The relationship between having an HIV partner and a patient's HIV severity

Statistical report prepared by

Fan Yang, Jia Liu, Josh Wiltsie, Monica Patrin,
Rodrigo Lovaton, Seongkwon Lyu

Instructor: Aaron Rendahl

Class: Statistics 8801

April 29th, 2013

Executive summary

In this report, we examine if there is a relationship between a partner's HIV status and three different blood measurements for patients with HIV. The data used for this study includes patient's gender, patient's partner status (HIV or not), and three blood measurements (CD4, CD8, and RNA). The measurements were obtained from 278 participants. In order to apply statistical procedures and for easier visualization, we perform a statistical transformation to the values of the blood measurements. The hypothesis of a relationship is tested both using Multivariate Analysis of Variance (MANOVA) and individual linear regression models for each of the blood measurements. Results indicate that having a partner without HIV has a positive relationship with CD8 (i.e. healthier response) and the partner status is not statistically significant for CD4. In addition, we observe that males having a partner with HIV have higher values for RNA (i.e. less healthy response), but no significant difference in RNA values for females.

Keywords: AIDS, Partner, Blood Measurements, Health Response, RNA, CD4, CD8, Statistical Transformation, MANOVA, Linear Regression.

I. Introduction

The objective of this report is to determine whether a patient's partner's HIV status is associated with the severity of the patient's HIV virus infection, as it is diagnosed through three different blood measurements. The data was collected in a single location over a one month period, it corresponds to information from 278 HIV infected patients (and their partners), and it was provided by the client for statistical analysis. There is no previous evidence in the literature for the research question proposed in this study. Possibly, two other unmeasured factors could affect the blood measurements: type of and compliance with treatments for the disease and the length of time since diagnosis. However, given the procedures used for the selection of patients and blood testing, it is not expected that they will affect the outcomes analyzed in this study.

The rest of the report is organized as follows. The data used is described in detail in section II, the methods to explore the research question in section III, and the main results of the statistical analysis in section IV. Finally, the discussion of results and conclusions of the study are included in section V. Ancillary tables are shown as an appendix of this report.

II. Data

The data used in this study corresponds to blood measurements performed on 278 HIV infected patients, the HIV status of their partners, as well as the patient's gender. Participants were self-selected into the sample: patients from the clinic were asked to participate voluntarily in the study and others knew about the study through advertisements placed around the city. All participants were given a small economic incentive for agreeing to be part of the study. The sample is expected to be similar to the overall HIV infected population of interest for the clinic. The data complies with the privacy regulations and has been anonymized.

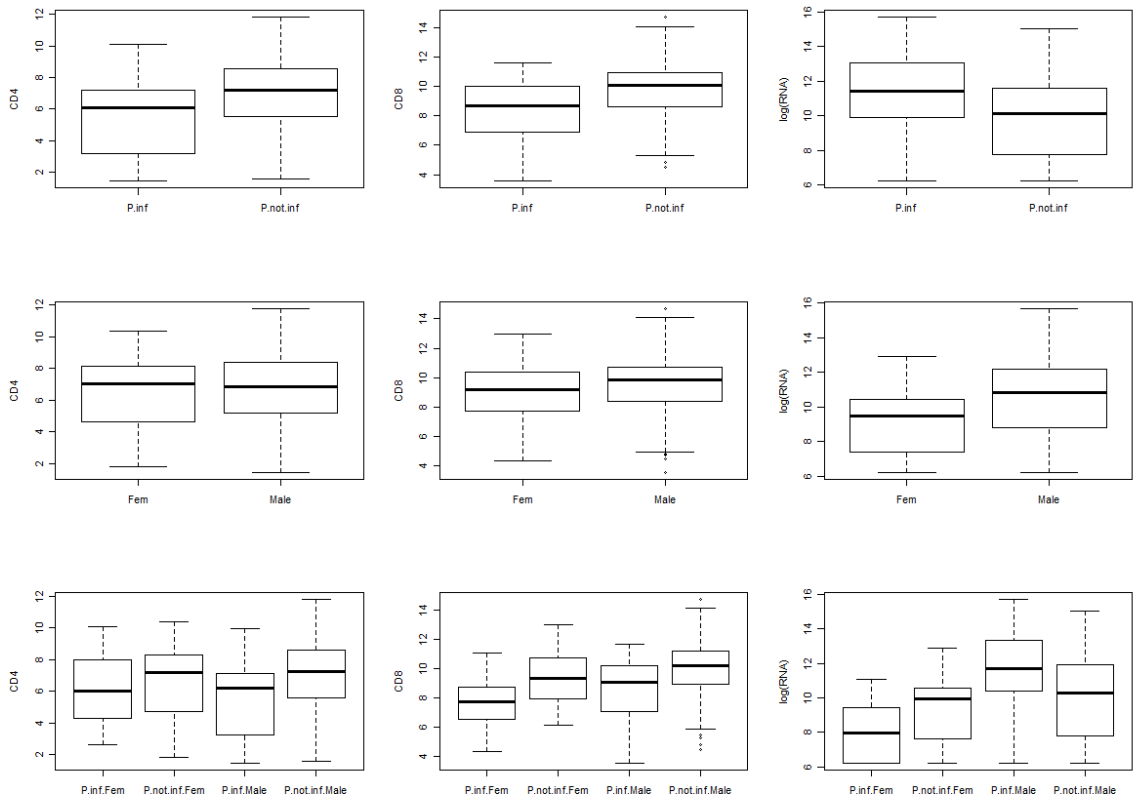
Three blood measurements were applied using standard procedures:

- CD4 or white blood cell count, which measures the severity of sickness, given that the virus uses these cells as host.
- CD8 cell count, which measures the patient's ability to fight infections.
- RNA test, which measures how much virus is in the patient's blood.

Lower counts of CD4 and CD8, and higher counts of RNA imply higher infection of the virus in the patient. Also, we were given a "type" variable which was categorized in CP (i.e. both infected) and DP (i.e. partner is not infected) and a variable of "sex".

We first explore the values of the outcomes through box plots. We transformed RNA to log(RNA) to see the differentiation in the low values. In Graph 1, we present the values for CD4 (first column), CD8 (second column), and log(RNA) (third column) by the partner's HIV status (first row), gender (second row), and both factors (third row). The first row of plots indicate that patients whose partner is not HIV positive tend to have healthier response values (i.e. higher CD4 and CD8 values, and lower RNA values). As seen in the second row, females seem to have healthier response values for RNA (lower). In the bottom right plot, we observe that males whose partners are infected have higher RNA values (i.e. less healthy) than females and than males whose partners do not have HIV.

Graph 1: Box plots for CD4, CD8, and RNA by partner's HIV status and gender



III. Methods

We used SAS to get the contents of the AIDS data, and it is shown below in Figure 1.

Figure1: Content of Data

Variable	N	Mean	Standard Deviation	Minimum	Maximum
ID	278	N/A	N/A	1	278
CD4	275	402.13	325.06	3.00	1776.00
CD8	275	1061.37	616.67	46.00	3466.00
RNA	195	229026.67	659085.49	500.00	6600000.00

After obtaining a general idea of the given data, we plot the kernel density curves which are a fundamental method to check the probability density function of a random variable. The graphs are in Appendix 1. Based on these graphs, RNA data is very skewed, and the multivariate

normality seems violated, which is addressed later.

There is some missing data in each response. The specific missing numbers are listed below in Figure 2.

Figure 2: Missing Data Summary

Variable	N Missing: Female, Male	Total N: Female, Male
ID	0, 0	61, 217
CD4	0, 3	61, 214
CD8	0, 3	61, 214
RNA	25, 58	39, 159

We inquired as to why there is a large portion of missing data for RNA, and we were informed that the missing values did not depend on the true values, so we could treat them as missing at random. Therefore, we decided that after applying the Multivariate Regression and Multivariate Analysis of Variance, we would use a statistical data imputation as a robustness check.

MANOVA--MANOVA was used to determine if there was a relationship between partner's status and at least one of the patient's blood measurements (CD4, CD8, and/or RNA). However, there are several assumptions that need to be fulfilled before this analysis can be performed. MANOVA requires that the response variables have multivariate normality and that they are somewhat correlated. We checked these assumptions graphically by viewing a scatterplot matrix of our blood measures (see appendix).

From the scatterplot matrix, we can see that the distribution of each of our measurements is clearly not normal; a transformation is in order. We used powerTransform() in R in order to find the appropriate transformations required to achieve multivariate normality. This function suggested the following transformations: $CD4^{0.33}$, $CD8^{0.33}$, and log(RNA). These transformed responses are used for the remainder of our analysis. We can see that log(RNA) is bimodal, it appears that the minimum values are censored at a value of 500. This violates the normality assumption, but we are not sure how much this affects our results. The scatterplot matrices also indicate that our variables are relatively correlated, as the confidence ellipses appear to collapse diagonally.

Linear Regression-- Producing three separate linear regression models was the next step

in this analysis. These models tell us exactly which measures are associated with the predictors. For the RNA analysis, Tukey's HSD was used to to further investigate the significant interaction term.

Data imputation -- The data received includes some missing values for the blood measurements (83 cases for RNA and 3 cases each for CD4 and CD8). In order to avoid discarding any case with a missing value that may introduce bias or affect the representativeness of the results, we use statistical imputation: missing data is substituted with a probable value based on other available information. The linear regression results with the imputed data was not different from the one with the original data. The imputed data and results are shown in the appendix.

IV. Results

The results presented in this section consider all three responses under analysis. We apply the techniques described in the methods section.

In Figure 3, we show the results for a multivariate analysis of variance on the three outcomes. All variables are statistically jointly significant, which implies that considering all three outcomes, there is a statistically significant relationship between the partner's HIV status and CD4, CD8, and/or RNA. The assumptions for a valid MANOVA hold and these results are also the same with or without imputation of missing values (as shown in the appendix). However, we also need to test for the effect of these predictors on each of the outcomes separately.

Figure 3: Multivariate analysis of variance (MANOVA)

	Pillai's trace test	Approx. F test	Den. degrees of freedom	P-value
Partner without HIV	0.137	9.826	186	0.000 ***
Male	0.077	5.177	186	0.002 **
Interaction	0.053	3.475	186	0.017 *
Residuals	188			

Statistical significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In Figure 4, we show the results for the linear regression for CD4, CD8, and RNA. From this table, males have increased RNA values (i.e. less healthy individual). Having a partner without HIV has a positive effect on CD8 (i.e. healthier response) and the partner status is not

statistically significant for CD4 or RNA. The interaction between these factors is significant for RNA. Further investigation of the significant interaction term for log(RNA) shows that males that have a partner with HIV have higher values for rna (i.e. less healthy response), which agrees with the qualitative assessment of the box plots in the data section. See "Tukey multiple comparisons of means" in Appendix for interaction term tests. The results are similar if we use data imputation to replace the missing values for RNA.

Figure 4: Linear regression results

Outcome	$CD4^{0.33}$	$CD8^{0.33}$	log(RNA)
Intercept	**6.191** ***	**7.718** ***	**8.066** ***
Std. Er	[0.758]	[0.639]	[0.820]
Partner without HIV	0.327	**1.658** *	1.225
Std. Er	[0.821]	[0.692]	[0.929]
Male	-0.784	0.795	**3.489** ***
Std. Er	[0.817]	[0.689]	[0.881]
Interaction	1.185	-0.089	**-2.929** **
Std. Er	[0.894]	[0.754]	[1.009]
R^2	0.063	0.12	0.14
Observations	271	271	191

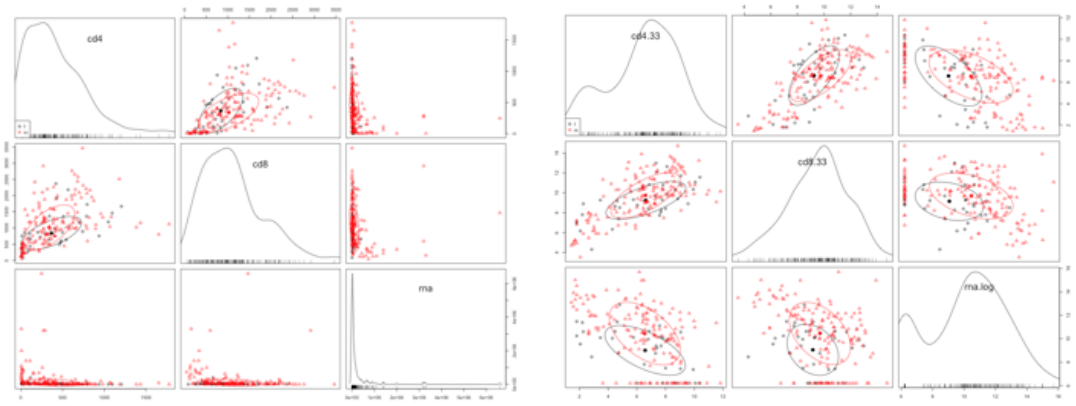
Statistical significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

V. Discussion and conclusions

In this report, we examined whether an AIDS carrier partner's HIV status has an association with a patient's progression of the HIV virus using three different blood measurements. The sample is expected to be similar to the overall HIV infected population of interest. The results from MANOVA indicate that there is a jointly significant relationship between the partner's HIV status and CD4, CD8, and/or RNA. This finding confirms the hypothesis suggested by the client. We further tested for the numeric relationship of the partner's HIV status on each of the measurements separately. Based on the regression analysis, patients who have a partner without HIV have higher CD8 values (i.e. healthier response) and the partner status is not statistically significant for CD4. In addition, there is an interaction effect suggesting that males that have a partner without HIV have lower values for RNA (i.e. healthier response) with respect to males that have a partner with HIV. Given RNA has several missing values, results were tested using data imputation methods and we confirmed that the effects previously described hold.

Appendix

Scatterplot matrices to assess correlation and multivariate normality of the 3 responses (before and after transformation):



> summary(p1<-powerTransform(cbind(cd4,cd8,rna) ~1,AIDS))

bcPower Transformations to Multinormality

	Est.Power	Std.Err.	Wald Lower Bound	Wald Upper Bound
cd4	0.3141	0.0471	0.2218	0.4064
cd8	0.2753	0.0744	0.1294	0.4211
rna	0.0492	0.0274	-0.0046	0.1030

Likelihood ratio tests about transformation parameters

	LRT	df	pval
LR test, lambda = (0 0 0)	63.221509	3	1.204592e-13
LR test, lambda = (1 1 1)	1162.409489	3	0.000000e+00
LR test, lambda = (0.33 0.33 0)	3.695433	3	2.962856e-01

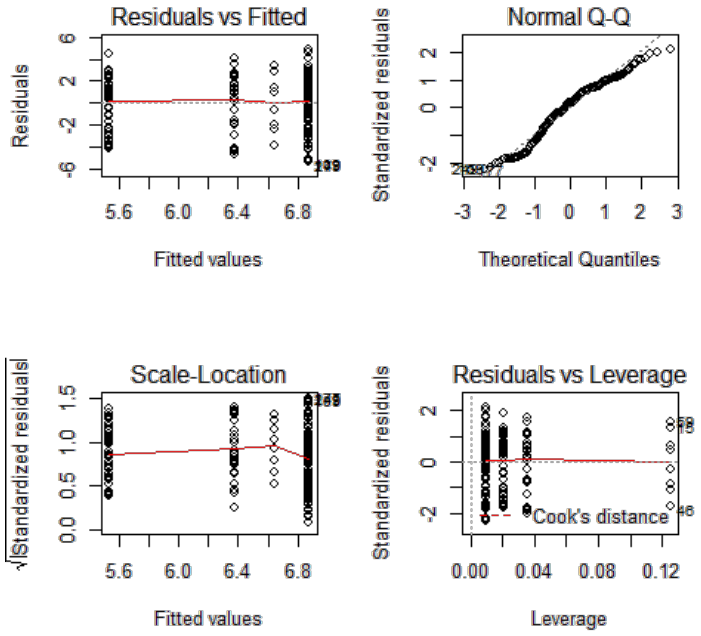
Tukey multiple comparisons of means (95% family-wise confidence level)

Fit: aov.default(formula = log(rna) ~ sex * type)

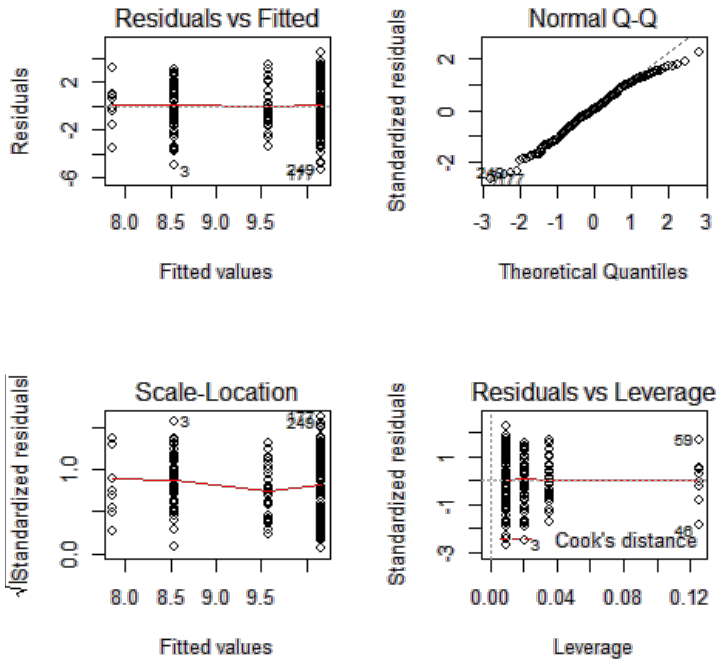
sex:type	diff	lower	upper	p adj
male infected minus female infected	3.49	1.21	5.77	0.0006
female not infected minus female infected	1.23	-1.19	3.64	0.5530
male not infected minus female infected	1.79	-0.42	3.99	0.1569

female not infected minus male infected	-2.26	-3.67	-0.85	0.0003
male not infected minus male infected	-1.7	-2.72	-0.69	0.0001
male not infected minus female not infected	0.56	-0.72	1.84	0.6668

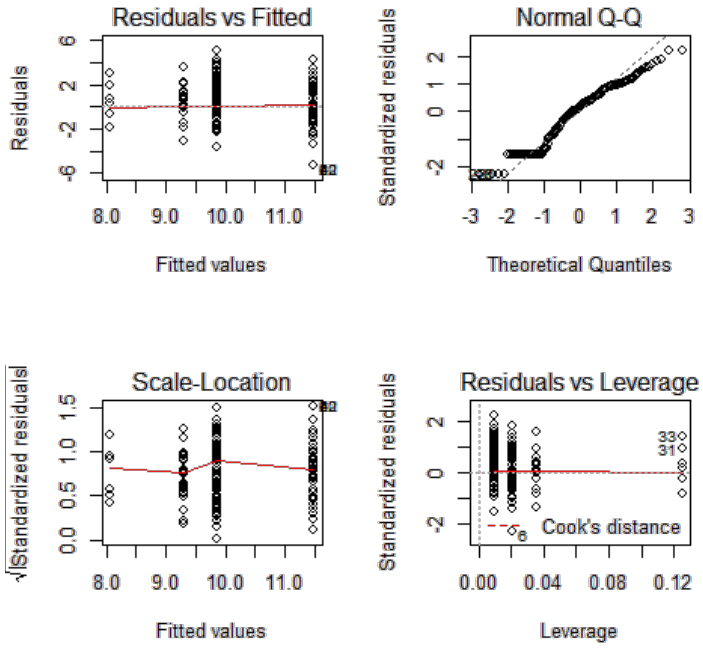
The following graph is for the diagnostics of the linear regression of cd4t~type*sex.



The following graph is for the diagnostics of the linear regression of cd8t~type*sex.

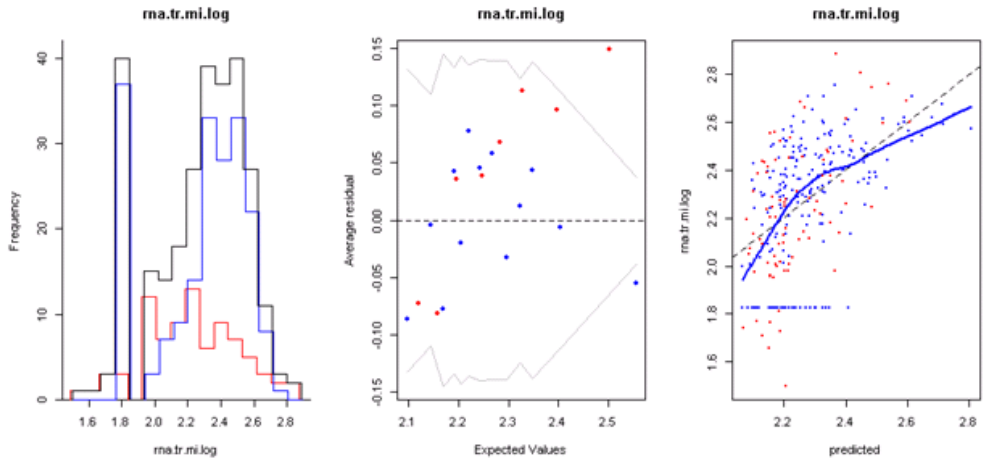


The following graph is for the diagnostics of the linear regression of rnat~type*sex.



These graphs show our imputed data. In the first graph, the blue line is the histogram from the original data

and the red one is from imputed data. The black line is the histogram of new rna data with the imputed data. In the third graph, the blue dots are the original rna values and the red dots are the imputed rna values. The expected values are calculated through the mi packages.



This is another multivariate regression result. We removed the lowest data points to achieve normality and imputed the missing data. Still the imputed data with 'mi' package gives better result.

Outcome	Log(RNA)		Log(RNA.pred)		log(RNA.mi)	
	Estimate	Std. Er	Estimate	Std. Er	Estimate	Std. Er
Intercept	9.177***	[0.718]	9.686***	[0.566]	9.014***	[0.636]
partnerwithoutHIV	0.613	[0.799]	0.744	[0.601]	0.540	[0.675]
Male	2.793***	[0.764]	2.225***	[0.599]	3.038***	[0.670]
Interaction	-1.608 ·	[0.862]	-1.783**	[0.643]	-1.697*	[0.721]
R2	0.14		0.10		0.12	
Observations	191		271		274	

The following is the MANOVA analysis using imputed data. We can see every term is significant.

```
predictrnat<-predict(m1,AIDS[mindex,7:8])

AIDS2<-AIDS[,c(1,2,3,7,8,9)]
```

```
AIDS2[mindex,'rnat']=predictrnat

m2<-manova(cbind(cd4t,cd8t,rnat)~type*sex,data=AIDS2)

> summary(m2)

          Df   Pillai approx F num Df den Df       Pr(>F)

type       1 0.119212  12.1361      3    269  1.79e-07 ***

sex        1 0.062395   5.9671      3    269 0.0005951 ***

type:sex   1 0.046856   4.4079      3    269 0.0047840 **

Residuals 271

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```