Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

casebase: An Alternative Framework For Survival Analysis

Sahir Bhatnagar * McGill Univeristy

Maxime Turgeon * McGill University

James Hanley McGill Univeristy

Olli Saarela University of Toronto

Abstract

The abstract of the article. * joint co-authors

Keywords: keywords, not capitalized, Java.

1. Code formatting

Don't use markdown, instead use the more precise latex commands:

- Java
- **plyr**
- `print("abc")`

2. Introduction

- Motivation
  - Flexible
  - Flexible
  - Flexible

3. Theoretical details

4. Implementation details

1. Population-tim plots
2. Sampling
3. Fitting
4. Absolute Risks

5. Population-time plots

6. Case study 1: Veteran data (or ERSPC if we can)

- First example
- Show how we can test for non-proportional hazard?

7. Case study 2: Bone-marrow transplant

The next example shows how case-base sampling can also be used in the context of a competing risk analysis. For illustrative purposes, we will use the same data that was used in Scrucca *et al* (2010). The data was downloaded from the main author's website, and it is also available as part of the **casebase** package.

```
R> library(casebase)
R> data(bmtcrr)
```

The data contains information on 177 patients who received a stem-cell transplant for acute leukemia. The event of interest is relapse, but other competing causes (e.g. transplant-related death) were also recorded Several covariates were also captured at baseline: sex, disease type (acute lymphoblastic or myeloblastic leukemia, abbreviated as ALL and AML, respectively), disease phase at transplant (Relapse, CR1, CR2, CR3), source of stem cells (bone marrow and peripheral blood, coded as BM+PB, or only peripheral blood, coded as PB), and age. A summary of these baseline characteristics appear in Table 1. We note that the statistical summaries were generated differently for different variable types: for continuous variables, we gave the range, followed by the mean and standard deviation; for categorical variables, we gave the counts for each category.

In order to try and visualize the incidence density of relapse, we can look at the corresponding population-time plot. In Figure 1, failure times associated with relapse are highlighted on the plot using red points, while Figure 2 provides a similar population-time plot for competing events.

Our main objective is to compute the absolute risk of relapse for a given set of covariates. First, we fit a smooth hazard to the data; for the sake of this example, we opted for a linear term for time:

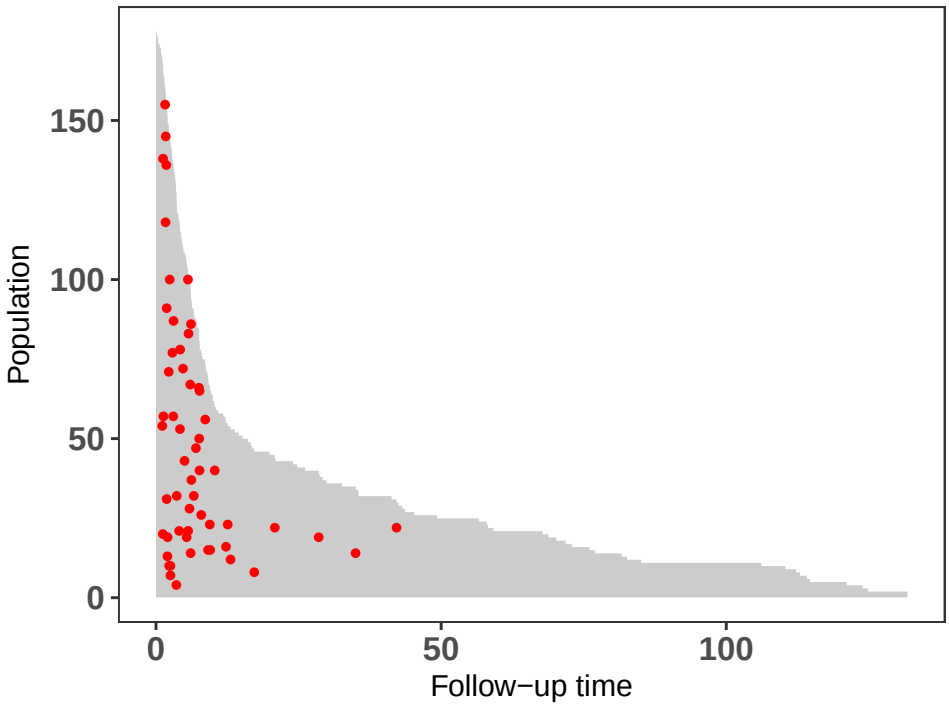

Figure 1: Population-time plot for the stem-cell transplant study. The points represent the event of interest (i.e., relapse).

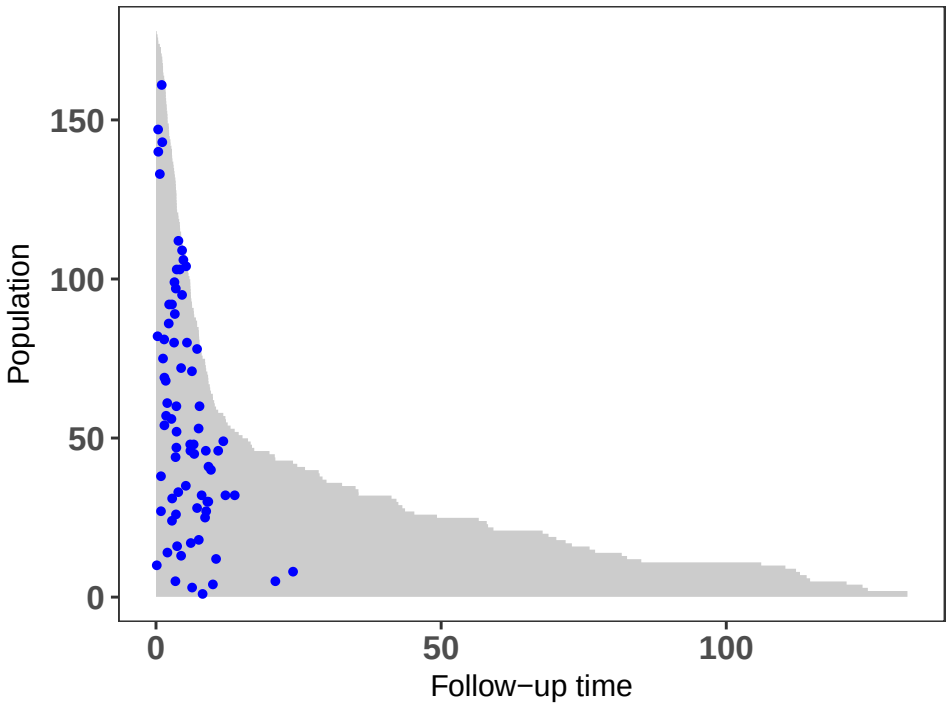

Figure 2: Population-time plot for the stem-cell transplant study. The points represent the competing events.

Variable	Description	Statistical summary
Sex	Sex	M=Male (100)
		F=Female (77)
D	Disease	ALL (73)
		AML (104)
Phase	Phase	CR1 (47)
		CR2 (45)
		CR3 (12)
		Relapse (73)
Source	Type of transplant	BM+PB (21)
		PB (156)
Age	Age of patient (years)	4–62
		30.47 (13.04)
Ftime	Failure time (months)	0.13–131.77
		20.28 (30.78)
Status	Status indicator	0=censored (46)
		1=relapse (56)
		2=competing event (75)

Table 1: Baseline characteristics of patients in the stem-cell transplant study.

```
R> model_cb <- fitSmoothHazard(
R+     Status ~ ftime + Sex + D + Phase + Source + Age,
R+     data = bmtcrr,
R+     ratio = 100,
R+     time = "ftime")
```

From the fit object, we can extract both the hazard ratios and their corresponding confidence intervals:

Covariates	HR	95% CI
Sex	0.71	(0.41, 1.24)
Disease	0.52	(0.28, 0.93)
Phase (CR2 vs. CR1)	1.17	(0.47, 2.93)
Phase (CR3 vs. CR1)	1.78	(0.46, 6.91)
Phase (Relapse vs. CR1)	4.45	(2.07, 9.57)
Source	1.64	(0.54, 5)
Age	0.99	(0.97, 1.02)

As we can see, the only significant hazard ratio is the one associated with the phase of the disease at transplant. More precisely, being in relapse at transplant is associated with a hazard ratio of 3.92 when compared to CR1.

Given our estimate of the hazard function, we can compute the absolute risk curve for a fixed covariate profile. We performed this computation for a 35 year old woman who received a stem-cell transplant from peripheral blood at relapse. We compared the absolute risk curve for such a woman with acute lymphoblastic leukemia with that for a similar woman with acute myeloblastic leukemia. Figure 3 shows these two curves as a function of time. This figure also shows the Kaplan-Meier estimate fitted to the two disease groups (ignoring the

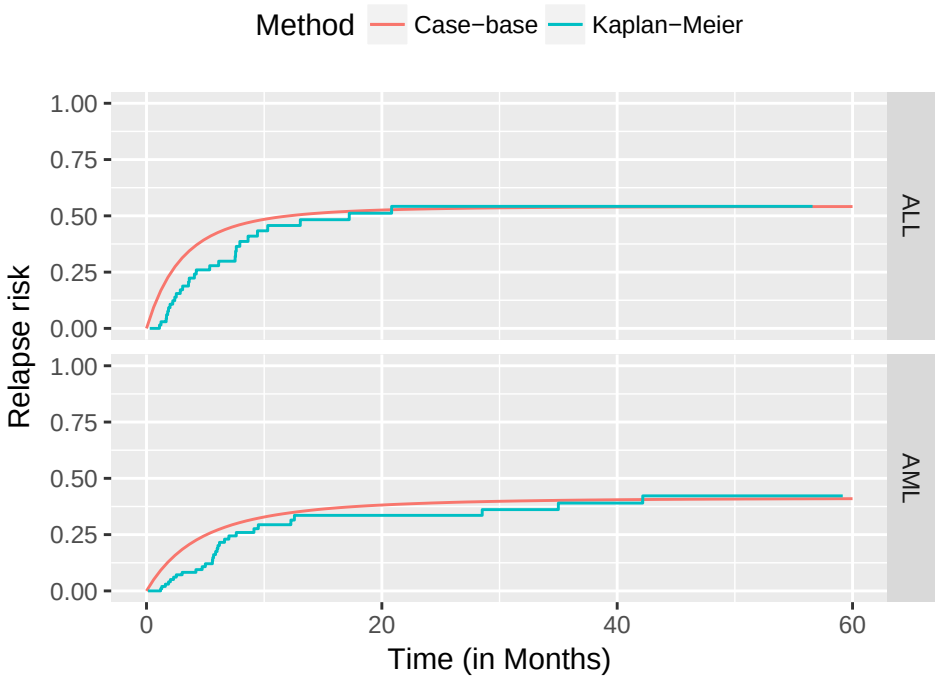

Figure 3: Absolute risk curve for a fixed covariate profile and the two disease groups. The estimate obtained from case-base sampling is compared to the Kaplan-Meier estimate.

other covariates).

```
R> # Pick 100 equidistant points between 0 and 60 months
R> time_points <- seq(0, 60, length.out = 100)
R>
R> # Data.frame containing risk profile
R> newdata <- data.frame("Sex" = factor(c("F", "F"),
R+                                       levels = levels(bmtcrr[,"Sex"])),
R+                       "D" = c("ALL", "AML"),
R+                       "Phase" = factor(c("Relapse", "Relapse"),
R+                                        levels = levels(bmtcrr[,"Phase"])),
R+                       "Age" = c(35, 35),
R+                       "Source" = factor(c("PB", "PB"),
R+                                         levels = levels(bmtcrr[,"Source"])))
R>
R> # Estimate absolute risk curve
R> risk_cb <- absoluteRisk(object = model_cb, time = time_points,
R+                         method = "montecarlo", newdata = newdata)
```

8. Case study 3: Vaccination study (recurrent events)

- Give a more complex example of sampling; time-dependent exposure

– Sampling needs to be done manually, but fitting function can still be used

9. Discussion

References

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

Affiliation:

Sahir Bhatnagar *
McGill Univeristy
1020 Pine Avenue West Montreal, QC, Canada H3A 1A2
E-mail: sahir.bhatnagar@mail.mcgill.ca
URL: http://sahirbhatnagar.com/

Maxime Turgeon *
McGill University
1020 Pine Avenue West Montreal, QC, Canada H3A 1A2
E-mail: maxime.turgeon@mail.mcgill.ca
URL: http://maxturgeon.ca/

James Hanley
McGill Univeristy
1020 Pine Avenue West Montreal, QC, Canada H3A 1A2
E-mail: james.hanley@mcgill.ca
URL: http://www.medicine.mcgill.ca/epidemiology/hanley/

Olli Saarela
University of Toronto
Dalla Lana School of Public Health, 155 College Street, 6th floor, Toronto, Ontario M5T 3M7, Canada
E-mail: olli.saarela@utoronto.ca
URL: http://individual.utoronto.ca/osaarela/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd