Variable Selection in Parametric Hazard Models

Sahir Rai Bhatnagar
McGill University

sahirbhatnagar.com

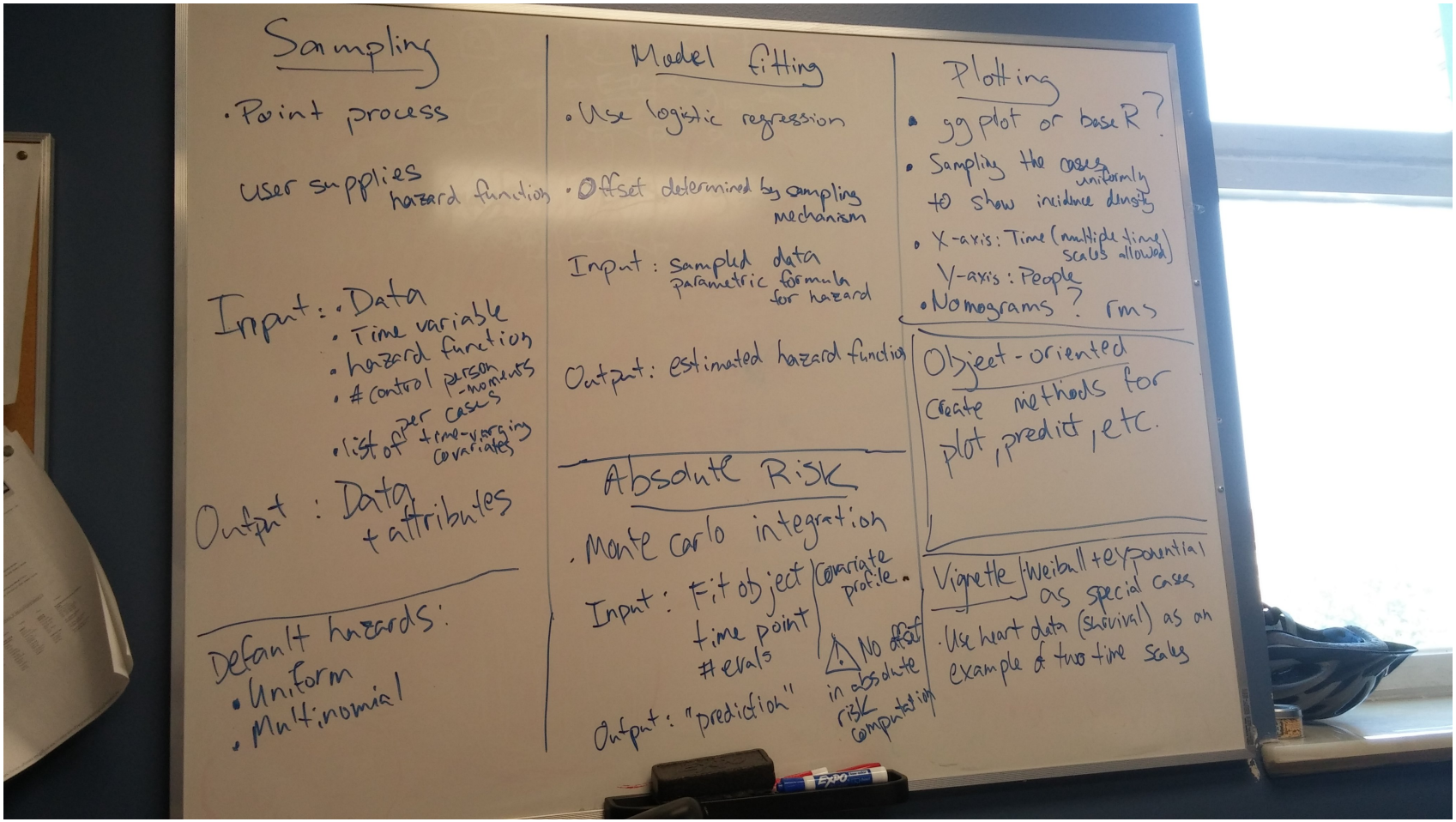
July 9, 2022

Jesse Islam
McGill

Jim Hanley
McGill

Olli Saarela
U. Toronto

Maxime Turgeon
U. Manitoba



August 19, 2015 in Dr. Celia Greenwood's lab at the Lady Davis Institute.

Outline

1. Overview of case-base sampling
2. Live coding demo
3. Extension to variable selection

Summary

Survival analysis

Survival analysis

Cox regression and absolute risk

- Time matching/risk set sampling (including Cox partial likelihood) eliminates the baseline hazard from the likelihood expression for the hazard ratios.

$$\lambda(t) = \lambda_0(t) \exp(\beta X)$$

Reid N (1994). *A Conversation with Sir David Cox.* Statistical Science.

Cox regression and absolute risk

- Time matching/risk set sampling (including Cox partial likelihood) eliminates the baseline hazard from the likelihood expression for the hazard ratios.

$$\lambda(t) = \lambda_0(t) \exp(\beta X)$$

Reid: So if you had a set of censored survival data today, you might rather fit a parametric model, even though there was a feeling among the medical statisticians that that wasn't quite right.

Cox: That's right, but since then various people have shown that the answers are very insensitive to the parametric formulation of the underlying distribution [see, e.g., Cox and Oakes, Analysis of Survival Data, Chapter 8.5]. **And if you want to do things like predict the outcome for a particular patient, it's much more convenient to do that parametrically.**

Reid N (1994). *A Conversation with Sir David Cox*. Statistical Science.

Linear and logistic first, survival last

Linear/logistic model	**Survival model**
Lasso (1996)	Coxnet (2011)
SCAD (2001)	Cox+SCAD (2011)
Elastic net (2005)	
Group lasso (2006)	
Hierarchical penalties (2006)	Penalized Cox for interactions (2010)
Neural Netwoks (2010)	DeepHit, DeepSurv (2018)

`casebase`: An alternative framework for survival analysis

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.

`casebase`: An alternative framework for survival analysis

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.

- This enables easy fitting of smooth-in-time and non-proportional hazard models with multiple time scales.

`casebase`: An alternative framework for survival analysis

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.
- This enables easy fitting of smooth-in-time and non-proportional hazard models with multiple time scales.
- Extensions to penalized models and neural networks.

`casebase`: An alternative framework for survival analysis

- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods.
- This enables easy fitting of smooth-in-time and non-proportional hazard models with multiple time scales.
- Extensions to penalized models and neural networks.
- Provides an alternative to Kaplan-Meier-based methods for estimating discrimination/calibration statistics (e.g. ROC, AUC, risk reclassification probabilities, Brier score) from censored survival data.

casebase R package

casebase: Fitting Flexible Smooth-in-Time Hazards and Risk Functions vi

Fit flexible and fully parametric hazard regression models to survival data with single event type or multip
its interactions with other predictors for time-dependent hazards and hazard ratios. From the fitted hazard r
This approach accommodates any log-linear hazard function of prognostic time, treatment, and covariates,
plots. Based on the case-base sampling approach of Hanley and Miettinen (2009) <doi:10.2202/1557-4679

Version:	0.10.1
Depends:	R (≥ 3.5.0)
Imports:	data.table, ggplot2, methods, mgcv, stats, survival, VGAM
Suggests:	colorspace, eha, glmnet, knitr, progress, rmarkdown, splines, testthat (≥ 3.0.0), visreg
Published:	2021-10-20
Author:	Sahir Bhatnagar [aut, cre] (http://sahirbhatnagar.com/), Maxime Turgeon [aut], Jes (http://www.medicine.mcgill.ca/epidemiology/hanley/)
Maintainer:	Sahir Bhatnagar <sahir.bhatnagar at gmail.com>
BugReports:	https://github.com/sahirbhatnagar/casebase/issues
License:	MIT + file LICENSE
URL:	http://sahirbhatnagar.com/casebase/
NeedsCompilation:	no
Citation:	casebase citation info
Materials:	README NEWS
In views:	Survival
CRAN checks:	casebase results

Documentation:

Reference manual: casebase.pdf

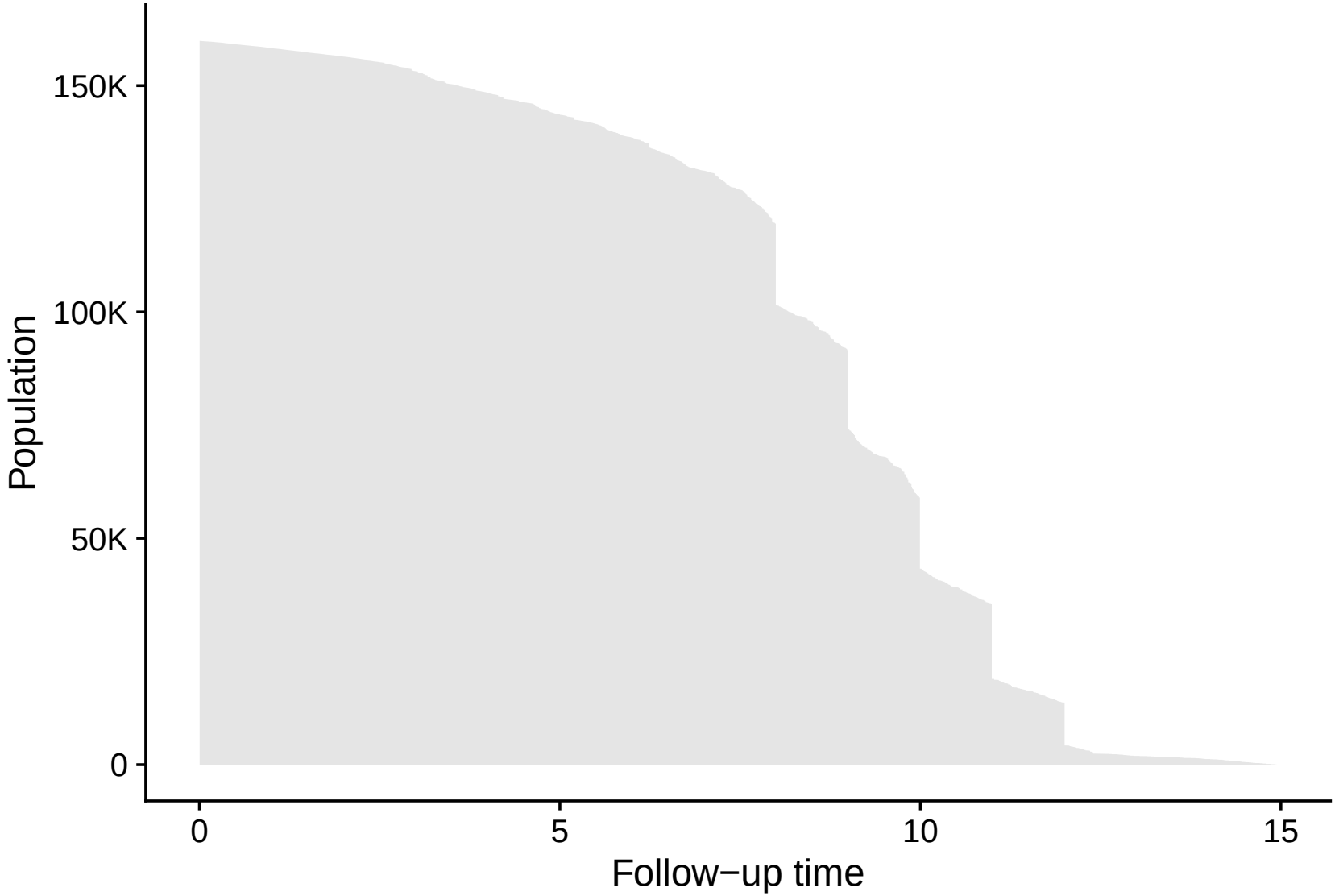
Vignettes:
- Competing risk analysis
- Customizing Population Time Plots
- Plot Cumulative Incidence and Survival Curves
- Plot Hazards and Hazard Ratios
- Population Time Plots
- Introduction to casebase sampling

---

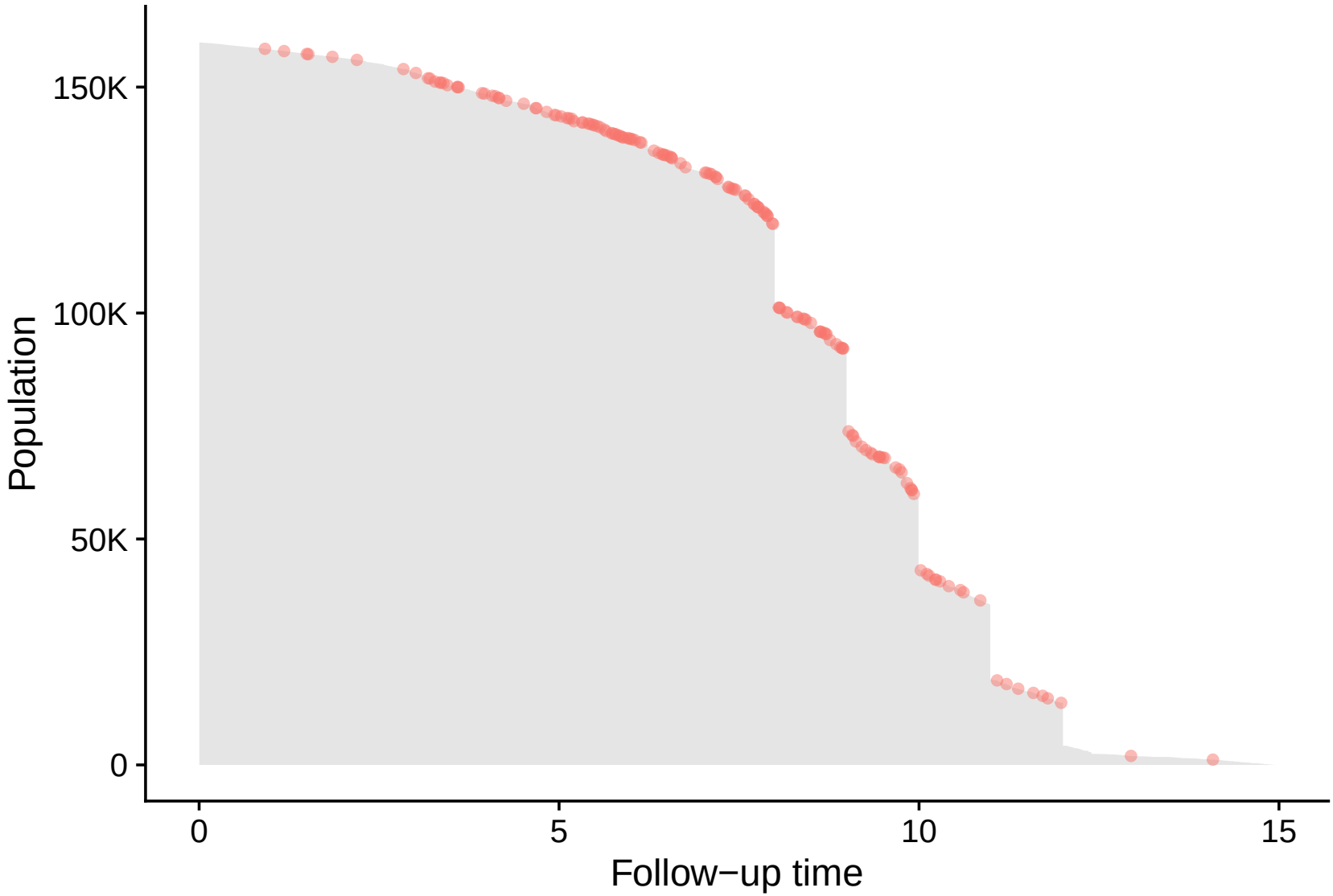
https://arxiv.org/abs/2009.10264 *accepted in* R Journal (2022+), https://cran.r-project.org/package=casebase. 55k downloads (as of July 2022).

Case-base sampling

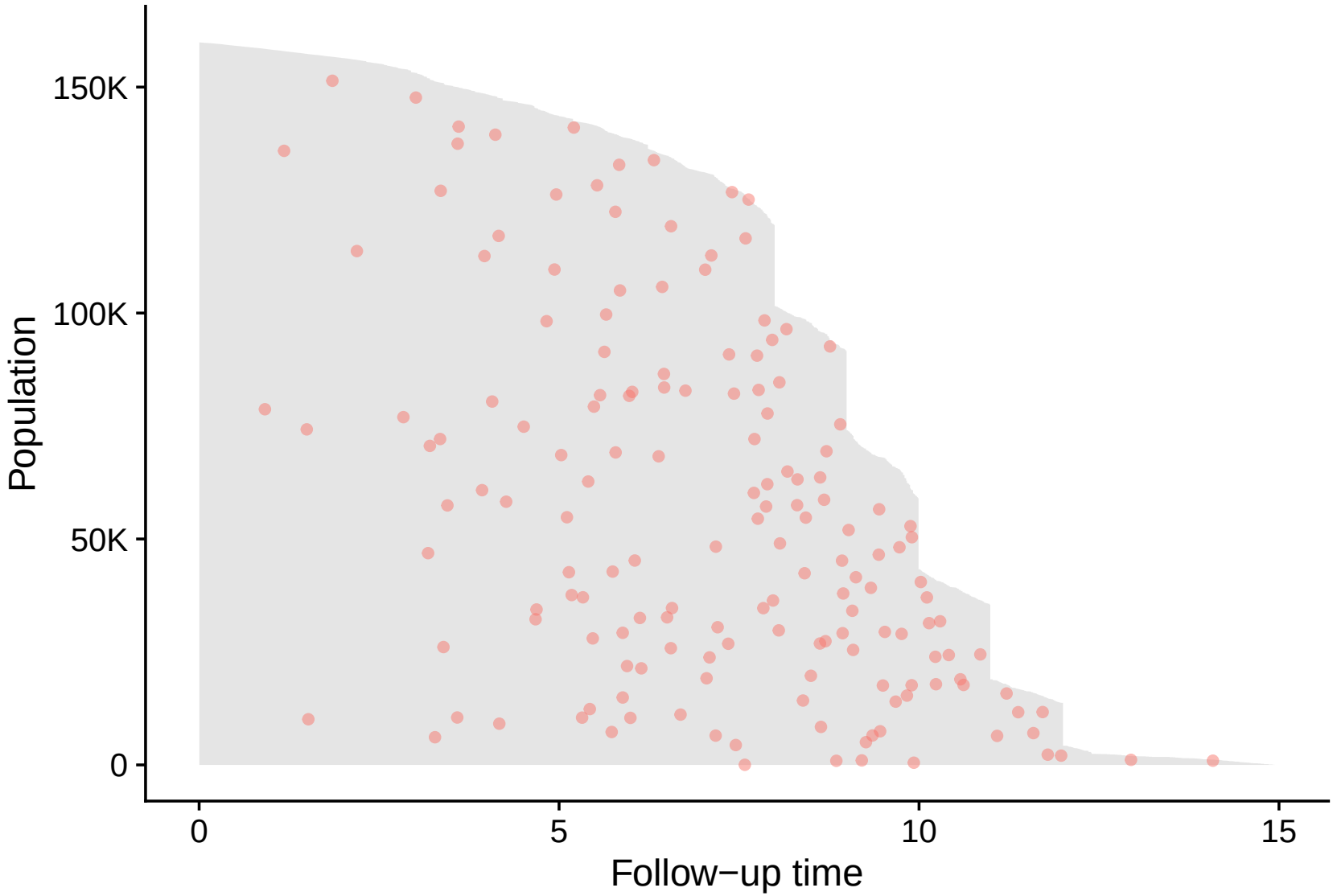
Study base



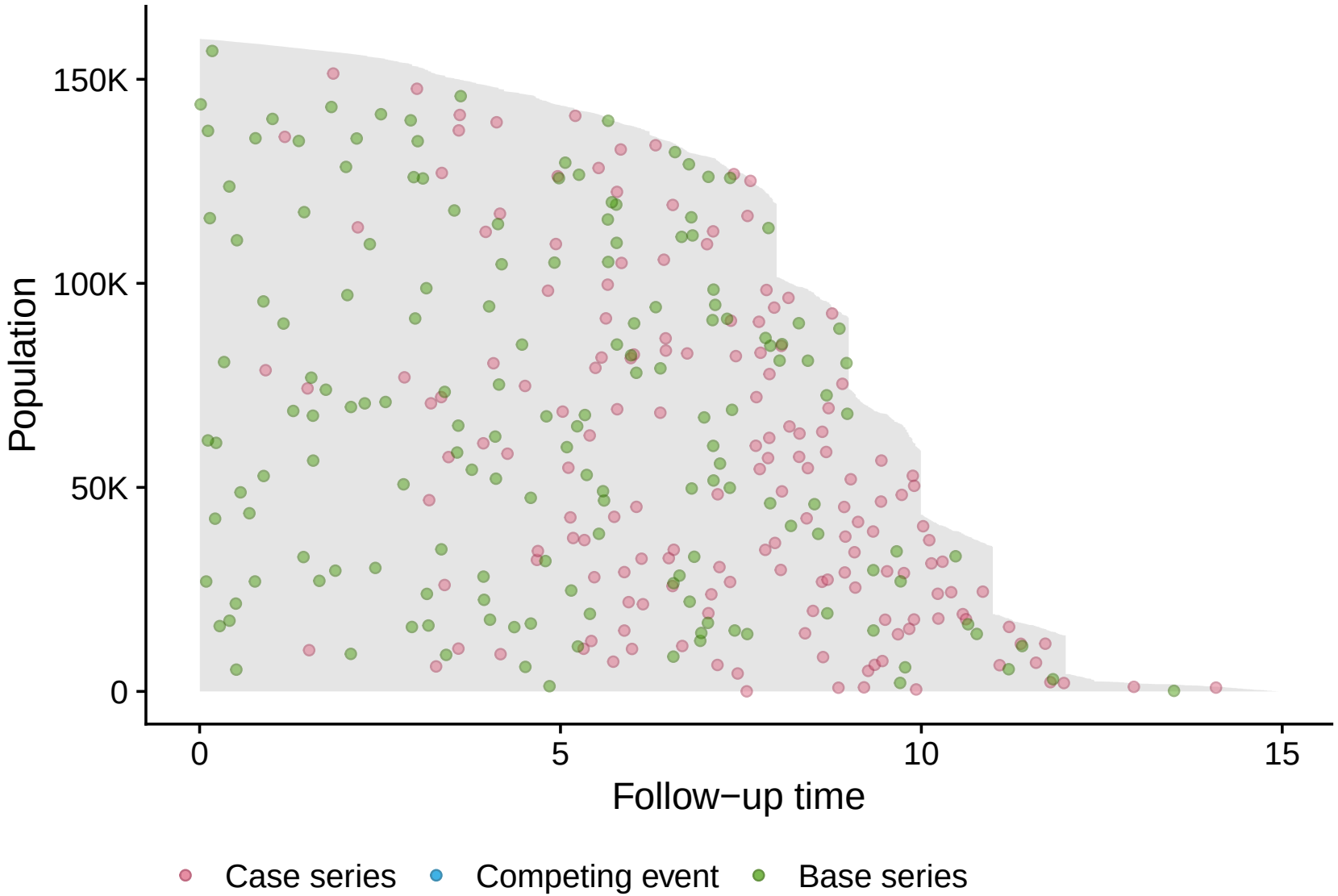
Case series



Case series



Case and base series



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 $\log(1/\rho_i(t))$**.**

O. Saarela (2015). *A case-base sampling method for estimating recurrent event intensities.* Lifetime data analysis.

Asymptotic properties

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.

Asymptotic properties

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.

Live coding demo

Variable Selection

R packages for survival analysis

Package	Competing Risks	Allows Non PH	Penalized Regression	Splines	Parametric	Semi Parametric	Interval/Left Censoring	Risk Estimates
casebase	✓	✓	✓	✓	✓			✓
CFC	✓	✓			✓			✓
cmprsk	✓					✓		✓
crrp	✓		✓			✓		
fastcox			✓			✓		
flexrsurv		✓		✓	✓			✓
flexsurv	✓	✓		✓	✓			✓
glmnet			✓			✓		✓
glmpath			✓			✓		
mets	✓			✓		✓		✓
penalized			✓			✓		
riskRegression	✓		✓			✓		✓
rstpm2		✓		✓	✓	✓	✓	✓
SmoothHazard		✓		✓	✓		✓	
survival	✓	✓			✓	✓	✓	✓

Penalized logistic regression

- To perform variable selection on the regression parameters $\theta \in \mathbb{R}^p$ of the hazard function, we can add a penalty to the likelihood and optimise the following equation:

$$\min_{\theta \in \mathbb{R}^p} \; -\log L(\theta) + \sum_{j=1}^{p} w_j \; p_{\lambda,\alpha}(\theta_j)$$

Penalized logistic regression

- To perform variable selection on the regression parameters $\theta \in \mathbb{R}^p$ of the hazard function, we can add a penalty to the likelihood and optimise the following equation:

$$\min_{\theta \in \mathbb{R}^p} \; -\log L(\theta) + \sum_{j=1}^{p} w_j \; p_{\lambda,\alpha}(\theta_j)$$

- $p_{\lambda,\alpha}(\theta_j)$ is a penalty term controlled by λ and α

Penalized logistic regression

- To perform variable selection on the regression parameters $\theta \in \mathbb{R}^p$ of the hazard function, we can add a penalty to the likelihood and optimise the following equation:

$$\min_{\theta \in \mathbb{R}^p} \; -\log L(\theta) + \sum_{j=1}^{p} w_j \; p_{\lambda,\alpha}(\theta_j)$$

- $p_{\lambda,\alpha}(\theta_j)$ is a penalty term controlled by λ and α

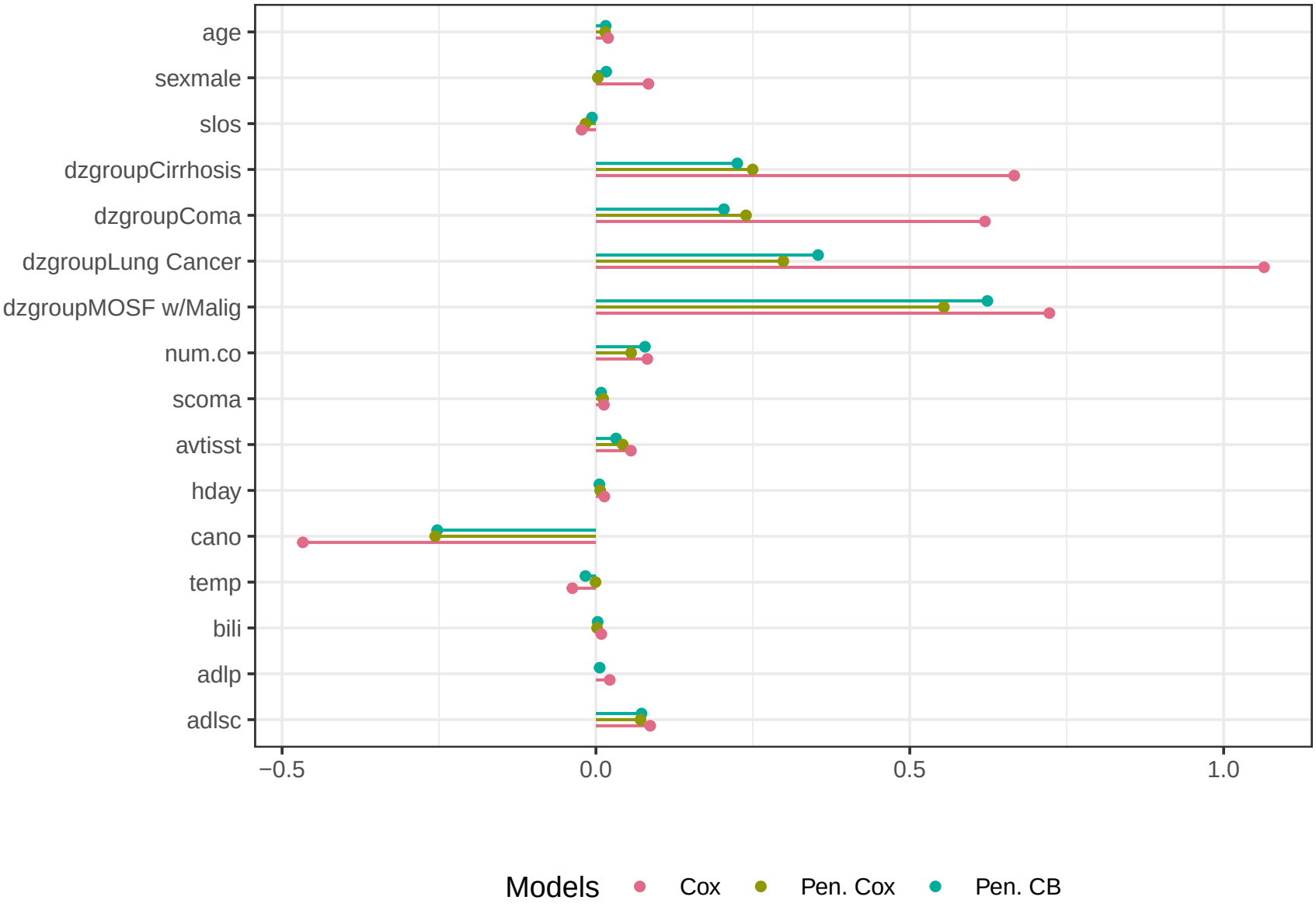
- w_j is the penalty factor for the jth covariate

Variable selection with casebase

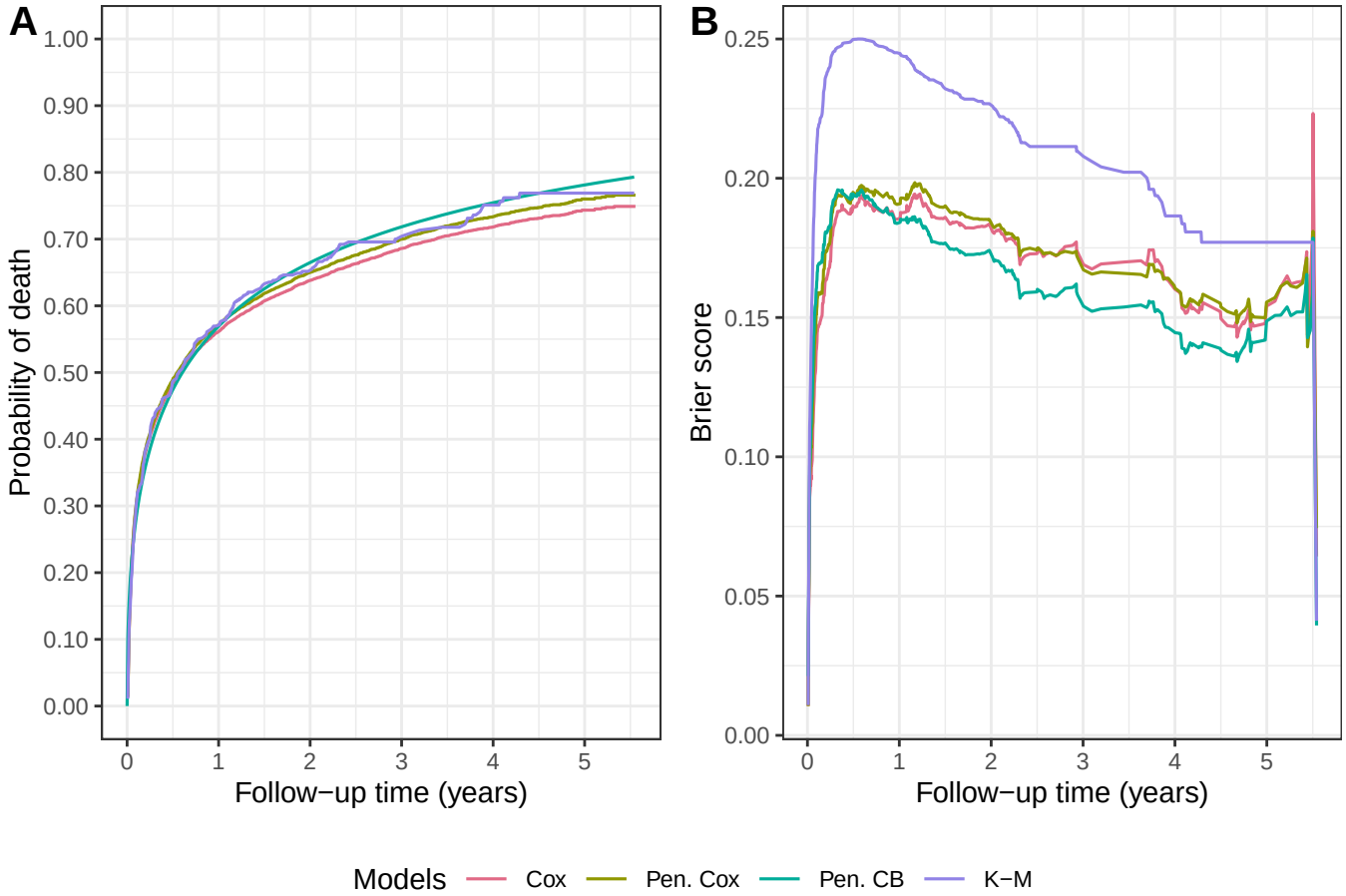
```
# casebase
pen_cb <- fitSmoothHazard.fit(x, y, family = "glmnet", time = "d.time",
                              event = "death", ratio = 10,
                              formula_time = ~ log(d.time),
                              alpha = 1, standardize = TRUE,


# coxnet
u <- with(train, survival::Surv(time = d.time, event = death))
coxNet <- glmnet::cv.glmnet(x = x, y = u, family = "cox",
                            alpha = 1, standardize = TRUE)
```

Variable selection with casebase



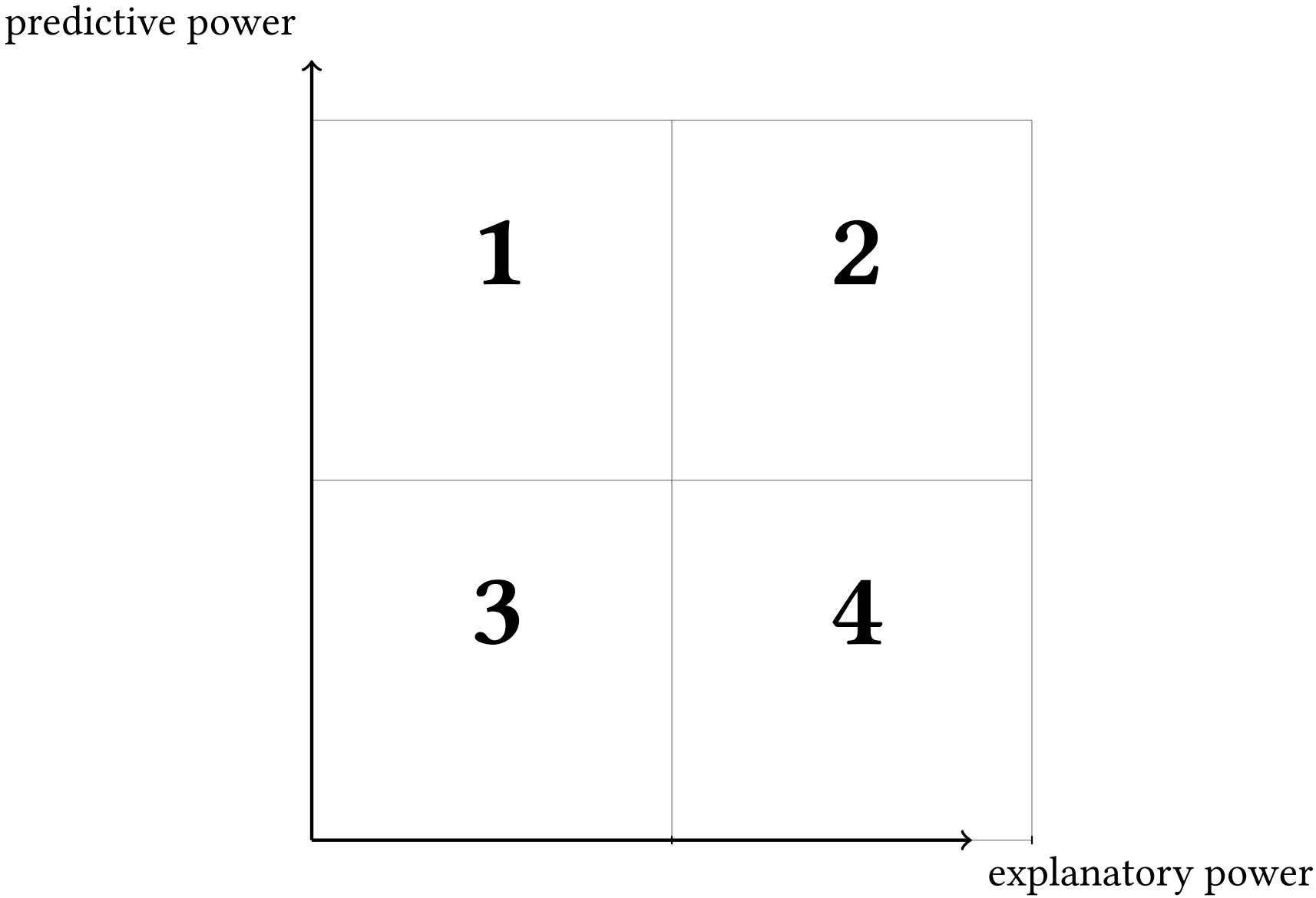
Brier score



$$\text{Brier Score}(t) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left(\widehat{F}(X_i,t)-1\right)^2 \cdot I_{T_i\leq t,\delta_i=1}}{\widehat{G}(T_i)} + \frac{\left(\widehat{F}(X_i,t)\right)^2 \cdot I_{T_i>t}}{\widehat{G}(t)}\right)$$

Future Directions

To explain or predict?

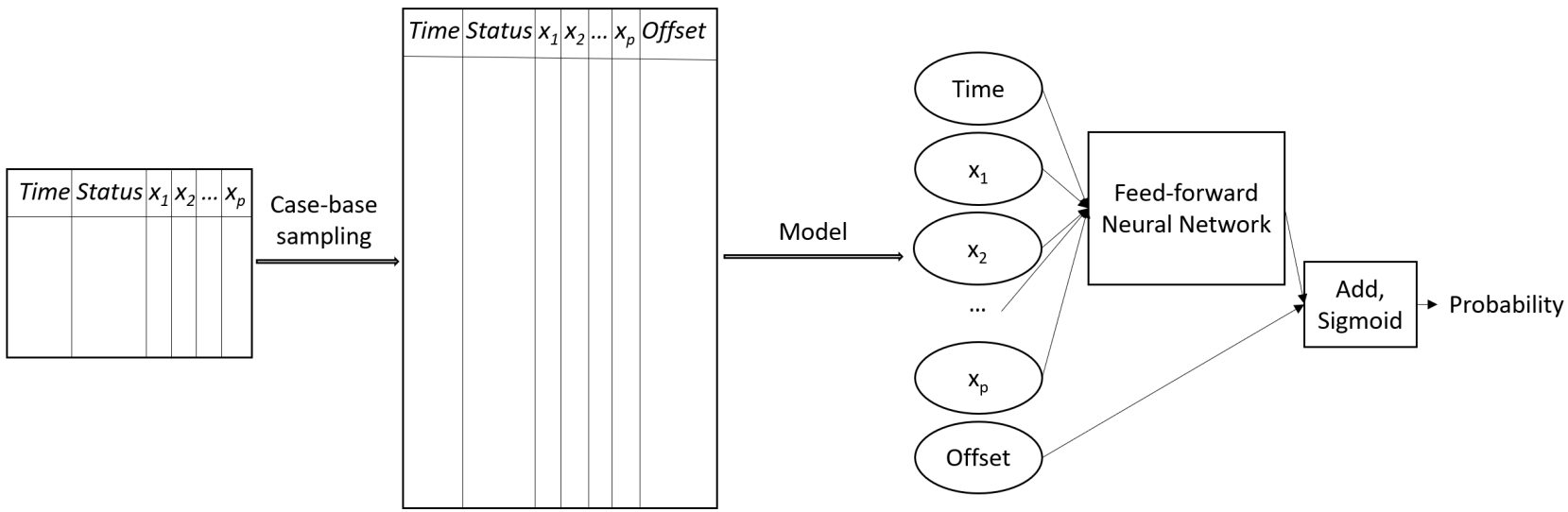


Extension: Higher-order interactions and flexible baseline

RESEARCH ARTICLE

Case-Base Neural Networks: survival analysis with time-varying, higher-order interactions

Jesse Islam[1] | Maxime Turgeon[2] | Robert Sladek[3] | Sahir Bhatnagar[4]



Jesse Islam, PhD (c), Quantitative Life Sciences

https://github.com/Jesse-Islam/pmnn

http://sahirbhatnagar.com/casebase/

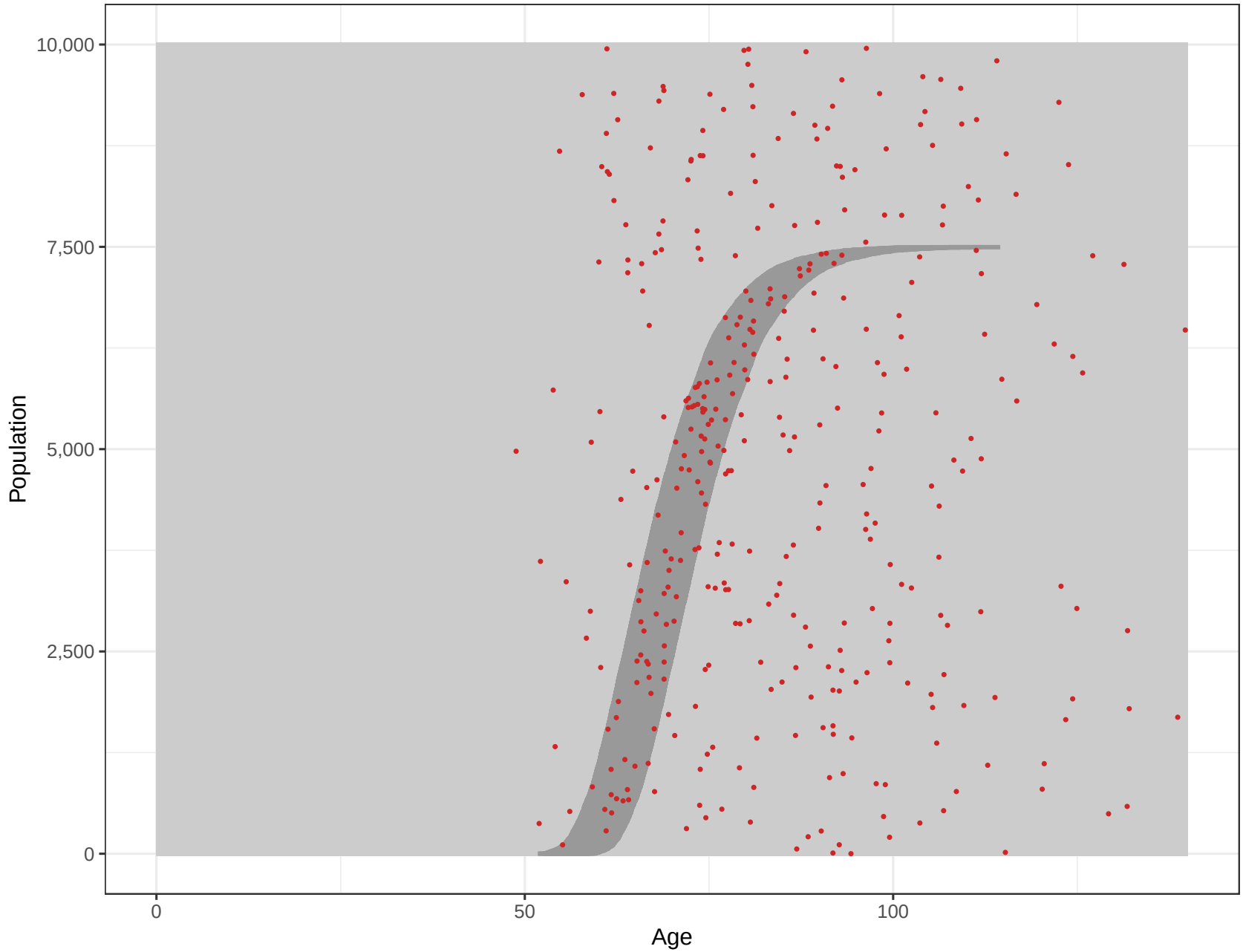
Remarks

- 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.
- Case-base sampling combined with logistic/multinomial regression provides an alternative to risk set sampling-based semi-parametric survival analysis methods
- Similarly, this provides an alternative to Kaplan-Meier-based methods for estimating discrimination statistics (e.g. ROC, AUC, risk reclassification probabilities) from censored survival data.
- The `R` package `casebase` provides convenient functions for the different parts of the analysis.

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.

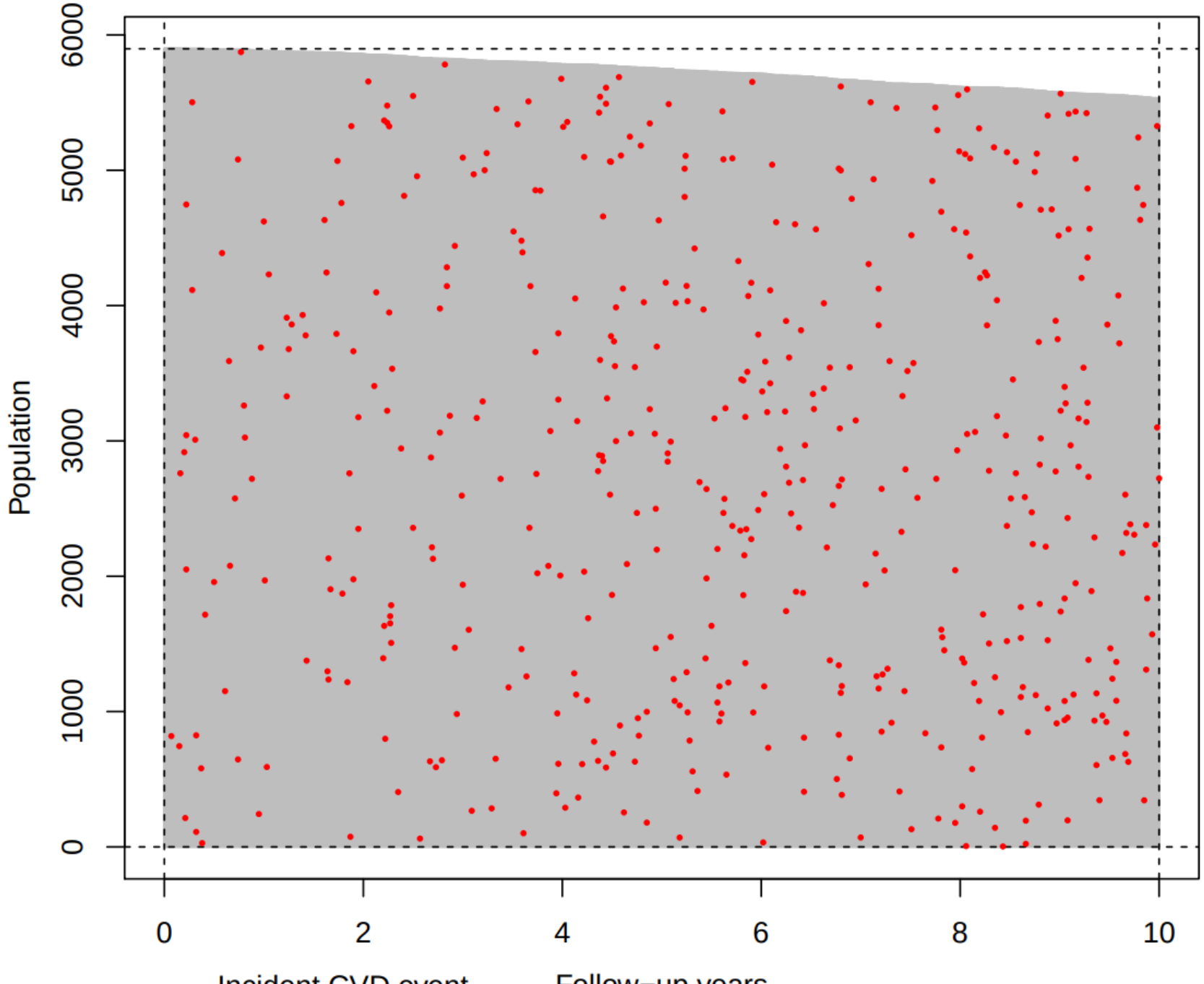
Vaccination safety (Saarela & Hanley, 2015)



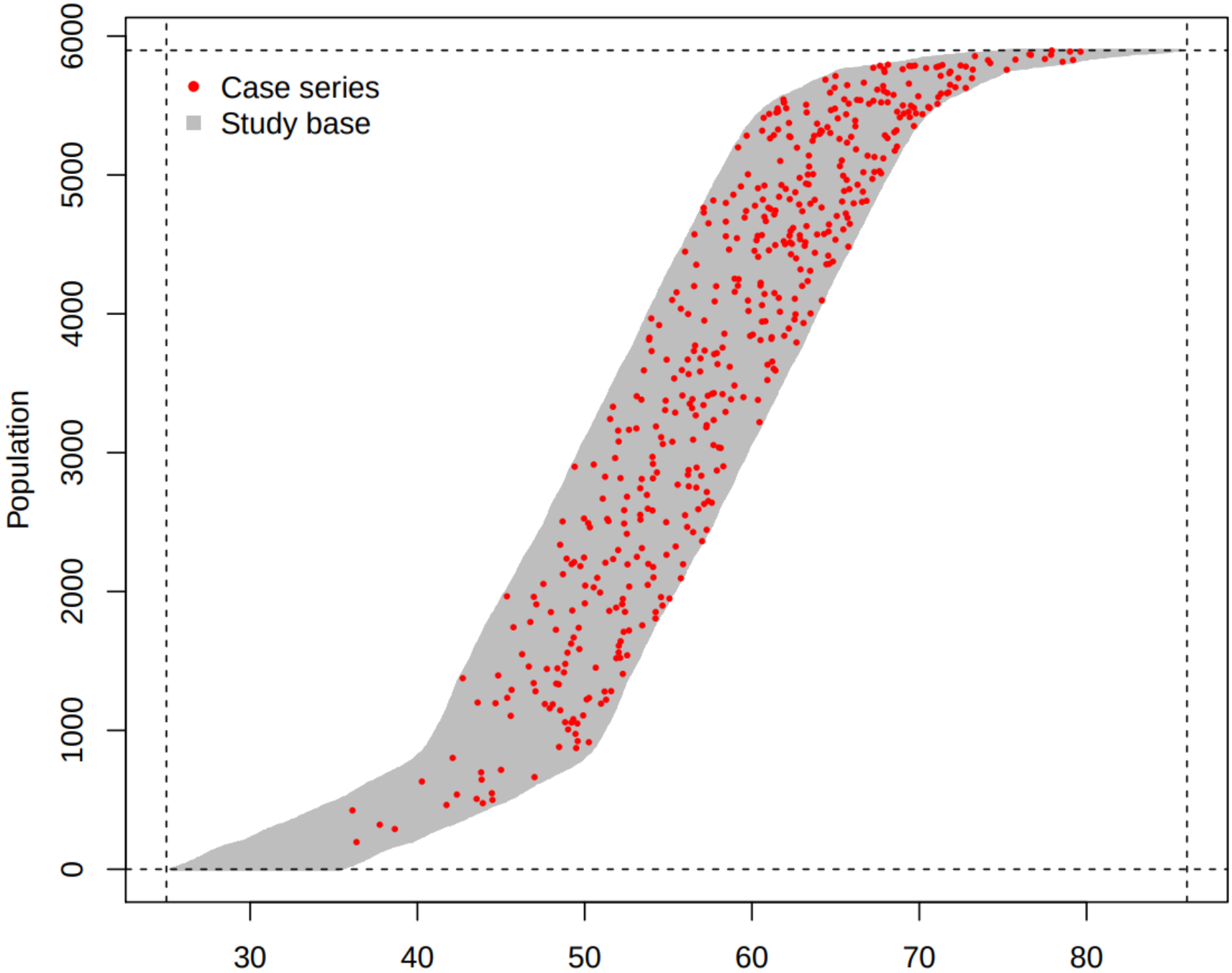
Different time scales

- 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.

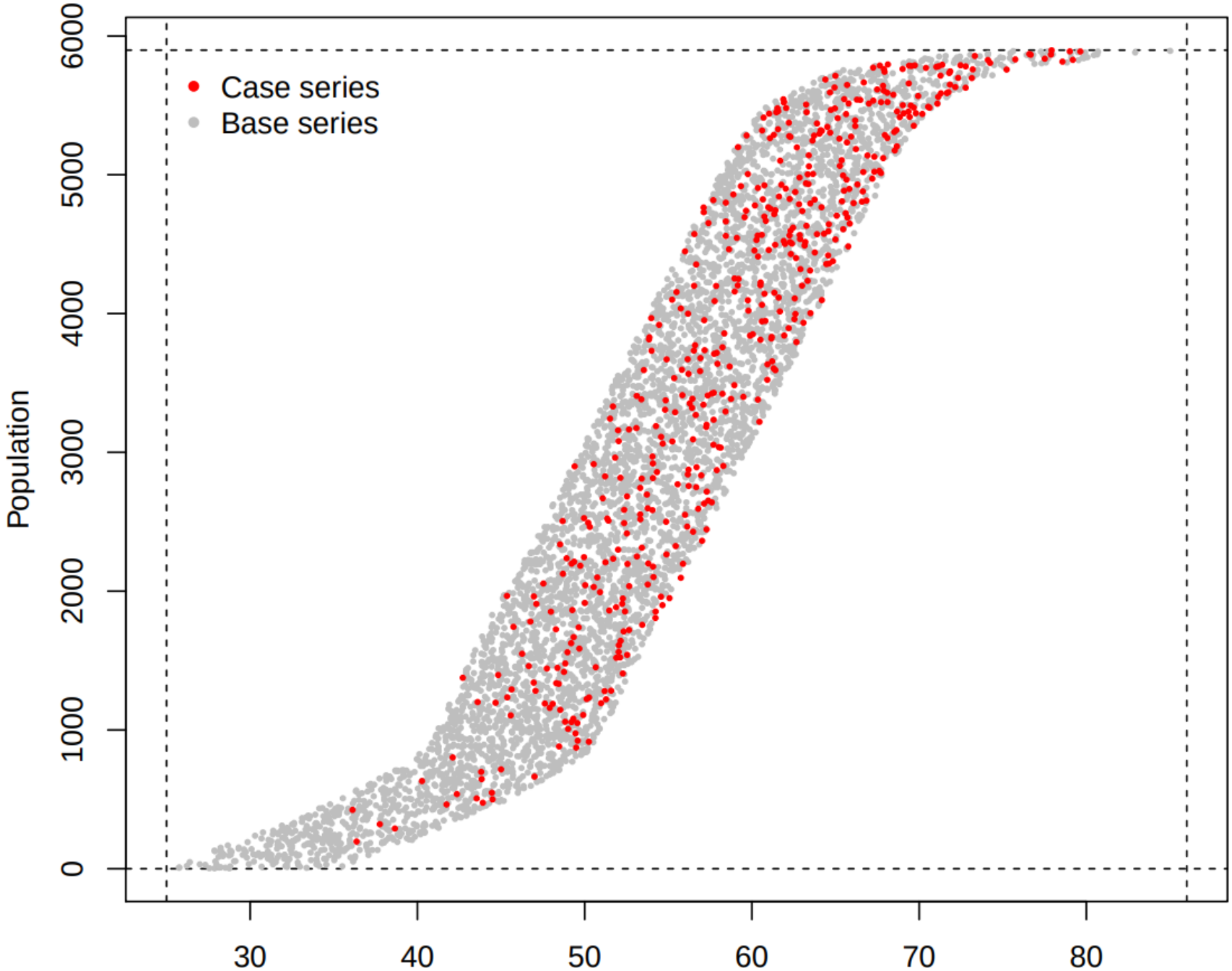
Different time scales



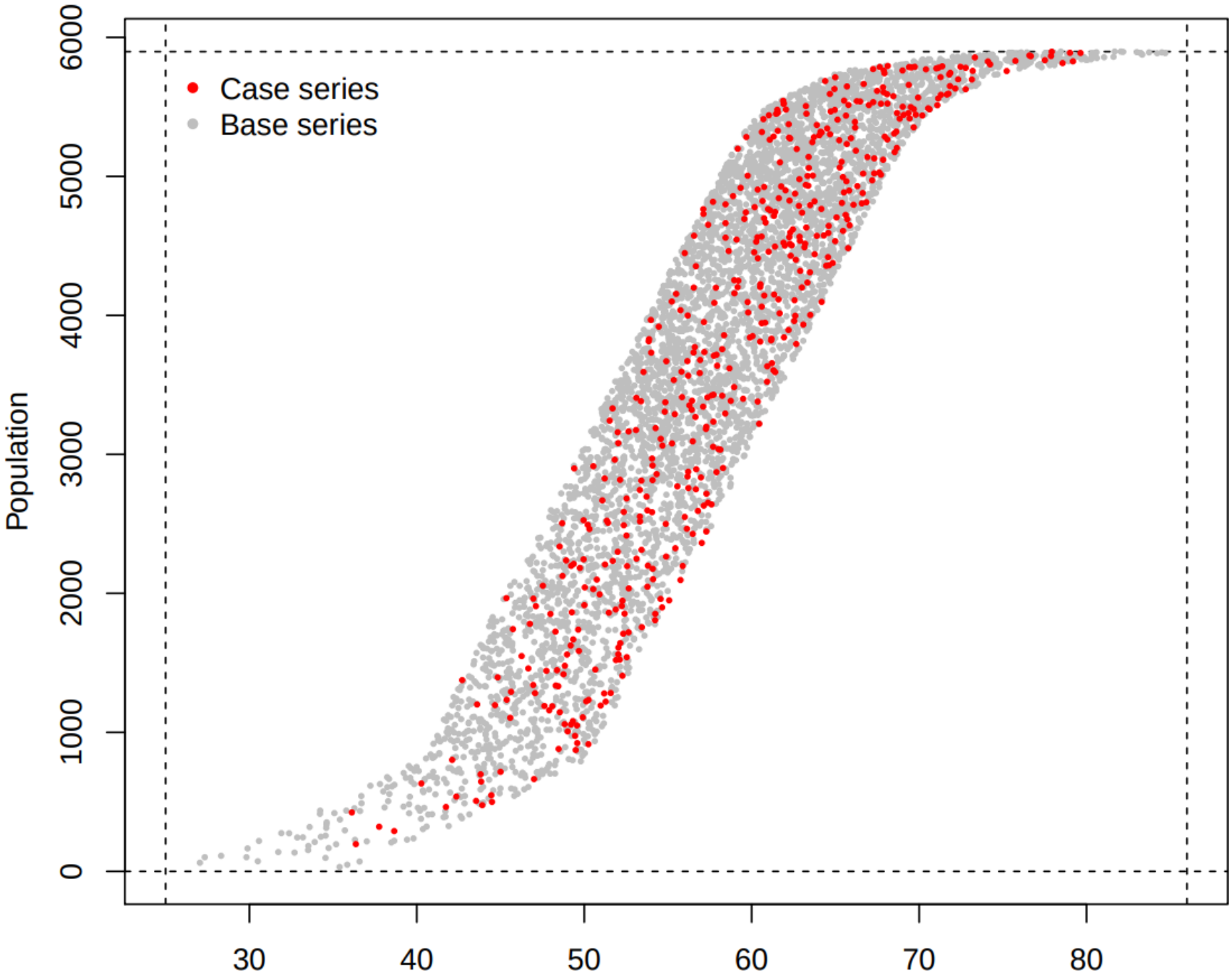
Different time scales



Different time scales



Different time scales



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

```
^^I^^I^^IpopTime(data, time, event, censored.indicator, exposure
^^I^^I
```

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

sampleCaseBase

```
^^I^^I^^IsampleCaseBase(data, time, event, ratio = 10,
^^I^^I^^Icomprisk = FALSE, censored.indicator)
^^I^^I
```

- `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

```
^^I^^I^^IfitSmoothHazard(formula, data, time,
^^I^^I^^Ifamily = c("glm", "gam", "gbm", "glmnet"),
^^I^^I^^Icensored.indicator, ratio = 100, ...)
^^I^^I^^I
^^I^^I^^IfitSmoothHazard.fit(x, y, formula_time, time, event,
^^I^^I^^Ifamily = c("glm", "gbm", "glmnet"),
^^I^^I^^Icensored.indicator, ratio = 100, ...)
^^I^^I
```

- 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

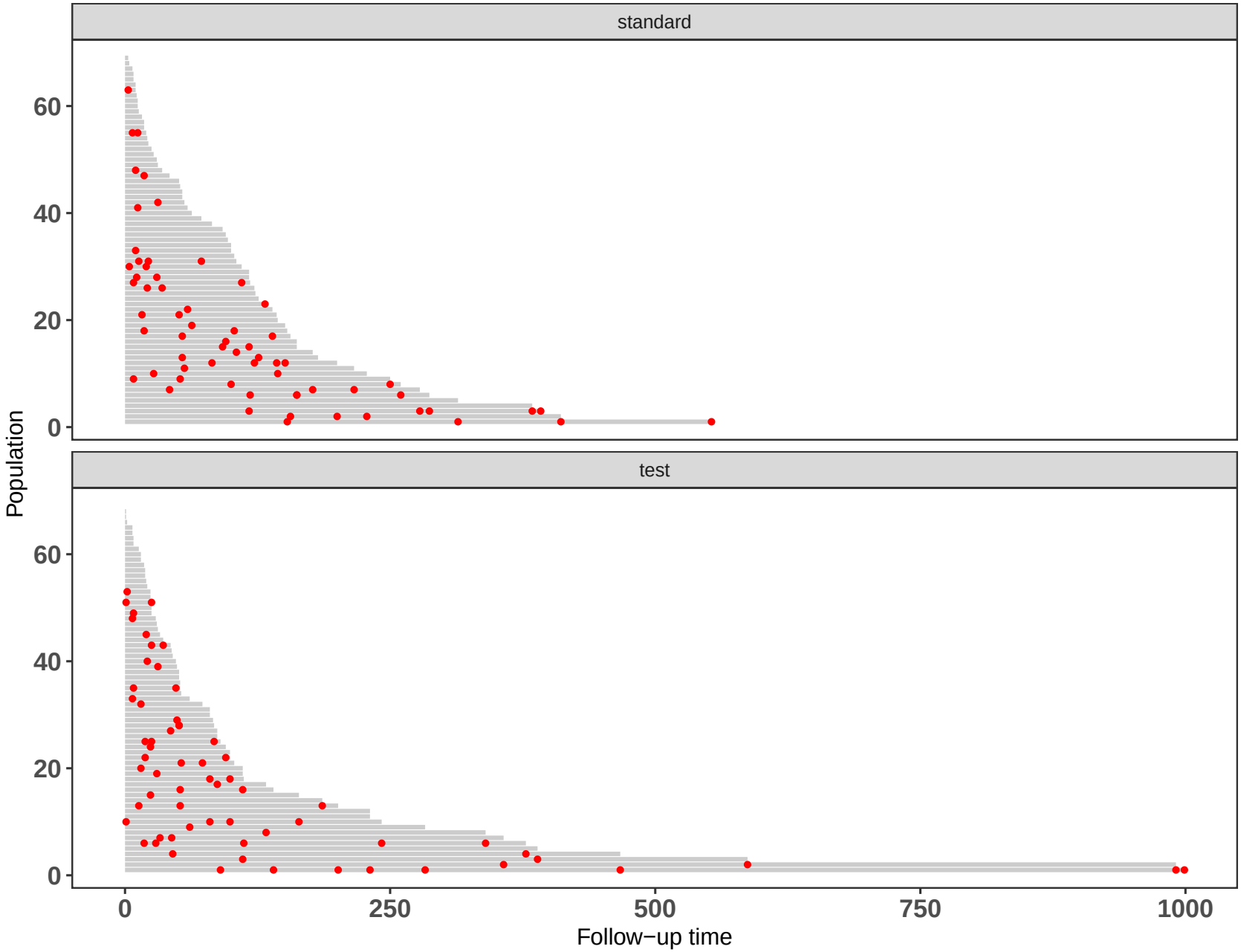
```
^^I^^I^^IabsoluteRisk(object, time, newdata,
^^I^^I^^Imethod = c("numerical", "montecarlo"),
^^I^^I^^Insamp = 100, onlyMain = TRUE, ...)
^^I^^I^^I
^^I^^I
```

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

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

```
^^I^^I^^Iphreg(Surv(time, status) ~ karno + diagtime + age +
^^I^^I^^Iprior + celltype + trt,
^^I^^I^^Idata = veteran, shape = 0, dist = "weibull")
^^I^^I^^I
^^I^^I^^IfitSmoothHazard(status ~ log(time) + karno + diagtime +
^^I^^I^^Iage + prior + celltype + trt,
^^I^^I^^Idata = veteran)
^^I^^I^^I
^^I^^I^^Icoxph(Surv(time, status) ~ karno + diagtime + age +
^^I^^I^^Iprior + celltype + trt, data = veteran)
^^I^^I
```

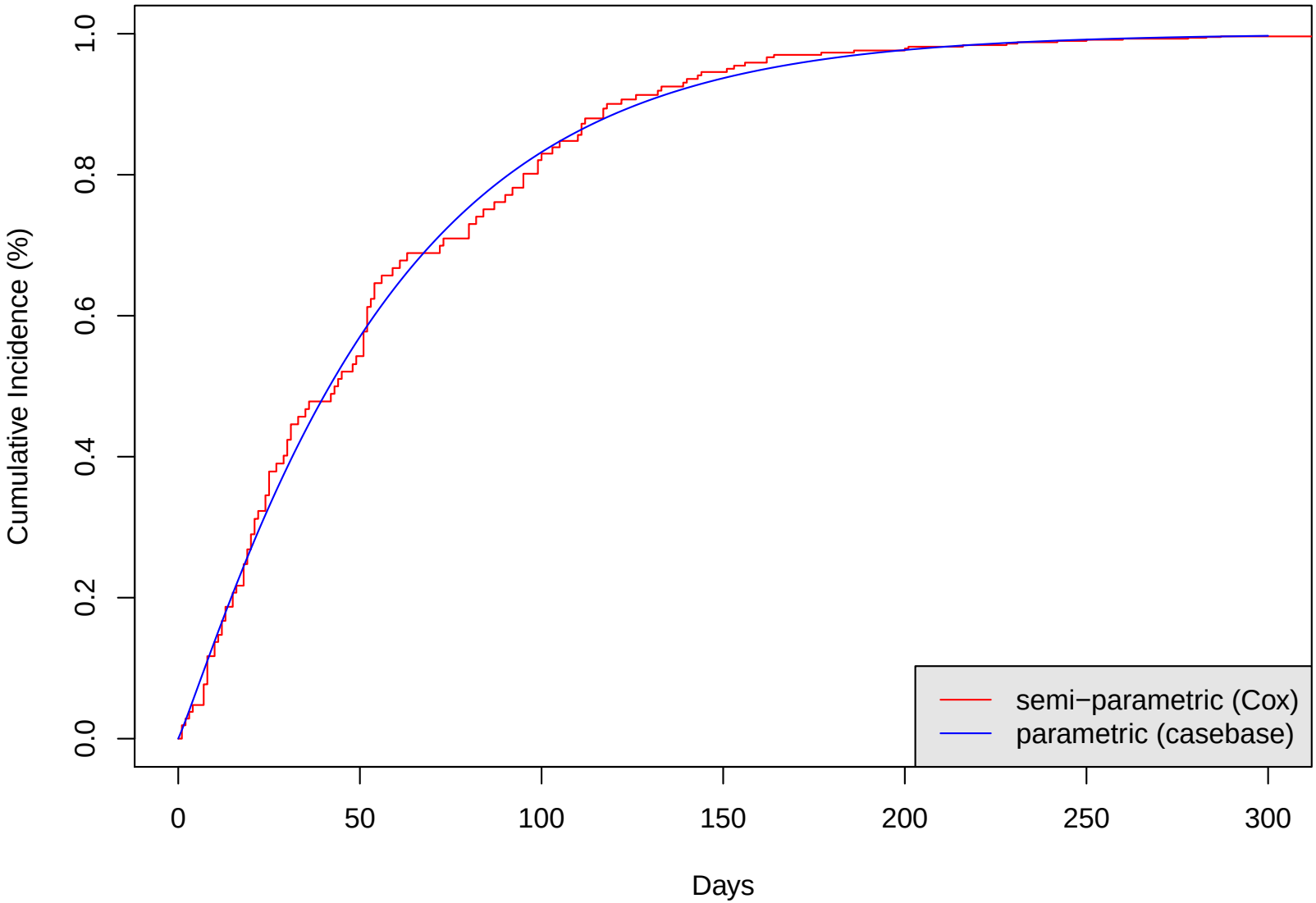
Veteran data–Estimates

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)

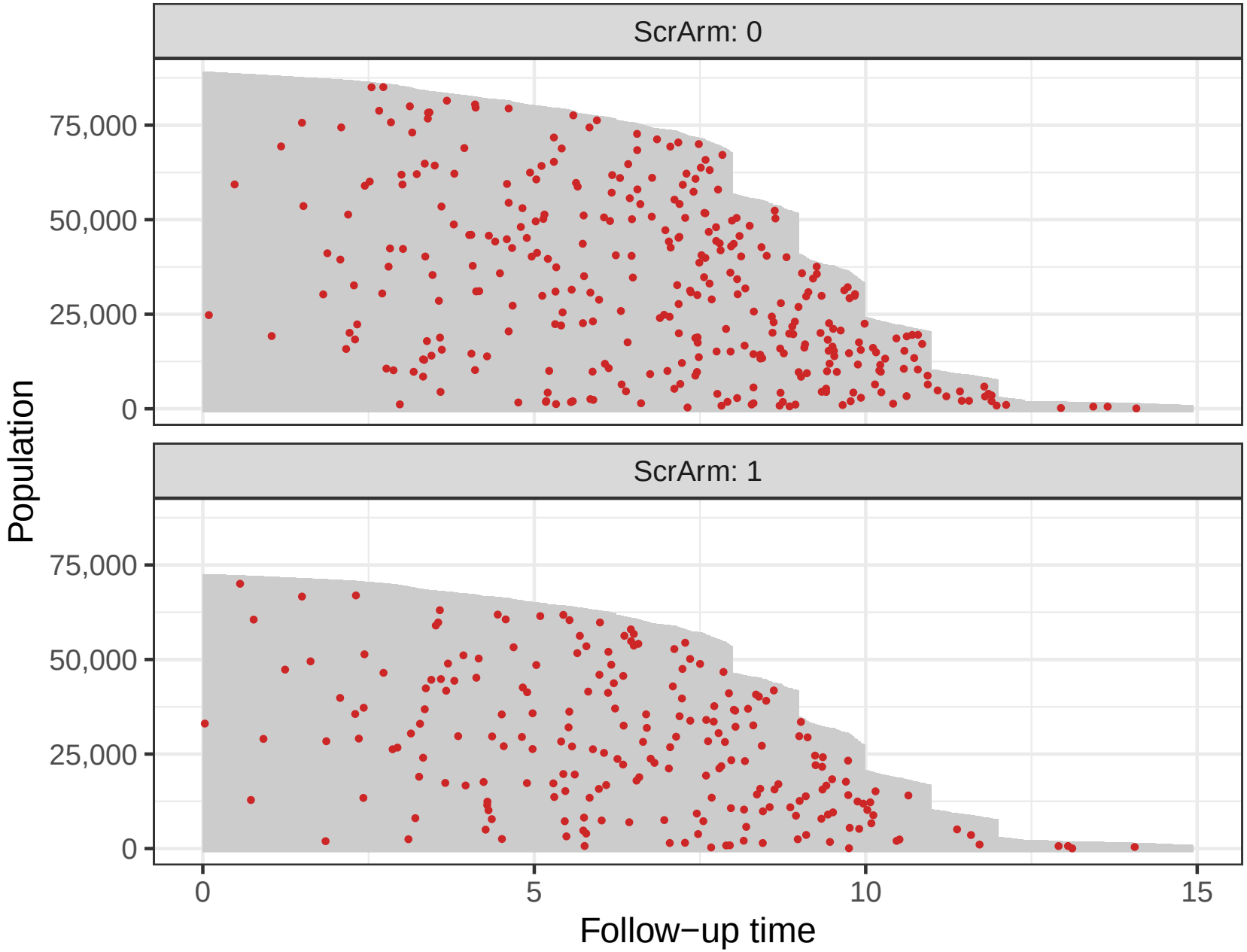
Veteran data–Risk plot



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



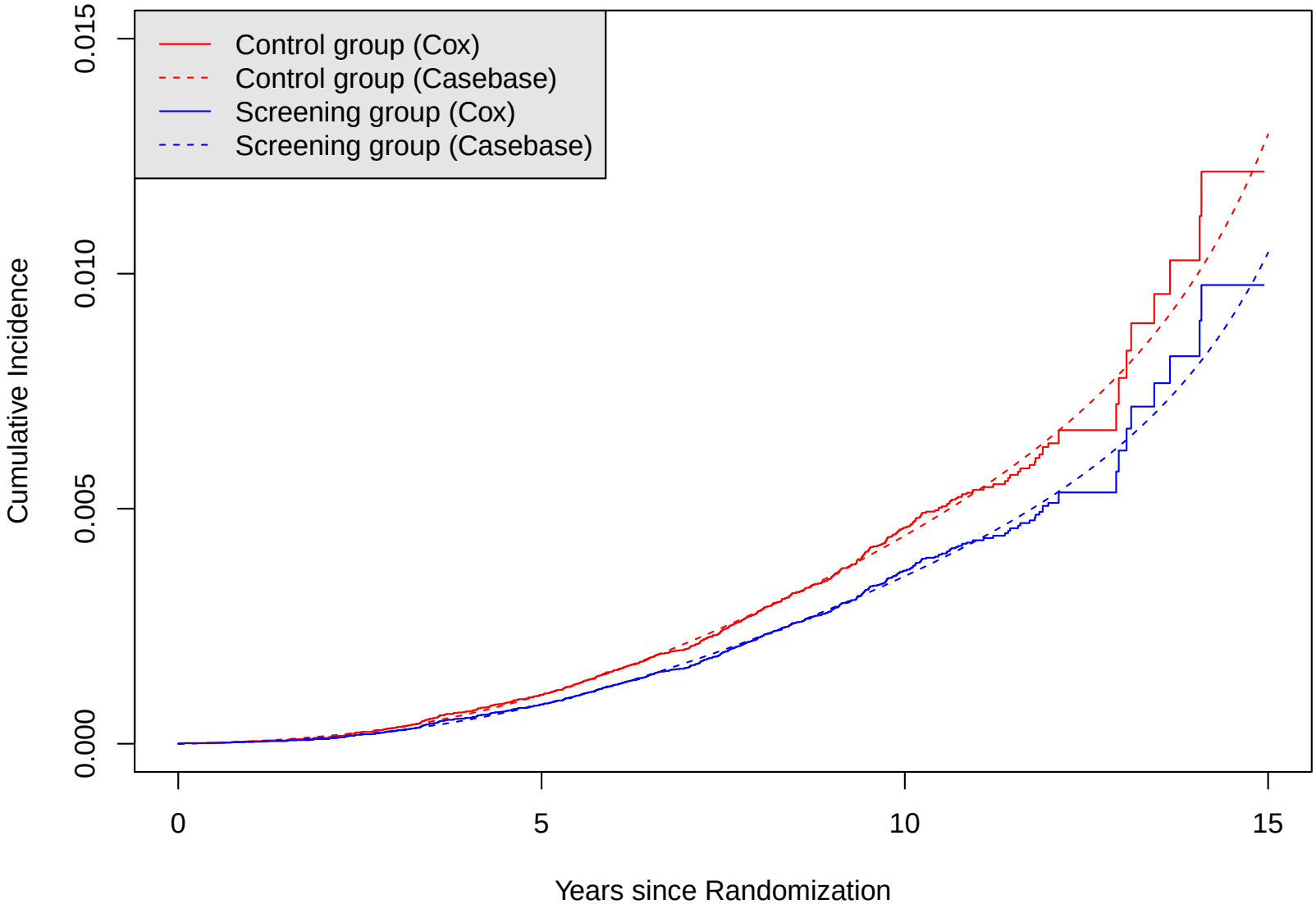
ERSPC data–Model fit

```
^^I^^I^^Ilibrary(splines)
^^I^^I^^I
^^I^^I^^Icoxph(Surv(Follow.Up.Time, DeadOfPrCa) ~ ScrArm,
^^I^^I^^Idata = ERSPC)
^^I^^I^^I
^^I^^I^^IfitSmoothHazard(DeadOfPrCa ~ bs(Follow.Up.Time) + ScrAr
^^I^^I^^Idata = ERSPC)
^^I^^I
```

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



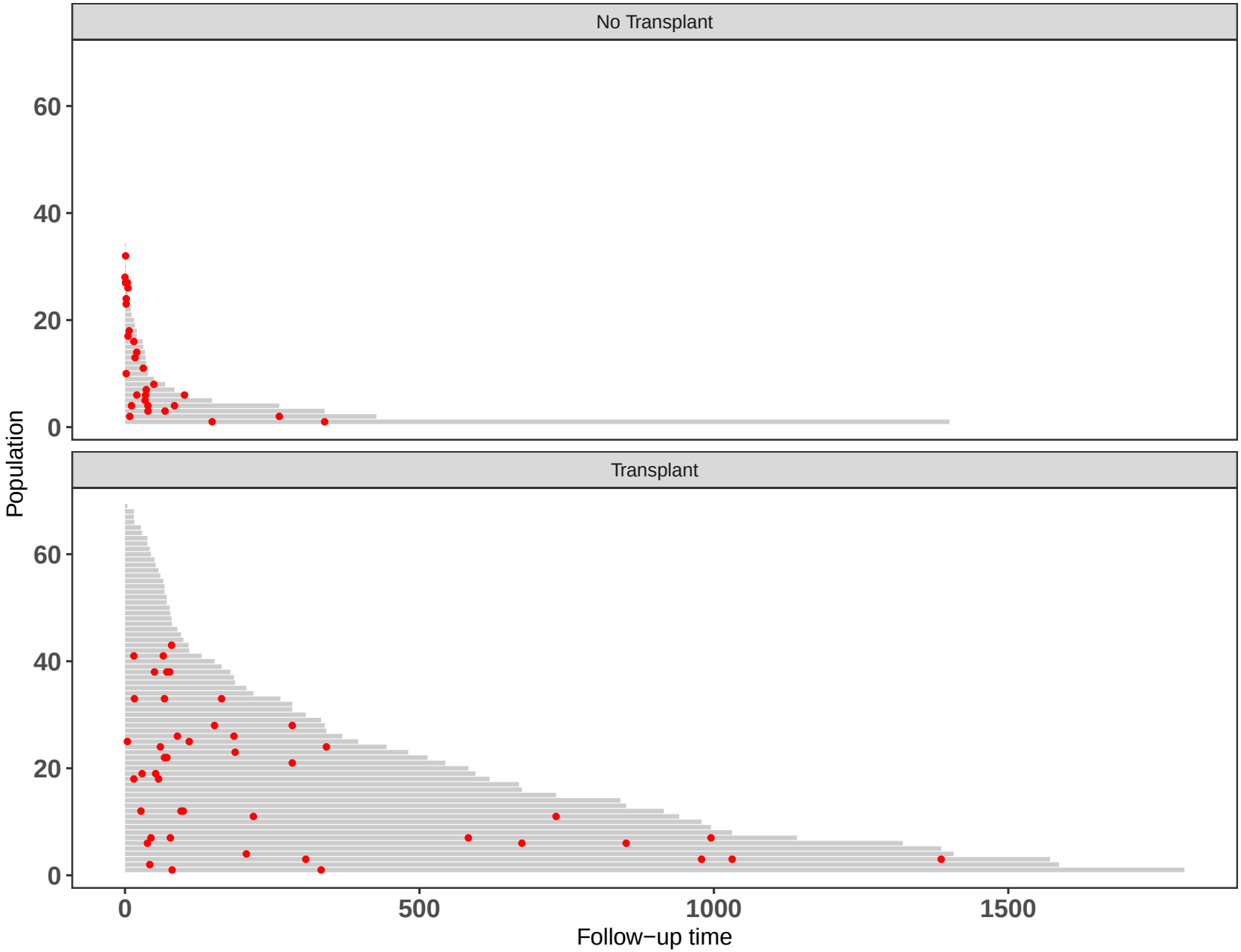
Non-proportional hazard

- 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



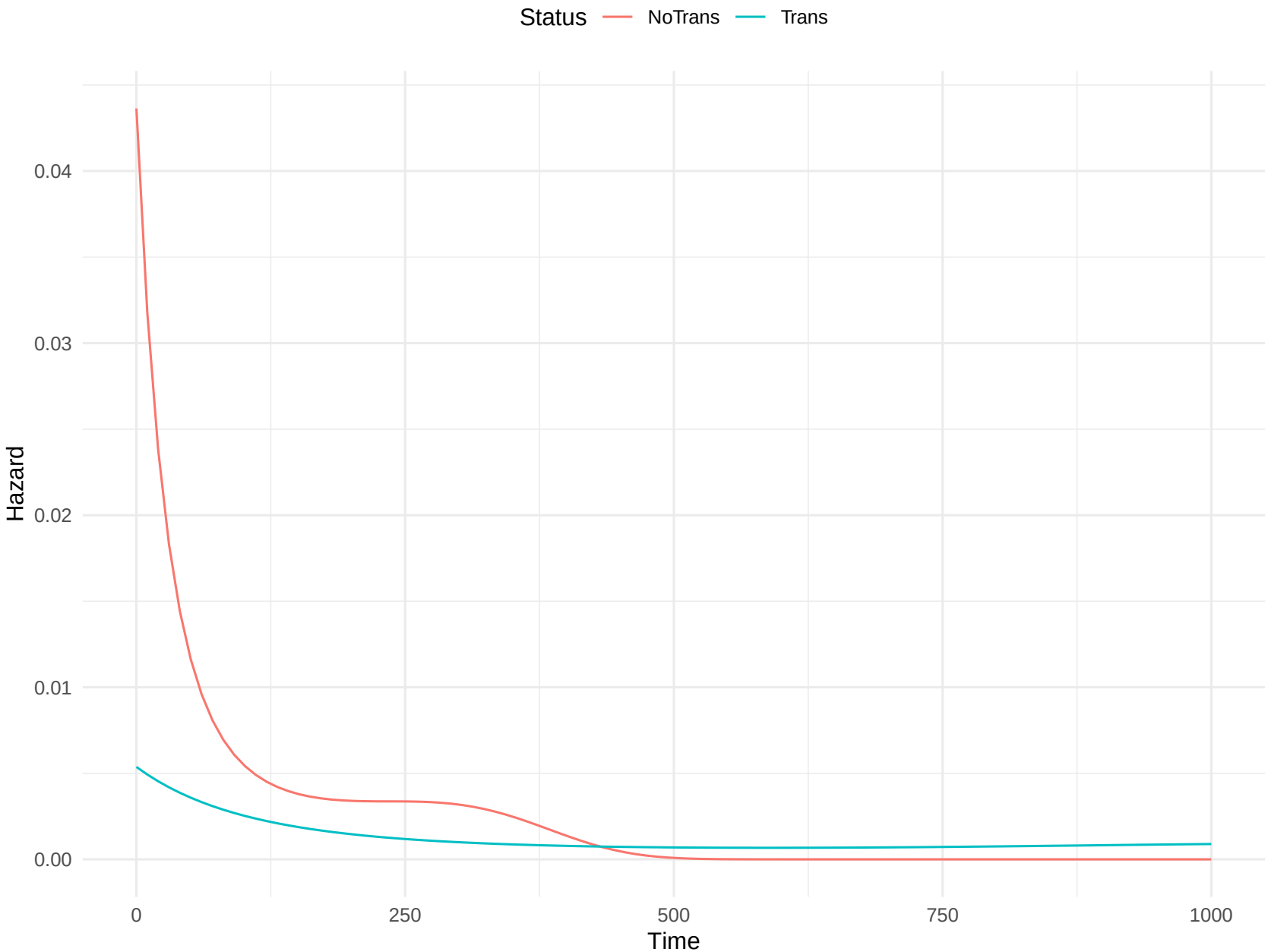
Stanford data–Model fit

```
^^I^^I^^Ifit1 <- fitSmoothHazard(fustat ~ transplant,
^^I^^I^^Idata = jasa, time = "futime")
^^I^^I^^I
^^I^^I^^Ifit2 <- fitSmoothHazard(fustat ~ transplant + futime,
^^I^^I^^Idata = jasa, time = "futime")
^^I^^I^^I
^^I^^I^^Ifit3 <- fitSmoothHazard(fustat ~ transplant + bs(futime
^^I^^I^^Idata = jasa, time = "futime")
^^I^^I^^I
^^I^^I^^Ifit4 <- fitSmoothHazard(fustat ~ transplant*bs(futime),
^^I^^I^^Idata = jasa, time = "futime")
^^I^^I
```

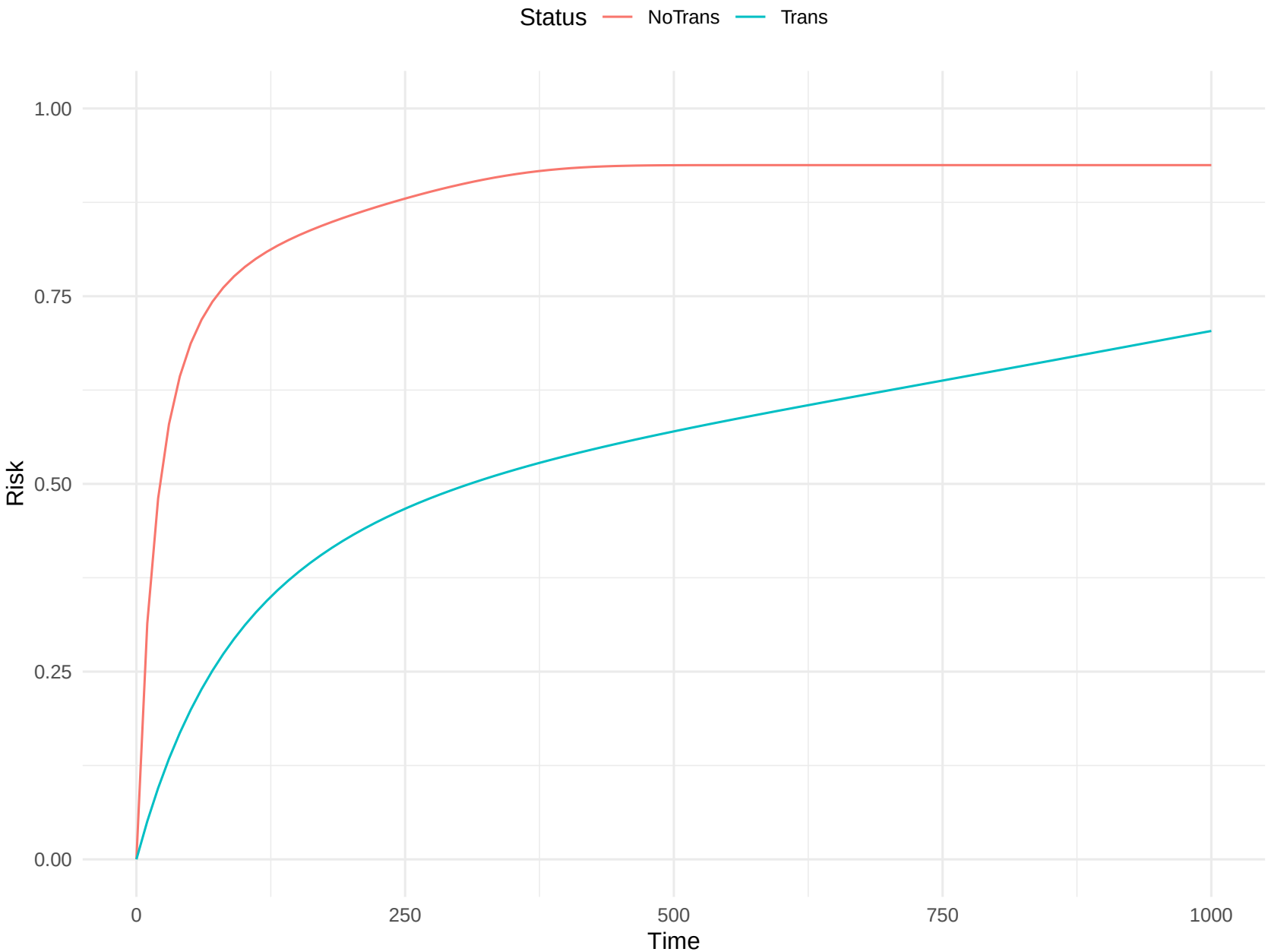
Stanford data–Model selection

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

```
^^I^^I^^IfitSmoothHazard(Status ~ bs(ftime, df = 5) + Sex + D +
^^I^^I^^IPhase + Source + Age,
^^I^^I^^Idata = bmtcrr, time = "ftime")
^^I^^I^^I
^^I^^I^^Icomp.risk(Event(ftime, Status) ~ const(Sex) + const(D)
^^I^^I^^Iconst(Phase) + const(Source) + const(Age),
^^I^^I^^Idata = bmtcrr, cause = 1, model = "fg")
^^I^^I^^I
^^I^^I^^Icoxph(Surv(ftime, Status == 1) ~ Sex + D + Phase +
^^I^^I^^ISource + Age, data = bmtcrr)
^^I^^I
```

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