Time to publication protocol Survival Analysis

|  |  |
| --- | --- |
| Julien CAPITAINE | Katarina LECHNER |
| Chidiac Kieffer NKWAMEN NANA | Louis BARANZELLI |

Authors:

# **Goal of the Project**

The goal of this project is to perform a survival analysis study. The following investigation has been performed on a [Time to publication dataset] (https://figshare.com/articles/dataset/Time\_to\_publication\_data/4054878) represented the time spent between the agreement for a public funding of a clinical trial and the related publications. All the trials gathered in this dataset have been made in Australia and funded by the Australian government only.

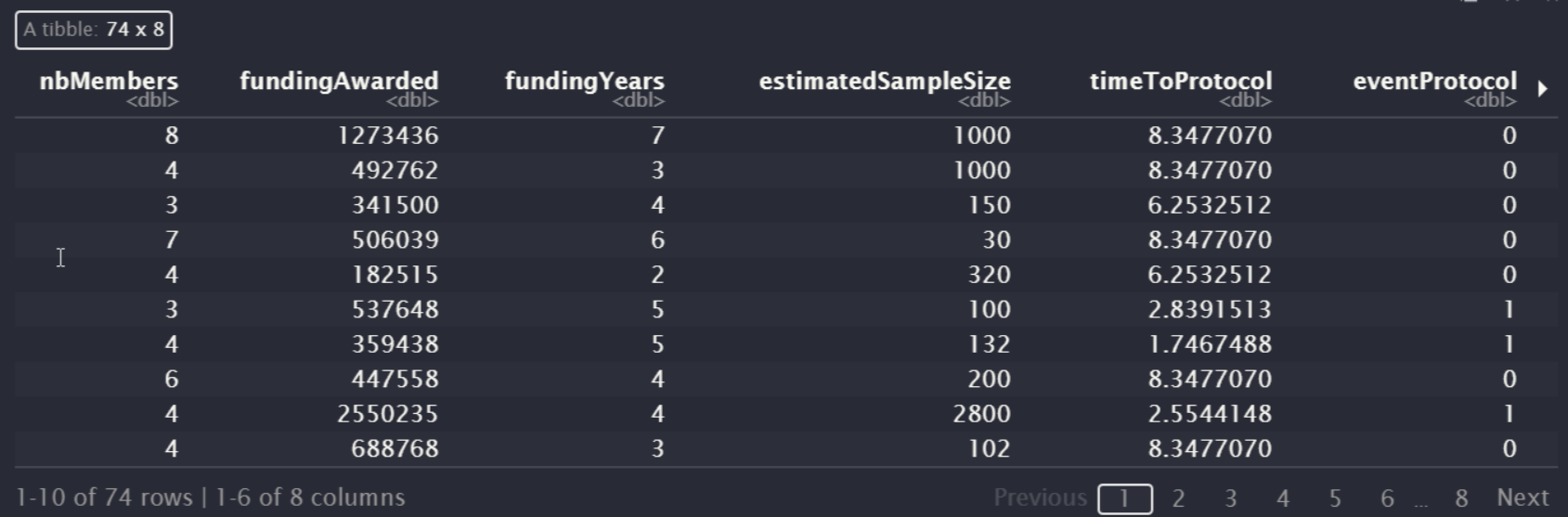
The special feature of this dataset is that it contains two consecutive publication events. The first possible event is the protocol paper, the second is the main paper which is also the main event of interest. We will refer to the publication of the main paper event as the `main event` to make it short.

The purpose of the analysis is to evaluate whether (1) the funding received from the government is well used and (2) the end-goal of all scientific researches is met (i.e. sharing of the knowledge with the community). The business goal is to measure the efficiency of public fundings and to provide means to leverage in order to progress toward scientific excellence.

## Data: loading & cleaning

In our case the data provided is already relatively clean.

myData



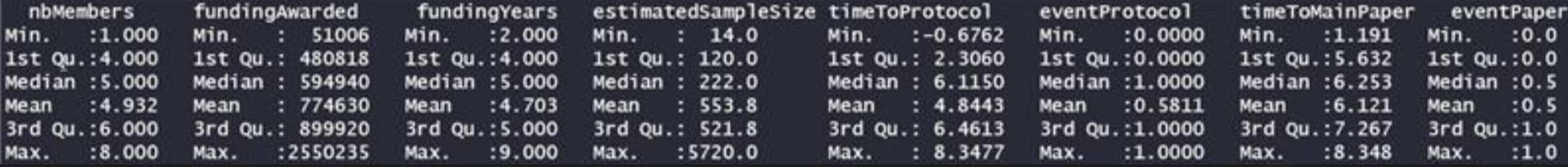
The meaning of the columns in the dataset is as follows:

* nbMembers: number of investigators;
* fundingAwarded: funding awarded ($AUD); scrambled by -/+ $1000;
* fundingYears: length of funding in years;
* estimatedSampleSize: estimated sample size (some missing);
* timeToProtocol: time in years from funding until protocol paper was published (or censored);
* eventProtocol: protocol paper published (1=yes, 0=censored). The goal of the protocol is to communicate methods and primary outcomes in order to avoid outcome switching and duplicate research;
* timeToMainPaper: time in years from funding until main paper was published (or censored);
* eventPaper: (1=yes, 0=censored).

Note: Main paper presents the results of a clinical trial, a protocol paper outlines the plan for conducting the trial.

In our analysis we consider the $eventPaper$ as the main event we want to predict the time to with $eventProtocol$, $timeToProtocol$, $estimatedSampleSize$, $fundingYears$, $fundingAwarded$ and $nbMembers$ as covariates.

summary(myData)



From the summary, we see the following:

* There are 77 observations.
* In the timeToProtocol column the minimum value is negative -> we assume that the protocol in this case was written before the funding was granted, so we keep the value as is.
* The estimatedSampleSize is of type character -> this is because there are three NA values. We remove these observations.
* Based on the eventPaper mean which is 0.5, we see that half of the observations have a value 1 and half a value 0 -> The sample is balanced. The same observation applies to the variable eventProtocol.

**Cleaning of the data**

Remove NA values

Change $estimatedSampleSize$ from char to numeric

myData$estimatedSampleSize <- as.numeric(myData$estimatedSampleSize)

myData <- myData[complete.cases(myData),]

Note: Our dataset now has 74 observations.

# Exploratory Data Analysis

## Understanding the events

**Code for the violin box plot**

ggstatsplot::ggbetweenstats(

data = myData,

x = eventProtocol,

y = timeToProtocol,

xlab = "Event outcome",

ylab = "Time to protocol (years)",

title = "Protocol distributions",

plottype = "box", type = "p", conf.level = 0.95, results.subtitle = FALSE,

) + ##C> - Modifying the plot further

ggplot2::scale\_y\_continuous(

limits = c(-1, 9),

breaks = seq(from = -1, to = 9, by = 1)

)

plotMainPaper <- ggstatsplot::ggbetweenstats(

data = myData,

x = eventPaper,

y = timeToMainPaper,

xlab = "Event outcome",

ylab = "Time to main Paper (years)",

title = "Paper distributions",

plottype = "box", type = "p", conf.level = 0.95, results.subtitle = FALSE,

) +

ggplot2::scale\_y\_continuous(

limits = c(-1, 9),

breaks = seq(from = -1, to = 9, by = 1)

)

grid.arrange(plotProtocol, plotMainPaper, ncol=2)

**Code for confusion matrix**

cm <- confusionMatrix(factor(myData$eventPaper), factor(myData$eventProtocol), dnn=c("eventPaper", "eventProtocol"))  
fourfoldplot(as.table(cm),color=c("green","red"),main = "Confusion Matrix")

|  |  |
| --- | --- |
| 1. Events outcomes violin plots | 1. Matrice de confusion |

From the plots we see the following:

* It is not mandatory to publish a protocol before publishing the main paper. In these cases (eventProtocol = 0), the timeToProtocol has a median of 6.99y, i.e. almost 7 years;
* The distribution of eventProtocol = 1 is slightly right-skewed. We don't see any outliers in terms of timeToProtocol;
* Same as before, for eventPaper = 0, the mean and median are much higher than for cases eventPaper = 1;
* The distribution of eventPaper = 1 is slightly left-skewed. Same as before, we don't see any outliers in the data;
* The data in terms of eventProtocol is 'relatively' balanced (31 `0`s vs 43 `1`s);
* In terms of eventPaper, data is balanced (37 samples for both values of eventPaper `0`s and `1`s).

## Understanding the covariates

In the following of the EDA we do not show the code we wrote, we only display some figures (figures 3 to 6 grouped after) which we thought were interesting (or not) and explain our results based on them.

In **figure 3**, we plotted the (linear) correlation matrix between all the explanatory variables. It appears that most of the variables are positively correlated with one another (either 0.2 or 0.4). Only estimatedSampleSize is not correlated with nbMembers and fundingYears at all.

In **figure 4**, we display a scatterplot of the amount awarded (in $AUD) w.r.t. the duration of the funding, by distinguishing between positive and censored outcomes. We do recover the linear relationship (corcoef = 0.34) between this variable. Obviously, the longer the duration of the funding, the larger the amount provided.

**In**

|  |  |
| --- | --- |
| 1. Correlation matrix | 1. **Relationship** between **funding years and funding awards** |
| 1. Relationship between number of Members and time to main event | 1. Distribution of funding awarded group by event paper |

In **Figure 5** we display a scatterplot of the number of team-members w.r.t. the time to the main event, and distinguishing positive and censored outcomes. We observe that all censored outcomes are older than six years and spread over three values. This should indicate that the funding are granted during comission which take place yearly. We assume that all observations were taken from a set of three consecutive years and that's all. Then, if we focus on the positive outcomes, the trials may yield a publication at any time following these grants. And by studying the bulk of the points we see that most trials lead to a publication after five years.

In **Figure 6** we display some violin plots of the funding awarded w.r.t. to main event outcome, with some statistics. These distributions are alike, whatever the outcome. This plot is representative of the relation between all covariates. No specific relationship.

# Survival analysis

## Plot survival function

To study the evolution of the main event, we plot the survival function a.f.o. time.

It is observed that the function does not reach 0 since only 50% of the sample are able to publish a paper.

fit <- survfit(Surv(myDatatimeToMainPaper, myDataeventPaper) ~ 1, data = myData)

ggsurvplot(fit,

pval = TRUE, conf.int = TRUE,

risk.table = TRUE, # Add risk table

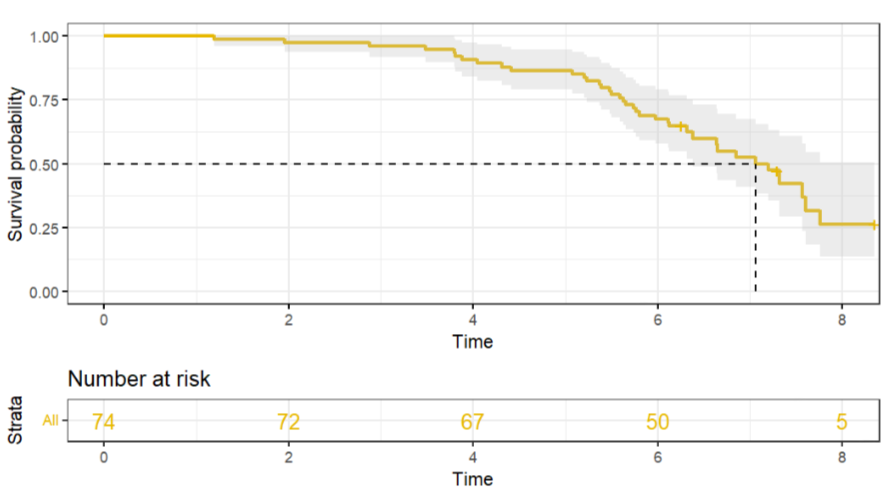
risk.table.col = "strata", # Change risk table color by groups

linetype = "strata", # Change line type by groups

surv.median.line = "hv", # Specify median survival

ggtheme = theme\_bw(), # Change ggplot2 theme

palette = c("#E7B800", "#2E9FDF"))



1. Survival function

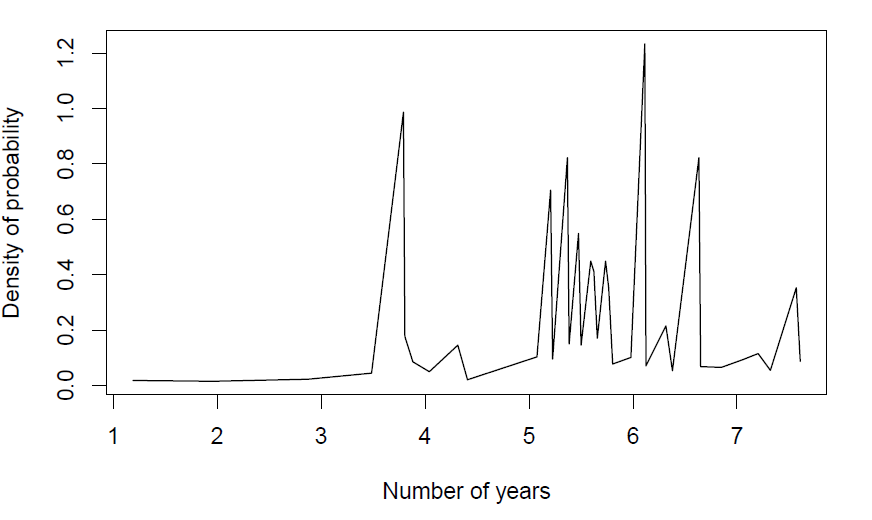
## Create the Hazard Function

In order to compute the Hazard Function we need to define the PDF.

PDF <- diff(cum\_distribution) / diff(time\_sorted)

time <- time\_sorted[-length(time\_sorted)]

plot(time , PDF, main="PDF", type = "l", ylab="Density of probability", xlab="Number of years")



1. PDF

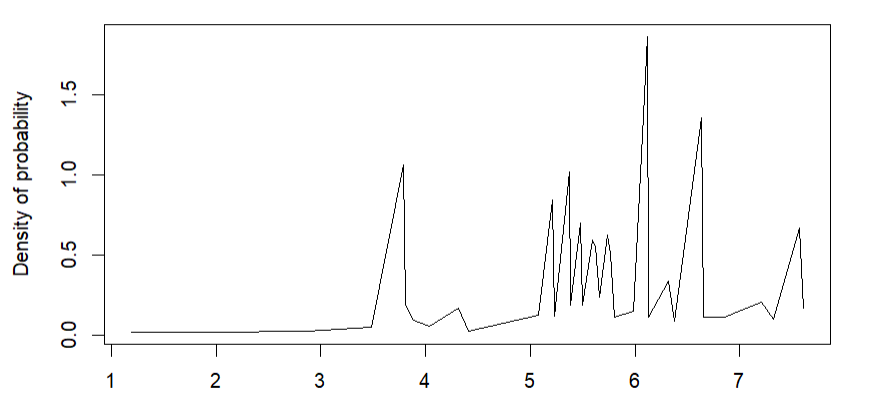
The PDF shows the instantaneous probabilities of publishing a paper.

Note: We observed that the graphical integration of the PDF is not equal to 1 due to the "sawtooth" shape of the curve, which introduces significant inaccuracies. Indeed, the data are sparse and not evenly distributed across the time axis.

## Compute the Hazard Function

hasard\_function = PDF / survival[-length(survival)]

plot(time , hasard\_function, main="Hazard Function", type = "l", ylab="Density of probability", xlab="Number of years")

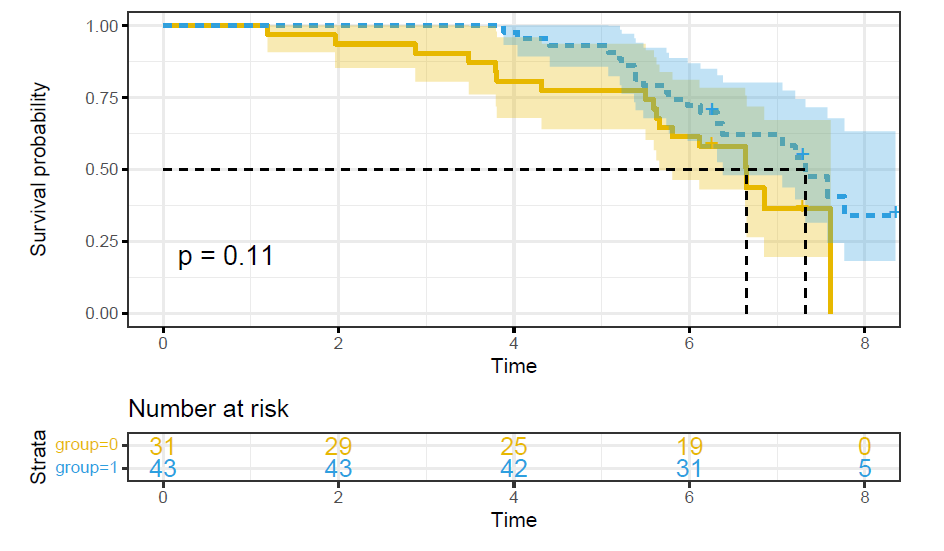


1. Hazard Function

The HF provides the instantaneous probability at any given time of the main event. The curve shares similar limitations as the PDF itself. Having more data would likely result in a smoother curve and provide a more accurate representation.

## Deep-dive in 'eventProtocol' and how it influences the main event

Does publishing an experimental protocol absolutely increase the chances of publishing a paper? This is the question we address here by comparing the two survival curves: one for those who have published a protocol (Group 1) and the other for those who haven't (Group 0).



1. Survival distribution of the main event depending on the event protocol

dat <- data.frame(time=myDatatimeToMainPaper, status=myDataeventPaper, group=myDataeventProtocol)

survdiff(Surv(time, status) ~ group, data = dat)

With a p-value of 0.1, we cannot reject this hypothesis, so we can consider that whether or not someone has submitted a protocol has no influence on the likelihood of the corresponding main event.

# Cox model

## Univariate Cox model

For the different variables, we check if they impact individually the main event by checking the p-value (statistically meaningful or not). If the p-value is bigger than a threshold (here threshold = 0.05), we cannot reject the H0.

The null hypothesis in this Cox model is that there is no significant relationship between the explanatory variable (in the different cases) and the survival time of the main event.

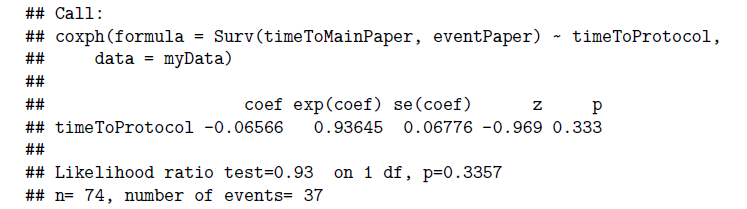
As usual we show the details of the calculation for only one variable. Then, we will present only the conclusion.

**Impact of** timeToProtocol

res.cox <- coxph(Surv(timeToMainPaper, eventPaper) ~ timeToProtocol, data = myData)

res.cox

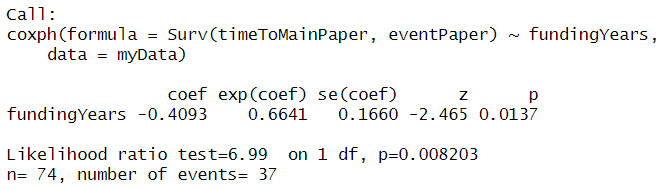
The p-value is 0.33. timeToProtocol has individually no impact on eventPaper.



1. Result of the univariate Cox model

**For the other variables**

* The p-value is 0.0068. fundingYears has individually an impact on eventPaper.



1. Result of the univariate Cox model

The coef -0.4 indicates that the higher the values of the variable fundingYears, the less significant the chance of an eventPaper. With an exp(coef) = 0.66, a decrease of fundingYears by 1 induces an increase of the chance of the main event by 34%. And by the way, we can observe this effect in the **figure 4**, in which there is almost no publication after 6 years.

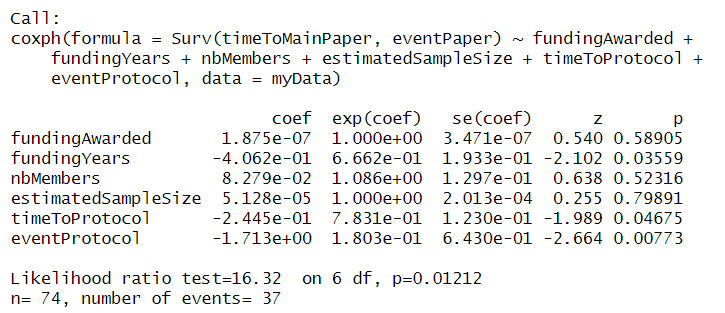
* The p-value is 0.68. nbMembers has individually no impact on eventPaper.
* The p-value is 0.62. estimatedSampleSize has individually no impact on eventPaper
* The p-value is 0.95. fundingAwarded has individually no impact on eventPaper.

## Multivariate Cox model

Now we consider the combination of variables.

res.cox <- coxph(Surv(timeToMainPaper, eventPaper) ~ fundingAwarded + fundingYears + nbMembers + estimatedSampleSize + timeToProtocol + eventProtocol, data = myData)

res.cox

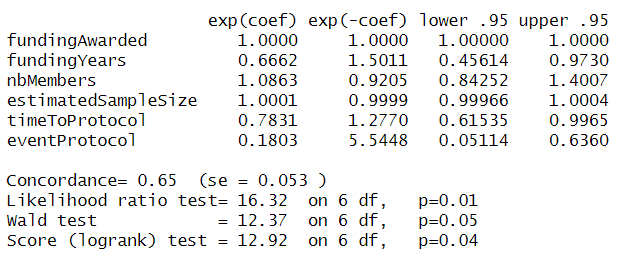


1. Results of the multivariate Cox-Model

We get these results:

* In terms of variables, fundingYears, timeToProtocol and eventProtocol are significant, with respective p-values equal to 0.035, 0.046 and 0.007.
* All 3 of them with negative coef values.
* It means that the higher the values (of the variables fundingYears, timeToProtocol and eventProtocol) associated with these coefficients, the less significant the chance of the main event.
* It means that, in case of a positive eventProtocol, the longer it took to get published, the lower the chance to get a positive outcome for the main event; and we observe that fundingYears, timeToProtocol are positively correlated.

summary(res.cox)



1. Summary of the res.cox

The p-value for all three overall tests (Likelihood, Wald, and Score) are significant, indicating that the model is significant: null hypothesis is soundly rejected.

In the multivariate Cox analysis, the covariates fundingYears, timeToProtocol and eventProtocol remain significant (p < 0.05).

The p-value for fundingYears is 0.035 (statistically meaningful), with HR = exp(coef) = 0.67, indicating a relationship between the fundingYears variable and the main event. Decrease of the fundingYears by 1 means an increase of the chance of the main event by 33%.

The p-value for timeToProtocol is 0.045 (statistically meaningful), with HR = exp(coef) = 0.78, indicating a relationship between the timeToProtocol variable and the main event. A decrease of timeToProtocol by 1 year means an increase of the chance of the main event by 22%.

The p-value for eventProtocol is 0.007 (statistically meaningful), with a HR = exp(coef) = 0.18, indicating a strong relationship between the eventProtocol and the main event. In the frame of the multivariate Cox-Model, with the inclusion of fundingYears and timeToProtocol as explanatory variables, belonging to the censored group for the protocol event increases the chance of the main event by 82%.

# Conclusions

After the analysis, we can get the following conclusions:

1. The dataset contains information about various variables, including the number of investigators, funding awarded, funding years, estimated sample size, time to protocol paper publication, event of protocol paper publication, time to main paper publication, and event of main paper publication. The dataset consists of 74 observations.
2. The data was relatively clean, but some steps were required to handle missing values and convert certain variables from character to numeric format (estimatedSampleSize).
3. Exploratory Data Analysis: The analysis includes several visualizations and summaries to understand the data and relationships between variables. We can say that there is homogenous correlation between almost all variables.
4. The publication of a protocol paper does not seem to have a significant impact to the main event.
5. Survival analysis techniques were applied to understand the time to chance of main event (i.e. publication of the main paper) and the factors influencing it. We notice that:

* The survival function indicates that only 50% of the sample publishes a paper within the observed time. The hazard function provides the instantaneous probability of publishing a paper at any given time.
* The impact of individual covariates on the main event was assessed using univariate Cox models. Only fundingYears was shown to have a significant impact.
* A multivariate Cox model was built, including fundingYears, timeToProtocol and eventProtocol. These variables remained significant, indicating that longer fundingYears, longer timeToProtocol, and negative eventProtocol increase the risk of not publishing a main paper.
* FundingYears, timeToProtocol, and eventProtocol have were shown to have significant impact on the main event.
* Longer FundingYears, timeToProtocol are associated with a decreased likelihood of publishing a main paper.
* The publication of a protocol paper seems to have a positive impact on the likelihood of publishing a main paper, although results are not statistically significant.
* The number of investigators, the amount of the funding awarded, and the estimated sample size did not show a significant impact on the time to main event.