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 University

Maxime Turgeon *
University of Manitoba

Jesse Islam
McGill University

James Hanley
McGill University

Olli Saarela
University of Toronto

Abstract

The abstract of the article. * joint co-authors

Keywords: keywords, not capitalized, Java.

1. Introduction

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. 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.

2. 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*.

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-homogeneous Poisson processes, we eventually reach an expression where each person-moment's contribution is of the form

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

where $N(t)$ is the counting process associated with the event of interest, $h(t)$ 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 $\log(h(t)) = g(t; X)$ 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 sampling *uniformly*, i.e. to sample the base series uniformly from the study base. In terms of Poisson processes, this sampling strategy corresponds essentially 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.

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

$$h(t; X) = \exp[g(t; X)]$$

where t denotes time and X, the individual's covariate profile. 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.

$$h(t; X) = \exp(\beta_0 + \beta_1 X)$$

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 in the study. This is also known as the *memoryless property* (Kalbfleisch and Prentice, 2002).

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

$$h(t; X) = \exp(\beta_0 + \beta_1 t + \beta_2 X)$$

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

$$h(t; X) = \exp(\beta_0 + \beta_1 \log(t) + \beta_2 X)$$

For competing-risk analyses with J possible events, we can show that each person-moment's contribution of the likelihood is of the form

$$\frac{h_j(t)^{dN_j(t)}}{\rho(t) + \sum_{j=1}^{J} h_j(t)},$$

where $N_j(t)$ is the counting process associated with the event of type j and $h_j(t)$ is the corresponding hazard function. As may be expected, this functional form is similar to the terms appearing in the likelihood function for multinomial regression.[1]

3. Existing packages

Survival analysis is an important branch of applied statistics and epidemiology. Accordingly, there is a vast ecosystem of `R` packages implementing different methods. In this section, we describe how the functionalities of **casebase** compare to these packages.

At the time of writing, a cursory examination of CRAN's task view on survival analysis reveals that there are over 250 packages related to survival analysis (2019). For the purposes of this article, we restricted our description to packages that implement at least one of the following features: parametric modeling, non-proportional hazard models, competing risk analysis, penalized estimation, and cumulative incidence curve estimation. By searching for appropriate keywords in the `DESCRIPTION` file of these packages, we found 60 relevant packages. These 60 packages were then manually examined to determine which ones are comparable to **casebase**. In particular, we excluded packages that were focused on a different set of problems, such as frailty and multi-state models. The remaining 14 packages appear in Table 1, along with a description of some of the functionalities they offer.

Several packages implement penalized estimation for the Cox model: **glmnet** (2011), **glmpath** (2018), **penalized** (2010), **RiskRegression** (2019). Moreover, some packages also include penalized estimation in the context of Cox models with time-varying coefficients: **CoxRidge** (2015), **rstpm2** (2019) and **survival** (2015). On the other hand, **casebase** provides penalized estimation in the context of parametric hazards. To our knowledge, this is the only package to offer this functionality. Is this true? Because it seems to be a logical conclusion from your paragraph.

Parametric survival models are implemented in a handful of packages: **CFC** (2019), **flexsurv** (2016), **SmoothHazard** (2017), **rsptm2** (2019), **mets** (2014), and **survival**. The types of models they allow vary for each package. For example, **SmoothHazard** is limited to Weibull distributions (2017), whereas both **flexsurv** and **survival** allow users to supply any distribution of their choice. Also, **flexsurv**, **smoothhazard**, **mets** and **rstpm2** also have the ability to model the effect of time using splines, which allows flexible modeling of the hazard function. Moreover,

[1] Specifically, it corresponds to the following parametrization:

$$\log\left(\frac{P(Y = j \mid X)}{P(Y = J \mid X)}\right) = X^T \beta_j, \qquad j = 1, \ldots, J-1$$

flexsurv has the ability to estimate both scale and shape parameters for a variety of parametric families (see Table **??**). As discussed above, **casebase** can model any parametric family whose log-hazard can be expressed as a linear model of covariates (including time). Therefore, our package allows the user to model the effect of time using splines, and through interaction terms involving covariates and time, it also allows user to fit time-varying coefficient models. However, we do not explicitly model any shape parameter, unlike **flexsurv**.

Of the methods mentioned so far, only **CFC**, **flexsurv**, **mets** and **survival** contain implementations for competing risks. The differentiating factor between these packages and **casebase** is that **casebase** handles competing risks through multiple logistic regression. There's a literature review of packages for competing-risk analysis in the paper for **CFC**. Could you incorporate some of their discussion, with proper citations? The biggest omission right now is **cmprsk**, which we actually use below.

Finally, several packages include functions to estimate the cumulative incidence function. The corresponding methods generally fall into two categories: transformation of the estimated hazard function, and semi-parametric estimation of the baseline hazard. The first category broadly corresponds to parametric survival models, where the full hazard is explictly modeled. Using this estimate, the survival function and the cumulative incidence function can be obtained using their functional relationships (see Equations 1 and 2 below). Packages including this functionality include **CFC**, **Flexsurv**, **mets**, and **survival**. Our package **casebase** also follows this approach for both single-event and competing-event analyses. The second category outlined above broadly corresponds to Cox models. These models do not model the full hazard function, and therefore the baseline hazard needs to be estimated separately in order to estimate the survival function. This is achieved using semi-parametric estimators (e.g. Breslow's estimator). Packages that implement this approach include **RiskRegression**, **rstpm2**, and **survival**. As mentioned in the introduction, a key distinguishing factor between these two approaches is that the first category leads to smooth estimates of the cumulative incidence function, whereas the second category produces estimates in the form of step-wise functions. This was one of the main motivations for introducting case-base sampling in survival analysis.

Package	Competing risks	Non-proportional	Penalization	Splines	Parametric	Semi-parametric	Interval/left censoring	Absolute risk
casebasee Hanley and Miettinen (2009)	x	x	x	x	x			x
CFC Sharabiani and Mahani (2019)	x	x			x			x
coxRidge Perperoglou (2015)		x	x			x		
crrp Fu (2015)	x		x					
fastcox Yi and Zou (2017)			x			x		
Flexrsurv Clerc-Urmès, Grzebyk, and Hédelin (2017)		x		x	x			x
Flexsurv Jackson (2016)	x	x		x	x			x
glmnet Simon *et al.* (2011)			x			x		
glmpath Park and Hastie (2018)			x			x		
mets Scheike *et al.* (2014)	x			x		x		x
penalized Goeman, Meijer, Chaturvedi, and Lueder (2019)			x			x		
RiskRegression Gerds *et al.* (2019)			x			x		x
Rstpm2 Clements *et al.* (2019)		x	x	x	x	x	x	x
SmoothHazard Touraine *et al.* (2017)		x		x	x		x	
Survival Therneau (2015)	x	x			x	x	x	x

Table 1: Different features of interest in various survival packages.

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) parametric modeling. We explicitly 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 unit tests.

4.1. Population-time plots

4.2. Parametric modeling

The parametric modeling step was separated into three parts:

1. case-base sampling;
2. estimation of the smooth hazard function;
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.

To fit the hazard function, we provide several approaches that are available via the `family` parameter. These approaches are:

- `glm`: This is the familiar logistic regression.
- `glmnet`: This option allows for variable selection using Lasso or elastic-net. This functionality is provided through the `glmnet` package (Friedman, Hastie, and Tibshirani 2010).

- `gam`: This option provides support for *Generalized Additive Models* via the `gam` package (Hastie and Tibshirani 1987).
- `gbm`: This option provides support for *Gradient Boosted Trees* via the `gbm` package. This feature is still experimental.

In the case of multiple events, the hazard is fitted via multinomial regression as performed by 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

$$S(t) = \exp\left(-\int_0^t h(u;X)du\right). \tag{1}$$

The integral is computed using either the `stats::integrate` function or Monte-Carlo integration. The risk function (or cumulative incidence function) is then defined as

$$CI(t) = 1 - S(t). \tag{2}$$

For the case of a competing-event analysis, the event-specific risk is computed using the following procedure: first, we compute the overall survival function (i.e. for all event types):

$$S(t) = \exp\left(-\int_0^t H(u;X)du\right), \qquad H(t;X) = \sum_{j=1}^{J} h_j(t;X).$$

From this, we can derive the event-specific subdensities:

$$f_j(t) = h_j(t)S(t).$$

Finally, by integrating these subdensities, we obtain the event-specific cumulative incidence functions:

$$CI_j(t) = \int_0^t f_j(u)du.$$

We created `absoluteRisk` as an `S3` generic, with methods for the different types of outputs of `fitSmoothHazard`. The method dispatch system of R then takes care of matching the correct output to the correct methodology for calculating the cumulative incidence function, without the user's intervention.

In the following sections, we illustrate these functionalities in the context of three case studies.

5. Case study 1–European Randomized Study of Prostate Cancer Screening

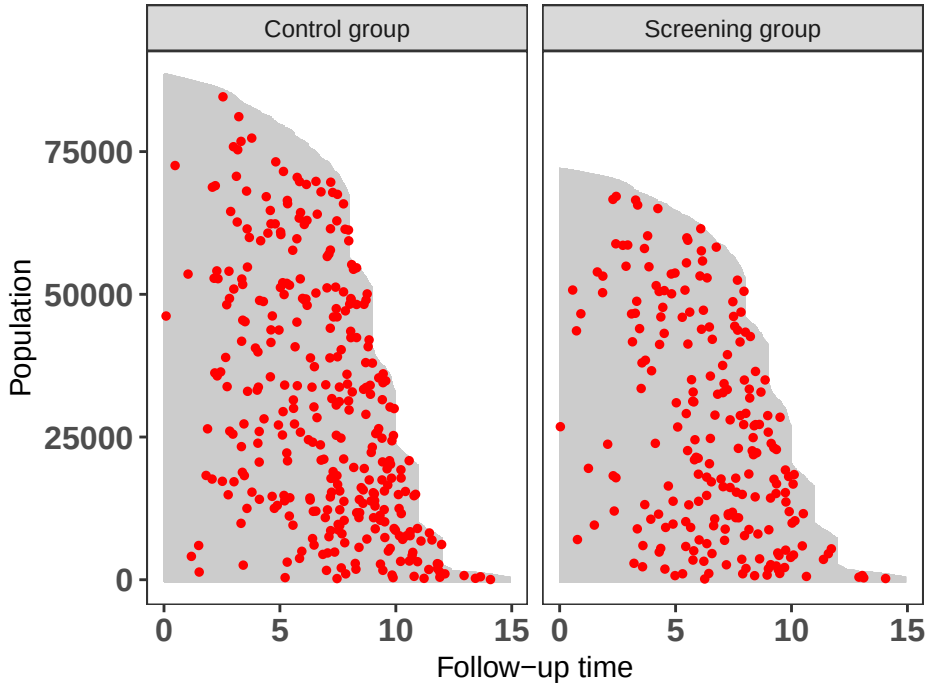
For the first case study, we will use of the European Randomized Study of Prostate Cancer Screening data. This dataset is available through the **casebase** package:

The results of this study were published by (Schröder, Hugosson, Roobol, Tammela, Ciatto, Nelen, Kwiatkowski, Lujan, Lilja, Zappa *et al.* 2009), and the dataset itself was obtained using the approach described in (Liu, Rich, and Hanley 2014).

Population time plots can be extremely informative graphical displays of survival data. They should be the first step in an exploratory data analysis. We facilitate this task in the **casebase** package using the `popTime` function. We first create the necessary dataset for producing the population time plots, and we can generate the plot by using the corresponding `plot` method:

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

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



ADD PARAGRAPH ABOUT WHAT WE CAN CONCLUDE FROM THESE GRAPHS.

Next, we investigate the differences between the control and the screening arms. A common choice for this type of analysis is to use a Cox regression model and estimate the hazard ratio for the screening group (relative to the control group). In R, it can be done as follows:

```
#> 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 *
#> ---
```

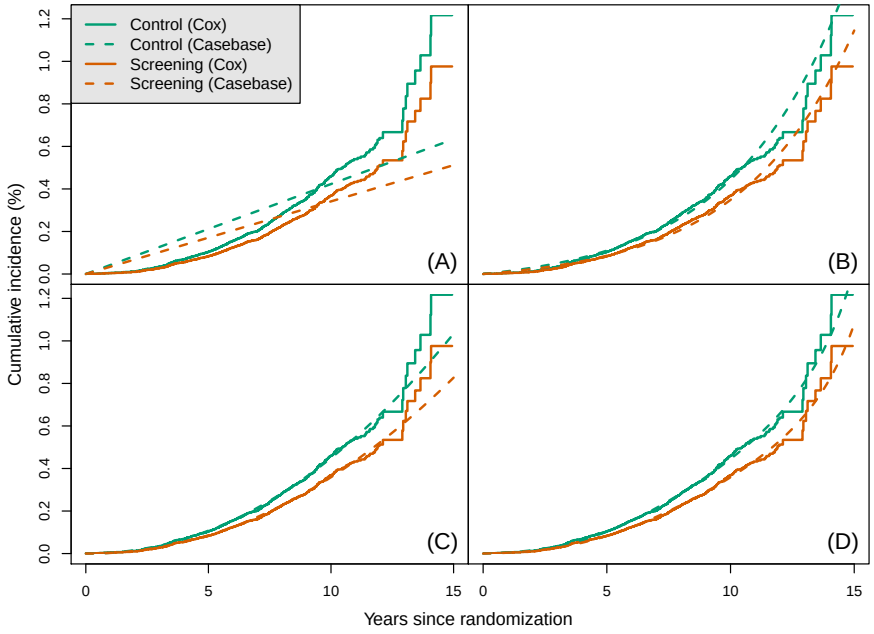
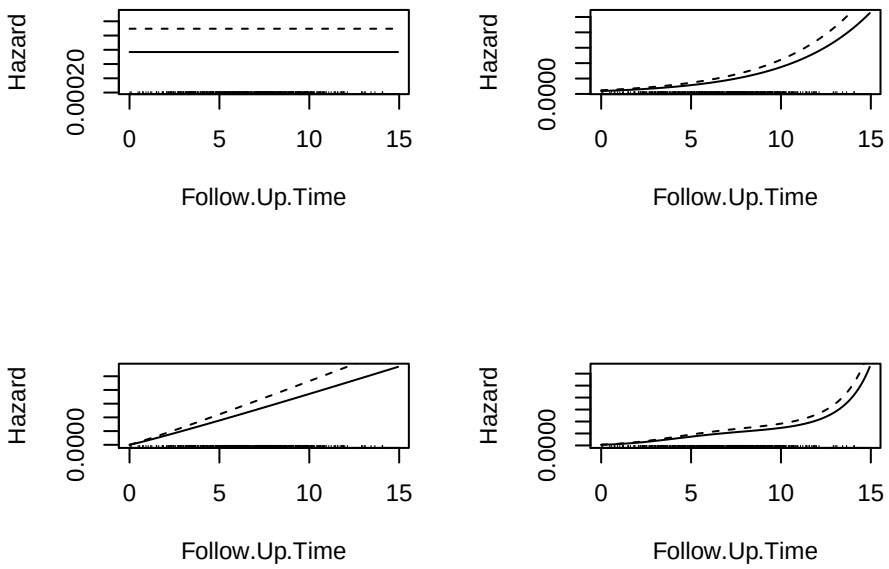


Figure 1: Cumulative incidence functions for control and screening groups in the ERSPC data. In each of the panels, we plot the CIF from the Cox model (solid line) and the CIF from the caseabse sampling scheme (dashed line) with different functional forms of time in a logistic regression. (A) The time variable is excluded so that the hazard function is constant over time which is equivalent to the one-parameter exponential distribution. (B) Linear function of time which is equivalent to a Gompertz distribution. (C) The natural logarithm of time which is equivalent to a Weibull distribution. (D) Cubic b-spline expansion of time.

```
#> 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 )
#> Likelihood ratio test= 6.45  on 1 df,   p=0.01
#> Wald test            = 6.37  on 1 df,   p=0.01
#> Score (logrank) test = 6.39  on 1 df,   p=0.01
```

We can also plot the cumulative incidence function (CIF) for each group. However, this involves estimating the baseline hazard, which Cox regression treated as a nuisance parameter.



5.1. Equivalence between casebase and AFT

Next, we show how case-base sampling can be used to carry out a similar analysis. We fit several models that differ in how we choose to model time.

First, we will fit an exponential model. Recall that this corresponds to excluding time from the linear predictor.

We can then use the `absoluteRisk` function to get an estimate of the cumulative incidence curve for a specific covariate profile. In the plot below, we overlay the estimated CIF from the exponential model on the Cox model CIF:

As we can see, the exponential model gives us an estimate of the hazard ratio that is similar to the one obtained using Cox regression. However, the CIF estimates are quite different. Based on what we observed in the population time plot, where more events are observed later on in time, we do not expect a constant hazard would be a good description for this data. In other words, this poor fit for the exponential hazard is expected. A constant hazard model would overestimate the cumulative incidence earlier on in time, and underestimate it later on; this is what we see on the cumulative incidence plot. This example demonstrates the benefits of population time plots as an exploratory analysis tool.

For the next model, we include time as a linear term. Recall that this corresponds to a Weibull model.

Again, the estimate of the hazard ratio is similar to that obtained from Cox regression. We then look at the estimate of the CIF:

We see that the Weibull model leads to a better fit of the CIF than the Exponential model, when compared with the semi-parametric approach.

Finally, we can model the hazard as a smooth function of time using the `splines` package:

Table 2: Mean

Model	casebase	survreg
Exponential	0.81 (0.68, 0.96)	0.81 (0.68, 0.96)
Gompertz	0.78 (0.66, 0.93)	0.80 (0.67, 0.95)
Weibull	0.80 (0.67, 0.95)	0.80 (0.65, 0.96)
Splines	0.81 (0.68, 0.96)	–

Note:
Median (Inter-quartile range) is given for Model Size.

Once again, we can see that the estimate of the hazard ratio is similar to that from the Cox regression. We then look at the estimate of the CIF:

Qualitatively, the combination of splines and case-base sampling produces the parametric model that most closely resembles the Cox model. In the following table, we see that the confidence intervals are also similar across all four models. In other words, this reinforces the idea that, under proportional hazards, we do not need to model the full hazard to obtain reliable estimates of the hazard ratio. Of course, different parametric models for the hazards give rise to qualitatively different estimates for the CIF.

As noted above, the usual asymptotic results hold for likelihood ratio tests built using case-base sampling models. Therefore, we can easily test the null hypothesis that the exponential model is just as good as the larger (in terms of number of parameters) splines model.

References

Allignol A, Latouche A (2019). "CRAN Task View: Survival Analysis." URL https://cran.r-project.org/web/views/Survival.html.

Clements M, Liu XR, Lambert P, Jakobsen LH, Gasparini A, Smyth G, Alken P, Wood S, Ulerich R (2019). "Smooth Survival Models, Including Generalized Survival Models [R package rstpm2 version 1.5.1]." URL https://cran.r-project.org/web/packages/rstpm2/index.html.

Clerc-Urmès I, Grzebyk M, Hédelin G (2017). *flexrsurv: An R package for relative survival analysis*. R package version 1.4.1, URL https://CRAN.R-project.org/package=flexrsurv.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1). ISSN 1548-7660. doi:10.18637/jss.v033.i01.

Fu Z (2015). "Package crrp." URL https://cran.r-project.org/web/packages/crrp/index.html.

Gerds TA, Blanche P, Mortersen R, Tollenaar N, Mogensen UB, Ozenne B (2019). "Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks [R package riskRegression version 2019.11.03]." URL https://CRAN.R-project.org/package=riskRegression.

Goeman J, Meijer R, Chaturvedi N, Lueder M (2019). "Package penalized." URL https://cran.r-project.org/web/packages/penalized/index.html.

Goeman JJ (2010). "L1 penalized estimation in the Cox proportional hazards model." *Biometrical Journal*, (52), –14.

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

Hastie T, Tibshirani R (1987). "Generalized additive models: some applications." *Journal of the American Statistical Association*, **82**(398), 371–386.

Jackson C (2016). "flexsurv: A Platform for Parametric Survival Modeling in R." *Journal of Statistical Software*, **70**(8), 1–33. doi:10.18637/jss.v070.i08.

Liu Z, Rich B, Hanley JA (2014). "Recovering the raw data behind a non-parametric survival curve." *Systematic reviews*, **3**(1), 151.

Park MY, Hastie T (2018). "Package glmpath." URL https://CRAN.R-project.org/package=glmpath.

Perperoglou A (2015). "Package CoxRidge." URL https://CRAN.R-project.org/package=CoxRidge.

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.

Scheike TH, Holst KK, Hjelmborg JB (2014). "Estimating twin concordance for bivariate competing risks twin data." *Statistics in medicine*, **33**(7), 1193–1204.

Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, *et al.* (2009). "Screening and prostate-cancer mortality in a randomized European study." *New England Journal of Medicine*, **360**(13), 1320–1328.

Sharabiani MT, Mahani AS (2019). "Package CFC." URL https://cran.r-project.org/web/packages/CFC/index.html.

Simon N, Friedman J, Hastie T, Tibshirani R (2011). "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software*, **39**(5), 1–13. URL http://www.jstatsoft.org/v39/i05/.

Therneau TM (2015). *A Package for Survival Analysis in S*. Version 2.38, URL https://CRAN.R-project.org/package=survival.

Touraine C, Gerds TA, Joly P (2017). "SmoothHazard: An R Package for Fitting Regression Models to Interval-Censored Observations of Illness-Death Models." *Journal of Statistical Software*, **79**(7), 1–22. doi:10.18637/jss.v079.i07.

Yi Y, Zou H (2017). "Package fastcox." URL https://cran.r-project.org/web/packages/fastcox/index.html.

Affiliation:

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

Maxime Turgeon *
University of Manitoba
186 Dysart Road Winnipeg, MB, Canada R3T 2N2
E-mail: max.turgeon@umanitoba.ca
URL: https://maxturgeon.ca/

Jesse Islam
McGill University
1020 Pine Avenue West Montreal, QC, Canada H3A 1A2
E-mail: jesse.islam@mail.mcgill.ca

James Hanley
McGill University
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