Modeling Risk in HIV Care: A Data-Driven Approach to Patient Profiling

M. Filomena Teodoro[1,2(✉)] and Teresa Oliveira[3,4]

[1] CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, 2810-001 Almada, Portugal

[2] CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, 1048-001 Lisboa, Portugal
mteodoro64@gmail.com, maria.teodoro@tecnico.ulisboa.pt

[3] CEAUL, Faculdade de Ciências, Universidade de Lisboa, 1649-014 Lisboa, Portugal
Teresa.Oliveira@uab.pt

[4] Universidade Aberta, Departamento de Ciências e Tecnologia, LE@D – Laboratório de Educação a Distância e eLearning, Lisboa, Portugal

Abstract. A statistical framework is developed to analyze and model the censored survival times of one hundred HIV-positive individuals. Censored data refers to observations where the exact value of the variable of interest (typically time until an event) is partially known. This occurs when the event (e.g., death, relapse, failure) has not yet happened for some individuals during the study period, or its occurrence is not observed for other reasons. As the observation period is limited, many lifetimes are right-censored, requiring appropriate statistical treatment. The analysis begins with the estimation of the survival function through nonparametric methods, providing an initial overview of T, the distribution of lifetime. In a second phase, the authors investigate how individual characteristics, such as age and type of medication, affect survival, employing the Cox proportional hazards model to quantify these relationships. The results contribute to a deeper understanding of the factors influencing life expectancy in HIV patients and offer valuable insights for public health planning.

Keywords: survival analysis · hazards · computational methods · AIDS

1 Introduction

The survival analysis has developed significantly since the last two decades of last century, mainly applied to medicine [1–5]. In literature, we can find a wide number of applications of survival models, namely in engineering [6], ecology [7], biometrics [8], physics [6], etc. For a first approach, we can find in [2,9,10] the basics of survival analysis techniques. More recent work, using Machine learning techniques, the authors of [11] explored deep Cox mixtures for survival regression. In [6] it is applied the censored expectation maximization algorithm for mixtures to intertrade waiting times. Also, the authors of [12] talk about interference and sensitivity analysis. In [13], the authors outcome risk model development for heterogeneity of treatment effect analysis, performing a comparison of non-parametric machine learning methods and semi-parametric

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 99–114, 2027.
https://doi.org/10.1007/978-3-032-30521-3_7

statistical methods. In the study proposed in [14], a comparison of the Cox model with the Fine-Gray model to take competing risks into account for survival analyzes of re-fracture rates is performed.

To evaluate some intervention strategies, such as predicting outbreaks or understanding avian influenza dynamics, Huang and their collaborators have searched and organized literature [15] classifying mathematical, statistical, and machine learning-based models as a tool to play a crucial role in risk assessment, understand transmission and predict outbreak, allowing to reduce the mortality index of birds and mamals, to improve the food security, reduce the rate transmission, and so on. In [16], a systematic review of statistical methods to evaluate the performance of survival prediction models from conventional to deep learning approaches.

The main objective of this work is to model the life time T of a group of one hundred of individuals with AIDS. The provided data, like most cases in survival analysis, are censored (right censored) since the study is usually carried out over a fixed period of time (or a fixed number of deaths is reached).

This work is divided in five sections. An introduction starts the manuscript; Sect. 2 details some aspects about survival analysis. Follows the description of the applied methodology in Sect. 3. In Sect. 4 some issues about the empirical application are described, starting with the basic statistical techniques, including the evaluation of some tests about the normality of data, followed by the construction and interpretation of the results. In Sect. 5 takes place some discussion and it is issued some final remarks.

2 Survival Analysis Some Concepts

In the present work, the author intend to determine and investigate the lifetime (T) distribution of one hundred individuals with HIV. The available data, similarly with the majority of cases in survival analysis, is censoured (on right) once the study is usually performed during a certain period or until there exists a certain number of deaths.

2.1 Survival Function

In this section we present some basic concepts of survival analysis [4,5,7,17–19].

A survival function $S(t)$ is defined as the individual surviving at an instant $t > 0$, and it is being given by

$$S(t) = 1 - F(t) = P(T > t),\ t > 0.$$

The failure rate function or hazard, $h(t)$, is defined as the probability of an individual dying in the next instant knowing that he survived until $t > 0$, and it is given by:

$$h(t) = \frac{f(t)}{S(t)}.$$

The cumulative hazard function $H(t)$ is defined as

$$H(t) = \int_0^t h(t)dt = -\log S(t).$$

It is important to understand whether the data to be studied correspond to moments of death or censorship, given that can be considered different approaches. At an early stage it is important to sketch some estimate of the survival function. Only after a qualitative analysis should models be estimated according to Cox (proportional hazards) or others.

2.2 Survival Function of Kaplan-Meier

The estimate of the nonparametric survival function, for censored data, called Kaplan-Meier, is given by

$$\hat{S}(t) = \prod_{i:t_i<t} \left(\frac{n_i - d_i}{n_i} \right),$$

with

- n_i the number individuals in risk immediately before t_i,
- d_i the number dead in t_i,
- l_i the number individuals censoured in $[t_i, t_{i+1}[$,

being $n_i = n - (d_1 + \cdots + d_{i-1}) - (l_0 + \cdots + l_{i-1})$.

One can also construct confidence intervals for $S(t)$ knowing the estimate of its variance.

$$\hat{S}(t) \pm 1.96\sqrt{\widehat{var}\left(\hat{S}(t)\right)}.$$

In this work, the confidence intervals obtained in some survival function plots used the Greenwood estimate for the variance.

2.3 Cox Proportional Hazard Models

The Cox proportional hazards model is one of the classes of regression models. These are models with proportional failure rate (hazard proportional). In these models it is assumed that given two individuals, with associated covariate vectors $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})$, with $i = 1, 2$, the ratio $h(t, \mathbf{z}_1)/h(t, \mathbf{z}_2)$ does not depend on t. With this premise the failure rate function T, given z, can be written in the form

$$h(t, \mathbf{z}) = h_0(t) g(\mathbf{z}).$$

A possible structure for $g(\mathbf{z})$ is given by

$$g(\mathbf{z}) = \exp(\beta^T \mathbf{z})$$

ensuring positivity. Note that the vector β^T measures the effect of the covariates. Each coefficient β_i associated with each covariate is often called the risk score associated with the variable z_i of $\mathbf{z}$. One question that can be asked is: what is the structure of the survival function for this model? It can be deduced obtaining:

$$S(t,\mathbf{z}) = \underbrace{S_0(t)}_{\text{associated to } h_0(t)}^{g(\mathbf{z})}$$

Coming back to the failure rate, how to estimate it? In the Cox model, it is non-parametric, obtaining a partially parametric model due to the influence of the covariates. Cox found a function that behaves like the likelihood function, not depending on the nonparametric $h_0((t)$. This function is given by

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\left(\mathbf{z}_i^T \beta\right)}{\sum_{l \in R_i} \exp\left(\mathbf{z}_l^T \beta\right)},$$

where

- $R_i = R_I(t)$ is the set of indices of individuals at risk at instant t,
- k the number of deaths in $t_1 < \cdots < t_k$,
- n the number of individuals under study.

A generalization of the Cox model is to admit the existence of non-proportional hazards between strata $h(t,\mathbf{z}) = h(t_i)g(\mathbf{z})$.

As with any regression model, we have all the inherent model validation steps: variable selection, model comparison, residual analysis (identification of outliers and influential observations), verification of Cox proportional hazards hypotheses, etc. Note that the tests used in linear models are again used in this context, highlighting t-ratios, likelihood ratio, etc.

Since we are estimating several survival functions, it is important to mention the existence of comparison tests between the different functions. The test used for this purpose is the Mantel-Haenzel-Savage (log-rank) test which, despite adding non-independent frequency tables, is quite powerful when the failure rate functions are approximately proportional.

3 Empirical Approach

3.1 Data

The data used in this article are reported to the first decade of this century, concern AIDS patients, a total of 100, and include the following variables:

- **ID**- identification code,
- **Time** - survival time (in months),
- **Age** - age in years,
- **Drug** - intravenous use (0 = “no”, 1 = “yes”),
- **Status** - 1 = “death due to AIDS”, 0 = “Obs. Censored”.

3.2 Survival Function of Kaplan-Meier

In the initial step we obtained the Kaplan-Meier estimate considering the censoured data, using the SPLUS software.

```
> KM<-survfit(Surv(Tempo, Status), conf.type="plain",
 type="kaplan-meier",data = Dados2)
> plot(KM)
```

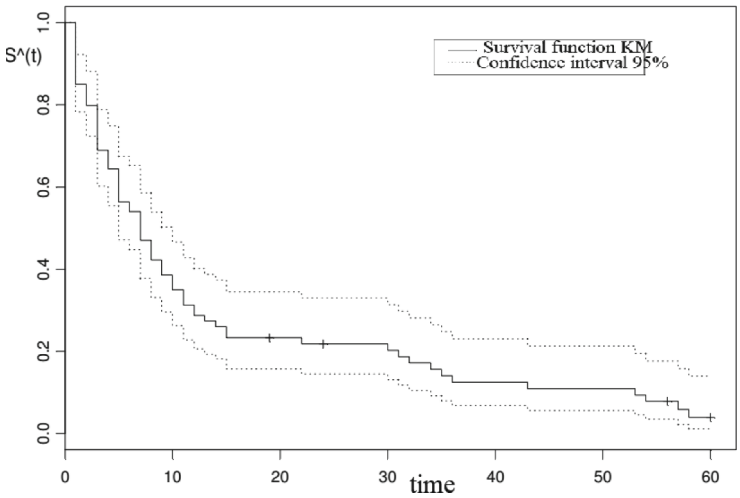

Fig. 1. Empirical Survival graph. Kaplan-Meier estimate.

Figure 1 represents the empirical survival curve using the Kaplan-Meier estimate due the fact that it is decreasing with concavity facing upwards it is similar to the survival function of the exponential distribution.

A more detailed analysis can be done by determining the Kaplan-Meier estimate, disaggregating by type of drug (intravenous - no, intravenous - yes).

```
   >KMdroga<-survfit(Surv(Tempo, Status) ~as.factor(Droga),conf.type="plain"
  ,type="kaplan-meier",data = Dados2)
  > plot(KMdroga, lty=2:3, xlab="T", ylab="Prob.  de sobrev." )
  > legend(30, 0.80, legend=c(" intra-venosa-nao"," intra-venosa-sim"
), lty=2:3)
```

Figure 2 indicates distinct survival empirical functions per kind of drug. Besides both curves be similar to exponential distribution (concavity facing upward), the curve associate to "*no intravenous* drug evidences a longest time of survival. This fact evidences that drug kind hall be considered explanatory variable in the model. Other variables in the data were age and gender, they were also considered as covariates in the model to estimate.

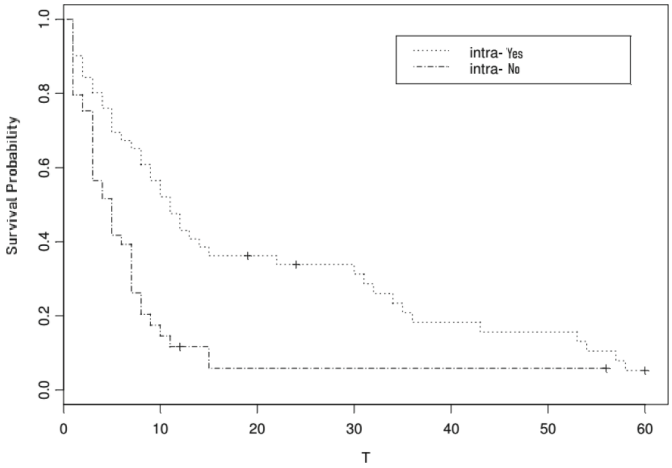

Fig. 2. Empirical Survival graph per level of drug (intravenous - yes, intravenous - no). Kaplan-Meier estimate.

3.3 Model 1

It was estimated the model of Cox Proportional Hazards, including all possible information as covariates, the type of drug and age respectively. The output reveals good indicators, the "*t-ratios*" associated to covariates coefficients are significant, meaning that this explanatory variables shall be considered. Other indication in the same way are the confidence intervals obtained for the exponential of each coefficient: none contains the value one. The confidence interval [20] for the age is farm from of 1:

```
      *** Cox Proportional Hazards ***
Call:
coxph(formula = Surv(Tempo, Status) ~ Idade + Droga, data = Dados2)

  n= 100

        coef exp(coef) se(coef)     z        p
Idade 0.0971      1.10   0.0186 5.21 1.9e-007
Droga 1.0167      2.76   0.2562 3.97 7.2e-005

      exp(coef) exp(-coef) lower .95 upper .95
Idade      1.10      0.907      1.06      1.14
Droga      2.76      0.362      1.67      4.57

Rsquare= 0.324   (max possible= 0.997 )
Likelihood ratio test= 39.1  on 2 df,   p=3.18e-009
Wald test            = 36.1  on 2 df,   p=1.43e-008
Score (logrank) test = 38.4  on 2 df,   p=4.6e-009
```

$$\begin{aligned} \exp(\beta_{\text{age}}) &: [1.06, 1.14], \\ \exp(\beta_{\text{drug}}) &: [1.67, 4.65]. \end{aligned}$$

The tests that allow to compare the estimated model and the model with explanatory variables, namely the likelihood ratio test, the wald test and the logrank test lead to largely significant statistical values.

The estimate of the survival function (model 1) is given by the following code:

```
> summary(survfit(cox.fit1))
Call: survfit.coxph(object = cox.fit1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    100      15 0.885870 0.02969    0.8295436       0.9460
    2     83       5 0.842263 0.03518    0.7760556       0.9141
    3     73      10 0.740745 0.04557    0.6566098       0.8357
    4     61       4 0.692477 0.04944    0.6020504       0.7965
    5     56       7 0.597255 0.05501    0.4986109       0.7154
    6     49       2 0.567512 0.05620    0.4673935       0.6891
    7     46       6 0.472812 0.05854    0.3709334       0.6027
    8     39       4 0.401388 0.05911    0.3007497       0.5357
    9     35       3 0.349514 0.05830    0.2520481       0.4847
   10     32       3 0.300375 0.05642    0.2078603       0.4341
   11     28       3 0.252605 0.05370    0.1665260       0.3832
   12     25       2 0.220606 0.05144    0.1396798       0.3484
   13     21       1 0.200386 0.05060    0.1221582       0.3287
   14     20       1 0.180206 0.04949    0.1051961       0.3087
   15     19       2 0.143205 0.04612    0.0761761       0.2692
   22     16       1 0.124105 0.04412    0.0618267       0.2491
   30     14       1 0.105184 0.04169    0.0483721       0.2287
   31     13       1 0.088282 0.03870    0.0373887       0.2085
   32     12       1 0.072868 0.03531    0.0281872       0.1884
   34     11       1 0.058942 0.03160    0.0206068       0.1686
   35     10       1 0.046215 0.02762    0.0143236       0.1491
   36      9       1 0.035166 0.02347    0.0095066       0.1301
   43      8       1 0.023454 0.01880    0.0048757       0.1128
   53      7       1 0.015072 0.01416    0.0023902       0.0950
   54      6       1 0.008552 0.00968    0.0009294       0.0787
   57      4       1 0.003895 0.00557    0.0002364       0.0642
   58      3       1 0.000872 0.00192    0.0000117       0.0649
```

The graphic of the survival function (model 1) is displayed in Fig. 3.

3.4 Model 2

The second model considers only one explanatory variable, the age.

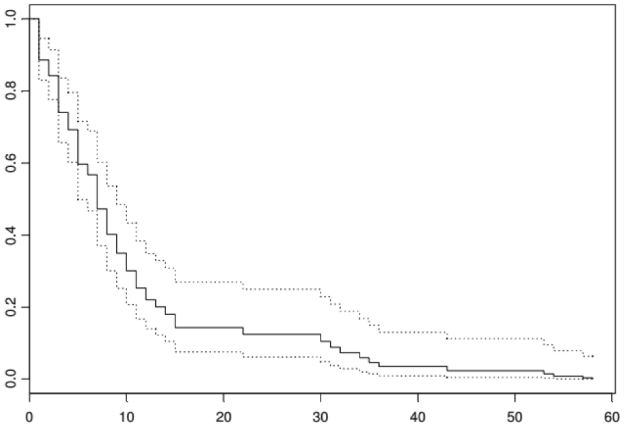

Fig. 3. Model 1-Cox Proportional Hazards. Explanatory variables. Age, Drug type.

```
> summary(survfit(cox.fit1))
Call: survfit.coxph(object = cox.fit1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    100      15 0.885870 0.02969    0.8295436       0.9460
    2     83       5 0.842263 0.03518    0.7760556       0.9141
    3     73      10 0.740745 0.04557    0.6566098       0.8357
    4     61       4 0.692477 0.04944    0.6020504       0.7965
    5     56       7 0.597255 0.05501    0.4986109       0.7154
    6     49       2 0.567512 0.05620    0.4673935       0.6891
    7     46       6 0.472812 0.05854    0.3709334       0.6027
    8     39       4 0.401388 0.05911    0.3007497       0.5357
    9     35       3 0.349514 0.05830    0.2520481       0.4847
   10     32       3 0.300375 0.05642    0.2078603       0.4341
   11     28       3 0.252605 0.05370    0.1665260       0.3832
   12     25       2 0.220606 0.05144    0.1396798       0.3484
   13     21       1 0.200386 0.05060    0.1221582       0.3287
   14     20       1 0.180206 0.04949    0.1051961       0.3087
   15     19       2 0.143205 0.04612    0.0761761       0.2692
   22     16       1 0.124105 0.04412    0.0618267       0.2491
   30     14       1 0.105184 0.04169    0.0483721       0.2287
   31     13       1 0.088282 0.03870    0.0373887       0.2085
   32     12       1 0.072868 0.03531    0.0281872       0.1884
   34     11       1 0.058942 0.03160    0.0206068       0.1686
   35     10       1 0.046215 0.02762    0.0143236       0.1491
   36      9       1 0.035166 0.02347    0.0095066       0.1301
   43      8       1 0.023454 0.01880    0.0048757       0.1128
   53      7       1 0.015072 0.01416    0.0023902       0.0950
   54      6       1 0.008552 0.00968    0.0009294       0.0787
   57      4       1 0.003895 0.00557    0.0002364       0.0642
   58      3       1 0.000872 0.00192    0.0000117       0.0649
```

All tests were largely significant. A last test to perform is the comparison this model with model 1. To test the best model, we simply consider the null hypothesis the coefficient of the drug variable as null against all other alternatives. By the likelihood ratio

we have that

$$\Lambda_{1,2|H_0} = -2\left(\ln L\left(\widetilde{\beta}\right)_{\text{mod }2} - \ln L\left(\widehat{\beta}\right)_{\text{mod }1}\right) \overset{a}{\sim} \chi^2_{(1)}$$

The null hypothesis is rejected when

$$\Lambda_{1,2obs} > \chi^2_{(1,1-\alpha)}.$$

When comparing the two models we have There is a significant difference between the two models as expected from the analysis already done in model 1. Another model was studied. It only presented the drug variable as an explanatory variable. The conclusions are similar: model 1 should be considered.

$$\Lambda_{1,2obs} = 39.1 - 23.3 = 15.8$$
$$p - value = 1 - \chi^2_{(15.8,1)} = 7.0403 \times 10^{-5}.$$

3.5 Model 3

The next step was to estimate models for each level of the drug variable, that is, to assume the existence of a stratification factor that allows the existence of non-proportional hazards between the strata. The obtained model (model 3) was

```
> cox.2<-coxph(Surv(Tempo, Status)~Idade+strata(Droga), data=Dados2)
> summary(cox.2)
Call:
coxph(formula = Surv(Tempo, Status) ~ Idade + strata(Droga), data = Dados2)

  n= 100

        coef exp(coef) se(coef)    z        p
Idade 0.0923       1.1   0.0188 4.91 9.2e-007

      exp(coef) exp(-coef) lower .95 upper .95
Idade       1.1      0.912      1.06      1.14

Square= 0.216   (max possible= 0.992 )
Likelihood ratio test= 24.3  on 1 df,   p=8.22e-007
Wald test            = 24.1  on 1 df,   p=9.2e-007
Score (logrank) test = 24.5  on 1 df,   p=7.41e-007
```

In this model, like before, all tests are significant. To compare the survival curves in the two strata, we used the logrank test. Since the test is largely significant, there is a clear difference in survival functions between strata. There is also statistical evidence that the age variable is statistically significant.

```
survdiff(formula = Surv(Tempo, Status) ~ Idade + strata(Droga), data = Dados2)

         N Observed Expected (O-E)^2/E (O-E)^2/V
Idade=20 1        0    2.409   2.40947   3.30534
Idade=21 1        1    2.142   0.60853   0.67873
Idade=22 1        1    2.725   1.09190   1.27708
Idade=25 2        1    4.500   2.72206   3.29818
Idade=26 2        2    1.914   0.00391   0.00498
Idade=28 3        2    1.089   0.76191   0.85593
Idade=29 2        1    3.644   1.91861   2.24068
Idade=30 6        4    4.847   0.14813   0.18045
Idade=31 5        4    5.666   0.48993   0.67233
Idade=32 9        8    6.238   0.49767   0.65192
Idade=33 6        5    5.473   0.04083   0.04928
Idade=34 8        8    7.292   0.06878   0.08471
Idade=35 5        2    3.523   0.65842   0.80318
Idade=36 9        8   10.342   0.53050   0.68169
Idade=37 3        2    0.830   1.64923   1.86199
Idade=38 3        3    0.769   6.46638   7.83106
Idade=39 5        5    3.923   0.29546   0.38819
Idade=40 3        2    1.457   0.20240   0.25347
Idade=41 4        3    2.344   0.18386   0.22059
Idade=42 4        3    1.842   0.72812   0.89297
Idade=43 2        2    1.795   0.02334   0.02552
Idade=44 4        2    1.080   0.78478   0.99088
Idade=45 1        1    0.508   0.47611   0.60322
Idade=46 3        2    1.340   0.32503   0.39086
Idade=47 4        4    1.464   4.39550   4.93894
Idade=48 1        1    0.212   2.92819   3.17114
Idade=50 1        1    0.204   3.10408   3.90000
Idade=51 1        1    0.163   4.28858   4.68661
Idade=54 1        1    0.265   2.04298   2.20111

 Chisq= 54.7  on 28 degrees of freedom, p= 0.00183
```

```
> survdiff(Surv(Tempo, Status)~Droga, data=Dados2)
Call:
survdiff(formula = Surv(Tempo, Status) ~ Droga, data = Dados2)

          N Observed Expected (O-E)^2/E (O-E)^2/V
Droga=0  51       42     54.9      3.02      11.9
Droga=1  49       38     25.1      6.60      11.9

 Chisq= 11.9  on 1 degrees of freedom, p= 0.000575
```

4 Analysis of Residuals

The residuals were analyzed for both model 1 and model 3. Since these analyzes are very similar, this article only focuses on the analysis of model 3.

In the first phase, the covariate age was investigated to ensure it contributed in the correct functional form. Graphical analysis of the residues is essential. We can find in next figures some plots with interest.

From observation of Fig. 4, we tried models where the covariate age had various shapes close to line in Fig. 4: square root, polygonal line, polynomial shape. The best models obtained, despite being more complex, did not have results that were significantly different from model 3.

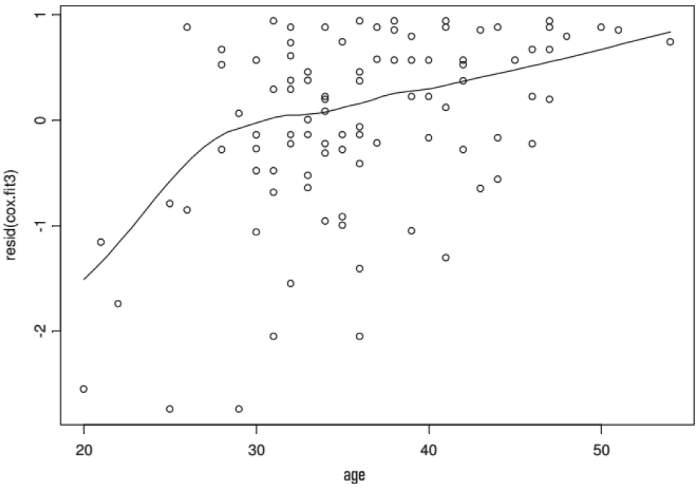

Fig. 4. Model 3. Residuals versus Age.

Also, no poorly adjusted individuals were identified (seen Fig. 5). In identifying influential observations, "dfbetas" plots were made, which give us the difference in the estimate when observation is omitted (see Fig. 6). Notice that an observation is considered influential if the absolute value of this difference is greater than 0.5.

To verify the proportional hazards assumption, a graphical analysis of the estimated values for the coefficients of the covariates is important: whether they can be considered constant or not. As many graphs are made as there are explanatory variables considered. In model 3 there is only one covariate (age), we only need to analyze a single graph. Figure 7 displays the estimates of coefficient beta of age along the time.

Estimates of the beta coefficient over time lead to confidence bands, which, through graphical analysis, appear to limit a horizontal line, of the type $\beta(t) = \beta$. The relative linearity test for beta consists of investigating whether the β estimates are correlated with time. The following output was obtained:

```
> cox.zph(cox.2)
          rho  chisq     p
Idade -0.0176 0.0223 0.881
>
```

The test of null correlation with the time conduced to $pvalue = 0.881$, meaning that there exists evidence to not reject the horizontal linearity. The model 3 (model the Cox is not rejected and we obtain the following curves of survival. By the observation of Fig. 8, it is clear that the survival curve relative to the stratum of individuals who consume intravenous drugs is lower.

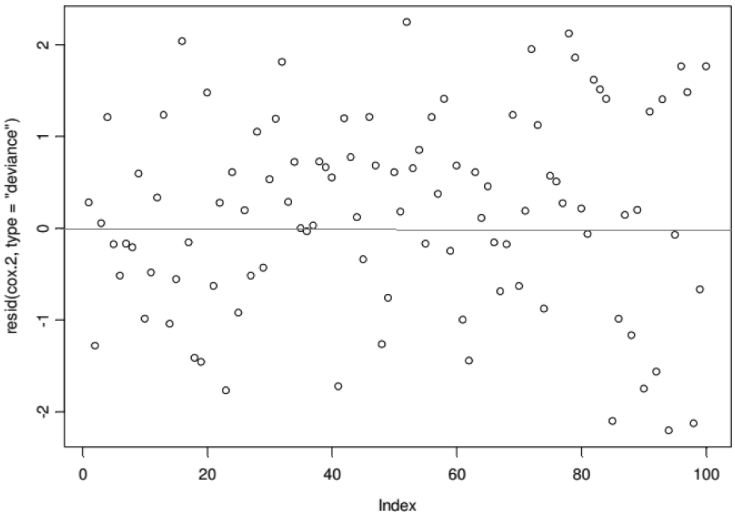

Fig. 5. Model 3. Residuals versus index.

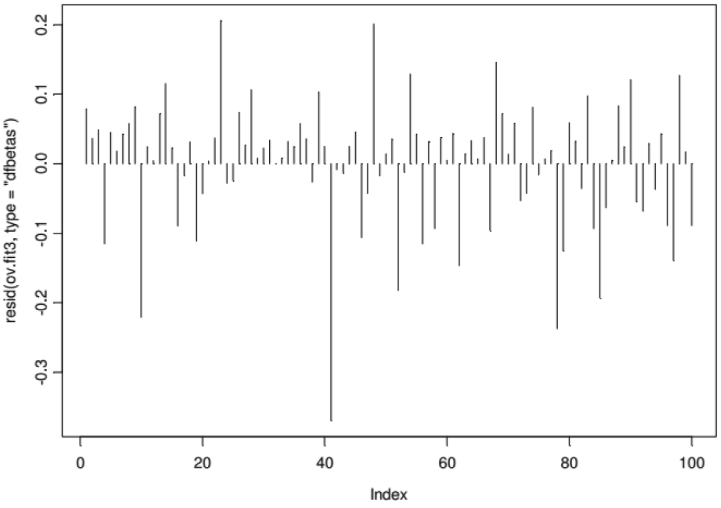

Fig. 6. Model 3. DFbetas versus index.

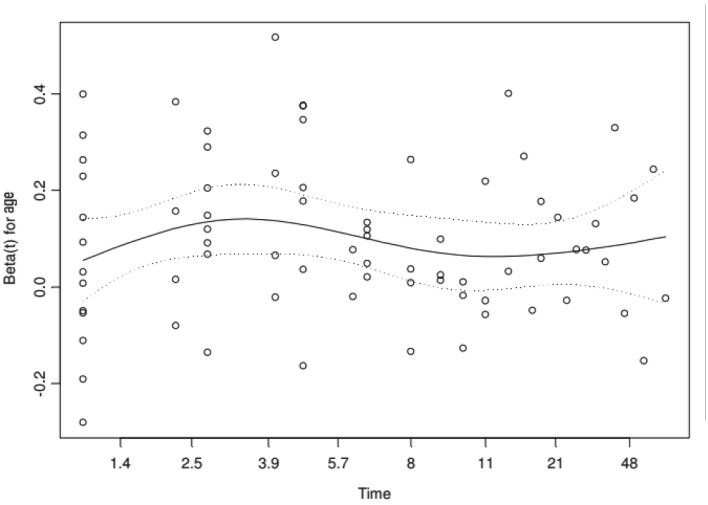

Fig. 7. Model 3. Estimates of coefficient β along the time.

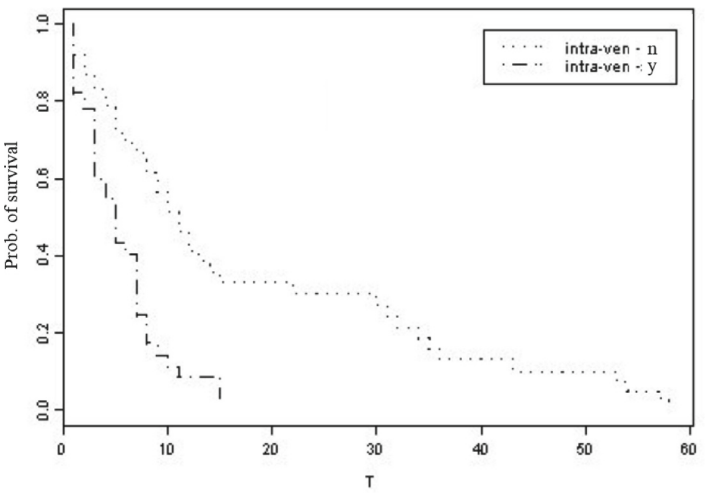

Fig. 8. Model 3. Curves of Cox survival function per drug strata.

5 Discussion and Final Remarks

Model 3 is a hazard proportional by strata depending on the drug values. The estimated model comes in the form

$$\widehat{h}(t,\ \text{age}) = \widehat{h}_{drug}(t)\exp(\widehat{\beta}\times \text{age}),\ \text{so}$$

$$\exp(\widehat{\beta}\times \text{age}) = \exp(0.0923\times \text{age}).$$

It can be concluded that the risk factor increases $exp(0.0923) = 1.1$ when age increases by 1 year, at each drug level. The 95% risk confidence interval for the age is $(1.06, 1.14)$. The survival functions obtained accordingly with model 3 clearly confirm this fact. Survival of individuals taking intravenous drugs is lower. This behavior is identical to what we already had in the empirical survival functions of Kaplan-Meier. It is clear that age is an important explanatory factor in the survival of the individual.

In model 1 the model is a hazard proportional without stratification. The estimate of the model is given by

$$\begin{aligned}\widehat{h}(t,\ \text{age,drug}) &= \widehat{h}(t)\exp(\widehat{\beta}_1\times \text{age} + \widehat{\beta}_2\times \text{drug})\\ &= \widehat{h}(t)\exp(\widehat{\beta}_1\times \text{age})\exp\left(\widehat{\beta}_2\times \text{drug}\right)\end{aligned}$$

onde

$$\begin{aligned}&\exp(\widehat{\beta}_1\times \text{age})\exp\left(\widehat{\beta}_2\times \text{drug}\right)\\ =\ &\exp(0.0971\times \text{age})\exp(1.0167\times \text{drug}).\end{aligned}$$

In model 1 we can conclude that the risk factor increases $exp(0.0971) = 1.1$ when the age increases by 1 year, when the drug variable remains the same. Note that the risk interval obtained at 95% does not contain the value 1, $(1.06, 1.14)$, despite being very close at its left end. Age is an important explanatory variable. Regarding the drug variable, if this changes from 0 to 1, the group of individuals who use intravenous drugs has a higher risk factor given by $exp(1.10167) = 2.76$. The relative risk interval obtained at 95% does not contain the value 1 and has an extreme well above the interval obtained for age: $(1.67, 4.57)$.

We highlight that both models agree on their conclusions: age is an important variable for an individual's survival, the kind of drug is an important factor for survival. In both models this explanatoru variable was highly significant. In future work, we intend to explore the same data, but knowing that database was limited, the next step would be to add more information such as gender, some familial diseases, etc. We would also like to apply other techniques that have emerged more recently, which are extensions of the methods used in this article [11, 13, 14, 21–23].

Acknowledgements. The first author was supported by Portuguese funds through the *Center of Naval Research* (CINAV), Naval Academy, Portugal and *The Portuguese Foundation for Science and Technology* (FCT), through the *Center for Computational and Stochastic Mathematics* (CEMAT), University of Lisbon, Portugal, project DOI: (https://doi.org/10.54499/UID/04621/2025). The second author was supported by LE@D – Laboratory of Distance Education and eLearning, Universidade Aberta, Portugal, funded by National Funds through the FCT, within the scope of the projects UID/4372/2025 e UID/PRR/04372/2025, and of CEAUL Research Unit, UID/00006/2025.

References

1. Gill, R.D.: Understanding cox's regression model: a martingale approach. J. Am. Stat. Assoc. **79**(386), 441–447 (1984). http://www.jstor.org/stable/2288288. Accessed 7 May 2026
2. Gill, R.D.: Lectures on Survival Analysis, pp. 115–241. Springer, Heidelberg (1994). https://doi.org/10.1007/BFb0073873. Accessed 9 May 2026
3. Lee, E., Go, O.: Survival analysis in public health research. One Health **18**, 105–134 (1997). https://doi.org/10.1146/annurev.publhealth.18.1.105. Accessed 9 May 2026
4. Harrell, F., Lee, K., Mark, D.: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. **15**, 361–387 (1996)
5. Christensen, E.: Multivariate survival analysis using cox's regression model. Hepatology **7**, 1346–1358 (1987)
6. Kreer, M., Kizilersu, A., Thomas, A.W.: Censored expectation maximization algorithm for mixtures: application to intertrade waiting times. Physica A: Stat. Mech. Appl. **587**, 126456 (2022). https://www.sciencedirect.com/science/article/pii/S0378437121007299. Accessed 9 May 2026
7. Kenneth, H.P., Scott, R.W., Christine, M.B., Paul, D.C.: Survival analysis in telemetry studies: the staggered entry design. J. Wildlife Manage. **53**(1), 7–15 (1989). http://www.jstor.org/stable/3801296
8. David, A.S.: Sample-size formula for the proportional-hazards regression model. Biometrics **39**(2), 499–503 (1983). http://www.jstor.org/stable/2531021
9. Lawless, J.: Statistical Models and Methods for Lifetime Data, 2nd edn. Wiley, New York (2003)
10. Rocha, C., Papoila, A.L.: Análise de Sobrevivência. Sociedade Portuguesa de Estatística., Lisboa (2009), https://www.spestatistica.pt/storage/app/uploads/public/600/481/dc3/600481dc32725998871922.pdf. Accessed 9 May 2026
11. Nagpal, C., Yadlowsky, S., Rostamzadeh, N., Heller, K.: Deep cox mixtures for survival regression. In: Simos, T., et al. (eds.) Proceedings of Machine Learning Research, vol. 126, pp. 1–27. AIP (2021, in press)
12. Tyler, J.V., Eric, J.T.T., Elizabeth, H.M.: Interference and sensitivity analysis. Stat. Sci. **29**(4), 687–706 (2014)
13. Xu, E., et al.: Outcome risk model development for heterogeneity of treatment effect analyses: a comparison of non-parametric machine learning methods and semi-parametric statistical methods. BMC Med. Res. Methodol. **24**(1), 1–9 (2024). https://doi.org/10.1186/s12874-024-02265-8
14. Nolan, E.K., Chen, H.-Y.: A comparison of the Cox model to the Fine-Gray model for survival analyses of re-fracture rates. Arch. Osteoporos. **15**(1), 1–8 (2020). https://doi.org/10.1007/s11657-020-00748-x

15. Huang, S., et al.: A systematic review of mathematical and machine learning models of avian influenza. One Health **21**, 2109–2123 (2025). https://www.sciencedirect.com/science/article/pii/S2352771425002393. Accessed 9 May 2026
16. Park, S., Park, J., Kim, H., Park, S.: Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). Korean J. Radiol. **22**(10), 1697–1707 (2021). https://doi.org/10.3348/kjr.2021.0223. Accessed May 9th 2026
17. Rosner, B.: Fundamentals of Biostatistics, 6th edn. Pacific Grove, Duxbury (2006)
18. Long, J.: Regression Models for Categorical and Limited Dependent Variables. Sage Publications, Thousand Oaks (1987)
19. Peduzzi, P., Concato, J., Feinstein, A., Holford, T.: Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J. Clin. Epidemiol. **48**, 1503–1510 (2004)
20. Pencina, M., D'Agostino, R.: Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat. Med. **23**(23), 2109–2123 (2004)
21. Goerdten, J., Carrière, I., Muniz-Terrera, G.: Comparison of cox proportional hazards regression and generalized cox regression models applied in dementia risk prediction. Alzheimers Dement **6**(1), e12041 (2020). http://www.jstor.org/stable/2288288. Accessed 9 May 2026
22. Ahmed, H.S.: Advanced statistical methods for hazard modeling in cardiothoracic surgery: a comprehensive review of techniques and approaches. Indian J. Thoracic Cardiovascular Surgery **40**(5), 633–644 (2020). http://www.jstor.org/stable/2288288. Accessed 9 May 2026
23. Hélio, A.M.: Survival analysis in clinical and experimental studies. J. Vasc. Bras. **16**(4), 267–269 (2017). https://doi.org/10.1590/1677-5449.001604. Accessed 8 May 2026