Survival Analysis of Myocardial Infarction Data

Claire DAVAT, FONDZEWONG WIYFENGLA Anslem, Andrew WIEBER

25 August 2024

Introduction

Myocardial infarction (MI), generally referred to as a *heart attack*, is a frequently encountered health problem that often leads to death. It has many risk factors, including genetics and diet. This data is generally complicated to obtain. The presence of heart conditions, diabetes, sex, and age are easier to obtain. The TRACE dataset in the R *timereg* package contains such information. In this dataset, just over half of the patients perish from the event. There are competing hazards (other causes of death) that represent a tiny fraction of overall data. These other causes are unspecific. Unfortunately, there is no data regarding non-lethal myocardial infarctions that may have occurred during the study period.

This data is approached via two survival analysis methods: Kaplan-Meier + log-rank test and Cox Proportional Hazards (CoxPH). The appropriateness of each method is analyzed. Cross-validation is applied to the most appropriate CoxPH models to select the best among them.

Methods

1. Exploratory data analysis (EDA)

The data set contains 1878 unique patients, each with a single sampling time. Each line of data contains the following variables: *id*, *status*, *time*, *age, sex, wmi* (wall motion index)*, chf* (clinical heart failure), *diabetes*, and *vf* (ventricular fibrillation). There is no missing or incoherent data.

The *status* variable refers to survival status at the time of measurement: 0 = alive, 9 = dead from MI, 7 & 8 = dead from other causes. The *time* (linked to status) and *age* variables are continuous and measured in years. Sex is a binary variable where 1 is female and 0 is male. The *wmi* variable is a measure of heart pumping effect based on ultrasound measurements where 2 is normal and 0 is the worst. All other variables are binary: 1= present and 0 = absent.

The continuous variables of *age*, *time*, and *wmi* each have different distributions (*figure 1*). There is a bias in the age of patients due to few people living above 80 in the general population and few people suffering from heart disease before 50 years. There are also very few exceptionally healthy patients, as viewed through *wmi*, which is expected. The *time* variable represents the time at which the status is measured (either death or survival). It is non-linear, as in most survival data.

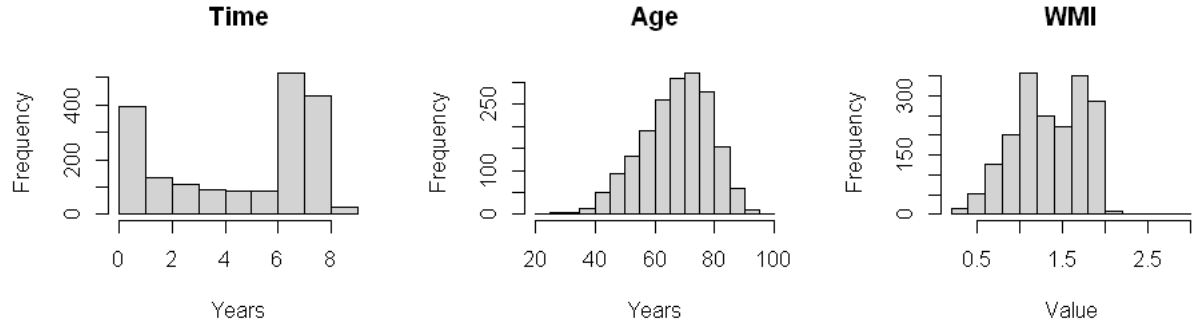


Figure 1: *Histograms of continuous variable*

Among the categorical variables, only *status* and *chf* have equal representation between groups. The survival fraction at the end of the study is just under 50%. There are approximately twice as many females as males, which may or may not represent a sampling bias. The variables *diabetes and vf* have a low frequency, which makes the predictions for these populations more difficult. See *figure 2* for details.

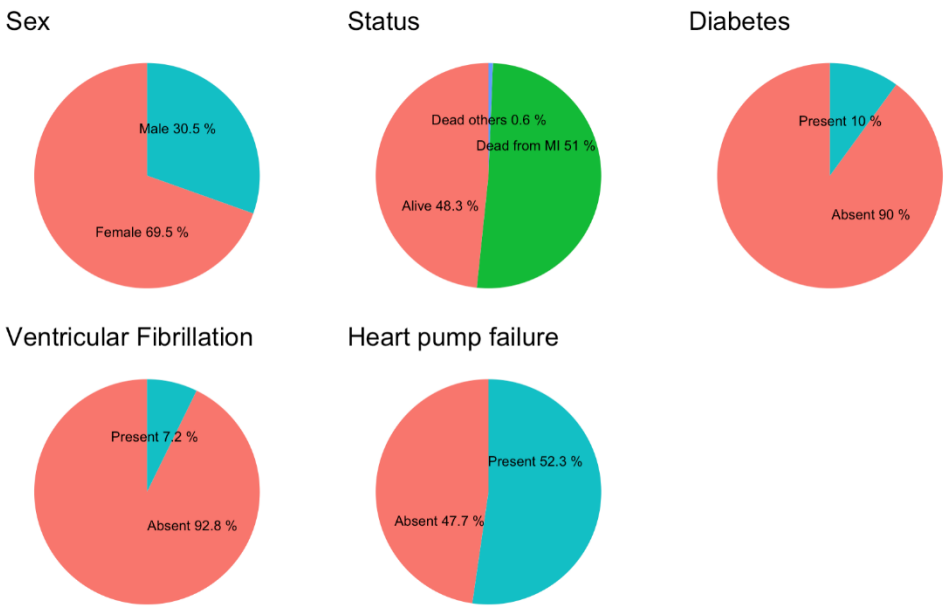


Figure 2: *Pie charts of categorical variables*

2. Kaplan Meier

The Kaplan-Meier (KM) estimator is useful for estimating survival functions. Nevertheless, this estimator assumes non-informative censoring, and identical distribution of samples, and independence of censoring from survival time[(2)]. While the Kaplan-Meier estimator allows for analysis of multiple groups it has limitations including its inability to account for covariates, handle competing risks, and potential unreliability in small sample sizes. The log-rank test is used to examine whether the groups are statistically different; a low p-value (<0.05) indicates this. A visual analysis of the KM survival curves is used to check for crossed paths, which would provide clear proof of non-proportionality. As such, the only way to account for other variables is to split the population into groups representing all the possible combinations. In the case of the Myocardial Infarction data, this represents many groups and is not manageable. This approach would create groups with few patients and the results would be difficult to interpret in some of the groups obtained.

3. Cox PH (Semi-Parametric Regression)

3.1. Details

The Cox Proportional Hazards (Cox PH) Model is applied to the myocardial infarction data to avoid the constraints of the Kaplan-Meier method. It has the benefit of being able to use multiple variables simultaneously, treat both categorical and continuous variables, and handle right censoring. The Cox Proportional Hazards Model applies four main hypotheses:

1) The observations are independent from each other
2) Censoring is independent from survival
3) Hazard functions are proportional between the strata
4) The hazard ratios are constant (independent of time)

This dataset contains right-censored data (status=1) and competing risks data (status=7&8). For the Cox model the response variable must be binary so they could either be merged with the right censored data or excluded. Given that there are only 11 observations with competing risk data (<1% observation), they are excluded them from this analyze.

3.2. Variable selection and transformations

Starting with the model that includes all variables, apply various methods are applied to identify which variables should be retained and which may require transformation. The selection of variables is guided by the p-values from the Cox model summary, which indicates the statistical significance of each variable within the model.

To assess the proportional hazards assumption, Schoenfeld residuals are analyzed. The null hypothesis for this test states that the residuals are independent of time and centered around a zero-mean. A low p-value in this context suggests that the proportional hazards assumption is violated, indicating that the effect of a variable may change over time. Ideally, the residuals should show no significant pattern over time, confirming the stability of the hazard ratio.

The linearity of continuous variables is also assessed. Non-linear relationships can lead to non-constant regression coefficients, which could violate the model's assumptions. Transformations, such as logarithms or powers, may be applied to achieve linearity.

Based on these analyses, different models are tested to refine the Cox proportional hazards model, ensuring it is both robust and adheres to the necessary assumptions for valid interpretation.

3.3. Cross validation of CoxPH models

Survival analyses often suffer from small data sets. This makes traditional cross validation difficult, since even fewer data points are present in the test set. A typical cross validation generates 10 folds, with only 1 used as the test set. The myocardial infarction data set contains a single data point for 1878 patients, meaning that a test set would include only 188 patients for the 8.5-year trail period. Since small test sets can lead to instability in the traditional approach, Verweij and Van Houwelingen (1993) proposed an alternative approach for the calculation of the cross-validation error [(1)]:

$$CVE = -2\sum_{k=1}^{K}(L_{all}(\beta_{train}) - L_{train}(\beta_{train})),$$

where the partial likelihood of the training dataset is subtracted from the partial likelihood of the entire dataset, with both calculations using the regression coefficients (betas) from the training dataset.

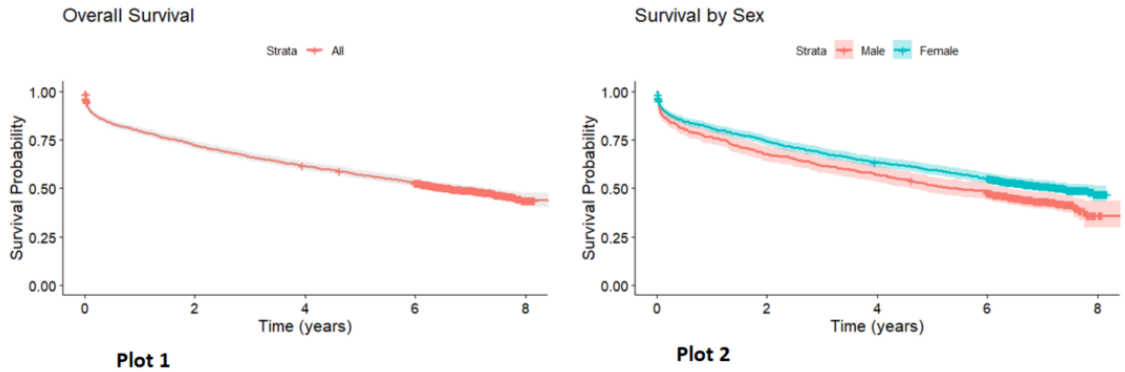
Cross-validation is used to compare CoxPH models once a final set of candidates is chosen. In effect, proper models must not violate the hypothesis of the Cox PH model and are useless to cross-validate; selection of strata and modeling of non-linear time dependance are necessary first steps.

Results

1. Kaplan-Meier

1.1. Single categorical variables

A comprehensive explanation of how different factors and their interactions influence survival probabilities (as visualized through the Kaplan-Meier plots in the images) are presented below. Each plot reveals important insights into the risk factors affecting patient survival, highlighting the necessity of considering multiple variables in survival analysis.



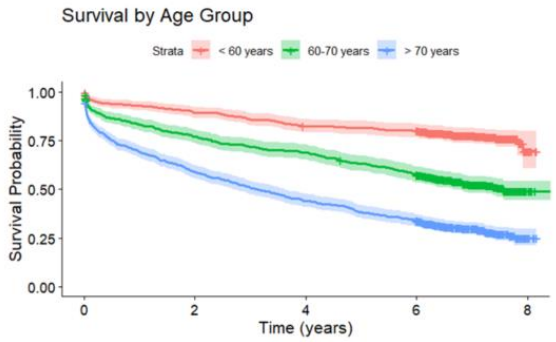
Plot 1 **Plot 2**

The Kaplan-Meier survival curve in Plot 1 illustrates the overall survival probability for the patient population. After an initial phase where the death rate (slope) is fast, the curve reaches a slower, almost linear phase. Towards the end, the death rate increases slightly and then flattens. The median survival time is approximately 6.64 years, meaning that half of the patients are expected to survive beyond this point. The confidence interval around the curve (represented by the shaded area) provides a measure of uncertainty around the estimated survival probabilities, which remain relatively stable over the observed period.

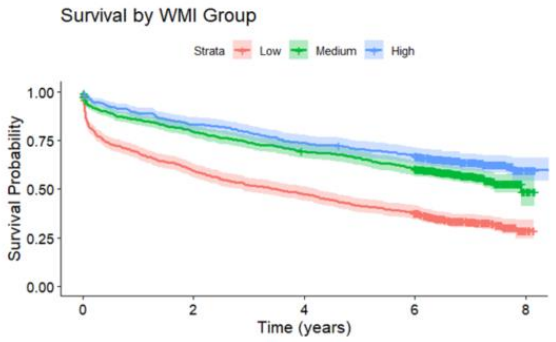
In Plot 2, the curves are stratified by *sex*, showing a difference between males and females. It can thus be concluded that males exhibit a lower survival probability compared to females throughout the

study period. The divergence of the curves is clear, with males facing a higher risk of mortality. This is statistically confirmed by the log-rank test (p = 5e-04), showing that *sex* is an important factor that influences survival outcomes, with females generally having a survival advantage over males.

Plot 3 shows the survival by age group for three age categories: under 60 years, 60-70 years, and over 70 years. The analysis reveals that the older age, precisely those over 70, have lower survival probabilities. The youngest group (<60 years) shows the highest survival rate, with a more gentle decline in the survival curve. The log-rank test (p= <2e-16) results gives the importance of age as a determinant of survival, with older age being associated with a higher risk of mortality.



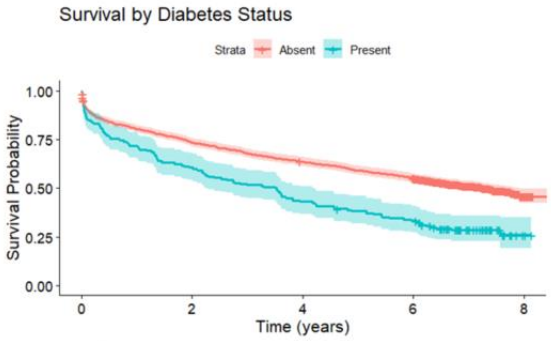
Plot 3



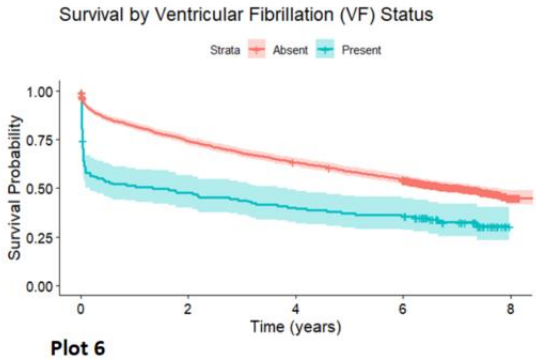
Plot 4

Plot 4 presents the survival probabilities based on Wall Motion Index (WMI) groups, classified into low, medium, and high. These groups are divided almost equally into 33% fractions, giving cutoffs of ≤ 1.2 for low and ≥ 1.8 for high. The steeply declining survival curve indicate that patients with a lower WMI face a much higher risk of death. In contrast, those with a higher WMI show better survival outcomes. The significant difference across these groups is highlighted by the log-rank test (p= <2e-16), which confirms that WMI is a strong predictor of survival, with better cardiac function correlating with improved survival chances.

In Plot 5 the Kaplan-Meier curves demonstrate that diabetic patients have significantly lower survival probabilities over time compared to non-diabetic patients. This separation of the curves reflects the detrimental impact of diabetes on survival, which is statistically validated by the log-rank test (p = 2e-10).



Plot 5

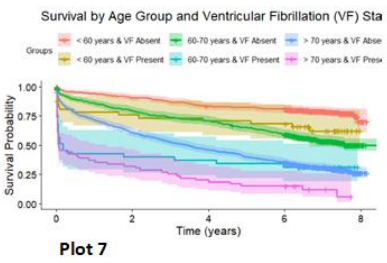


Plot 6

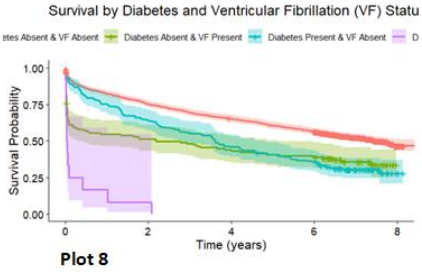
Plot 6 survival curves clearly show that patients with VF have a much lower probability of survival compared to those without the condition. This is seen from the rapid decline in the survival curve for VF patients. The significant log-rank test result (p = 7e-11) further supports the fact that VF is a critical factor in determining survival outcomes.

1.2. *Cross Group Analysis*

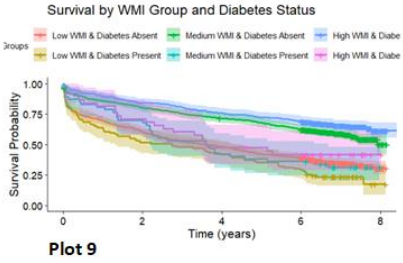
The Kaplan-Meier (KM) estimator is applied to different combinations of groups to determine if there is a KM model that is appropriate. As can be seen, the survival curves in this section cross each other. This shows a violation of proportional hazards (constant hazard rations over time) and is complicated to interpret. The size of some resulting groups is also very small, in particular in the case of *diabetes + vf* (notice the large confidence interval). This indicates in that an alternative statistical method may be needed to analyze the data appropriately. Methods that allow for non-proportional hazards, such as Cox models.



Plot 7



Plot 8



Plot 9

2. *Cox Proportional Hazards*

2.1. *Variable selection*

The first step in selecting an appropriate CoxPH model is to run it with all the initial features. The first conclusion is that *sex* isn't relevant, so it's removed from the model. This is in contradiction to the Kaplan-Meier estimator, which may be a result of including more features. The retained variables are thus *age*, *wmi*, *vf*, *diabetes* and *chf*. Schoenfeld and Martingale residuals are used to assess whether the model is acceptable.

The Schoenfeld residuals show a strong violation of the hypothesis of proportionality assumption across all features, particularly with *vf*. This result confirms the need for transforming variables or other approaches to capture the time effects of *wmi*, *chf*, *vf* and *diabetes*.

Risk

	chisq	df	p
chf	8.5523895	1	3.450690e-03
vf	40.5392113	1	1.927097e-10
age	0.8635397	1	3.527500e-01
wmi	8.8932976	1	2.862193e-03
diabetes	3.3338519	1	6.786776e-02
GLOBAL	54.8433043	5	1.405934e-10

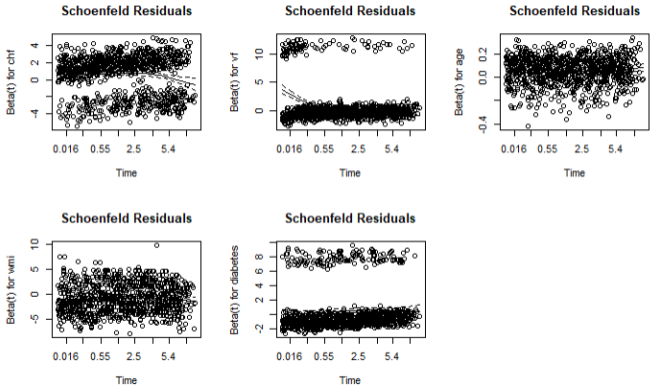


Figure 3: *CoxPH with all base variables except sex*

The Martingale residuals show a proportionality issue with *age* and *wmi*. Thus, a transformation is applied on *age* (power 2.6) and *wmi* (natural logarithm) to linearize them.

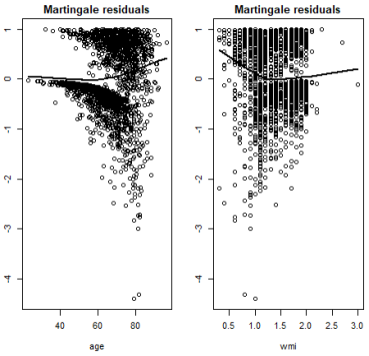


Figure 4a: *Martingale residuals for age & wmi*

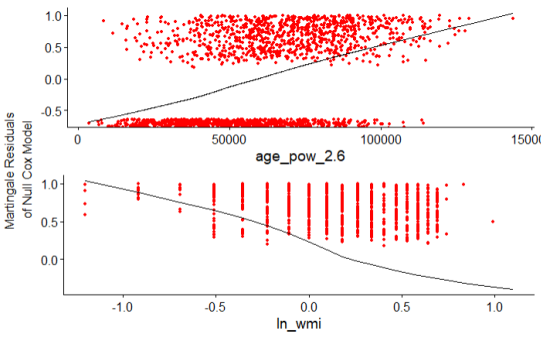


Figure 4b: *Martingale residuals for linearized age & wmi*

2.2. *Interaction terms*

Some features show significative interactions but due to the issue of non-proportionality of the risk with the concerned covariate, they are not directly developed this model. This aspect is addressed through strata.

2.3. *Non proportionality of the risk*

There are 4 features that show issue of proportionality of the risk: *diabetes*, *vf*, *chf*, and *wmi* (without transformation). To address this, different approaches are attempted:

- Adding a parametric function (xlog(t+1)) to the feature using the integrated time transformation (tt) option of the coxph function
- Adding of a time interaction (sqrt(t), t..) to the features
- Stratification
- Truncation

2.4. *Adding of a parametric function using integrated tt option of coxph*

In order to evaluate the overall effect and how its effect may evolve over time for time dependent features (*vf*, *diabetes* and *chf*), a log(t+1) function is applied to the variable. The model is run with these time transformations and all the base variables. In this way, both linear and non-linear aspects of these variables may taken into account. As the cox.zph function does not have a built-in tt option, the Schoenfeld residual is manually plotted.

The different results show that *diabetes* is not relevant and the residuals of *chf* and tt(*chf*) do not seem constant. The other variables are kept and the resulting model is *age_pow_2.6 + ln_wmi + vf + tt(vf) + tt(diabetes)*.

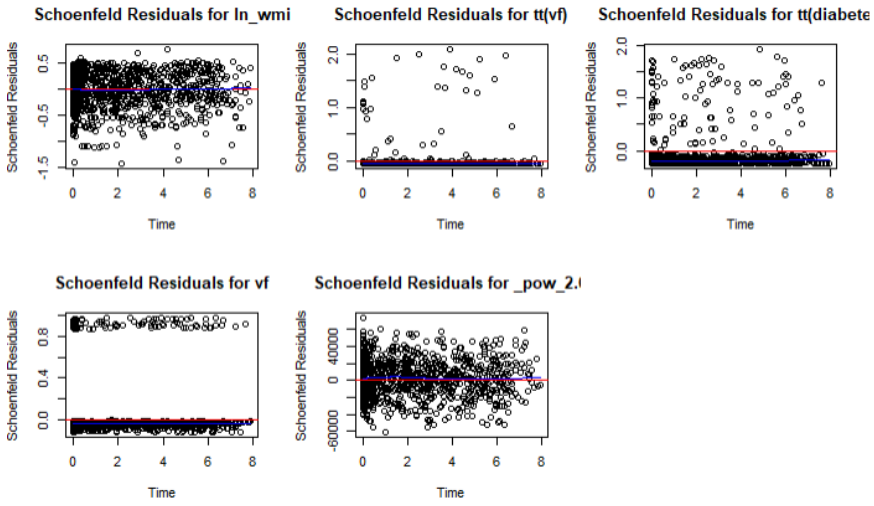


Figure 5: *CoxPH with age_pow_2.6 + ln_wmi + vf + tt(vf) + tt(diabetes)*

2.5. *Adding of a time interaction*

Different functions for time interactions are examined, for example *vf* * sqrt(*time*) or *wmi* * log(*time*), to capture the time interaction but none of them give a good result. Thus, this approach is not developed further. *The code is available in the R file.*

2.6. *Stratification*

Observed previously, *wmi* does not respect the hypothesis of proportionality but its transformed version does not. Therefore, both strata(*wmi*) and ln(*wmi*) are tested. For the other time-dependent features, given their interaction with *wmi*, several combinations of stratification are examined (both individual and combined). Since the effect of the stratified variable cannot be directly estimated, all models provide a correct fit (based on concordance results) and are considered acceptable with regards to the model hypotheses. They are compared using AIC and cross-validation. An example is presented in figure 6.

Cox PH

```
                coef  exp(coef)   se(coef)      z Pr(>|z|)
age_pow_2.6  2.455e-05  1.000e+00  1.513e-06  16.23   <2e-16 ***
ln_wmi      -9.951e-01  3.697e-01  9.706e-02 -10.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Concordance= 0.678  (se = 0.01 )
Likelihood ratio test= 367.9  on 2 df,   p=<2e-16
Wald test            = 364.5  on 2 df,   p=<2e-16
Score (logrank) test = 375.5  on 2 df,   p=<2e-16
```

Risk

```
                chisq df          p
age_pow_2.6 1.494801  1 0.22147308
ln_wmi      2.797400  1 0.09441731
GLOBAL      4.239394  2 0.12006798
```

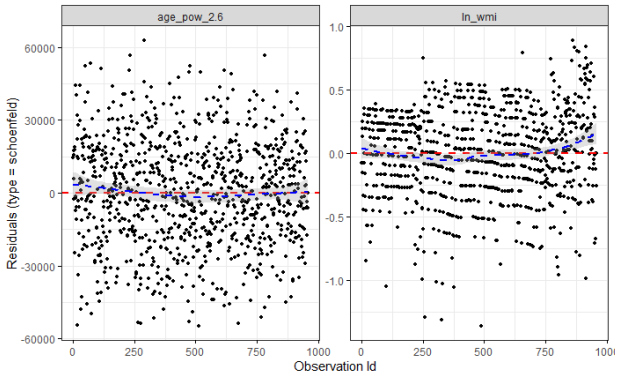


Figure 6: *CoxPH with age_pow_2.6 + ln_wmi + strata(diabetes,vf) + strata(chf)*

2.7. *Truncation*

The survival curve is plotted to visually identify the truncation period. It shows a large decrease in the first 1.8 years followed by a linear trend until 6.65 years. A new *status_truncated* variable is created, with values outside the interval [1.8, 6.65] set to 0, and values within the interval retained.

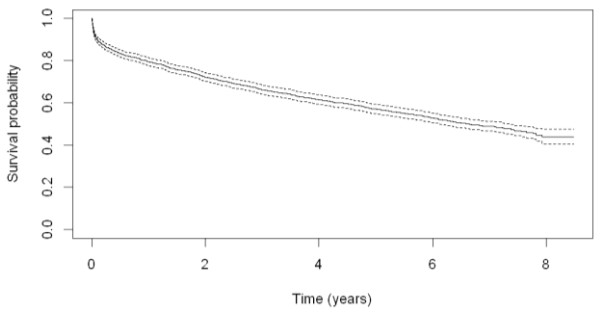


Figure 7a: *General survival curve for all patients*

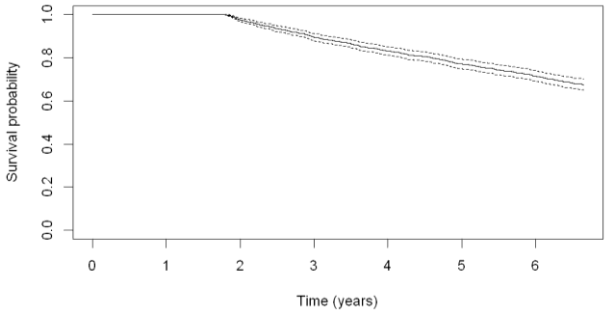


Figure 7b: *Survival Probability for truncated data*

With this truncated data set, 5 features are significant: *sex*, *chf*, *age, diabetes*, and *wmi*. This model does not exhibit a proportionality violation.

Cox PH coefficients

```
              coef exp(coef)  se(coef)      z Pr(>|z|)
sex       0.199150  1.220364  0.106135  1.876   0.0606 .
chf       0.482619  1.620312  0.105057  4.594 4.35e-06 ***
age       0.059838  1.061665  0.005434 11.011  < 2e-16 ***
wmi      -0.604011  0.546615  0.131453 -4.595 4.33e-06 ***
diabetes  0.587062  1.798697  0.142686  4.114 3.88e-05 ***

Concordance= 0.719  (se = 0.012 )
Likelihood ratio test= 275  on 5 df,   p=<2e-16
Wald test            = 254.4  on 5 df,   p=<2e-16
Score (logrank) test = 267.7  on 5 df,   p=<2e-16
```

Risk proportionality

```
              chisq df          p
sex       0.673441288  1 0.4118544
vf        0.009641083  1 0.9217822
chf       2.332337183  1 0.1267115
age       0.016097299  1 0.8990392
wmi       0.626871351  1 0.4285053
diabetes  0.434849924  1 0.5096187
GLOBAL    5.353554008  6 0.4993303
```

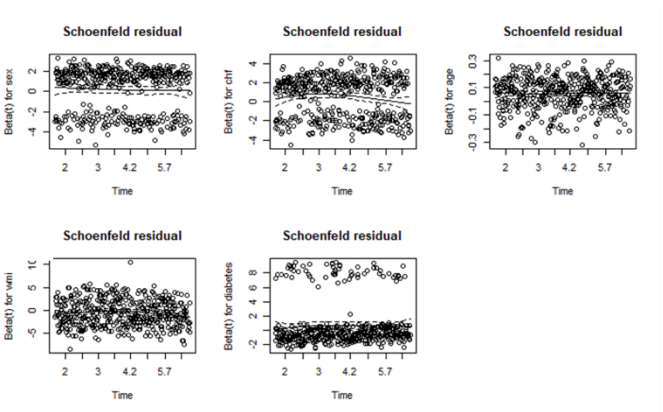


Figure 8: *CoxPH of truncated data*

2.8. *AIC and Cross Validation comparison*

A total of 7 models are compared via AIC and cross validation: a model with the time transformation function (tt), 5 models using stratification, and the truncated model.

Model	Variables or coxph model	AIC	CV error	Std deviation
Truncated Model				
1	sex + chf + age + diabetes + wmi	5951	6790	58.6
Stratification Models				
2	strata(chf) + age_pow_2.6 + ln_wmi + strata(diabetes) + strata(vf)	10577	12382	69.2
3	age_pow_2.6 + ln_wmi + strata(diabetes,vf) + strata(chf)	10577	12383	66.1
4	strata(chf) + age_pow_2.6 + strata(wmi) + strata(diabetes) + strata(vf)	5555	7189	45.4
5	strata(chf) + age_pow_2.6 + strata(wmi,diabetes) + strata(vf)	5555	7192	39.0
6	strata(chf) + age_pow_2.6 + strata(wmi,diabetes,vf)	5555	7190	43.2
Time Transformation (tt) Model				
7	age_pow_2.6 + ln_wmi + vf + tt(vf) + tt(diabetes)	13056	14875	94.1

Table 1: *AIC & Cross-validation results for best CoxPH models*

The model with the truncated data (model 1) from years 1.8 to 6.65 has the second lowest AIC and the lowest CV error and shows the best fit to the available data.

The models with strata shows the second lowest AIC and are the most suitable for the whole period. They demonstrate the best predictive performance and will be preferred to explain. Their results with cross validation and standardization error are close. Model 5 could be the best choice due to its stability (especially with difference cross-validation error not significant with others strata models).

Conclusion

Kaplan-Meier gives insights, but is not adapted to the Myocardial Infarction data due to the large number of groups, small group size, inability to handle covariates and continuous variables, non-proportionality of hazards and nonlinearity of variable effect through time.

The Cox Proportional Hazards model is shown to be well adapted to the data, with the hypotheses respected after transformations of the features. Applying strata is essential to this result, as they allow the hazard proportionality constraint to be avoided. Linearization of age is also proven useful. Time transformations are not particularly useful here. There is no single best model since models 4, 5, and 6 all perform equally well.

The truncated model has a good predictive result, but cannot be used on the full data set. It is shown that non-proportionalities essentially come from the extremes in time and that strata are thus no longer necessary in this time period. Considering that several non-truncated CoxPH models perform well in comparison (CV error), the truncated model is not the best overall model.

An accelerated failure time model is not presented in this paper. It may prove better than CoxPH, as it is capable of handling non-proportional hazards and is more flexible.

Link to code

https://github.com/awieber-france/Survival-Analysis-for-Myocardial-Infarction

References

(1) Dai, B., Breheny, P. (2019). Cross Validation Approaches for Penalized Cox Regression. *AirXiv*. https://arxiv.org/pdf/1905.10432.

(2) Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.2307/2281868.