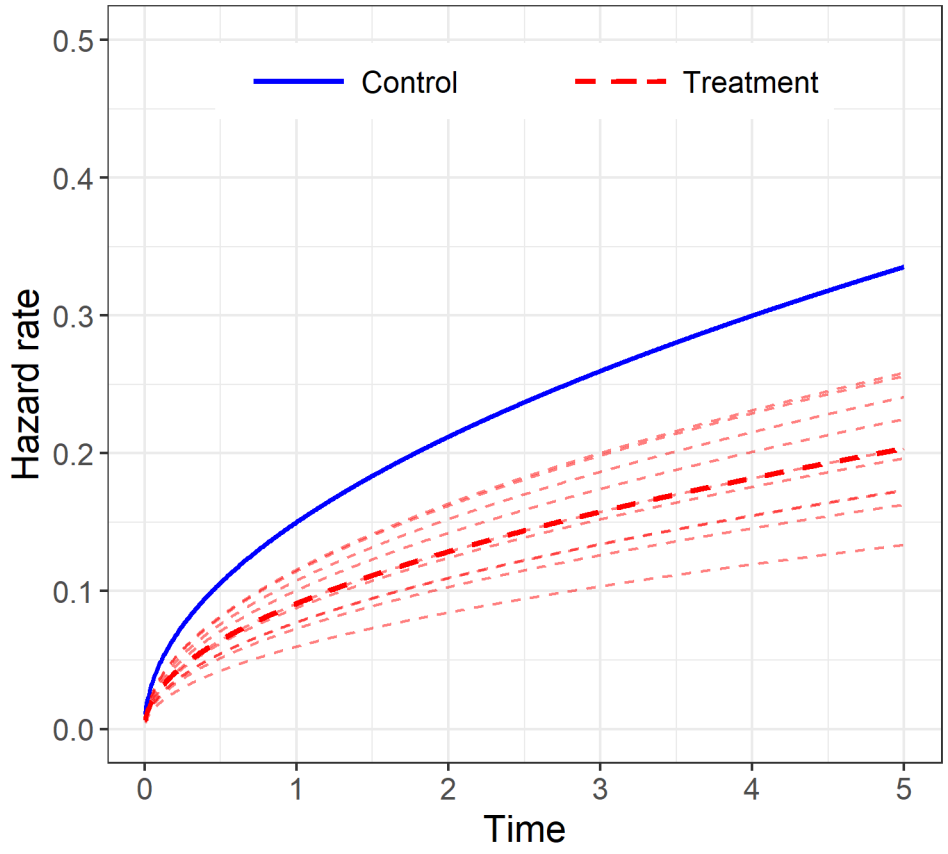
$\lambda = 0.1$ (scale parameter)

$\gamma = 1.5$ (shape parameter)

$\beta = -0.5$ (population average treatment effect)

$b_j \sim N(0, 0.2)$ (study-specific deviation)

Example 4: Clustered survival times




```
# Dimensions
n <- 50      # number of patients per study
J <- 200     # total number of studies
N <- n * J   # total number of patients

# Define covariate data
covs <- data.frame(id    = 1:N,
                   study = rep(1:J, each = n),
                   trt   = rbinom(N, 1, 0.5))

# Define true coefficients
trt_j <- -0.5 + rnorm(J, 0, 0.2)
pars  <- data.frame(trt = rep(trt_j, each = n))

# Simulate the event times
times <- simsurv(dist    = 'weibull',
                 lambdas = 0.1,
                 gammas  = 1.5,
                 x       = covs,
                 betas   = pars)
```

Summary

- The method **only requires that we can specify the hazard** for the data generating model

Summary

- The method **only requires that we can specify the hazard** for the data generating model
- I showed examples for some common scenarios, for which ‘simsurv’ has convenient arguments the user can specify

Summary

- The method **only requires that we can specify the hazard** for the data generating model
- I showed examples for some common scenarios, for which ‘simsurv’ has convenient arguments the user can specify
- I did not demonstrate “user-defined” hazard functions, which can allow even more flexibility
  - e.g. time-varying covariates, joint longitudinal-survival models, Royston-Parmar models, etc

Summary

- The method **only requires that we can specify the hazard** for the data generating model
- I showed examples for some common scenarios, for which ‘simsurv’ has convenient arguments the user can specify
- I did not demonstrate “user-defined” hazard functions, which can allow even more flexibility
  - e.g. time-varying covariates, joint longitudinal-survival models, Royston-Parmar models, etc
- Computation times are “relatively” fast, e.g.
  - 10,000 event times under a standard Weibull distribution (< 1 sec)
  - 10,000 event times under a user-defined hazard function (~10 sec)

Summary

- The method **only requires that we can specify the hazard** for the data generating model
- I showed examples for some common scenarios, for which ‘simsurv’ has convenient arguments the user can specify
- I did not demonstrate “user-defined” hazard functions, which can allow even more flexibility
  - e.g. time-varying covariates, joint longitudinal-survival models, Royston-Parmar models, etc
- Computation times are “relatively” fast, e.g.
  - 10,000 event times under a standard Weibull distribution (< 1 sec)
  - 10,000 event times under a user-defined hazard function (~10 sec)
- Future work: competing risks, vectorisation of ‘uniroot’

Thank you!

Acknowledgements

- My supervisors: Rory Wolfe, Margarita Moreno-Betancur, Michael J. Crowther
- CRAN and useR volunteers!

References

[1] Leemis LM. Variate Generation for Accelerated Life and Proportional Hazards Models. Operations Research, 1987: 35(6); 892–894.

[2] Bender R et al. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine. 2005: 24(11); 1713–1723.

[3] Crowther MJ, Lambert PC. Simulating Biologically Plausible Complex Survival Data. *Statistics in Medicine*, 2013: **32**(23); 4118–4134.

[4] Crowther MJ, Lambert PC. Simulating Complex Survival Data. *The Stata Journal*, 2012: **12**(4); 674–687.

[5] Brilleman S. (2018) simsurv: Simulate Survival Data. R package version 0.2.2. https://CRAN.R-project.org/package=simsurv