Raheela Charania, Anmol Anchala, Prabhu Shankar

Team B2: Executive Summary

Problem

The human immunodeficiency virus (HIV) is a virus that weakens the immune system. Specifically, it attacks CD4 T-helper cells which are normally involved in activating other immune cells, releasing cytokines, and developing into memory T cells for future infections.[1] According to the World Health Organization, there are about 39.9 million individuals infected with HIV in 2023.[1] Although, there is no known cure for HIV; antiretroviral treatments (ART) are able to suppress viral replication and therefore viral load, preventing the transmission of HIV and allowing individuals to live relatively normal lives. An effective ART also results in an increase in CD4 levels, allowing for the immune system to recover. As such, determining which ARV regimen to prescribe a patient is vital. However, there exists more than 30 ARV FDA-approved drugs.[2] Therefore, there is an increasing need to standardize the criteria for prescribing specific regimens. Today, we review multiple combinations of ARV treatments prescribed to 187,236 patients to create a model that can best determine which regimen will be most effective for a particular patient.

Method

In order to determine the best way to analyze this dataset, we first familiarized ourselves with the data. This includes reviewing the data dictionary, searching for missing values and reviewing the various treatment options as well as the other variables given. We then conducted exploratory data analysis to better understand the dataset.

Based on the guidelines, we needed to create a model that could determine the best regimen to decrease the viral load to at least ≤ 250 copies/mL, preferably ≤ 50 copies/mL and have a CD4 of 500 cells/mm^3 or better at first, second and first instances, respectively. First, we conducted a join to obtain the specific months and corresponding data for each patient that met this criteria. To designate the patients that met this criteria, a new column called “All Criteria Met” was created with “yes” indicating that all three criteria were met and “no” indicating that the three criteria were not met. At this point, we observed an intriguing statistic- the percentages of “yes” and “no” responses vary widely across the different ethnicities (Fig.#1).

We then set the seed at 43 and divided the dataset into a training (80%) and testing (20%) dataset. This was in order to ensure that the model we create can be applied to a previously untouched dataset. We found that our dataset was unbalanced based on the percentages of Yes vs. No in the All Criteria Met column and balanced our training dataset. Based on our earlier observations, we also subset the dataset into the four different ethnic groups, Asian, Black, White, and others. From here, we created a random tree model for each of the ethnic groups, determined the accuracy on the test dataset and analyzed our results. We also created a logistic model for the overall dataset, again with “All Criteria Met” as the target variable and determined the accuracy on the test dataset.

Solution

We created four random forest models separated by ethnicity. Below are the accuracy levels of each model (Fig#2). We also created one overall logistic model. Both of these models can be used to predict the effective treatment regimen for a particular patient. However, the accuracy of each of the four random forest models is higher than the accuracy of the logistic model, indicating the importance of ethnicity for determining the effectiveness of the treatment.

Benefit

These new random forest models can potentially reduce the monthly expenses for newly diagnosed patients. Currently, these patients spend an average of $2,567 per month, with ART costs ranging from $1,800 to $4,500.[3] By decreasing this cost, we can decrease financial burdens on patients and also decrease healthcare resources utilized from changes in drug regimen.

Future Recommendations & Analysis

Although this is an initial look into the data, we believe ethnicity is correlated with the type of regimen that will be effective for a particular patient. In accordance with this, we believe it will be beneficial to provide providers training on how ethnicity can impact treatment outcomes for patients. Furthermore, to continue developing the model, we would like to include other demographic variables that may affect HIV treatments.

Fig.#1

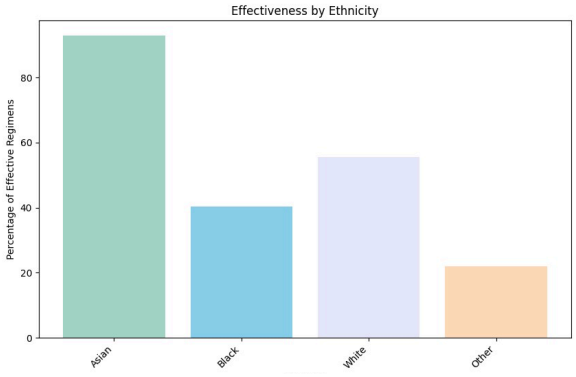

Fig.#2

Model	Accuracy
Random Forest with Asian Ethnicity	91.78%
Random Forest with Black Ethnicity	76.05%
Random Forest with White Ethnicity	70.08%
Random Forest with Other Ethnicity	81.80%
Logistic Regression with All Data	69.86%

References:

1. World Health Organization. (2024, July 22). *HIV and AIDS*. World Health Organization. Retrieved February 8, 2025, from https://www.who.int/news-room/fact-sheets/detail/hiv-aids.
2. **U.S. Department of Health and Human Services.** *What to Start: Initial Combination Regimens for the Antiretroviral-Naive Patient*. ClinicalInfo.HIV.gov. Accessed February 8, 2025. https://clinicalinfo.hiv.gov/en/guidelines/hiv-clinical-guidelines-adult-and-adolescent-arv/what-start-initial-combination.
3. Journal of Health Economics and Outcomes Research. (2023). *Economic burden of HIV in a commercially insured population in the United States*. Journal of Health Economics and Outcomes Research. Retrieved February 5, 2025, from https://jheor.org/article/56928-economic-burden-of-hiv-in-a-commercially-insured-population-in-the-united-states