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

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.frame`s and `data.table`s. 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–European Randomized Study of Prostate Cancer Screening Data

This vignette introduces the main functions in the `casebase` package. The methods implemented in this package are based on the method developped in Fitting Smooth-in-Time Prognostic Risk Functions via Logistic Regression (Hanley and Miettinen, 2009). A rigorous treatment of the theory is developed in A case-base sampling method for estimating recurrent event intensities (Saarela, 2015) and Non-parametric Bayesian Hazard Regression for Chronic Disease Risk Assessment. The motivation for this work is nicely summarised by Cox:

> Particularly notable are Cox's own reflections (Reid, 1994) on the uses of his model:
>
> > **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.

Figure 1:

The purpose of the `casebase` package is to provide practitioners with an easy-to-use software tool to predict the risk (or cumulative incidence (CI)) of an event, for a particular patient. The following points should be noted:

1. Time matching/risk set sampling (including Cox partial likelihood) eliminates the baseline hazard from the likelihood expression for the hazard ratios
2. If, however, the absolute risks are of interest, they have to be recovered using the semi-parametric Breslow estimator
3. Alternative approaches for fitting flexible hazard models for estimating absolute risks, not requiring this two-step approach? Yes! Hanley and Miettinen, 2009

   Hanley and Miettinen, 2009 propose a fully parametric hazard model that can be fit via logistic regression. From the fitted hazard function, cumulative incidence and, thus, risk functions of time, treatment and profile can be easily derived.

The `casebase` package fits the family of hazard functions of the form

$$h(x,t) = exp[g(x,t)]$$

where t denotes the numerical value (number of units) of a point in prognostic/prospective time and x is the realization of the vector X of variates based on the patient's profile and intervention (if any). Different functions of t lead to different parametric hazard models.

The simplest of these models is the one-parameter exponential distribution which is obtained by taking the hazard function to be constant over the range of t.

$$h(x,t) = exp(\beta_0 + \beta_1 x)$$

The instantaneous failure rate is independent of t, so that the conditional chance of failure in a time interval of specified length is the same regardless of how long the individual has been on study a.k.a the memoryless property (Kalbfleisch and Prentice, 2002).

The Gompertz hazard model is given by including a linear term for time:

$$h(x,t) = exp(\beta_0 + \beta_1 t + \beta_2 x)$$

Use of $log(t)$ yields the Weibull hazard which allows for a power dependence of the hazard on time (Kalbfleisch and Prentice, 2002):

$$h(x,t) = exp(\beta_0 + \beta_1 \log(t) + \beta_2 x)$$

Recall that the relative risk model (Cox, 1972)

$$\lambda(t;x) = \lambda_0(t) exp(\mathbf{X}\boldsymbol{\beta})$$

where $\lambda_0(\cdot)$ is an arbitrary unspecified baseline hazard function for continous time.

In the following table we provide a comparison between the Cox model and case-base sampling:

Slides from Olli Saarela

We fist load the required packages:

Throughout this vignette, we make use of the European Randomized Study of Prostate Cancer Screening data which ships with the `casebase` package:

```
R> data(ERSPC)
R> head(ERSPC)
```

```
  ScrArm Follow.Up.Time DeadOfPrCa
1      1         0.0027          0
2      1         0.0027          0
3      1         0.0027          0
4      0         0.0027          0
5      0         0.0027          0
6      0         0.0027          0
```

```
R> ERSPC$ScrArm <- factor(ERSPC$ScrArm,
R+                        levels = c(0,1),
R+                        labels = c("Control group", "Screening group"))
```

The results of this study were published by Schroder FH, et al. N Engl J Med 2009. There's a really interesting story on how this data was obtained. See `help(ERSPC)` and Liu Z, Rich B, Hanley JA, Recovering the raw data behind a non-parametric survival curve. Systematic Reviews 2014 for further details.

Population time plots can be extremely informative graphical displays of survival data. They should be the first step in your exploratory data analyses. We facilitate this task in the `casebase` package using the `popTime` function. We first create the necessary dataset for producing the population time plots:

```
R> pt_object <- casebase::popTime(ERSPC, event = "DeadOfPrCa")
```

```
'Follow.Up.Time' will be used as the time variable
```

We can see its contents and its class:

```
R> head(pt_object)
```

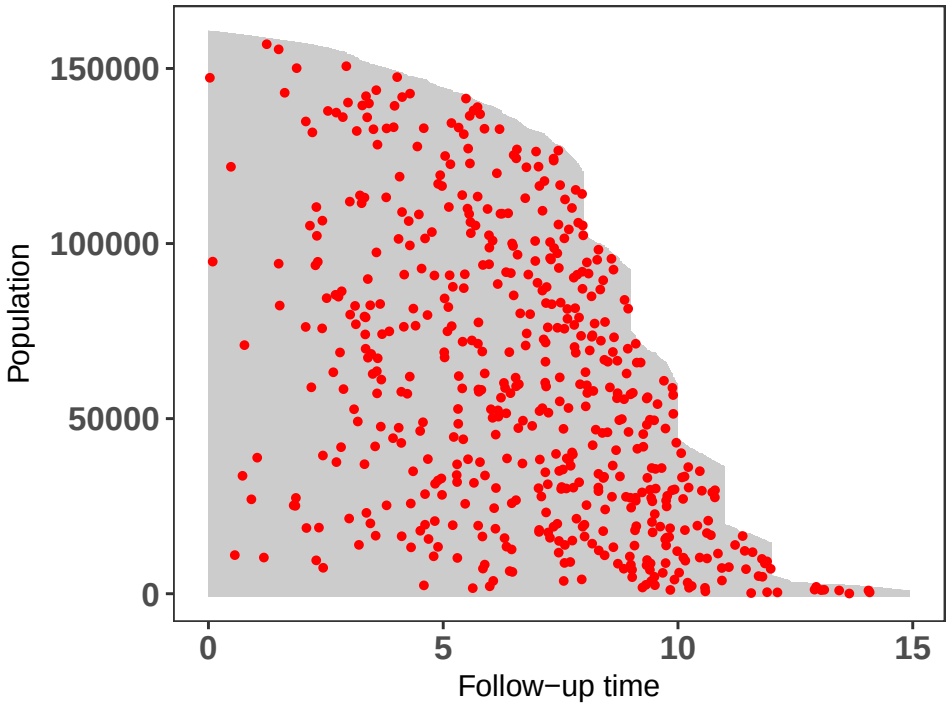
```
            ScrArm   time event original.time original.event event status
1: Screening group 0.0027     0        0.0027              0     censored
2: Screening group 0.0027     0        0.0027              0     censored
3: Screening group 0.0027     0        0.0027              0     censored
4:   Control group 0.0027     0        0.0027              0     censored
5:   Control group 0.0027     0        0.0027              0     censored
6:   Control group 0.0027     0        0.0027              0     censored
   ycoord yc n_available
1: 159893  0           0
2: 159892  0           0
3: 159891  0           0
4: 159890  0           0
5: 159889  0           0
6: 159888  0           0
```

```
R> class(pt_object)
```

```
[1] "popTime"    "data.table" "data.frame"
```

The `casebase` package has a `plot` method for objects of class `popTime`:

```
R> plot(pt_object)
```



We can also create exposure stratified plots by specifying the `exposure` argument in the `popTime` function:

```
R> pt_object_strat <- casebase::popTime(ERSPC,
R+                                      event = "DeadOfPrCa",
R+                                      exposure = "ScrArm")

'Follow.Up.Time' will be used as the time variable
```

We can see its contents and its class:

```
R> head(pt_object_strat)

$data
                ScrArm    time event original.time original.event
     1:   Control group  0.0027     0        0.0027              0
     2:   Control group  0.0027     0        0.0027              0
     3:   Control group  0.0027     0        0.0027              0
     4:   Control group  0.0027     0        0.0027              0
     5:   Control group  0.0137     0        0.0137              0
    ---
159889: Screening group 14.9405     0       14.9405              0
159890: Screening group 14.9405     0       14.9405              0
159891: Screening group 14.9405     0       14.9405              0
159892: Screening group 14.9405     0       14.9405              0
159893: Screening group 14.9405     0       14.9405              0
```

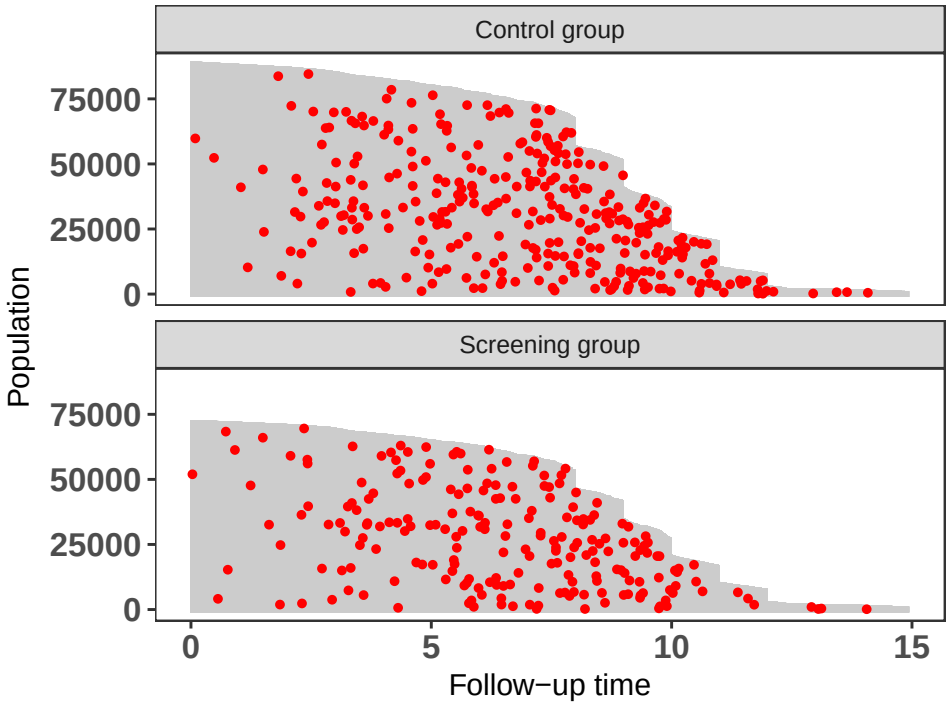
```
        event status ycoord yc n_available
     1:     censored  88232  0           0
     2:     censored  88231  0           0
     3:     censored  88230  0           0
     4:     censored  88229  0           0
     5:     censored  88228  0           0
    ---
159889:     censored      5  0           0
159890:     censored      4  0           0
159891:     censored      3  0           0
159892:     censored      2  0           0
159893:     censored      1  0           0

$exposure
[1] "ScrArm"
```

```
R> class(pt_object_strat)
```

```
[1] "popTimeExposure" "list"
```

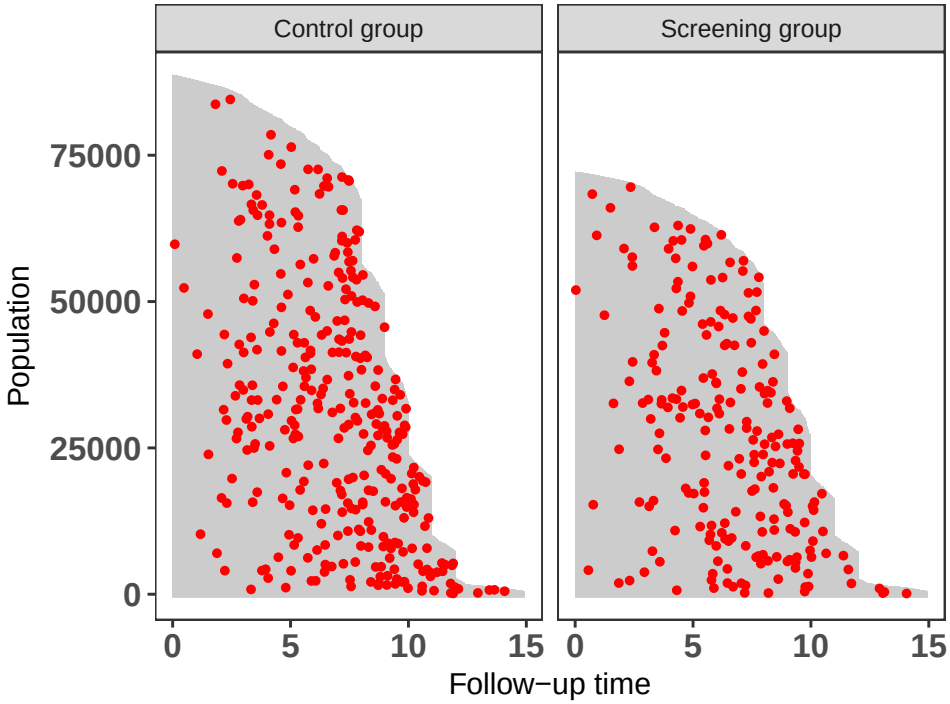
The `casebase` package also has a `plot` method for objects of class `popTimeExposure`:

```
R> plot(pt_object_strat)
```



We can also plot them side-by-side using the `ncol` argument:

```
R> plot(pt_object_strat, ncol = 2)
```



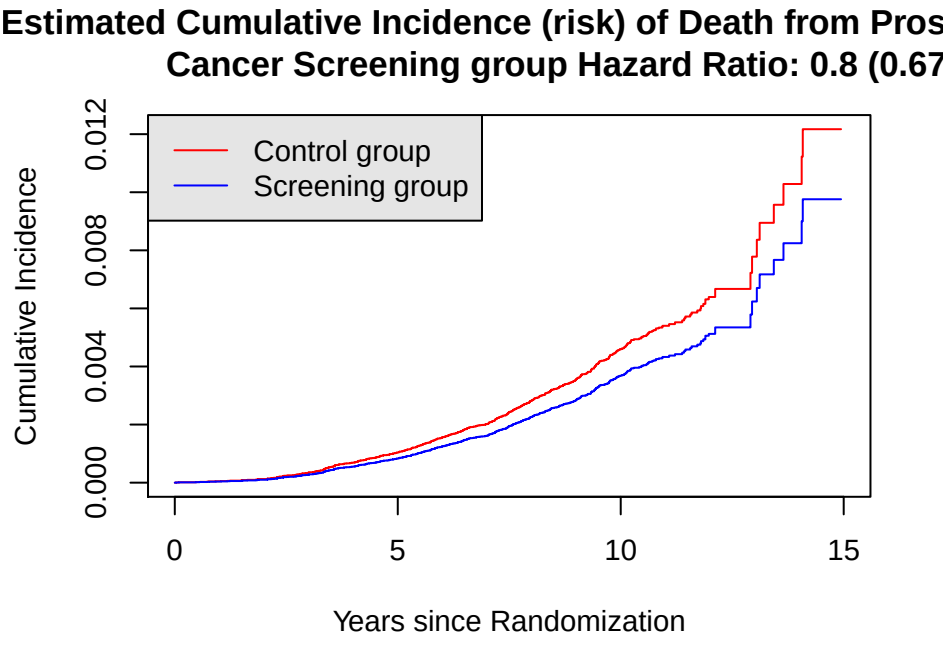
We first fit a Cox model, examine the hazard ratio for the screening group (relative to the control group), and plot the cumulative incidence function (CIF).

```
R> cox_model <- survival::coxph(Surv(Follow.Up.Time, DeadOfPrCa) ~ ScrArm,
R+                              data = ERSPC)
R> (sum_cox_model <- summary(cox_model))

Call:
survival::coxph(formula = Surv(Follow.Up.Time, DeadOfPrCa) ~
    ScrArm, data = ERSPC)

  n= 159893, number of events= 540

                        coef exp(coef) se(coef)     z Pr(>|z|)
ScrArmScreening group -0.222     0.801    0.088 -2.52    0.012 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                      exp(coef) exp(-coef) lower .95 upper .95
ScrArmScreening group     0.801       1.25     0.674     0.952

Concordance= 0.519  (se = 0.011 )
Rsquare= 0   (max possible= 0.075 )
Likelihood ratio test= 6.45  on 1 df,   p=0.0111
```

```
Wald test            = 6.37  on 1 df,   p=0.0116
Score (logrank) test = 6.39  on 1 df,   p=0.0115
```

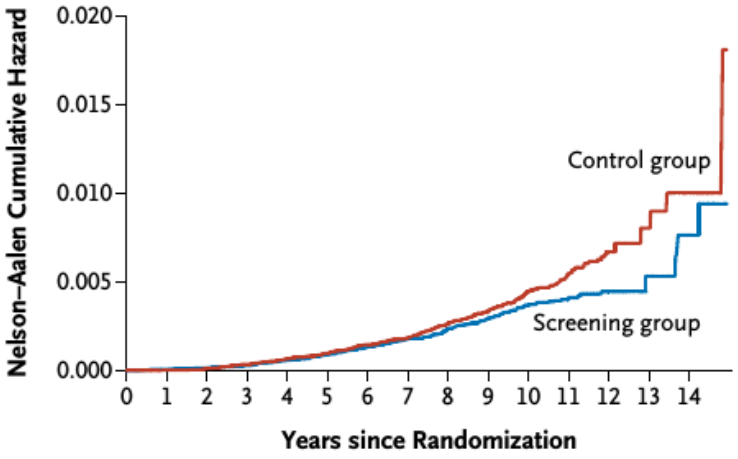
We can plot the CIF for each group:

```
R> new_data <- data.frame(ScrArm = c("Control group", "Screening group"),
R+                        ignore = 99)
R>
R> plot(survfit(cox_model, newdata = new_data),
R+      xlab = "Years since Randomization",
R+      ylab = "Cumulative Incidence",
R+      fun = "event",
R+      xlim = c(0,15), conf.int = F, col = c("red","blue"),
R+      main = sprintf("Estimated Cumulative Incidence (risk) of Death from Prostate
R+                     Cancer Screening group Hazard Ratio: %.2g (%.2g, %.2g)",
R+                     exp(coef(cox_model)),
R+                     exp(confint(cox_model))[1],
R+                     exp(confint(cox_model))[2]))
R> legend("topleft",
R+        legend = c("Control group", "Screening group"),
R+        col = c("red","blue"),
R+        lty = c(1, 1),
R+        bg = "gray90")
```



We compare it to the figure in Schroder FH, et al. N Engl J Med 2009 and see that the plots are very similar, as is the hazard ratio and 95% confidence interval:

Next we fit several models using case-base sampling. The models we fit differ in how we choose to model time.



No. at Risk

Screening group	65,078	58,902	20,288
Control group	80,101	73,534	23,758

Figure 2. Cumulative Risk of Death from Prostate Cancer.

As of December 31, 2006, with an average follow-up time of 8.8 years, there were 214 prostate-cancer deaths in the screening group and 326 in the control group. Deaths that were associated with interventions were categorized as being due to prostate cancer. The adjusted rate ratio for death from prostate cancer in the screening group was 0.80 (95% CI, 0.65 to 0.98; P=0.04). The Nelsen–Aalen method was used for the calculation of cumulative hazard.

Figure 2:

The `fitSmoothHazard` function provides an estimate of the hazard function $h(x,t)$ is the hazard function, t) denotes the numerical value (number of units) of a point in prognostic/prospective time and x is the realization of the vector X of variates based on the patient's profile and intervention (if any).

```
R> # set the seed for reproducible output
R> set.seed(1234)
R>
R> casebase_exponential <- casebase::fitSmoothHazard(DeadOfPrCa ~ ScrArm,
R+                                                   data = ERSPC,
R+                                                   ratio = 100)

'Follow.Up.Time' will be used as the time variable

R> summary(casebase_exponential)


Call:
glm(formula = formula, family = binomial, data = sampleData)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.458  -0.458  -0.409  -0.409   2.246

Coefficients:
                     Estimate Std. Error z value            Pr(>|z|)
(Intercept)           -7.7553     0.0584 -132.87 <0.0000000000000002 ***
ScrArmScreening group -0.2382     0.0921   -2.59              0.0097 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3619.1  on 5939  degrees of freedom
Residual deviance: 3612.3  on 5938  degrees of freedom
AIC: 3616

Number of Fisher Scoring iterations: 5

R> exp(coef(casebase_exponential)[2])

ScrArmScreening group
                 0.79

R> exp(confint(casebase_exponential)[2,])

Waiting for profiling to be done...
```

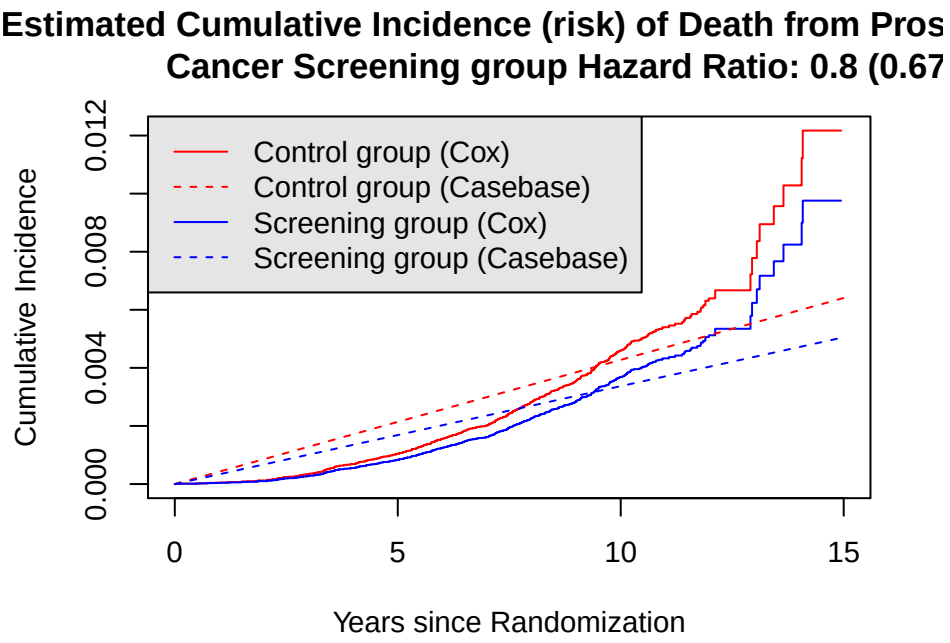
```
2.5 % 97.5 %
 0.66   0.94
```

The `absoluteRisk` function provides an estimate of the cumulative incidence curves for a specific risk profile using the following equation:

$$CI(x,t) = 1 - exp\left(-\int_0^t h(x,u)\mathrm{d}u\right)$$

In the plot below, we overlay the estimated CIF from the casebase exponential model on the Cox model CIF:

```
R> smooth_risk_exp <- casebase::absoluteRisk(object = casebase_exponential,
R+                                           time = seq(0,15,0.1),
R+                                           newdata = new_data)
R>
R> plot(survfit(cox_model, newdata = new_data),
R+      xlab = "Years since Randomization",
R+      ylab = "Cumulative Incidence",
R+      fun = "event",
R+      xlim = c(0,15), conf.int = F, col = c("red","blue"),
R+      main = sprintf("Estimated Cumulative Incidence (risk) of Death from Prostate
R+                      Cancer Screening group Hazard Ratio: %.2g (%.2g, %.2g)",
R+                      exp(coef(cox_model)),
R+                      exp(confint(cox_model))[1],
R+                      exp(confint(cox_model))[2]))
R> lines(smooth_risk_exp[,1], smooth_risk_exp[,2], col = "red", lty = 2)
R> lines(smooth_risk_exp[,1], smooth_risk_exp[,3], col = "blue", lty = 2)
R>
R>
R> legend("topleft",
R+        legend = c("Control group (Cox)","Control group (Casebase)",
R+                   "Screening group (Cox)", "Screening group (Casebase)"),
R+        col = c("red","red", "blue","blue"),
R+        lty = c(1, 2, 1, 2),
R+        bg = "gray90")
```



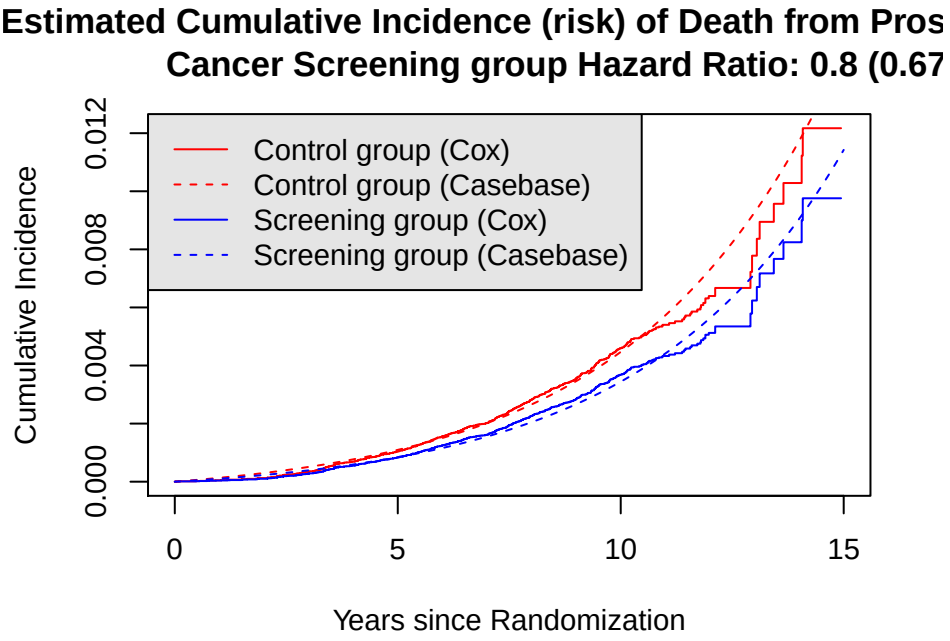
As we can see, the exponential model is not a good fit. Based on what we observed in the population time plot, where more events are observed later on in time, this poor fit is expected. A constant hazard model would overestimate the cumulative incidence earlier on in time, and underestimate it later on, which is what we see in the cumulative incidence plot. This example demonstrates the benefits of population time plots as an exploratory analysis tool.

Next we enter time linearly into the model:

```
R> casebase_time <- fitSmoothHazard(DeadOfPrCa ~ Follow.Up.Time + ScrArm,
R+                                  data = ERSPC,
R+                                  ratio = 100)

'Follow.Up.Time' will be used as the time variable

R> summary(casebase_time)


Call:
glm(formula = formula, family = binomial, data = sampleData)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.043  -0.487  -0.365  -0.275   2.744

Coefficients:
                    Estimate Std. Error z value            Pr(>|z|)
(Intercept)          -9.0421     0.1183  -76.45 <0.0000000000000002 ***
```

```
Follow.Up.Time          0.2254     0.0158   14.24 <0.0000000000000002 ***
ScrArmScreening group  -0.2611     0.0940   -2.78              0.0055 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3619.1  on 5939  degrees of freedom
Residual deviance: 3394.9  on 5937  degrees of freedom
AIC: 3401

Number of Fisher Scoring iterations: 5

R> exp(coef(casebase_time))

          (Intercept)        Follow.Up.Time ScrArmScreening group
              0.00012               1.25285               0.77023

R> exp(confint(casebase_time))

Waiting for profiling to be done...

                         2.5 %  97.5 %
(Intercept)           0.000093 0.00015
Follow.Up.Time        1.214840 1.29263
ScrArmScreening group 0.639934 0.92534

R> smooth_risk_time <- casebase::absoluteRisk(object = casebase_time,
R+                                            time = seq(0,15,0.1),
R+                                            newdata = new_data)
R>
R> plot(survfit(cox_model, newdata = new_data),
R+      xlab = "Years since Randomization",
R+      ylab = "Cumulative Incidence",
R+      fun = "event",
R+      xlim = c(0,15), conf.int = F, col = c("red","blue"),
R+      main = sprintf("Estimated Cumulative Incidence (risk) of Death from Prostate
R+                     Cancer Screening group Hazard Ratio: %.2g (%.2g, %.2g)",
R+                     exp(coef(cox_model)),
R+                     exp(confint(cox_model))[1],
R+                     exp(confint(cox_model))[2]))
R> lines(smooth_risk_time[,1], smooth_risk_time[,2], col = "red", lty = 2)
R> lines(smooth_risk_time[,1], smooth_risk_time[,3], col = "blue", lty = 2)
R>
R> legend("topleft",
R+        legend = c("Control group (Cox)","Control group (Casebase)",
```

```
R+                   "Screening group (Cox)", "Screening group (Casebase)"),
R+        col = c("red","red", "blue","blue"),
R+        lty = c(1, 2, 1, 2),
R+        bg = "gray90")
```



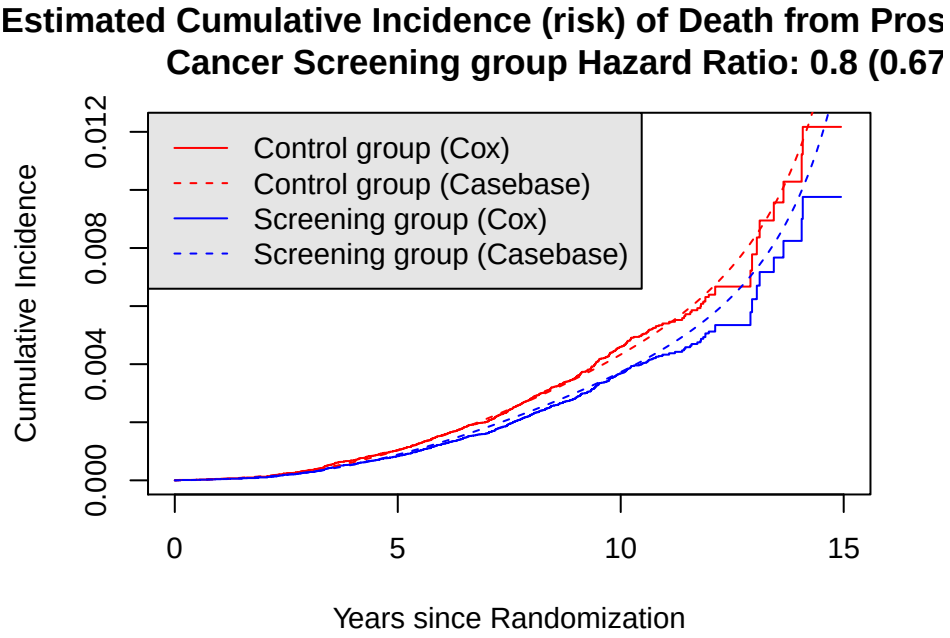
We see that the Weibull model leads to a better fit.

Next we try to enter a smooth function of time into the model using the `splines` package

```
R> casebase_splines <- fitSmoothHazard(DeadOfPrCa ~ bs(Follow.Up.Time) + ScrArm,
R+                                     data = ERSPC,
R+                                     ratio = 100)

'Follow.Up.Time' will be used as the time variable

R> summary(casebase_splines)


Call:
glm(formula = formula, family = binomial, data = sampleData)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.197  -0.522  -0.398  -0.212   3.187

Coefficients:
                     Estimate Std. Error z value             Pr(>|z|)
(Intercept)           -10.5045     0.3342  -31.43 < 0.0000000000000002 ***
```

```
bs(Follow.Up.Time)1     4.8348     0.8400    5.76     0.00000000862087 ***
bs(Follow.Up.Time)2     1.8367     0.4859    3.78              0.00016 ***
bs(Follow.Up.Time)3     5.0478     0.6998    7.21     0.00000000000055 ***
ScrArmScreening group  -0.1503     0.0941   -1.60              0.11013
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3619.1  on 5939  degrees of freedom
Residual deviance: 3356.5  on 5935  degrees of freedom
AIC: 3367

Number of Fisher Scoring iterations: 6

R> exp(coef(casebase_splines))

          (Intercept)   bs(Follow.Up.Time)1   bs(Follow.Up.Time)2
             0.000027            125.811140              6.275672
  bs(Follow.Up.Time)3 ScrArmScreening group
           155.673063              0.860457

R> exp(confint(casebase_splines))

Waiting for profiling to be done...

                          2.5 %     97.5 %
(Intercept)            0.000014   0.000051
bs(Follow.Up.Time)1   25.404219 687.891148
bs(Follow.Up.Time)2    2.425559  16.399777
bs(Follow.Up.Time)3   39.667300 626.079069
ScrArmScreening group  0.714859   1.033843

R> smooth_risk_splines <- absoluteRisk(object = casebase_splines,
R+                                     time = seq(0,15,0.1),
R+                                     newdata = new_data)
R>
R> plot(survfit(cox_model, newdata = new_data),
R+      xlab = "Years since Randomization",
R+      ylab = "Cumulative Incidence",
R+      fun = "event",
R+      xlim = c(0,15), conf.int = F, col = c("red","blue"),
R+      main = sprintf("Estimated Cumulative Incidence (risk) of Death from Prostate
R+                     Cancer Screening group Hazard Ratio: %.2g (%.2g, %.2g)",
R+                     exp(coef(cox_model)),
R+                     exp(confint(cox_model))[1],
```

```
R+                      exp(confint(cox_model))[2]))
R> lines(smooth_risk_splines[,1], smooth_risk_splines[,2], col = "red", lty = 2)
R> lines(smooth_risk_splines[,1], smooth_risk_splines[,3], col = "blue", lty = 2)
R>
R> legend("topleft",
R+        legend = c("Control group (Cox)","Control group (Casebase)",
R+                   "Screening group (Cox)", "Screening group (Casebase)"),
R+        col = c("red","red", "blue","blue"),
R+        lty = c(1, 2, 1, 2),
R+        bg = "gray90")
```



It looks like the best fit.

Since we are in the GLM framework, we can easily test for which model better fits the data using a Likelihood ratio test (LRT). The null hypothesis here is that the linear model is just as good as the larger (in terms of number of parameters) splines model.

```
R> anova(casebase_time, casebase_splines, test = "LRT")

Analysis of Deviance Table

Model 1: DeadOfPrCa ~ Follow.Up.Time + ScrArm + offset(offset)
Model 2: DeadOfPrCa ~ bs(Follow.Up.Time) + ScrArm + offset(offset)
  Resid. Df Resid. Dev Df Deviance     Pr(>Chi)
1      5937       3395
2      5935       3357  2     38.4 0.0000000046 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that splines model is the better fit.

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.

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.

In order to try and visualize the incidence density of relapse, we can look at the corresponding population-time plot. In Figure 3, failure times associated with relapse are highlighted on the

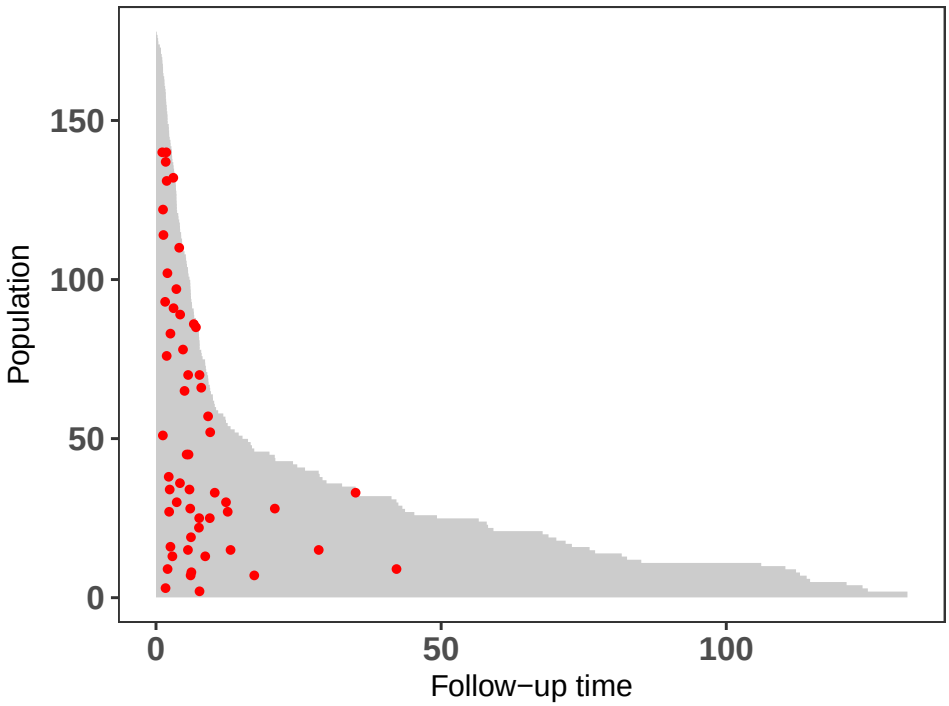


Figure 3: Population-time plot for the stem-cell transplant study. The points represent the event of interest (i.e., relapse).

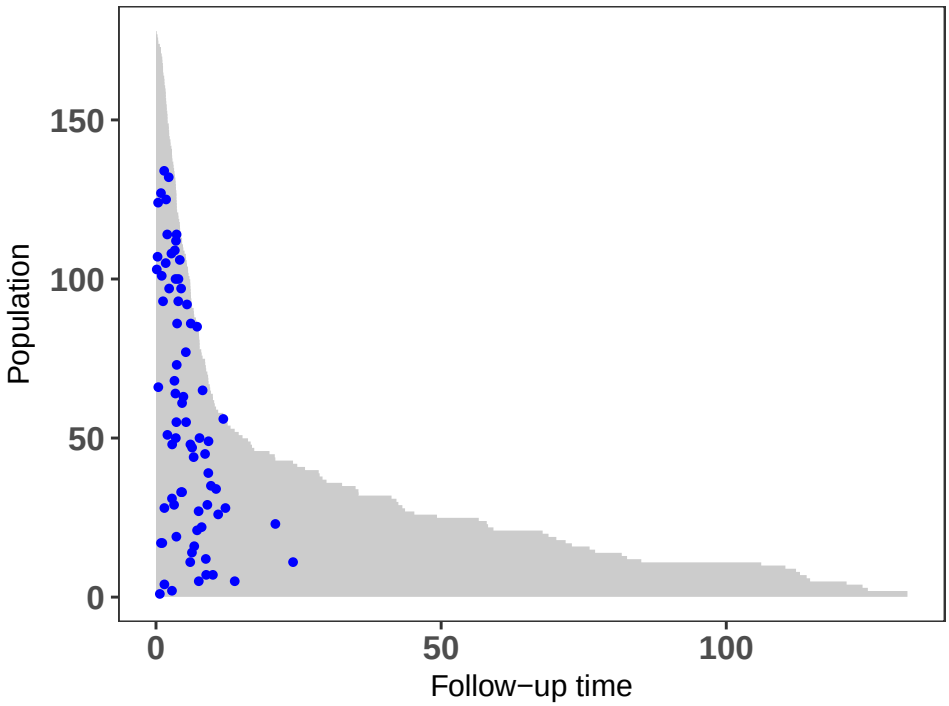


Figure 4: Population-time plot for the stem-cell transplant study. The points represent the competing events.

plot using red points, while Figure 4 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:

Covariates	HR	95% CI
Sex	0.78	(0.43, 1.41)
Disease	0.41	(0.22, 0.77)
Phase (CR2 vs. CR1)	1.27	(0.5, 3.23)
Phase (CR3 vs. CR1)	1.77	(0.44, 7.09)
Phase (Relapse vs. CR1)	5.17	(2.34, 11.43)
Source	1.87	(0.59, 5.88)
Age	1.00	(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 5 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"])),
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>
```

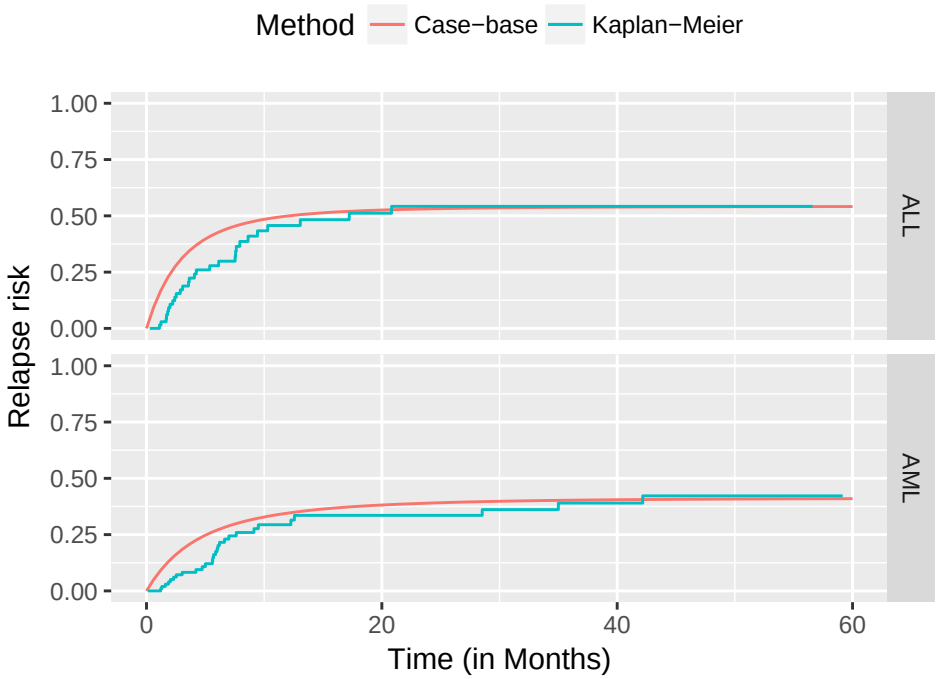


Figure 5: Absolute risk curve for a fixed covariate profile and the two disease groups. The estimate obtained from case-base sampling is compared to the Kaplan-Meier estimate.

```
R> # Estimate absolute risk curve
R> risk_cb <- absoluteRisk(object = model_cb, time = time_points,
R+                         method = "montecarlo", newdata = newdata)
```

7. Case study 3

- Glmnet, gam, and gbm

8. Discussion

- Discuss

References

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