README

Table of Contents

# 1 Overview of PiecewiseChangepoint package

The goal of PiecewiseChangepoint is to estimate the number and locations of change-points in piecewise exponential models.

## 1.1 Installation

You can install the released version of PiecewiseChangepoint from [GitHub](https://github.com/Anon19820/PiecewiseChangepoint) with:

devtools::install\_github("Anon19820/PiecewiseChangepoint")

In order to run some of the functions JAGS and Stan are required along with RTools??

## 1.2 Simulated Example

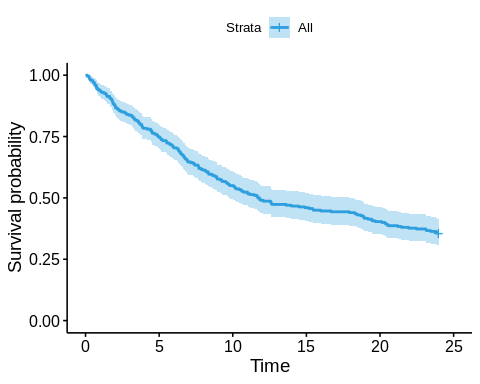
First we load the package and simulate some piecewise exponential data.

library("PiecewiseChangepoint")  
  
## simulated example  
set.seed(123)  
n\_obs =20  
n\_events\_req=20  
max\_time = 24 # months  
  
rate = c(0.75,0.25)/12 # we want to report on months  
t\_change =12 # change-point at 12 months  
  
df <- gen\_piece\_df(n\_obs = n\_obs,n\_events\_req = n\_events\_req,  
 num.breaks = length(t\_change),rate = rate ,  
 t\_change = t\_change, max\_time = max\_time)

We see the output of this dataframe below:

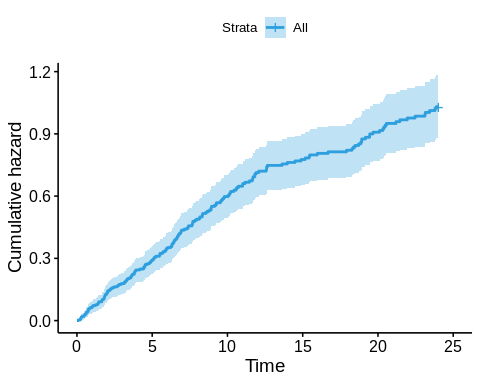
## time\_event status time  
## 24 0.09194727 1 0.09194727  
## 193 0.23141129 1 0.23141129  
## 87 0.24251702 1 0.24251702  
## 126 0.25450622 1 0.25450622  
## 297 0.28833655 1 0.28833655  
## 139 0.32615105 1 0.32615105

For this simulated dataset; *time\_event* represents the time the event would occur at in the absence of censoring, while *time* is minimum of the censoring time and the event time. *status* is an indicator variable if the event occurred at the corresponding time or if it was censored. Plotting this survival function we see a potential change in the hazard at around year 1.



As noted in ([Bagust and Beale 2014](#ref-Bagust.2014)), constant hazards are linear with respect to the cumulative hazard function, therefore, the change in hazards at approximately 12 months can be seen more clearly in this plot.

ggsurvplot(fit, palette = "#2E9FDF", fun = "cumhaz")



Next we fit the model noting that only the time and status columns are required. The timescale argument changes the prior for the hazards so that it is appropriate for the timescale. For example if the timescale is years then the a vague prior centered around 1 is appropriate (i.e.  of population having the event each year), while if the timescale is in months the equivalent prior should have an expected value of 1/12 (and days 1/365).

Collapsing\_Model <- collapsing.model(df,  
 n.iter = 20750,  
 burn\_in = 750,  
 n.chains = 2,  
 timescale = "months")

As we would expect the one change-point model has the highest posterior probability.

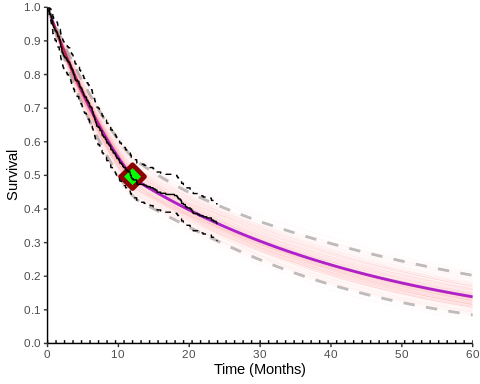
print(Collapsing\_Model)

print(Collapsing\_Model)

## Posterior Change-point Probabilities:  
## 0 1 2 3 4 5   
## 0.000575 0.809525 0.166000 0.020650 0.002850 0.000400   
##   
## Summary of 1 change-point model:  
##   
## changepoint\_1 lambda\_1 lambda\_2   
## Min. : 1.873 Min. :0.04206 Min. :0.01170   
## 1st Qu.:11.701 1st Qu.:0.05614 1st Qu.:0.02355   
## Median :11.900 Median :0.05940 Median :0.02652   
## Mean :11.857 Mean :0.05954 Mean :0.02682   
## 3rd Qu.:12.603 3rd Qu.:0.06277 3rd Qu.:0.02983   
## Max. :19.457 Max. :0.08456 Max. :0.04832

Once we are satisfied that there is good mixing and that we have run the model for long enough (20,000 simulations over 2 chains should be more than enough), we may want to look at a plot of the survivor function. In health economics we are typically interested in long term survival of our parametric models. In this situation we want a plot of the first 5 years which we can do using the *max\_predict* argument (in this case 60 months). The red lines show the individual posterior simulations and are a natural representation of the parameter uncertainty.

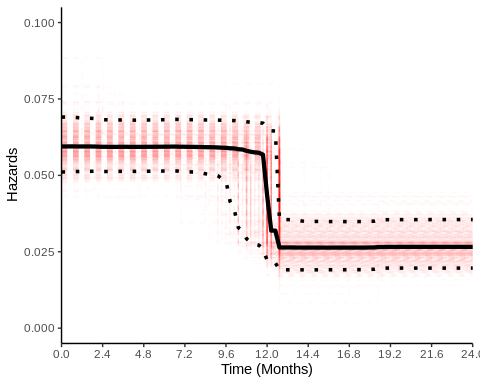
plot(Collapsing\_Model, max\_predict = 60, chng.num = 1)+xlab("Time (Months)")



Similarly we may also want to look at the hazard function. In this situation we only present the hazard up to the maximum time observed in the data. This is because by definition the hazard from the final interval will be the one which is extrapolated throughout the time horizon.

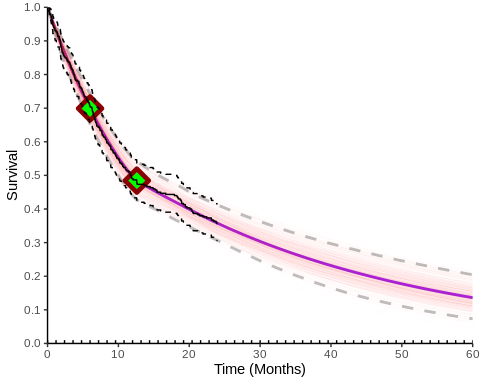
plot(Collapsing\_Model, type = "hazard")+xlab("Time (Months)")+ylab("Hazards")+ylim(c(0,.1))

## Scale for 'y' is already present. Adding another scale for 'y', which will  
## replace the existing scale.



By default the plot methods described above use all the posterior simulations. If for example, we were only interested in the 2 change-point model, we can specify this using the *chng.num* argument. The green points indicate the mean location of the change-points. When plotting “all” of the simulations there is no sensible mean location of the change-points as there are different numbers of change-points.

plot(Collapsing\_Model, max\_predict = 60, chng.num = 2)+xlab("Time (Months)")



## 1.3 Comparison with other survival models

### 1.3.1 Assessment of Goodness of Fit

In health economics we are typically interested in picking between one of a number of alternative parametric models, although it is also possible to combine all models using model averaging ([Jackson, Sharples, and Thompson 2010](#ref-Jackson.2010)). Model fit statistics can provide an assessment of fit to the **observed** data, although, they do not guarantee the best fitting model will be appropriate for extrapolation. Nevertheless, we can compare our fitted model with 6 commonly used parametric models along with Royston-Parmar spline models. We fit the models using the JAGS ([Plummer 2003](#ref-Plummer.2003)) and Stan ([Stan Development Team, n.d.](#ref-RStan.2023)) and compare the model fit using Widely Applicable Information Criterion (WAIC) ([Watanabe 2010](#ref-Watanabe.2010)).

### 1.3.2 Including General Population Mortality

Including General Population Mortality (GPM) is required to ensure that the extrapolated hazards are consistent with the increasing hazards associated with advanced ageing. Adjustments for GPM is typically done within the cost-effectiveness model, however, we can include them directly at the analysis stage so that we see their impact on the extrapolated survival.

In this example we consider GPM from a UK data source which provides mortality rates, defined as “the probability of that a person aged exactly will die before reaching . Therefore, this data source provides the conditional probability of death within a year at each age.

Assuming our population is male and female and the age at baseline is 55 years we have the following conditional probabilities of death at each age:

age\_baseline\_example <- 55  
prop\_male <- 0.5  
time\_horizon <- 100   
  
Conditional\_Death\_df <- read.xlsx(paste0(pathway, "Examples/Conditional\_Death\_UK.xlsx"), 1) %>%   
 filter(age >=age\_baseline\_example)  
head(Conditional\_Death\_df)

## age Males..2018.2020. Females.2018.2020  
## 1 55 0.005046 0.003283  
## 2 56 0.005593 0.003637  
## 3 57 0.006060 0.003928  
## 4 58 0.006695 0.004367  
## 5 59 0.007239 0.004707  
## 6 60 0.007912 0.005247

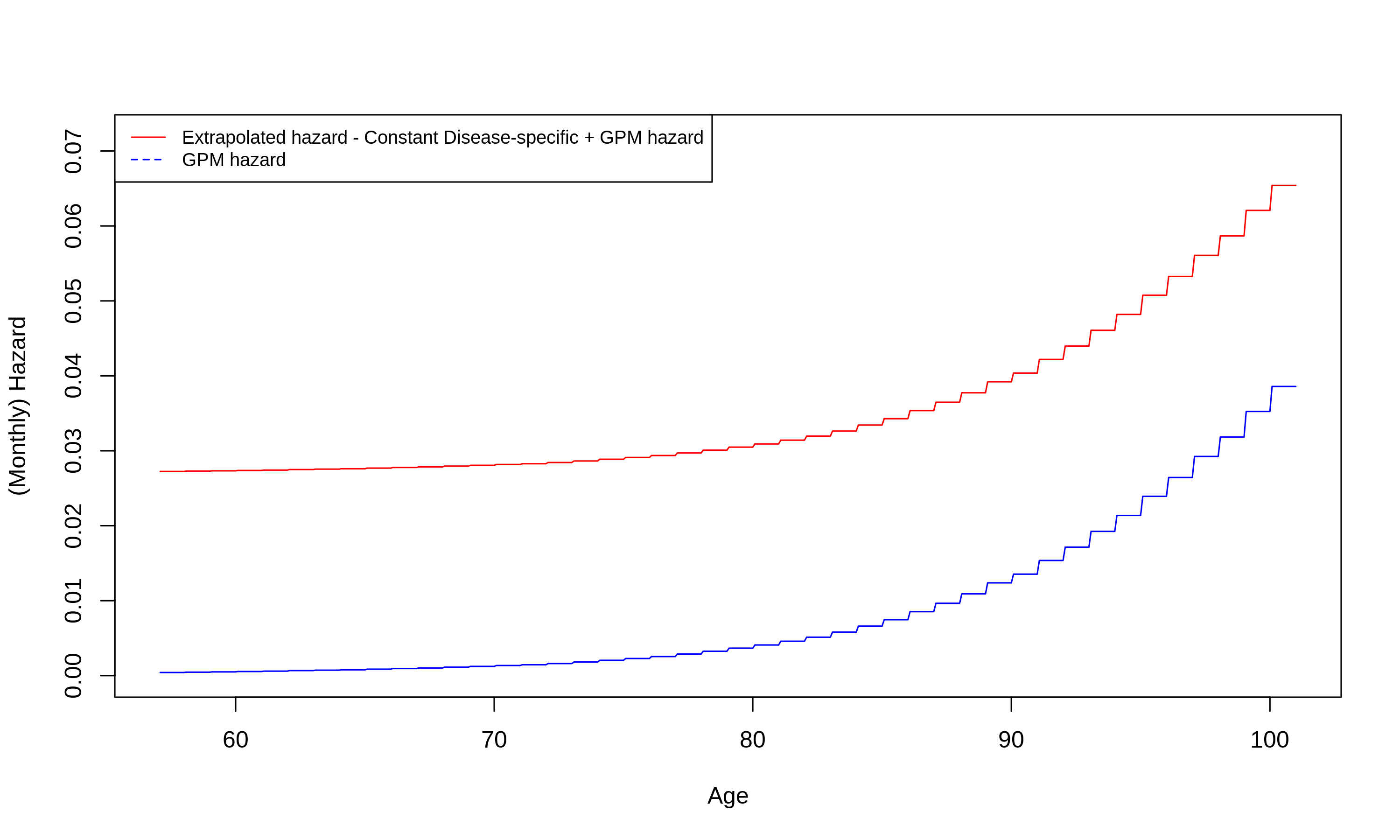
Our timescale is months and we need to convert this annual probability to a monthly rate which is done using the following formula (assuming a constant rate of mortality) ([Fleurence and Hollenbeak 2007](#ref-Fleurence.2007)):

Because there are 12 months in a year and is the specific (in our case annual) probability of death. With the below R code we now have the monthly rate of death for ages 55 (our assumed starting age of the cohort) up to 100 years of age, adjusted for distribution of males and females.

time\_factor <- 12  
df\_temp <- Conditional\_Death\_df  
df\_temp[, "mix\_prob"] <- df\_temp[,2]\*prop\_male + df\_temp[,3]\*(1-prop\_male)  
df\_temp <- df\_temp %>% filter(age >= age\_baseline\_example & age <= time\_horizon)  
  
df\_temp$mix\_haz <- -log(1-df\_temp$mix\_prob)/time\_factor  
  
gmp\_haz\_vec\_example = rep(df\_temp$mix\_haz,each = time\_factor)  
#We now have the hazard at each timepoint  
gmp\_haz\_df\_example <- data.frame(time = 1:length(gmp\_haz\_vec\_example),  
 hazard = gmp\_haz\_vec\_example)

Within the compare.surv.mods function the cumulative hazard of death (associated with GPM) and cumulative hazard of an event (from the parametric model) is added to obtain the overall cumulative hazard . The cumulative hazard is the sum (in the case of discrete hazards as above) of the individual hazards and the integral of the parametric hazards. Survival probabilities are obtained through the relation . By default the compare.surv.mods function only implements GPM hazards after follow-up as we observe survival from all causes up until then (although GPM hazards can be added from start of follow-up by using the gpm\_post\_data = FALSE).

We see in the plot below that including the GPM hazard ensures that the extrapolated hazard exhibits the characteristic increasing hazards associated with ageing.



Overview of all TAs investigated

### 1.3.3 Fitting of Standard Parametric models and Plot of Extrapolated Survival

The fitting of other parametric models is accomplished by the compare.surv.mods and general population mortality is adjusted for by including a gmp\_haz\_df as described above. Fitted models include:

* Exponential
* Weibull
* Gamma
* Gompertz
* Generalized Gamma
* Royston-Parmar Spline (best fitting by WAIC between 1 and 2 knot)

Model fit to the observed data and a plot of the extrapolated survival are available from within the mod\_comp object along with the posterior samples from all of the fitted models.

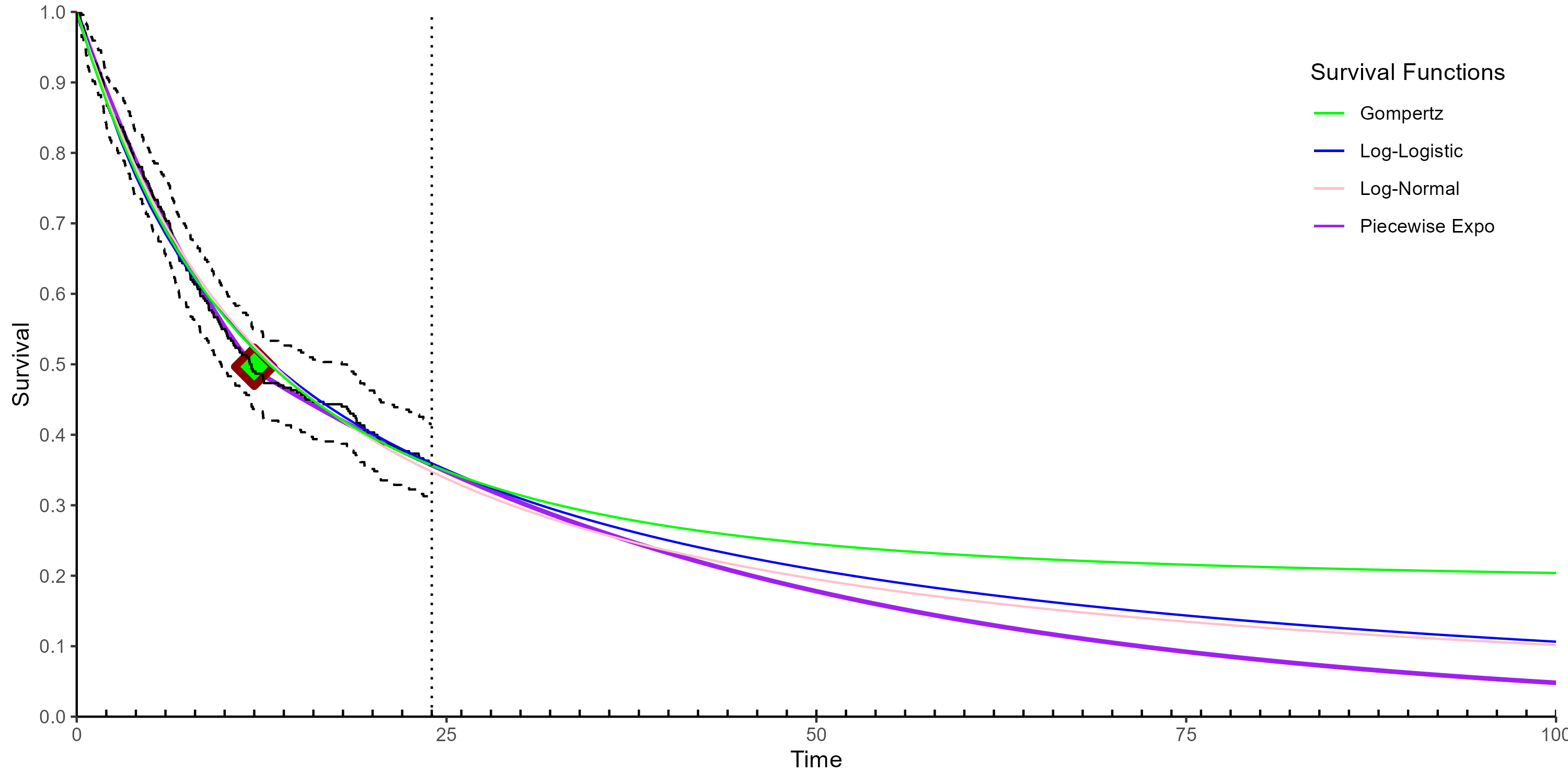
#This can take a number of minutes   
set.seed(123)  
mod\_comp <- compare.surv.mods(Collapsing\_Model,  
 max\_predict = 100, #100 months  
 n.iter.jags = 5000, #Run JAGS/Stan for 5000 samples  
 n.thin.jags = 1,  
 n.burnin.jags = 500,  
 chng.num = 1, #Using results from 1 change-point PEM  
 gmp\_haz\_df =gmp\_haz\_df\_example) #GPM dataset   
  
mod\_comp$mod.comp[,c(1,3)]

#Returns a dataframe with the model fit results  
mod\_comp$mod.comp[,c(1,3)] %>% arrange(WAIC)

## Model WAIC  
## 1 Piecewise Exponential 1547.599  
## 2 Log-Normal 1552.492  
## 3 Log-Logistic 1553.213  
## 4 Gompertz 1553.266  
## 5 Royston-Parmar 2 knot 1553.743  
## 6 Generalized Gamma 1556.875  
## 7 Weibull 1561.850  
## 8 Gamma 1564.098  
## 9 Exponential 1568.053

mod\_comp$plot\_Surv\_all

## NULL



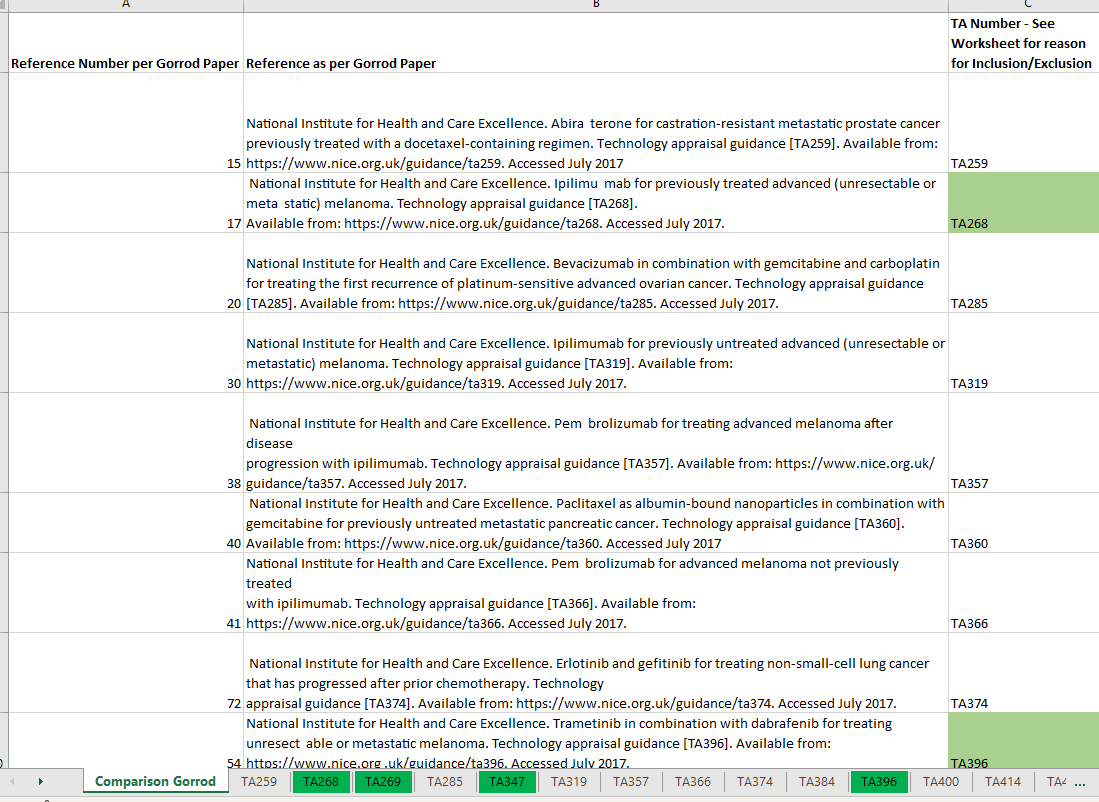
Extrapolated Survival Probability

# 2 Replication of Results Presented in Manuscript

## 2.1 Details of Data Extraction from Technology Appraisals

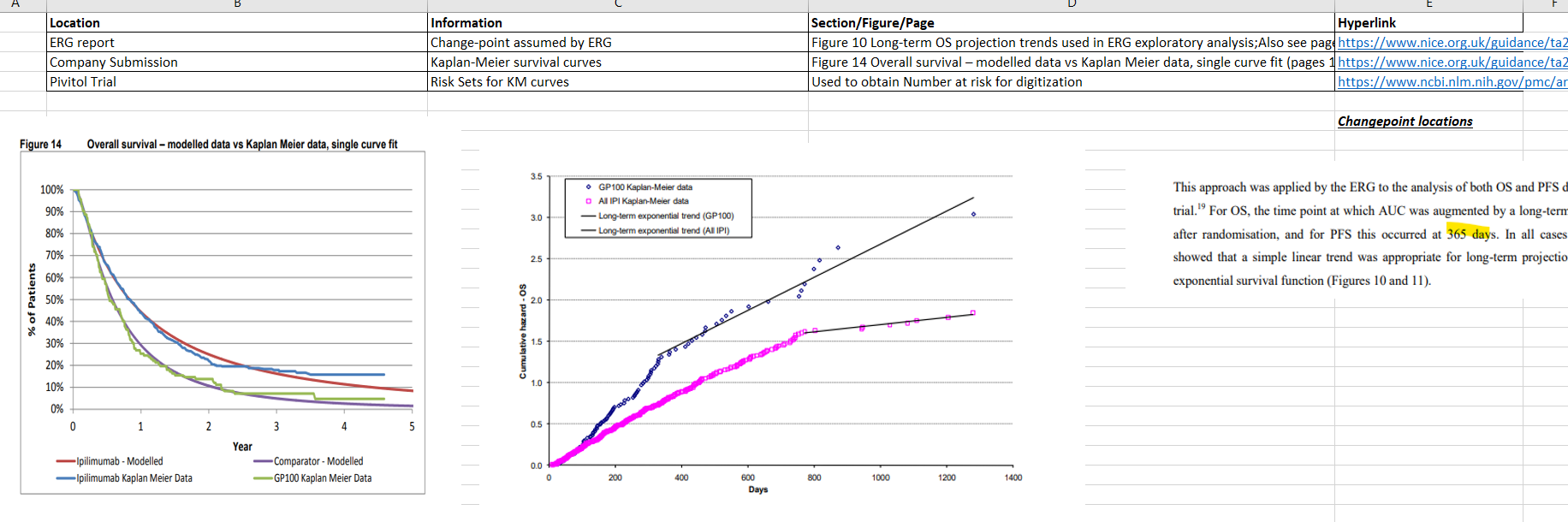
Information used for the identification of the relevant Technology Appraisals (TAs) from the review by ([Gorrod et al. 2019](#ref-Gorrod.2019)) along with the relevant information extracted from them can be found in the Excel file located here.

The first worksheet of this file lists the TAs investigated (Figure 3).



Overview of all TAs investigated

For each of the TAs listed in the first worksheet a separate worksheet provides further information relating to whether or not it used the Bagust and Beale (B&B) approach along with the relevant location in the TA and an associated screengrab of the relevant information. In situations where B&B was confirmed to have been used, further information including the Kaplan-Meier survival curves and location of the assumed change-point are recorded along with the respective locations within the TA are recorded. Kaplan-Meier curves for any survival data made available after the original TA is along presented along with a link to the relevant data-source. For an example of some of the data extracted from TA268 ([TA268 2012](#ref-TA268)).



Overview of all TAs investigated

## 2.2 Analysis of Extracted data using PiecewiseChangepoint package

All results from the manuscript can be replicated using the folder located here.

The folder contains a R script titled Digitizing\_R\_code\_Final\_Share.R which will produce relevant plots and tables in the folder named pub-plots-tabs, using the R functions described previously. A number of sub-folders are also contained within the main folder and provide pseudo-patient data created from the Kaplan-Meier curves presented in the TAs (and in publications providing later datacuts). These are named as follows TA\_Treatment\_Outcome\_Datacut and are read in and digitized from the survival.txt and nrisk.txt files in these folders.



Overview of File Structure

# References

Bagust, Adrian, and Sophie Beale. 2014. “Survival Analysis and Extrapolation Modeling of Time-to-Event Clinical Trial Data for Economic Evaluation: An Alternative Approach.” *Medical Decision Making* 34 (3): 343–51. <https://doi.org/10.1177/0272989X13497998>.

Fleurence, Rachael, and Christopher Hollenbeak. 2007. “Rates and Probabilities in Economic Modelling: Transformation, Translation and Appropriate Application.” *PharmacoEconomics* 25 (February): 3–6. <https://doi.org/10.2165/00019053-200725010-00002>.

Gorrod, Helen Bell, Ben Kearns, John Stevens, Praveen Thokala, Alexander Labeit, Nicholas Latimer, David Tyas, and Ahmed Sowdani. 2019. “A Review of Survival Analysis Methods Used in NICE Technology Appraisals of Cancer Treatments: Consistency, Limitations, and Areas for Improvement.” *Medical Decision Making* 39 (8): 899–909.

Jackson, C. H., L. D. Sharples, and S. G. Thompson. 2010. “Structural and parameter uncertainty in Bayesian cost-effectiveness models.” Journal Article. *J R Stat Soc Ser C Appl Stat* 59 (2): 233–53. <https://doi.org/10.1111/j.1467-9876.2009.00684.x>.

Plummer, Martyn. 2003. “JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling.” *3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria* 124 (April).

Stan Development Team. n.d. “RStan: The R Interface to Stan.” <https://mc-stan.org/>.

TA268. 2012. “Ipilimumab for previously treated advanced (unresectable or metastatic) melanoma: Technology Appraisal Guidance.” *NICE*. <https://www.nice.org.uk/guidance/ta268>.

Watanabe, Sumio. 2010. “Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory.” *J. Mach. Learn. Res.* 11 (December): 3571–94.