# Absolute Risk integration using penalized logistic regression

Jesse Islam

2/16/2020

# Popular methods in time-to-event analysis

- In disease etiology, we tend to make use of the proportional hazards hypothesis .
  - Cox Regression
- When we want the absolute risk:
  - Parametric models
  - Breslow estimator

# Motivations for a new method

- Julien and Hanley found that survival analysis rarely produces prognostic functions, even though the software is widely available in cox regression packages. [1]
- They believe the stepwise nature is the reason, as it reduces interpretability. [1]
- Want to easily model non-proportional hazards. [1]
- A streamlined approach for reaching a **smooth absolute risk** curve. [1]

# Dr. Cox's perspective

**Reid**: How do you feel about the cottage industry that's grown up around it [the Cox model]?
**Cox**: Don't know, really. In the light of some of the further results one knows since, I think I would normally want to tackle problems parametrically, so I would take the underlying hazard to be a Weibull or something. I'm not keen on nonparametric formulations usually.
**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.

---

PatientID ScrArm Follow.Up.Time DeadOfPrCa

# Recall

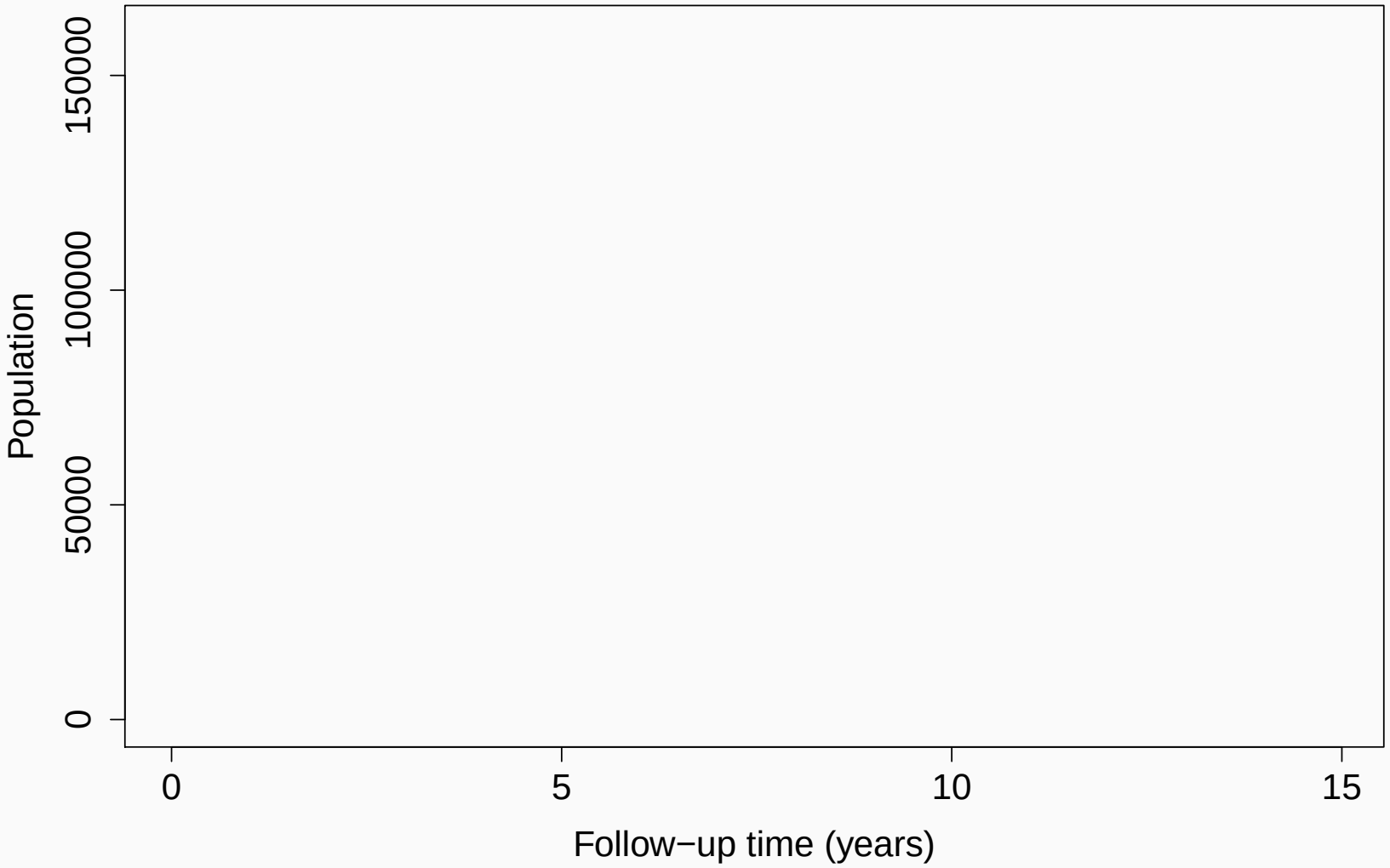
- **Using the ERSPC dataset and casebase, we will determine Justin's absolute risk for death by prostate cancer.**

# Casebase Overview

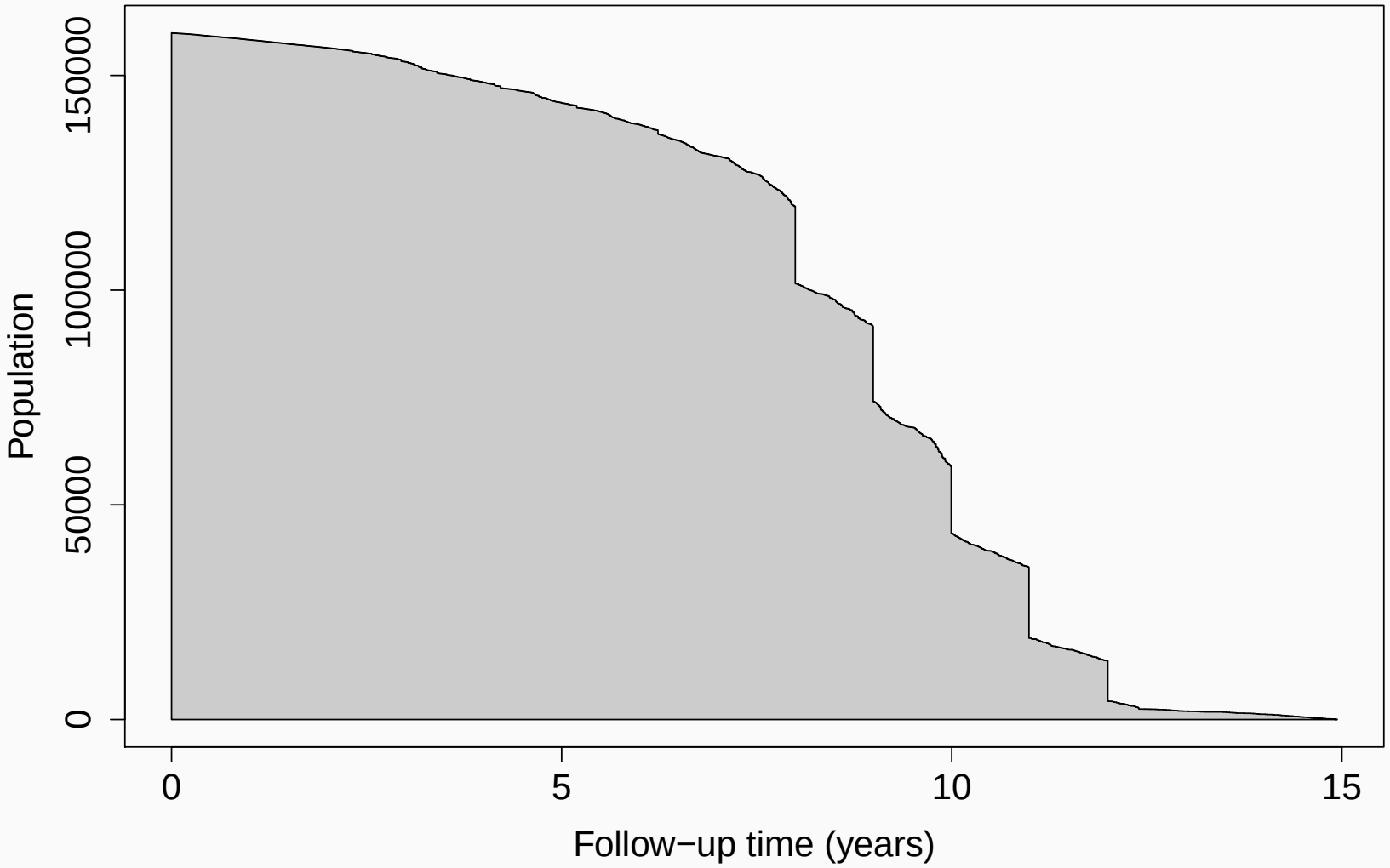
1. Clever sampling.
2. Implicitly deals with censoring.
3. Allows a parametric fit using *logistic regression*.

- Casebase is parametric, and allows different parametric fits by incorporation of the time component.
- Package contains an implementation for generating *population-time* plots.

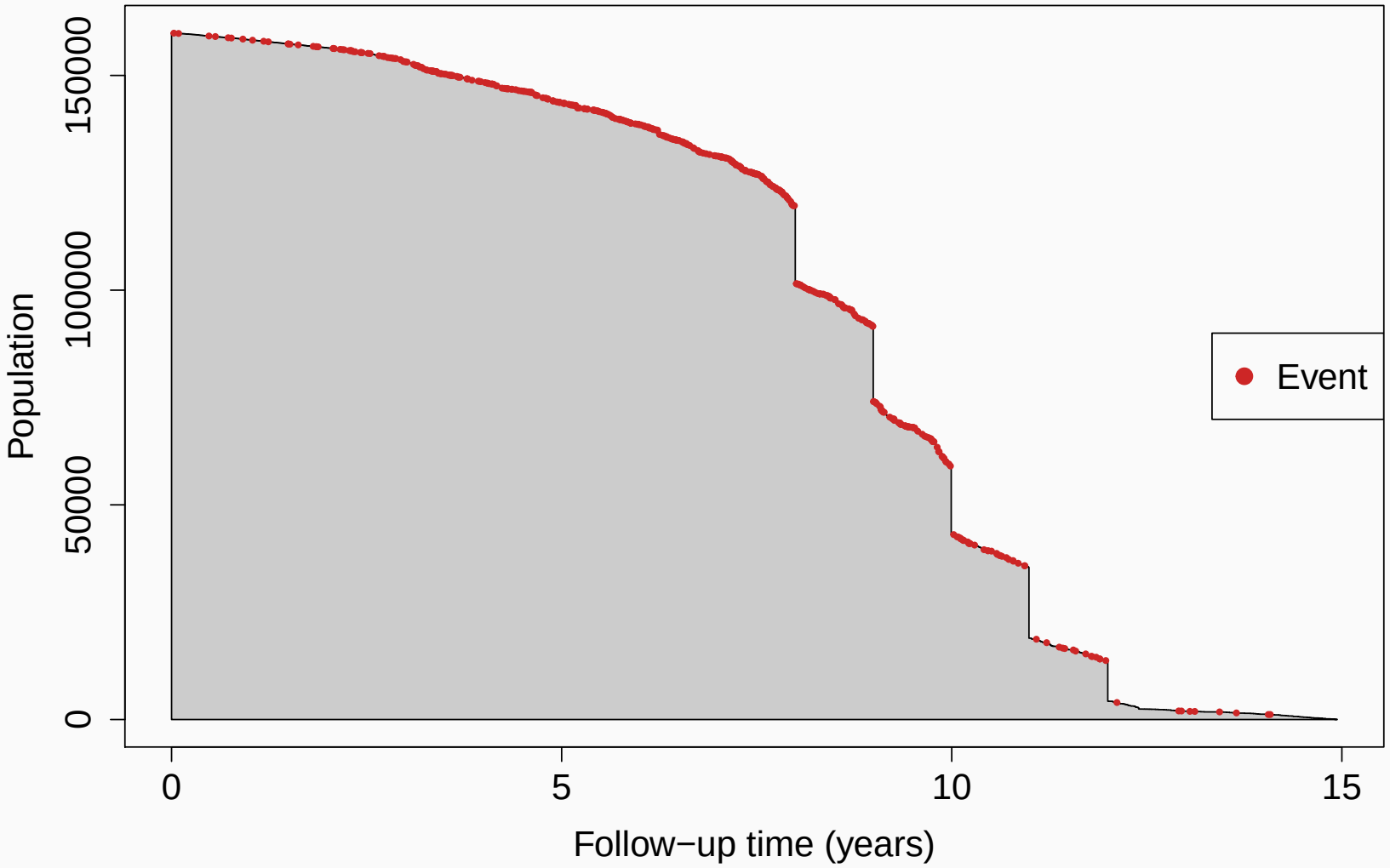
# Casebase: Sampling



# Casebase: Sampling

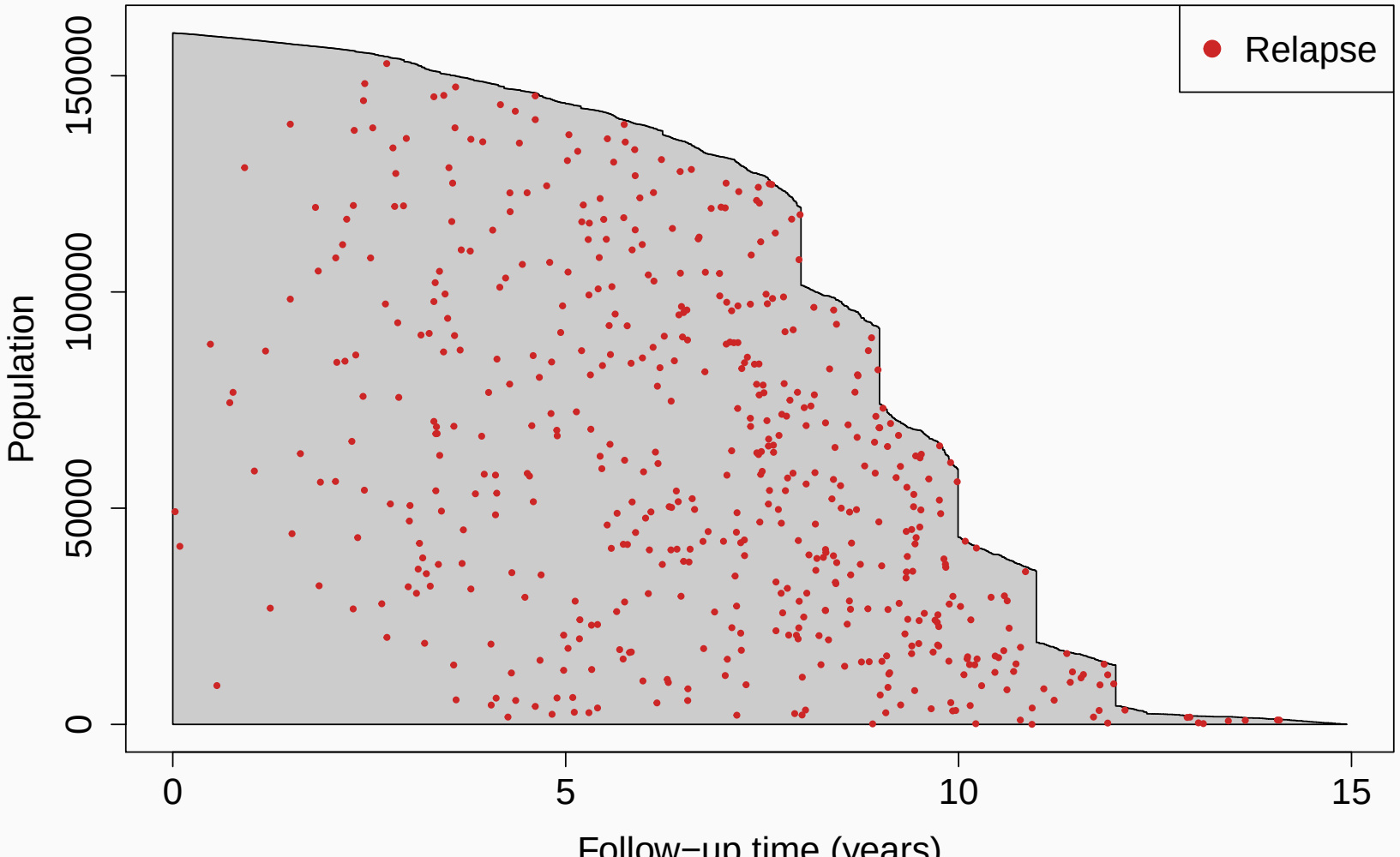


# Casebase: Sampling

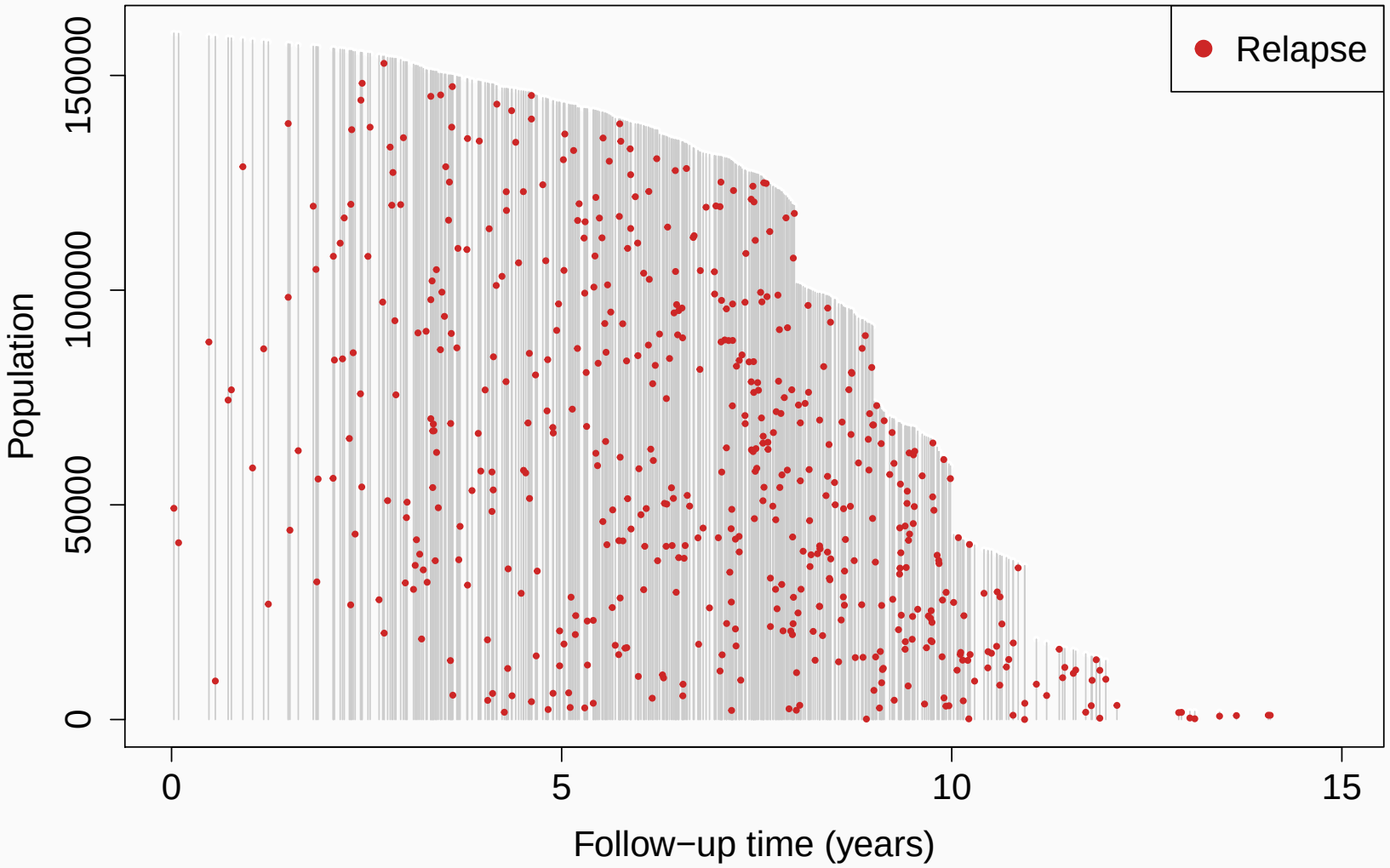


# Casebase: Sampling

```
casebase::popTime(Data,Event,Time)
```



# Casebase: Sampling



# Casebase: Parametric families

- We can now fit models of the form:

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

- By changing the function $g(t; \alpha)$, we can model different parametric families easily:

# Casebase: Parametric models

*Exponential*: $g(t; \alpha)$ is equal to a constant

```
casebase::fitSmoothHazard(status ~ X1 + X2)
```

*Gompertz*: $g(t; \alpha) = \alpha t$

```
casebase::fitSmoothHazard(status ~ time + X1 + X2)
```

*Weibull*: $g(t; \alpha) = \alpha log(t)$

```
casebase::fitSmoothHazard(status ~ log(time) + X1 + X2)
```

# Death by prostate cancer: hazard ratios

```
casebase::fitSmoothHazard(DeadOfPrCa~ log(Follow.Up.Time)+
                                ScrArm, data=ERSPC, ratio = 100)
```

# ERSPC Hazard comparison

| Model | Hazard Ratio | Std.Error |
|---|---|---|
| Cox | 0.801 | 1.092 |
| Gompertz | 0.802 | 1.093 |
| Exponential | 0.810 | 1.092 |
| Weibull | 0.797 | 1.093 |

# Absolute Risk

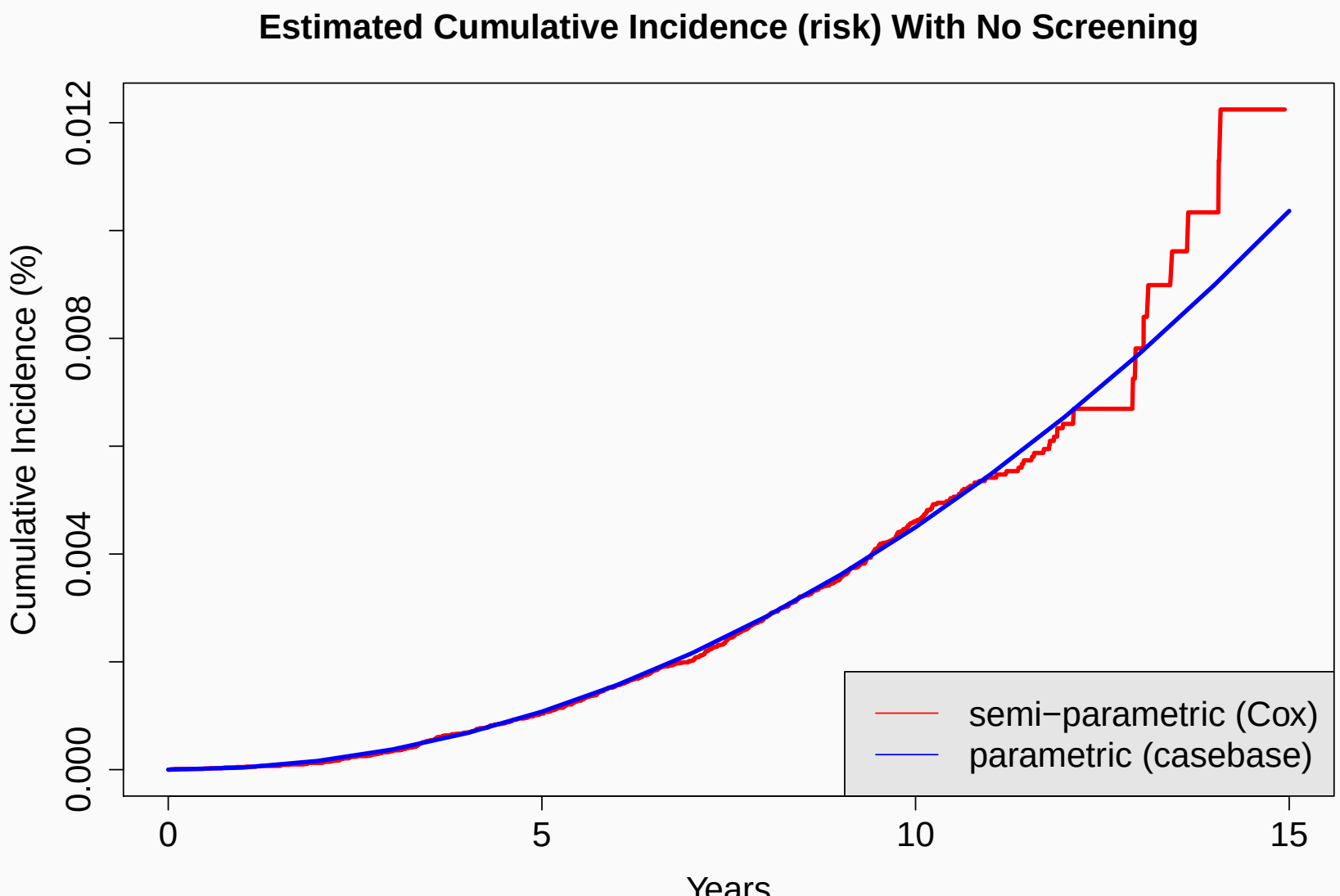
- We have a bunch of different parametric hazard models now.
- To get the absolute risk, we need to evaluate the following equation in relation to the hazard:

$$CI(x,t) = 1 - e^{-\int_0^t h(x,u)du}$$

- $CI(x,t)$= Cumulative Incidence (Absolute Risk)
- $h(x,u)$= Hazard function
- Lets use the weibull hazard

# Casebase: Absolute Risk comparison

```
casebase::absoluteRisk(fit, time=5, covariate_profile)
```



# Summary

- Casebase sampling implicitly incorporates censoring and permits the use of GLMs and the tools associated with them
- The casebase package contains tools to generate:
  - Population-Time plots
  - Hazard functions
  - Absolute Risk
  - Casebase can deal with competing risks.

# References 1

1.Hanley, James A, and Olli S Miettinen. 2009. "Fitting Smooth-in-Time Prognostic Risk Functions via Logistic Regression." *The International Journal of Biostatistics 5 (1)*.

2.Saarela, Olli, and Elja Arjas. 2015. "Non-Parametric Bayesian Hazard Regression for Chronic Disease Risk Assessment." Scandinavian Journal of Statistics 42 (2). Wiley Online Library: 609–26.

3.Saarela, Olli. 2015. "A Case-Base Sampling Method for Estimating Recurrent Event Intensities." *Lifetime Data Analysis*. Springer, 1–17

# References 2

4.Schroder FH, et al., for the ERSPC Investigators.Screening and Prostate-Cancer Mortality in a Randomized European Study. *N Engl J Med* 2009;360:1320-8.

5.Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. *Bone Marrow Transplant*. 2007 Aug;40(4):381-7. doi: 10.1038/sj.bmt.1705727.

6.Turgeon, M. (2017, June 10). Retrieved May 05, 2019, from https://www.maxturgeon.ca/slides/MTurgeon-2017-Student-Conference.pdf

# Tutorial and Slides

Tutorial:

http://sahirbhatnagar.com/casebase/

Slides:

https://github.com/Jesse-Islam/UseR–CaseBase-Presentation

Questions?