**Final Project**

**Survival Analysis in Comparing Therapies in HIV-infected Adults**

*Name: Jennifer Chan Hei Nam*

## Introduction/Motivation:

This project focuses on implementing survival analysis to compare different therapies in treating human immunodeficiency virus (HIV) infected adults, using a clean dataset provided in UC Irvine Machine Learning Repository1. The chosen topic is interesting since Acquired immunodeficiency syndrome (AIDS) is known as a fatal disease progressed from HIV infection. Using statistical method to compare different clinical treatments will provide a useful insight on selecting a superior therapy in treating HIV-infected patients as early as possible, which will significantly impact patient outcomes and quality of life.

There are similar works done by others. In particular, we reimplement the works done in a bioengineering paper, “A Trial Comparing Nucleoside Monotherapy with Combination Therapy in HIV-Infected Adults with CD4 Cell Counts from 200 to 500 per Cubic Millimeter”2. Creative extensions are also made on top of the reimplemented works.

## Problem definition

The topic in this project is to compare therapies in HIV-infected adults. To achieve this purpose, we must define measurements for evaluation. According to the selected clinical paper, CD4 cell counts is an important marker in evaluating clinical conditions of HIV-infected patients. It is a common measurement to check the condition of immune system when infected by HIV, which keeps attacking and destroying CD4 cells, and can lead to fatality3.

This project focuses on statistical analytics with regard to CD4 cell counts in HIV-infected adults. In general, statistical analysis is commonly used in the clinical context for comparing different treatments. This will let clinicians to draw conclusions in a statistical way and potentially influencing future practices in the field.

## Methods

**Algorithms:**

In short, this project consists of reimplementation of hypothesis testing and survival analysis to assess various treatments towards HIV-infected adults. First, analysis of variance and two-sample t-tests are used to compare sample means of changes in CD4 cell counts.

Second, methods of Kaplan-Meier (KM), log-rank test and Cox proportional-hazards model are performed to estimate survival curves and make comparisons among different therapy groups. In the survival analysis, the event is defined by a primary end point, when there is over 50 percent decline in the CD4 cell count, development of AIDS, or death.

Third, a creative extension is made to train a Cox proportional-hazards model with regularization, focusing on examining how regularization affects weights of coefficients and the process of hyper-parameter tuning.

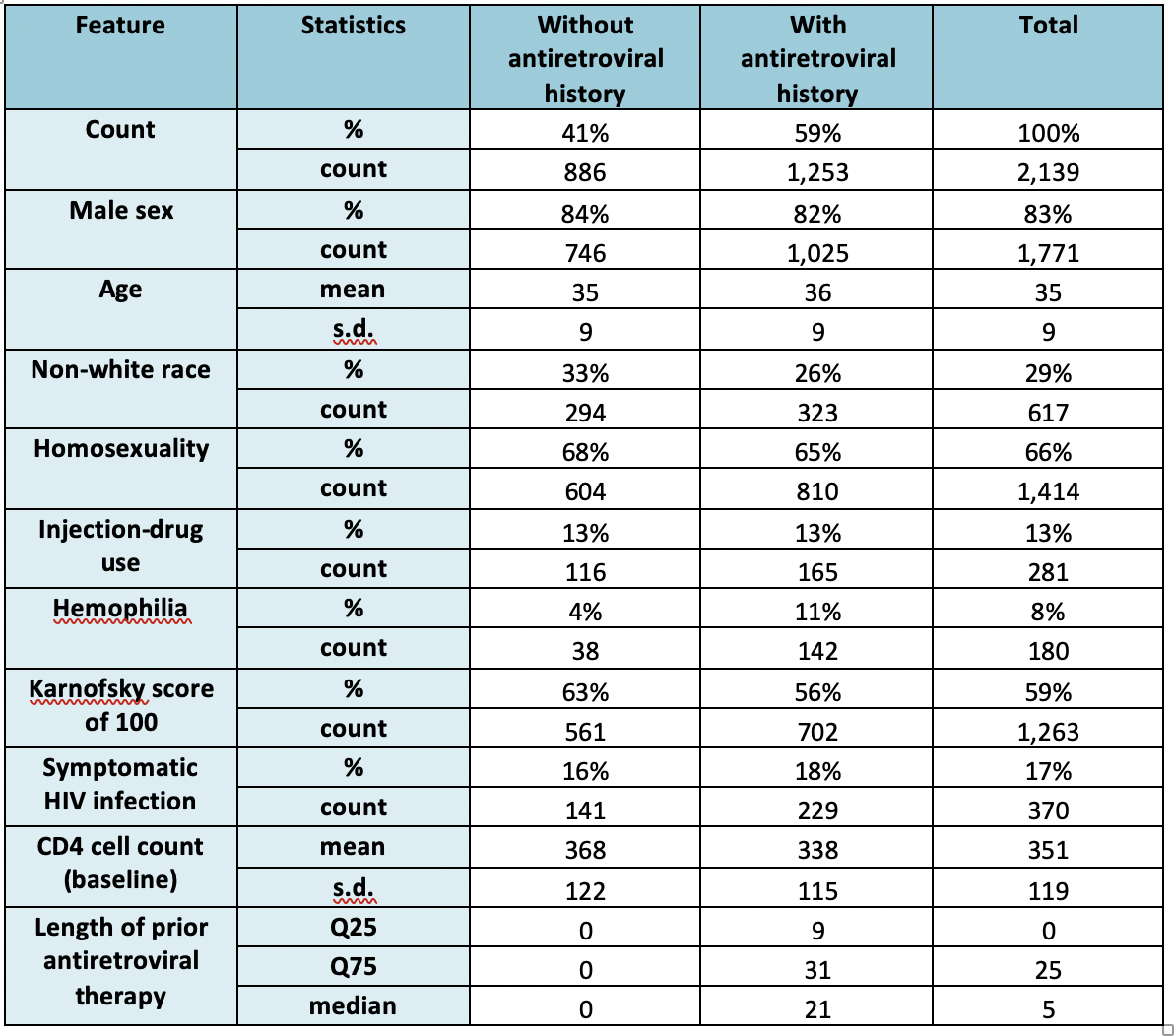
**Software package:**

For the technical part, Python programming is used to implement the above methodologies, using library of lifelines for survival analysis; bioinfokit, scipy, statsmodel for hypothesis testing and relevant analysis. Common libraries such as pandas, numpy, sklearn, matplotlib, and seaborn are also used for general purpose of data pre-processing, manipulation and plotting.

## Results

1. **Checking with base-line summary statistics in the paper**

To begin with, summary statistics are computed and compared with the original paper, as illustrated below:

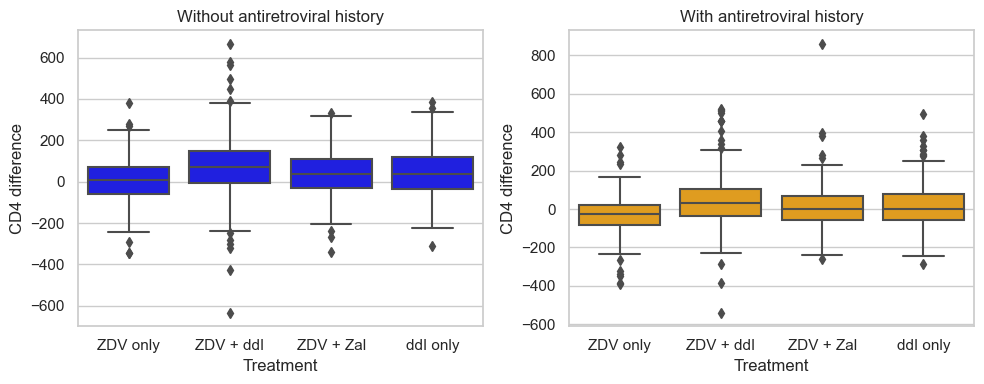


There are some minor differences with the base-line characteristics table of the paper since the dataset provided (N=2,139) is different from the original dataset (N=2,467), but the differences in above summary statistics are at a small extent.

1. **Hypothesis Testing of CD4 Cell Count**

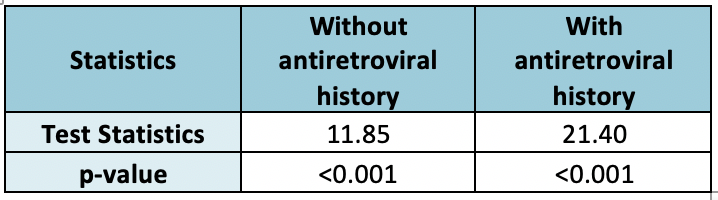
In this statistical testing, we aim to conclude whether changes in CD4 cell count from base-line to 20 (+/-5) weeks are significantly different, when comparing various therapies with the basic treatment of monotherapy with zidovudine (ZDV). Note that the paper compares the base-line with 8 weeks instead, which is unavailable in the provided dataset.

Biologically, an increase in CD4 cell counts indicates that the treatment is effective to help keeping the patient healthy3. Below shows a summary plot of changes in CD4 cell by treatments, in two subgroups of patients (with or without antiretroviral history), where we can observe that monotherapy with ZDV have the smallest change on average in both patient groups:



In this project, we follow the paper to conduct Analysis of variance (ANOVA) first for calculating a global p-value to conclude whether the means are significantly the same for all treatment groups, followed by two-sided t-tests. The reason of doing ANOVA first to compare among all treatment groups is to circumvent the accumulation in error when performing multiple pairwise t-tests, while t-test is used for comparing pairs of treatment groups later. A significant level of 95% is used.

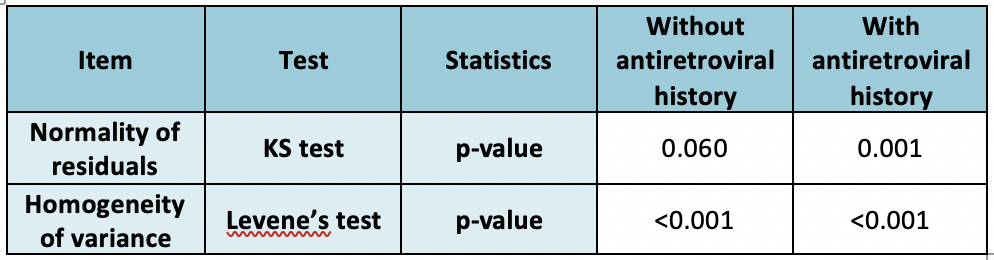
Below shows results of ANOVA tests:



These result show that the mean changes of CD4 cell count are not equal for at least one pair of the treatment groups.

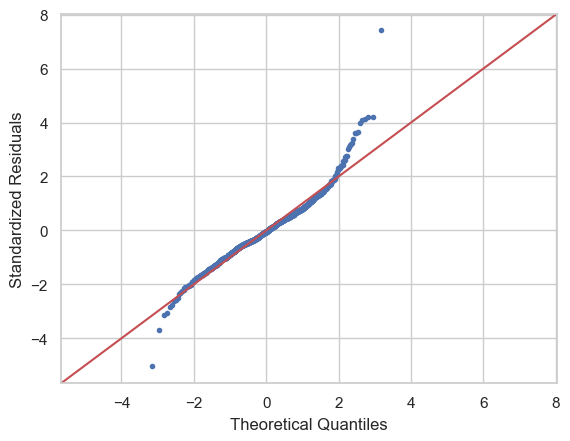
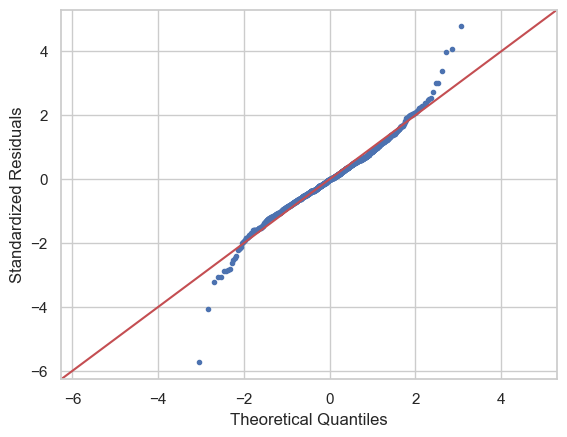
During the project, a challenge was encountered when we further checked assumptions for ANOVA test. When employing ANOVA, it is assumed the data groups have approximately equal variances and the residuals are normally distributed.

Results are shown below:

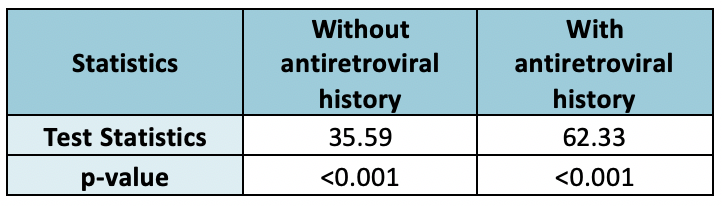


Supplementary QQ plots for normality of residuals are shown below, where we can observe fat tails in the distribution:

Without antiretroviral history With antiretroviral history

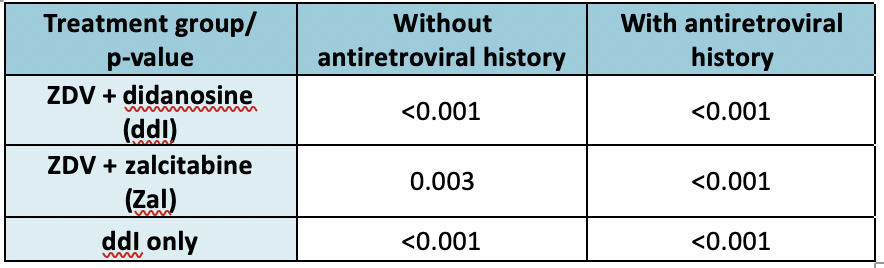


In overall, assumptions of ANOVA are not followed very well in both groups of data. Therefore, Kruskal-Wallis test, a non-parametric test and an alternative to ANOVA when assumptions are not met, is performed to supplement the results:



These result show that the median changes of CD4 cell count are not equal for at least one pair of the treatment groups. Therefore, similar conclusion is drawn that the central measure are significantly different for at least one pair of treatment groups.

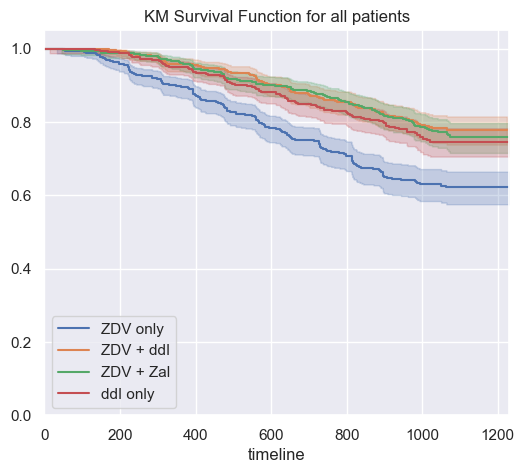
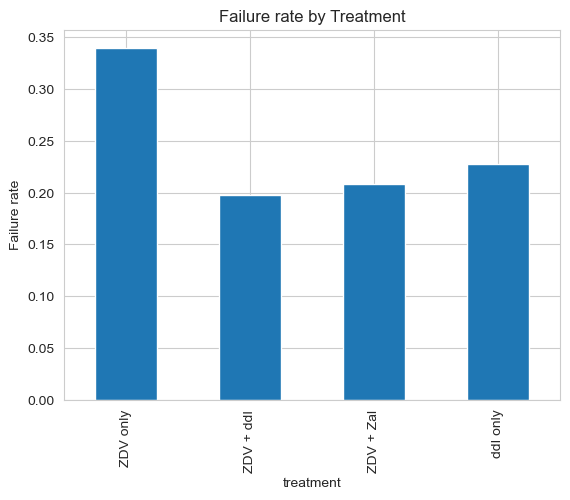
In the next step, pairwise t-tests are performed for comparison between pairs of treatment groups. Since Levene’s test suggests that data groups do not share same variances, for simplicity, Welch Two Sample t-test are used for all pairwise tests against monotherapy with ZDV. Results are shown below:



This suggests a similar result with the paper that the mean changes of CD4 cell count of ZDV only treatment group is significantly different from other treatment groups. Combined with visualization from the previous box plots, we can conclude that other treatments are more effective than monotherapy with ZDV.

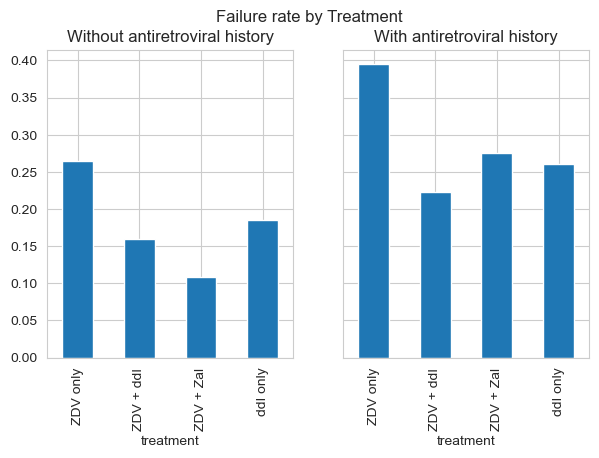
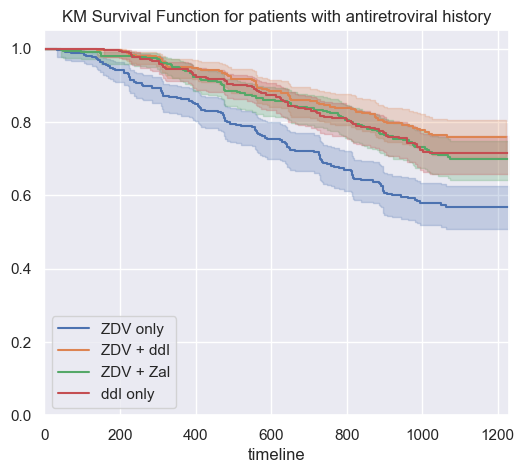
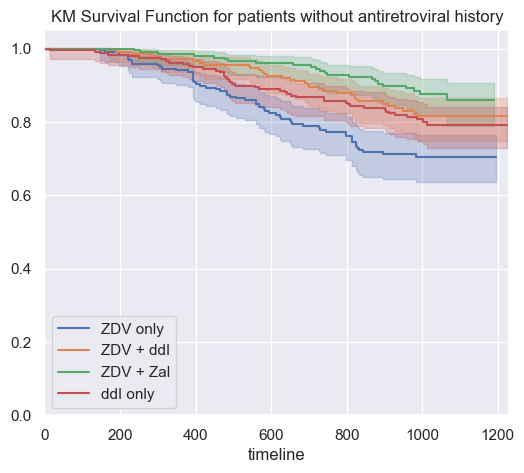
1. **Survival Analysis**

First, KM estimates of the proportion of patients not reaching the primary end point is performed. Below shows the KM survival function plot and corresponding summary statistics of failure rate:

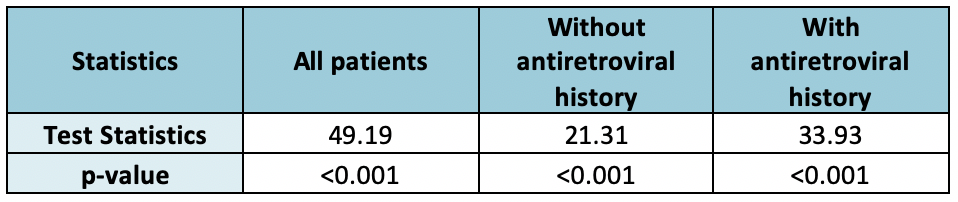
 

From both plots, we can observe that patients treated with monotherapy with ZDV have a lower survival probability in general.

We further plot KM survival function plots and summary statistics of failure rate for the two subgroups, where we can observe patients without antiretroviral history have a higher survival probability in general:



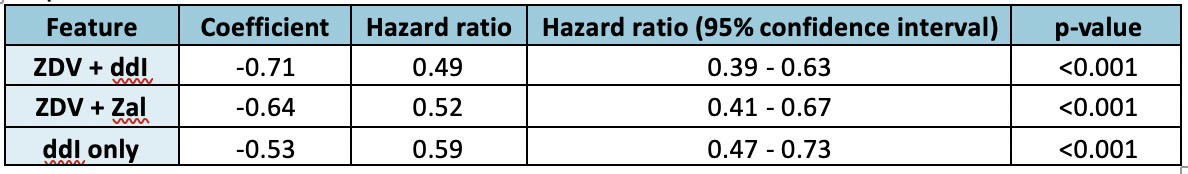
Then, we follow the paper to perform log-rank test to compute global p-values. Results are shown below:



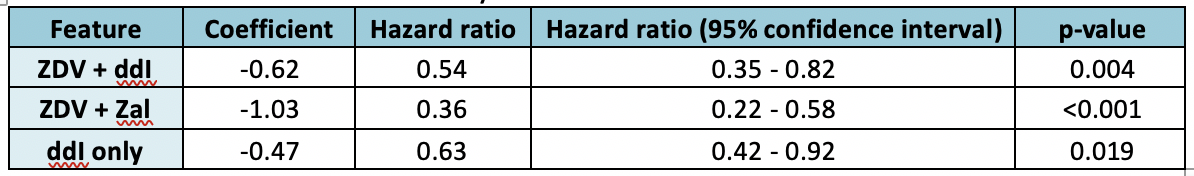
The results indicate that null hypothesis of all treatment groups having the same survival distribution is rejected, which aligns with conclusions in the paper.

Next, we reimplemented cox proportional-hazards model in the paper to compute hazard ratios, which is an estimate of the ratio of the hazard rate between two groups4, of different treatments. One-hot encoding of treatment group feature is performed before modelling. Results are shown as follows:

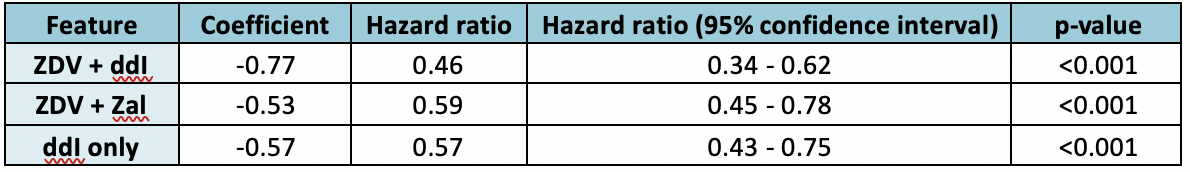
All patients:



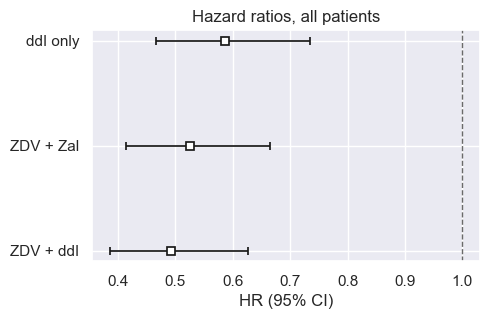
Patients without antiretroviral history:

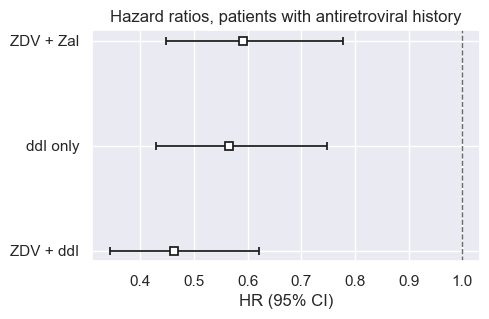
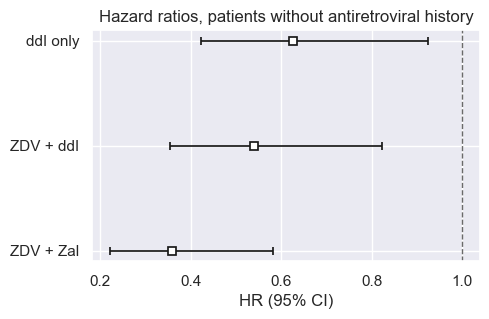


Patients with antiretroviral history:



Their corresponding plots of hazard ratios are as follows:

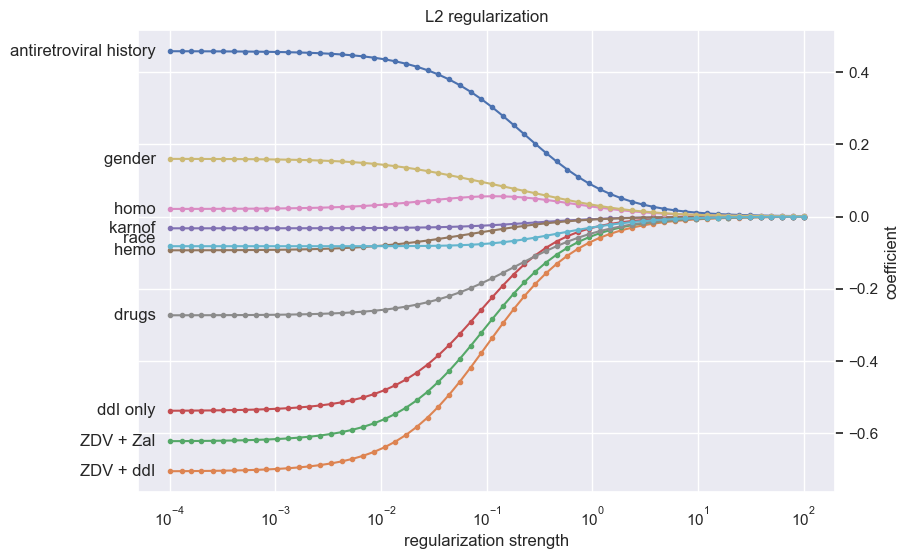


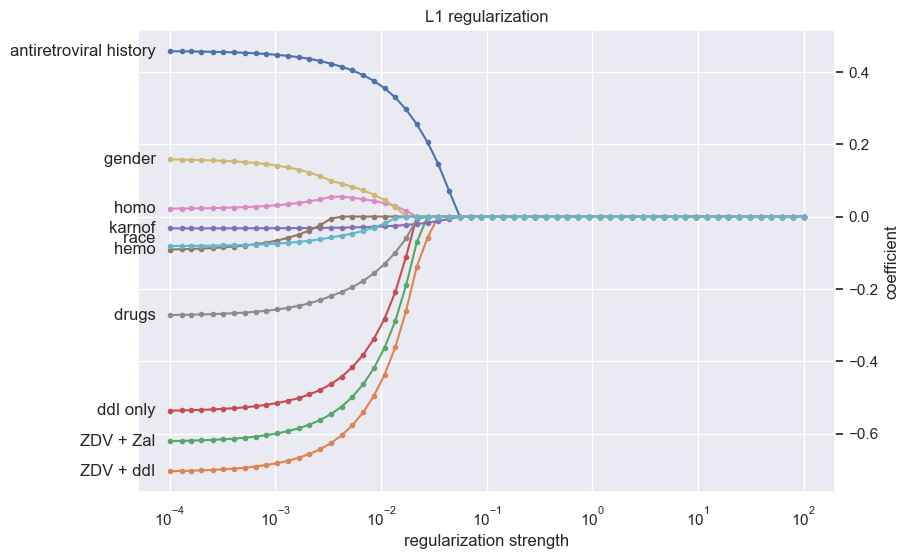


The results align with the results in the paper to a large extent, with minor deviations due to the fact that the provided dataset is not completely identical. Similar conclusions can be made from the above results that ZDV alone was significantly inferior to the other three treatments, with risk reductions of 51%, 48%, and 41% for ZDV with ddI, ZDV with ZaI, and ddI alone respectively. The results for the two subgroups are similar, except that ZDV with ZaI is more effective in patients without antiretroviral history, with the lowest hazard ratio of 0.36.

1. **Creative extension – penalized cox proportional-hazards model**

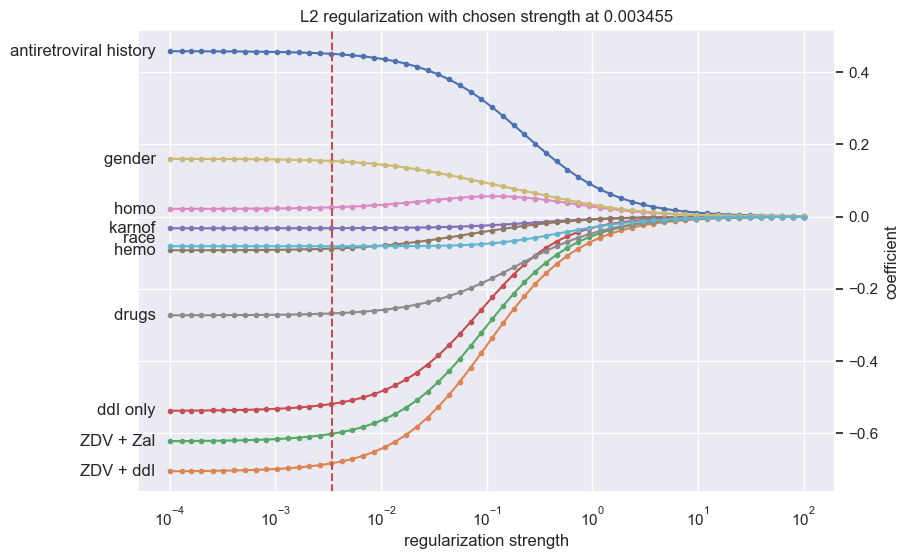
A creative extension is made to develop a cox proportional-hazards model with more variables, consisting of patient demographics and risk factors, applied with regularization to penalize large coefficients. First, we study how regularization affects weights of variables:



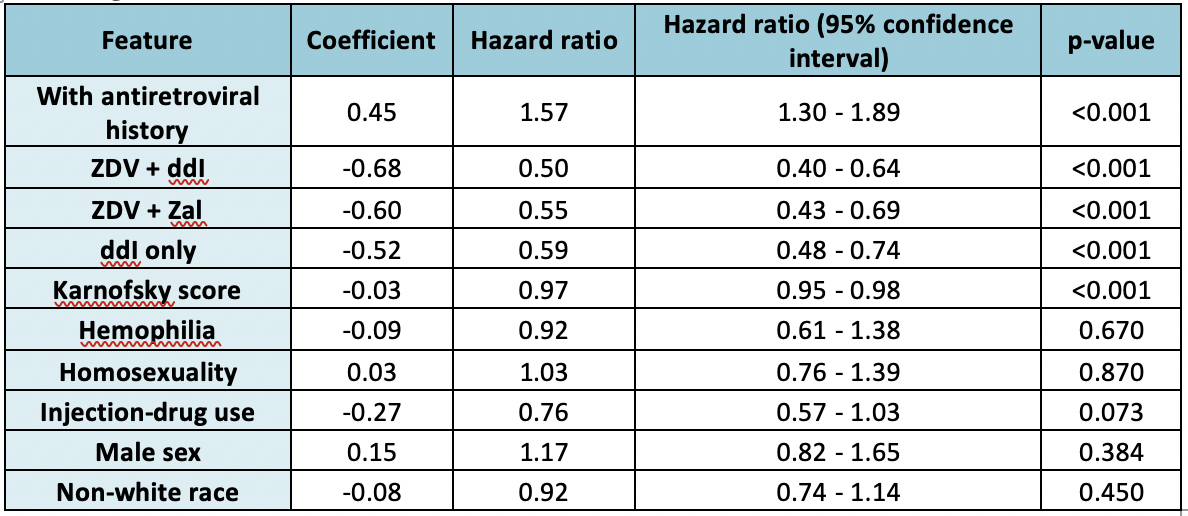


We can observe that in L2 regularization, all coefficients shrink more smoothly and gradually to zero at a very large regularization strength; in L1 regularization, coefficients of variables which are believed to be more important shrinks at a later stage as compared to less important variables.

Then, hyper-parameter tuning of L1 ratio and regularization strength is achieved through a grid search using 5-fold cross validation, with concordance index, which assesses the model’s ability to provide a reliable ranking of the survival times based on the individual risk scores5, as the evaluation metric. It is found that regularization strength with 0.003455 (indicated by red dotted line) and zero L1 ratio yields the best cross validation score of 0.6166:



A final model is trained on the full dataset with the tuned hyper-parameters, which gives following results:



It can be observed that treatments and antiretroviral history features have the most impactful hazard ratios at a statistically significant level, which align with our understanding. Karnofsky score reaches a significance level, but with a nearly neglectable hazard ratio, which may be a result of very similar Karnofsky scores mostly distributed around 100 among all patients.

## References

