A Randomized Clinical Trial's Insights on HIV Treatment

Betsy Fridman and Vaibhav Jha

December 13, 2025

1 Introduction

Human immunodeficiency virus (HIV) is a sexually-transmitted disease that is the focus of much clinical research today. This paper extracts insights from a randomized clinical trial (RCT) comparing therapies in adults infected with HIV. The RCT focused on potential treatment options to improve patient health, with data on patients' prior HIV treatment history, demographics, and performance over a time period of 96 weeks. This analysis focuses on patient response to therapy related to stage of disease, prior drug exposure, and duration of therapy.

CD4 cells, or helper T cells, coordinate immune responses and are the primary target of HIV infection, which leads to immune suppression as these cells are depleted. CD8 cells are the cytotoxic T cells that detect and destroy infected cells, including those carrying HIV. These cell types work together, where CD4 T cells help activate the CD8 T cells, and the CD8 T cells in turn control viral replication by eliminating infected cells. Monitoring CD4 and CD8 levels provides important insight into disease progression and the effectiveness of therapy.

Given the study, we asked the following questions. Are there significant relationships between stage of disease and prior HIV therapy history with response to the drug intervention? Further, among less advanced HIV disease patients, is there evidence of prolonged clinical benefit with drug intervention?

We found that a patient's stage of disease greatly influenced their estimated response to HIV therapy with drug intervention. On the other hand, a patient's prior HIV drug exposure was not a significant indicator of the patient's response to HIV therapy. Finally, among patients with less advanced HIV disease, there was evidence of prolonged clinical benefit from drug intervention. This paper explores each of these findings in depth, providing insight into the efficacy of the experimental drugs used to treat HIV.

2 Exploratory Data Analysis and Summary

The data set used to perform this analysis involves 2139 observations with 27 variables. 776 of the total patients were taken off treatment by 96 weeks. These patients were removed in order to assess the true effect of the therapy in our analysis, leaving the data with 1,363 patients. Further, there are censored patients at week 96, with 797 observations having a missing cd4.96 value. These observations were excluded from the question 3 of the analysis assessing prolonged clinical benefit, for which remained 1075 patients. After removing the patients with missing cd4.96 values, the distribution of the data looked fairly similar, indicating that there is no apparent reason for the missing data. The following variables were chosen to be included in the analysis because we believe they have a theoretical effect on the response variable.

Table 1: Variable Descriptions for HIV Trial Dataset.

cd4, cd4.20, cd4.96	CD4 cell count at baseline, 20 weeks, 96 weeks	Continuous
cd8	CD8 cell count at baseline	Continuous
age	Age in years at baseline	Continuous
wt	Weight in kg at baseline	Continuous
race	Race (0=white, 1=non-white)	Discrete Nominal
gender	Gender (0=female, 1=male)	Discrete Nominal
hemo	Hemophilia (0=no, 1=yes)	Discrete Nominal
drugs	History of intravenous drug use (0=no, 1=yes)	Discrete Nominal
symptom	Symptomatic indicator at baseline (0=asymptomatic, 1=symptomatic)	Discrete Nominal
treat2	Treatment arm (0=Drug A, 1=Drugs A+B, 2=Drugs A+C, 3=Drug B Only)	Discrete Nominal
oprior	HIV therapy other than Drug A prior to initiation of study treatment (0=no, 1=yes)	Discrete Nominal
a30	Drug A use in the 30 days prior to treatment initiation (0=no, 1=yes)	Discrete Nominal
predays	Number of days of previously received HIV therapy	Continuous

2.1 Exclusion of Variables

The Karnofsky score, although commonly used as a clinical measure of patient functioning, was removed due to its subjective nature and its possible overlap with CD4. Including it risked collinearity with the CD4 cell count at baseline covariate of interest, without adding much benefit. The HIV therapy history stratification variable was also excluded. After cross-checking its values with predays, its categorical bins did not accurately reflect patients' true duration of prior therapy. We used predays, the number of days of prior HIV therapy, as the sole measure of prior HIV treatment exposure, binning the values to show the different effects for the different levels of prior exposure.

Additional variables were omitted when they either lacked variation, duplicated information in other variables, or had no direct relevance to the individual's response to therapy. For example, the variable indicating homosexual activity relates to likelihood of HIV acquisition rather than treatment response. The variable aprior was excluded because it provided no insight; all patients received Drug A HIV therapy prior to the start of the study. If there was variation in this measure, we would have included the variable due to its confounding nature in explaining the response variable.

2.2 One-Dimensional Exploratory Data Analysis

Our exploratory data analysis began with identification of the response variable and covariates of interest. Change in CD4 cell count was the response variable in this analysis, as we feel that it most accurately represents the change in HIV level. The exact weeks used to assess the difference varied by type of analysis; baseline to 20 weeks was used to assess its relationships with stage of disease and prior HIV therapy history, while 20 to 96 weeks was used to look at the prolonged effect of the trial for less advanced HIV patients. While the sample size was smaller when using data at 96 weeks, this gave us the best understanding of the long-term impact of the stage of the disease on the response variable. To measure long-term response, we chose to use a response variable of the change from 20 to 96 weeks because the response in the first 20 weeks is highly variable, and would not help us isolate the sustained treatment effect.

The covariates of interest in this analysis were CD4 cell count and the number of days patients received HIV therapy prior to the study. All other variables in Table 1 were evaluated as important confounding variables in the context of modeling the response variable with the covariates of interest. Figures 1 and 2 show the spreads of our response variables and the main covariates of interest.

Figure 1: The distributions of response variables are relatively normal.

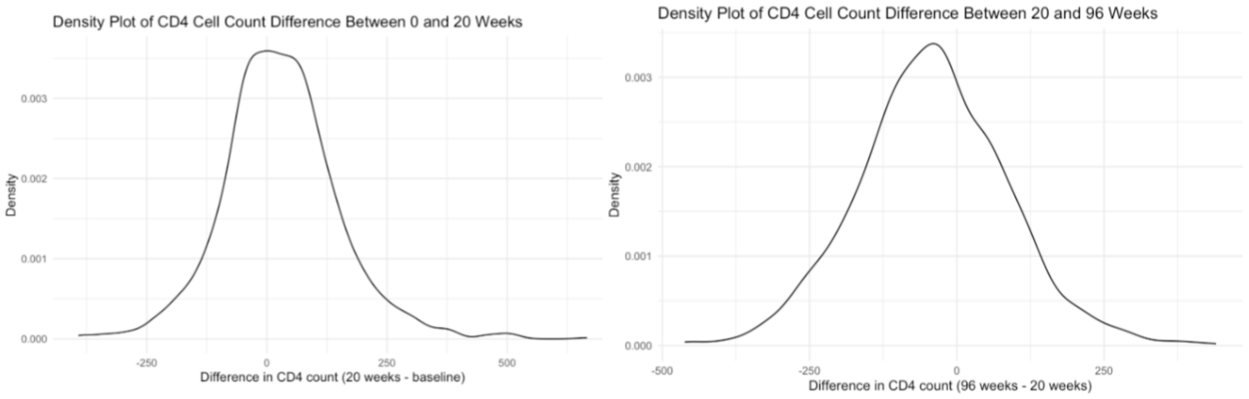
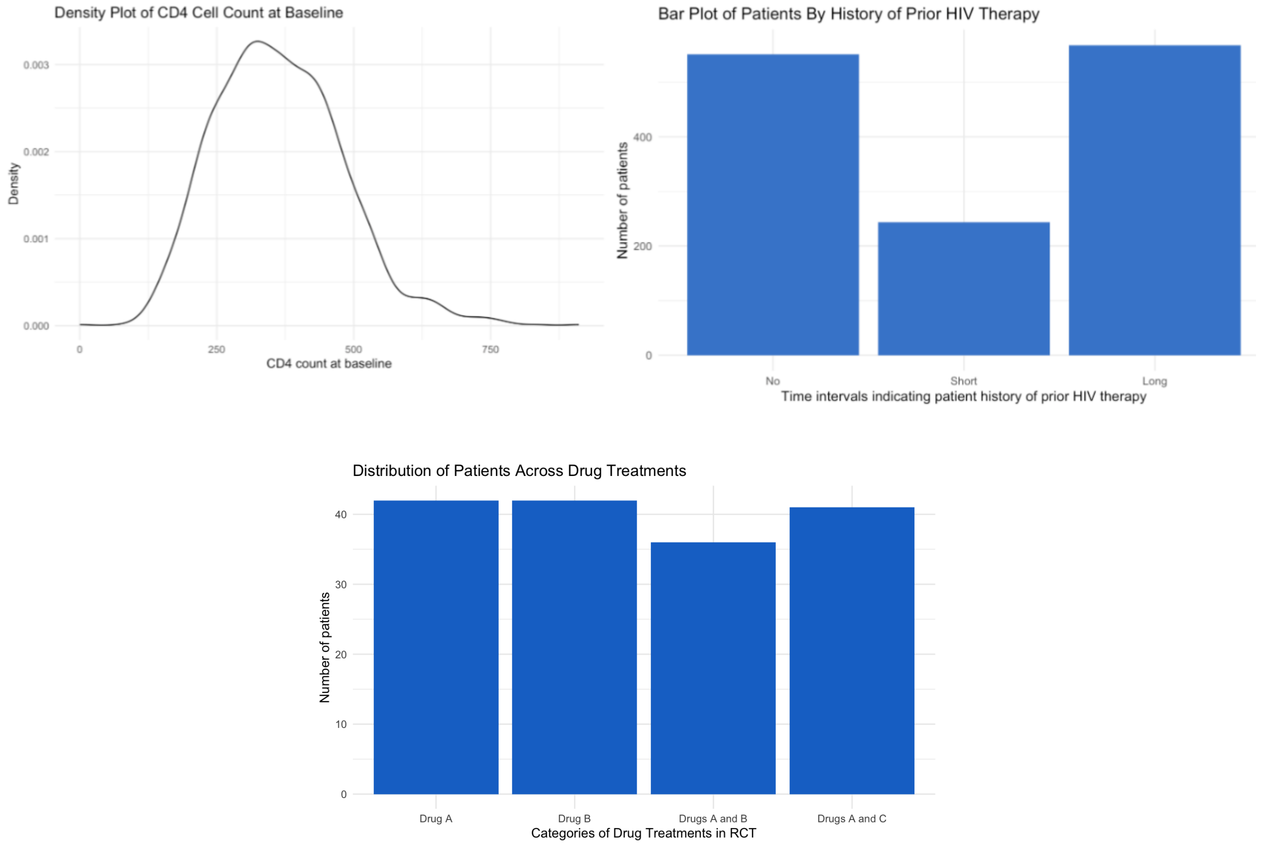


Figure 2: The treatment arm variable (treat2) shows a relatively uniform distribution.



2.3 Two-Dimensional Exploratory Data Analysis

The scatterplots and boxplots in Figures 3 and 4 illustrate the relationship between the change in CD4 counts after 20 weeks of therapy and each covariate. Among the continuous variables, baseline CD4 counts show the strongest association, with a slight negative correlation of -0.28, indicating that patients with higher initial CD4 counts tended to experience slightly smaller increases. From these visuals, we can infer that the change in CD4 counts is influenced by a multitude of factors, none of which alone explain the variation.

CD8 count was included because it captures cytotoxic T-cell activity. It is highly correlated with immune activation, and we have strong reason to believe that it causes our response variable. Although CD8 did not show a strong relationship with CD4 change, we know that it is clinically important, and its potential confounding influence justifies its inclusion in the multivariable model.

Median change in CD4 differed slightly across treatment arms, with the highest median improvement in the group with Drug A and B and the lowest in the group with Drug A only. There is no visible strong association between the main covariates and the response variable. Although it is not fully visible in our sample, we believe that there is a relationship between the covariates and the response variable. Despite this, we believe that they have a theoretical relationship with the response variable.

Figure 3: The important discrete covariates and their relationship with cd4.20 - cd4

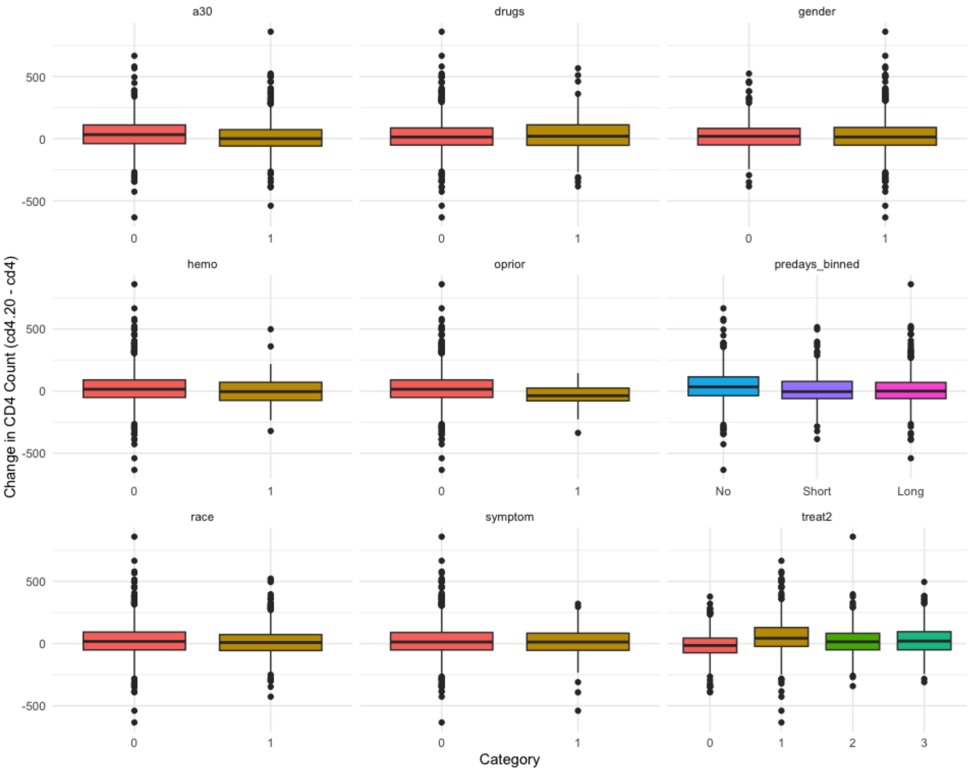
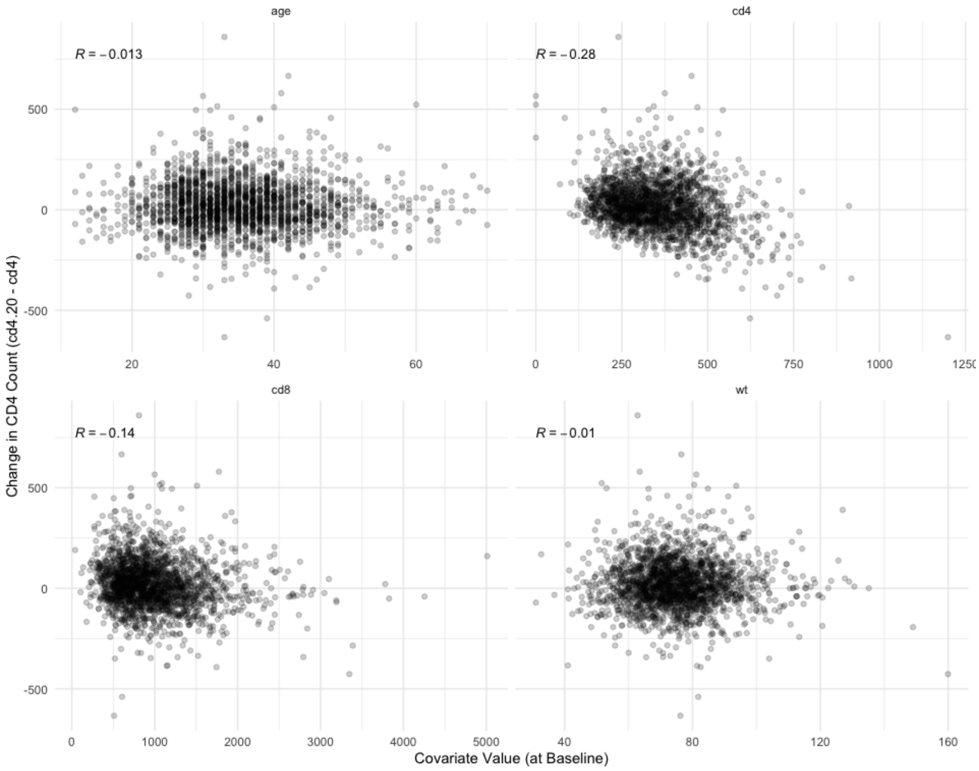


Figure 4: The important continuous covariates and their relationship with cd4.20 - cd4



2.4 Creating New Covariates

The sole covariate created was a categorical version of the days of prior HIV history variable, predays. Collected as a continuous variable, the granularity ensured that data collection was correct. We initially decided to use the HIV therapy history stratification variable to measure prior HIV therapy history. Upon exploratory data analysis with predays, there were several inaccuracies with its data. Patients were bucketed into categories that were inconsistent with the true number of days of prior HIV therapy they received.

We recognized that these intervals were likely created with intentionality on the clinicians' side at 52 week markers. Therefore, we decided to use the predays data and bin the values according to the interval lengths determined by the original covariate of interest. This allowed us to maintain accuracy with the data analysis process and group data points in important intervals according to domain experts. The predays variable was binned following the original clinical categories, which used 52-week thresholds (0 days = 'None', 1-364 days = 'Short', ≥365 days = 'Long').

2.5 Limitations

Several limitations should be noted. There is a substantial number of missing values for the CD4 count at 96 weeks, 797, which may reduce the reliability of the long-term analysis. Measurements such as CD4 and CD8 counts, as well as self-reported data such as history of drug use, may be subject to measurement error. Additionally, the patients included in this study may not be representative of the broader HIV-positive population, and 776 participants dropped therapy during the study, which could bias treatment effect estimates. Finally, some variables, such as age, baseline CD8 counts, and prior therapy history, may be correlated with CD4 counts. This collinearity can affect the stability and interpretability of regression models by making it harder to distinguish the independent effect of each variable on CD4 outcomes.

3 Methods

To study the influence of stage of disease on response to HIV therapy, we analyzed the significance of CD4 cell count at baseline when modeling the difference in CD4 cell count between baseline and 20 weeks. We fit a linear model regressing the change in CD4 after 20 weeks on the baseline CD4 cell count, controlling for age, weight, hemophilia, drug use, oprior, a30, short therapy history, long therapy history, race, gender, symptom, Drug A and B treatment, Drug A and C treatment, and Drug B only treatment. To answer our question we construct a 95% confidence interval for the coefficient of the baseline CD4 count.

$$\text{cd4}_{20} - \text{cd4} = \beta_0 + \beta_1 cd4 + \beta_2 age + \beta_3 weight + \beta_4 hemo + \beta_5 drugs + \beta_6 oprior + \beta_7 a30 + \\ \beta_8 short\_therapy\_history + \beta_9 long\_therapy\_history + \beta_{10} race + \beta_{11} gender + \beta_{12} symptom + \\ \beta_{13} drug\_B\_treatment + \beta_{14} drug\_AandB\_treatment + \beta 15 drug\_AandC\_treatment + \epsilon$$

The second research question involved studying how prior HIV therapy status influenced response to drug intervention, if at all. Using the same model fitted to answer the first research question on stage of disease, we conducted a partial F test to test the significance of prior HIV therapy patient history in modeling response to drug intervention.

Finally, the third research question used a similar model to the previous one to assess the effect of drug intervention on prolonged clinical benefit for less advanced HIV patients, keeping all of the covariates the same but with a response variable of the change in CD4 count after 96 weeks. By looking at the coefficient of the baseline CD4 count and its 95% confidence interval, we are able to check if there is a prolonged benefit from the therapy.

$$\mathrm{cd4}_{96} - \mathrm{cd4}_{20} = \beta_0 + \beta_1 cd4 + \beta_2 age + \beta_3 weight + \beta_4 hemo + \beta_5 drugs + \beta_6 oprior + \beta_7 a30 + \beta_8 short\_therapy\_history + \beta_9 long\_therapy\_history + \beta_{10} race + \beta_{11} gender + \beta_{12} symptom + \beta_{13} drug\_B\_treatment + \beta_{14} drug\_AandB\_treatment + \beta 15 drug\_AandC\_treatment + \epsilon$$

Following model building, a residual analysis was conducted to ensure all assumptions about the error term ϵ were met prior to inference. In a linear regression model, ϵ is assumed to have a mean of zero, constant variance, independence, and to follow a normal distribution. These assumptions are crucial because violations can lead to biased standard errors, invalid confidence intervals, or incorrect p-values, which therefore can compromise the inference.

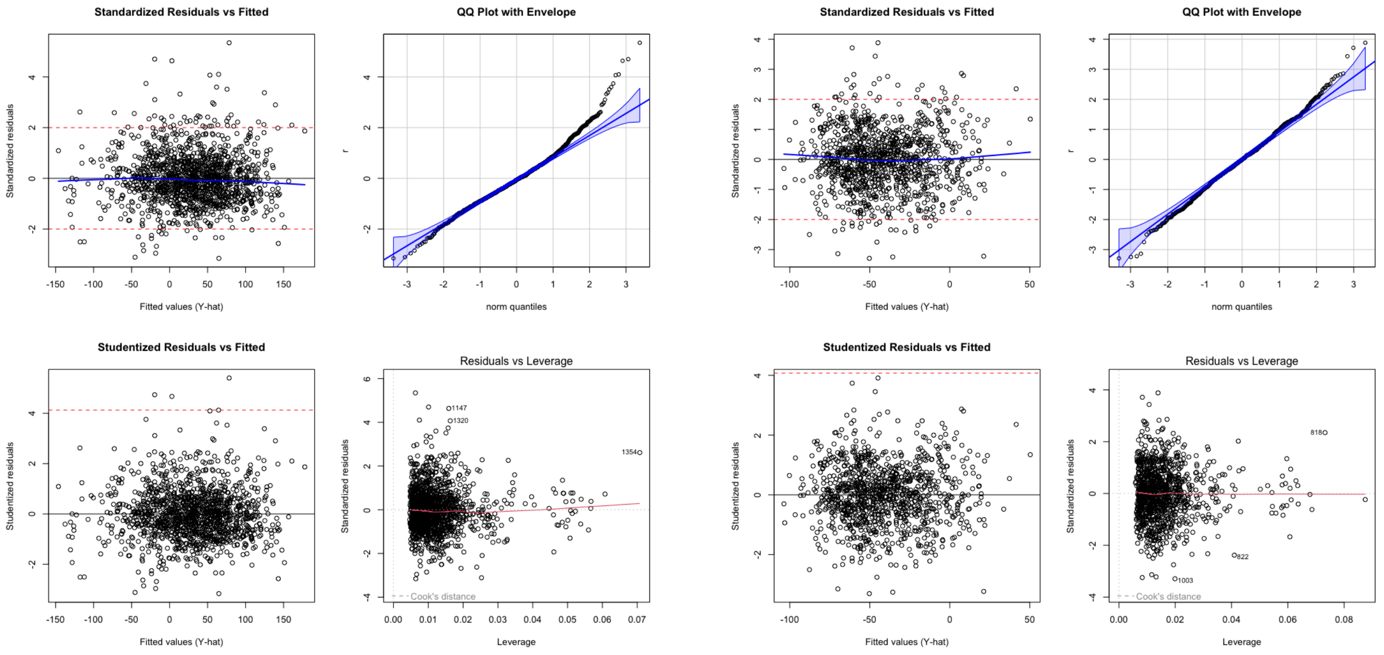
Figure 5a shows the diagnostics for the models of research questions 1 and 2, and Figure 5b shows the diagnostics for the model of research question 3. Linearity and constant variance were assessed using standardized residuals vs. fitted values (top left plots). There is a random scatter of points around zero and there is no systematic pattern or fanning of the datapoints, which indicates that the linearity and homoscedasticity assumptions are reasonable.

The independence assumption was satisfied because the data were collected from a randomized clinical trial. Each observation can be treated as independent, with limited or no influence on other observations' outcomes, assuming proper trial implementation.

Normality of the error term was assessed using Q–Q plots (top right plots). The residuals approximately follow a straight line along the theoretical quantiles, with only slight deviations, supporting the assumption that ϵ is normally distributed. This validation allows us to do reliable hypothesis testing and confidence interval estimation.

From looking the standardized residuals vs. leverage plots, we can see that none of the points have a large Cook's distance or are too influential, so there is no need to remove any points from the dataset.

Figure 5: All assumptions for linear regression are met for Research Questions 1, 2, and 3.



(a) Models 1 and 2

(b) Model 3

4 Results

After fitting the model for research question 1, the CD4 cell count at baseline parameter's coefficient was estimated to be -0.295 (Table 2). Its estimate produced a 95% confidence interval [-0.34, -0.24].

Holding all other covariates constant in the model for research question 1, we are 95% confident that for every one unit increase in CD4 cell count at baseline, there is a decrease in change of CD4 cell count between baseline and 20 weeks between 0.24 and 0.34 units.

To assess the influence of prior HIV therapy on response to drug therapy, the partial F test did not provide evidence to suggest a significant relationship. Specifically, the category of prior HIV therapy – none, short duration, and long duration – did not have a significant influence in modeling the change in CD4 cell count between baseline and 20 weeks. The ANOVA table in Table 3 shows this. The test produced $\Pr(> F) = 0.2303$.

For the final analysis, we produced Bonferroni-adjusted 95% confidence intervals to assess the drug treatments in comparison to Drug A—a treatment that was determined to not confer prolonged benefits for less advanced patients. Table 4 shows that adding treat2 significantly improves the model. The confidence intervals for Drug B, Drug A and B, and Drug A and C interventions were determined to be [-6.6904, 50.6504], [-16.3376, 42.0976], and [18.6540, 77.5260], respectively. For each of these drug interventions, their 95% estimated effect on CD4 cell count between 20 and 96 weeks in comparison to having only Drug A falls between these ranges.

Table 2: The coefficients of cd4, cd8, and the factors hemo, oprior, race, treat2 are significant.

term	estimate	std error	statistic	p-value
(Intercept)	137.112	24.864	5.514	0.000
cd4	-0.295	0.027	-10.957	0.000
age	-0.558	0.362	-1.540	0.124
wt	0.239	0.252	0.952	0.341
as.factor(hemo)Yes	-25.032	11.529	-2.171	0.030
as.factor(drugs)Yes	8.815	9.874	0.893	0.372
as.factor(oprior)Yes	-82.301	23.234	-3.542	0.000
as.factor(a30)Yes	-18.439	15.715	-1.173	0.241
as.factor(predays_binned)Short	-26.254	16.309	-1.610	0.108
as.factor(predays_binned)Long	-27.903	16.546	-1.686	0.092
as.factor(race)White	21.735	7.141	3.044	0.002
as.factor(gender)Male	-3.049	9.049	-0.337	0.736
as.factor(symptom)Symptomatic	-16.093	8.516	-1.890	0.059
cd8	-0.018	0.007	-2.744	0.006
as.factor(treat2)Drugs A + B	72.248	8.607	8.394	0.000
as.factor(treat2)Drugs A + C	30.650	8.745	3.505	0.000
as.factor(treat2)Drug B alone	45.294	8.438	5.368	0.000

Table 3: The hypothesis for question 2 has failed to reject, predays does not significantly influence HIV outcomes from treatment.

Model	Res.Df	RSS	Df	Sum Sq	F	Pr(>F)
Model 1	1348	16,391,396	–	–	–	–
Model 2	1346	16,355,668	2	35,728	1.470	0.2303

Table 4: ANOVA Comparison of the Models for Question 3.

Model	Res.Df	RSS	Df	Sum Sq	F	Pr(>F)
Model 1	1061	17,042,630	–	–	–	–
Model 2	1058	16,725,243	3	317,387	6.6924	0.0001782

Table 5: Drug Treatment Performance Over Prolonged Time Period

term	estimate	std error	statistic	p-value
as.factor(treat2)Drug B alone	21.98	10.86	2.025	0.0432
as.factor(treat2)Drugs A + B	12.88	11.09	1.161	0.2459
as.factor(treat2)Drugs A + C	48.09	11.15	4.313	1.76e-05

5 Discussion

This randomized clinical trial assigned HIV drug therapies to assess relationships of stage of disease and prior HIV therapy history with response to drug intervention. The resulting confidence interval for the stage of disease does not include 0. Specifically, for every unit increase in CD4 cell count at baseline, we are 95% confident that there is a decrease in change of CD4 cell counts after 20 weeks between 0.24 and 0.34 cells. This demonstrates a significant relationship between stage of disease and response to drug intervention, holding all other confounding variables constant. Moreover, in studying prior HIV therapy history, we find that there is not a significant relationship between prior HIV therapy status and the response to drug intervention. We cannot estimate any significant impact on response to drug intervention given prior HIV therapy status.

In studying the prospect of prolonged clinical benefit for less advanced HIV patients, we find that there is significant prolonged clinical benefit for patients that received both Drugs A & C and Drug B. Given information that Drug A does not confer prolonged therapy in less advanced HIV patients, this study shows that the aforementioned treatments give patients significant prolonged clinical benefit in comparison over a 96 week span. Over a span of 96 weeks, we are 95% confident that Drugs A & C increase CD4 cell count between 27 and 69 cells and that Drug B increases CD4 cell count between 1 and 43 cells.

This study contains several limitations, constraining potential implications. As referred to earlier, there are several inconsistencies with data collection across the experiment. This involves variables that were of main interest like HIV therapy history stratification, forcing us to settle for secondary options to answer our research questions.

Another key limitation surrounds the third research question. Missing data is a problem in studying prolonged clinical benefit; many patients were censored at 96 weeks. It is not clear whether these patients left the study due to death or other unrelated factors. The lack of specification forced us to remove all observations with missing data. This resulted in optimistic interpretations of drug interventions, only involving those who remained in the study at 96 weeks.

To address these limitations, this experiment would have to be repeated with several modifications. Clearer data collection practices would have to be instituted in order to ensure precise data analysis. Further, more information would have to be collected as to how patients leave the study and why. A given reason for patient withdrawal from the study, whether due to death or unrelated factors, is important to study drug intervention's potential prolonged benefit.