**Summary Report**

## 1.0 Introduction

Globally, chronic liver disease causes many complications that lead to death each year. One of such chronic liver diseases is Hepatitis C which is a contagious liver disease caused by infection with Hepatitis C virus. The severity of this illness can range from mild illness that lasts a few weeks to a serious, lifelong condition. People are usually infected with Hepatitis C by receiving blood from an infected person. Hepatitis C is one of the most common liver infections with a record of 3-4 million infected people yearly (El Houby, 2014). Approximately 150 million people live with this chronic infection, and some may develop liver cancer and cirrhosis and every year, more than 350,000 people die of liver disease related to hepatitis C. Furthermore, the rates of new hepatitis C virus is on the rise which is the motivation behind this study.

## 1.1 Research Question

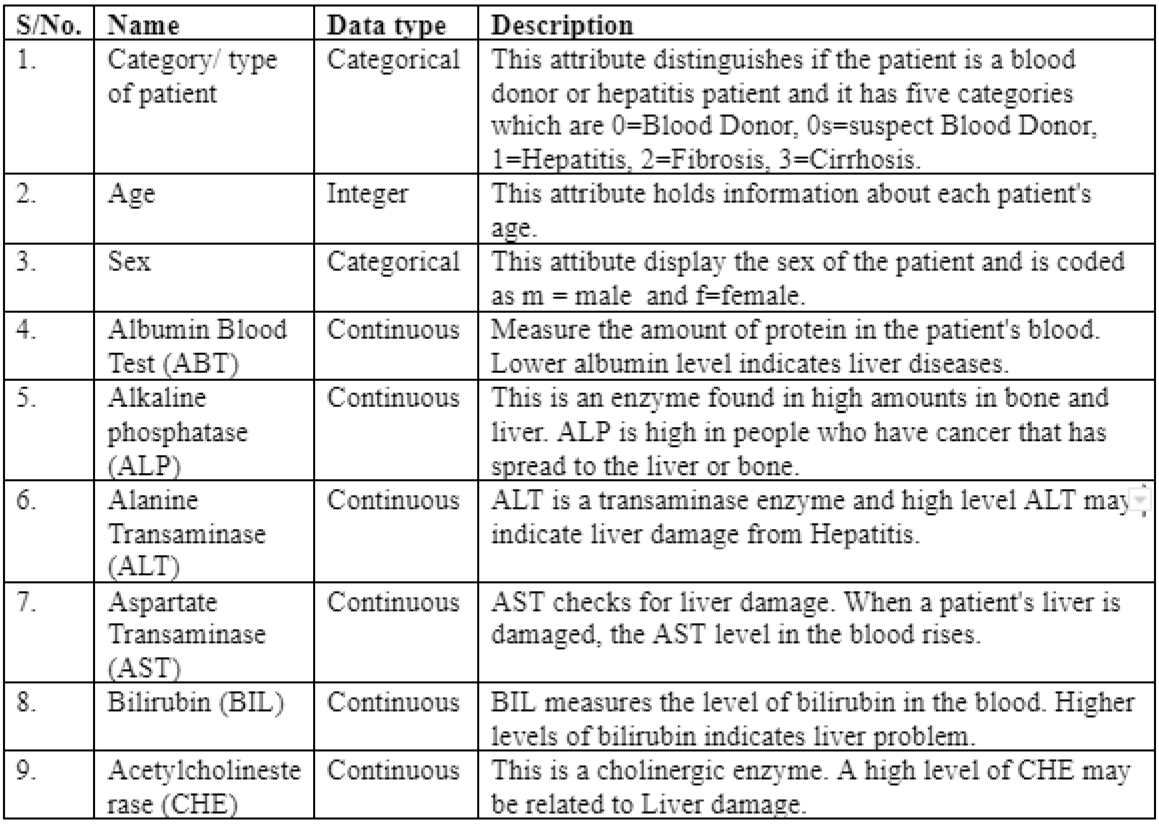
As the infection of Hepatitis C around the world continues to rise (Kokki *et al*, 2016), researchers have continued to assess demographic influence on hepatitis patients. This study uses laboratory values of blood donors and hepatitis patients at different stages which includes its progress (suspected blood donors, Hepatitis C, Fibrosis, Cirrhosis). This study aims to assess if demographic features have any influence on the category (blood donor/hepatitis) of a patient, if demographic features have any influence on the laboratory values, and if there is any difference between the laboratory values of blood donors and hepatitis C patients.

To achieve the research question, the following statistical analysis/test will be conducted;

1. Tests for significant differences between the age distribution of blood donors and hepatitis C patients in the dataset.
2. Explore the distribution of sex between blood donors and hepatitis patients and test the proportion of females in each of the two groups of blood donors and hepatitis patients.
3. Test for significant difference between the Alkaline phosphatase (ALP) for male and female patients.
4. Test for significant difference between the Alkaline phosphatase (ALP) for blood donors and hepatitis patients.
5. Test for significant difference between the Acetylcholinesterase (CHE) for male and female patients.
6. Test for significant difference between the Acetylcholinesterase (CHE) for blood donors and hepatitis patients.
7. Test for significant difference between the Alanine Transaminase (ALT) for male and female patients.
8. Test for significant difference between the Alanine Transaminase (ALT) for blood donors and hepatitis patients.

## 1.2 Dataset Description

The hepatitis C dataset contains laboratory values of blood donors and hepatitis patients. The dataset has 615 observations and 14 attributes with the present of some missing values. The dataset is sourced from the kaggle online database. The dataset was compiled by Ralf Lichtinghagen, Frank Klawonn and Georg Hoffmann. The table below gives a detailed description of each attribute in the dataset.

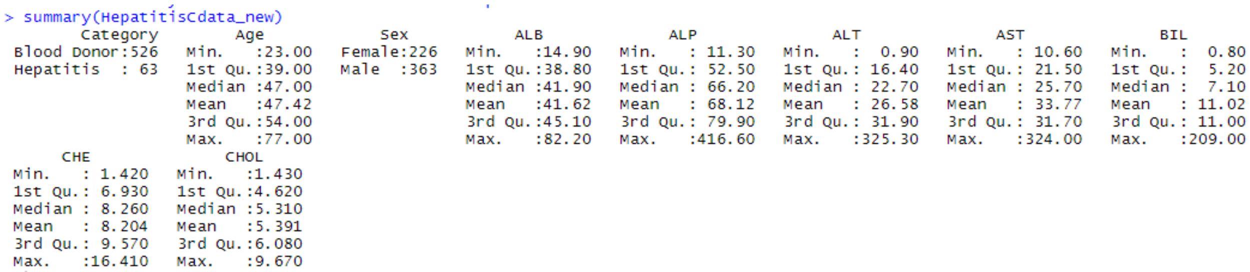


| S/No. | Name | Data type | Description |
|---|---|---|---|
| 1. | Category/ type of patient | Categorical | This attribute distinguishes if the patient is a blood donor or hepatitis patient and it has five categories which are 0=Blood Donor, 0s=suspect Blood Donor, 1=Hepatitis, 2=Fibrosis, 3=Cirrhosis. |
| 2. | Age | Integer | This attribute holds information about each patient's age. |
| 3. | Sex | Categorical | This attibute display the sex of the patient and is coded as m = male and f=female. |
| 4. | Albumin Blood Test (ABT) | Continuous | Measure the amount of protein in the patient's blood. Lower albumin level indicates liver diseases. |
| 5. | Alkaline phosphatase (ALP) | Continuous | This is an enzyme found in high amounts in bone and liver. ALP is high in people who have cancer that has spread to the liver or bone. |
| 6. | Alanine Transaminase (ALT) | Continuous | ALT is a transaminase enzyme and high level ALT may indicate liver damage from Hepatitis. |
| 7. | Aspartate Transaminase (AST) | Continuous | AST checks for liver damage. When a patient's liver is damaged, the AST level in the blood rises. |
| 8. | Bilirubin (BIL) | Continuous | BIL measures the level of bilirubin in the blood. Higher levels of bilirubin indicates liver problem. |
| 9. | Acetylcholinesterase (CHE) | Continuous | This is a cholinergic enzyme. A high level of CHE may be related to Liver damage. |

## 2.0 Methodology and Result

The Hepatitis C dataset contains 615 observations, and 14 attributes which later reduced to 589 observations after preprocessing technique of dropping all missing was performed on the dataset. Descriptive statistics were carried out on the final data to understand the distribution of our dataset. The category attribute in the data was recategorized to having only two levels: blood donor and Hepatitis. The Hepatitis level is a summation of the different stages of Hepatitis, including suspected blood donors. To answer the research questions, statistical tests were used, including the Shapiro Wilk test, Wilcoxon sign rank test, and two-way analysis of variance test. All data manipulations, statistical tests, and graph plotting were done using R programming with related libraries like dplyr, plyr, tidyr, tidyverse and ggplot2. The Shapiro Wilk test is used to check if the data follows a normal distribution. This will better inform us on the appropriate test statistics to employ in testing for significant influence. p-value<0.05 suggests a significant difference in the data. All tests carried out in this study check for significant effect but do not provide the effect's cause.

## 2.1 Descriptive statistics

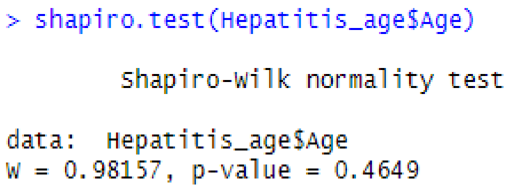


```
> summary(HepatitisCdata_new)
        Category        Age              Sex           ALB             ALP              ALT              AST              BIL        
 Blood Donor:526   Min.   :23.00   Female:226   Min.   :14.90   Min.   : 11.30   Min.   :  0.90   Min.   : 10.60   Min.   :  0.80  
 Hepatitis  : 63   1st Qu.:39.00   Male  :363   1st Qu.:38.80   1st Qu.: 52.50   1st Qu.: 16.40   1st Qu.: 21.50   1st Qu.:  5.20  
                   Median :47.00                Median :41.90   Median : 66.20   Median : 22.70   Median : 25.70   Median :  7.10  
                   Mean   :47.42                Mean   :41.62   Mean   : 68.12   Mean   : 26.58   Mean   : 33.77   Mean   : 11.02  
                   3rd Qu.:54.00                3rd Qu.:45.10   3rd Qu.: 79.90   3rd Qu.: 31.90   3rd Qu.: 31.70   3rd Qu.: 11.00  
                   Max.   :77.00                Max.   :82.20   Max.   :416.60   Max.   :325.30   Max.   :324.00   Max.   :209.00  
      CHE              CHOL      
 Min.   : 1.420   Min.   :1.430  
 1st Qu.: 6.930   1st Qu.:4.620  
 Median : 8.260   Median :5.310  
 Mean   : 8.204   Mean   :5.391  
 3rd Qu.: 9.570   3rd Qu.:6.080  
 Max.   :16.410   Max.   :9.670  
```

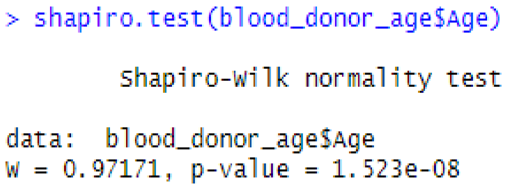
**Comment**

The figure above shows the descriptive statistics for all attributes in the data frame. Descriptive statistics for numerical attributes are represented in terms of mean, median, maximum and interquartile range, while categorical attributes are represented in terms of frequency of each level in the attribute.

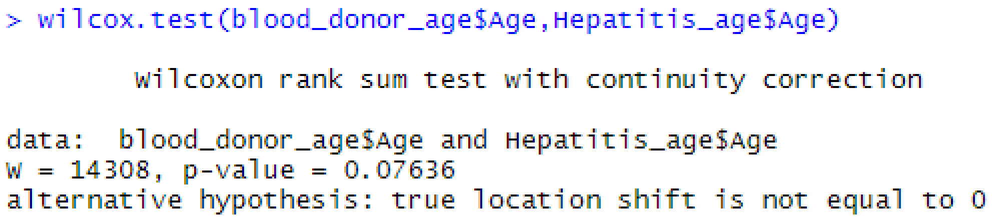
### 2.2 Tests for significant differences in age distribution of blood donors and hepatitis C patients



```
> shapiro.test(Hepatitis_age$Age)

        Shapiro-Wilk normality test

data:  Hepatitis_age$Age
W = 0.98157, p-value = 0.4649
```



```
> shapiro.test(blood_donor_age$Age)

        Shapiro-Wilk normality test

data:  blood_donor_age$Age
W = 0.97171, p-value = 1.523e-08
```

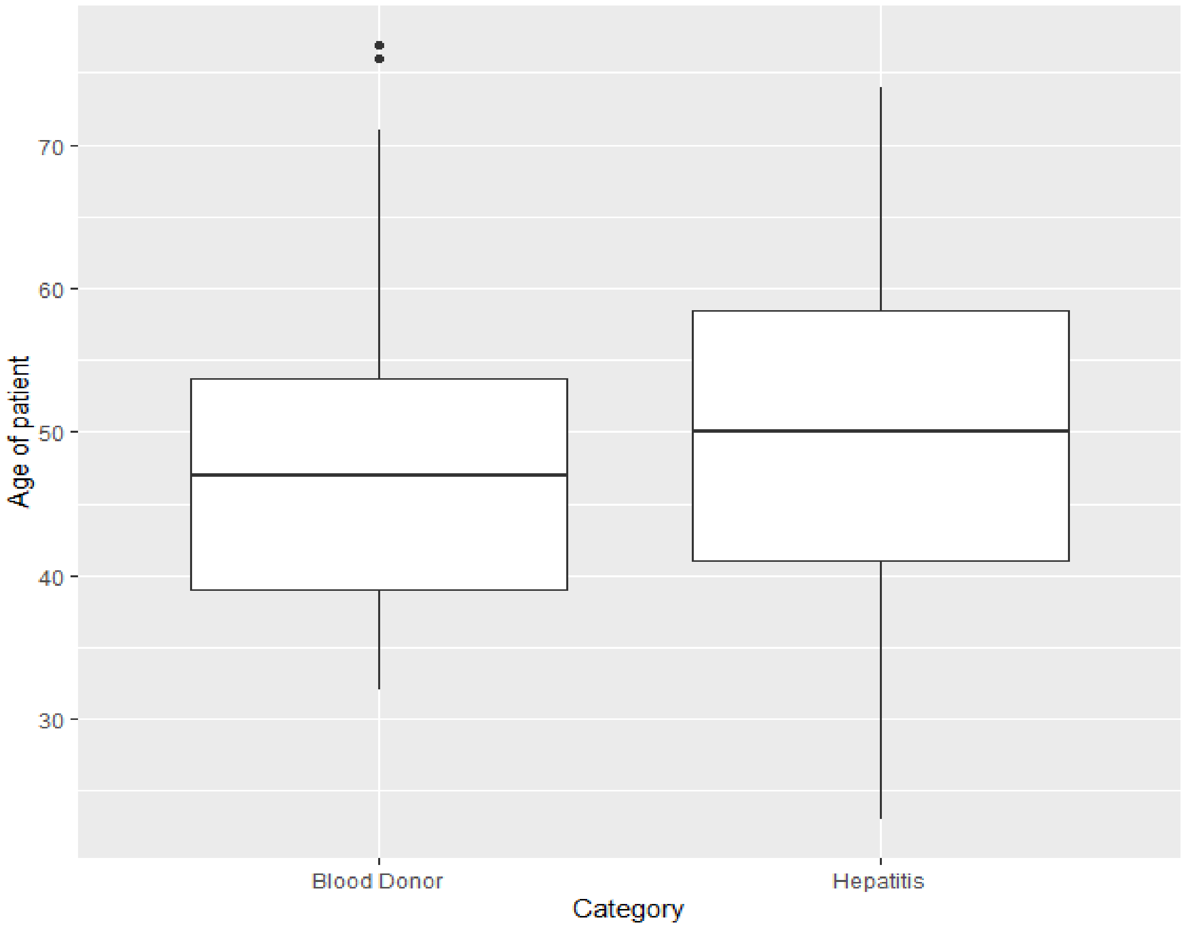


```
> wilcox.test(blood_donor_age$Age,Hepatitis_age$Age)

        Wilcoxon rank sum test with continuity correction

data:  blood_donor_age$Age and Hepatitis_age$Age
W = 14308, p-value = 0.07636
alternative hypothesis: true location shift is not equal to 0
```

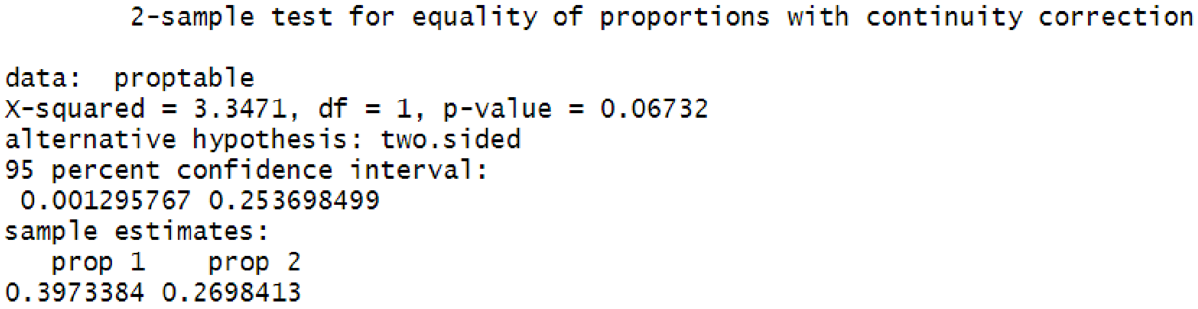
**Comment**

The Shapiro Wilk test on the ages of hepatitis patients suggests that the data comes from a normal distribution (p>0.05) which is the inverse for the ages of blood donor patients. To test for significant effect using the Wilcoxon rank test suggests that the age of a patient does not have any influence (p-value>0.05)on the category (blood donor or hepatitis) of a patient. This can further be seen in figure 1 below, as the box plot for each category does not suggest any significant difference.

**Figure 1: Boxplot showing the age distribution of a patient by category**



## 2.3 Test for Proportion of Females

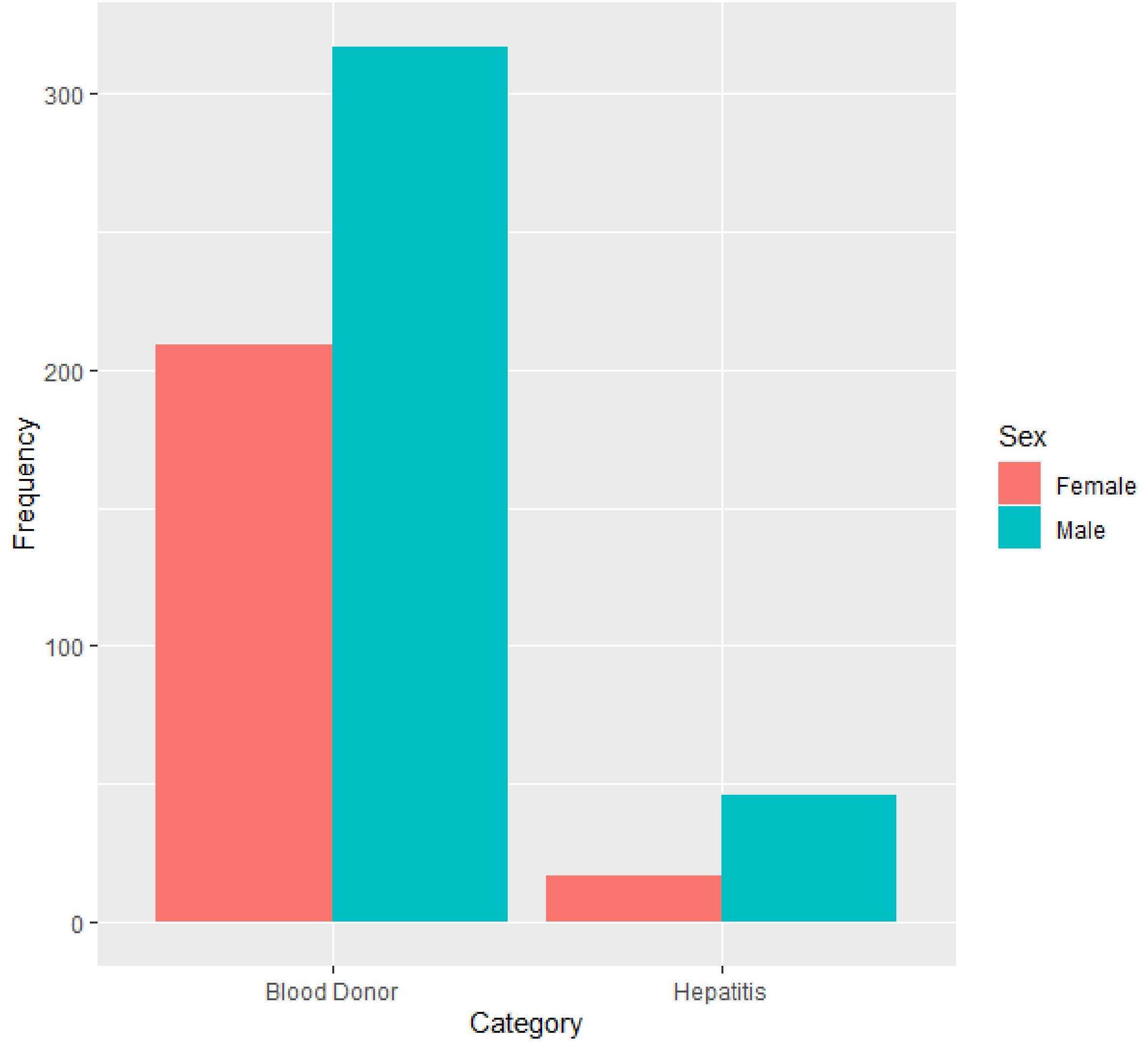


```
        2-sample test for equality of proportions with continuity correction

data:  proptable
X-squared = 3.3471, df = 1, p-value = 0.06732
alternative hypothesis: two.sided
95 percent confidence interval:
 0.001295767 0.253698499
sample estimates:
   prop 1    prop 2 
0.3973384 0.2698413 
```

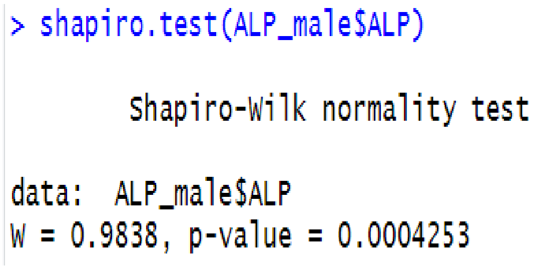
**Comment**

The test for the proportion of females in each category suggests no significant difference in the proportion of females in the two categories (blood donor and hepatitis), which can be generalized by inferencing that the sex of a patient does not have any influence on the category a patient.

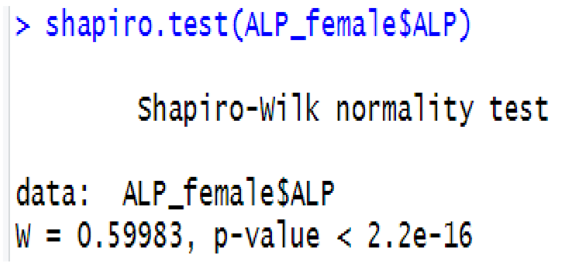
**Figure 2: Bar chart showing the frequency of category by sex**



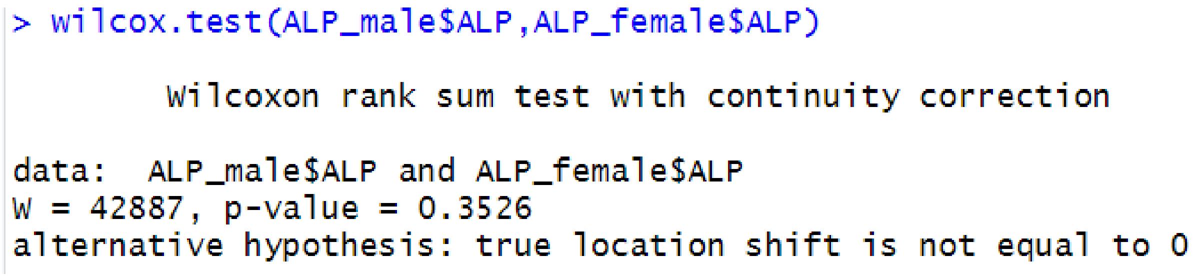
### 2.4 Test for difference in ALP for male and female patients



```
> shapiro.test(ALP_male$ALP)

        Shapiro-Wilk normality test

data:  ALP_male$ALP
W = 0.9838, p-value = 0.0004253
```



```
> shapiro.test(ALP_female$ALP)

        Shapiro-Wilk normality test

data:  ALP_female$ALP
W = 0.59983, p-value < 2.2e-16
```

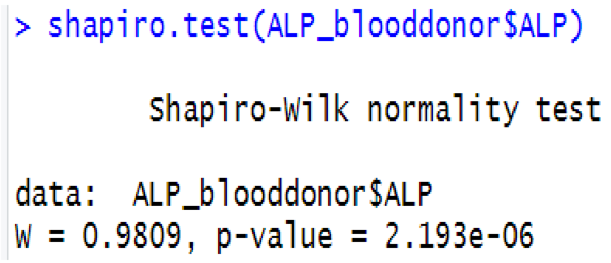


```
> wilcox.test(ALP_male$ALP,ALP_female$ALP)

        Wilcoxon rank sum test with continuity correction

data:  ALP_male$ALP and ALP_female$ALP
W = 42887, p-value = 0.3526
alternative hypothesis: true location shift is not equal to 0
```

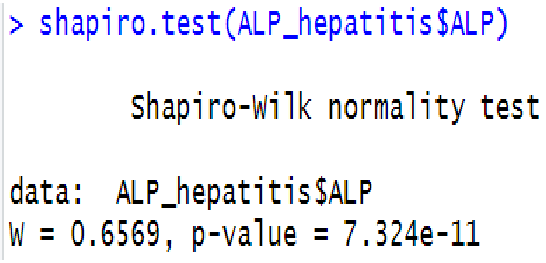
**Comment**

The Shapiro Wilk test suggests that both data do not come from a normal distribution (p-value<0.05). Testing for significant effect using the Wilcoxon rank test suggests that the sex of a patient does not have any significant influence (p-value>0.05) on the ALP value. And can be seen in figure 3 as the box plot for male and female in each category do not show any significant difference.

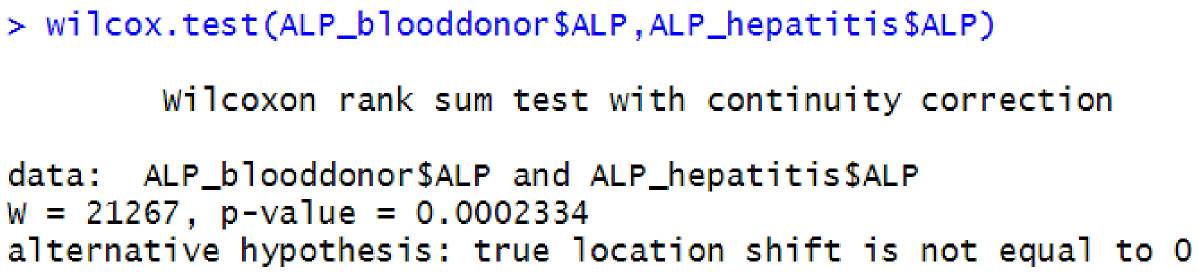
### 2.5 Test for difference in ALP for blood donor and hepatitis patients



```
> shapiro.test(ALP_blooddonor$ALP)

	Shapiro-Wilk normality test

data:  ALP_blooddonor$ALP
W = 0.9809, p-value = 2.193e-06
```



```
> shapiro.test(ALP_hepatitis$ALP)

	Shapiro-Wilk normality test

data:  ALP_hepatitis$ALP
W = 0.6569, p-value = 7.324e-11
```

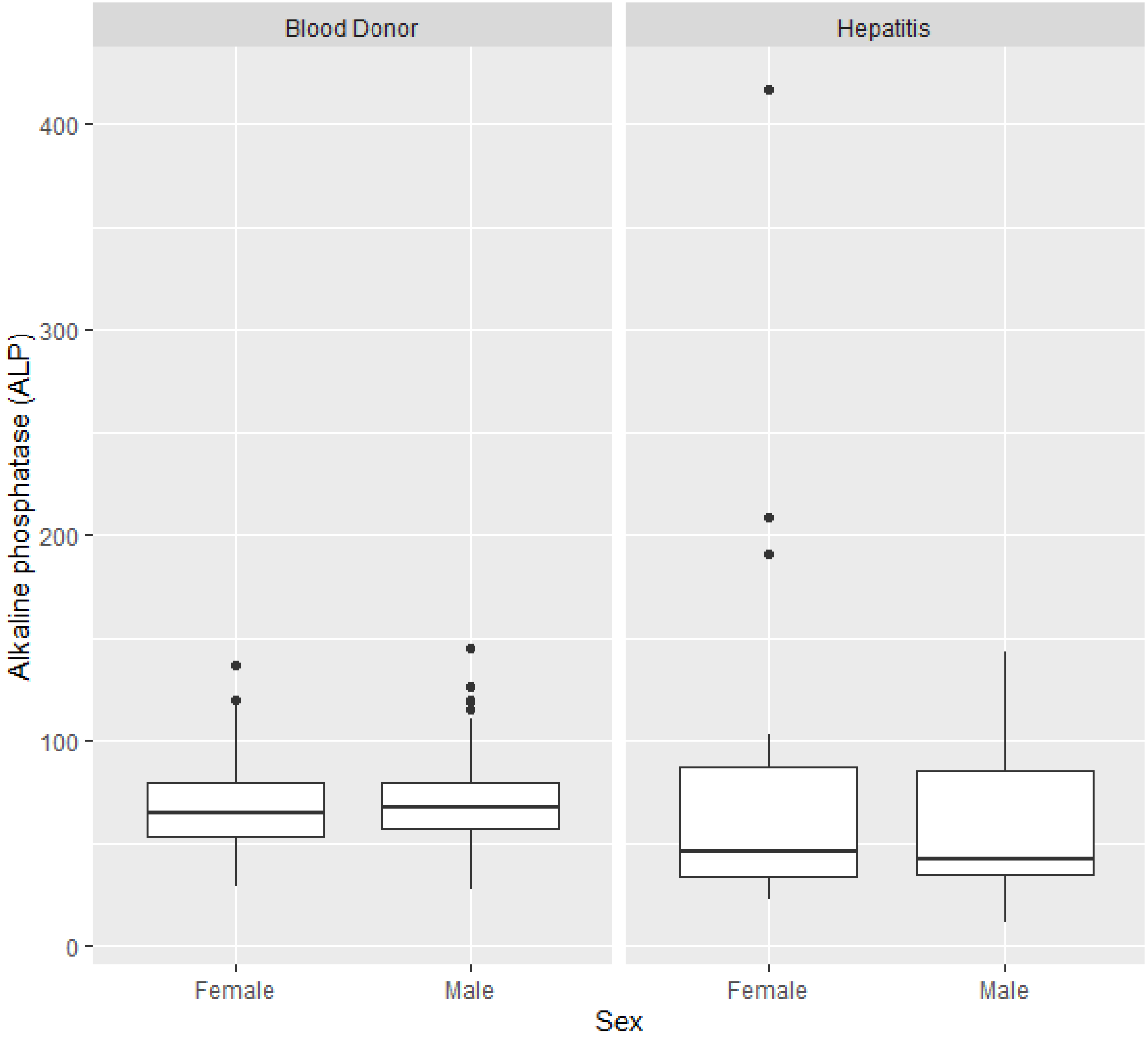


```
> wilcox.test(ALP_blooddonor$ALP,ALP_hepatitis$ALP)

	Wilcoxon rank sum test with continuity correction

data:  ALP_blooddonor$ALP and ALP_hepatitis$ALP
W = 21267, p-value = 0.0002334
alternative hypothesis: true location shift is not equal to 0
```

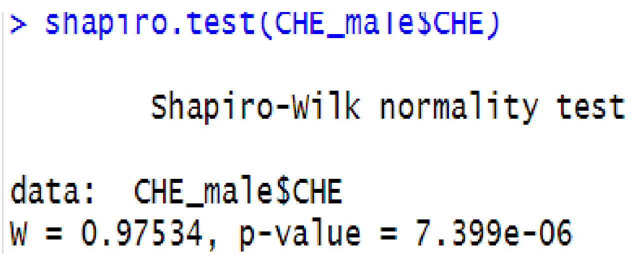
**Comment**

The Shapiro Wilk test suggests that both data do not come from a normal distribution (p-value<0.05). Using the Wilcoxon rank test to test for significant effect suggests that the patient category has a significant influence (p-value<0.05) on the ALP value. This can be further seen through the box plot in figure 3 below, which shows a significant difference in the ALP value between a blood donor and a hepatitis patient.

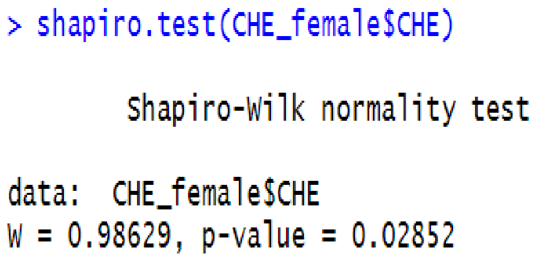
**Figure 3: Boxplot showing the ALP distribution of a patient by sex and category**



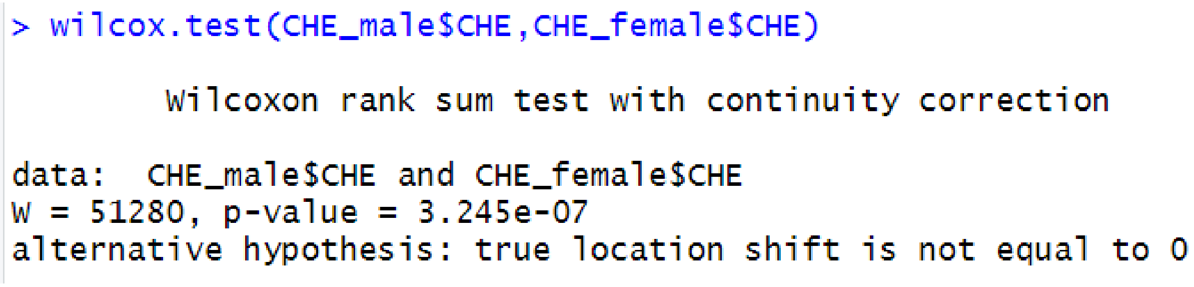
### 2.4 Test for difference in CHE for male and female patients



```
> shapiro.test(CHE_male$CHE)

        Shapiro-Wilk normality test

data:  CHE_male$CHE
W = 0.97534, p-value = 7.399e-06
```



```
> shapiro.test(CHE_female$CHE)

        Shapiro-Wilk normality test

data:  CHE_female$CHE
W = 0.98629, p-value = 0.02852
```

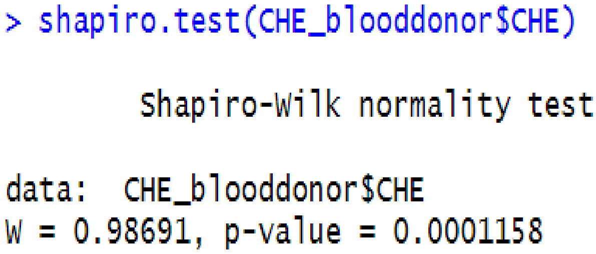


```
> wilcox.test(CHE_male$CHE,CHE_female$CHE)

        Wilcoxon rank sum test with continuity correction

data:  CHE_male$CHE and CHE_female$CHE
W = 51280, p-value = 3.245e-07
alternative hypothesis: true location shift is not equal to 0
```

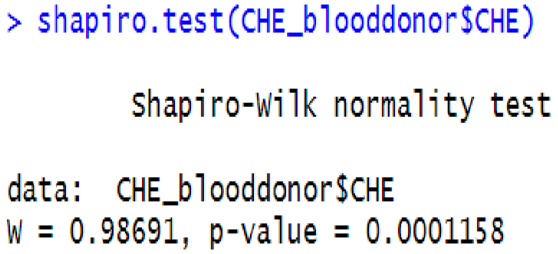
**Comment**

The Shapiro Wilk test indicates that both data do not come from a normal distribution (p-value<0.05). Hence, testing for significant effect using the Wilcoxon rank test suggests that the sex of a patient has a significant influence (p-value<0.05) on a patient's CHE value. This can be further seen through the box plot in figure 4 below, which shows a significant difference in the CHE value between male and female patients.

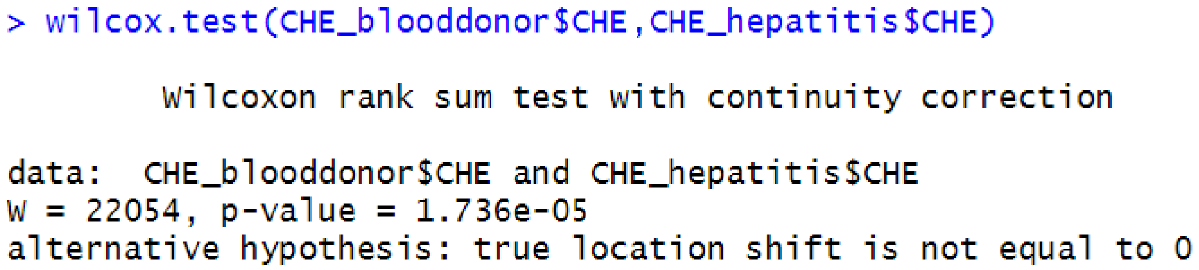
### 2.7 Test for difference in CHE for blood donor and hepatitis patients



```
> shapiro.test(CHE_blooddonor$CHE)

        Shapiro-Wilk normality test

data:  CHE_blooddonor$CHE
W = 0.98691, p-value = 0.0001158
```



```
> shapiro.test(CHE_blooddonor$CHE)

        Shapiro-Wilk normality test

data:  CHE_blooddonor$CHE
W = 0.98691, p-value = 0.0001158
```

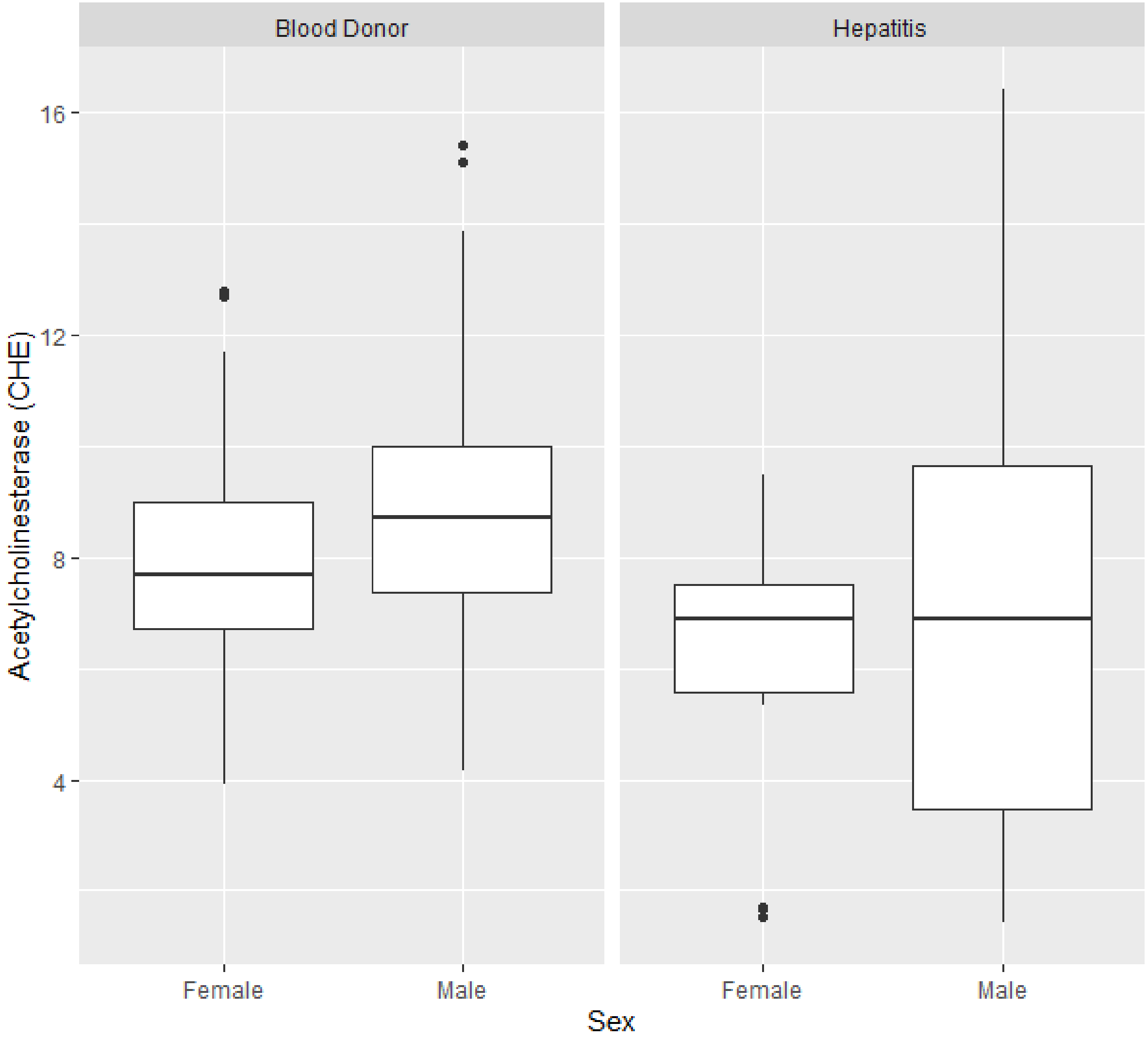


```
> wilcox.test(CHE_blooddonor$CHE,CHE_hepatitis$CHE)

        Wilcoxon rank sum test with continuity correction

data:  CHE_blooddonor$CHE and CHE_hepatitis$CHE
W = 22054, p-value = 1.736e-05
alternative hypothesis: true location shift is not equal to 0
```

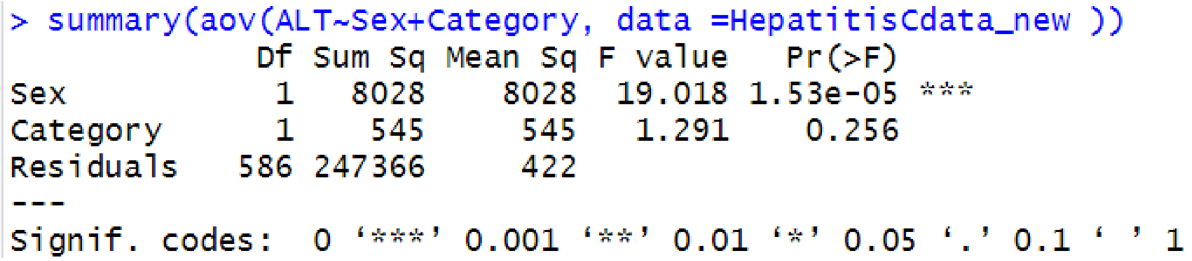
**Comment**

The Shapiro Wilk test indicates that both data do not come from a normal distribution (p-value<0.05). Hence, testing for significant effect using the Wilcoxon rank test suggests that the category of a patient has a significant influence (p-value<0.05) on a patient's CHE value. This can be further seen through the box plot in figure 4 below, which shows a significant difference in the CHE value between blood donor and hepatitis patients.

**Figure 4: Boxplot showing the CHE distribution of a patient by sex and category**



### 2.8 Test for difference in ALT for blood donor and hepatitis patients

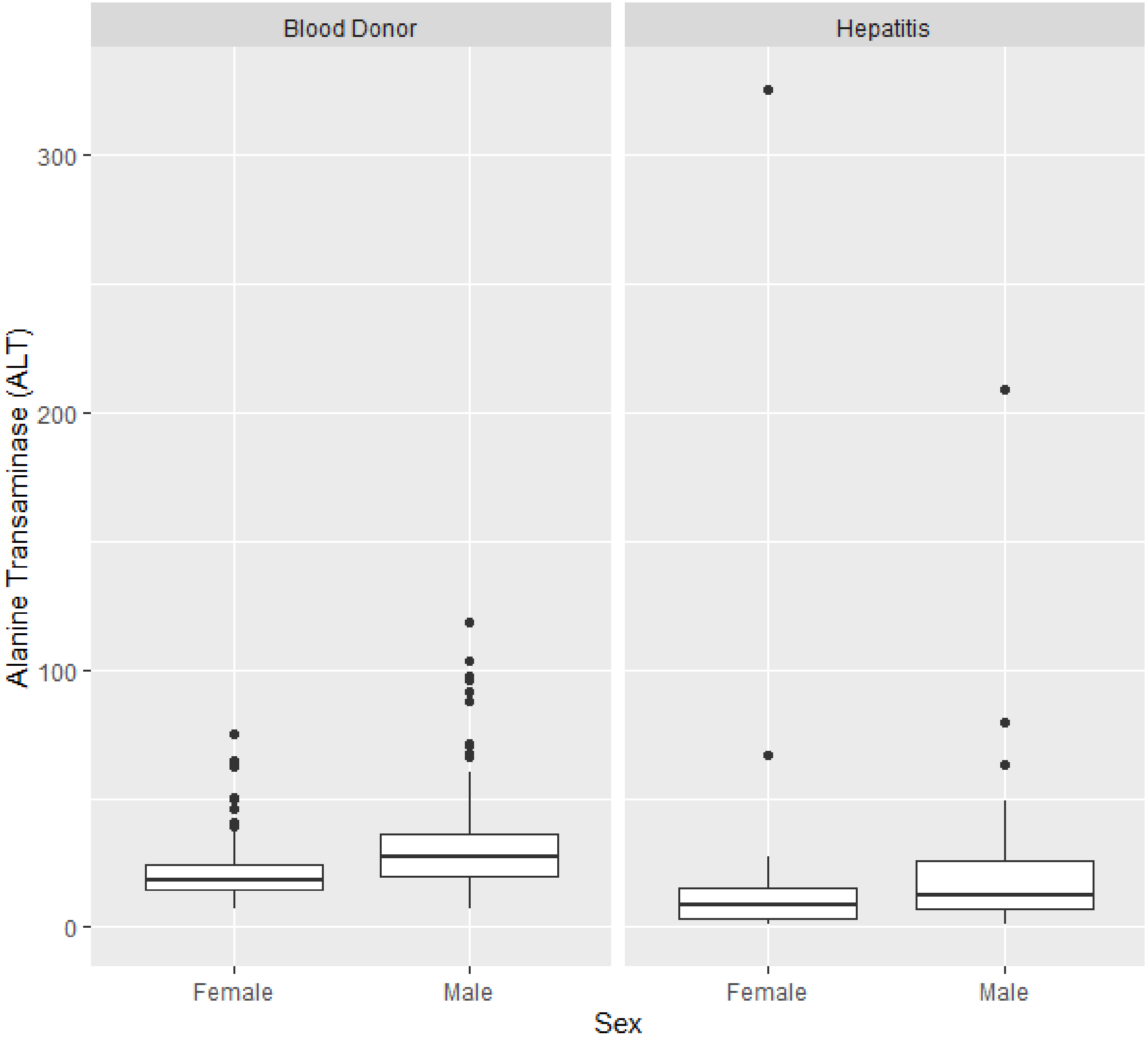


```
> summary(aov(ALT~Sex+Category, data =HepatitisCdata_new ))
             Df Sum Sq Mean Sq F value   Pr(>F)    
Sex           1   8028    8028  19.018 1.53e-05 ***
Category      1    545     545   1.291    0.256    
Residuals   586 247366     422                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Comment**

In this section, two-way analysis of variance (ANOVA) is use to test if the category and sex attribute influence the ALT laboratory values. The result above suggests that the sex of a patient have a significant influence (p-value<0.05) on the ALT value. In contrast, the category of a patient (blood donor or hepatitis) does not suggest any significant influence (p-value>0.05) on the ALT value of a patient. This can be seen in figure 5 below, which shows a difference in the box plot between males and females but shows a similar distribution for blood donors and hepatitis patients.

**Figure 5: Boxplot showing the ALT distribution of a patient by sex and category**



### 3.0 Conclusion

This study investigates demographic influence on some laboratory values of hepatitis and blood donor patients. It also tries to assess the influence of the category of a patient on their laboratory values. We discovered that demographic features have an influence on CHE and ALT, but ALP suggests the opposite. Similarly, while the category of a patient influences CHE and ALP, ALT suggests the opposite. This study builds on previous literature by (El-Salam *et al.,* 2019; Konerman *et al.*, 2019 and Hoffmann *et al.* 2018 ), where machine learning algorithms were used in investigating laboratory values and predicting diagnostics pathways for hepatitis patients.

**Reference**

Abd El-Salam, S.M., Ezz, M.M., Hashem, S., Elakel, W., Salama, R., ElMakhzangy, H. and ElHefnawi, M., 2019. Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients. *Informatics in Medicine Unlocked*, *17*, p.100267.

El Houby, E.M., 2014. A framework for prediction of response to HCV therapy using different data mining techniques. *Advances in bioinformatics*, *2014*.

Hoffmann, G., Bietenbeck, A., Lichtinghagen, R. and Klawonn, F., 2018. Using machine learning techniques to generate laboratory diagnostic pathways—a case study. *J Lab Precis Med*, *3*, p.58.

Konerman, M.A., Beste, L.A., Van, T., Liu, B., Zhang, X., Zhu, J., Saini, S.D., Su, G.L., Nallamothu, B.K., Ioannou, G.N. and Waljee, A.K., 2019. Machine learning models to predict disease progression among veterans with hepatitis C virus. *PloS one*, *14*(1), p.e0208141.

Kokki, I., Smith, D., Simmonds, P., Ramalingam, S., Wellington, L., Willocks, L., Johannessen, I. and Harvala, H., 2016. Hepatitis E virus is the leading cause of acute viral hepatitis in Lothian, Scotland. *New microbes and new infections*, *10*, pp.6-12.

World Health Organization (WHO), July 2012, http://www.who.int/mediacentre/factsheets/fs164/en/.

**Appendix**

Link to hepatitis dataset on kaggle - https://www.kaggle.com/fedesoriano/hepatitis-c-dataset