A novel approach to competing risks analysis using case-base sampling

Maxime Turgeon

June 10th, 2017

McGill University
Department of Epidemiology, Biostatistics, and Occupational Health

Acknowledgements

This project is joint work with:

- Sahir Bhatnagar
- Olli Saarela (U. Toronto)
- Jim Hanley

Introduction

Motivation

Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia

Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
- "What is my 5-year risk of relapse?"

Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
- "What is my 5-year risk of relapse?"
    - $P(\text{Time to event} < 5, \textbf{Relapse} \mid \text{Covariates})$

Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
- "What is my 5-year risk of relapse?"
  - $P(\text{Time to event} < 5, \textbf{Relapse} \mid \text{Covariates})$
- "What about 1-year? 2-year?"

Motivation

- Jane Doe, 35 yo, received stem-cell transplant for acute myeloid leukemia
- "What is my 5-year risk of relapse?"
  - $P(\text{Time to event} < 5, \textbf{Relapse} \mid \text{Covariates})$
- "What about 1-year? 2-year?"
  - A **smooth** absolute risk curve.

Current methods

Current methods

- Proportional hazards hypothesis

Current methods

- Proportional hazards hypothesis
    - Disease etiology

Current methods

- Proportional hazards hypothesis
  - Disease etiology
  - E.g. Cox regression.

Current methods

- Proportional hazards hypothesis
  - Disease etiology
  - E.g. Cox regression.
- Proportional subdistribution hypothesis

Current methods

- Proportional hazards hypothesis
  - Disease etiology
  - E.g. Cox regression.
- Proportional subdistribution hypothesis
  - Absolute risk

Current methods

- Proportional hazards hypothesis
  - Disease etiology
  - E.g. Cox regression.
- Proportional subdistribution hypothesis
  - Absolute risk
  - E.g. Fine-Gray model.

Summary

Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.

Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.
    - Our approach relies on Hanley & Miettinen's **case-base sampling** method [1].

Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.
    - Our approach relies on Hanley & Miettinen's **case-base sampling** method [1].
- Smooth hazards give rise to smooth absolute risk curves.

Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.
    - Our approach relies on Hanley & Miettinen's **case-base sampling** method [1].
- Smooth hazards give rise to smooth absolute risk curves.
- Our approach allows for a **symmetric** treatment of all time variables.

Summary

- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.
    - Our approach relies on Hanley & Miettinen's **case-base sampling** method [1].
- Smooth hazards give rise to smooth absolute risk curves.
- Our approach allows for a **symmetric** treatment of all time variables.
- Finally, it also allows for **hypothesis testing** and **variable selection**.

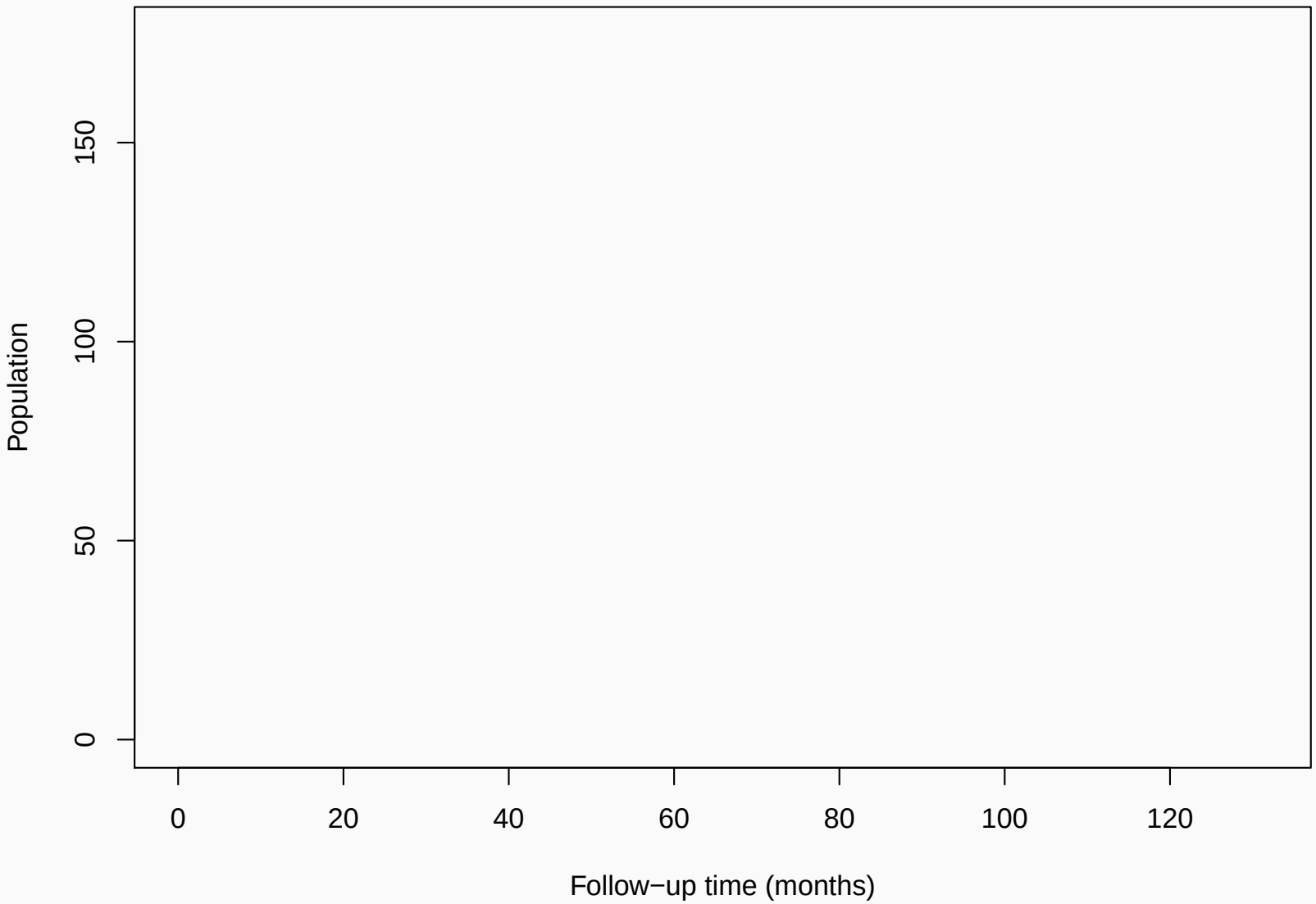
Summary

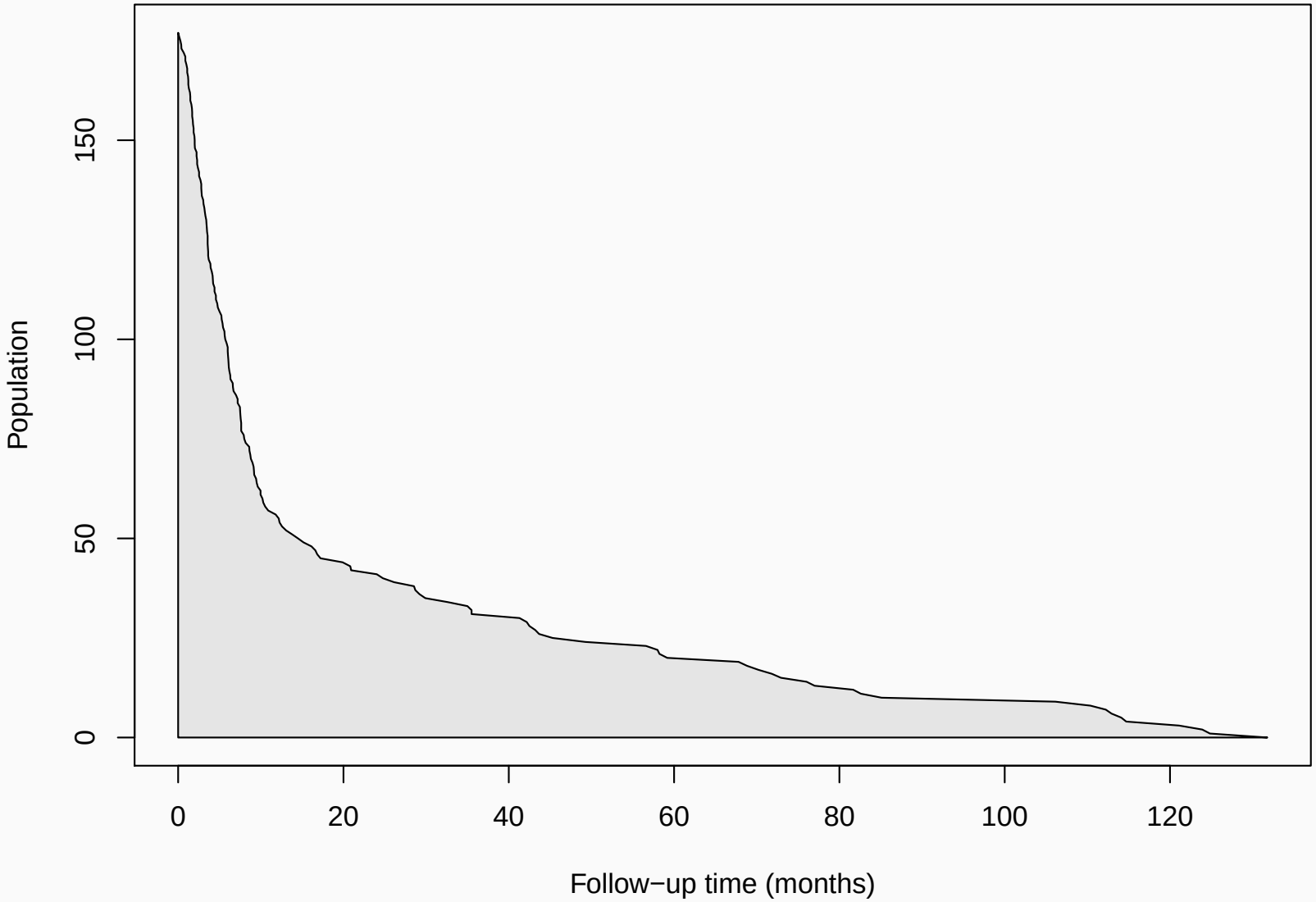
- We propose a **simple** approach to modeling **directly** the cause-specific hazards using (smooth) parametric families.
  - Our approach relies on Hanley & Miettinen's **case-base sampling** method [1].
- Smooth hazards give rise to smooth absolute risk curves.
- Our approach allows for a **symmetric** treatment of all time variables.
- Finally, it also allows for **hypothesis testing** and **variable selection**.

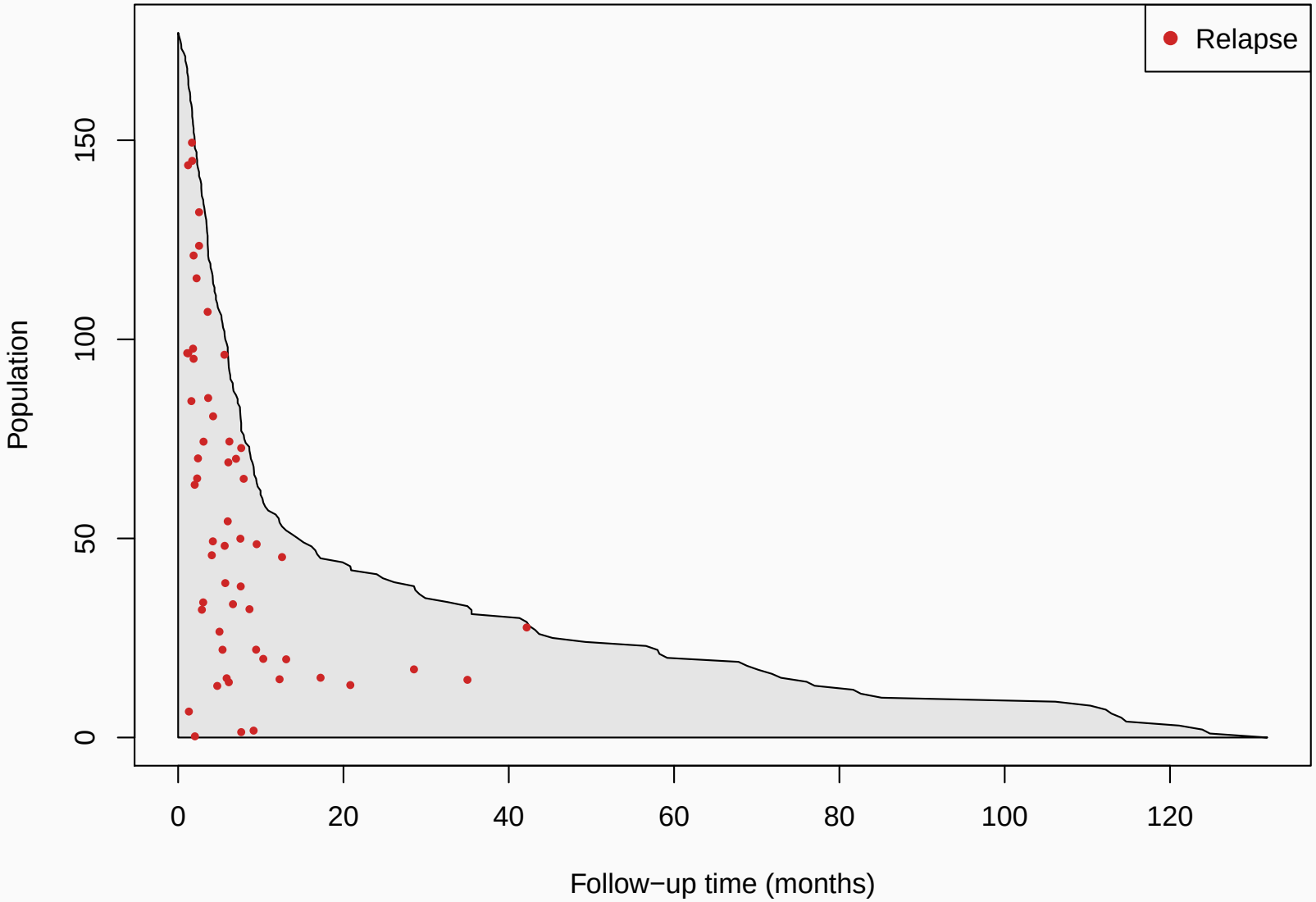
This method is currently available in the R package `casebase` on CRAN.

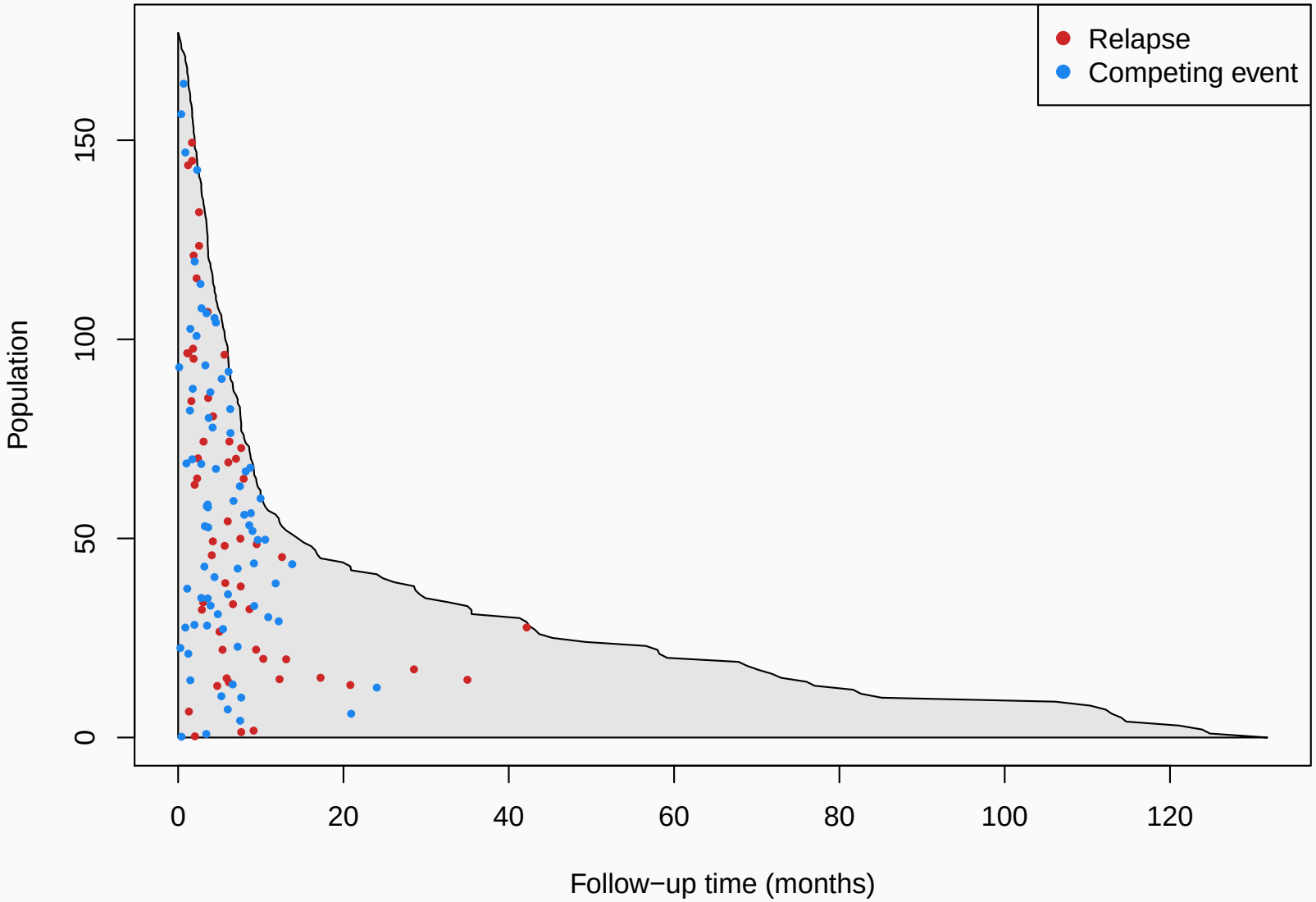
See also our website: http://sahirbhatnagar.com/casebase/

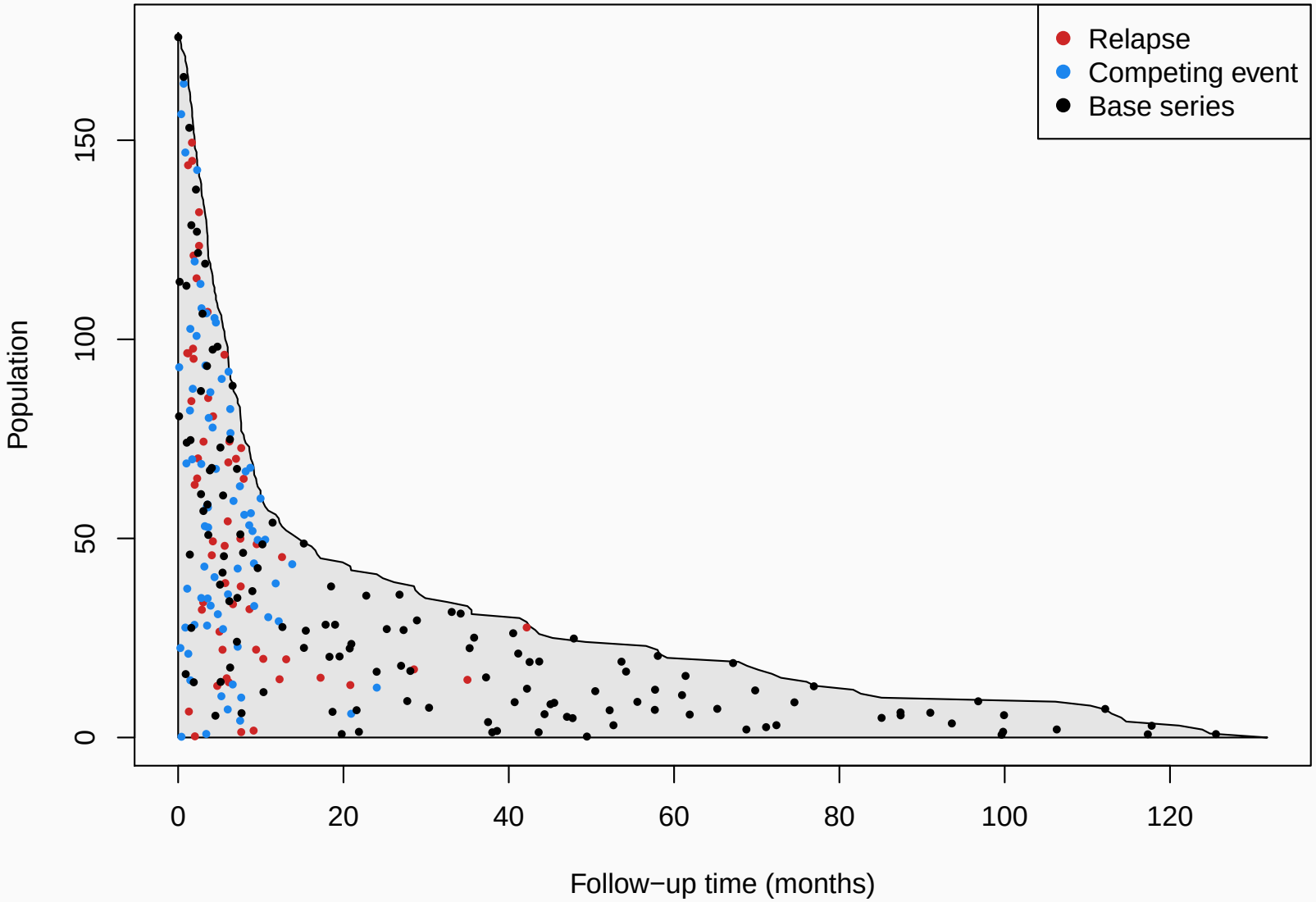
Case-base sampling


Population
0
50
100
150
0
20
40
60
80
100
120
Follow−up time (months)


Population
0
50
100
150
0
20
40
60
80
100
120
Follow−up time (months)


Relapse
Population
0
50
100
150
0
20
40
60
80
100
120
Follow−up time (months)


Relapse
Competing event
Population
0
50
100
150
0
20
40
60
80
100
120
Follow−up time (months)


Relapse
Competing event
Base series
Population
0
50
100
150
0
20
40
60
80
100
120
Follow−up time (months)

Case-base sampling

Case-base sampling

- The unit of analysis is a *person-moment*.

Case-base sampling

- The unit of analysis is a *person-moment*.
- Case-base sampling reduces the model fitting to a familiar multinomial regression.

Case-base sampling

- The unit of analysis is a *person-moment*.
- Case-base sampling reduces the model fitting to a familiar multinomial regression.
    - The sampling process is taken into account using an offset term.

Case-base sampling

- The unit of analysis is a *person-moment*.
- Case-base sampling reduces the model fitting to a familiar multinomial regression.
    - The sampling process is taken into account using an offset term.
- By sampling a large base series, the information loss eventually becomes negligible.

Case-base sampling

- The unit of analysis is a *person-moment*.
- Case-base sampling reduces the model fitting to a familiar multinomial regression.
    - The sampling process is taken into account using an offset term.
- By sampling a large base series, the information loss eventually becomes negligible.
- This framework can easily be used with time-varying covariates (e.g. time-varying exposure).

Theoretical details

Assumptions

We make the following assumptions:

Assumptions

We make the following assumptions:

- For each event type $j = 1, \ldots, m$, a non-homogeneous Poisson process with hazard $\lambda_j(t)$.

Assumptions

We make the following assumptions:

- For each event type $j = 1, \ldots, m$, a non-homogeneous Poisson process with hazard $\lambda_j(t)$.
    - At most one event type can occur.

Assumptions

We make the following assumptions:

- For each event type $j = 1, \ldots, m$, a non-homogeneous Poisson process with hazard $\lambda_j(t)$.
  - At most one event type can occur.
- Non-informative censoring.

Assumptions

We make the following assumptions:

- For each event type $j = 1, \ldots, m$, a non-homogeneous Poisson process with hazard $\lambda_j(t)$.
  - At most one event type can occur.
- Non-informative censoring.
- Case-base sampling occurs following a non-homogenous Poisson process with hazard $\rho(t)$.

Likelihood

Each person-moment's contribution to the likelihood is of the form:

$$\prod_{j=1}^{m} \frac{\lambda_j(t)^{dN_j(t)}}{\rho(t) + \sum_{j=1}^{m} \lambda_j(t)}.$$

Likelihood

Each person-moment's contribution to the likelihood is of the form:

$$\prod_{j=1}^{m} \frac{\lambda_j(t)^{dN_j(t)}}{\rho(t) + \sum_{j=1}^{m} \lambda_j(t)}.$$

This is reminiscent of a multinomial likelihood, with offset $\log(1/\rho(t))$**.**

Main Theorem

Likelihood

Main Theorem

The likelihood defined above has mean zero and is asymptotically normal.

Likelihood

Main Theorem

The likelihood defined above has mean zero and is asymptotically normal.

Implication: All the GLM machinery (e.g. deviance tests, information criteria, regularization) is available to us.

Parametric families

Parametric families

We can fit any model of the following form:

$$\log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X.$$

Parametric families

We can fit any model of the following form:

$$\log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X.$$

Different choices of the function g leads to familiar parametric families:

Parametric families

We can fit any model of the following form:

$$\log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X.$$

Different choices of the function g leads to familiar parametric families:

- Exponential: g is constant.

Parametric families

We can fit any model of the following form:

$$\log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X.$$

Different choices of the function g leads to familiar parametric families:

- Exponential: g is constant.
- Gompertz: $g(t; \alpha) = \alpha t$.

Parametric families

We can fit any model of the following form:

$$\log \lambda(t; \alpha, \beta) = g(t; \alpha) + \beta X.$$

Different choices of the function g leads to familiar parametric families:

- Exponential: g is constant.
- Gompertz: $g(t; \alpha) = \alpha t$.
- Weibull: $g(t; \alpha) = \alpha \log t$.

Simulation study

Simulation scenario

- We simulate 1000 datasets from an exponential and a Gompertz family.

Simulation scenario

- We simulate 1000 datasets from an exponential and a Gompertz family.
- Binary covariate

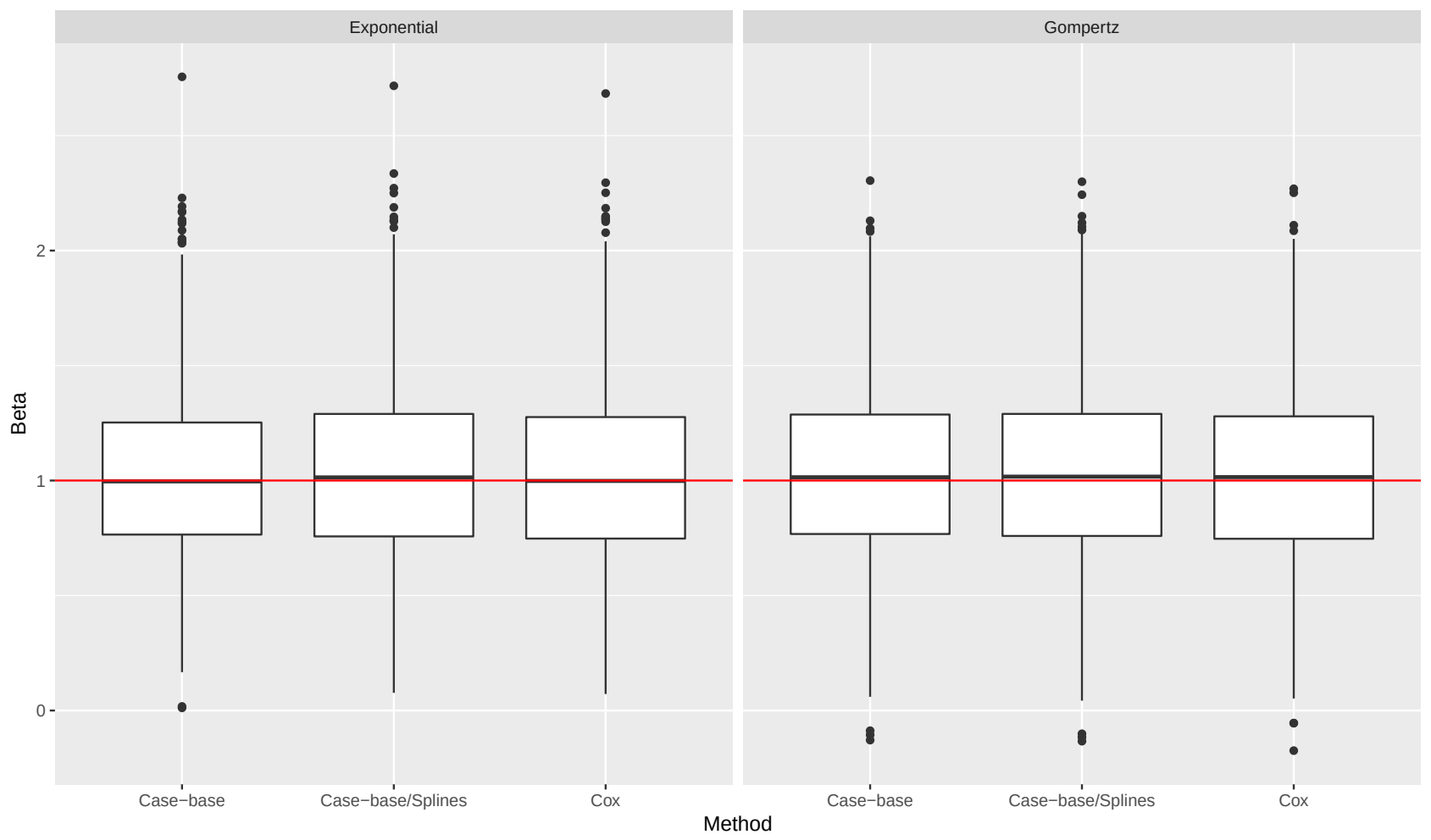
Simulation scenario

- We simulate 1000 datasets from an exponential and a Gompertz family.
- Binary covariate
- Random censoring

Simulation scenario

- We simulate 1000 datasets from an exponential and a Gompertz family.
- Binary covariate
- Random censoring
- We compare case-base with a correctly specified family, case-base with splines, and Cox regression.

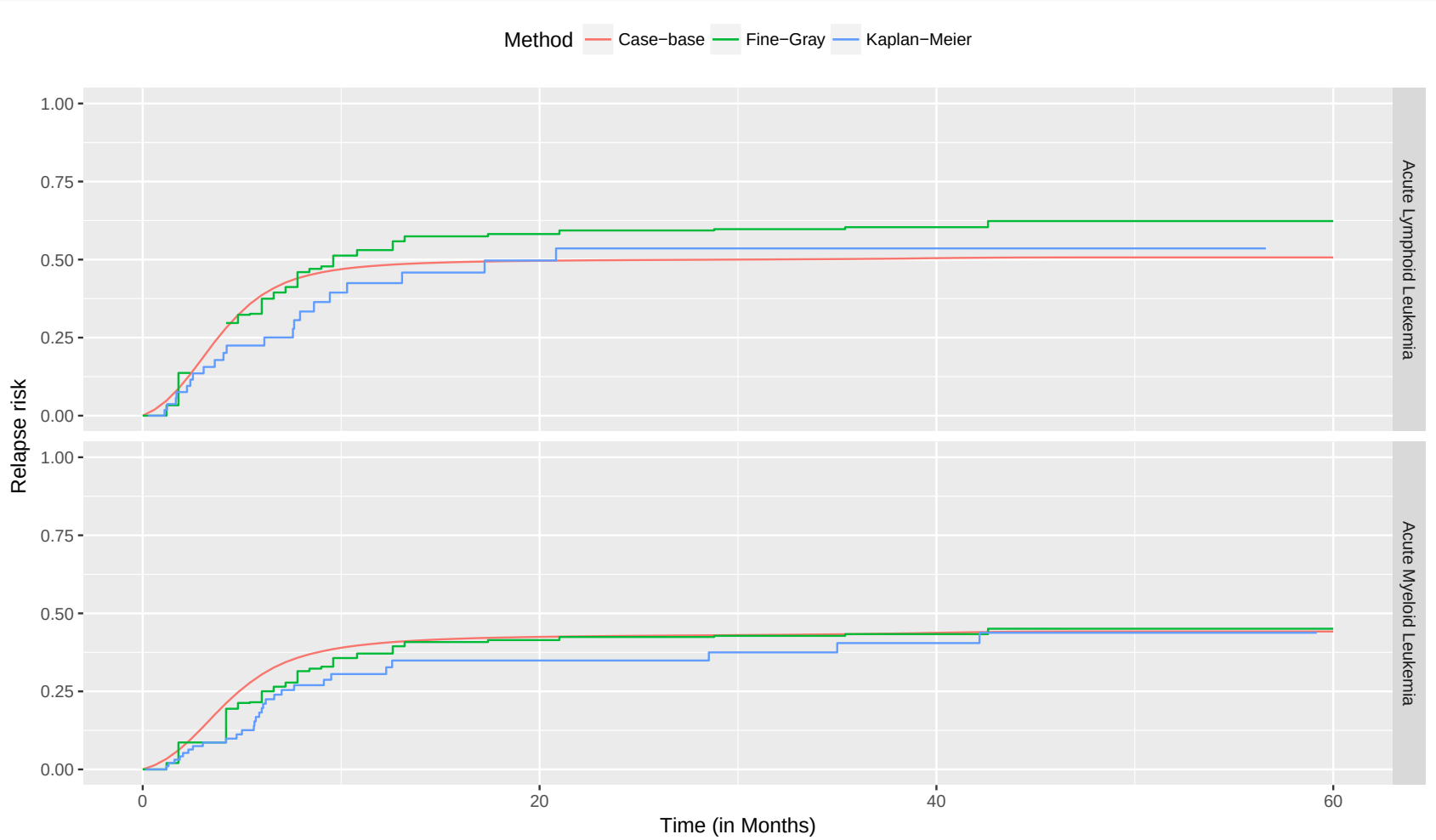
Simulation results



Data analysis

Data

Variable description	Statistical summary
Sex	M=Male (87)
	F=Female (72)
Disease	ALL (59)
	AML (100)
Phase	CR1 (43)
	CR2 (40)
	CR3 (10)
	Relapse (65)
Type of transplant	BM+PB (15)
	PB (144)
Age of patient (years)	16–62
	33 (IQR 19.5)
Failure time (months)	0.13–131.77
	20.28 (30.78)
Status indicator	0=censored (40)
	1=relapse (49)
	2=competing event (70)



Absolute risk for female patient, median age, in relapse at transplant (stem cells from peripheral blood).

Model fit

	Case-base		Cox regression	
Variable	Hazard ratio	95% CI	Hazard ratio	95% CI
Sex	0.64	(0.35, 1.20)	0.75	(0.42, 1.35)
Disease	0.54	(0.27, 1.07)	0.63	(0.34, 1.19)
Phase CR2	1.00	(0.37, 2.70)	0.95	(0.36, 2.51)
Phase CR3	1.25	(0.24, 6.53)	1.38	(0.28, 6.76)
Phase Relapse	4.71	(2.11, 10.54)	4.06	(1.85, 8.92)
Source	1.89	(0.40, 8.99)	1.49	(0.32, 6.85)
Age	0.99	(0.97, 1.02)	0.99	(0.97, 1.02)

Discussion

Discussion

Discussion

- We proposed a simple and flexible way of directly modeling the hazard function, using **multinomial regression**.

Discussion

- We proposed a simple and flexible way of directly modeling the hazard function, using **multinomial regression**.
    - This leads to smooth estimates of the absolute risks.

Discussion

- We proposed a simple and flexible way of directly modeling the hazard function, using **multinomial regression**.
    - This leads to smooth estimates of the absolute risks.
- We are explicitely modeling time.

Discussion

- We proposed a simple and flexible way of directly modeling the hazard function, using **multinomial regression**.
    - This leads to smooth estimates of the absolute risks.
- We are explicitely modeling time.
- We can test the significance of covariates.

References I

J. A. Hanley and O. S. Miettinen.
Fitting smooth-in-time prognostic risk functions via logistic regression.
The International Journal of Biostatistics, 5(1), 2009.

O. Saarela.
A case-base sampling method for estimating recurrent event intensities.
Lifetime data analysis, pages 1–17, 2015.

O. Saarela and J. A. Hanley.
Case-base methods for studying vaccination safety.
Biometrics, 71(1):42–52, 2015.

References II

L. Scrucca, A. Santucci, and F. Aversa.
Regression modeling of competing risk using R: an in depth guide for clinicians.
Bone marrow transplantation, 45(9):1388–1395, 2010.

Questions or comments?

For more details, visit
http://sahirbhatnagar.com/casebase/