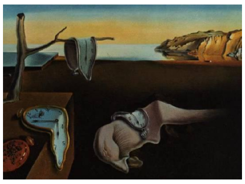

Journal of Statistical Software

MMMMMMM 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

As discussed in Hanley & Miettinen (2009), the key idea behind case-base sampling is to discretize the study base into an infinite amount of *person moments*. These person moments are indexed by both an individual in the study and a time point, and therefore each person moment has a covariate profile, an exposure status and an outcome status attached to it. We note that there is only a finite number of person moments associated with the event of interest (what Hanley & Miettinen call the *case series*). The case-base sampling refers to the sampling from the base of a representative finite sample called the *base series*.

Popular survival analysis methods, like Kaplan-Meier and Cox regression, rely on the notion of risk-set sampling. Case-base sampling can be seen as an alternative.

As shown by Saarela & Arjas (2015) (and further expanded in Saarela (2016)), writing the likelihood arising from this data-generating mechanism using the framework of non-homogenous Poisson processes, we eventually reach an expression where each person-moment's contribution is of the form

$$\frac{\lambda(t)^{dN(t)}}{\rho(t) + \lambda(t)},$$

where $N(t)$ is the counting process associated with the event of interest, $\lambda(t)$ is the corresponding hazard function, and $\rho(t)$ is the hazard function for the Poisson process associated with case-base sampling. This parametric form suggests that we can readily estimate log-hazards of the form $\log(\lambda(t)) = g(t; X)$ using logistic regression, where each observation corresponds to a person moment, the function $g(t; X)$ is linear in a finite number of parameters, and where we treat $\log(\rho(t))$ as an offset.

In Hanley & Miettinen (2009), the authors suggest performing case-base samping *uniformly*, i.e. to sample the base series uniformly from the study base. In terms of Poisson processes, this sampling strategy corresponds to a time-homogeneous Poisson process with intensity equal to b/B, where b is the number of sampled observations in the base series, and B is the total population-time for the study base. More complex examples are also available; see for example Saarela & Arjas (2015), where the probabilities of the sampling mechanism are proportional to the cardiovascular disease event rate given by the Framingham score.

4. Implementation details

The functions in the casebase package can be divided into two categories: 1) data visualization, in the form of population-time plots; and 2) data analysis.

We explicitely aimed at being compatible with both `data.frames` and `data.tables`. This is evident in some of the coding choices we made, and it is also reflected in our testing units.

4.1. Population-time plots

4.2. Data analysis

The data analysis step was separated into three parts: 1) case-base sampling; 2) estimation of the smooth hazard function; and 3) calculation of the risk function. By separating the sampling and estimation functions, we allowed the possibility of users implementing more complex sampling scheme, as described in Saarela (2016).

The sampling scheme selected for `sampleCaseBase` was described in Hanley and Miettinen (2009): we first sample along the "person" axis, proportional to each individual's total follow-up time, and then we sample a moment uniformly over their follow-up time. This sampling scheme is equivalent to the following picture: imagine representing the total follow-up time of all individuals in the study along a single dimension, where the follow-up time of the next individual would start exactly when the follow-up time of the previous individual ends. Then the base series could be sampled uniformly from this one-dimensional representation of the overall follow-up time. In any case, the output is a dataset of the same class as the input, where each row corresponds to a person-moment. The covariate profile for each such person-moment is retained, and an offset term is added to the dataset. This output could then be used to fit a smooth hazard function, or for visualization of the base series.

The fitting function `fitSmoothHazard` starts by looking at the class of the dataset: if it was generated from `sampleCaseBase`, it automatically inherited the class `cbData`. If the dataset supplied to `fitSmoothHazard` does not inherit from `cbData`, then the fitting function starts by calling `sampleCaseBase` to generate the base series. In other words, the occasional user can bypass `sampleCaseBase` altogether and only worry about the fitting function `fitSmoothHazard`.

The fitting function retains the familiar formula interface of `glm`. The left-hand side of the formula should be the name of the column corresponding to the event type. The right-hand side can be any combination of the covariates, along with an explicit functional form for the time variable. Note that non-proportional hazard models can be achieved at this stage by adding an interaction term involving time. The offset term does not need to be specified by the user, as it is automatically added to the formula.

Finally, the hazard function is fitted to the data using the function `glm`, unless there is more than one type event (i.e. in a competing-risk analysis), in which case it uses the multinomical regression capabilities of the **VGAM** package. This package was selected for its ability to fit multinomial regression models with an offset.

Once a model-fit object has been returned by `fitSmoothHazard`, all the familiar summary and diagnostic functions are available: `print`, `summary`, `predict`, `plot`, etc. Our package provides one more functionality: it computes risk functions from the model fit. For the case of a single event, it uses the familiar identity

$$S(t) = \exp\left(-\int_0^t h(u;X)du\right).$$

The integral is computed using either the `stats::integrate` function or Monte-Carlo integration. For the case of a competing-event analysis, the event-specific risk is computed using a nested double integral; in this setting, Monte-Carlo integration is faster and more flexible than `stats::integrate`. This is due to the fact that the computation involves a double loop over the selected time points; Monte-Carlo integration performs this step by using the vectorized `rowSums` and `colSums` functions.

To decide between a single-event and a competing-event analysis, we created `absoluteRisk` as an `S3` generic, with methods for both `glm` and `CompRisk` objects (the latter inherits from `vglm` as well). The method dispatch system of `R` then takes care of matching the correct input to the correct methodology, without the user's intervention.

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

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

6. 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:

```
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:

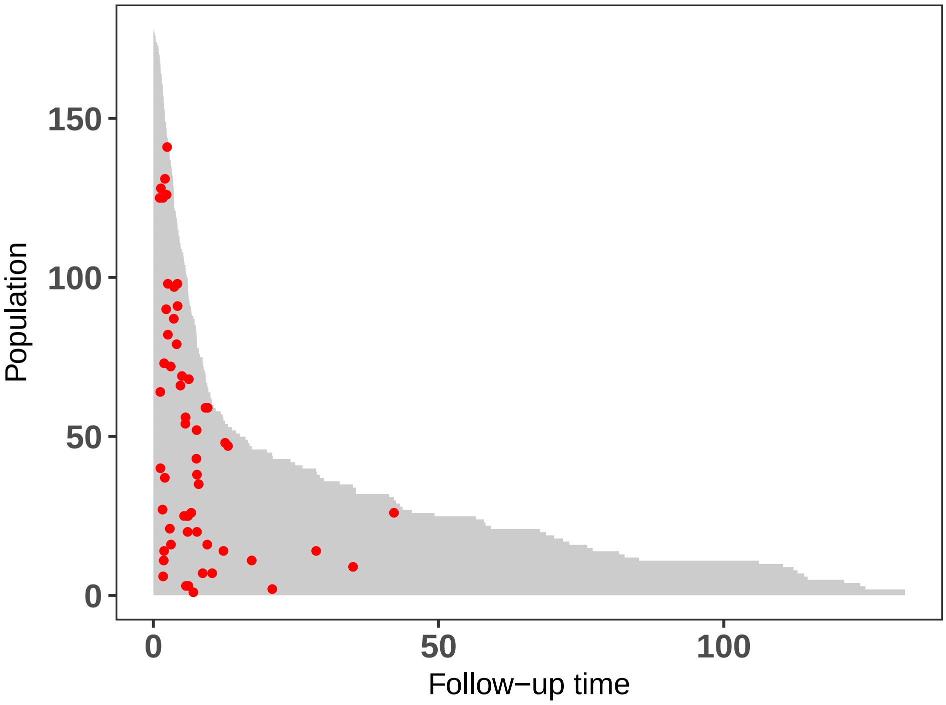

Figure 1: Population-time plot for the stem-cell transplant study. The points represent the event of interest (i.e., relapse).

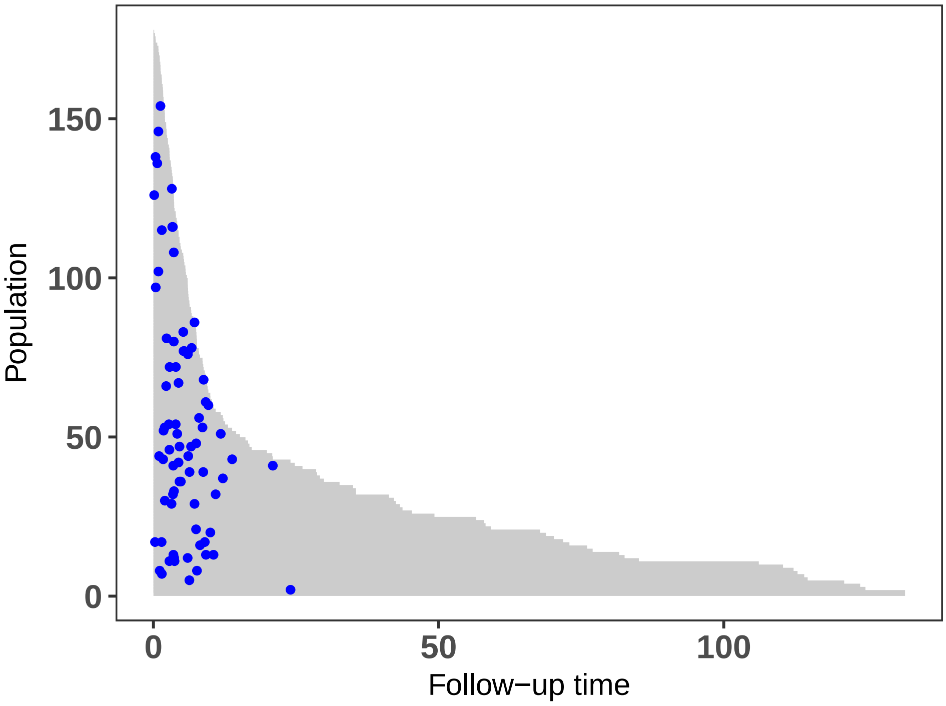

Figure 2: Population-time plot for the stem-cell transplant study. The points represent the competing events.

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.

Covariates	HR	95% CI
Sex	0.82	(0.46, 1.48)
Disease	0.48	(0.25, 0.92)
Phase (CR2 vs. CR1)	1.43	(0.56, 3.66)
Phase (CR3 vs. CR1)	1.63	(0.41, 6.45)
Phase (Relapse vs. CR1)	4.73	(2.13, 10.48)
Source	1.51	(0.49, 4.69)
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 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"])),
```

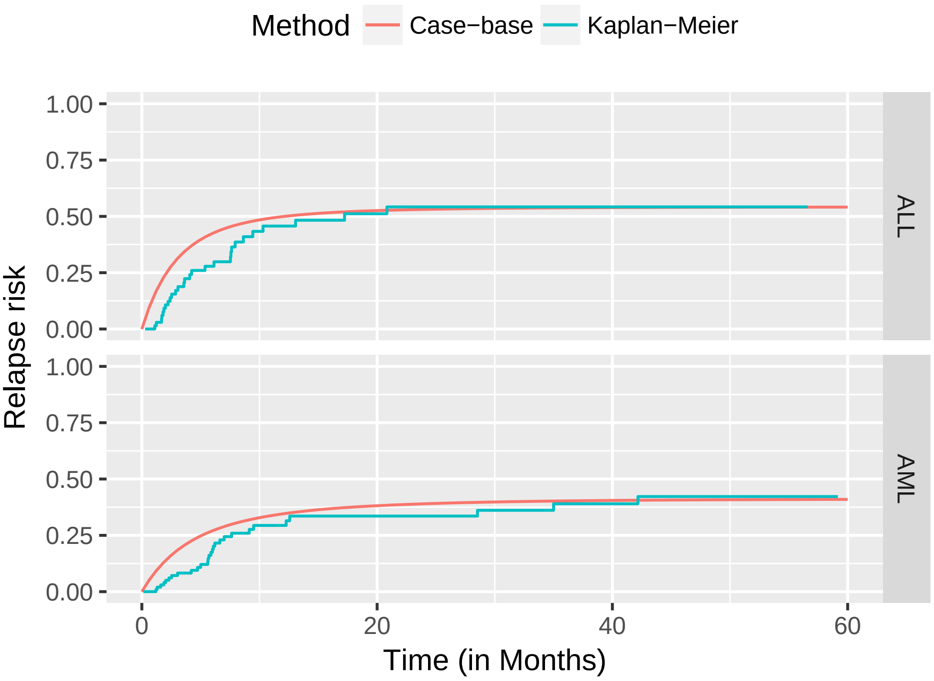

Figure 3: Absolute risk curve for a fixed covariate profile and the two disease groups. The estimate obtained from case-base sampling is compared to the Kaplan-Meier estimate.

```
R+                       "D" = c("ALL", "AML"),
R+                       "Phase" = factor(c("Relapse", "Relapse"),
                                          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)
```

7. 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

Discussion

Hanley JA, Miettinen OS (2009). “Fitting smooth-in-time prognostic risk functions via logistic regression.” *The International Journal of Biostatistics*, **5**(1).

Saarela O (2016). “A case-base sampling method for estimating recurrent event intensities.” *Lifetime data analysis*, **22**(4), 589–605.

Saarela O, Arjas E (2015). “Non-parametric Bayesian Hazard Regression for Chronic Disease Risk Assessment.” *Scandinavian Journal of Statistics*, **42**(2), 609–626.

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