

casebase: An Alternative Framework For Survival Analysis and Comparison of Event Rates by Sahir Rai Bhatnagar*, Maxime Turgeon*, Jesse Islam, James A Hanley and Olli Saarela. (*joint co-authors)

Response to the reviewers¹

We thank the reviewers for their constructive comments, which we believe has significantly improved our manuscript. In this document, we reproduce the reviewers comments, and provide our response to each of them below. For a quick reference, here are the main changes to the manuscript:

- We re-wrote the entire introduction to focus on the motivation for our package.
- We clarified the “Theoretical Details” section by expanding our discussion of person moments, case series, and base series.
- We refocused the first Case Study to highlight what we believe is the strength of the case-base sampling framework, which is flexible hazard modelling (e.g. non-linear time effect, non-proportional hazard) within a single framework.
- We added a section on how we can readily compute confidence bands of various estimators derived from the hazard function.
- We clarified the use of “cumulative incidence” throughout and restricted it to the competing-risk setting.
- We added a discussion of Poisson regression for survival analysis, and how it differs from case-base sampling, to our manuscript.

Reviewer 1

A primary critique is that after reading pages 1 and 2, I still had no idea what this package is doing nor a good feel for why it needs to exist. This is coming from someone who has worked in survival analysis for over 3 decades. This is not meant to be mean: if we assume that the package is worthwhile, and you want people to use it, this needs to be addressed up front. Let me break down the issue into more practical advice.

Reviewer Point P 1.1 — The introduction put me off with things that were incorrect and/or “fluff”. Overdone salesman pitches sometimes show up in the innovation and significance sections of grant applications, but even there they tend to decrease rather than increase the reader’s enthusiasm.

- a. “... stepwise estimates of the survival function that can be difficult to interpret”. Since the entire research world seems to handle Kaplan-Meier curves just fine, this is a statment that has no face validity. I’ve never heard that complaint from an actual user. Actually, the KM may sometimes be preferred to smooth estimates, since the ‘bumps’ provide a visual estimate

¹November 17, 2021

of precision; when there are multiple curves explicit confidence bands often make a plot too busy.

- b. "... opens the door to an extensive array of modeling tools. Indeed, lasso and elastic-net regression can be used..." Both of these apply directly to survival data/ Cox models. Read the help page for `glmnet`.

Turn down the rhetoric.

Reply: We have substantially re-written the introduction to focus on the motivation for our package. Overall, the ultimate goal of our package is to make fitting flexible hazards accessible to more end users who favour reporting absolute risks over hazard ratios (this is also mentioned in your comment 1.3 below). We frame our package as an alternative framework for survival analysis. More specifically, case-base sampling is a parametric approach which directly models the hazard function in continuous time, using the well-understood and familiar logistic regression. The asymptotic normality of the coefficient estimates allow for an efficient way to estimate the variance of various estimators derived from the hazard function. It also allows one to model flexible functions of time and their interactions with covariates. For example, in our first case study, we fit the interaction between `survival::psplines(time)` and a binary treatment variable. The following code shows how we can fit the hazard function using the formula interface:

```
fit <- fitSmoothHazard(DeadOfPrCa ~ pspline(Follow.Up.Time, df = 2) * ScrArm,
                      data = ERSPC)
```

From the fitted object, we can readily obtain the hazard function by treatment group, the time-dependent hazard ratio, and the absolute risk with standard errors using `plot`, `absoluteRisk`, and `confint` methods:

```
# hazard function for each arm
plot(fit, type = "hazard", hazard.params = list(xvar = "Follow.Up.Time", by = "ScrArm"))

# hazard ratio
new_data <- data.frame(ScrArm = factor("Control group",
                                       levels = c("Control group", "Screening group")),
                      Follow.Up.Time = seq(1, 12, by = 0.1))

plot(fit, type = "hr", newdata = new_data, var = "ScrArm", xvar = "Follow.Up.Time")

# absolute risk and confidence intervals
new_data <- data.frame(ScrArm = c("Control group", "Screening group"))
new_time <- seq(0, 14, 0.1)
risk <- absoluteRisk(fit, time = new_time, newdata = new_data)
conf_ints <- confint(risk, fit)
```

An experienced user can likely do this with existing packages, but even so, it is not so straightforward to obtain these quantities with their standard errors. Related to your comment 1.4 below,

we contrast the Whitehead (1980) approach with the casebase approach and show that Poisson regression can lead to some computational issues when the censoring fraction is large and the events are sparse.

As for the second point, a similar question was raised by the other reviewer (Reviewer Point 2.5). We have clarified the relationship between casebase and `cv.glmnet` within the manuscript. In Example 3, we are comparing two approaches: case-base sampling, where the hazard is estimated using penalized logistic regression; and Coxnet, which is a regularized version of the Cox model. As in Examples 1 and 2, the function `fitSmoothHazard` starts by sampling the case series and base series and calculating the offset term. With `family = "glmnet"`, the data is then transformed to match the expected matrix input of `cv.glmnet`, before calling `cv.glmnet` with the offset term. In other words, `fitSmoothHazard` performs most of the necessary data processing. If a user fits a model using `cv.glmnet` and the Cox family, then they are fitting a semi-parametric model that is different from the case-base model.

Reviewer Point P 1.2 — The package is based on a 2009 paper by Hanley and Miettinen, and here lies one of the issues. It is difficult to argue with the strong impact OM has had on the epidemiologic literature, for the good, but a significant downside to his work is the use of an alternate and unique vocabulary. For instance, all of us are familiar with the cumulative distribution function $F(t) = Pr(X \leq t)$ and its complement the survival curve $S(t) = 1 - F(t)$. In the OM world $F(t)$ is now the “cumulative incidence function” $CI(t)$; a completely unnecessary substitution. (Also a very confusing one, since the CIF is not the integral of the incidence). The words “case”, “base”, “person-moments” are likewise something quite peculiar. The overall title of casebase will possibly be confused with “case only” study designs in genetics (someone reads it as “case based”). If you want people to know what’s up you will need to provide a translation service.

I took the time to read the 2009 paper, and I still don’t know what this sentence means (page 2 of submission) “with the person-moments where $dR_i(t) = 1$ constituting the base series.” What is R counting? When would dR be zero? Nor is ρ or Q clear to me. The author’s need to give actual, clear definitions.

Reply: We clarified the use of the term “cumulative incidence”, restricting it to the competing risk setting. We also expanded our introduction of the terminology (including person moments, case series, and base series) at the beginning of the section Theoretical Details to provide more detail and clarity.

To summarize, there are two main counting processes: the “usual” one $N(t)$, which counts the events of interest and whose hazard we want to model; and $R(t)$, which arises from the case-base sampling process. These broadly correspond to “cases” and “controls” in case-control sampling. Then $Q(t)$ is simply their sum. The purpose of the process $R(t)$ is to allow us to write down a (partial) likelihood that can be maximized. We are **not** interested in modelling the hazard $\rho(t)$ for the process $R(t)$, and in fact it is completely user-defined (and so there is no point in trying to model it). Currently, our package only allows one sampling mechanism, corresponding to $R(t)$ being a homogeneous Poisson process over the whole study base; however, as pointed out in the discussion, future work will look at providing a user-interface for specifying the case-base sampling mechanism.

Reviewer Point P 1.3 —

The H and M paper includes a valid justification, which is missing here. That is that the users of Cox models rarely provide the necessary information to compute absolute risk. This is a major gap in reporting; absolute risk is as important as relative risk, maybe more so in fact. There are 2 solutions. i. Train users to include that information (it’s an extra computing step for a Cox model,

but not difficult) or ii. use an alternative model where the baseline hazard has a simple parametric form (and train authors to report those coefficients).

Aside: Short of the journals making it a requirement, I have low expectations for any "train users" plan. Page limit are often so severe that any sentence which can be omitted is leapt upon like a press gang seizing a drunk. (This particular personal pessimism should not count against the paper.)

Reply: We agree and have addressed this issue up front in our introduction, as well as our response to your first comment 1.1. One step towards this goal would be to give users tools that allow them to easily compute and visualize the hazard function, time-dependent hazard ratios, and absolute risks. Our first case study highlights these features of the casebase framework. We would like to re-iterate however, that we are not saying this is a better approach than other frameworks (e.g. Cox model), but one that is perhaps more familiar and within reach of a junior level data analyst.

Reviewer Point P 1.4 — On a statistical front, there already exist simple methods for substituting in a smooth hazard. See Whitehead (Applied Statistics 1980, 268-275) for instance: A quick look at the KM normally allows one to break it into 3-4 segments over which $\log(S(t))$ is approximately linear. Split the follow-up time into epochs based on those cutpoints (`survSplit` will do this), and fit simple Poisson regression to the result with one intercept per epoch. One can even use many splits and model time as a spline, but this conflicts with the "simple reporting" goal. In what ways is this approach better? (There may be several, I don't know.)

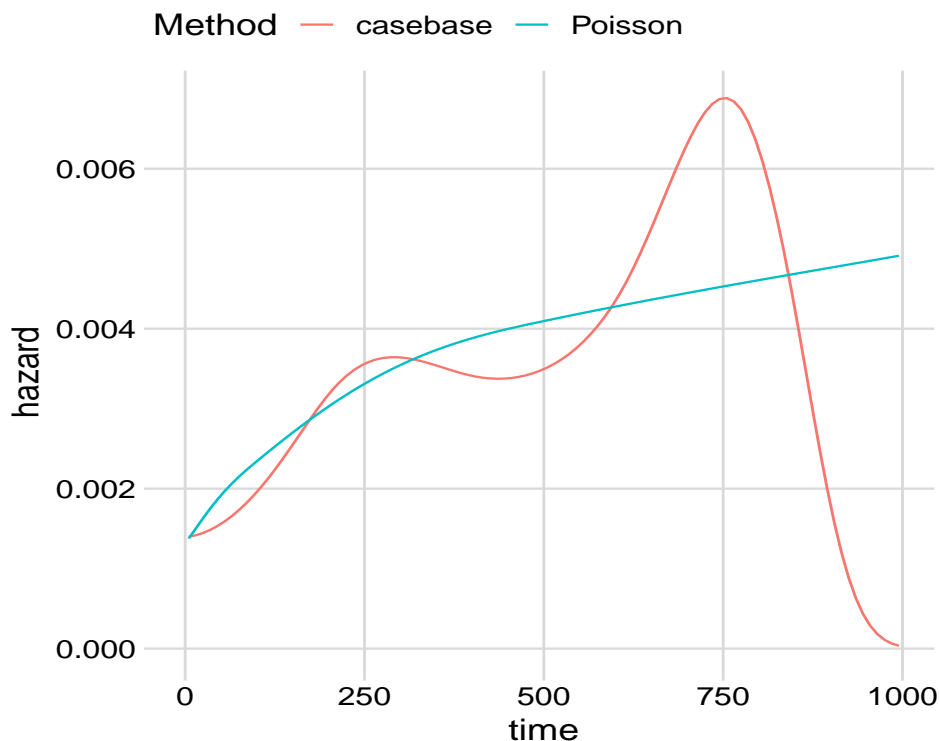
Reply: We added a discussion of Poisson regression for survival analysis, and how it differs from case-base sampling, to our manuscript. Case-base sampling differs from what the reviewer describes in two ways: it assumes a continuous-time model (unlike Poisson regression), and it does not rely on the user making choices about cut points (unlike using `survSplit`).

In practice, we find that Poisson regression leads to issues when the censoring fraction is large and events are rare (e.g. the ERSPC dataset available with our package). For illustration, we provide a comparison of these two approaches here. We follow the [online textbook by Carstensen](#) for fitting spline-based Poisson models. The code to reproduce Figure R1 below is available in a public gist ([link here](#)). In Figure R1, we are plotting the hazard as estimated using case-base sampling with splines (using `pspline`), and a splines-based Poisson regression model on the lung dataset. The hazard function is estimated for a 60-year-old male patient. Unfortunately, we couldn't produce a similar Figure with the ERSPC dataset, since the estimates obtained using Poisson regression were off by several orders of magnitude compared to case-base sampling.

In Table R1, we present the hazard ratio estimates for both datasets and three different approaches: Poisson regression with splines, Cox regression, and case-base sampling with splines. For the lung dataset, we can see that all three methods provide nearly identical hazard ratios and confidence intervals. However, for the ERSPC dataset, we can see that the estimates obtained from the splines-based Poisson model are very different from those obtained using Cox regression or case-base sampling.

We attempted to diagnose this issue, but we didn't reach any satisfying conclusions. Changing the number of time bins led to different estimates of the hazard ratio, but none were similar to the case-base sampling estimates. We obtained similar discrepancies when we artificially reduced the censoring fraction of the ERSPC dataset. As Poisson regression is not the focus on this manuscript, we didn't pursue this investigation further.

Figure R1: Estimates of the hazard function estimate from the lung dataset in the `survival` package. The red line corresponds to a case-base approach with splines; the blue line corresponds to a Poisson approach using splines.



	Lung				ERSPC	
	HR (Age)	95% CI	HR (Sex)	95% CI	HR (ScrArm)	95% CI
Poisson	1.016	(0.998, 1.035)	0.600	(0.438, 0.844)	0.999	(0.996, 1.002)
Cox	1.017	(0.999, 1.036)	0.599	(0.431, 0.831)	0.801	(0.674, 0.952)
case-base	1.016	(1.000, 1.037)	0.604	(0.431, 0.835)	0.800	(0.656, 0.929)

Table R1: Hazard ratio estimates and 95% confidence interval for two datasets. We compare Poisson regression using splines, Cox regression, and case-base sampling using splines. The grey cells highlight the discrepancy between Poisson regression and the other two methods. **Lung**: lung dataset from the `survival` package. **ERSPC**: ERSPC dataset from the `casebase` package.

Reviewer Point P 1.5 — In the comparisons with other packages, the authors overlooked that the standard `survival::survfit` function provides Aalen-Johansen estimates; both the KM and competing risks are special cases of the AJ.

More importantly, the competing risks literature defines "cumulative incidence" as something quite different than Hanley and Miettinen's definition (equation 2 of 2009 paper.) If casebase is implementing H&M, then this whole discussion of the CI is very confusing.

Reply: As noted above, we clarified throughout the use of the term "cumulative incidence", restricting it to the competing-risk setting. We also included Cox regression and the Aalen-Johansen estimator to

our comparison of cumulative incidence estimates (Case Study #2).

Reviewer Point P 1.6 — Page 5. "The case based approach described in section 2.2 can be visualized as ..." There is no section 2.2. (Other than figures and tables there are no numbers in this reviewer's copy at all.)

Reply: We removed the mention of a numbered section.

Reviewer Point P 1.7 — Page 6. I am very confused by this plot. In the ERSPC data there is no date of enrollment, and a subject's follow-up times ends at prostate cancer (PCA) death, other death, or last follow-up. How then can a data point for a case be within the gray area. For example, died of PCA at 5 years, but the gray for this person extends out to 10. Labeling a subject as a case before the actual occurrence of PCA death is an example of immortal time bias, a source of many false inference schemes.

Reply: We updated Figure 1 in the manuscript to explain the process of constructing a population-time plot. In particular, we discuss why some red points appear within the grey area.

A population-time plot helps to visualize incidence density. The grey area can be thought of as infinitely thin horizontal rectangles ordered by length of followup (Figure 1 A). Then, we label each event using a red point; as the reviewer points out, these events occur at the end of follow-up, so at the end of a grey horizontal line (Figure 1 B). Visually, these labelled points along the edge of the grey space are not visually distinct enough to demonstrate incidence density. To help the visualization, these points are randomly moved along their vertical axis, resulting in these points appearing in the grey space (Figure 1 C). Finally, we plot a sample of the base series, uniformly from the entire grey space (Figure 1 D) creating the complete population time plot. We have updated our manuscript to include this information.

Reviewer 2

This article provides an overview of the `casebase` package and a comparison with other survival packages. The package's main goal is to analyze survival data allowing users to estimate smooth baseline hazards over time. Their claim is that these results are easier to interpret. The methods for this approach have been published previously and are sound. Whether this approach is useful in practice is less clear based on the examples shown in the paper.

Reviewer Point P 2.1 — In several spots the authors claim that a smoothed estimate of absolute risk is easier to interpret, however by smoothing the data, they are also losing information. Specifically, the height of the steps provides a quick way to determine a crude estimate of the variability. Additionally, in Figure 5 the smoothed curve overestimates the relapse risk at the beginning of the time interval (this was not mentioned as a possible problem with their approach).

Reply: The reviewer brings up an important point that we overlooked. We took the opportunity to add a function that computes confidence bands for survival and risk functions. In the Theoretical Details section, we discuss the validity of our approach and how it relies on the asymptotic normality of the case-base estimator. We also added an illustration of this approach to the first case study. In Figure

4, we can now see a comparison of the (stratified) Kaplan-Meier curves and the case-base confidence bands.

Reviewer Point P 2.2 — A small point, but in the paper the code uses categorical values for ScrArm but in the dataset it is coded as 0/1

```
new_data <- data.frame(ScrArm = c("Control group", "Screening group"))
```

Reply: We updated the dataset in the package to use categorical values.

Reviewer Point P 2.3 — In Example 1 the code shows how to estimate the hazard ratio as function of time. If part of the goal of the paper is to provide comparisons with other survival packages, it might be useful to note that it is easy to estimate the HR within periods of time using the coxph function. This also “fixes” the problem that they are trying to solve.

```
ERSPC2 <- survSplit(Surv(Follow.Up.Time, DeadOfPrCa) ~ ScrArm, data= ERSPC,  
  cut=c(4,8), episode='period')
```

```
coxph(Surv(tstart,Follow.Up.Time,DeadOfPrCa)~strata(period)/ScrArm, data=ERSPC2)
```

Reply: Although a user can estimate piece-wise constant hazard ratios using the function coxph, doing so requires the user to select the cut points. On the other hand, with case-base can model the hazard ratio as a smooth function of time. Figure [R2](#) below shows a comparison of the case-base approach and the piece-wise approach based on Cox regression.

Reviewer Point P 2.4 — Table 4 states that it is difficult to estimate absolute risk from a Cox model if there are competing risks, and perhaps that was true at one point, but the survival package has been updated for several years now and the statement is no longer correct. This approach should be shown in Example 2 and the statement in Table 4 should be modified. In Example 2 it is unclear what “newdata” values were used for the curves, but the following code illustrates the necessary steps.

```
newdat <- expand.grid(D=c('ALL','AML'), Sex='F', Phase='Relapse',Age=30, Source='PB')
```

```
## Cox model for competing risk
```

```
cfit <- coxph(Surv(ftime, factor(Status)) ~ Sex + D + Phase +  
  Source + Age, data=bmtcrr, id=id)
```

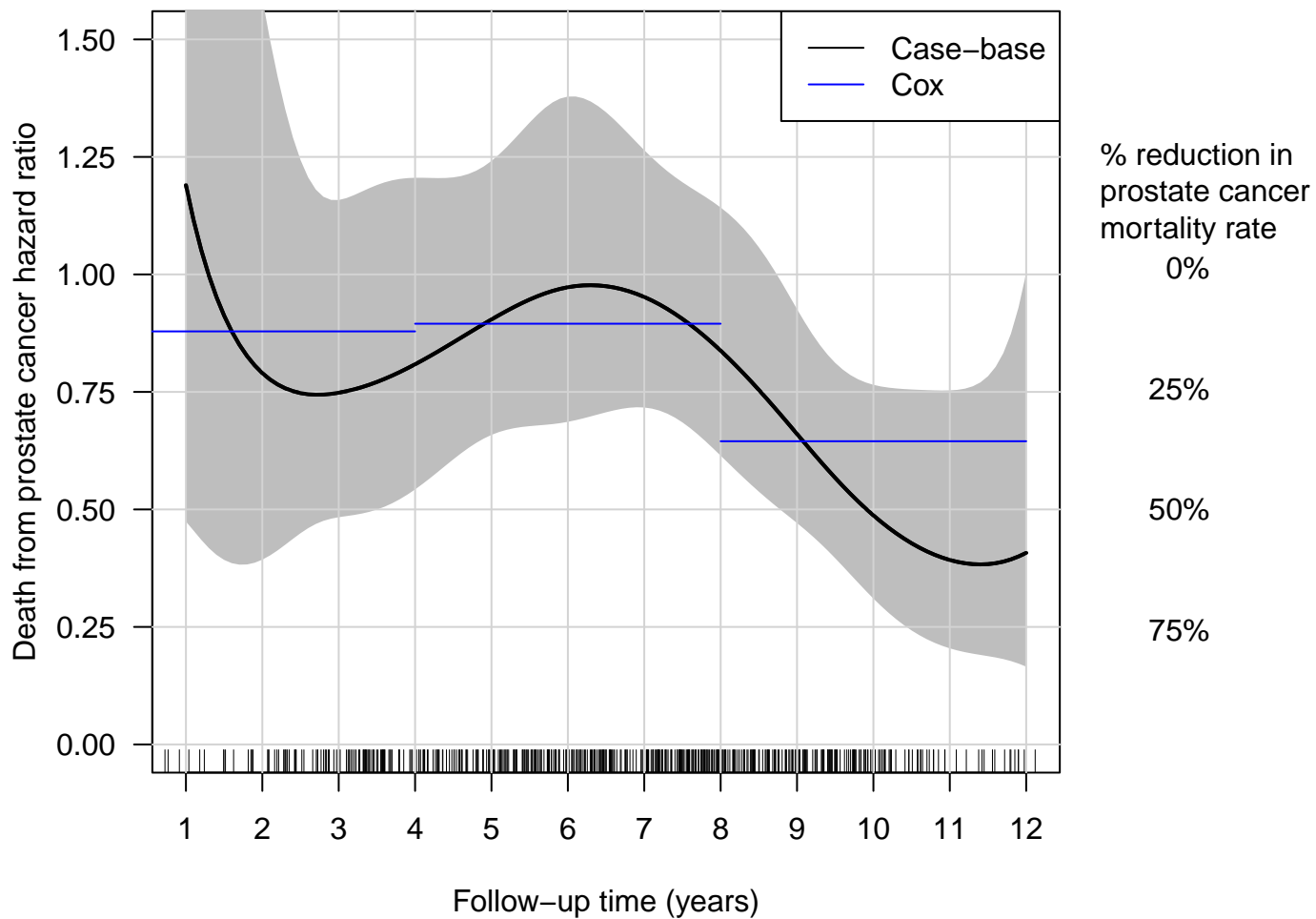
```
## Aalen-Johansen estimate of absolute risk
```

```
plot(survfit(cfit, newdata=newdat)[,2], ylim=c(0,1), col=1:2, xmax=60)
```

```
## Fine-Grey estimate
```

```
fgdata <- finegray(Surv(ftime, factor(Status)) ~ Sex + D + Phase +  
  Source + Age, data=bmtcrr, id=id)
```

Figure R2: Figure 2 from the manuscript, with piece-wise HRs estimated using `coxph` (in blue)



```
fgfit <- coxph(Surv(fgstart,fgstop,fgstatus) ~ Sex + D + Phase +
              Source + Age, data=fgdata, weight=fgwt)

plot(survfit(fgfit, newdata=newdat), ylim=c(0,1), col=1:2, fun='event', xmax=60)

## casebase approach
model_cb <- fitSmoothHazard(Status ~ ftime + Sex + D + Phase + Source + Age,
                           data=bmtcrr, time='ftime')
cbfit <- absoluteRisk(object = model_cb,newdata = newdat,
                    time=0:60)
```



```
matplot(cbf[,1], cbfit[,2:3], col=1:2)
```

Reply: We removed any mention of this limitation, and we added the Cox model and the Aalen-Johansen estimate to our comparison of cumulative incidence function estimates (Figure 6).

Reviewer Point P 2.5 — In Example 3, it isn't clear that the casebase approach calls `cv.glmnet` behind the scenes - that would be easy to mention. Based on this example, the results are virtual identical to just directly calling `cv.glmnet` so it is unclear what the benefit is to using casebase here, except to show that the code runs.

Reply: We have clarified the relationship between casebase and `cv.glmnet` within the manuscript.

In Example 3, we are comparing two approaches: case-base sampling, where the hazard is estimated using penalized logistic regression; and Coxnet, which is a regularized version of the Cox model. As in Examples 1 and 2, the function `fitSmoothHazard` starts by sampling the case series and base series and calculating the offset term. With `family = "glmnet"`, the data is then transformed to match the expected matrix input of `cv.glmnet`, before calling `cv.glmnet` with the offset term. In other words, `fitSmoothHazard` performs most of the necessary data processing.

If a user fits a model using `cv.glmnet` and the Cox family, then they are fitting a semi-parametric model that is different from the case-base model.

Reviewer Point P 2.6 — Notes about the software code:

I think inclusion of `gbm` is extremely dangerous given all the hyperparameters that need to be evaluated. A better approach, if possible, would be to create the appropriate dataset that can then be analyzed with `gbm`.

Reply: The inclusion of `gbm` to the casebase package was done at an early stage but never properly evaluated. Moreover, the `gbm` package, on which this implementation relies, is no longer actively maintained. Therefore, we have removed mentions of `gbm` from the manuscript and from the package documentation. Proper testing and implementation of this approach is now considered future work.