Mathematical Modeling of HCV Viral Kinetics

Robert Baraldi, Karissa Cross, Christina McChesney, Laura Poag, Emma Thorpe,
Kevin Flores and H.T. Banks
Center for Research in Scientific Computation
Center for Quantitative Sciences in Biomedicine
North Carolina State University
Raleigh, NC

July 15, 2013

Abstract

Hepatitis C virus (HCV) is a single-stranded RNA virus that infects the hepatocytes of the liver. Infected cells become damaged and release alanine aminotransferase (ALT), which exists normally within liver cells. Our main goal in this project was to determine the usefullness of ALT levels to determine the outcome of HCV treatment and viral load levels. As part of our investigations, we first looked for correlations between ALT levels and viral load in order to examine the hypothesis that ALT levels could have the capability to estimate viral load. Since we did not have quantitative viral load data, we created a mathematical model and simulated data to test these correlations. Next, we created boxplots for each month of our simulated and collected ALT data to observe the differences in their distributions over time. This was also to determine whether or not our simulated data could be useful in drawing conclusions about how ALT levels and viral load are correlated in reality. The next objective was to use principal component analysis to identify the principal components responsible for the most significant portion of variation in the data. Finally, the simulated data was generated for different longitudinal schedules of observation to compare the parameter estimates and standard errors. By doing this, we were able to suggest alternative time points for which to collect data that would allow more efficient estimation of parameters in the future.

1 Introduction

Hepatitis C virus (HCV) is a positive sense, single-stranded RNA virus [4]. Eighty percent of HCV infections in humans are chronic, which often leads to cirrhosis of the liver. Multiple genotypes exist for HCV due to the highly error prone RNA polymerase. This results in a wide range of genetic diversity that permits the virus to escape multiple treatments. Our data involves genotype 1, which is the most common genotype and has lower levels of response to standard treatment than the other genotypes [4]. The treatment that has proved successful for Hepatitis C involves dual therapy with Ribavirin (RBV) and pegylated interferon (Peg-IFN); recent protease inhibitors have been added to increase response rates among individuals infected.

Suggested chemical markers for liver damage produced by an HCV infection include aspartate aminotranserase (AST) and alanine aminotransferase (ALT). AST is found in multiple organs including the liver, heart, skeletal muscles, kidney, and brain, whereas ALT is mostly associated with the liver [13]. AST and ALT are involved in various metabolic pathways. When liver cells are damaged high levels of these markers will enter the blood serum. We investigated a possible relationship between serum ALT levels and viral load, as well as the potential time-dependence of such a relationship.

1.1 Data Received

Important data received from our collaborators (Cassia Mendes Correa, Aluisio Segurado, and colleagues) in Brazil includes viral load before treatment, virus detection or lack thereof during treatment, ALT level before treatment, ALT levels 1-17mo, and type of response.

Viral load before treatment is measured in international units per milliliter (IU/ml), with >850,000 representing anyone with such a high viral load that it exceeds that threshold. Viral load during treatment is qualitative and is represented with a positive (P) or negative (N). The lower limit of detection used was 400 international units. ALT levels are measured in units per liter (U/L), with levels ranging from 13 to 534.

Different responses to therapy include sustained virologic response (SVR), breakthrough (B), no response (NR), and relapse (R). *Sustained virologic response* is the goal of hepatitis C treatment, and means that a patient achieved undetectable levels of virus at the end of treatment and did not have detectable levels 6 months after the treatment had ended. A *breakthrough* means that a patient achieved SVR at the time of treatment but had the virus return during treatment. *No Response* means a patient did not achieve SVR at any time during treatment, and a *relapse* patient is one that achieved SVR at the end of treatment but had virus levels return during the first 6 months after treatment has ended.

2 Model Description

2.1 Biological model

We began our analysis of the data by creating a biological model to represent hepatitis C viral dynamics in the blood stream. In previous research, such as in [2] and [8], scientists have used viral kinetic models that incorporate infected cells (I), target cells (T), and viral load (V). We have included an additional compartment for ALT (A) in our model, since one of our goals was to evaluate the relationship between ALT and viral load. A schematic representation of the HCV infection is given in Figure 1.

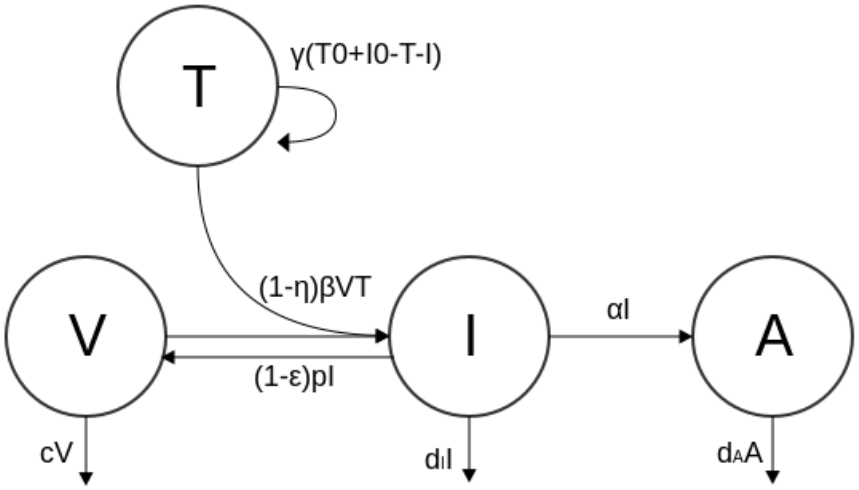

Figure 1: A biological model depicting the basic dynamics of HCV. Parameters are explained in Table 1.

2.2 Mathematical model

To create our first mathematical model, we used the viral kinetic model given in [8] with a few additional components. Our additions account for the relationship between the V, I, and T populations and the serum ALT level, A, and additional $(1-\eta)\beta VT$ terms in the $\frac{dV}{dt}$ and $\frac{dT}{dt}$ equations to account for basic balance laws. This resulted in the following system of equations:

$$\begin{aligned}
\frac{dV}{dt} &= (1-\varepsilon)pI - (1-\eta)\beta VT - cV \\
\frac{dI}{dt} &= (1-\eta)\beta VT - d_I I \\
\frac{dT}{dt} &= \gamma(T_0 + I_0 - T - I) - (1-\eta)\beta VT \\
\frac{dA}{dt} &= \alpha I - d_A A,
\end{aligned} \tag{1}$$

where V, I, T, and A are viral load, infected cells, uninfected cells, and ALT levels, respectively. The parameters are described in Table 1.

Within the mathematical model there are a series of terms that represent various kinetic behaviors within the system. Most of the reactions included in our model are based on the law of mass action, which assumes that the rate at which a population grows or decays is directly dependent on the size of the population at any given time. The $\gamma(T_0 + I_0 - T - I)$ term given in [8] accounts for homeostasis of the liver cells. Cells in the body undergo apoptosis, or programmed cell death, to remove any damaged or redundant cells [3]. Apoptosis is necessary to remove unhealthy cells as well as keep the liver from producing excessive numbers of cells [12]. The term $(1-\eta)\beta VT$ represents the treatment-related infection rate and $(1-\varepsilon)pI$ represents the treatment-related viral production rate [8].

parameter	description of parameter
α	the rate of ALT excretion by infected cells
β	the rate of viral infection of target cells in the absence of treatment
c	the clearance of virions by the immune system
d_A	the rate of ALT degradation in the blood
d_I	the elimination/removal of infected cells by the immune system
ε	the inhibition of virion production/replication by treatment
η	the inhibition of viral infection of target cells by treatment
p	the rate of virion production in the absence of treatment
γ	the rate of target cell regeneration
I_0	the initial value of I (in billions)
T_0	the initial value of T (in billions)

Table 1: A table including all of the parameters used in our model and a brief description of each.

In order to simplify our model, we examined whether or not these additional $(1-\eta)\beta VT$ terms in the $\frac{dV}{dt}$ and $\frac{dT}{dt}$ equations were necessary. Our simplified model was the following:

$$\begin{aligned}
\frac{dV}{dt} &= (1-\varepsilon)pI - cV \\
\frac{dI}{dt} &= (1-\eta)\beta VT - d_I I \\
\frac{dT}{dt} &= \gamma(T_0 + I_0 - T - I) \\
\frac{dA}{dt} &= \alpha I - d_A A.
\end{aligned} \tag{2}$$

To find the difference between these models, we solved the differential equation systems in Matlab using *ode45* with the set of parameters given in Table 2, and compared the solutions of each component by plotting their difference over time. These parameters were estimated based on SVR patient data and therefore represent, to the best of our ability, a realistic biological scenario. The resulting plots are given below in Figures 2-3.

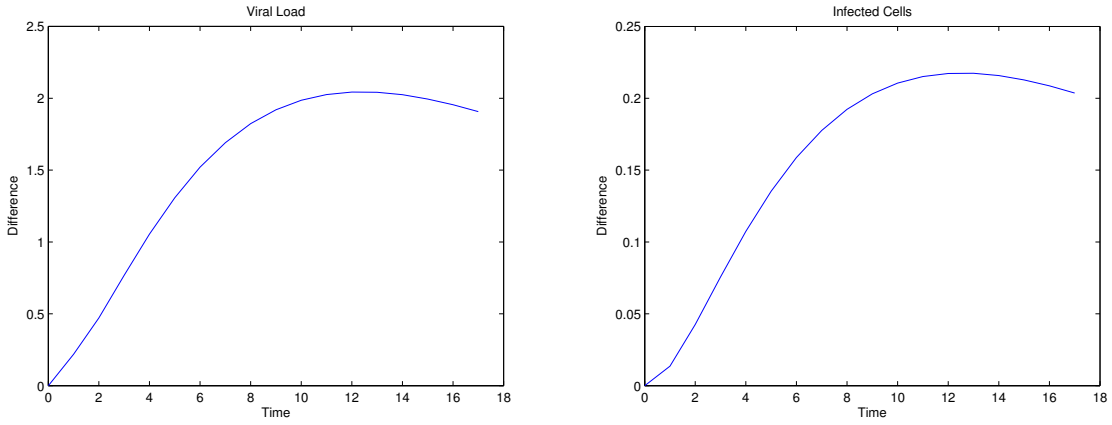

Figure 2: A plot of months 0 through 17 of the difference between the initial model (1) and the simlified model (2) for the viral load in UI/ml (left) and billions of infected cells (right). Our simulations yielded viral load levels between 232 and 895 IU/ml and infected cell values between 25 billion and 100 billion cells, as can be seen in Figure 4.

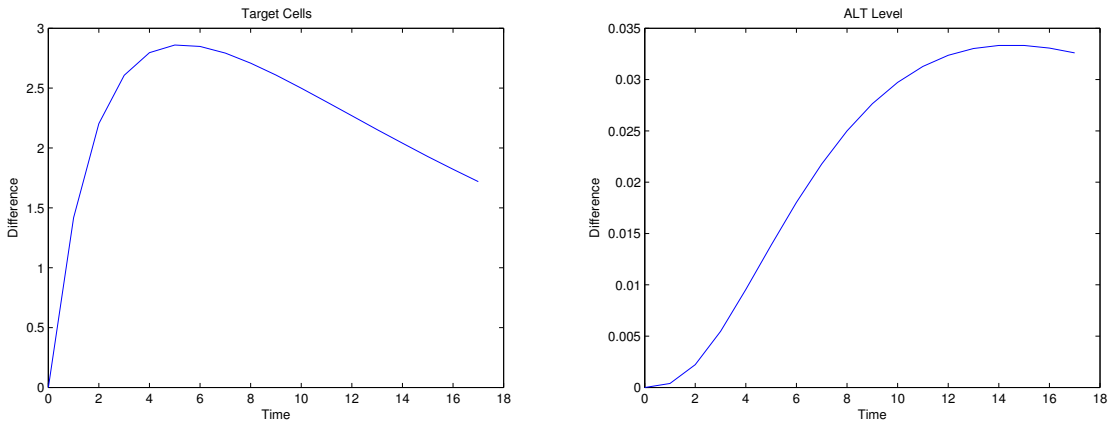

Figure 3: A plot of months 0 through 17 of the difference between the initial model (1) and the simlified model (2) for billions of target cells (left) and ALT levels in U/L (right). Our simulations yielded target cell values between 100 billion and 136 billion cells and ALT levels between 4 and 33 U/L, as can be seen in Figure 4.

As one can see in the figures, the difference between these models is very small, especially as compared to the magnitude of the solutions. To quantify the difference between these models, we calculated the largest percent difference for each variable over all time points. To do this, we used the following equation:

$$\max(\%\text{ difference}) = \max(\frac{Y_i - X_i}{X_i} * 100),\ i = 0, 1, ..., 17, \tag{3}$$

where X is the matrix of solutions to the initial model and Y is the matrix of solutions to the simplified model. The largest percent differences for V, I, T, and A were 0.7694%, 0.7503%, 2.3963%, 0.6793%, respectively. The average of the residuals between the model and patient ALT data is 152.64%, proportionally. Most of the variation occurs towards the end of treatment and this variability easily encompasses the percent differences between the two models. Since the change between the initial model and simplified model is very small for each component, we can justify using the simplified model. The simplified model, given in equation (2), is the only model used in the remainder of this paper.

A simulation of this model using parameters corresponding to an SVR patient can be seen in Figure 4 and an NR patient simulation is given in Figure 5. The parameter values for the SVR and NR simulations are given in Tables 2 and 3, respectively. The process of attaining these values is described in greater detail in Section 3.1. These values were obtained by following the inverse problem methodology outlined in [9]. We ran a Simbiology "parameter fitting" task to get estimates of the parameters in our model. This utilizes an Ordinary Least Squares (OLS) framework with a constant error model.

A main concern with these simulations is that the SVR simulation shows viral load levels approaching a value higher than that of the NR simulation. This is most likely a problem with our chosen parameter values. These parameters were estimated when fitting our model to collected ALT data, but since we have no quantitative viral load data on which to base our estimates, we were less successful in modeling viral load. Because each set of parameters were based on a single individual, discrepancies in results could simply reflect variability between individuals, despite the intention that each person would represent a specific outcome.

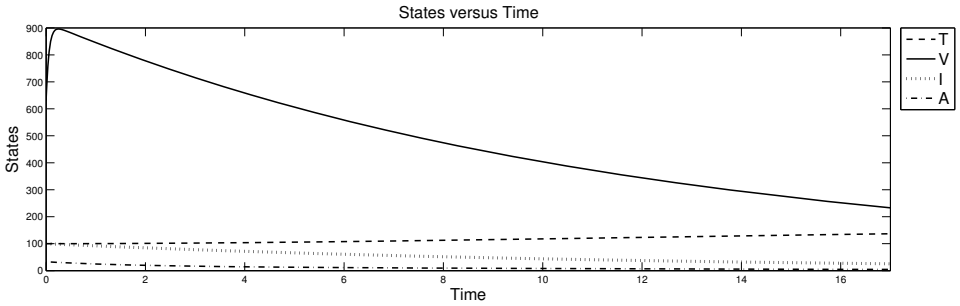

Figure 4: A simulation of our model using SVR parameter values given in Table 2.

Parameter	Value	Standard error
α	0.093533	0.000507
β	0.0000419	0.00000109
c	17	.477832
δ_I	0.102951	0.000449
d_A	0.6005	0.001856
ε	0.85	
γ	0.07	
η	0.5	
p	1037.142	29.328

Table 2: Fixed parameter values and their corresponding standard errors when initially fitting our model to ALT data for an SVR patient. T, I, V, and A were given initial values of 100, 100, 630.091, and 33, respectively. The standard errors were not computed for η, γ, and ε because these were held constant. These parameter values were used in simulating SVR data.

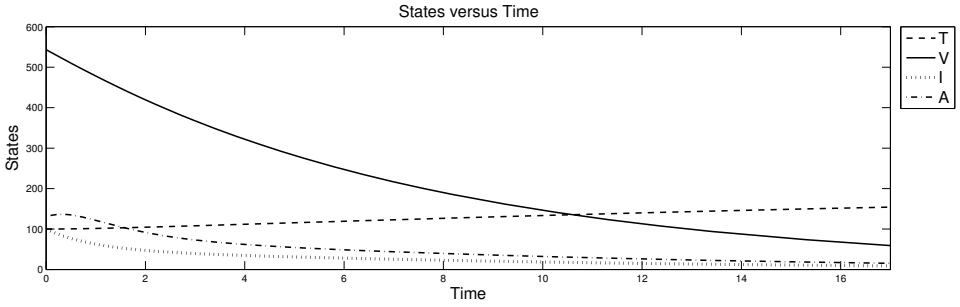

Figure 5: A simulation of our model using NR parameter values given in Table 3.

Parameter	Estimate	Standard error
α	2.555029	0.24405
β	0.00190696	0.00007028
c	0.142626	0.028894
δ_I	1.103802	0.223093
d_A	1.578334	0.204545
ε	0.85	
γ	0.07	
η	0.5	
p	0.636801	0.527088

Table 3: Fixed parameter values and their corresponding standard errors when initially fitting our model to ALT data for an NR patient. T, I, V, and A were given initial values of 100, 100, 543.443, and 130, respectively. The standard errors were not computed for η, γ, and ε because these parameters were held constant. These parameter values were used in simulating NR data.

3 Correlation Between Viral Load and ALT Levels

3.1 Correlation Procedure

One of our objectives was to assess the relationship between viral load and ALT levels throughout treatment. Since viral load was lacking from our data, we simulated data for each of these populations. To do this, we needed reasonable parameter estimates, which we obtained through fitting our model to ALT data for a patient who achieved SVR (patient ID: JMSC) and for a patient who was a NR (patient ID:AFN). We did this in Simbiology using parameter values from [2] and [8] as initial guesses. The parameter estimates we gathered through this process are given in Tables 2 and 3.

It should be noted that at this point, some parameters were fixed. These include ε, η, γ, T_0, and I_0. The parameters ε and η represent the efficacy of the drugs used against the virus and are set at 0.85 and 0.5, respectively. T_0 and I_0 are set at 100, which is estimated to be how many billions of hepatocytes there are, and should remain relatively constant for any given person [15]. The parameter γ, which describes the rate of target cell regeneration is held constant at 0.07. Initial viral load and ALT levels are taken directly from the SVR and NR patient data.

We then generated parameter sets intended to resemble 50 unique patients achieving SVR by using the 95% confidence interval around the estimates in the Table 2 to create a range of parameters that reasonably represent patients achieving SVR. We randomly chose parameter values from these ranges and simulated V and A data for 50 patients. We repeated this process using the non-responder parameter estimates in Table 3. The ALT levels and viral load for these data sets were then graphed and analyzed according to R^2 values in order to determine if there is a relationship between viral load and ALT levels. R values were computed in Matlab with the *corrcoef* function and then squared. The matrix R=corrcoef(X) is given by

$$R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}}, \tag{4}$$

where C is the covariance matrix of the data X. If the R^2 value was greater than 0.50, we accepted there was a correlation. If R^2 is between 0.25 and 0.5, we agreed that there is a slight correlation, and if R^2 is between 0 and 0.25, there is no correlation.

3.2 Correlation Results

This section outlines the results of examining the relationship between ALT and viral load for the SVR and NR patient simulated data.

3.2.1 Results for SVR data

We observe the trends of simulated ALT and viral load over time, and then examine the relationships between ALT and viral load for the SVR simulated data. The resulting plots are given in Figures 6 - 10.

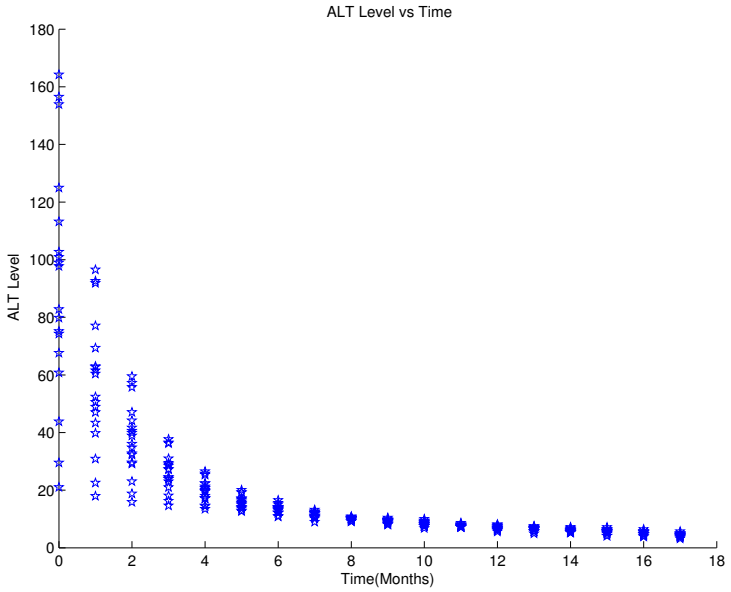

Figure 6: Time vs. ALT: it was observed that ALT decreased over time for 18 simulated SVR patients. This trend is expected as it is shown in literature that a decrease in ALT levels is associated with HCV treatment [5, 11, 16]. Each point represents the ALT level for a single simulated patient for the given month.

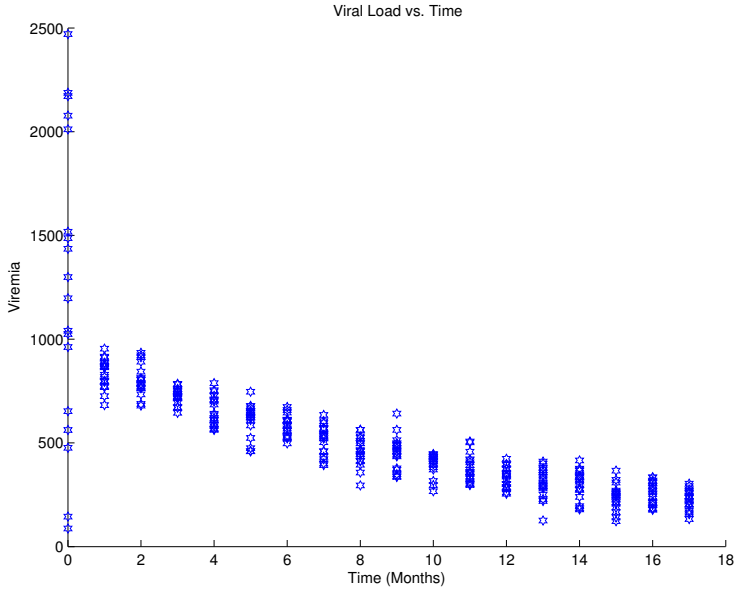

Figure 7: Time vs. Viral Load: it was observed that viral load decreased over time for 18 simulated SVR patients. We expected this trend, since SVR patients have undetectable levels of virus by the end of treatment.

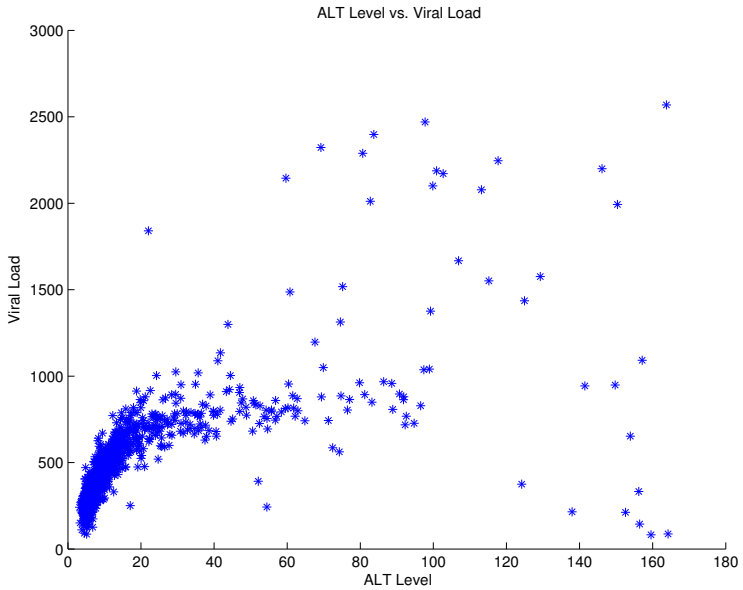

Figure 8: ALT vs. Viral Load: for the 50 simulated SVR patients there appears to be a slight positive relationship between these two factors when the values of both are relatively small. However, this is deteriorated when the values are larger. For comparison, the R^2 value for ALT levels less than 40 and their corresponding viral load values is 0.4821. There is a less significant correlation, with R^2 value of 0.3931, for the entire ALT range. This is possibly due to the negative trend found in both items over time. However, the correlation does not mean that trends in ALT can be used to accurately predict and describe trends in viral load.

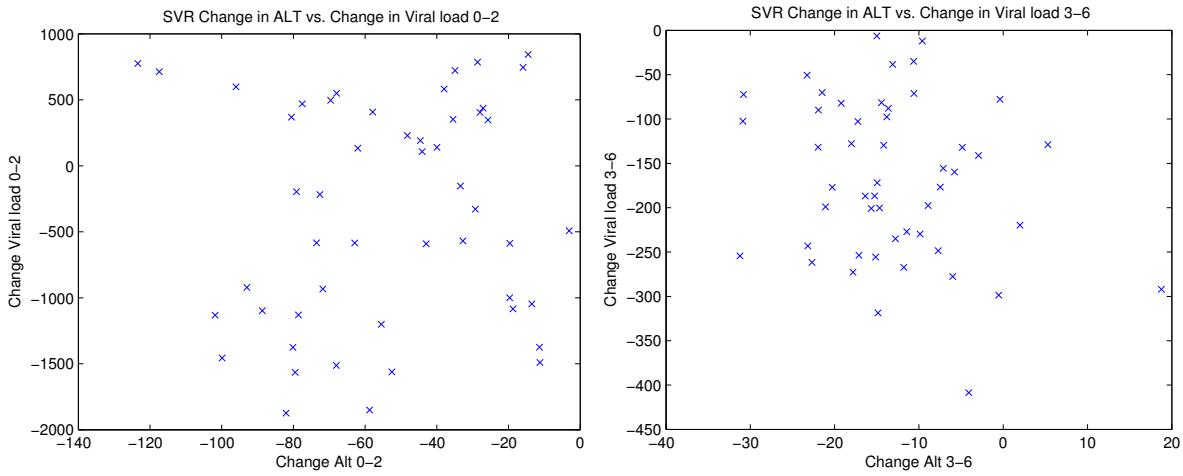

Figure 9: Left: change in ALT vs. change in Viral Load for 50 simulated SVR patients between zero and the first two months: there was no observable correlation (R^2=0.0196, p-value=0.3316). Right: change in ALT vs. change in Viral Load for 50 simulated SVR patients between three and six months: there was no observable correlation (R^2=0.0095, p-value=0.5000).

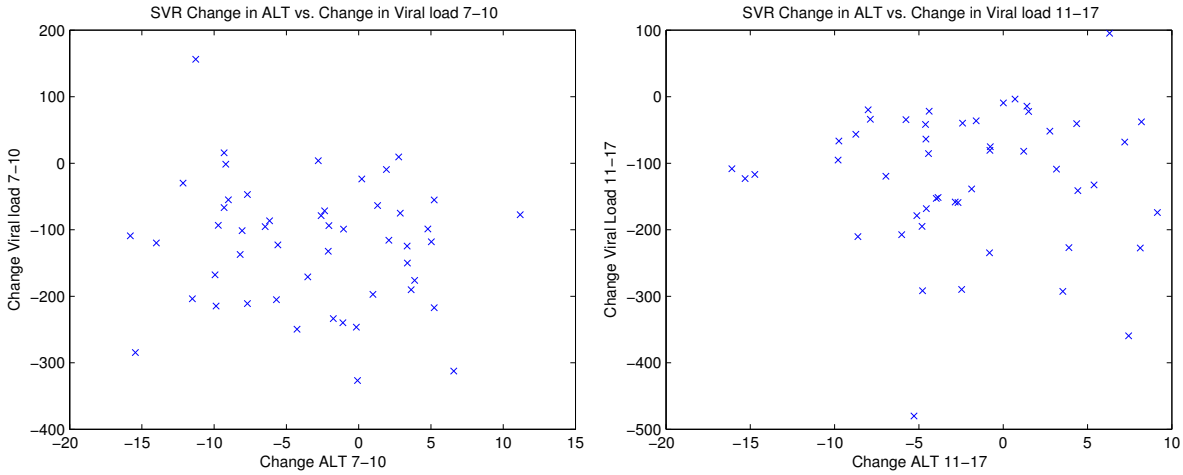

Figure 10: Left: change in ALT vs. change in Viral Load for 50 simulated SVR patients between months seven and ten: there was no observable correlation (R^2=5.1793e-4, p-value=0.8753). Right: change in ALT vs. change in Viral Load for 50 simulated SVR patients between months eleven and seventeen: there was no observable correlation (R^2=2.0657e-4, p-value=0.9211).

According to these R^2 values, there is no correlation between change in viral load and change in ALT levels during any of the time segments chosen. However, there is a slight correlation between viral load and ALT levels over the 17 month time frame. This shows that although there are negative trends over the 17 month time frame for both ALT and viral load, the two factors do not necessarily correlate at all time points.

3.2.2 Results for NR data

The same procedure was repeated for a patient who was a non-responder (NR) and we used the same methods as with the SVR patient to estimate parameters (given in Table 3). The resulting plots are given in Figures 11 - 15.

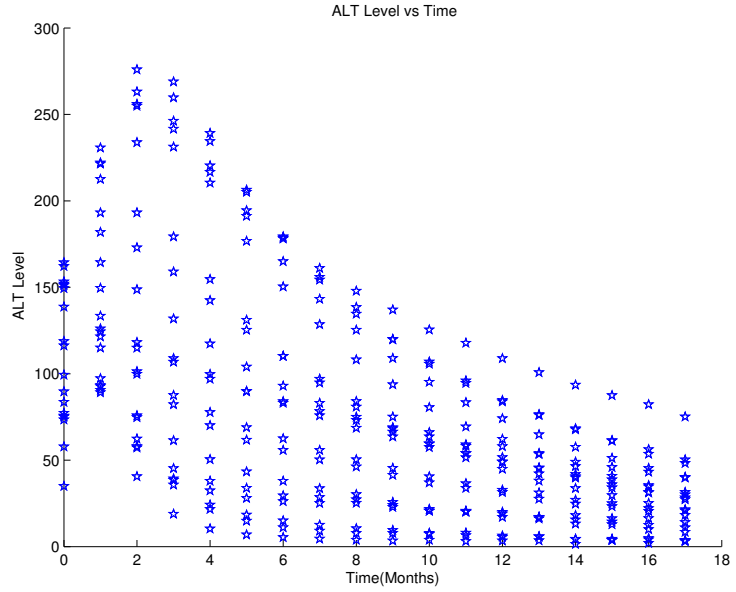

Figure 11: Time vs. ALT: it was observed that ALT decreased over time for 18 simulated NR patients. The range of ALT levels was observed to be larger for NR patients than it was for SVR patients. The increased range could be due to the fact that these simulated patients do not reach SVR and therefore have varying responses to treatment.

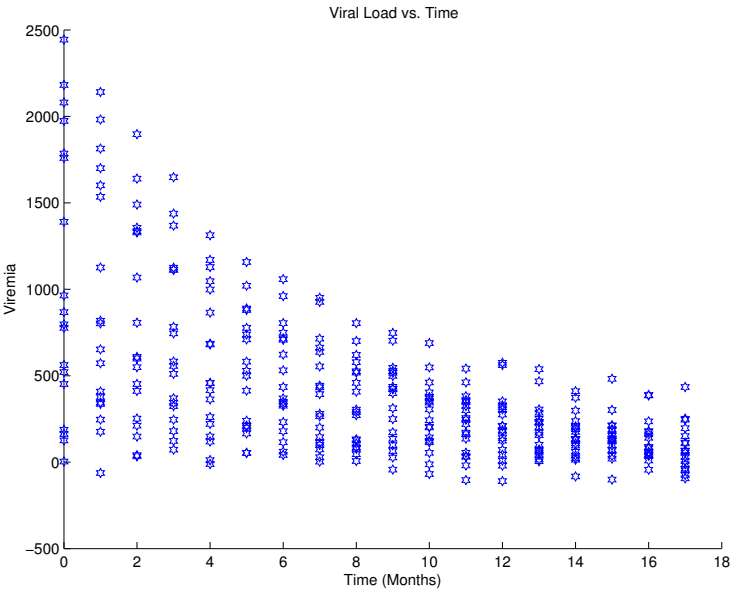

Figure 12: Time vs. Viral Load: it was observed that viral load decreased over time for 18 simulated NR patients. The span of values in this figure had a wider range than for patients who obtained SVR.

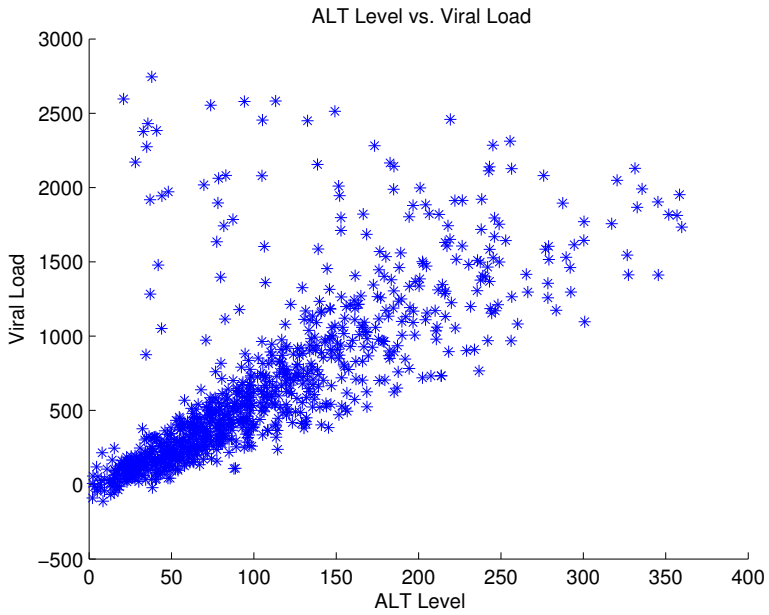

Figure 13: ALT vs. Viral Load: there is a strong correlation between the 50 simulated NR patients ALT and viral load, with an R^2 value of 0.7722. The correlation deteriorates as levels increase, which was also seen in the simulated SVR patients.

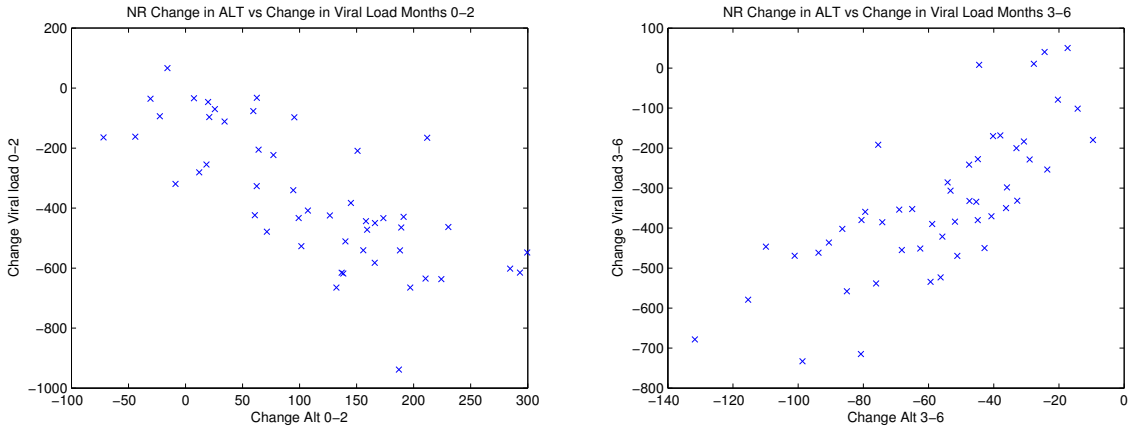

Figure 14: Left: change in ALT vs. change in Viral Load before treatment and in the first two months for the 50 simulated NR patients: there was an observable correlation (R^2=0.5357, p-value=1.5520e-9). Right: change in ALT vs. change in Viral Load for months three through six for the 50 simulated NR patients: there was an observable correlation (R^2=0.5658, p-value=3.0228e-10).

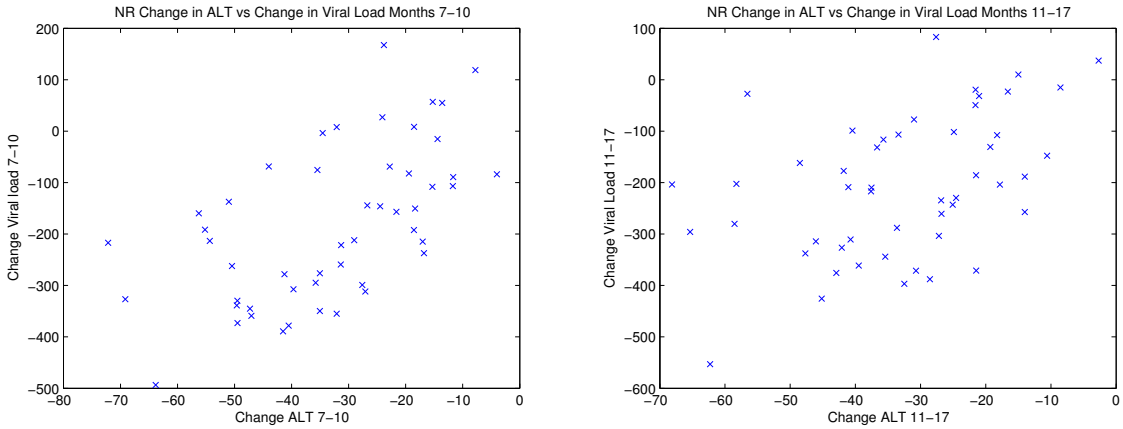

Figure 15: Left: change in ALT vs. change in Viral Load for months seven through ten for the 50 simulated NR patients: there was a slight correlation (R^2=0.3410, p-value=8.5270e-6). Right: change in ALT vs. change in Viral Load for months eleven through seventeen for the 50 simulated NR patients: the R^2 value indicates there is no correlation while the p-value indicates the correlation is significant (R^2=0.1993, p-value=0.0012).

For every time frame examined, the simulated NR data consistently exhibited either a correlation or a slight correlation between ALT levels and viral load, and between the change in ALT levels and change in viral load. This suggests that ALT levels could be a useful indicator of viral load for NR patients, but it is important to note that the majority of these correlations were slight and were found in simulated data based on one NR patient. Also, viral load may not be realistically estimated in these data sets since apart from initial viral load, we did not have quantitative viral load data with which to base our estimates.

3.2.3 Correlation Between Initial Viral Load and Initial ALT Levels

Earlier in this section, we assessed the correlation between viral load and ALT levels, depicted in Figures 8 and 13, for all those values throughout the span of treatment. Here we used the programing language R to calculate the Pearson's product-momentum correlation for viral load and ALT levels specifically before treatment. The alternative hypothesis was defined as: correlation$\neq$ 0. Signficance was established at p-value$<$ 0.05. We did this for all patients, SVR patients, and NR patients. We did not find a significant correlation in any of these three assessments. The results of these tests can be seen in Figure 16.

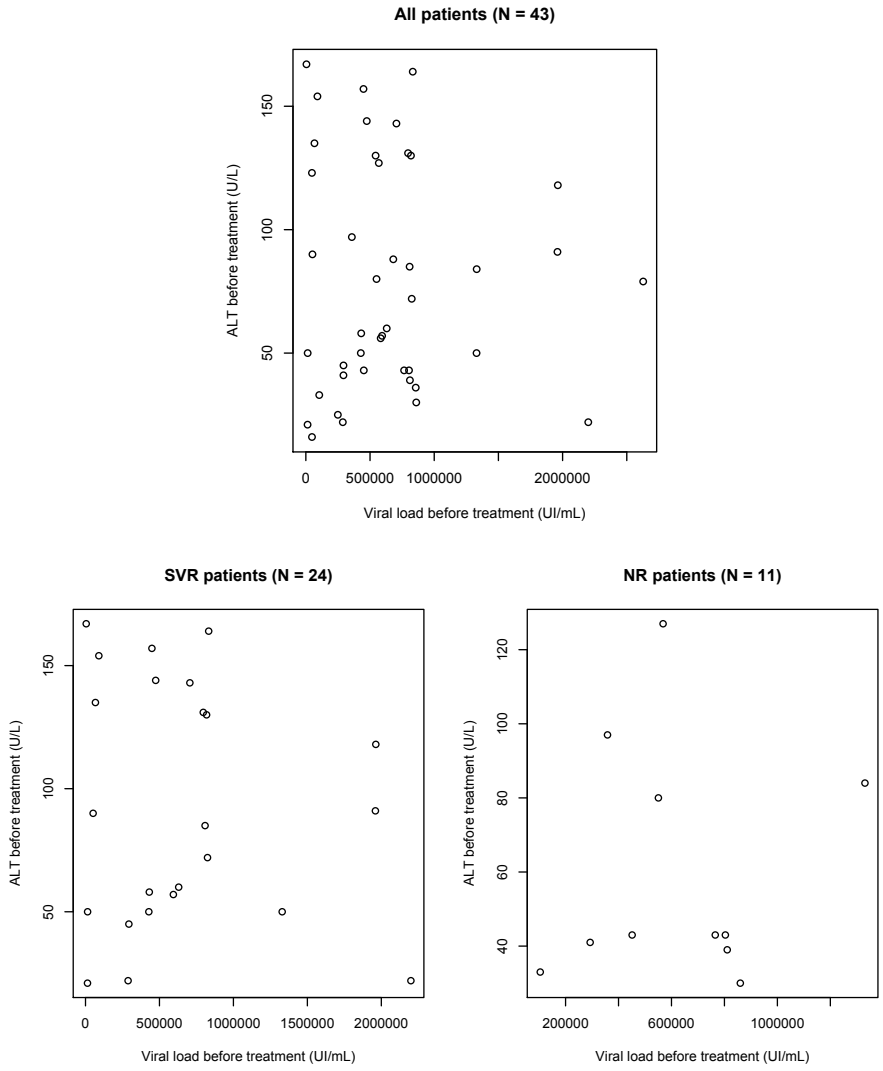

Figure 16: For all patients, SVR patients, and NR patients, no significant correlation was found between these values using a Pearson's product-moment correlation. All patients: p-value=0.8392 and $R^2 = 0.001016266$, SVR patients: p-value=0.6171 and $R^2 = 0.01155767$, NR patients: p-value=0.8115 and $R^2 = 0.006660334$.

3.2.4 Change in ALT Level Over the First Month of Treatment for NR and SVR Patients

We assessed whether there was a significant difference between change in ALT between months 0 and 1 using the Wilcoxson signed-ranked test. Signficance was established at p-value < 0.05. Figure 17 is a boxplot of the changes with outliers included and Figure 18 is the same boxplot with outliers removed. Only the NR patients had outliers removed. The program R was used to remove outliers and does so by computing the interquartile range, i.e., the absolute value of the distance between the 25^{th} and 75^{th} percentile, multiplying that number by 1.5 to get a new number, say X, and then removing any data points outside the interquartile range $\pm X$. The whiskers are the furthest data point outside the interquartile range that is not outside the interquartile range $\pm X$.

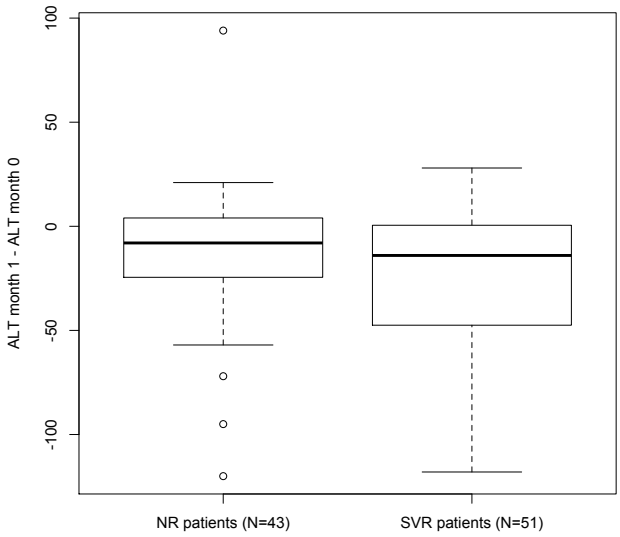

Figure 17: Change in ALT level in the first month of treatment (including outliers): The Wilcoxson signed-rank test on the ALT month 1 - ALT month 0 values for NR and SVR patients produced a test statistic, $W = 1307.5$ and a p-value=0.05503.

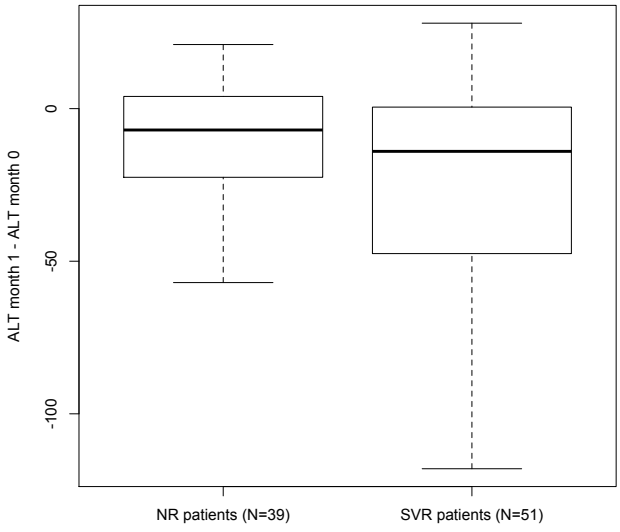

Figure 18: Change in ALT level in the first month of treatment (excluding outliers): Once outliers were removed, a signficant difference was found between the ALT month 1 - ALT month 0 values for NR and SVR patients using the Wilcoxson signed-rank test; test statistic, $W = 1247.5$ and p-value=0.01987.

3.3 Correlation Discussion

At this point we can only postulate as to the accuracy of the model we created when compared with real-world data, because quantitative viral load data is lacking. According to the model and resulting graphs for the simulated data, there is no apparent relationship between change in ALT levels and change in viral load for the simulated SVR data, but there is a correlation between change in ALT levels and change in viral load for the simulated NR data. However correlation does not equal causation. Although both ALT levels and viral load decrease over time with treatment, which leads to a positive correlation between the two, this does not necessarily mean that the two are dependent upon each other. In fact, although ALT and viral load appear to correlate for small values, this correlation deteriorates for larger values of these components, as can be seen by the fan shape in Figures 8 and 13. This may suggest that although trends in ALT may be potential indicators of trends in viral load, it should not be used as the only indicator, because the correlation can deteriorate or even be nonexistent in some cases. For example, some patients have normal ALT levels but elevated viremia. We also found that there was no correlation between initial viral load and initial ALT level for SVR and NR patients. However, we found that the change in ALT levels from month 0 to month 1 of SVR patients is significantly different from that of NR patients. This suggests change in ALT level immediately following initiation of treatment may give insight to end response.

It should also be noted that gathering ALT data in addition to viral load can still be useful in conjunction with other data, since a decrease in ALT can be used to indicate that viral load may have also changed. Changes in ALT levels, or the levels themselves, are not necessarily indicative of SVR, whereas levels of viral load determine the end result of SVR or NR. In conclusion, although it may be useful to collect ALT data, this data is best used when quantitative and time-dynamic viral load data is also collected.

4 Box Plots

4.1 Box Plot Procedure

To create a visual comparison between ALT data that we simulated and real patient data, we created box plots of each month of data with the corresponding month of simulated data. Box plots organize data into quartiles, with the median of the data represented by a line through the middle of the box and the 25th and 75th percentiles represented by the bottom and top of the box respectively. The box encompasses the middle 50% of the data and has two "whiskers" that extend above and below the box that encompass the bottom and top 25% (which accounts for the remaining 50%) [7]. Each month is represented by a different figure to show the change in the relationship between the simulated data and the experimental data over time. Crosses represent outliers of the data. Matlab determines outliers by taking the subset of data that will be included in the plot and creating a box and whiskers that cover 99.3% of the data using a default whisker length. Any points lying beyond the default whisker length are plotted as outliers. If it was found that an excessive number of points were being plotted as outliers the default whisker length could be changed in order to include more of these points in the box or whiskers if one so desired.

4.2 Box Plot Results

Box plots comparing simulated and reported data for ALT levels at months 0, 5, 10, and 17 for NR and SVR patients are given below in Figures 19-22. As can be seen in the figures, our simulated SVR and NR ALT data do not match the reported data exactly. It is more similar to the real patient data in earlier months, as can be seen in Figure 19, but as time passes the distributions are quite different from each other, especially by months 10 (Figure 21) and 17 (Figure 22). These box plots could suggest that our simulated data is less accurate for later months.

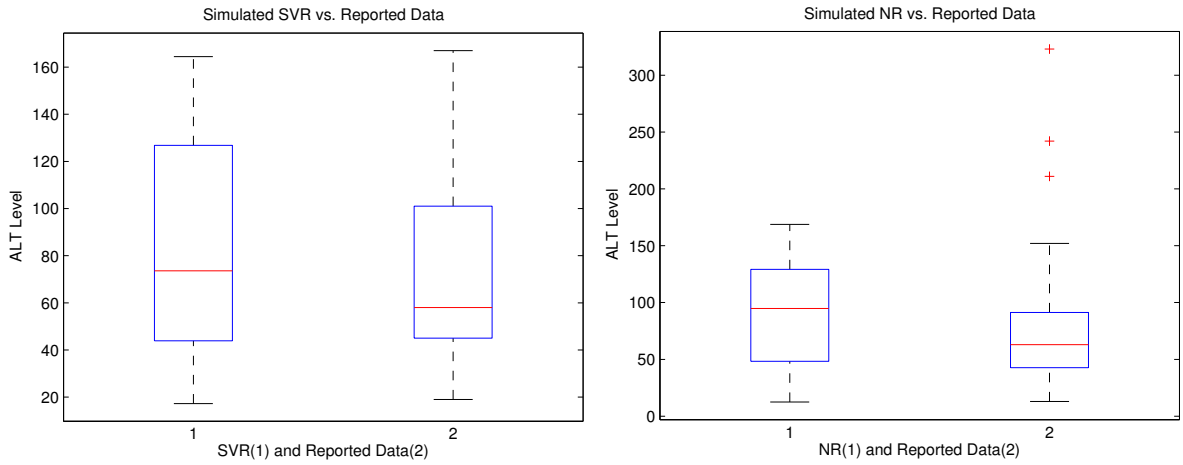

Figure 19: Box plot of the initial ALT levels for simulated and reported SVR (left) and NR (right) data.

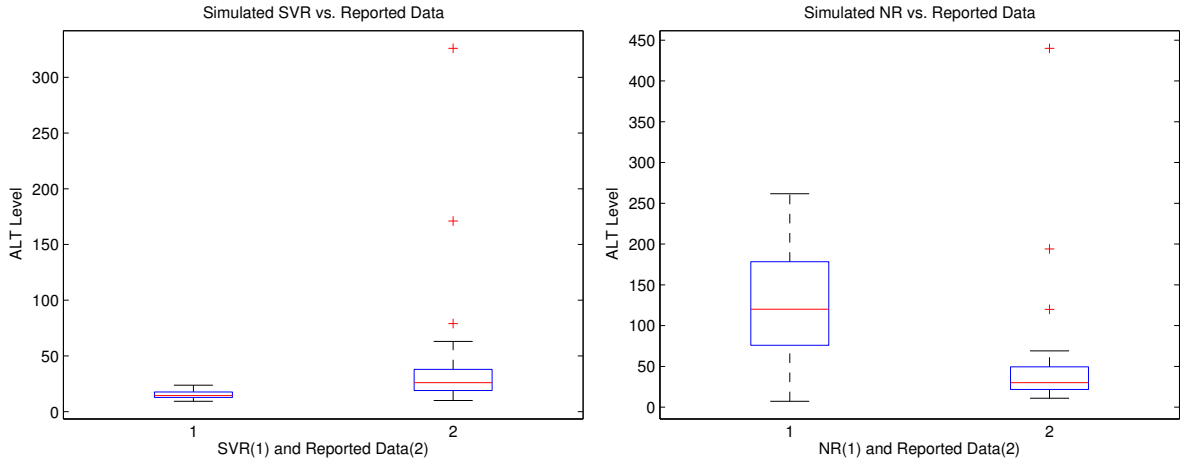

Figure 20: Box plot of the month 5 ALT levels for simulated and reported SVR (left) and NR (right) data.

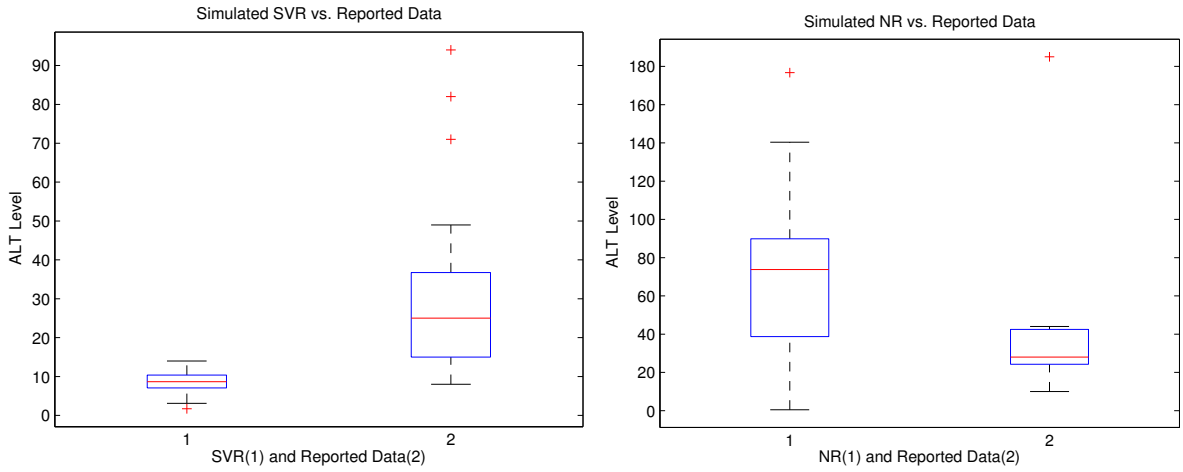

Figure 21: Box plot of the month 10 ALT levels for simulated and reported SVR (left) and NR (right) data.

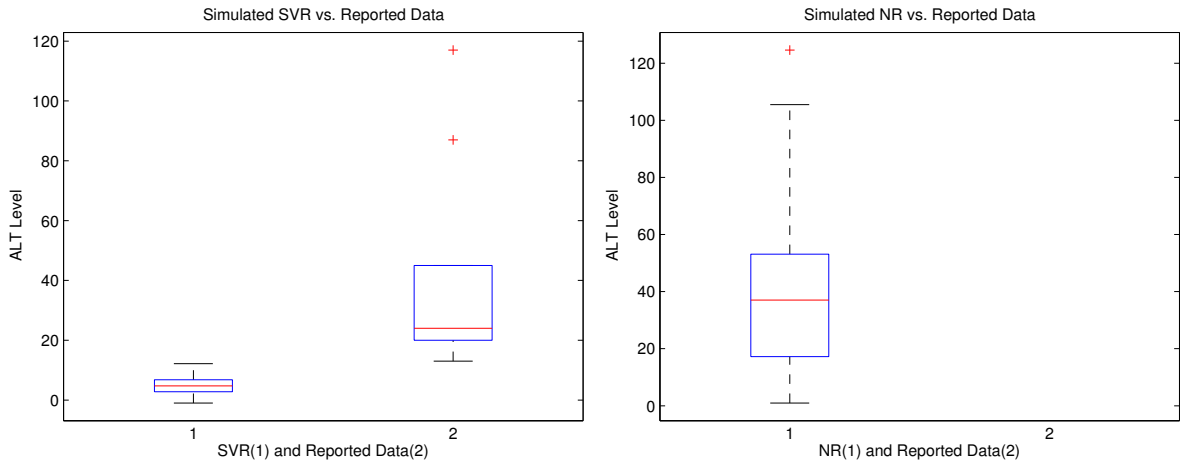

Figure 22: Box plot of the month 17 ALT levels for simulated and reported SVR (left) and NR (right) data. There was no reported data for NR patients in month 17.

4.3 Box Plot Discussion

Our boxplots show that our simulated SVR and NR data closely matches the experimental data for the first few months, but the similarity lessens as time goes on. This could potentially be due to the fact that although the simulated data was created using the experimental data as a baseline, the simulated data uses an ideal model, which could lead it to differ from the real-world experimental data. Real-world situations are rarely ideal, so we should not expect our simulated data to exactly match reported data.

5 Principal Component Analysis

5.1 PCA Procedure

Principal Component Analysis (PCA) is a way to find patterns in data and highlight similarities and differences. It is a tool to transform a set of data components, some of which may be correlated, into a set of uncorrelated,

principal components [10]. When the principal components are plotted, related data will cluster together. An excellent example of this clustering can be seen in a study on the Tyrolean Iceman [1]. For our project, patterns in the data were not immediately evident so we performed PCA to see which variables were responsible for most of the variability in the data. With this test, we would be able to see if variability in ALT levels can explain the end results of NR vs SVR. One main concern is that, according to [6], "any subject who fails to answer just one item will not provide usable data for the principal component analysis, and will therefore be dropped from the final sample." Many of the patients were missing data at some point in the study and therefore we interpolated missing data points, which may distort results.

We began with a data set $X \in \mathbb{R}^{mxn}$, which has n measured variables and m samples (patients). We performed PCA using qualitative viral load data for weeks 4, 12, and 24 as well as ALT data for months 0-6, which formed the n component of our matrix. We filtered the data to only include patients with viral load data for the aforementioned weeks and at least 4 of the first 7 data points for ALT levels. We then used linear interpolation to fill in the missing ALT values where necessary. After making these changes, we used the PCA function in Matlab to calculate the PCA coefficients. The dot products of each column of these coefficients and the data matrix were then computed, producing the corresponding principal components. Principal components 1 and 2 were then plotted against each other to observe clustering of the data. We plotted patients with end results SVR, NR, R, and B with distinct markers in order to see if these types of patients cluster together.

5.2 PCA Results

The PCA plot corresponding to the data is given in Figure 23. This plot shows the PCA results using viral load data for weeks 4, 12, and 24 and the ALT levels for months 0-6. No distinct clusters were observed in the PCA plot.

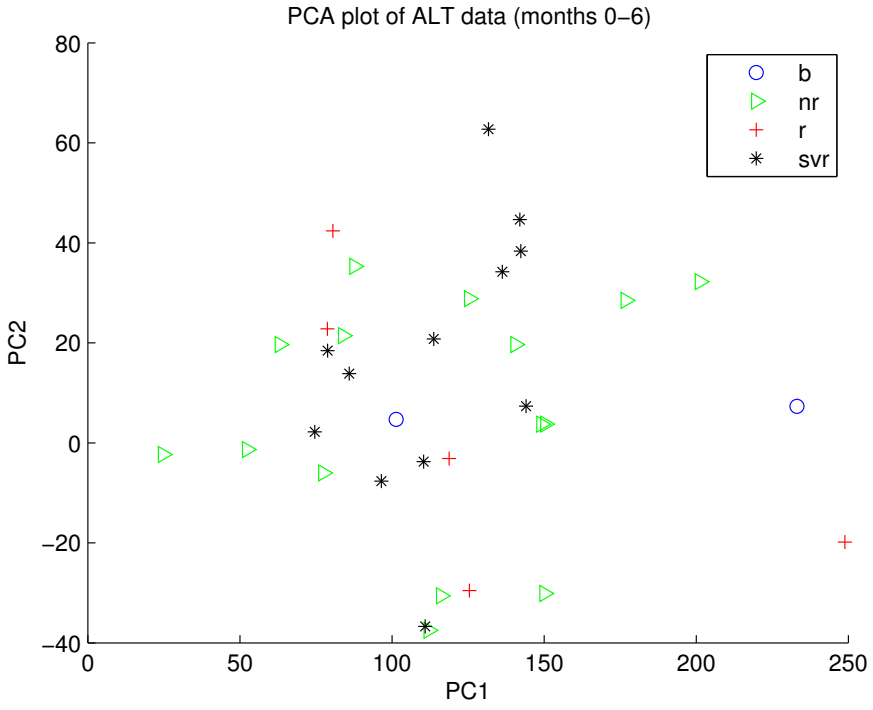

Figure 23: PCA plot for the first six months of ALT data

5.3 PCA Discussion

Principal components 1 and 2 were plotted against each other to observe any potential clustering of the data. Apart from the general lack of clustering, when patients were plotted with distinct markers according to the end results SVR, NR, R, and B, there were no distinct, observable patterns. After finding that there was no clustering, it seems that variation in end response is not explained by ALT levels and qualitative viral load data taken at these times. However, it should be noted that the lack of clustering could have also been due to holes in the ALT data. It should also be noted that the viral load data used was from very early in treatment and isn't as indicative of end response as later points would be. We chose these points because they had the least missing data points. Missing ALT points were estimated using cubic interpolation. However cubic interpolation could lead to simulated data points that were inaccurate. Holes in the data have a large impact on the accuracy of PCA as a whole.

6 Optimal Design

6.1 Optimal Design Procedure

The data we received from our collaborators in Brazil has ALT and AST levels for each month the patient was given treatment. The treatment lasted up to 17 months. This monthly collection of data could overlook possibly important data that occurs at the beginning of treatment when the ALT levels changed drastically and when it is most important for researchers to understand the progression of Hepatitis C in the patient. If it was known at an earlier time that a patient was approaching being a non-responder then maybe earlier data would allow for treatment to be altered so that a response could be attained. We had this in mind when we tried to determine when it would be best to gather viral load and ALT data. Our initial hypothesis was that more frequent time points at the beginning of treatment would better allow us to model the dynamics of an HCV infection. We tested this hypothesis according to SVR and NR simulated data for 17 months, and then for the first 6 months.

To test our hypothesis, we generated data by running a forward simulation in Matlab using the parameters given in Tables 2 and 3 for SVR and NR patients, respectively. Random noise was computed using the *randn* function in Matlab, and was added to each component, with a maximum of 63 for viral load and 0.6 for ALT level. We created four simulated data sets, each differing by its time collection points. We simulated data according to two different time vectors for 17 months, with t=[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17] and t=[0 .25 .5 .75 1 1.5 2 2.5 3 4 5 6 7 9 11 13 15 17], and then simulated two different data sets for the first 6 months of treatment with t=[0 1 2 3 4 5 6] and t=[0 .5 1 1.5 2 4 6]. For each of these estimates, one parameter was estimated at a time, while holding each of the remaining parameters fixed to the values given in Table 2. We then were able to estimate these parameters using each data set and compare the standard errors. In this way, we were able to draw conclusions about the usefulness of collecting data at different time points.

6.2 Optimal Design Results

6.2.1 Simulated SVR data

We first decided to compare simulated SVR data corresponding to two time vectors with 18 time points, starting at 0 and ending at 17. We simulated ALT and viral load data based on a patient with SVR for 17 months with t=[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17] and t=[0 .25 .5 .75 1 1.5 2 2.5 3 4 5 6 7 9 11 13 15 17]. The results for these two longitudinal schedules were inconclusive. The monthly schedule had better estimates of α, β, and d_I, while the schedule with more frequent time points at the beginning of treatment estimated p, d_A, and c better. Because of this even split in the quality of parameter estimates, the better schedule cannot be determined. The resulting parameter estimates and standard errors are given in Table 4 and 5.

Parameter	Estimate	Standard error
α	0.09422622	0.1060171
β	0.00002582	0.00000525
c	17.75256823	0.2795333
d_I	0.11075517	0.00263461
d_A	0.59867825	0.37553467
p	992.8577164	15.71387571

Table 4: Parameter estimates and standard errors computed when fitting the model to simulated data corresponding to t=[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17].

Parameter	Estimate	Standard error
α	0.0940033	0.11879669
β	0.00003225	0.00000793
c	17.5792962	0.25747574
d_I	0.10771144	0.0039108
d_A	0.59941936	0.35473695
p	1002.81003882	14.74933465

Table 5: Parameter estimates and standard errors computed when fitting the model to simulated data corresponding to t=[0 .25 .5 .75 1 1.5 2 2.5 3 4 5 6 7 9 11 13 15 17].

We then simulated SVR data for the first 6 months, using t=[0 1 2 3 4 5 6] and t=[0 .5 1 1.5 2 4 6]. The standard errors were smaller using the second longitudinal schedule, which suggests that more frequent data initially during the first few months is more effective than monthly data during the first six months. These estimates and standard errors are given in Tables 6 and 7.

Parameter	Estimate	Standard error
α	0.09234106	0.22121178
β	0.00002927	0.00002209
c	16.66059375	0.52078317
d_I	0.10854282	0.01031512
d_A	0.60146903	0.65628063
p	1058.34249707	33.20048305

Table 6: Parameter estimates and standard errors computed when fitting the model to simulated data corresponding to t=[0 1 2 3 4 5 6].

Parameter	Estimate	Standard error
α	0.09478808	0.20055546
β	0.00005338	0.00002147
c	17.48572831	0.40716323
d_I	0.09799568	0.01000518
d_A	0.59949907	0.52914983
p	1008.1947708	23.57673919

Table 7: Parameter estimates and standard errors computed when fitting the model to simulated data corresponding to t=[0 .5 1 1.5 2 4 6].

6.2.2 Simulated NR data

Next, we compared simulated NR data corresponding to the same longitudinal schedules used for the SVR simulated data. For the data using 18 time points, the monthly collected data produced smaller parameter standard errors for p, c, and d_I, while the data collected with higher frequency at the beginning produced smaller standard errors for α, β, and d_A. Like the SVR estimates for 18 timepoints, the fact that both schedules have estimate 3 parameters better than the other schedule means that it is difficult to conclude which is better. The resulting parameter estimates and standard errors are given in Table 8 and 9.

Parameter	Estimate	Standard error
α	2.54669835	0.6067866
β	0.00204077	0.0005753
c	0.13478959	0.00830845
d_I	1.04500714	0.24103959
d_A	1.58287035	0.33689903
p	1.08043859	0.4618001

Table 8: Parameter estimates and standard errors computed when fitting the model to simulated data corresponding to t=[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17].

Parameter	Estimate	Standard error
α	2.54861729	0.46105142
β	0.00178383	0.0005573
c	0.15516084	0.01119533
d_I	1.1663598	0.26453455
d_A	1.58177716	0.25858983
p	0.01783054	0.58135127

Table 9: Parameter estimates and standard errors computed when fitting the model to simulated data corresponding to t=[0 .25 .5 .75 1 1.5 2 2.5 3 4 5 6 7 9 11 13 15 17].

Then, we simulated NR data for the first 6 months, using t=[0 1 2 3 4 5 6] and t=[0 .5 1 1.5 2 4 6]. The standard errors of the monthly collected data had smaller standard errors than the data collected with higher frequency in the beginning, which suggests that it is better to collect data monthly during the first 6 months for NR treatments. These estimates and standard errors are given in Tables 10 and 11.

Parameter	Estimate	Standard error
α	2.5534548	0.29953831
β	0.00184177	0.00032544
c	0.14947762	0.00724914
d_I	1.13109689	0.14784317
d_A	1.57921948	0.1635361
p	0.31069286	0.38809015

Table 10: Parameter estimates and standard errors computed when fitting the model to simulated data corresponding to t=[0 1 2 3 4 5 6].

Parameter	Estimate	Standard error
α	2.54362121	0.64598497
β	0.00203373	0.00091533
c	0.11919558	0.01814801
d_I	1.0517985	0.3445589
d_A	1.58517655	0.36072952
p	1.95386695	0.9933885

Table 11: Parameter estimates and standard errors computed when fitting the model to simulated data corresponding to t=[0 .5 1 1.5 2 4 6].

6.3 Optimal Design Discussion

Given the results, we can conclude that collecting data more frequently at the beginning is not necessarily more effective than collecting monthly data for seventeen months. When collecting six months of data, higher frequency at the beginning is better than monthly data for SVR patients, but not for NR patients. It is important to note that this analysis is very surface-level, and we have only compared a few time vectors. More work can be done with more sophisticated optimal design techniques to determine the most effective times when to collect data.

7 Future Data

Additional experimental data that could possibly be obtained in the future is ethnicity of patients and the presence of host genotype IL28B. A polymorphism upstream of the host IL28B gene is the best known baseline predictor of SVR in HCV genotype 1 infected individuals. Individuals who are CC for this gene are much more likely to achieve SVR than those who are genotypically TT. In one study with patients receiving Peg-IFN/RBV treatment, 64% of CC individuals achieved SVR while 23% and 25% of individuals did in CT and TT cohorts, respectively. There is a discrepancy between the frequency of the C allele in African American and European American populations, which accounts for roughly half the discrepancy between the SVR rates of these two populations. While the C allele is less common in the African American population, SVR rates are also less common. [14]

8 Concluding Remarks

Based on our results, we have found that quantitative viral load data would be very useful, or even essential, in gaining further knowledge about HCV viral kinetics. We observed correlations between our simulated data for change in ALT levels and change in viral load for NR patients, but not SVR patients. This suggests that ALT levels may not always be a useful indicator of viral load or end response. Even when there was a correlation found between ALT levels and viral load, this does not mean that higher ALT levels are indicative of high viral load. The only way to effectively predict the end response of a patient is through collecting quantitative viral load data throughout treatment. Our box plots showed that our simulated data is not similar to the collected data after the first 6 months. This suggests that simulating data based on ALT levels from a single patient does not create a data set that accounts for enough variation between a real population of patients. In further research we could alter the noise level used in simulating data, or base our simulated data on a different patient. The PCA plot suggests variation in end response is not explained by ALT levels and qualitative viral load data taken at the points used. This technique can be used more effectively in the future with more useful data, such as time-dynamic quantitative viral load data. We also found, through simulating data with different time vectors, that it is unclear whether it would be more useful to collect data more frequently during the first few months of treatment. Since our results were not consistent between simulated SVR and NR data, it would be best to further examine these simulations with more sophisticated optimal design techniques in the future.

9 Acknowledgments

This research was supported in part by the Undergraduate Biomathematics grant number NSF DBI-1129214 from the National Science Foundation.

References

[1] Andreas Keller et al., New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing, *Nature Communications* **1701**, (2012), 3-6.

[2] Avidan U. Neumann et al., Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-α therapy, *Science* **282** (1998), 103-107.

[3] Blake A. Jones and Gregory J. Gores, Physiology and pathophysiology of apoptosis in epithelial cells of the liver, pancreas, and intestine, *American Journal of Physiology* **273(6)**, (1997), G1174-G1188.

[4] Chang Wook Kim, Kyong-Mi Chang, Hepatitis C virus: virology and life cycle, *Clin Mol Hepatol* **19(1)**, (2013), 17-25.

[5] Davis GL, Lindsay K, Albrecht J, et al., Clinical predictors of response to recombinant interferon-alpha treatment in patients with chronic non-A, non-B hepatitis (hepatitis C). The hepatitis interventional therapy group, *Journal of Viral Hepatitis* **1**, (1994), 55-63.

[6] Edward F. Vonesh, Generalized linear and nonlinear models for correlated data: theory and applications using SAS, *SAS Institute* **Principle Component Analysis** (2012), 01-55

[7] Elizabeth Stapel, Box-and-whisker plots: quartiles, boxes, and whiskers, *Purplemath* **1** (2013), 1-3.

[8] Eva Herrmann, Jung-Hun Lee, George Marinos, Marlene Modi, and Stefan Zeuzem, Effect of ribavirin on hepatitis C viral kinetics in patients treated with pegylated interferon, *Hepatology* **37**, (2003), 1351-1358.

[9] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton London New York, 2009.

[10] H.T. Banks, Keri L. Rehm, Karyn L. Sutton, Christine Davis, Lisa Hail, Alexis Kuerbis, Jon Morgenstern, Model formulation of drinking behavior using longitudinal data, CRSC-TR10-19, N.C. State University, Raleigh, NC, December, 2010.

[11] Hung CH, Lee CM, Lu SN, Wang JH, Tung HD, Chen TM, Chen CH, Changchien CS, Is delayed normalization of alanine aminotransferase a poor prognostic predictor in chronic hepatitis C patients treated with a combined interferon and ribavirin therapy?, *J Gastroenterol Hepatol* **(17)12**, (2002), 1307-11.

[12] John J. Lemasters, Dying a thousand deaths: redundant pathways from different organelles to apoptosis and necrosis, *Gastroenterology* **129(1)** (2005), 351-360.

[13] King, J.: *Practical Clinical Enzymology*, London, D. Van Nostrand Company, 1965. <http://www.aaltoscientific.com/purifiedhumanproteins/AlanineAminotransferase.php>

[14] Kwo P., Phase III results in Genotype 1 naive patients: predictors of response with boceprevir and telaprevir combined with pegylated interferon and ribavirin, *Liver International* **1** (2011), 39-43.

[15] Palmer, Chris, Livers Created from Stem Cells, *The Scientist*, (2013).

[16] Yun Jung Kim, Rapid normalization of alanine aminotransferase predicts viral response during combined peginterferon and ribavirin treatment in chronic hepatitis C patients, *Korean J Hapatol* **18**, (2012), 41-47.