IL-2 Follow-Up Study Longitudinal Analysis Report

Wei Dai and Joshua Wu

**Introduction:**

The dataset that was provided to us came from a previous paper of a Randomized Control Trial by Kovacs et al. Interleukin-2 (IL-2) is a cytokine that facilitates the creation and differentiation of white blood cells. The study intervention is infusing IL-2 into the body to increase the CD4 count, which is a proxy measurement to assess the strength of the immune system. The prevailing theory is that increasing CD4 count will decrease the HIV viral load, and so injecting IL-2 should lead to a decrease in HIV viral load. Our objective is to assess whether there is a difference in mean viral load profiles between the two groups over the follow-up period and to assess whether there is a significant difference in odds of not obtaining a detectable HIV viral load count. Our hypothesis is that the IL-2 group will have a lower mean viral load profile than the control group and that the IL-2 group will have a higher odds of not obtaining a detectable HIV viral load count.

**Methods:**

The dataset we obtained was from a previous study in 1996 by Kovacs et al. called *Controlled Trial of Interleukin-2 Infusions In Patients Infected With The Human Immunodeficiency Virus*. Our primary outcome was “rtcprnih”, which was our HIV viral load count. To assess for normality of the outcome, we drew a histogram and a QQ plot, and performed a natural log transformation since it was not normal. After peeking into the data, we also found that there were missing data and to address that we performed imputation. Before peeking into the data, we first determined a threshold for which to consider an exclusion criteria. We knew that there were 9 time points and so we decided that each observation needed at least 3 time observations to be included in our data. We also assumed that for those observations that ended in 49 (the detectability limit) for “rtcprnih” before lost to follow-up would continue to have 49 for the rest of the study. We then performed a log transformation of the data to create a new variable called “LVLoad” and performed an FCS multiple imputation using predicted mean matching since it looked like we had some non-Monotone missing in our data. To create the outcome variable for our second objective, we used the imputed log viral load data and created a new variable: 1 - the observation had a log viral load less or equal to log(e)[49], 0 - the observation had a log viral load greater than log(e)[49].

To assess our first objective, we first used simpler methods of analysis to see if there were any difference. We performed a t-test on the overall group mean by treatment group, a t-test on the overall mean group slope by treatment group, and a Repeated Measures ANOVA assessing whether the group means were different over time. We then performed more difficult analyses, such as MANOVA, and a Linear Mixed Effects Model. It must be noted that for the linear mixed effects model, time is transformed by subtracting 14 from the all time points. This is done so that the intercept in the model will match the first time recording. For our second objective we decided on a GEE model and used a binomial distribution, a logit link, and an exchangable log odds ratio regression structure. All the analyses were done in either SAS or in R.

Due to us log transforming our outcome, we can assume normality in our outcome variable. This helps us for our t-tests and ANOVA’s, although studies have shown that ANOVA is robust against non-normal data. We also conducted simpler and more difficult analyses that roughly answer the same question. This is to assess whether our conclusions will change depending on the results. A MANOVA allows us to account for time and the Linear Mixed Effects model allows us to incorporate potential random effects in the study population. The GEE model as an alternative logistic regression model for longitudinal data was used since we wanted to calculate odds of obtaining non-detectable HIV viral loads.

**Results:**

Table 1: Descriptive Statistics\*

|  |  |
| --- | --- |
| **Variable** | **Mean (SD) / Freq (%)** |
| Log HIV Viral Load (lvload) | 7.85(2.74) |
| IL-2 Intervention Group | 28 (53.85%) |
| Detectable HIV Load | 368 (78.63%) |

\* All Descriptive Statistics were run after performing multiple imputation

In Table 1, we show the descriptive statistics of our whole study. The mean log HIV viral load count was 7.85 with a standard deviation of 2.74. We had 53.85% of our study population in the IL-2 intervention group and had 78.63% of our time observations that had a detectable HIV load accounting for all subjects.

Table 2: T-test on Overall Group Mean, Slope, and Repeated Measures ANOVA

|  |  |
| --- | --- |
| **Test** | **p-value** |
| Difference in Overall Group Means | 0.39 |
| Difference in Overall Group Slope | 0.45 |
| Repeated Measures ANOVA | 0.39 |

In Table 2, we have various simpler tests to see if there is a difference in the group mean profiles between the IL-2 and the Control group. First, was the t-test and with a p-value of 0.39, we conclude that there is no difference in mean log viral load count by treatment group. The IL-2 group had a mean of 7.63 with a 95% CI of (6.95, 8.31). The Control group had a mean of 8.11 with a 95% CI of (7.16, 9.06). For the t-test for the group slopes, we conclude that there is no difference in group slopes between treatment groups. The IL-2 group had a slope of -0.0778 and a 95% CI of (-0.0950, -0.0606). The Control group had a slope of -0.0682 with a 95% CI of (-0.0874, -0.0490). So far, for both t-tests, we find that the IL-2 group had a lower group mean and a steeper slope, indicating that it would have a better outcome than the Control group. While this may be true, the difference between the IL-2 treatment group and the Control group was not significantly different. Finally, we have a repeated measures ANOVA test to assess whether there is a difference in mean log viral load counts over the length of the follow-up time. We see that there is no group effect on log viral load counts with a p-value of 0.39. Figure 1, below, shows the mean profiles in log viral load counts over the follow-up period. As shown, we can conclude either through numbers or through a visualization that there does not seem to be a difference in mean log viral load counts over time by treatment group.

Figure 1: Group Means of Log Viral Load Counts Over Each Follow-Up Visit

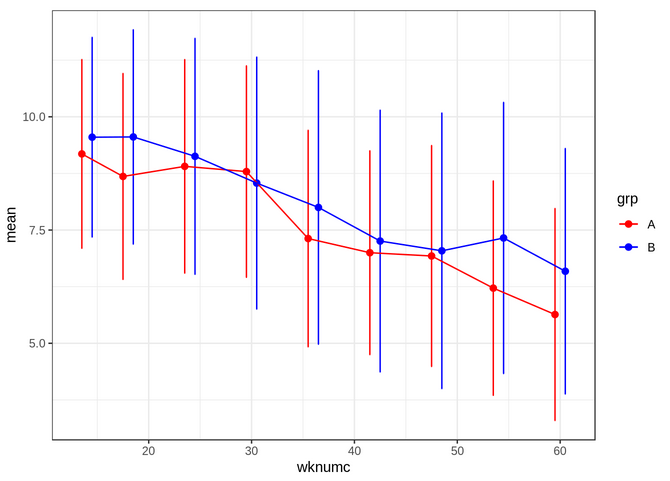


Table 3: MANOVA, Linear Mixed Effects Model, and GEE

|  |  |
| --- | --- |
| **Test** | **p-value** |
| MANOVA | 0.39 |
| Linear Mixed Effects Model | 0.68 |
| GEE | 0.64 |

The MANOVA provides a p-value of 0.3906. The results of the RMANOVA and the MANOVA should be similar and indeed they are. We conclude that there is no difference in the group mean profiles between the two treatment groups, IL-2 and Control.

Table 4: -2 Log Likelihoods for Mixed Effects Models:

|  |  |
| --- | --- |
| With Random Intercept Only | 1917.8 |
| With Random Intercept and Slope | 1909.2 |

It’s clear we want to include group effect, time effect, and group, time interaction into the fixed effect but we need to choose the best mixed model for the data. We started with a random intercept model and then fitted a model with random intercept and slope. The LRT showed a difference of 8.554 at the cost of 2 degrees of freedom. Running a chi-square test showed that the p value of the test is 0.014, which is significant. The test suggested that a more complex model (with random intercept and slope) is desired. So, we chose that model as our model for linear mixed effect model.

Table 5: Fixed Effects Parameter Estimates

|  |  |
| --- | --- |
| **Parameter** | **Estimate** |
| Intercept | 9.36 |
| Time | -0.08 |
| Group(B) | 0.27 |
| Time\*Group(B) | 0.01 |

From the fixed effects from our mixed effects model, the value of log viral load for Group A at baseline (14 months after randomization) is 9.36. For Group B it is 9.63 (9.36 + 0.27). The p-value for the group difference, as shown in Table 3, is 0.68, which is not significant. It was also found that the group-time interaction was also not significant. We conclude that we fail to reject the null hypothesis and that there is no evidence to suggest that there is a difference in log viral load count between the two treatment groups. Although the evidence shows that the IL-2 group may have a better outcome than the Control group, the difference is not significant.

Finally, we performed a GEE to see if there was a difference in odds between the two groups on obtaining a non-detectable log HIV viral load. If a subject was in the IL-2 treatment group, they are estimated to have 1.27 times higher odds of having a non-detectable log HIV viral load than those in the Control group. However, this was not significant with a p-value of 0.6411. So consistent with all our analyses, it seems that the IL-2 group yields better results, however, the differences between it and the Control group seem to be non-significant. We also obtained results of GEE in R, which gave us the same conclusion as in SAS. The probability of getting a lower viral load for group A is higher than group B as treatment goes on. However, the difference is not significant enough. Here, we attach the probability plot with time to visualize the results.

Figure 2: Probability of Obtaining an Undetectable Viral Load Count 

**Discussion:**

Strengths of our method is that we used multiple tests to corroborate our main MANOVA test and that we used a mixed effects model to help predict log viral load by subject. We also used a GEE for a logistic regression which is a method to calculate odds ratios for longitudinal data.

Our study, however, has various issues that need to be addressed. Even though we log transformed our original outcome variable (HIV viral load), the data was more normal, but was still not as normal as we would like. There was many subjects with a measurement at the lower bound of the measuring machine (<49) and they stayed in our dataset after log transformation. This could be addressed by only including cases that did not have the lower bound measurement in our analysis. That would ensure that our continuous outcome was normal, fulfilling a major assumption. In addition, we could also treat the lower bound as “0” and perhaps use a zero-inflated model to address non-normality at the lower end of our data. Our imputation method is not standard, and biases could be introduced using the method we chose. For instance, we removed 6 patients for having less than 3 visits. But it has been pointed out that these severe cases are very important and meaningful in the clinical sense. Also, the removal is unbalanced. We removed 5 out of 6 severe patients from group B and that removal could have led to a significant finding in group differences. For these reasons, these patients might be better to be included in the analysis. To see if our conclusions for our imputed data is robust, it might be appropriate to run a sensitivity analysis by running our analysis again on the raw data for all our analyses.