Statistical Analysis Plan

Background and Rationale

A trial was completed to assess the effects of different treatments on the count of CD4 cells in individuals with acquired immunodeficiency syndrome (AIDS). CD4 cells are white blood cells that fight infection in the body, and kill the human immunodeficiency virus (HIV) that can lead to AIDS. When HIV progresses, the number of CD4 cells declines.

Covariates Analysis

Categorical	**Continuous**	**Fixed**	**Random**	**Target**
TRT, SEX, TIME, AGE	CD4	TRT, SEX, TIME, AGE	ID	CD4

No factors are nested with the given data structure.

The age factor was binned using equal frequency distribution to assess the impact on age groups rather than on discrete age values. This was done to make age categorical with as balanced a number of subjects in each bin as possible. This also helps generalise the result of the study considering ages not represented in data.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
age	BIN_age	age < 30.0027	114	0.20879121
		30.0027 <= age < 35.0035	120	0.21978022
		35.0035 <= age < 39.0032	105	0.19230769
		39.0032 <= age < 45.0004	129	0.23626374
		45.0004 <= age	78	0.14285714

The target variable, CD4, was transformed into the difference of CD4 count at each time period (TIME=2,3) [ref - appendix] from the baseline (TIME=1). This was done to assess the *change* of CD4 cell count, and normalises the changes in CD4 for each participant.

$$\texttt{CD4\_diff08} = \texttt{CD4}_{time2} - \texttt{CD4}_{time1}$$
$$\texttt{CD4\_diff16} = \texttt{CD4}_{time3} - \texttt{CD4}_{time1}$$

Parameter Estimation Method

ANOVA Type III method was used to estimate parameters in the statistical analysis. Type I was not chosen as the order of terms is not known. ML was not used as, for some subsets, the degrees of freedom are quite small. REML was not used as the dataset is relatively small. As we do not have nested data, and we assume there is an interaction of effects, Type III is chosen as the most appropriate estimation method.

Testing Methodology

1. **Data randomization performance assessment**
2. **Hypothesis testing:** A null hypothesis is formed related to each research question of interest. Statistical analysis is completed to test each hypothesis and report the results (accepted/rejected) against these hypotheses. Testing of each hypothesis is completed in two steps:
    a. A 95% confidence interval is used to deem a result significant (alpha = 0.05). The p-value for each factor is evaluated to determine if there is 95% likelihood of the results occurring by chance and not actually impacting CD4 levels. With a p-value less than or equal to 0.05 means, the results are deemed statistically significant. In this case the null hypothesis (that the factor has no impact) is rejected, and the alternative one is accepted.
    b. If the alternative hypothesis is accepted, contrasts are assessed by analysing their correlation using a Pearson correlation coefficient analysis.
3. **Assumption verification:** Complete residual analysis to test if the underlying data is normally distributed to justify model parameter estimation.

The research questions that will be assessed test for significant interactions between treatments, sex, age, and time period.

Results

Data Randomization Performance

All participants have CD4 count results for each time point. There are no missing values. The data set is balanced for time (n=180) and unbalanced on treatment type (1=40, 2=48, 3=41, 4=53) and sex (Female=15, Male=167). Further, data is unbalanced within each treatment type for sex and age.

Treatment	Males	Female	F/M Ratio
1	37	3	0.08
2	46	2	0.04
3	37	4	0.11
4	47	6	0.13

Treatment	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
1	8	7	11	9	5
2	9	10	14	8	7
3	8	12	11	10	0
4	13	11	8	7	14
Total	**38**	**40**	**44**	**34**	**26**

Hypothesis Testing

The following table provides a summary of all significant findings that fall within a 95% confidence interval. Results are reported using format (F(between groups df, within groups df) = [F-value], p=[p-value]). Significance means there is at least one group with significantly different results from the reference factors.

Research Question	Significant Findings	Statistical Values
Is there a difference between treatments?	Yes	F(3, 327) = 5.96, p=0.0006
Is there an effect over time and is this different for treatments?	No	F(3, 327) = 0.06, p=0.9791
Is an age effect mediated by treatment?	Yes	F(11, 325) = 2.01, p=0.0271
Is an age effect mediated by time?	No	F(4, 327) = 0.84, p=0.5020
Is a sex effect mediated by treatment?	Yes	F(3, 328) = 4.48, p=0.0042
Is a sex effect mediated by time?	No	F(1, 327) = 0.23, p=0.6346
Is a sex effect mediated by age?	Yes	F(4, 324) = 3.66, p=0.0062

Assumption Verification

Assumptions for using ANOVA

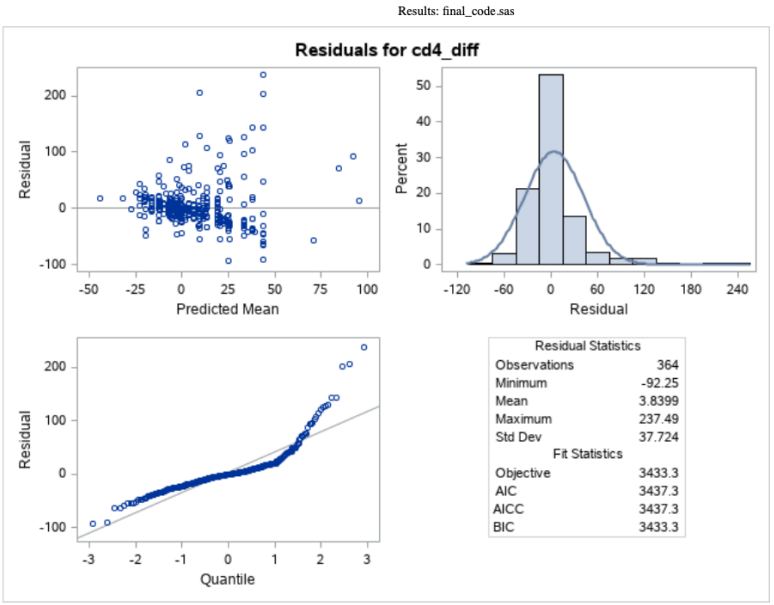
1. The independent variable is a categorical variable, and the dependent variable is a continuous variable : Our independent variables, treatment, sex, time, age, are categorical variables; dependent variable, CD4, is continuous.
2. Statistical population must be a normal distribution : According to the central limit theorem, the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution. Most of our data in different categories is over 30, except sex=female. Therefore, gender-related results may be biased.
3. Variance homoscedasticity, assumption of equal variance. Which can be tested with our code.

Testing assumptions

1. Normality of target variable:

   Test performed (verification using Residual Analysis)

   We checked the residual distribution for cd4_diff and observed that it is normally distributed.



2. Fixed and random effects - We assumed all factors are fixed except id.

Alternatives Approaches

1. REML estimator can be used if we add more data in future.
2. If Normality Assumption fails - Converging target variable to normal distribution using log, inverse, box-cox transformations.
3. If factors are random - We can also use linear mixed model to do the deeper analysis with random effects and their correlations with REML estimators.

Below is an analysis of Pearson correlation which is assess in the discussion and conclusion.

Pearson Correlation Coefficients, N = 364
Prob > |r| under H0: Rho=0

	id	trt	age	sex	BIN_age	cd4_diff
id	1.00000	0.00987 0.8512	-0.08863 0.0913	-0.13960 0.0076	-0.11942 0.0227	0.10677 0.0418
trt	0.00987 0.8512	1.00000	0.00289 0.9562	-0.07430 0.1572	-0.03110 0.5543	0.22839 <.0001
age	-0.08863 0.0913	0.00289 0.9562	1.00000	0.13829 0.0082	0.94039 <.0001	-0.00503 0.9238
sex	-0.13960 0.0076	-0.07430 0.1572	0.13829 0.0082	1.00000	0.13619 0.0093	-0.06884 0.1900
BIN_age	-0.11942 0.0227	-0.03110 0.5543	0.94039 <.0001	0.13619 0.0093	1.00000	-0.06843 0.1927
cd4_diff	0.10677 0.0418	0.22839 <.0001	-0.00503 0.9238	-0.06884 0.1900	-0.06843 0.1927	1.00000

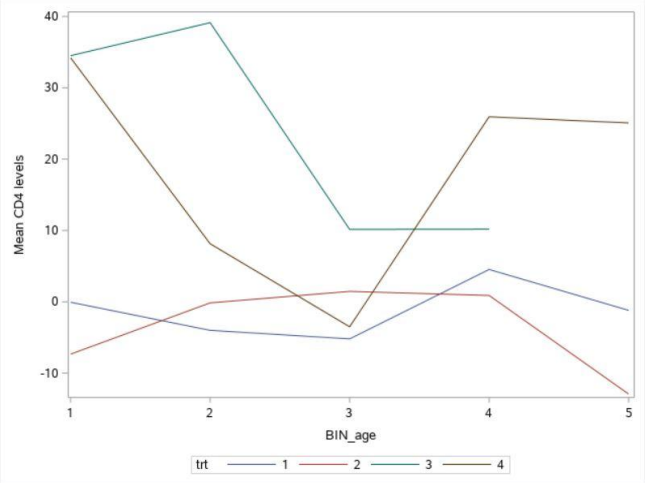
Conclusions & Discussion

The analysis of variance showed that the level of change in subject CD4 levels can be significantly impacted by certain treatments, ages and sex categories.

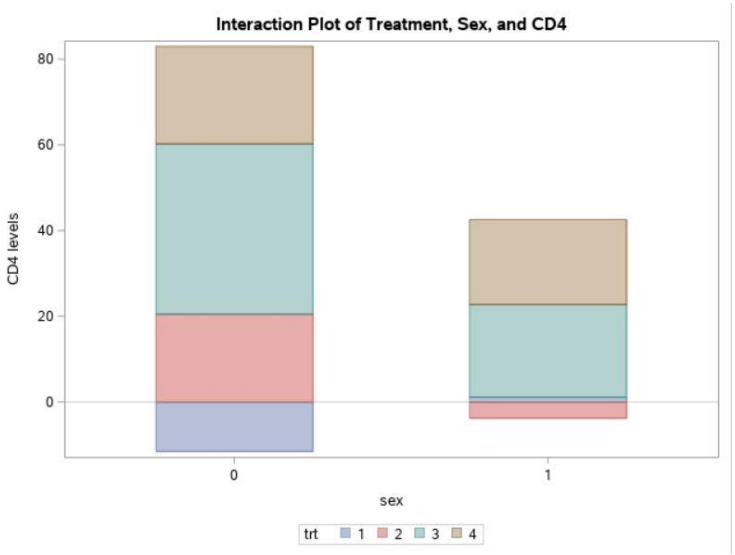
For factors that were found to have significant correlation coefficient between the change in CD4 and treatment of the patient was obtained to be 0.23. This means that the measure of the linear relationship between the CD4 levels and treatment is weak. However, the p-value of 0.0001 suggests that the linear relationship between the two factors is significantly different from 0.

Regarding the correlation of sex/age category to the CD4 levels, we notice that the correlation coefficient is almost the same =-0,07, which is very close to 0. The p-value of 0.19 also suggests that the linear relationship between them is not significantly different from zero. However, since correlation coefficient measures the strength of the **linear** relationship, we conclude that there might be some other forms of relationship (logarithmic, cubic, inverse, exponential and so on) between the sex/age and the level of CD4 of the patients.

1. Treatment 2 is less effective than Treatment 4.
2. Treatment 3 is more effective for subjects between ages 30 and 35 than subjects 39-45 years of age. This can be seen in the graph below.
   ***Note**: Conclusions cannot be drawn regarding effectiveness of treatment 3 for subjects aged 45+, as there were no data points for this age range.*



3. Treatments 2 and 3 are more effective than treatment 4 in increasing CD4 level for women.
   ***Note**: The number of females in the study is relatively small, therefore further studies should be conducted on female specific conclusions.*


Interaction Plot of Treatment, Sex, and CD4
CD4 levels
80
60
40
20
0
0
1
sex
trt 1 2 3 4

Appendix

SAS CODE -

```
/* import the assignment data */
/* ---------------------------------------------------------- */
libname SASDATA "/home/u62247656/LDA_PROJECT";

DATA assignment;
      SET SASDATA.assignment;
RUN;



/* ---------------------------------------------------------- */
/* Reshaping the dataset */
/* ---------------------------------------------------------- */


/* binning ages (equidepth) */

proc hpbin data=assignment output=assignment_bin pseudo_quantile;
   input age / numbin=5;    /* override global NUMBIN= option */
   id id;
run;

/* /sorting and merging the two datasets/ */
PROC SORT Data=assignment;
   BY id;
  RUN;

PROC SORT Data=assignment_bin;
   BY id;
  RUN;

DATA dataset;
   MERGE assignment assignment_bin;
   BY id;


/*  Transposing the dataset */

proc transpose data=dataset out=assignment_wide (drop=name) prefix=time;
```

```
    by id trt age sex bin_age;
    id time;
    var cd4;
run;

/* Add columns with differences */
Data assignment_wide;
 set assignment_wide;
 CD4_diff08 = time2 - time1;
 CD4_diff16 = time3 - time1;
run;

/* Reshape back from wide to long */
proc transpose data=assignment_wide (drop=time1 time2 time3)
out=assignment_long;
   by id trt age sex bin_age;
run;

data assignment_long;
    set assignment_long (rename=(_NAME_=time cd4=cd4_diff));
run;




/* ------------------------------------------------------------- */
/* The model: */
/* ------------------------------------------------------------- */

PROC MIXED DATA=assignment_long METHOD=Type3 COVTEST;
      CLASS trt sex bin_age time;
      MODEL cd4_diff = trt sex bin_age time time*trt bin_age*trt
bin_age*time sex*trt sex*time sex*bin_age
      /SOLUTION CL RESIDUAL DDFM=SAT;
      lsmeans trt sex bin_age time time*trt bin_age*trt bin_age*time
sex*trt sex*time sex*bin_age;
      RANDOM id; /* this keeps the individual subjects correlated to
each other */
      TITLE "ANOVA Model with Type3 Estimators";
RUN;
```

```
/* ---------------------------------------------------------------- */
/* /Checking correlation between factors/ */
/* ---------------------------------------------------------------- */
proc corr data = assignment_long;
run;
```

```
/* ---------------------------------------------------------------- */
/* Plotting the interactions
/* ---------------------------------------------------------------- */
```

```
/* /Plotting the interaction of age and treatment/ */
PROC SORT data=assignment_long out=back;
  by trt bin_age;
RUN;

PROC MEANS data=back noprint;
   by trt bin_age;
   var cd4_diff;
   output out=meaned mean=cd4;
RUN;

PROC PRINT;
   title 'CD4 for Treatment and Age Group combinations';
RUN;

PROC PLOT data=meaned;
   title 'Interaction Plot of Treatment, Age Group, and CD4';
   plot cd4*bin_age=trt;
RUN;

proc sgplot data=meaned;
series x=bin_age y=cd4 /group=trt;
yaxis label="Mean CD4 levels";
run;
```

```
/* /Plotting the interaction of age and treatment/ */
```

```
PROC SORT data=assignment_long out=back;
  by trt sex;
RUN;

PROC MEANS data=back noprint;
   by trt sex;
   var cd4_diff;
   output out=meaned mean=cd4;
RUN;

PROC PRINT;
   title 'CD4 for Treatment and Sex combinations';
RUN;

PROC PLOT data=meaned;
   title 'Interaction Plot of Treatment, Sex, and CD4';
   plot cd4*sex=trt;
RUN;

proc sgplot data=meaned;
  yaxis label="Mean CD4 levels";
  vbar sex / group=trt
                  response=cd4
             barwidth=0.5
             transparency=0.5;
run;




/* /Plotting the interaction of age and sex/ */

PROC SORT data=assignment_long out=back;
  by sex bin_age;
RUN;

PROC MEANS data=back noprint;
   by sex bin_age;
   var cd4_diff;
   output out=meaned mean=cd4;
RUN;

PROC PRINT;
   title 'CD4 for Sex and Age Group combinations';
RUN;
```

```
PROC PLOT data=meaned;
   title 'Interaction Plot of Sex, Age Group, and CD4';
   plot cd4*bin_age=sex;
RUN;

proc sgplot data=meaned;
  yaxis label="CD4 levels";
  vbar bin_age / group=sex
                 response=cd4
            barwidth=0.5
            transparency=0.5;
run;
```


ANOVA Model with Type3 Estimators

The Mixed Procedure

Model Information	
Data Set	WORK.ASSIGNMENT_LONG
Dependent Variable	cd4_diff
Covariance Structure	Variance Components
Estimation Method	Type 3
Residual Variance Method	Factor
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Satterthwaite

Class Level Information		
Class	Levels	Values
trt	4	1 2 3 4
sex	2	0 1
BIN_age	5	1 2 3 4 5
time	2	CD4_diff08 CD4_diff16

Dimensions	
Covariance Parameters	2
Columns in X	73
Columns in Z	1
Subjects	1
Max Obs per Subject	364

Number of Observations	
Number of Observations Read	364
Number of Observations Used	364
Number of Observations Not Used	0

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
trt	3	27319	9106.411491	Var(Residual) + Q(trt,trt*time,trt*BIN_age,trt*sex)	MS(Residual)	327	5.82	0.0007
sex	1	9818.118495	9818.118495	Var(Residual) + Q(sex,trt*sex,sex*time,sex*BIN_age)	MS(Residual)	327	6.27	0.0127
BIN_age	4	25341	6335.150785	Var(Residual) + Q(BIN_age,trt*BIN_age,BIN_age*time,sex*BIN_age)	MS(Residual)	327	4.05	0.0032
time	1	2827.518155	2827.518155	Var(Residual) + Q(time,trt*time,BIN_age*time,sex*time)	MS(Residual)	327	1.81	0.1799

Results: Stastitical_Modeling.sas

Type 3 Analysis of Variance								
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
trt*time	3	297.839897	99.279966	Var(Residual) + Q(trt*time)	MS(Residual)	327	0.06	0.9791
trt*BIN_age	11	34351	3122.845305	Var(Residual) + Q(trt*BIN_age)	MS(Residual)	327	2.00	0.0283
BIN_age*time	4	5244.288438	1311.072110	Var(Residual) + Q(BIN_age*time)	MS(Residual)	327	0.84	0.5020
trt*sex	3	20644	6881.167234	Var(Residual) + Q(trt*sex)	MS(Residual)	327	4.40	0.0047
sex*time	1	354.138969	354.138969	Var(Residual) + Q(sex*time)	MS(Residual)	327	0.23	0.6346
sex*BIN_age	4	22239	5559.839386	Var(Residual) + Q(sex*BIN_age)	MS(Residual)	327	3.55	0.0075
id	1	4373.390427	4373.390427	Var(Residual) + 4.3E7 Var(id)	MS(Residual)	327	2.79	0.0956
Residual	327	511798	1565.130100	Var(Residual)	.	.	.	.

Covariance Parameter Estimates				
Cov Parm	Estimate	Standard Error	Z Value	Pr Z
id	0.000065	0.000144	0.45	0.6499
Residual	1565.13	122.40	12.79	<.0001

Fit Statistics	
-2 Res Log Likelihood	3433.3
AIC (Smaller is Better)	3437.3
AICC (Smaller is Better)	3437.3
BIC (Smaller is Better)	3433.3

Solution for Fixed Effects												
Effect	NAME OF FORMER VARIABLE	trt	sex	BIN_age	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept					13.1744	10.2551	192	1.28	0.2005	0.05	-7.0530	33.4017
trt		1			-23.1065	15.9137	328	-1.45	0.1475	0.05	-54.4124	8.1994
trt		2			-32.8772	14.3156	328	-2.30	0.0223	0.05	-61.0393	-4.7152
trt		3			-16.9502	14.3716	328	-1.18	0.2391	0.05	-45.2223	11.3220
trt		4			0	.	.	.	.	.	.	.
sex			0		57.7859	30.1180	328	1.92	0.0559	0.05	-1.4630	117.03
sex			1		0	.	.	.	.	.	.	.
BIN_age				1	24.7347	13.5001	327	1.83	0.0678	0.05	-1.8232	51.2926
BIN_age				2	-19.8716	14.4204	328	-1.38	0.1691	0.05	-48.2399	8.4966
BIN_age				3	-20.6132	15.1399	328	-1.36	0.1743	0.05	-50.3968	9.1704
BIN_age				4	7.5153	14.5287	328	0.52	0.6053	0.05	-21.0658	36.0965
BIN_age				5	0	.	.	.	.	.	.	.

Results: Stastitical_Modeling.sas

Solution for Fixed Effects

Effect	NAME OF FORMER VARIABLE	trt	sex	BIN_age	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
time	CD4_diff08				6.3793	11.9351	327	0.53	0.5934	0.05	-17.1000	29.8586
time	CD4_diff16				0	.	.	.	.	.	.	.
trt*time	CD4_diff08	1			4.7522	12.0085	327	0.40	0.6926	0.05	-18.8715	28.3759
trt*time	CD4_diff16	1			0	.	.	.	.	.	.	.
trt*time	CD4_diff08	2			0.8672	11.3919	327	0.08	0.9394	0.05	-21.5435	23.2779
trt*time	CD4_diff16	2			0	.	.	.	.	.	.	.
trt*time	CD4_diff08	3			0.3103	12.1617	327	0.03	0.9797	0.05	-23.6148	24.2355
trt*time	CD4_diff16	3			0	.	.	.	.	.	.	.
trt*time	CD4_diff08	4			0	.	.	.	.	.	.	.
trt*time	CD4_diff16	4			0	.	.	.	.	.	.	.
trt*BIN_age		1		1	-17.6380	19.7531	328	-0.89	0.3726	0.05	-56.4969	21.2209
trt*BIN_age		1		2	7.2693	20.6653	328	0.35	0.7252	0.05	-33.3840	47.9226
trt*BIN_age		1		3	18.6038	22.0464	328	0.84	0.3994	0.05	-24.7667	61.9743
trt*BIN_age		1		4	2.2114	19.2481	327	0.11	0.9086	0.05	-35.6544	40.0771
trt*BIN_age		1		5	0	.	.	.	.	.	.	.
trt*BIN_age		2		1	-21.5972	18.3514	328	-1.18	0.2401	0.05	-57.6986	14.5043
trt*BIN_age		2		2	27.2050	18.5669	327	1.47	0.1438	0.05	-9.3204	63.7305
trt*BIN_age		2		3	36.7564	18.6986	328	1.97	0.0502	0.05	-0.02784	73.5407
trt*BIN_age		2		4	5.2727	19.0266	325	0.28	0.7819	0.05	-32.1581	42.7035
trt*BIN_age		2		5	0	.	.	.	.	.	.	.
trt*BIN_age		3		1	3.5889	18.7052	327	0.19	0.8480	0.05	-33.2089	40.3866
trt*BIN_age		3		2	48.8931	18.1278	327	2.70	0.0074	0.05	13.2313	84.5549
trt*BIN_age		3		3	26.1564	19.7169	327	1.33	0.1856	0.05	-12.6315	64.9442
trt*BIN_age		3		4	0	.	.	.	.	.	.	.
trt*BIN_age		4		1	0	.	.	.	.	.	.	.
trt*BIN_age		4		2	0	.	.	.	.	.	.	.
trt*BIN_age		4		3	0	.	.	.	.	.	.	.
trt*BIN_age		4		4	0	.	.	.	.	.	.	.
trt*BIN_age		4		5	0	.	.	.	.	.	.	.
BIN_age*time	CD4_diff08			1	-10.8224	14.5341	327	-0.74	0.4570	0.05	-39.4145	17.7697
BIN_age*time	CD4_diff16			1	0	.	.	.	.	.	.	.
BIN_age*time	CD4_diff08			2	11.5735	14.5502	327	0.80	0.4269	0.05	-17.0503	40.1973
BIN_age*time	CD4_diff16			2	0	.	.	.	.	.	.	.
BIN_age*time	CD4_diff08			3	-4.7611	15.0601	327	-0.32	0.7521	0.05	-34.3879	24.8657
BIN_age*time	CD4_diff16			3	0	.	.	.	.	.	.	.

om/SASStudio/sasexec/submissions/dc3105c2-204a-4c1b-8007-db919ab9fa06/results

Results: Stastitical_Modeling.sas

Solution for Fixed Effects

Effect	NAME OF FORMER VARIABLE	trt	sex	BIN_age	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
BIN_age*time	CD4_diff08			4	-1.6539	14.4562	327	-0.11	0.9090	0.05	-30.0927	26.7849
BIN_age*time	CD4_diff16			4	0	.	.	.	.	.	.	.
BIN_age*time	CD4_diff08			5	0	.	.	.	.	.	.	.
BIN_age*time	CD4_diff16			5	0	.	.	.	.	.	.	.
trt*sex		1	0		13.5094	25.6379	327	0.53	0.5986	0.05	-36.9263	63.9451
trt*sex		1	1		0	.	.	.	.	.	.	.
trt*sex		2	0		94.5787	31.6662	327	2.99	0.0030	0.05	32.2836	156.87
trt*sex		2	1		0	.	.	.	.	.	.	.
trt*sex		3	0		110.74	31.7893	328	3.48	0.0006	0.05	48.2057	173.28
trt*sex		3	1		0	.	.	.	.	.	.	.
trt*sex		4	0		0	.	.	.	.	.	.	.
trt*sex		4	1		0	.	.	.	.	.	.	.

sex*time	CD4_diff08		0		7.3063	15.3598	327	0.48	0.6346	0.05	-22.9101	37.5227
sex*time	CD4_diff16		0		0	.	.	.	.	.	.	.
sex*time	CD4_diff08		1		0	.	.	.	.	.	.	.
sex*time	CD4_diff16		1		0	.	.	.	.	.	.	.
sex*BIN_age			0	1	-100.55	35.0090	327	-2.87	0.0043	0.05	-169.42	-31.6743
sex*BIN_age			0	2	-53.6244	33.9506	327	-1.58	0.1152	0.05	-120.41	13.1647
sex*BIN_age			0	3	-157.50	44.4553	327	-3.54	0.0005	0.05	-244.95	-70.0420
sex*BIN_age			0	4	-115.18	48.4445	328	-2.38	0.0180	0.05	-210.48	-19.8816
sex*BIN_age			0	5	0	.	.	.	.	.	.	.
sex*BIN_age			1	1	0	.	.	.	.	.	.	.
sex*BIN_age			1	2	0	.	.	.	.	.	.	.
sex*BIN_age			1	3	0	.	.	.	.	.	.	.
sex*BIN_age			1	4	0	.	.	.	.	.	.	.
sex*BIN_age			1	5	0	.	.	.	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
trt	3	327	5.96	0.0006
sex	1	315	6.97	0.0087
BIN_age	4	325	4.21	0.0025
time	1	327	1.81	0.1799
trt*time	3	327	0.06	0.9791
trt*BIN_age	11	325	2.01	0.0271

a-4c1b-8007-db919ab9fa06/results

Results: Stastitical_Modeling.sas

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
BIN_age*time	4	327	0.84	0.5020
trt*sex	3	328	4.48	0.0042
sex*time	1	327	0.23	0.6346
sex*BIN_age	4	324	3.66	0.0062