BIS 687 Project Proposal

Group 2 Yufei Deng, Xinyi Di, Yihan Liu

Study Aim

Human immunodeficiency virus (HIV) and its subsequent stage, known as Acquired immunodeficiency syndrome (AIDS), remain a major global public health issue. By the end of 2022, there were an estimated 39.0 million people living with HIV. The global fight against AIDS remains a critical challenge, with ongoing efforts to optimize treatment strategies and increase the survival rate and life quality for those affected by the disease. Yet a lot remains undiscovered about the effectiveness of different treatment regimens across heterogeneous populations. This research project seeks to leverage a comprehensive clinical dataset to examine the performance of four different types of AIDS treatments. The dataset comprises health records of AIDS patients from the United States, offering a valuable resource for understanding treatment outcomes. Our primary aim is to develop predictive models and analytical frameworks that can improve treatment personalization and outcomes for AIDS patients.

Specific Aim 1: Compare the Effectiveness of Different AIDS Treatment Regimens

We will conduct statistical tests to compare the effectiveness of 4 different AIDS treatment regimens in this study. We will use Kaplan-Meier curves, median survival time, and 1-year survival rate as metrics for comparing the effectiveness of different treatment types. This aim will help identify which treatment regimen offers the best outcomes for specific patient demographics or disease characteristics, contributing to personalized treatment strategies.

Specific Aim 2: Identify Predictors of Disease Progression and Death

Our second aim is to identify key risk factors of disease progression in AIDS patients. Using techniques including univariate Cox model, Machine Learning feature selection and regularization through cross-validation, we will identify demographic, clinical, and treatment-related factors that significantly impact the effectiveness of AIDS treatments. Understanding these will enable early prevention and individualized treatment plans, potentially enhancing survival rates and improving patient quality of life.

Specific Aim 3: Survival Prediction Using Cox Proportional Hazards Model

We will employ the Cox Proportional Hazards (PH) model, using selected features from Aim 2, to model and predict survival outcomes in the AIDS patients. By fitting this model with the dataset, we intend to quantify the impact of different treatment approaches on patient survival

times, accounting for variables such as age, gender, and baseline health status.

Through these specific aims, our project intends to use the clinical dataset on AIDS treatments to advance personalized medicine approaches for AIDS patients, thereby improving treatment outcomes and enhancing the quality of care.

Research Strategy

Significance

This project is crucial for addressing knowledge gaps by evaluating AIDS treatment regimens, identifying success predictors, and using predictive modeling to improve patient outcomes. It promises significant clinical implications, enabling personalized treatment plans for AIDS patients to reduce morbidity and mortality, and enhance life quality.

Innovation

The project introduces advanced statistical and machine learning techniques to analyze AIDS treatment outcomes. By leveraging a comprehensive dataset, it moves beyond traditional research methods, offering novel insights into treatment effectiveness and personalization.

Research Plan

The study utilizes data from a U.S. and Puerto Rico AIDS clinical trial (December 1991 - October 1992), involving 2467 individuals across various treatment regimens. Methodologically, it incorporates Kaplan-Meier estimates, log-rank tests, and adjusted Cox models to analyze treatment efficacy. Key challenges include managing data on adverse effects and addressing participants lost to follow-up.

Building on this, our research intends to advance the understanding of AIDS treatment efficacy and patient outcomes by incorporating advanced statistical and machine learning methods. We plan to compare treatment regimens using stratified Cox proportional hazards models, enhanced by machine learning to tackle high-dimensional data and uncover complex treatment-patient interactions. To overcome data challenges like incompleteness and loss to follow-up, we'll employ multiple imputation and sensitivity analyses. Additionally, we aim to personalize treatment through unsupervised learning methods, like cluster analysis, to identify patient subgroups based on treatment responses. For survival prediction, we'll refine the Cox model with techniques like Lasso and Ridge regression to boost prediction accuracy and interpretability, addressing the study's challenges and expanding on its findings with sophisticated analytics.

Specific Aim 1: Compare the Effectiveness of Different AIDS Treatment Regimens

Hypothesis: We hypothesize that among the four different AIDS treatment regimens analyzed, certain regimens will demonstrate superior effectiveness in terms of patient survival rates, increase in CD4 cell counts, and other AIDS-related events across varied patient demographics and disease characteristics.

Rationale: Given the randomized assignment of treatments to patients, this study design enables a clear evaluation of the intrinsic effectiveness of each regimen without considering the specific characteristics of different patients. The identification of the most effective treatments across a randomized patient population provides a baseline understanding of treatment performance, essential for subsequent analysis aimed at personalization.

Experimental Approach Statistical analysis including Kaplan-Meier survival curves, along with the log-rank test will be employed to compare survival rates across different treatment regimens. Median survival time and 1-year survival rate will be used as primary metrics for comparison. Further, we will analyze increases in CD4 cell counts and explore the moderating effects of patient disease history and demographics on treatment outcomes.

Interpretation of Results The effectiveness of different AIDS treatment regimens evaluated by identifying treatments that statistically outperform others in terms of survival and immunological outcomes.

Potential Problems and Alternative Approaches The findings may not be generalizable to all patient populations, especially those outside of the United States or those not represented in the dataset. We will discuss the limitations of our study regarding generalizability and suggest areas for future research to validate and extend our findings to broader populations.

Specific Aim 2: Identifying Predictors of Disease Progression and Death

Hypothesis: We hypothesize that demographic (e.g., age), clinical risk factors (e.g., homosexuality, Karnofsky score), and other treatment-related factors (e.g., previous antiretroviral treatment) will significantly influence the effectiveness of AIDS treatments and AIDS progression. Moreover, different definitions of the clinical endpoint will impact the feature selection results.

Rationale: This analytical approach is designed to dissect the complex interplay between patient characteristics and treatment outcomes, thus discovering the key factors that predict disease

progression in AIDS patients. We propose to leverage the combining power of machine learning techniques, Cox regression models and regularization methods to provide a comprehensive understanding of the variables that drive treatment efficacy. Identifying these predictors will allow for the customization of treatment plans tailored to the individual patients, potentially enhancing treatment success rates and improving patient quality of life.

Experimental Approach: We will first explore the dataset using univariate Cox regression models to obtain the significance of each of the variables. Following this, machine learning techniques and regularization methods, integrated within a 5-fold cross-validation framework, will be utilized to perform feature selection on potential prognostic factors for disease progression. Separate analyses will be conducted on different endpoints (i.e., Disease progression into AIDS or death).

Interpretation of Results: The outcomes of this analysis will highlight the factors most significantly associated with HIV progression and death. This multifaceted approach ensures the robustness and generalizability of our findings. It also has the advantage of allowing for the evaluation of a wide range of predictors, potentially expanding the dataset to include genomics data and other biological metrics.

Potential Problems and Alternative Approaches: The primary challenge lies in the complex interactions between predictors that may not be readily apparent or easily modeled. To address this, we may employ advanced machine learning models capable of capturing non-linear relationships and interactions, such as gradient boosting models. Moreover, the heterogeneity in disease progression across different populations will necessitate the validation of our model across diverse patient groups to ensure the generalizability of our results.

Specific Aim 3: Survival Prediction Using Cox Proportional Hazards Model

Hypothesis: We hypothesize that using the Cox Proportional Hazards Model, we will be able to predict the survival outcomes in AIDS patients with particular treatment regimens and significant clinical characteristics revealed from Specific Aim 2 exerting a measurable impact on patient survival times.

Rationale: Employing the Cox Proportional Hazards Model enables the quantified assessment of the relative effect of each treatment regimen on survival, controlling for patient-related factors such as age, gender, and baseline health status. This approach allows for a nuanced analysis of how these variables interact to influence patient prognosis.

Experimental Approach: We will apply the Cox Proportional Hazards Model to our clinical trial data, with survival time as the primary outcome. Independent variables will include treatment regimen, demographic factors, and clinical characteristics from Specific Aim 2. We will evaluate the model's assumptions, including proportionality of hazards, and use time-dependent covariates if necessary to account for changes over time.

Interpretation of Results: The model will identify the most influential predictors of survival by its significance, providing a hazard ratio for each treatment and clinical characteristic. These results will help discern which regimens offer the best chance of prolonged survival and under what conditions.

Potential Problems and Alternative Approaches: The model assumes proportional hazards, which may not hold for all variables. If violated, we will explore stratified models or extended Cox models with time-varying covariates. Other powerful statistical models that can account for time-varying effects, including additive models, accelerated failure time models, regression splines models and fractional polynomials could also be considered. Additionally, the model's predictive accuracy may vary with patient demographics, necessitating validation across diverse populations.