

Final Project Report Template for Stats 170B, Spring 2020

Project Title: [Global Radiation Oncology Analysis](#)

Student Names:

[Shawna Tuli, tulis@uci.edu](#)

[Jessica George, jmgeorge@uci.edu](#)

Github: [Github](#) contains analysis code and instructions for which files to download and run
[Emailed sensitive data files](#)

Introduction and Problem Statement

Our project analyzes the survival of 1,000 female patients from a medical cohort in Botswana, Africa. The patients have cervical cancer with or without HIV co-infection. Their diseases have progressed beyond the point of surgery causing each patient to undergo treatment. Initialized in 2013, the medical cohort strives to analyze the differences between the survival for two main patient groups: cervical cancer patients with HIV and cervical cancer patients without HIV. This study is crucial because of Botswana's location in sub-Saharan Africa where the prevalence of both cervical cancer and HIV is extremely high.

Our main results are supported by different applications of statistical survival analysis. These results include findings that model both the global survival between patient groups with or without HIV co-infection and predict patient-specific survival. The statistically-tuned models that analyze the global survival between the specified groups are leveraged to make patient-specific survival predictions on new data with 90% accuracy. All of the methods that we apply are vital to explore for the patients from this medical cohort and these insights will be published in the Red Journal in the upcoming months.

Related Works

Cervical cancer and HIV is currently a topic of interest in the medical community. A global perspective on the associations between cervical cancer and HIV is discussed in the publication,

[1] Chirenje, Z. M. "HIV and cancer of the cervix." *Best practice & research Clinical obstetrics & gynaecology* 19.2 (2005): 269-276.

More specifically, between 2013 and 2015, the medical cohort in Botswana conducted a study on 143 patients with the purpose of comparing survival between HIV-infected versus HIV-uninfected cervical cancer patients in a limited resource setting. The details of the study can be found in the publication,

[2] Grover S, Bvochora-Nsingo M, Yeager A, et al. Impact of Human Immunodeficiency Virus Infection on Survival and Acute Toxicities From Chemoradiation Therapy for Cervical Cancer Patients in a Limited-Resource Setting. *Int J Radiat Oncol Biol Phys*. 2018;101(1):201-210.
doi:10.1016/j.ijrobp.2018.01.067.

Methods relevant to addressing the classical statistical survival analysis problems in our project include classical survival data analysis with Kaplan Meier estimates and survival curves, Cox proportional hazards regression models, and Penalized Cox Regression models. The Log-Rank Test is useful for testing potential differences between groups plotted in our Kaplan Meier survival curves. Similarly, Bayesian Information Criterion and Akaike's Information Criterion are computed to assess the robustness and accuracy of our proposed Cox models. In order to address the patient-specific survival

problems of our project, predictive analytics techniques from machine learning including Random Forest Ensemble Models and k-Fold Cross-Validation also prove to be useful methods.

References which describe relevant background on these algorithms include [Survival Analysis with R](#) by R Views, [Log Rank Test](#) by Boston University, [Regularized Cox Regression](#) by CRAN, [Cox Regression Model Evaluation](#) by Harvard University, [Glmnet Vignette](#) by Stanford University, and [Random Survival Forests](#) by Project Euclid.

Data Sets

Dr. Surbhi Grover, MD, MPH, an Assistant Professor of Radiation Oncology at the Hospital of the University of Pennsylvania provided us with all of our data sets. The data in these files were collected by a team of Penn Medicine researchers. The full patient data set is from the REDcap database.

The small patient data sets were used to develop the models for our project. It includes medical data about 182 patients from 2013 through 2020. There are several files including the [2013-2015 Data File](#), [2013-2018 Data File](#), [2019 Data File](#), and [2020 Data File](#) that were merged based on the unique patient omang ID.

The full patient data set was provided after our models had been developed and was used to see if increasing the number of patients changed the results. The [Full Data File](#) contains the medical data on 1,000 patients spanning 2013 to 2020. Our report analysis and results are based on this data set.

Data Snapshot

Data Set	Size	Years	Dimensionality	Types of Variables
Small	274 KB	2013-2020	182 patients 226 features	Numerical Categorical Binary Datetime
Full	5.6 MB	2010-2020	1,000 patients 293 features	Numerical Categorical Binary Datetime

Table 1: Snapshot of the main data sets.

Topics covered about the patients by the features include personal identification (patient ID, date of birth and geographical district), medical identification (age, height, weight, cancer stage and HIV status), important dates (enrollment date, treatment start and end dates and radiation start and end dates), chemotherapy (prescriptions, number of cycles received and chemotherapy completion), radiation therapy (prescriptions, doses and radiation completion), HIV medication prescriptions, blood test results (Hemoglobin, CD4 cells, Creatinine, etc.), and symptoms (nausea, vomiting, urinary incontinence, etc.).

The features captured in each data set are similar but can differ by definition and value mapping. We constructed a [Small Data Vocabulary Mapping](#) and a [Full Data Vocabulary Mapping](#) to keep track of the feature names, types, values, and definitions.

We explored a couple of questions to get a general understanding of the data. Of the 1,000 patients, 912 of them are in a cancer stage that is beyond the point of surgery and requires cancer treatment to reduce the size of the cancer. 588 of those patients are HIV-positive while 249 are

HIV-negative. The percentages of the patients' initial cancer stages are evenly distributed between HIV-infected and HIV-uninfected patients. The survival between HIV-infected and HIV-uninfected patients is compared to see if on average, one group had generally survived longer than another.

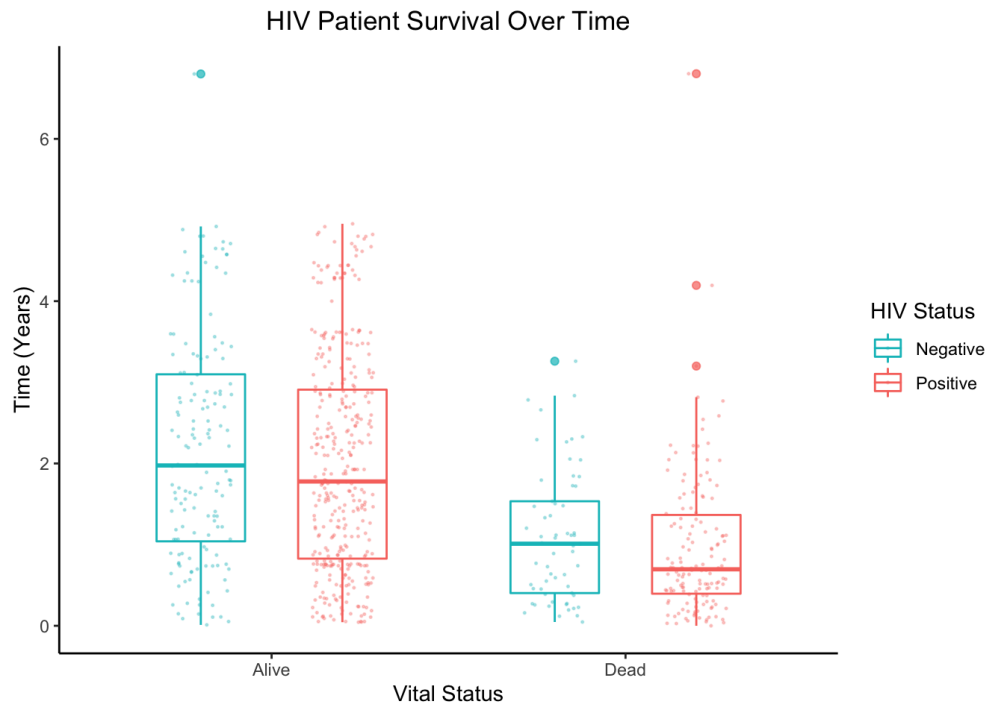


Fig 1: Boxplots Modeling Survival highlights the patient survival for patients that are alive based on followup times. The patient survival for patients that are dead is based on times from the patient starting treatment until death. For post mortem patients, the HIV negative group appears to survive longer in terms of years.

Overall Technical Approach



Fig 2: Project Pipeline summarizes the major points achieved to complete the project efficiently.

Data Collection

As mentioned previously, all of the data was provided by Dr. Surbhi Grover, MD, MPH, an Assistant Professor of Radiation Oncology at the Hospital of the University of Pennsylvania and her team of medical researchers. All of the data sets that were supplied are Stata files that are converted to CSV formats.

Data Cleaning

All of the raw CSV files are read into Jupyter Notebook to use libraries in Python to clean the data. One aspect of the data cleaning is adjusting the types of the variables, such as categorical, numerical, and datetime. The other aspect is padding the data with null values to account for inconsistencies and check for any abnormal patient recordings. All medical terminology is simultaneously defined in a table of variable mappings. Each data set is cleaned individually to ensure successful merging. The process of merging the files is semi-automated. The data sets are merged based on the patient's unique omang ID and validated by each patient's post mortem date or censorship date for those who had a followup visit or left the study. The small data set and full data set do not need to be merged because the patients in the small data set are already included in the full data set.

Data Preparation for Analysis

Additional feature generation is performed on the data sets to include features needed for survival analysis. Some of these features include the patients' vital statuses (Alive or Dead) and time alive (in years from start of treatment until death).

Exploratory Data Analysis

Exploratory data analysis is completed using packages in RStudio. Questions explored about the patients included topics such as age, location, cancer stages and treatment, and HIV status and treatment. These insights and visualizations are summarized in the Appendix.

Classical Survival Analysis

Classical survival analysis is used to create models for the patient's medical data. This is suitable for our data in order to analyze the time until the event of interest (death), the factors that can estimate the duration until the event of interest, and the differences between certain groups (HIV-infected and HIV-uninfected patients). All of the survival analysis is computed and visualized using packages in RStudio. The modeling methods include Kaplan Meier estimators and survival curves, Cox proportional hazards regression model, and Penalized Cox regression. Dr. Grover's team was crucial during this step because of their medical background for instructing us about the features and their impact on the patients' survival.

Kaplan Meier models are fit to analyze survival curves for HIV-infected and HIV-uninfected patients. Univariate Cox regression models are fit to determine which features are important in determining the global survival for the patient data. Multivariate Cox models are fit to model the global survival for the patient data with a set of covariates. Penalized Cox regression is another method used for model selection in the multivariate Cox models because the data is highly dimensional yet the sample size is small.

Model Assessment

The Log Rank Test is used to test for a difference between Kaplan Meier survival curves. Bayesian Information Criterion and Akaike's Information Criterion are computed to assess the robustness and accuracy of the Cox models.

Patient-Specific Survival Prediction

To validate our proposed Cox models beyond descriptive statistics, Random Forest Ensemble Models are also fit to model each patient's individual survival curve. The data is split into 75% train and 25% test sets and maintain an even distribution between HIV-infected and HIV-uninfected patients. After training the model to estimate the survival using the Cox models, the model predicts the survival on the test data.

Validation

To assess the accuracy of the Random Forest Ensemble model, k-Fold Cross Validation is applied to split the data, fit the model, and provide a performance metric for the proposed Random Forest Ensemble models.

Software

All of the source code is developed by us in RStudio and Jupyter Notebooks with R and Python respectively. Using outside sources/libraries was not necessary. For visualization purposes, we modified part of the ggsurvplot function to plot our Kaplan Meier survival curves with number of patients at risk and also the percentage at risk.

Data Cleaning Scripts

All of the cleaning scripts are coded using Python in Jupyter Notebook, an interface platform in which the Stata files are inputted, cleaned, wrangled, and output as CSV files. Python has libraries including NumPy (for dataframes), Pandas (for cleaning/manipulating dataframes), and Matplotlib (to create plots with pyplot). The files created to clean the data include 2data wrangling-cleaning.ipynb, 5data wrangling-cleaning.ipynb, merge.ipynb, and all-patients.ipynb. The first two files clean the originally provided small data sets individually. The merge.ipynb merges all of the files in the small data set. The last file, all-patients.ipynb cleans and prepares the full data set for analysis.

R Markdowns

The R Markdowns are created in RStudio. R provides packages including statVisual (for exploratory data analysis), ggplot2 (for data visualizations), survminer (for survival analysis), glmnet (for penalized Cox regression), and gg_rfrsc (for Random Forest Ensemble models). The R Markdown files created for the small data set include EDA.Rmd, Surv_GLM.Rmd, Kaplan.Rmd, Cox.Rmd, PCox.Rmd, and RF.Rmd. The R Markdown files created for the full data set include FullEDA.Rmd, FullKaplan.Rmd, FullCox.Rmd, FullPCox.Rmd, and FullRF.Rmd.

Experiments and Evaluation

We conducted several cervical cancer patient specific analyses and looked at the outcomes of this cancer in HIV infected and uninfected, female patients in Botswana. Detailed below are the models.

Models

Our covariates for the models include α 1 time_alive_years is the days, converted into years from when the patient initiates treatment until discovered to be dead, β 1 HIV Status is 0 if the patient tested negative for HIV and 1 if the patient tested positive for HIV, β 2 Vital Status is 0 if the patient is alive and 1 if the patient is dead, β 3 Age Cat is the age groupings of 25-39, 40-59, 60+, β 4 Cancer Stage is the patient's final cancer stage (1-4), β 5 Chemo Cycles is the number of completed chemo cycles from 1-8, β 6 HB (>10) is 0 if the HB is < 10 grams/deciliter and 1 if the HB is \geq 10 grams/deciliter, β 7 HB Final is

the patient's hemoglobin at the end of the treatment, β_8 EQD2 (>75) is 0 if the EQD2 is < 75 Gy and 1 if the EQD2 is ≥ 75 Gy and β_9 EQD2 Final is the patient's final EQD2 at the end of treatment, which spans 10 to 92.

Kaplan Meier estimates allowed us to measure the fraction of subjects who survived for a certain amount of survival time under the same circumstances depicting the probability of surviving in a given length of time where time is considered in small intervals. We estimated x-year survival and median survival time of all female patients in the cohort. Furthermore, we compared the survival times of these women by age groups of 20-39, 40-59, 60+ and by their HIV status. Finally, we summarized their survival analysis over time and looked at overall trends over time.

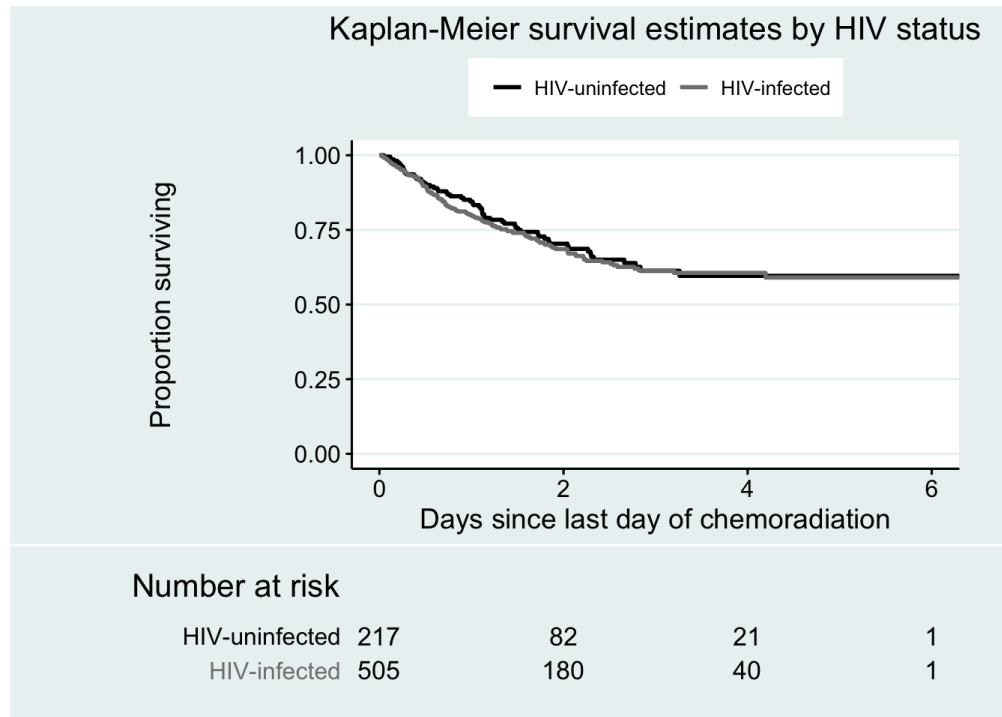


Fig 3: Kaplan Meier survival curves between HIV-infected and HIV-uninfected cervical cancer patients is used to measure the fraction of patients who survived for a certain amount of years under the same circumstances.

Our model for estimating all patients' survival probability is $Y = \alpha_1 + \beta_2 \sim 1$. Additionally, we looked at the survival probability of the patients by vital status, $Y = \alpha_1 + \beta_2 \sim \beta_2$, and HIV status, age for those who have HIV and age for those who do not have HIV and found that the HIV status is Negative with a mean of 1 year and the HIV status is Positive with a mean of 0.75 years. Secondly, we estimated the x-year survival of the patients, $Y = \alpha_1 + \beta_2 \sim 1$, times = 120, and estimated the median survival of the patients by HIV status, vital status, age and HIV status and age. Moreover, we compared the survival times of the patients by age groups of 20-39, 40-59, 60+, $Y = \alpha_1 + \beta_2 \sim \beta_3$ and the survival times of patients by HIV status $Y = \alpha_1 + \beta_2 \sim \beta_1$. Lastly, we summarized the survival analysis over time with $Y = \alpha_1 + \beta_2 \sim \beta_2$, times = c(1,30,60,90*(1:10)) and by 6 month intervals for vital status, age and HIV status.

The quantitative metrics of the Log-Rank Test provide information about the statistical differences between groups from the Kaplan Meier survival curves. For both the survival curves between patients' HIV status, $Y = \text{logrank_test}(\alpha_2 + \beta_2) \sim \text{as.factor}(\beta_1)$ and the survival curves between patients' age groups $Y = \text{logrank_test}(\alpha_2 + \beta_2) \sim \text{as.factor}(\beta_3)$, we failed to reject the null hypothesis that the two survival curves are identical.

Next, we used Multivariate Cox Proportional Hazards Regressions to derive a response from the interactions that occur when an independent variable has an effect on the outcome depending on the values of another independent variable. Model 1 $Y = \alpha_2 + \beta_2 \sim \beta_1 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_8$, model 2 $Y = \alpha_2 + \beta_2 \sim \beta_1 + \beta_3 + \beta_4 + \beta_5 + \beta_7 + \beta_9$ and model 3 $Y = \alpha_2 + \beta_2 \sim \beta_1 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_8 + \beta_4\beta_9$.

Multivariate Cox Proportional Hazards Regression with Interaction


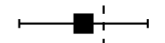

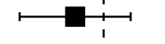






Variable	N		Hazard ratio	p	
HIV	Neg	144		Reference	
	Pos	357		0.87 (0.55, 1.37)	0.539
Age	20-39	108		Reference	
	40-59	288		0.82 (0.55, 1.21)	0.315
	60+	105		0.88 (0.49, 1.58)	0.669
Cancer Stage		501		1.39 (1.09, 1.76)	0.007
Chemo		501		0.73 (0.47, 1.14)	0.169
Hemoglobin		501		0.90 (0.83, 0.98)	0.015
EQD2		501		0.98 (0.97, 0.99)	0.001
Chemo:EQD2				1.00 (1.00, 1.01)	0.217

Fig 4: Summary of covariates in Model 3 includes the number of patients used to estimate each coefficient (disregarding missing data), the hazard ratios with confidence intervals, and the p-values.

Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) provide model assessment for both the Multivariate Cox regression models and the Penalized Cox regression models. BIC introduces a penalty term for the number of parameters in the model such that the model with a lower BIC value is a better model. AIC allows for testing how well the models fit the data set without overfitting it. In this case, the model with the lower AIC score has a superior balance between its ability to fit the data set and ability to avoid overfitting the data set. Hence, we prefer the models with both lower BIC and AIC values and conclude that the first model is the best.

The Random Forests Ensemble model uses bagging as the ensemble method and decision tree as the individual model to predict an individual patient's survival curve where bagging is training a bunch of individual models in a parallel way and each model is trained by a random subset of the data. The benefits of this model are that it is fully nonparametric, robust to outliers and does not suffer from a convergence problem, can be used for high-dimensional data, offers OOB (cross-validated) prediction that does not

overfit and provides a fully nonparametric VIMP measure of a variables' contribution to predicting survival.

The Random Forests model with HIV is $Y = \alpha_2 + \beta_2 \sim \beta_1 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_8$ with the multivariate cox proportional hazards regression model 1 data for HIV and the same model with the data for no HIV gave the patient-specific survival curves below with a prediction accuracy of 85-90% (10-15% Error). We got an indication of the errors by coloring the plots based on the patients' vital statuses. It would be assumed that with better accuracy, all post mortem patient survival curves would lie below patient survival curves for patients who are still alive.

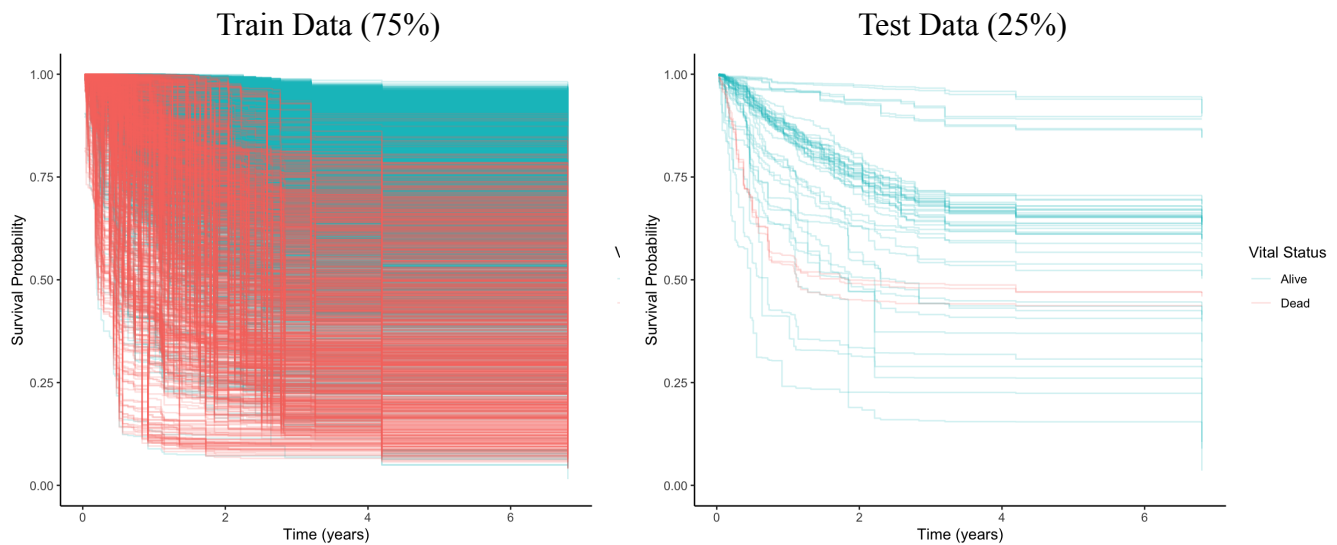


Fig 5: Patient specific survival curves include the trained model on the left and the predicted test model on the right. The lines are colored by patient vital status to give a small indication of how well the model is performing.

Notebook Description

The Github linked on the first page outlines definitions for each data set, Python script, and R Markdown. We have provided files for both the small data set and full data set that are condensed versions of the main topics covered in our analysis. Because the data sets contain confidential data about patients, the data sets required to run our analysis files are emailed to the professors.

Members Participation

The best approach in completing this project accurately was to jointly work on every task. Professor Minin's expertise and STATS 170B office hours together with weekly calls with Dr. Grover during the second half of the quarter, guided us in facilitating our data analyses.

Project Milestones

I.	Weeks 1-2	[Data Wrangling and Management]
II.	Weeks 3-4	[Data Analysis I, II]
III.	Weeks 5-6	[Prediction and Validation]
IV.	Weeks 7-8	[Data Visualization]
V.	Weeks 9-10	[Presentation, Report and Publication in IJROBP]

Task Breakdown

<u>Title</u>	<u>Date</u>	<u>Details</u>	<u>Participation</u>
Sync 1-7	2/29-6/7	Calls with Dr. Grover	Shawna - 50%, Jessica - 50%
Project Proposal	3/9	Drafted written proposal for the project outlining	Shawna - 50%, Jessica - 50%
Proposal Presentation	4/8	Data Wrangling and Management: Rename data features in order to map variables, clean datasets to account for post mortem dates or censorship dates, and merge datasets based in unique patient ID	Shawna - 35%, Jessica - 75%
First Presentation	4/13	Data Analysis I/II: Model/feature selection to reduce collinearity and statistical modeling without overfitting/underfitting	Shawna - 50%, Jessica - 50%
Second Presentation	5/11	Prediction and Validation: Model prediction in patient-specific survival and between HIV-infected and HIV-uninfected	Shawna - 50%, Jessica - 50%
Office Hours	5/18	Update the professors on project status and ask questions	Shawna - 50%, Jessica - 50%
Project Report Draft	5/22	Data Visualization: Organize all of the quarter's work into the template and set up github repository for code	Shawna - 75%, Jessica - 35%
Final Presentation	6/1	Presentation: Present all data analysis and insights	Shawna - 50%, Jessica - 50%
Final Report	6/12	Report and IJROBP Publication: Complete 10 pages detailing analyses and insights	Shawna - 50%, Jessica - 50%

Table 1: Task Breakdown outlines the tasks and members' participation respectively.

Discussion and Conclusion

Working with this data allowed us to gain exposure to cancer epidemiology in a global setting, consult with a data collection team concerning any questions, get familiar with REDCap software, clean and analyze patient's medical data, and practice survival analysis.

From the methods and algorithms we used, we learned how to accurately model survival data as a global average for all patients and patient-specific survival. In addition, we learned that the strengths of these methods and algorithms lie in their robust ability to use factors to predict the time until an event of interest. Nevertheless, their limitations stem from their age. Most methods that predict survival are still evolving and growing into their reliability, especially concerning topics of human life.

Verification of results ended up being harder than we expected for our project given small sample size, high dimensionality, and missing data. Provided that we did not have extensive knowledge in Global Health, we had a learning curve to understand medical terminology in order to further the context from Dr. Surbhi Grover's publications. Our findings regarding the Log Rank Test and failing to reject the null hypothesis that the survival curves were identical between HIV-infected and HIV-uninfected patients was surprising about our project because if a patient's immune system is compromised by HIV, it would be assumed that the patient's survival would be significantly different.

Other lessons we learned were expected and on the full scope and breadth necessary in executing a data science project end-to-end. In conjunction with this, we learned how to adjust the tool of Cox Regression to Penalized Cox Regression in our analysis. We did not use anything out of the ordinary.

If we were in charge of a research lab, we would invest in the next year or two in making major progress on this problem by exploring the ideas of how these insights specific to cervical cancer compare to similar insights discovered for women in Botswana with other cancers. Additionally, we would pursue the direction of cervical cancer in similar third-world countries to add to a more generalizable picture of Global Health.

Appendix

All insights use the full data set and were originally visualized with the small data set.

A. HIV Status

- a. How many patients are HIV-infected? 257
- b. How many patients are HIV-uninfected? 610

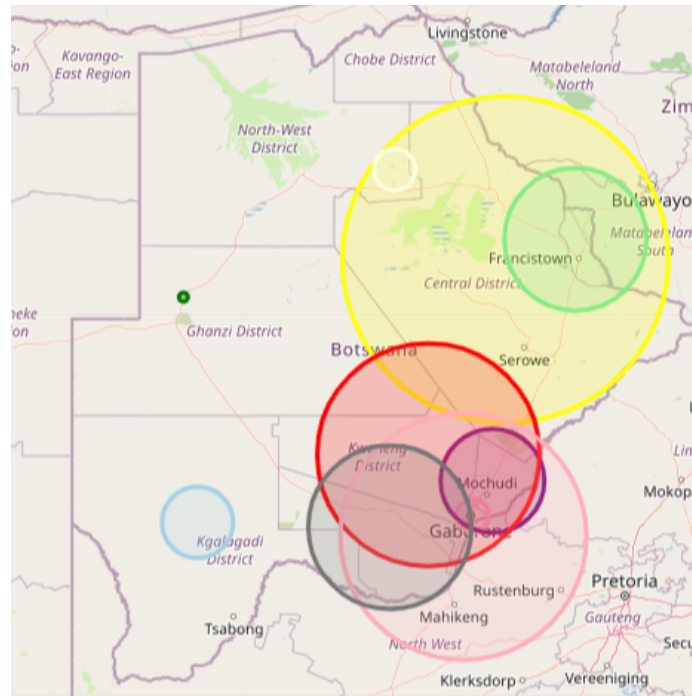
B. Age of Patients

- a. Minimum Age: 22 years
- b. Maximum Age: 95 years

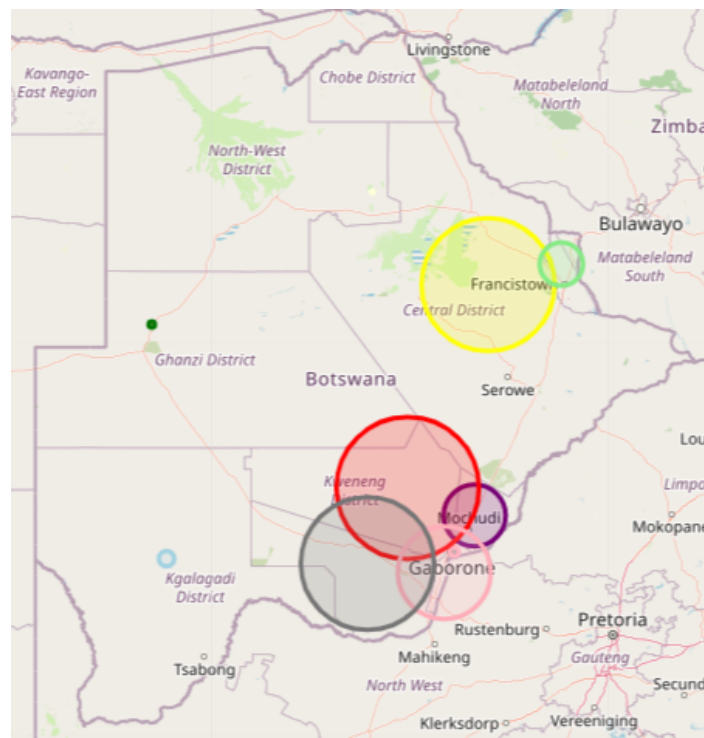
Age	Number of Patients
20-29	9
30-39	177
40-49	327
50-59	165
60-69	126
70-69	57
80-89	19
90+	1

A. Location of Patients by HIV Status

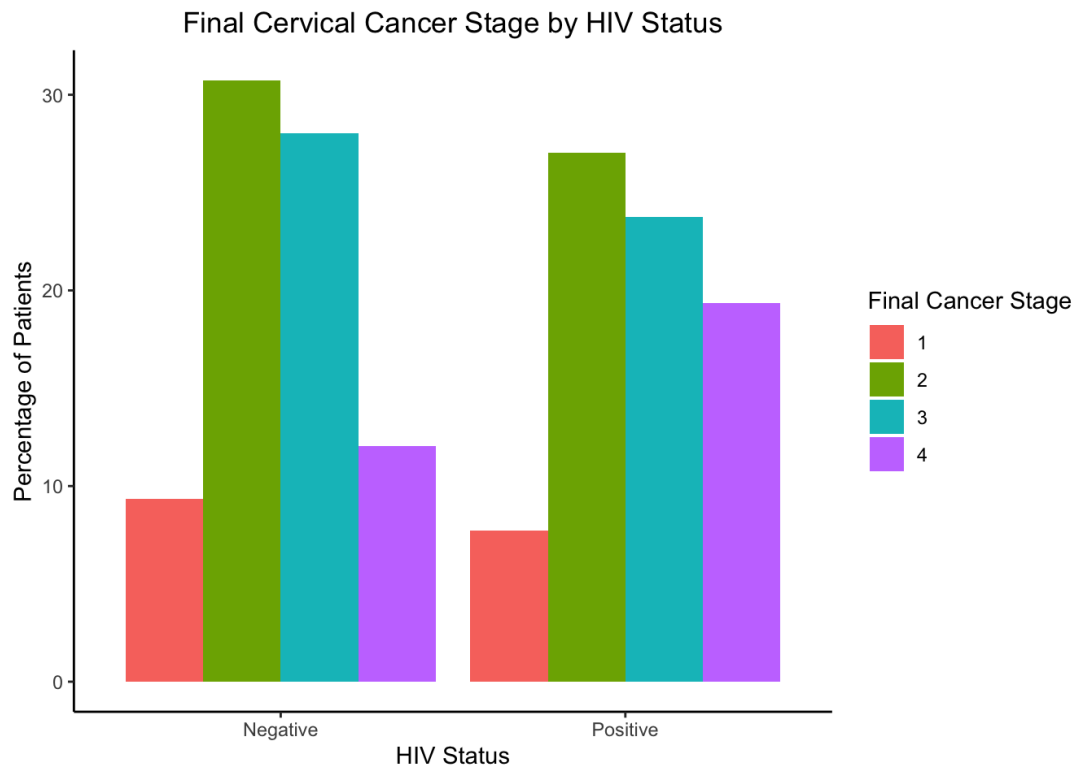
a. HIV-infected patients



b. HIV-uninfected patients

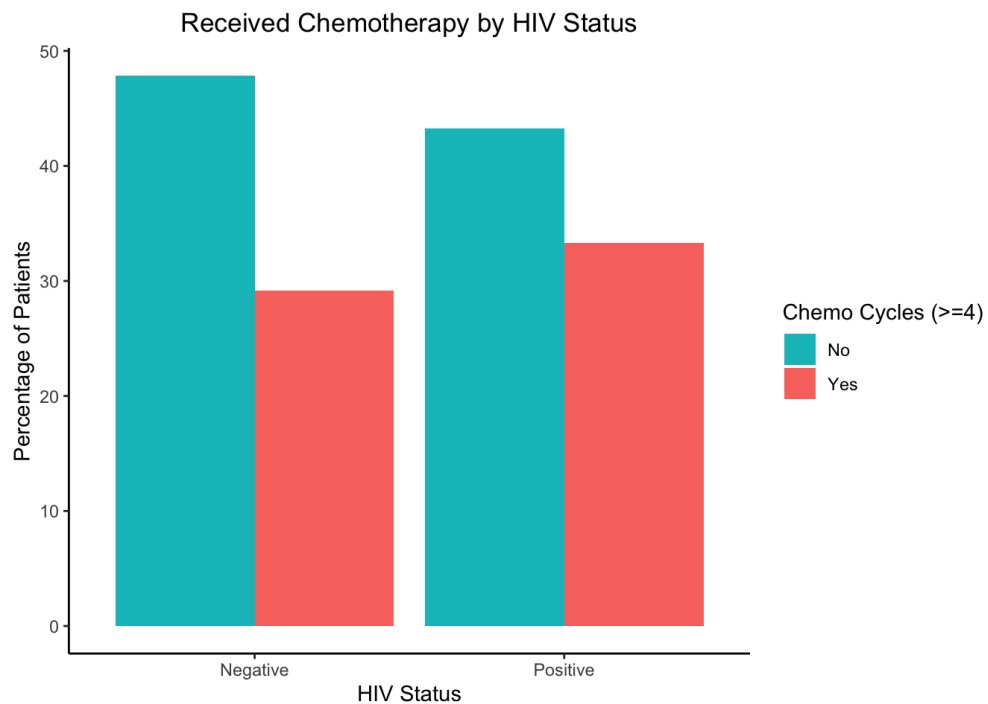


B. Final Cancer Stages by HIV Status



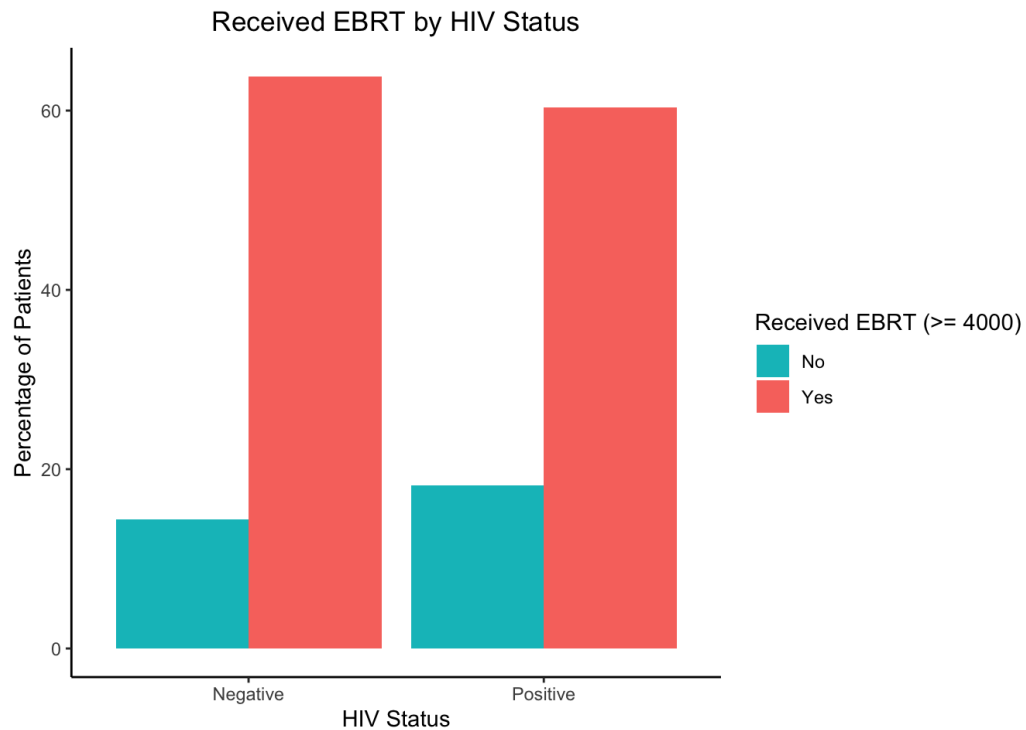
C. Treatment by HIV Status

- a. Chemotherapy: Considered all chemo received if patient received greater than or equal to 4 cycles of Cisplatin

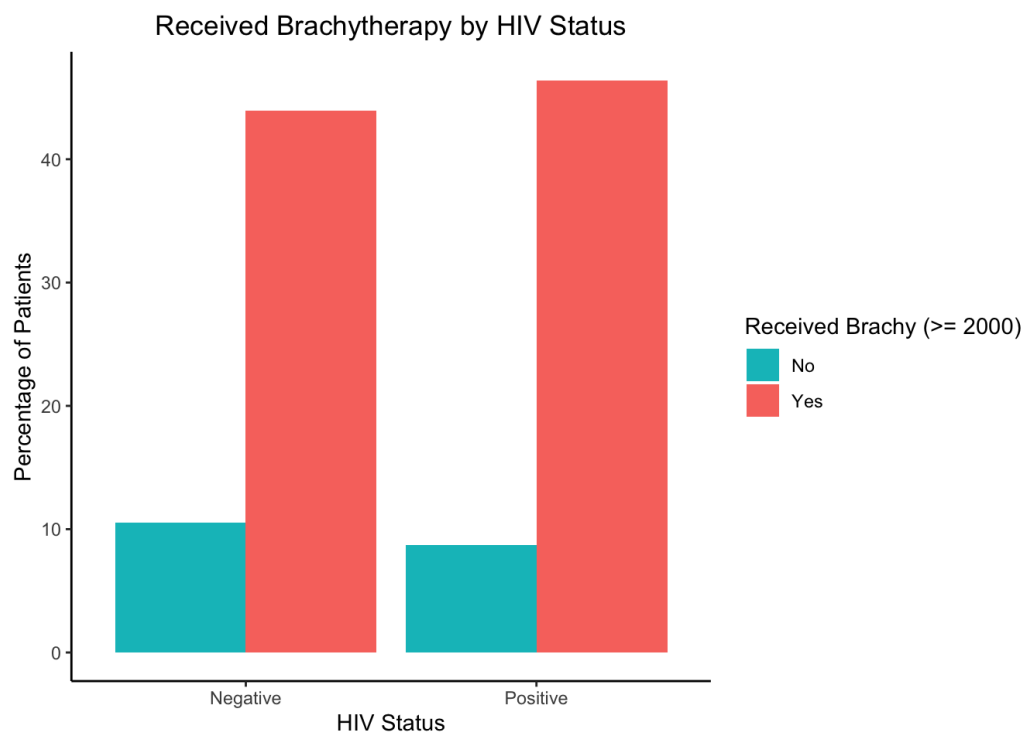


b. Radiation Therapy

- EBRT: External Beam Radiation Therapy; Considered received if patient received greater than or equal to 40 Gy



- Brachytherapy: Considered received if patient received greater than or equal to 20 Gy

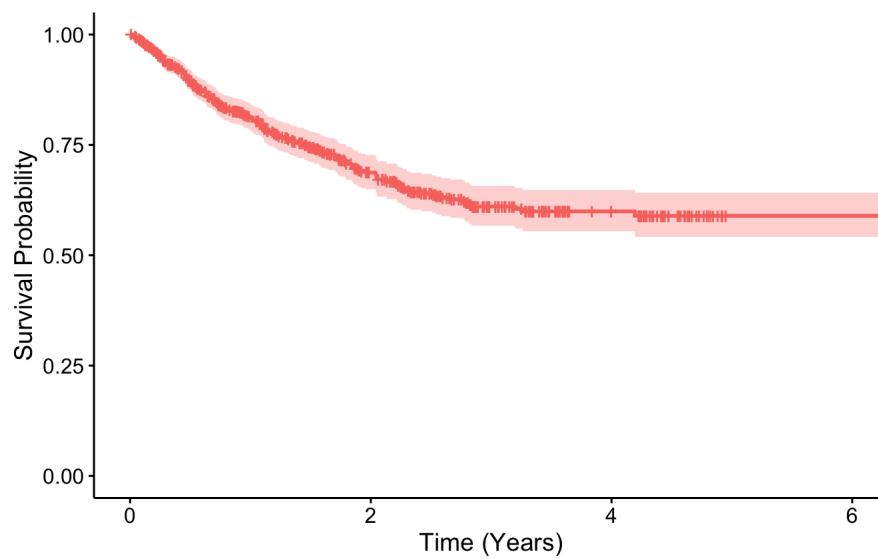


- EQD2: Combination radiation therapy: Considered received if patients received greater than or equal to 75 Gy



D. Kaplan Meier

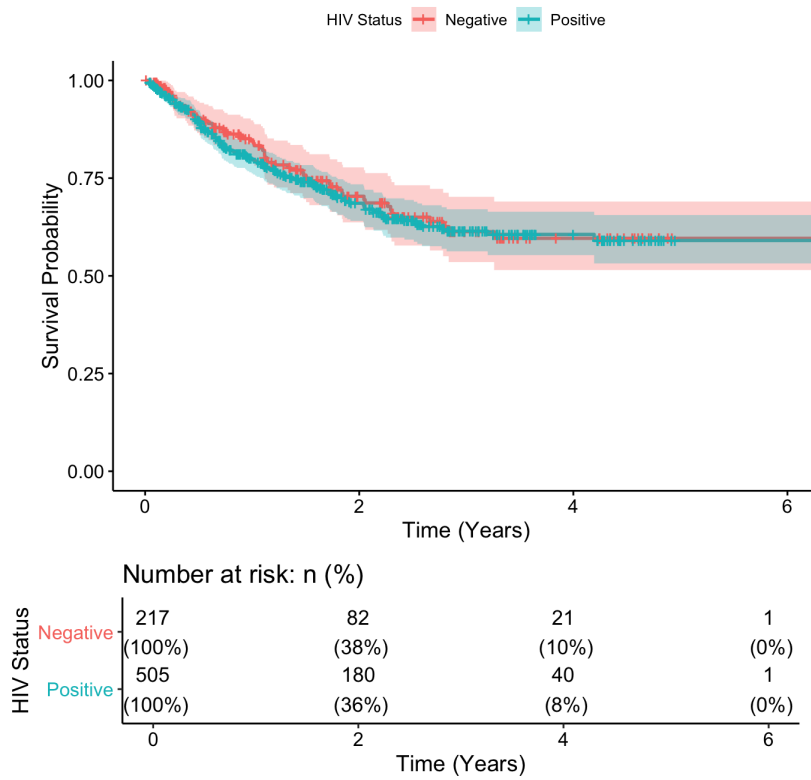
a. Overall Survival



Number at risk: n (%)

732	264	61	2
(100%)	(36%)	(8%)	(0%)

b. Survival by HIV Status



E. Random Forests

a. Patient-Survival Curves Grouped by HIV Status

