Survival and Longitudinal Data Analysis

APPLICATION TO ANALYSIS OF MORTALITY OF HIV PATIENTS

Compulsory Project

Jaad BELHOUARI

M2 Data Science : Santé, Finance et Assurance

December 11, 2024

Abstract

This study investigates the survival outcomes of HIV-infected patients who were either intolerant or had failed AZT (zidovudine) therapy and were subsequently treated with either didanosine (ddI) or zalcitabine (ddC). The survival data collected includes CD4 cell counts, which is a biomarker related to the disease progression, at multiple time points, patient demographics, and clinical variables. Survival analysis techniques were employed to identify factors influencing patient survival, compare the efficiency of the two treatments, and explore the relationship between CD4 cell count and mortality risk.

1 Introduction

HIV (human immunodeficiency virus) is a virus that progressively **weakens the immune system** by targeting and destroying **CD4 T-lymphocytes**, which are pivotal for immune defense. A decline in CD4 cell count is a significant indicator of deteriorating health and can signal the onset of AIDS (acquired immunodeficiency syndrome). Therefore, monitoring CD4 cell count is essential for assessing disease progression in HIV-infected individuals.

In a longitudinal study, HIV-infected patients who had either failed or were intolerant to AZT (zidovudine) therapy were randomly assigned to two alternative antiretroviral treatment groups: didanosine (ddI) and zalcitabine (ddC). The primary objective of this study was to evaluate and compare the efficacy and safety of these alternative treatments. CD4 cell counts were recorded at the start of the study (randomization) and subsequently at 2, 6, 12, and 18 months.

The dataset includes variables such as patient identifiers, time to death or censoring, death status, CD4 cell counts, observation times, treatment types, sex, previous infection status, and AZT intolerance or failure. This study aims to address key research questions regarding the factors influencing patient survival, differences in treatment effects, and the association between CD4 cell count and the risk of death.

Dataset Presentation

The dataset available for this study comprises 1405 observations, for 467 different patients. 9 variables describe each observation, which can be grouped based on their relevance to the survival analysis:

- **Survival Outcome Variables**:
  - The censoring `time`
  - A numeric vector denoting alive or dead, `death`
- **Biomarker Variables**:
  - The CD4 cell count, `cd4`
  - The time points at which the CD4 cell count was recorded, `time_obs`
- **Treatment Variables**:
  - A factor indicating the treatment group (ddC or ddI), `treatment`
  - A factor denoting AZT intolerance and AZT failure, `azt`
- **Patient Demographics and Clinical History**:
  - A factor with levels female and male, `sex`
  - A factor with levels AIDS (denoting previous opportunistic infection at study entry) and noAIDS (denoting no previous infection), `prev_infection`
- **Identification Variable**:
  - The patient identifier, `subject`

Feature Engineering

Feature engineering is a compulsory aspect of our analysis as it addresses the challenge of non-interpretable and less informative variables. The initial dataset lacked interpret ability and relevance for survival analysis. To enhance the dataset's utility, we performed feature engineering by transforming certain variables into more suitable formats. Consequently, we deemed it pertinent to modify the dataset by introducing a **start/stop format** linked to the variable `cd4` as it depends on the time. Indeed, in the original dataset, `cd4` cell counts were recorded at study entry, where randomization took place, as well as 2, 6, 12, and 18 months thereafter.

Thus, a significant aspect of our feature engineering for survival analysis involved the creation of temporal variables, notably capturing the duration between the onset of the disease progression and death. Additionally, we specify that our categorical variables to be considered such as, using the *as.factor* R function.

These transformations not only improved the interpretability of the dataset, but also ensured that our models were equipped with relevant and meaningful features for the analysis of survival data analysis of mortality of HIV patients.

Exploratory Data Analysis

Exploratory Data Analysis is determining in survival modeling, providing key insights into variable distributions. For the continuous variable, `cd4` [Figure 2a] exhibits a significant **right-skewed distribution**, emphasizing the concentration of CD4 cell counts observed between 1 and 10 for the majority. There are fewer observations at higher CD4 values. This is understandable, as this biomarker monitors disease progression: a decrease in CD4 levels indicates a decline in health. Therefore, the lower the CD4 value, the lower the probability that the person is still alive to measure this variable.

Examining discrete variables, the majority of patients exhibit a **censoring indicator** of 0, indicating death they are **still alive after the end of the study** [Figure 2b]. Regarding the other **binary** categorical variables, the distribution between the two treatment groups is balanced [Figure 3a]. However, this is not the case for the sex variable [Figure 3b], where males are significantly **over-represented** in our data. Due to this imbalance, one should be **cautious** when interpreting the model's findings on the **influence of sex** in this study.

The **correlation matrix** [Figure 5] confirms notable relationships, such as a positive correlation between the previous infection and the AZT status, indicating that patients with a **history of previous opportunistic infections** (AIDS diagnosis) are **more likely** to experience **AZT failure** rather than intolerance. This relationship suggests that the **severity of the disease** progression, as indicated by previous infections, **may impact** the **effectiveness** of AZT treatment. Besides, we also observe a positive correlation between the previous infection and the CD4 level, highlighting that patients with a history of previous opportunistic infections (AIDS diagnosis) tend to have **higher CD4 counts** at study entry. This might seem **counterintuitive**, as one would expect lower CD4 counts with disease progression. However, it **could indicate** that patients with previous infections were more likely to be **monitored and treated earlier**, leading to a temporary increase in CD4 levels due to effective management of their condition.

Then, we investigated the **influence of discrete categorical variables on survival outcomes**. Each variable within the list was examined to understand its influence on survival, employing statistical tests to assess potential differences in survival functions among groups defined by these variables. The subsequent application of the Kaplan-Meier estimator via the survfit function enabled the visualization of distinct survival curves for each group defined by the categorical variable [Figure 7a, 7b, 8a and 8b]. It shows directly that some variables such as

`prev_infection` or `azt` have a real **impact** (negative in this case) on the survival probability, and some others as the treatment given (*ddC* or *ddI*), seems to not change anything about it.

To delve deeper into the statistical significance of survival differences among these groups, we conducted a **log-rank test** using the `survdiff` function. The outcome of this test, gauging whether there are noteworthy disparities in survival distributions, was meticulously printed for each variable. This systematic approach not only furnishes a visual representation of survival nuances among categorical groups but also generates important **p-values** from the log-rank test. The obtained p-values provided interpretations consistent with the trends observed in our previous graph.

Finally, this methodological rigor aligns with our commitment to a comprehensive and nuanced exploration of the data, ensuring that our analyses are both **rigorous** and **interpretable** in the broader context of our research.

2 Statistical Modeling

2.1 Cox Model

The Cox Proportional Hazards model is a statistical technique used in survival analysis. It helps us understand how different factors - covariates influence the hazard or risk of an event occurring over time, such as death. The model is valuable for identifying variables that impact survival, estimating hazard ratios, and making predictions about the probability of events. It's particularly useful in medical research and other fields where analyzing the time until an event of interest is crucial.

The mathematical formulation of the Cox model is as follows:

$$h(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

where:

- $h(t)$ is the hazard function at time t
- $h_0(t)$ is the baseline hazard function
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients associated with covariates $X_1, X_2, \ldots, X_k$

2.1.1 Linear Relationship Cox model

The Cox model methodology selects variables based on a linear relationship, their significance in the log-rank test, and correlation with the censoring indicator `death` and the duration `time`. Chosen variables include: *cd4, time_obs, treatment, sex, prev_infection, azt.*

Results highlight key factors:

In the linear Cox Model, the coefficients $exp(\beta_i)$ for each covariate i represent the log hazard ratios for a one-unit increase in the covariate i and fixing the others covariates. The value $exp(\beta_i)$ is also called the **relative risk**. Therefore, our first model indicates the following:

- **cd4**: 0.88, suggesting that for each unit increase in CD4 count, the hazard of death decreases by 12%.
- **sex=male**: 0.73, indicating that males have a 27% lower hazard of death compared to females.

Variable	Relative Risk
cd4	0.88
sex=male	0.73
treatment=ddI	1.21
prev_infection=noAIDS	0.40
azt=intolerence	0.91

Table 1: Relative Risk for each variables of our linear Cox model

- **treatment=ddI**: 1.21, meaning that patients treated with didanosine (ddI) have a 21% higher hazard of death compared to those treated with zalcitabine (ddC).
- **prev_infection=noAIDS**: 0.40, showing that patients with no previous opportunistic infections (noAIDS) have a 60% lower hazard of death compared to the others.
- **azt=intolerance**: 0.91, indicating that patients intolerant to AZT have a 9% lower hazard of death compared to those who experienced AZT failure.

Model evaluation indicates **moderate accuracy** - *Concordance Index: 0.711.* Likelihood Ratio, Wald, and Score (Log-rank) Tests collectively confirm the **model's effectiveness** in providing accurate survival predictions.

2.1.2 Generalized Cox model

Transitioning to the generalized Cox model involves incorporating non linear effect of some covariates (in the exponential). Then, we will consider a non parametric effect of the covariate `cd4`, by replacing $X_{cd4}\beta_{cd4}$ by a function $f_{cd4}(X_{cd4})$. The [Figure 6] justify this choice as the plot of the **martingale residuals** of `cd4` against the values of the covariate **highights a non lineare behavior**. We thus decide to use the `pspline` function to add a non linear effect on this covariate. With this new model, we obtain almoast the same *Concordance Index: 0.712.* Thus, to differentiate our two Cox models, we will need other metrics to evaluate their performances.

2.2 Random Forest

Random Forest is an **ensemble learning** algorithm widely used for both classification and regression tasks. It constructs a multitude of decision trees during training and merges their predictions to provide a more accurate and stable result. Each tree is built using a subset of the training data and a random selection of features, introducing diversity in the learning process. A Random Forest model was constructed with the **same formula as the second Cox model** to account for non-linear relationships.

The initial model performed well, with an estimated error of **0.49**. Subsequently, a **grid search** optimized key hyperparameters, such as the number of trees, depth, node size, and random variable selection : it permits to make a second random forest model, with a better perfomance - estimated error of **0.45**.

2.3 Comparisons

The **comparison** between both Cox models and Random Forest can be conducted based on the **Brier scores**[1], and the results are presented in the table [Table 2].

We can notice that, the C-index **failed** to capture the differences between the two Cox models (linear vs non-linear), while have almost the same C-index, the Brier Score was able to **strongly differentiate** the two models to measure the accuracy of our probabilistic predictions.

Model	Integrating Brier Score
Cox Linear Model	5.45
Cox Generalized Model	4.16
Random Forest	**3.37**

Table 2: Comparison of results between the Random Forest and Cox models.

The [Figure 1] shows the computed Brier score for each models, given the **accuracy of a predicted survival function** at a given time `t`. The **Random Forest** model outperforms the Cox Proportional Hazards model with a lower *Integrating Brier score*, indicating improved predictive accuracy in forecasting survival outcomes. While the Brier plot indicates that the Random Forest is **generally the best model**, especially for predictions up to 14 months, the non-linear Cox model performs **better beyond** that period. Additionally, Random Forest models are **harder to interpret**, whereas the Cox model allows for **clearer interpretation** of covariate trends.

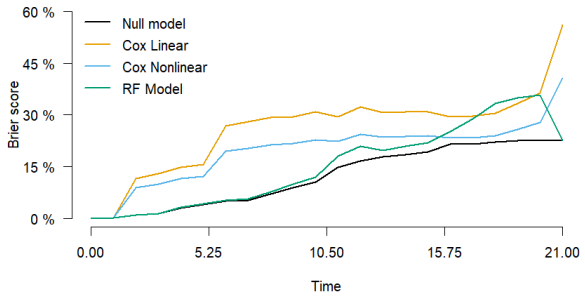


Figure 1: Brier plots for Cox and RF models

3 Discussions and Initiatives

For a male subject with **noAIDS** and **AZT intolerance**, the impact of ddC vs. ddI depends on **cd4** levels. When **cd4** $>$ **30**, no difference is observed, but for **cd4 between 0–10**, ddC slightly improves survival. For a male with **AIDS** and the same characteristics, ddC also shows an advantage when **cd4** $<$ **20**, with no significant differences for **cd4** $\geq$ **30**. Thus, ddC is preferable for **cd4** $<$ **20**, regardless of AIDS status, while both treatments perform equally well for **cd4** $\geq$ **30**. **[Figure 9]** illustrates our interpretations.

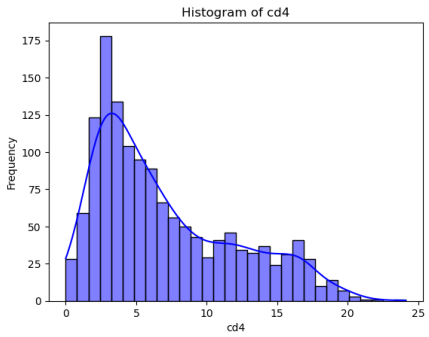
4 Conclusion

This study has made valuable contributions to understanding **HIV survival dynamics**. The mixed results regarding the efficacy of didanosine (ddI) versus zalcitabine (ddC) suggest that **both treatments can be considered** viable options. However, the **slight advantage of ddC** in the Cox model may guide clinicians to prefer ddC in certain scenarios, when the CD4 of the patient is under 20. Moreover, the **strong association** between lower CD4 cell counts and higher mortality risk underscores the critical role of regular CD4 **monitoring** in **managing HIV patients**. This can lead to **timely interventions** that may improve patient outcomes.

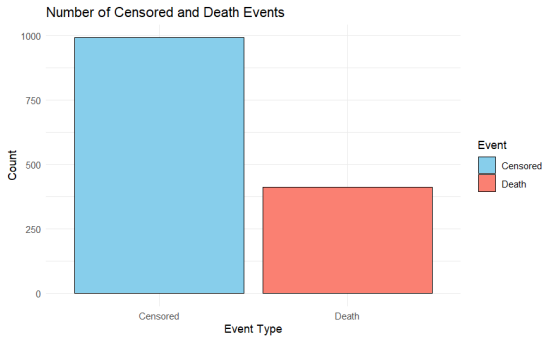
References

[1] Brandon Weathers. "Comparison of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis". In: *Spring 1920 to Spring 2023* (2017).

5 Annexes

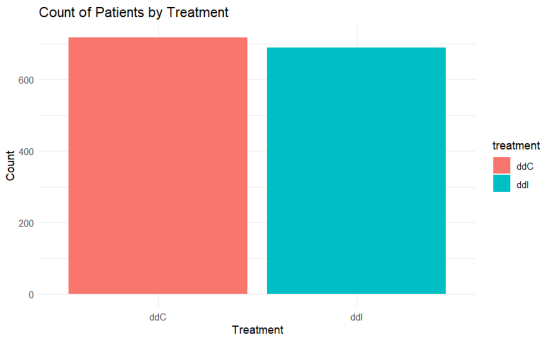


(a) CD4 frequency values histogram

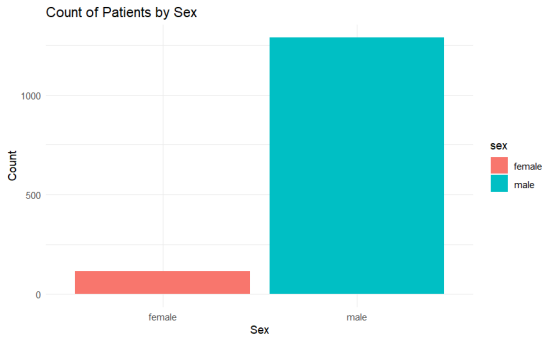


(b) Censored vs Death count in our data

Figure 2: Comparison of CD4 frequency values and Censored vs Death count

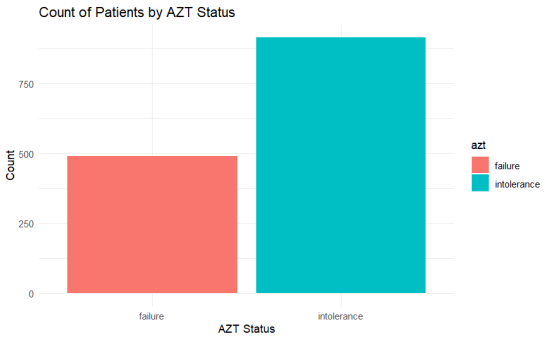


(a) Treatment distribution in our data

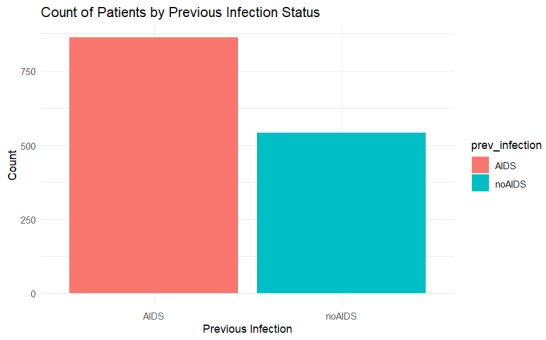


(b) Sex distribution in our data

Figure 3: Comparison of treatment and sex distribution



(a) AZT status distribution in our data



(b) Previous infection status distribution

Figure 4: AZT status and previous infection status

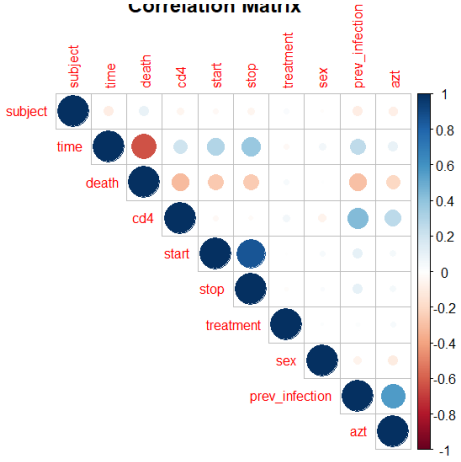


Figure 5: Correlation matrix of our covariates

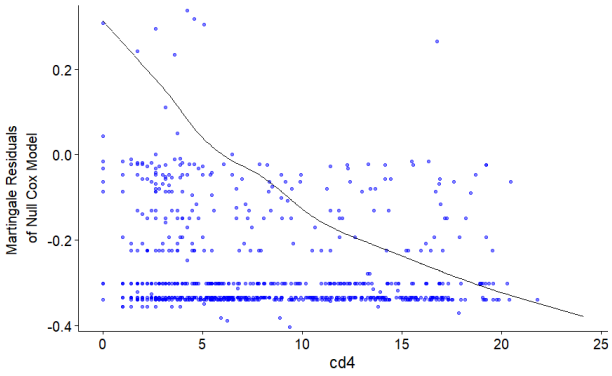
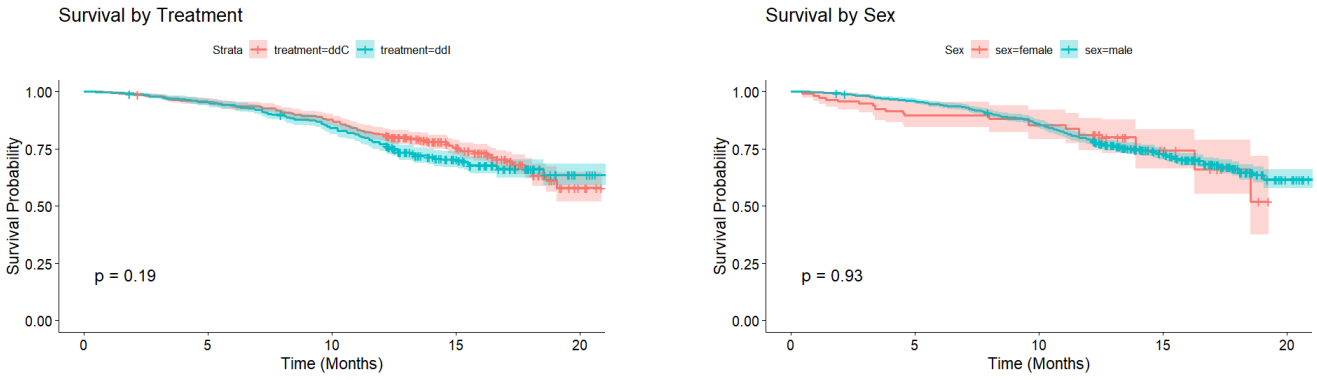
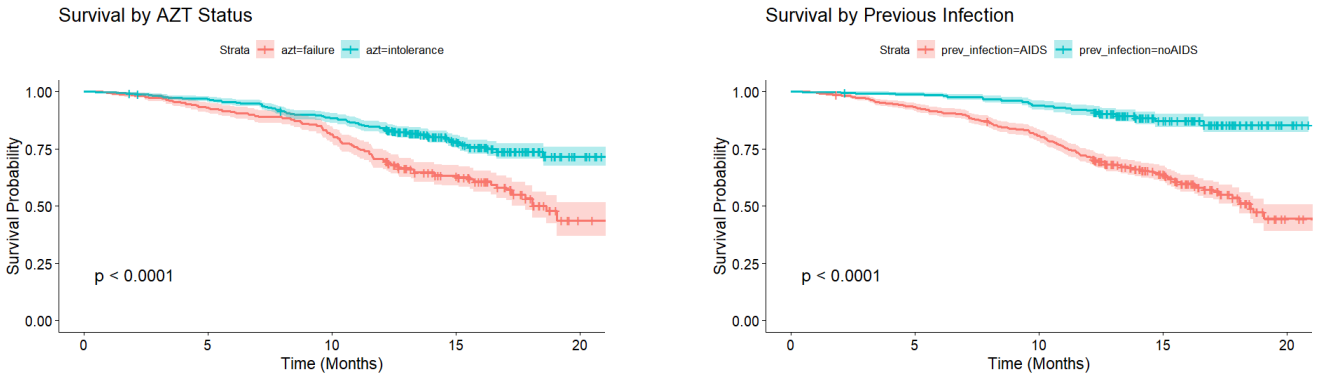


Figure 6: Martingale residuals of the linear Cox model



(a) **(b)**

Figure 7: Comparison of survival based on treatment and sex



(a) **(b)**

Figure 8: Comparison of survival based on AZT status and previous infection

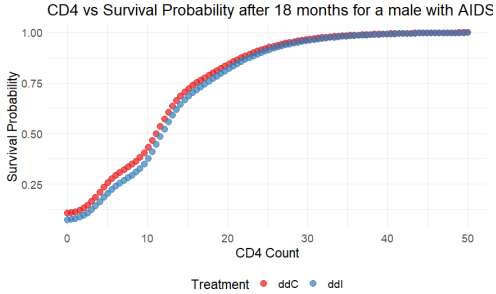


Figure 9: Comparison of Treatment Survival Probability Based on CD4 Levels