`casebase`: an alternative framework for survival analysis

Max Turgeon

November 26th, 2019

University of Manitoba
Departments of Statistics and Computer Science

Acknowledgements

This project is joint work with:

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

Outline

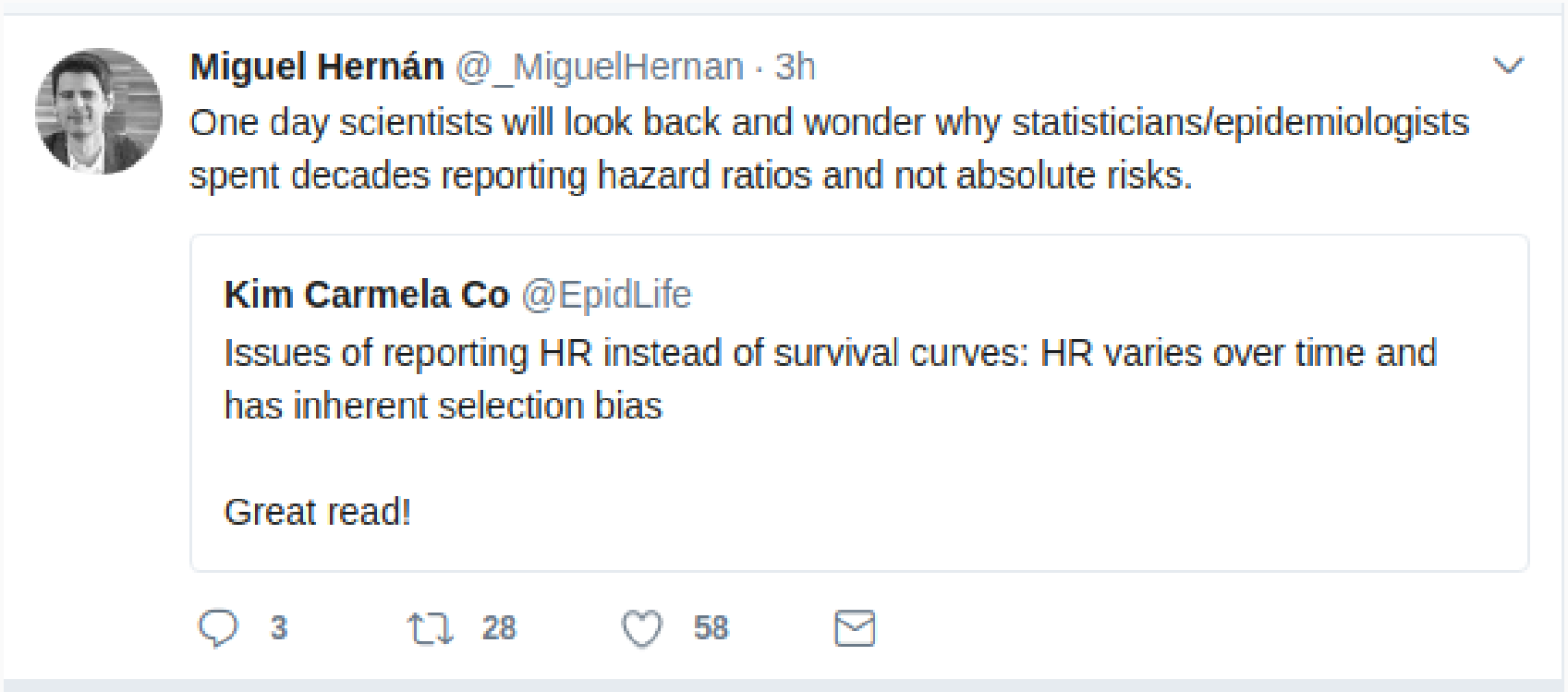
1. Overview of case-base sampling
2. Presentation of `R` package `casebase`
    - Case studies
3. Current and future directions

Introduction

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.

Motivation



Hernán, M.A. (2010). The hazards of hazard ratios.

Risk-set sampling

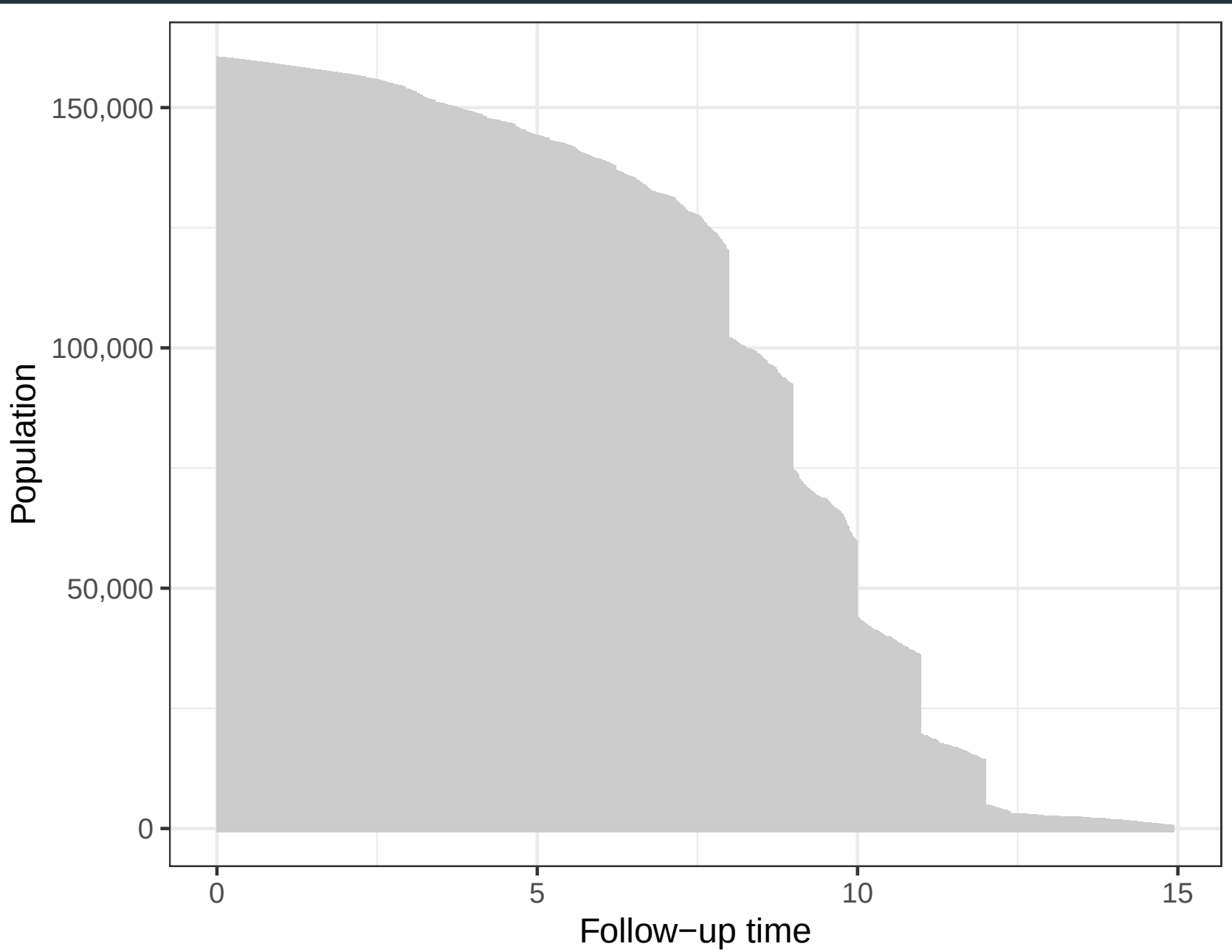
- Methods like Cox regression use **risk-set sampling**.
  - When an event occurs, we compare the individual with other individuals in our study that were "at risk" at that time point.
- By matching on time, these methods eliminate the baseline hazard (nuisance parameter)
- Absolute risks, cumulative incidence functions and survival functions can be recovered using the semi-parametric Breslow estimator.
  - This leads to **step functions**.

Summary

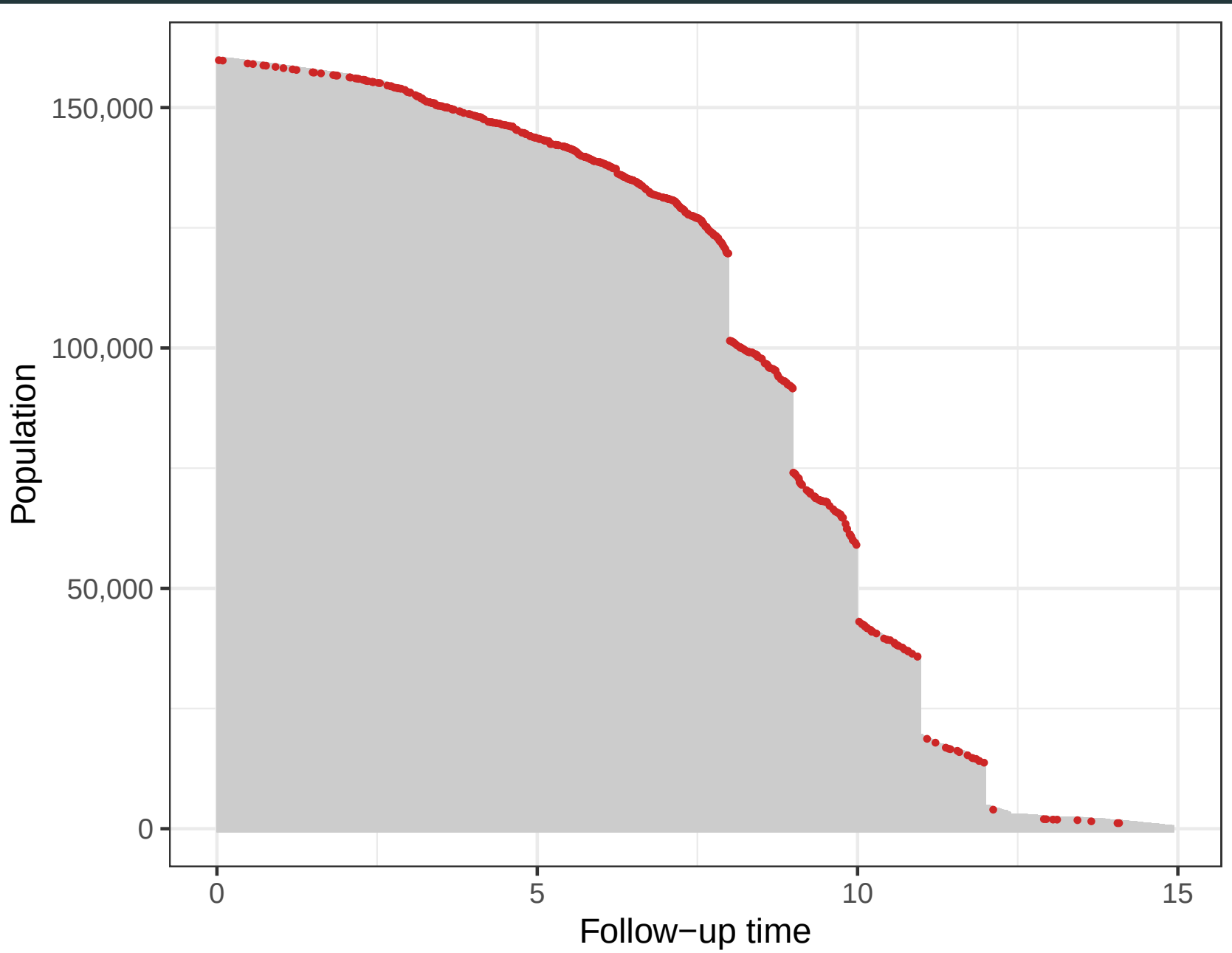
- Case-base sampling is a **simple** approach to modelling **directly** the hazards using (smooth) parametric families.
  - Introduced by Hanley & Miettinen [1], based on ideas by Mantel [2].
- Smooth hazards give rise to smooth absolute risk curves.
- This approach was implemented in the `R` package `casebase`.
  - Available on CRAN.
- See also our website: http://sahirbhatnagar.com/casebase/

Case-base sampling

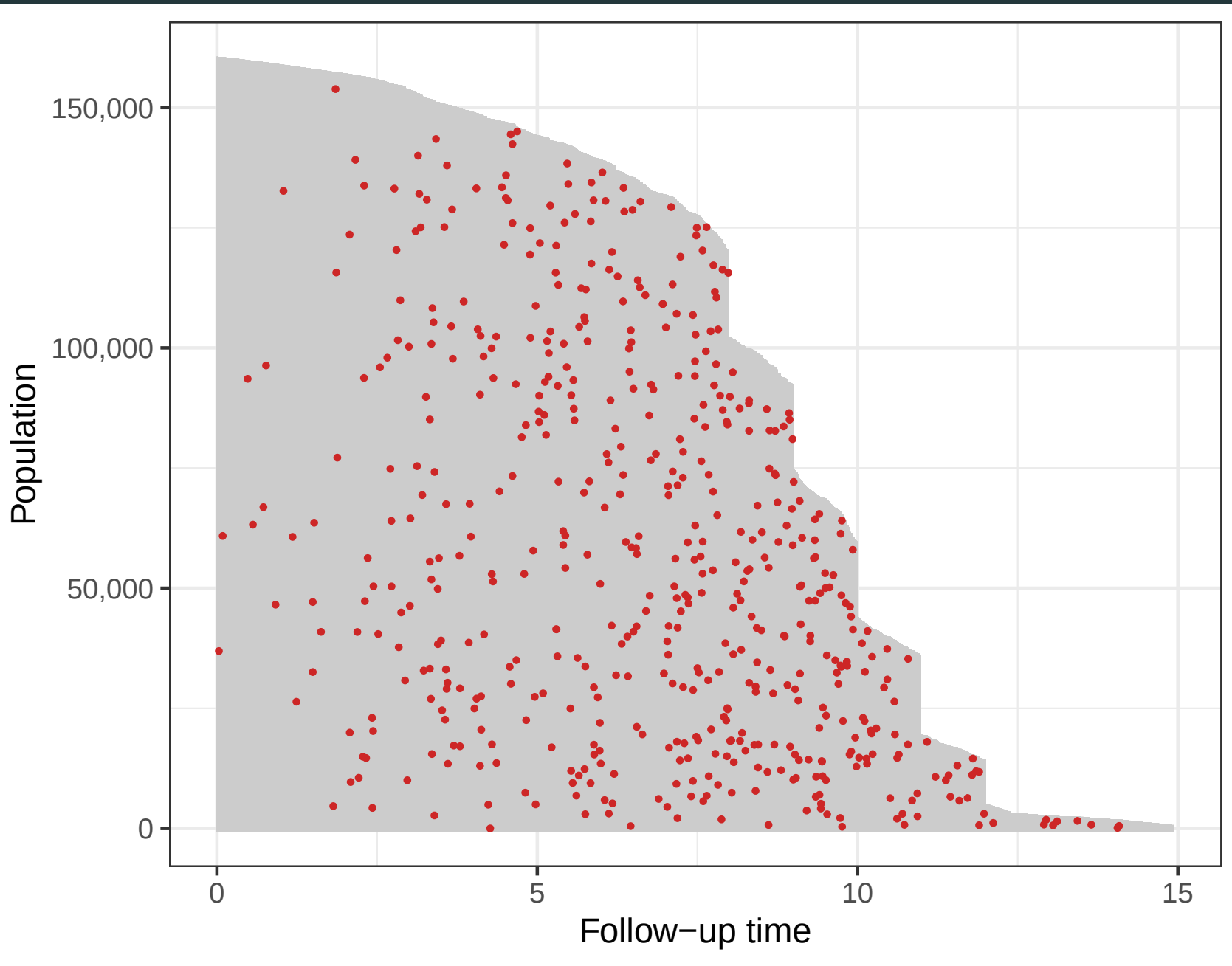
Population-Time plots



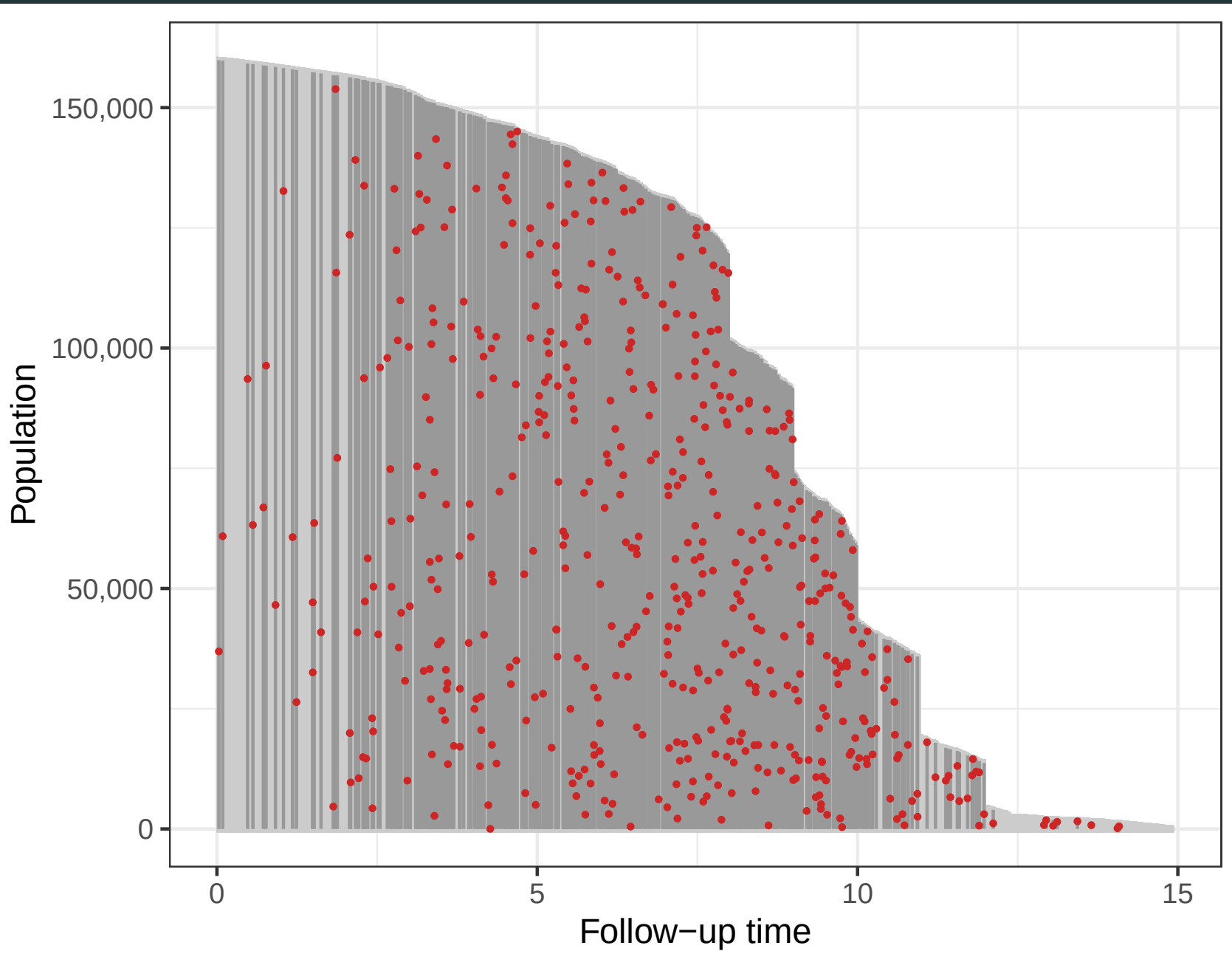
Population-Time plots



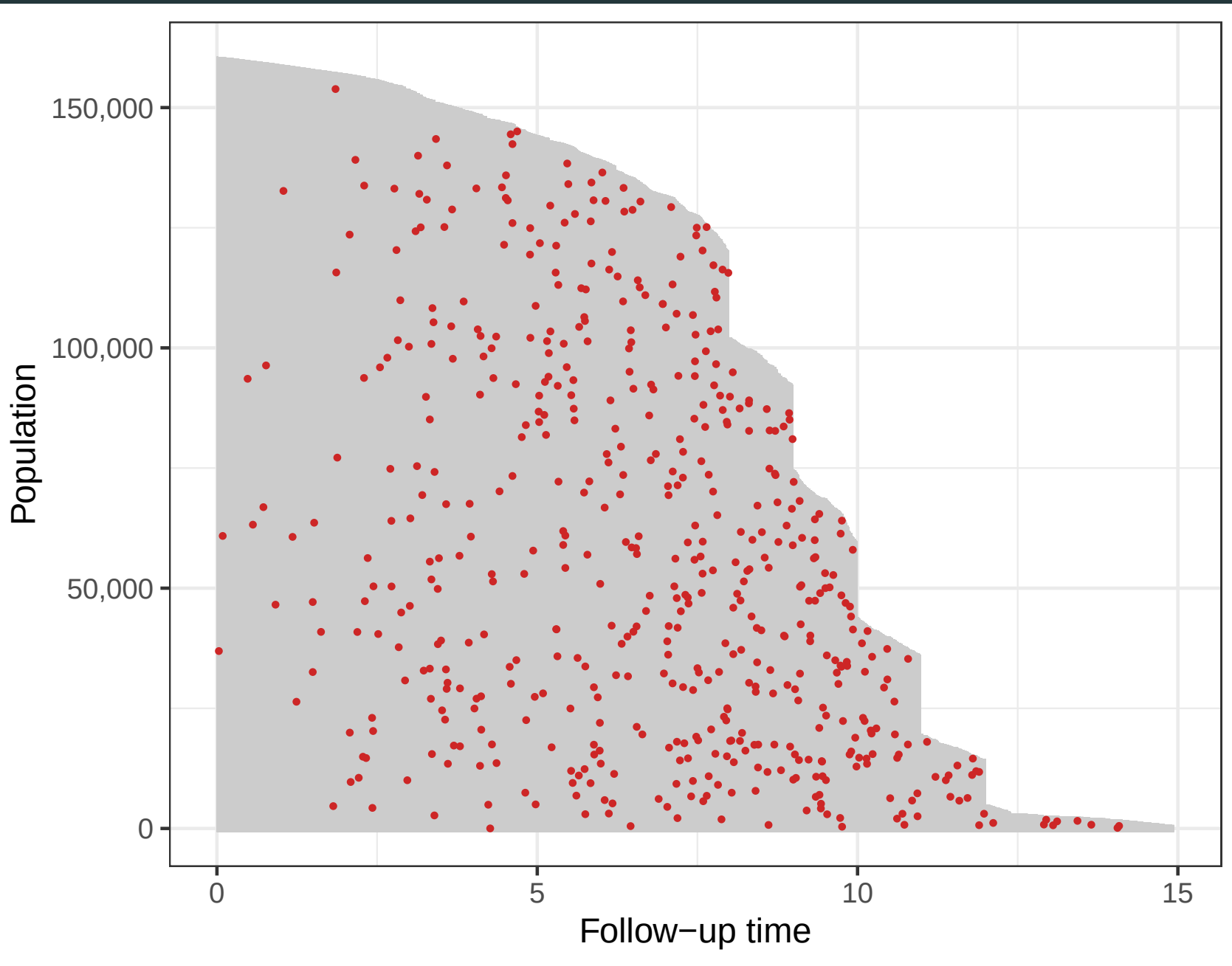
Population-Time plots



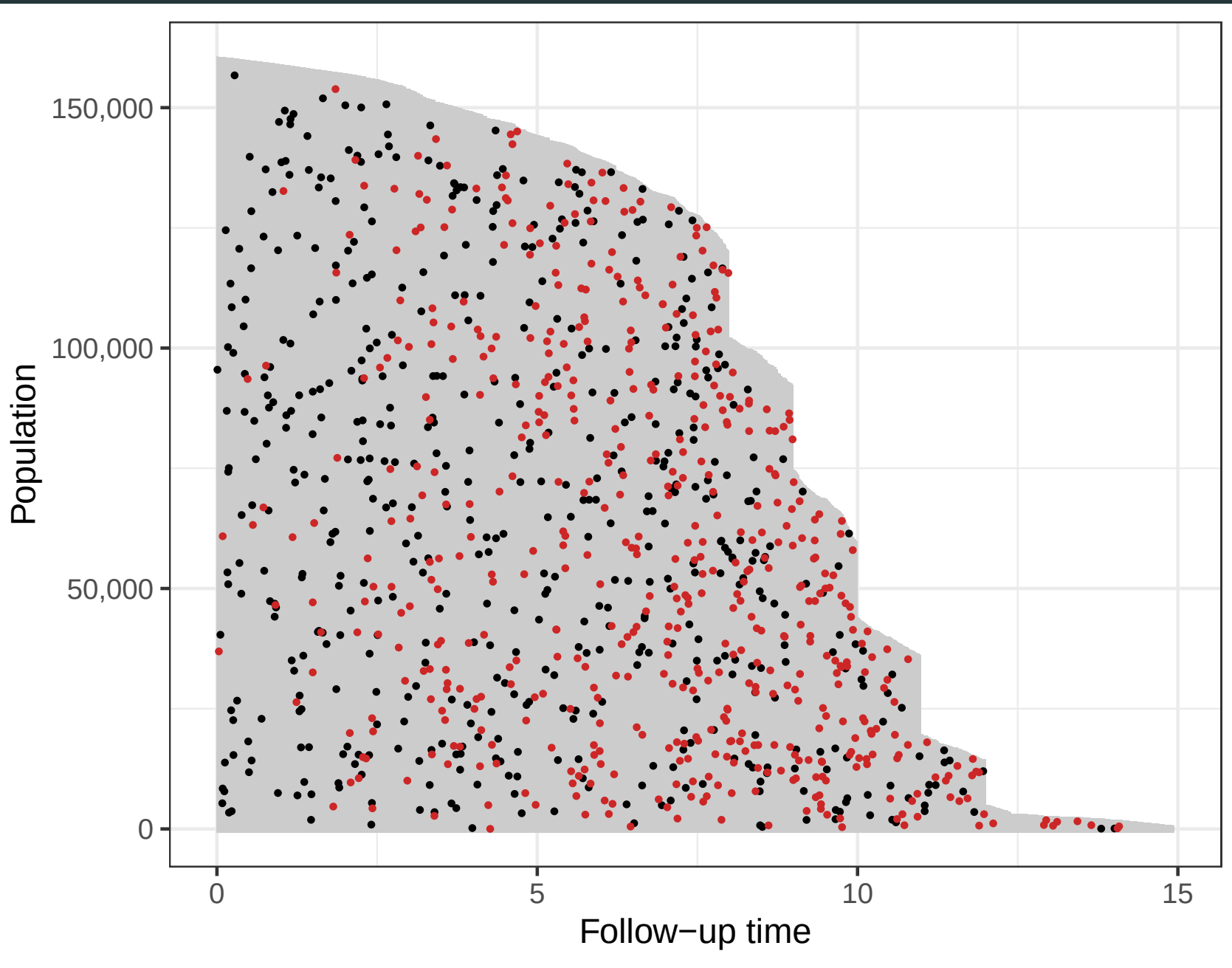
Population-Time plots



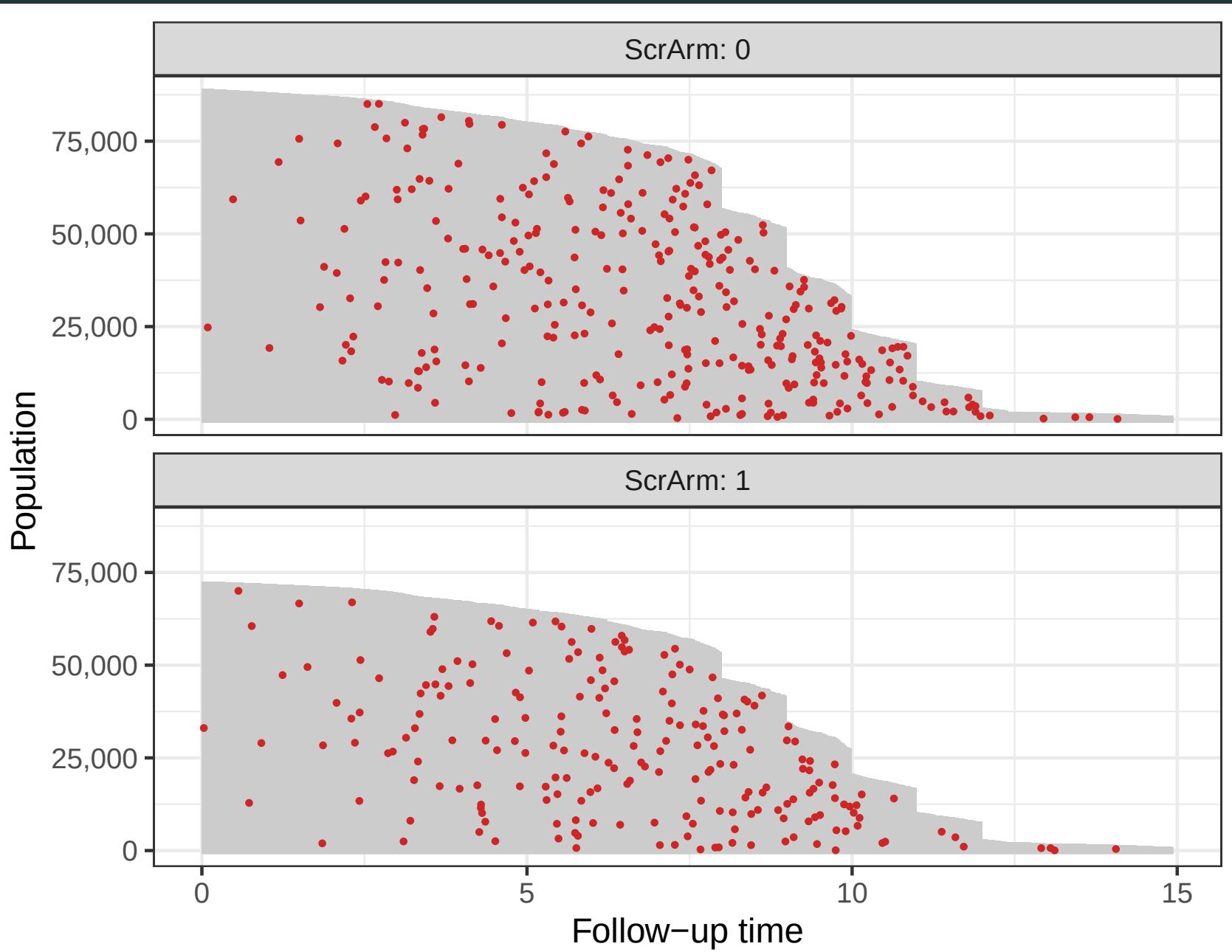
Population-Time plots



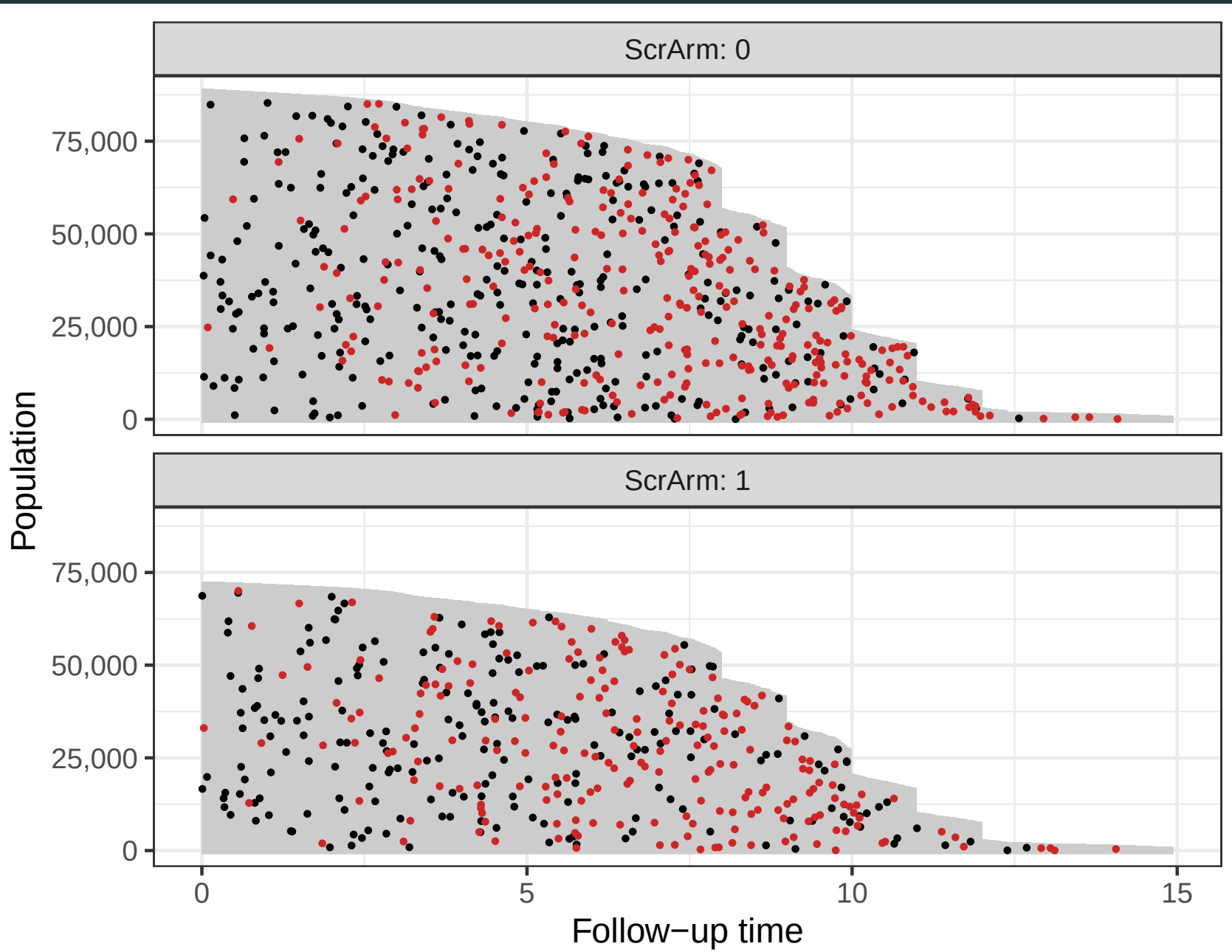
Population-Time plots



Population-Time plots



Population-Time plots



Case-base sampling–Overview

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

Parametric families

We can fit any hazard λ 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: $g(t; \alpha) = \alpha \log t$.

From hazards to risks

- Once we have an estimate $\hat{\lambda}(t)$ of the hazard, we can get an estimate of the survival function:

$$\hat{S}(t) = \exp\left(-\int_0^t \hat{\lambda}(u)du\right).$$

- Similarly, we can get an estimate of the cumulative incidence (i.e. CDF):

$$\widehat{CI}(t) = 1 - \hat{S}(t).$$

Theoretical details

Assumptions

For notational convenience, we will assume Type I censoring (e.g. every subject is followed until the event occurs or the end of the study).

We have two counting processes at play:

- **Event of interest**: A non-homogeneous Poisson process $N(t)$ with hazard $\lambda(t;\theta)$.
- **Case-base sampling**: A non-homogeneous Poisson process $M(t)$ with hazard $\rho(t)$.
    - In most examples, we will sample uniformly (i.e. *homogeneous* Poisson process).

Likelihood

The likelihood for this data-generating mechanism is given by

$$L(\theta) = \prod_{i=1}^{n} \prod_{t \in (0,\tau]} \left(\frac{\lambda_i(t;\theta)^{dN_i(t)}}{\rho_i(t) + \lambda_i(t;\theta)} \right)^{dM_i(t)}.$$

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

Likelihood

Theorem [Saarela (2015)]

- The above likelihood is a partial likelihood for the full data-generating mechanism.
- The corresponding score process has mean zero.
- The corresponding predictable variation process is equal to the observed information process in expectation.

Likelihood

Theorem [Saarela (2015)]

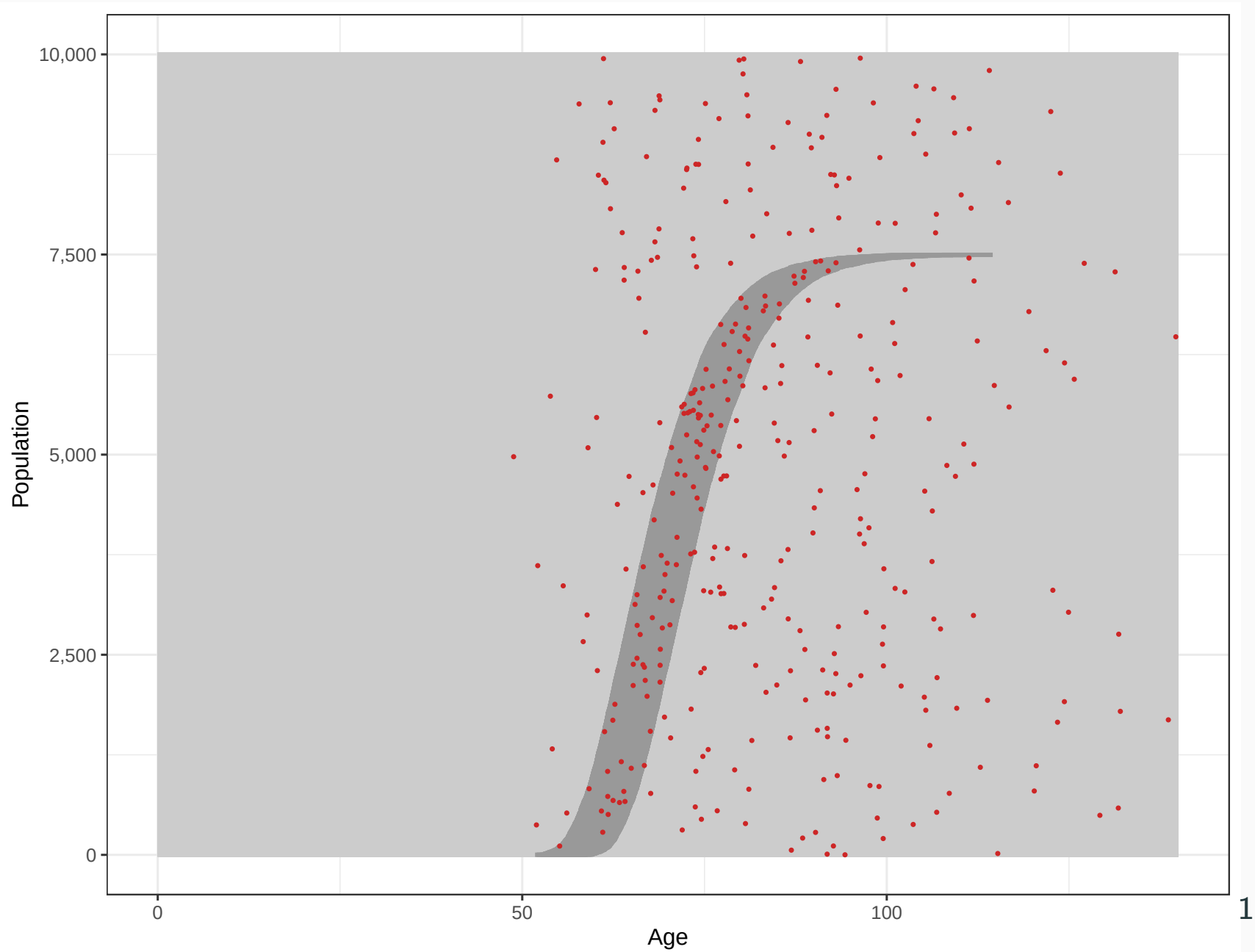
- The above likelihood is a partial likelihood for the full data-generating mechanism.
- The corresponding score process has mean zero.
- The corresponding predictable variation process is equal to the observed information process in expectation.

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

Vaccination safety (Saarela & Hanley, 2015)

- The motivation comes from Patel et al. (2011).
- They studied the potential effect of rotavirus vaccination on intussusception incidence in infants.
- Exposure period is one week after vaccination.

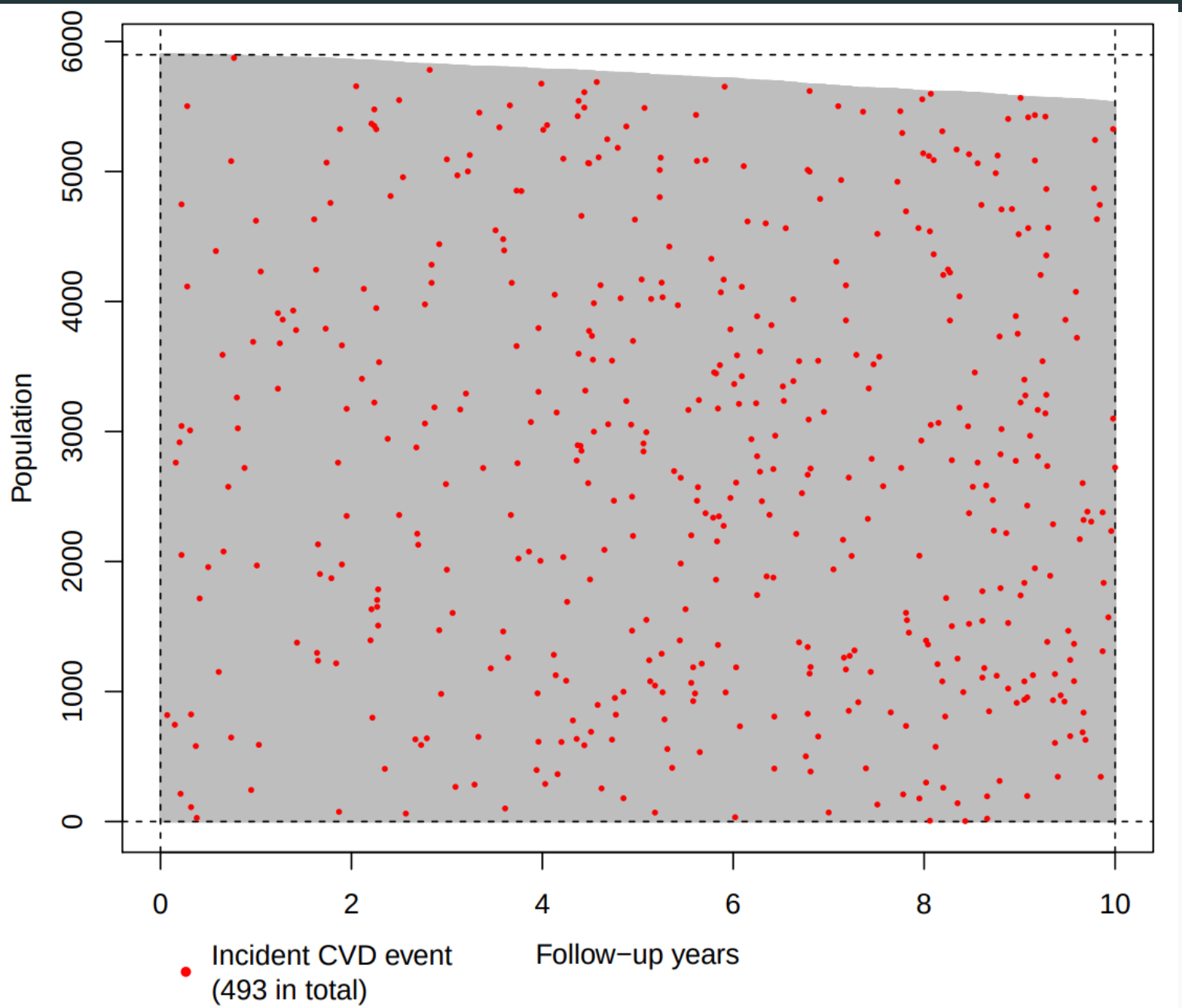
Vaccination safety (Saarela & Hanley, 2015)



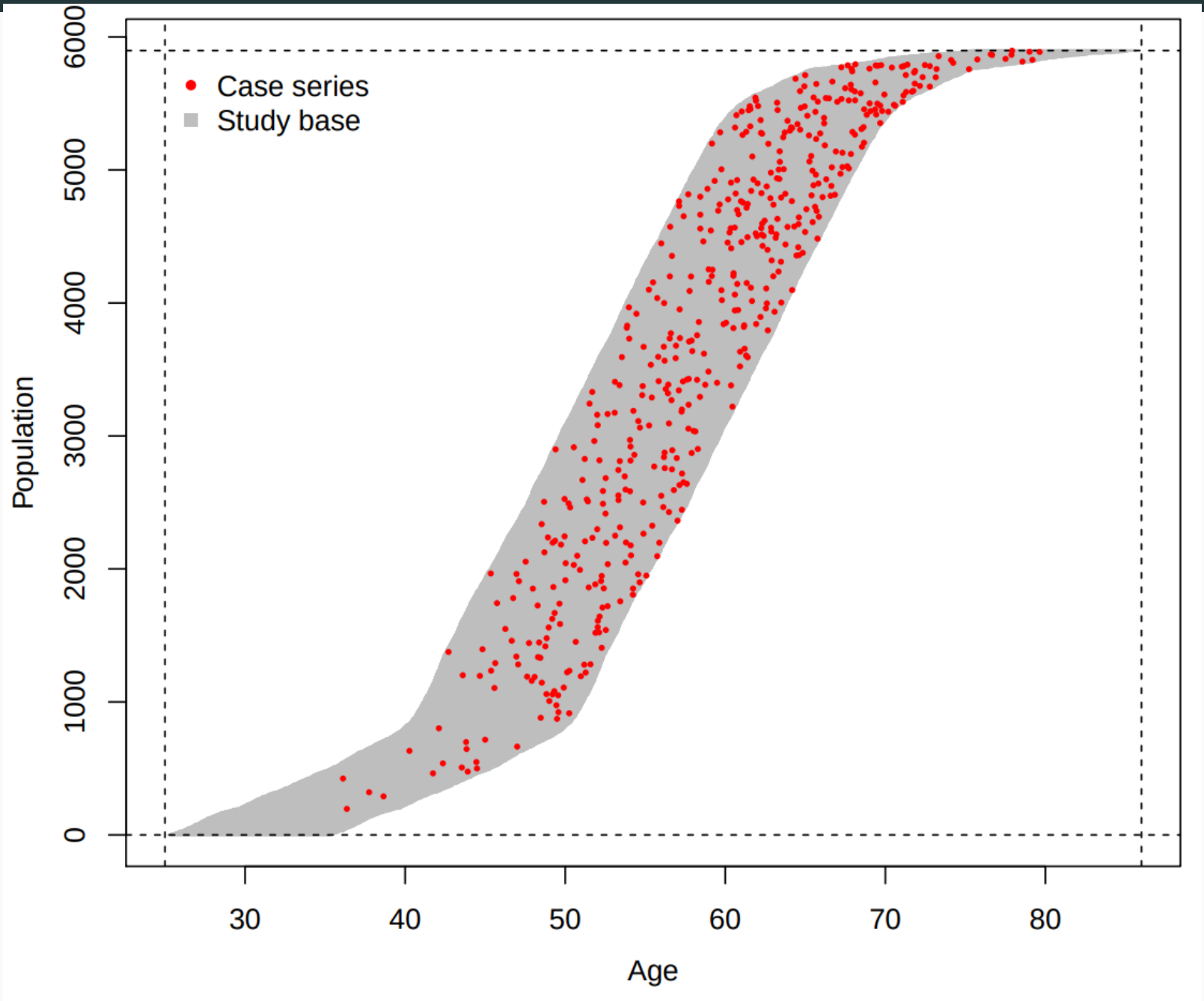
Different time scales

- Study on risk factors for cardio-vascular diseases (CVD)
- Time since enrolment does not have much clinical value...
- With case-base sampling, we can treat all time variables symmetrically.

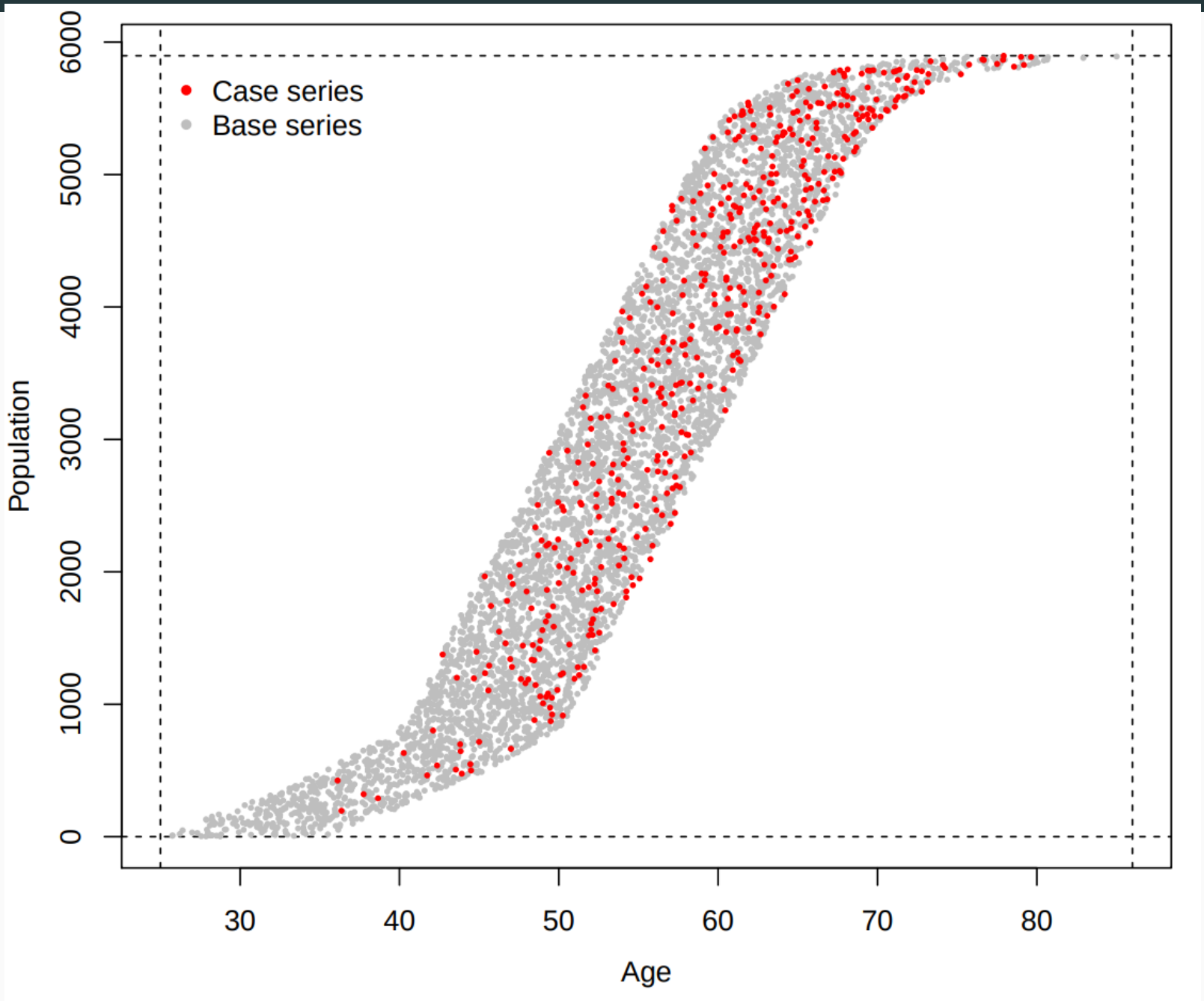
Different time scales



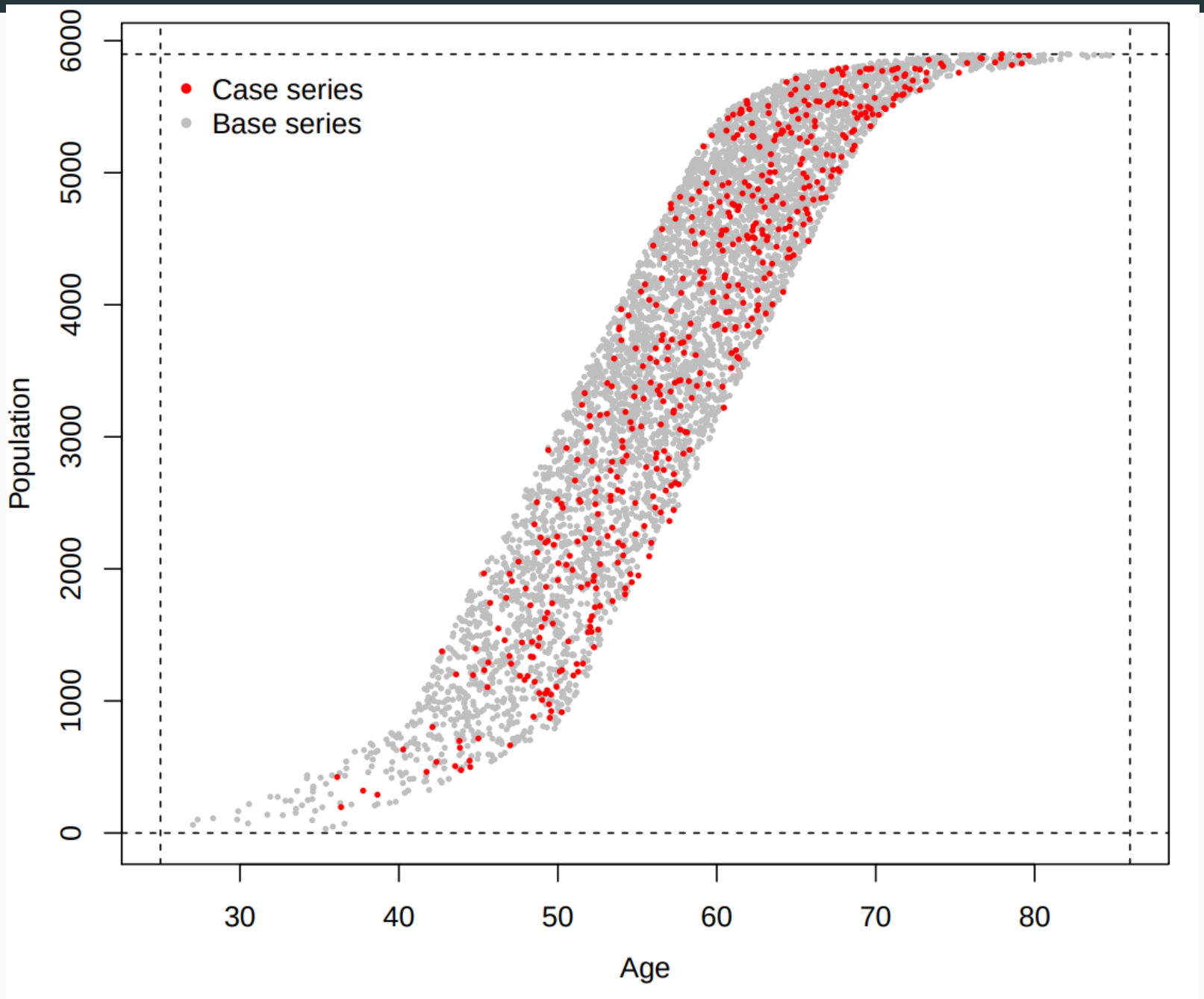
Different time scales



Different time scales



Different time scales



`casebase` **package**

Overview of main functions

There are essentially four main functions in the package:

- `popTime`: Creates `popTime` objects that can be plotted to create population-time plots.
- `sampleCaseBase`: Samples a base series uniformly from the study base.
- `fitSmoothHazard`: Fits a parametric hazard to the data using case-base sampling.
- `absoluteRisk`: Estimates absolute risks (or cumulative incidence functions) from a fitted hazard.

popTime

```
popTime(data, time, event, censored.indicator, exposure)
```

- `time`, `event`: Variable names representing these quantities. If not specified, we try to guess.
- `exposure`: To create stratified population-time plots.

sampleCaseBase

```
sampleCaseBase(data, time, event, ratio = 10,
               comprisk = FALSE, censored.indicator)
```

- `ratio`: Ratio of the size of the base series to the case series (i.e. how many controls for each case?)
- **Note**: Rarely need to call directly.

fitSmoothHazard

```
fitSmoothHazard(formula, data, time,
                family = c("glm", "gam", "gbm", "glmnet"),
                censored.indicator, ratio = 100, ...)

fitSmoothHazard.fit(x, y, formula_time, time, event,
                    family = c("glm", "gbm", "glmnet"),
                    censored.indicator, ratio = 100, ...)
```

- We allow both a formula and a matrix interface.
- We have four different model families:
  - `glm`: Vanilla case-base sampling.
  - `gam`: Generalized additive models.
  - `gbm`: Gradient boosted models (experimental!).
  - `glmnet`: Regularized logistic regression.

absoluteRisk

```
absoluteRisk(object, time, newdata,
             method = c("numerical", "montecarlo"),
             nsamp = 100, onlyMain = TRUE, ...)
```

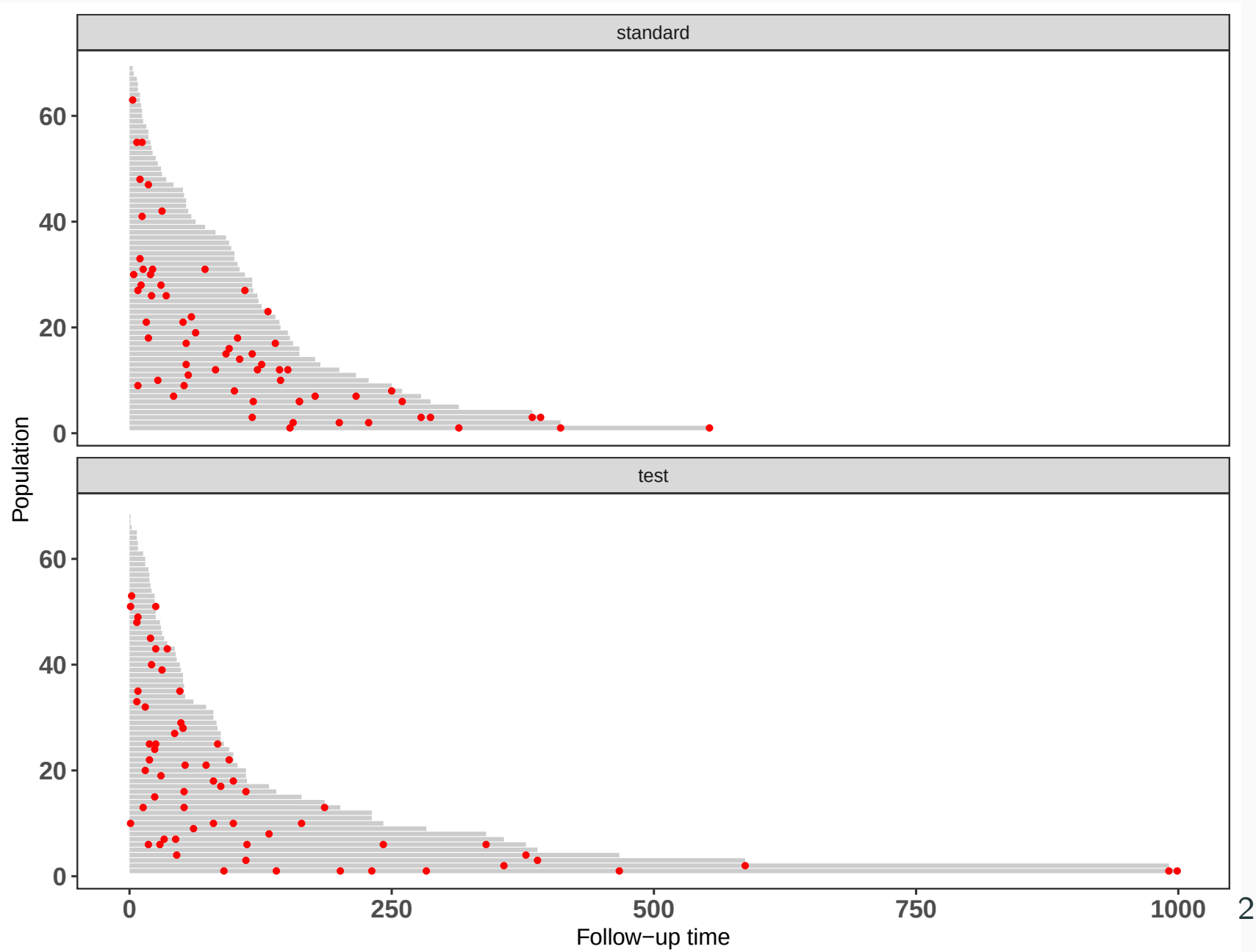
- `time`: Vector of time values at which we compute the risk.
- `method`: Should we use numerical or Montecarlo integration.

Case studies

Case Study I–Veteran data

- Survival data for 137 patients from Veteran's Administration Lung Cancer Trial.
- Patients were randomized to one of two chemotherapy treatments.

Veteran data–Population-Time plot



Veteran data–Model fit

```
phreg(Surv(time, status) ~ karno + diagtime + age +
      prior + celltype + trt,
      data = veteran, shape = 0, dist = "weibull")

fitSmoothHazard(status ~ log(time) + karno + diagtime +
                  age + prior + celltype + trt,
                  data = veteran)

coxph(Surv(time, status) ~ karno + diagtime + age +
      prior + celltype + trt, data = veteran)
```

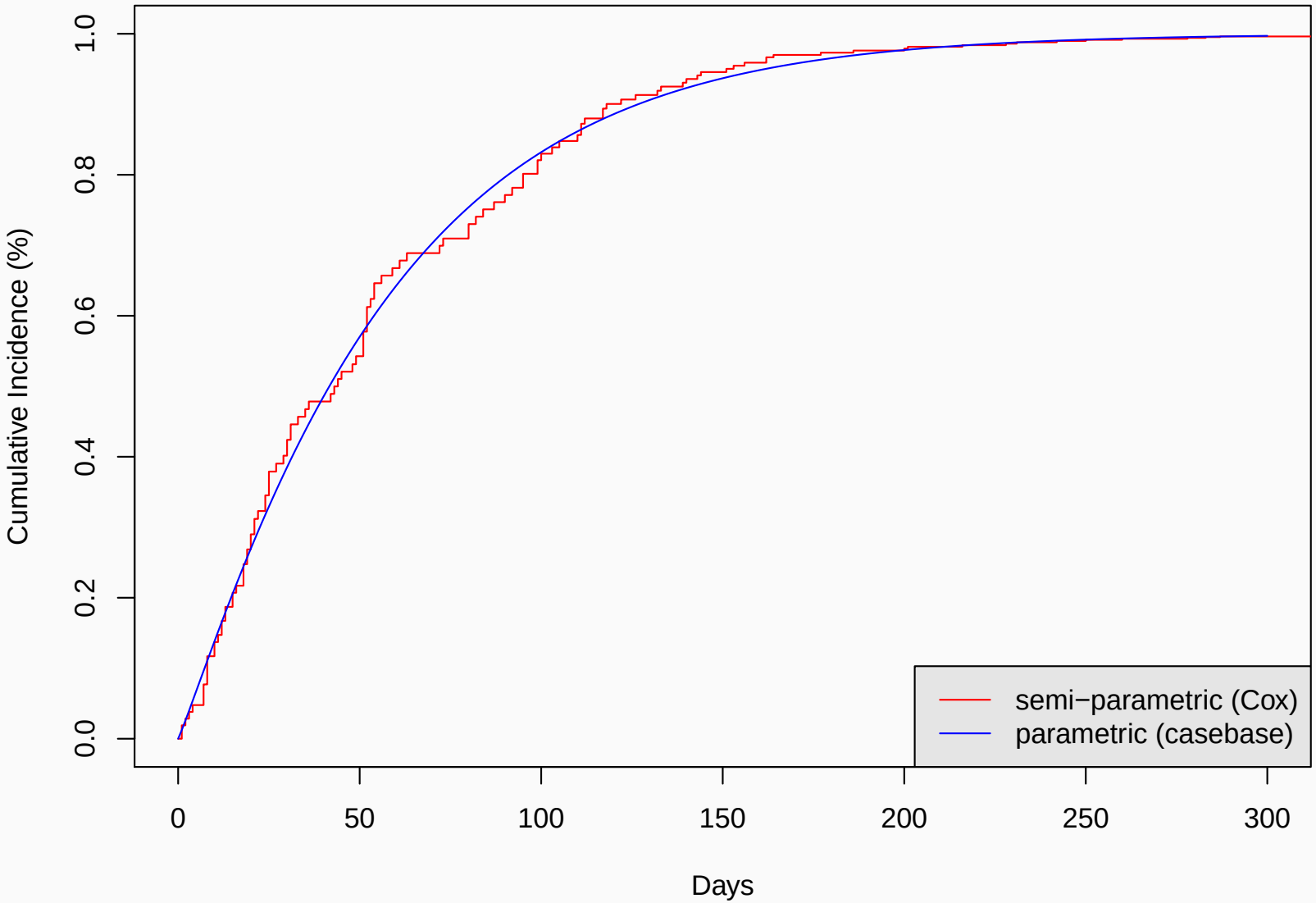
Veteran data–Estimates

Variables		Cox	Case-Base	Weibull
Karnofsky score		**0.97**	**0.97**	**0.97**
Time from diagnosis		1.00	1.00	1.00
Age		0.99	1.00	0.99
Prior therapy		1.07	1.06	1.05
	Squamous	0.67	0.66	0.65
Cell type	Small cell	1.58	1.56	1.59
	Adeno	**2.21**	**2.17**	**2.21**
Treatment		1.34	1.30	1.28

Veteran data–95% CI

Variables		Case-Base	Weibull
Karnofsky score		(0.96, 0.98)	(0.96, 0.98)
Time from diagnosis		(0.98, 1.02)	(0.98, 1.02)
Age		(0.98, 1.01)	(0.98, 1.01)
Prior therapy		(0.67, 1.66)	(0.67, 1.64)
	Squamous	(0.38, 1.15)	(0.38, 1.12)
Cell type	Small cell	(0.94, 2.64)	(0.95, 2.65)
	Adeno	(1.19, 3.94)	(1.23, 3.97)
Treatment		(0.87, 1.94)	(0.86, 1.90)

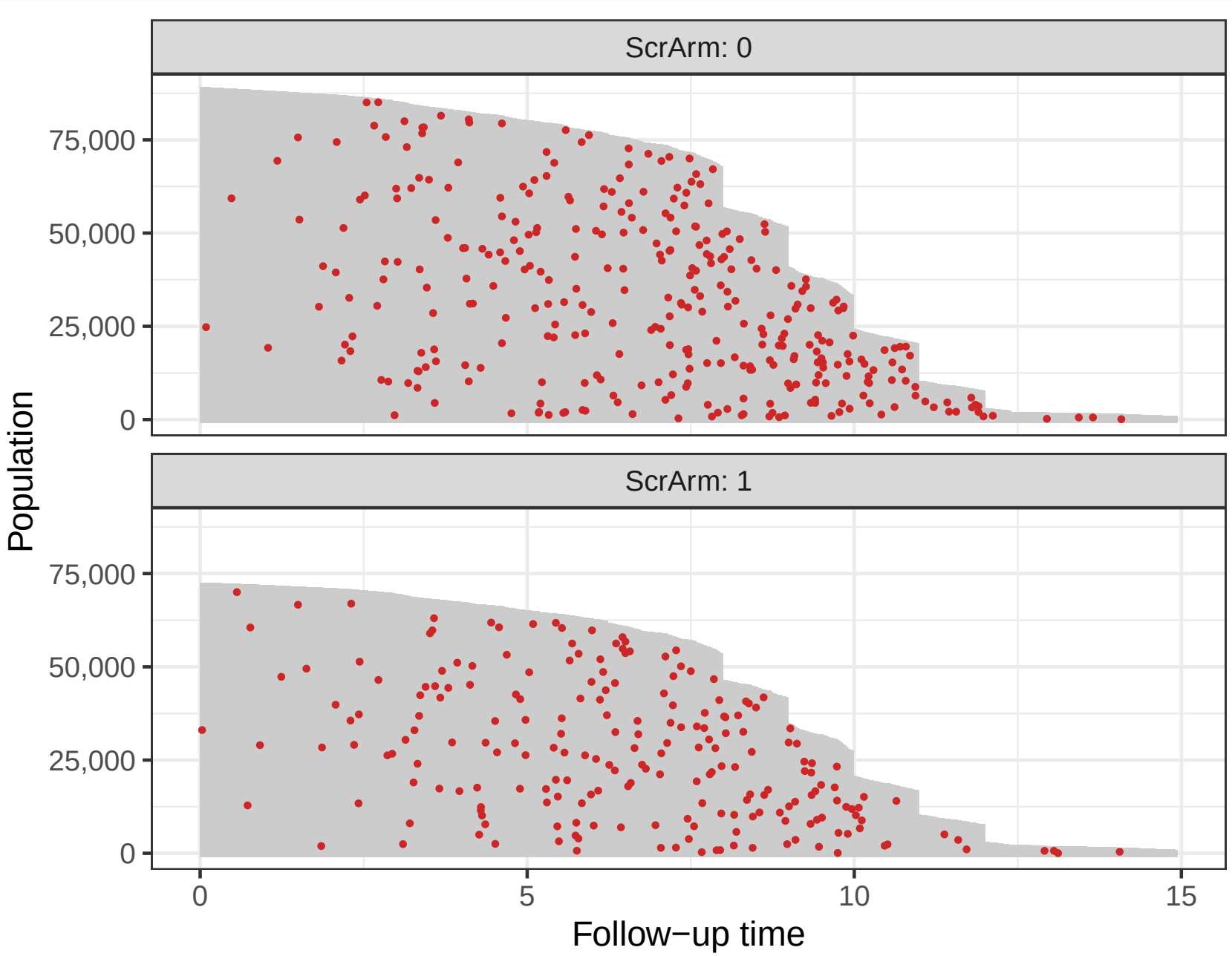
Veteran data–Risk plot



Case Study II–ERSPC data

- European Randomized Study of Prostate Cancer Screening (Schroeder *et al.*, 2009)
- 159,893 men between the ages of 55 and 69 years at entry.
- Recruited from seven European countries; recruitment started at different time.

ERSPC data–Population-Time plot



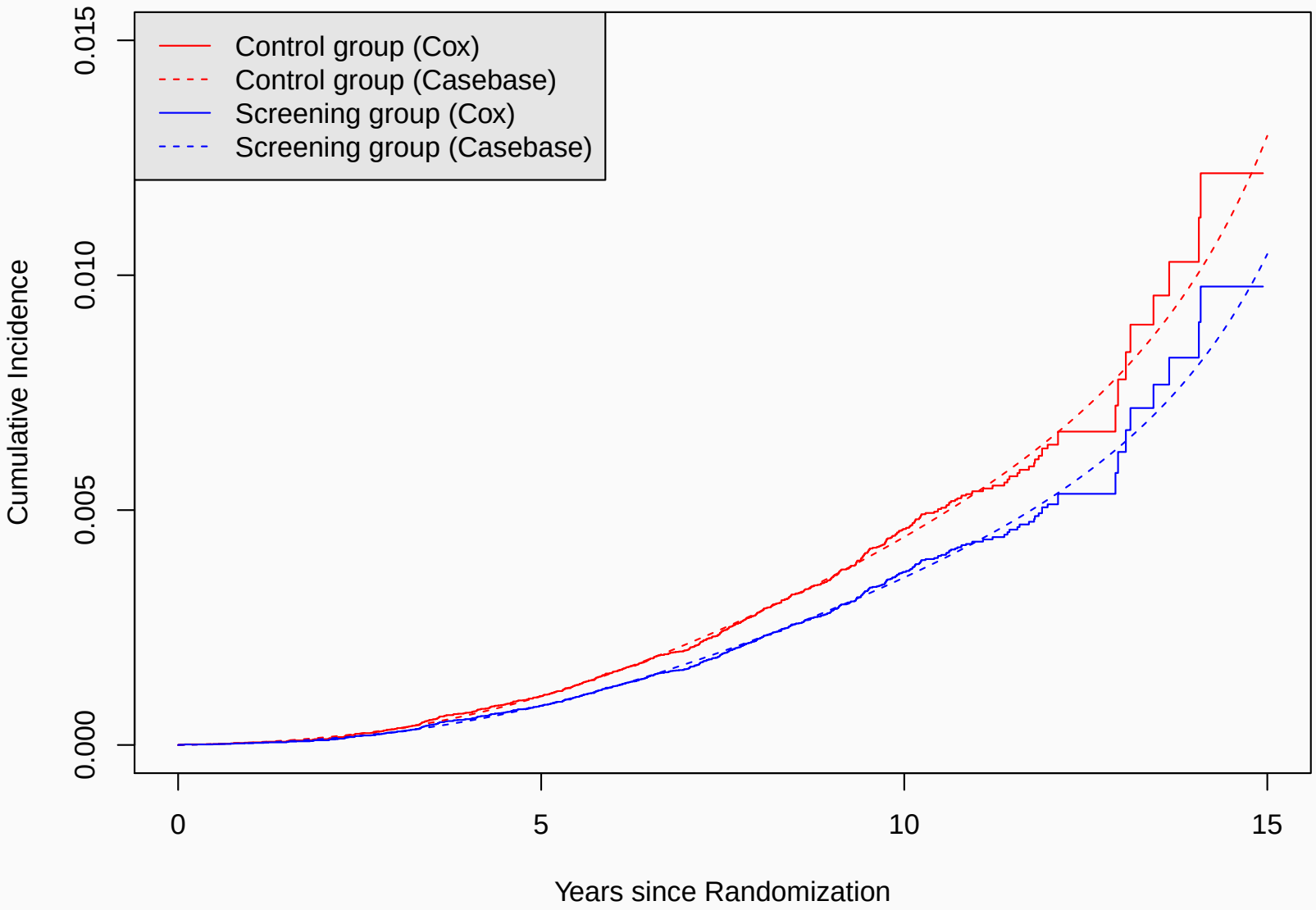
ERSPC data–Model fit

```
library(splines)

coxph(Surv(Follow.Up.Time, DeadOfPrCa) ~ ScrArm,
      data = ERSPC)

fitSmoothHazard(DeadOfPrCa ~ bs(Follow.Up.Time) + ScrArm,
                data = ERSPC)
```

ERSPC–Hazard ratio estimates

Model	HR	95% CI
Cox	0.80	(0.67 0.95)
Case-base	0.80	(0.68, 0.96)

ERSPC–Risk estimates



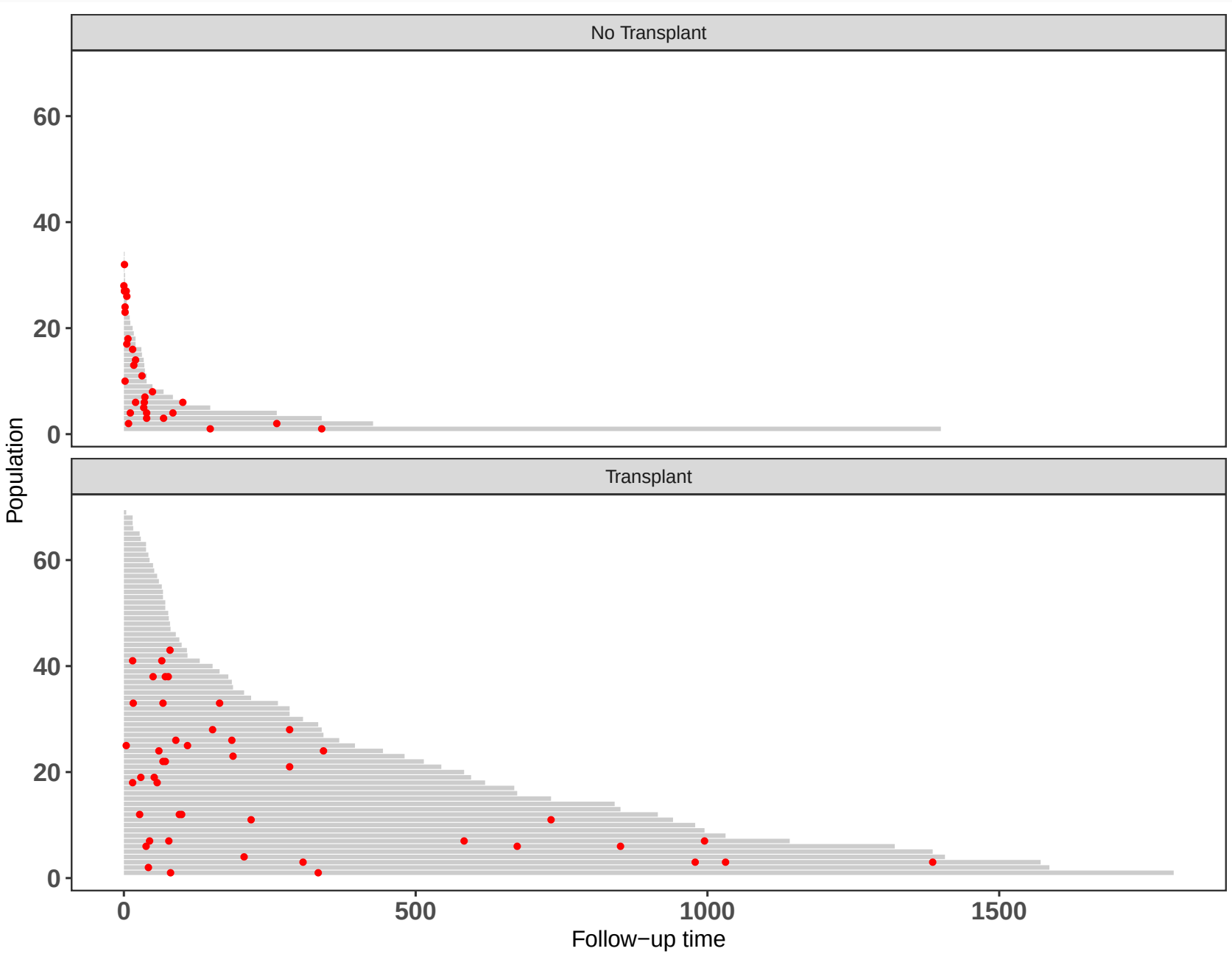
Non-proportional hazard

- Recall that we are explicitly modelling time.
- For this reason, we can fit non-proportional hazards using interaction terms
    - `Status ~ time * covariate`
- We will illustrate this approach using the Stanford Transplant data (available in the package `survival`).

Case Study III–Stanford transplant data

- Survival times of potential heart transplant recipients (Crowley & Hu, 1977).
- Evaluate the effect of transplant on subsequent survival
- For the purposes of this talk, we assume that exposure (i.e. transplant or no) is assessed at the **beginning of follow-up**.

Stanford data–Population-Time plot



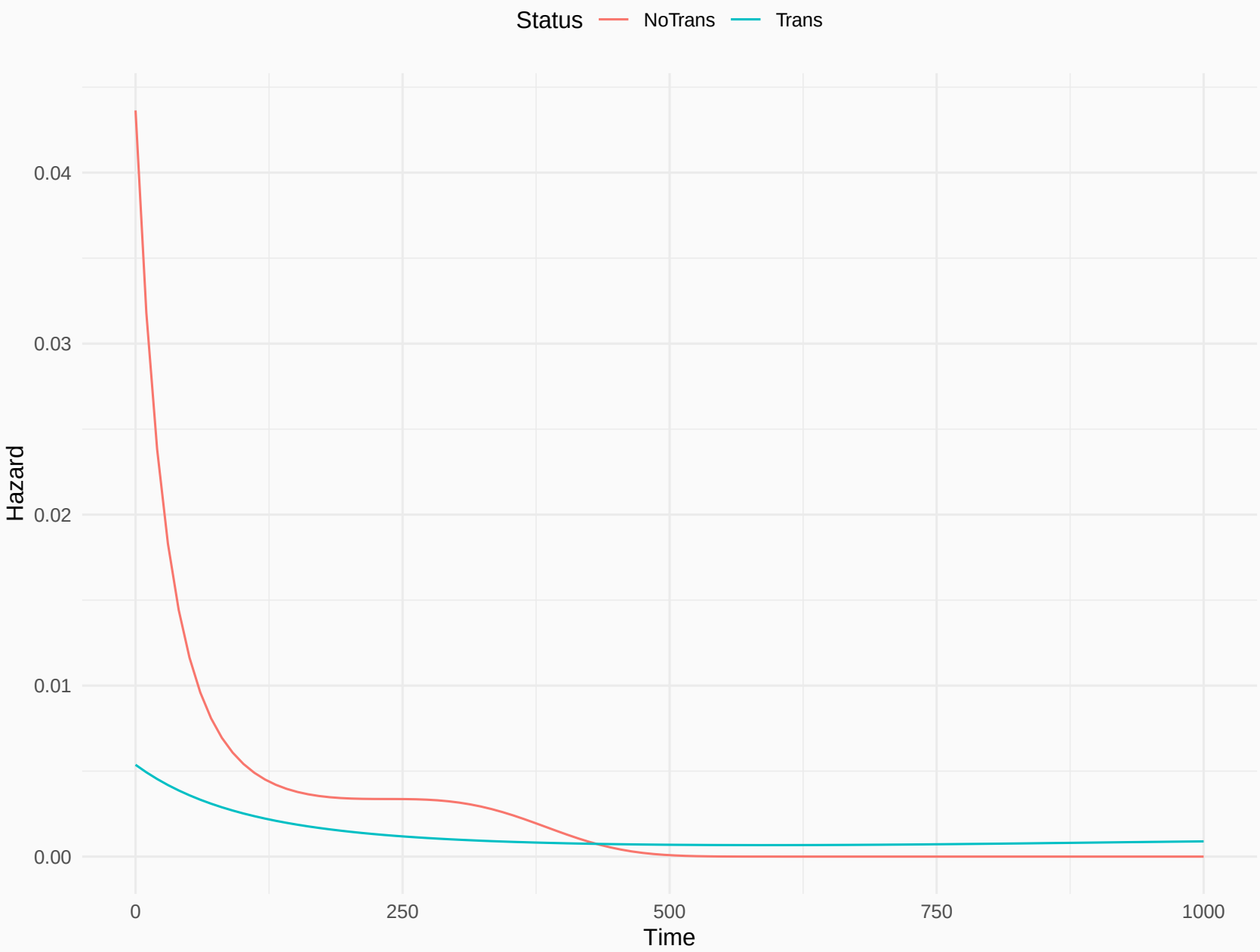
Stanford data–Model fit

```
fit1 <- fitSmoothHazard(fustat ~ transplant,
                        data = jasa, time = "futime")

fit2 <- fitSmoothHazard(fustat ~ transplant + futime,
                        data = jasa, time = "futime")

fit3 <- fitSmoothHazard(fustat ~ transplant + bs(futime),
                        data = jasa, time = "futime")

fit4 <- fitSmoothHazard(fustat ~ transplant*bs(futime),
                        data = jasa, time = "futime")
```

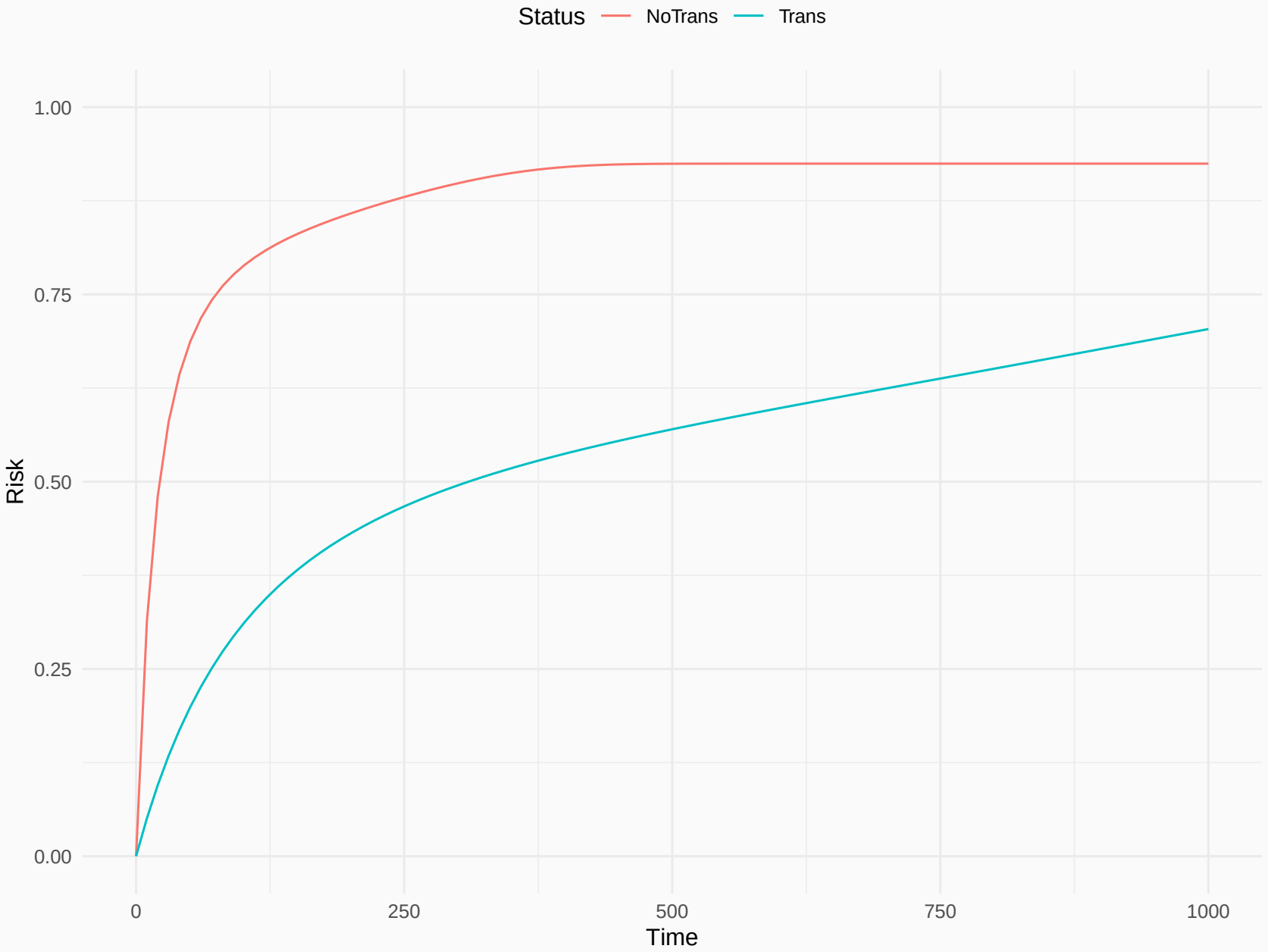
Stanford data–Model selection

Model	Predictors	PH	AIC
`fit1`	`transplant`	Yes	802.34
`fit2`	`transplant + time`	Yes	760.96
`fit3`	`transplant + bs(time)`	Yes	742.91
`fit4`	`transplant*bs(time)`	**No**	747.38

Stanford transplant data–Hazard and risk plots



Stanford transplant data–Hazard and risk plots



Case Study IV–Bone-marrow transplant study

- Data on patients who underwent haematopoietic stem cell transplantation for acute leukemia.
- Two types of stem-cell harvest:
  - Bone marrow and peripheral blood
  - Peripheral blood only
- Event of interest is relapse

Bone-marrow study–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)

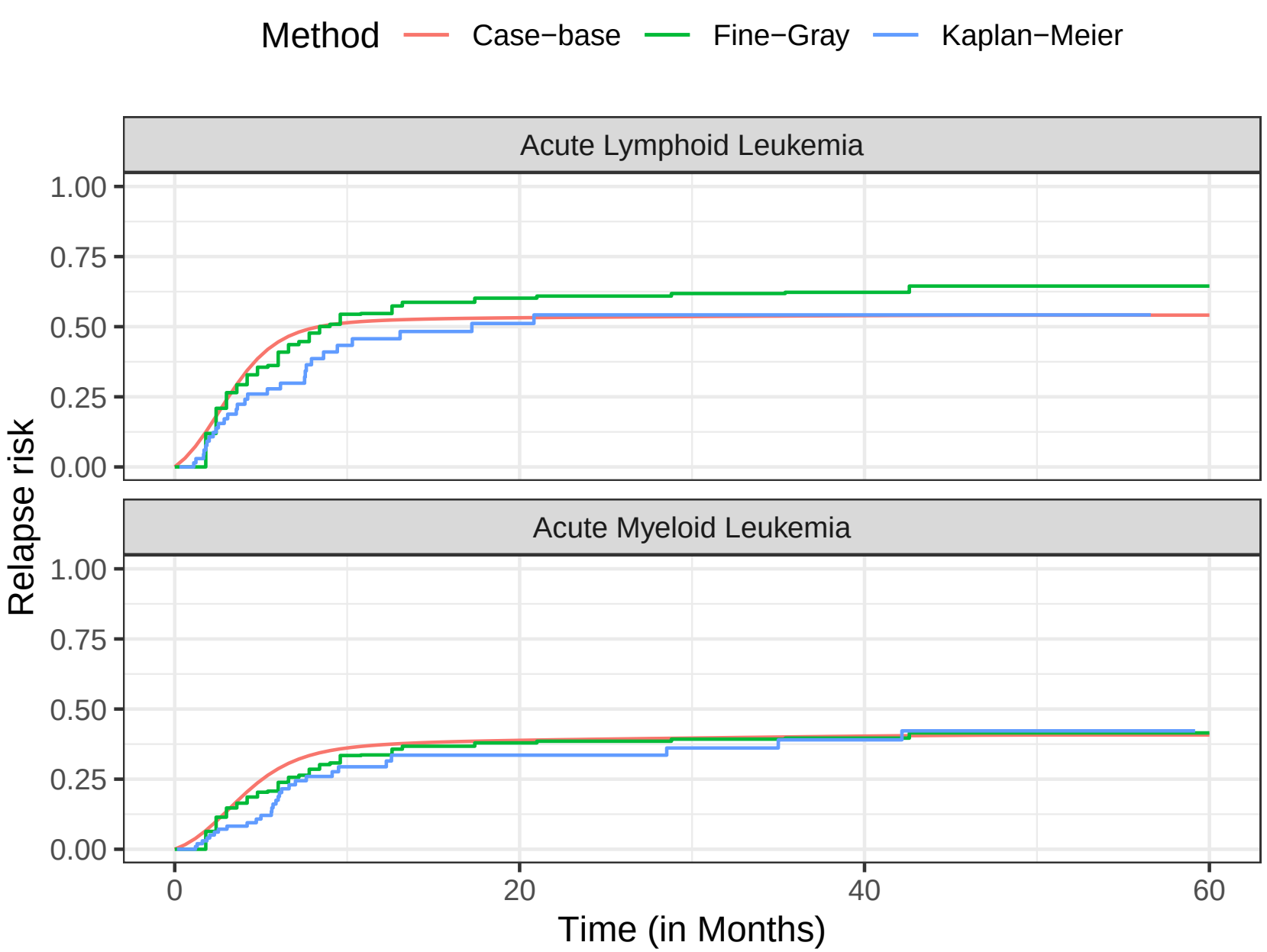
Bone-marrow study–Model fit

```
fitSmoothHazard(Status ~ bs(ftime, df = 5) + Sex + D +
                  Phase + Source + Age,
                  data = bmtcrr, time = "ftime")

comp.risk(Event(ftime, Status) ~ const(Sex) + const(D) +
          const(Phase) + const(Source) + const(Age),
          data = bmtcrr, cause = 1, model = "fg")

coxph(Surv(ftime, Status == 1) ~ Sex + D + Phase +
      Source + Age, data = bmtcrr)
```

Bone-marrow data–Hazard ratios and 95% CI

	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)

Bone-marrow data–Absolute risk plots



Discussion

Discussion

- We proposed a simple and flexible way of directly modelling the hazard function, using **logistic regression**.
    - This leads to smooth estimates of the absolute risks.
- We are explicitly modelling time.
- We can test the significance of covariates.
- The `R` package `casebase` provides convenient functions for the different parts of the analysis.

Ongoing projects

- In ongoing work, I extended the theory to the setting of **competing risks**.
  - Logistic regression is replaced by multinomial regression.
  - To get true absolute risks, we need to account for competing risks.
- Islam (PhD student of Bhatnagar) is working on combining case-base sampling and penalized regression.

Future directions

- **Methodology**: Combining dimension reduction and case-base sampling for survival analysis with high-dimensional covariates.
- **Software**: Add tools for diagnostic and model performance.
  - Martingale residuals
  - ROC, AUC, Brier score
  - Parametric bootstrap

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.

N. Mantel.

Synthetic retrospective studies and related topics.

Biometrics, pages 479–486, 1973.

O. Saarela.

A case-base sampling method for estimating recurrent event intensities.

Lifetime data analysis, pages 1–17, 2015.

References ii

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

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/