Parametric Analysis of Factor 8 (F8) Hemophilia A

A K M Raquibul Bashar[1] Chris P. Tsokos[2]

[1]University of South Florida, Department of Mathematics and Statistics
Tampa, FL, 33620, USA
[2]University of South Florida, Department of Mathematics and Statistics
Tampa, FL, 33620, USA

ABSTRACT: In this study, we are going to address some of the basic questions about the Factor 8 (F8) which is the reason behind hemophilia A. because hemophilia affects 1 in 5,000 male births. About 400 babies are born with hemophilia each year. The exact number of people living with hemophilia in the United States is not known. A CDC study conducted in six states in 1994 estimated that about 17,000 people had hemophilia at that time. Currently, the number of people with hemophilia in the United States is estimated to be about 20,000, based on expected births and deaths since 1994 [1]. In the United States, most people with hemophilia are diagnosed at a very young age. Based on CDC data, the median age at diagnosis is 36 months for people with mild hemophilia, 8 months for those with moderate hemophilia, and 1 month for those with severe Hemophilia.

So, after this study, we will be able to answer some critical basic questions from a statistical point of view, such as-does the level of severity depends on the races of the individuals? Is there any type of dependency among the inhibitor history, race and mutation type of the individuals? What could be the distribution of the level of severity? Etc.

Key Words: Inhibitor, F8, Hemophilia.

1 INTRODUCTION

Hemophilia is caused by a mutation or change, in one of the genes, that provides instructions for making the clotting factor proteins needed to form a blood clot. This change or mutation can prevent the clotting protein from working properly or to be missing altogether. These genes are located on the X chromosome. Males have one X and one Y chromosome (XY) and females have two X chromosomes (XX). Males inherit the X chromosome from their mothers and the Y chromosome from their fathers. Females inherit one X chromosome from each parent.

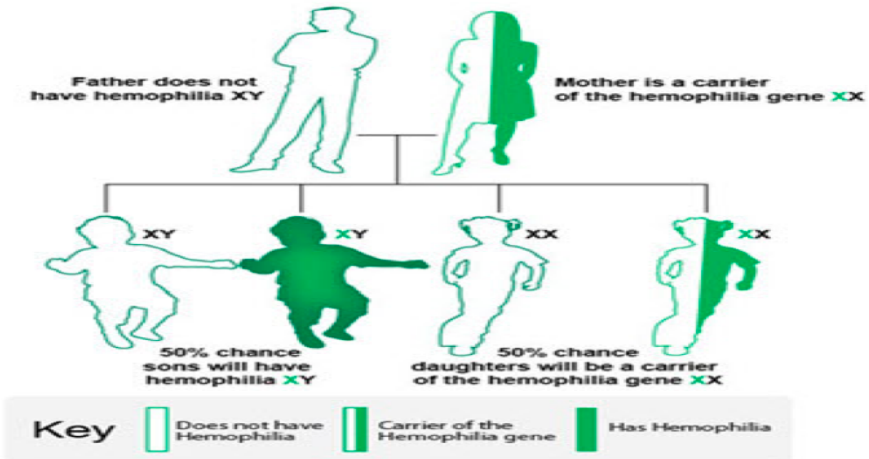

Figure 1: Parental relationship to the children with Hemophilia. (source: CDC)

The X chromosome contains many genes that are not present on the Y chromosome. This means that males only have one copy of most of the genes on the X chromosome, whereas females have 2 copies. Thus, males can have a disease like hemophilia if they inherit an affected X chromosome that has a mutation in either the factor VIII or factor IX gene. Females can also have hemophilia, but this is much rarer. In such cases both X chromosomes are affected or one is affected and the other is missing or inactive. In these females, bleeding symptoms may be similar to males with hemophilia.

Received August 15, 2018

A female with one affected X chromosome is a ”carrier” of hemophilia. Sometimes a female who is a carrier can have symptoms of hemophilia. In addition, she can pass the affected X chromosome with the clotting factor gene mutation on to her children [4].

2 DATA AND METHODOLOGY

2.1 DATA SOURCE

CHAMP F8 Mutations in the United States

Website: http://www.cdc.gov/ncbddd/hemophilia/champs.html

The Centers for Disease Control and Prevention (CDC) is conducting the largest survey to date of patients with hemophilia residing in the United States to identify the mutation in the gene that causes their hemophilia. From the CDC website the data has been extracted and analyzed for F8 mutation factor which causes the Hemophilia A.

2.2 DATA STRUCTURE AND DIAGRAM

The data set that has been extracted from the CDC database has a structure which is as follows in the

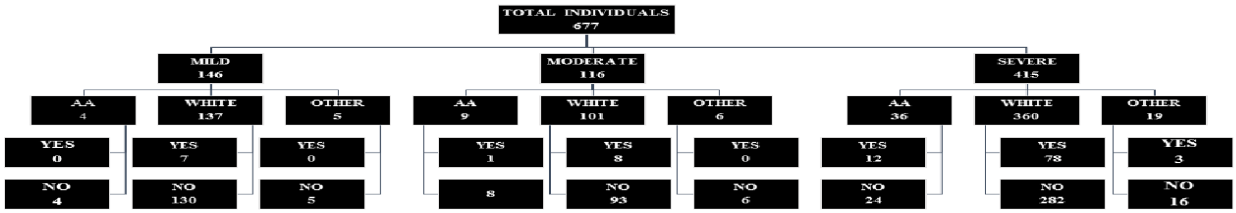

Figure 2: Data Diagram of the Hemophilia data structure

From the data diagram, we have total 677 individuals having Hemophilia A; out of this sample, there are three levels of Severity: Mild, Moderate and Severe. For each of these severity levels it includes three races categorized as- African American, White and Other. Those race categories are categorized by history of Inhibitors (Yes and No) with dichotomous categories. So, in this data set, we have one response variable named level of severity and two attributable variables named Race and History or Inhibitor. So, to answer the basic questions we are interested in, we have used some classical and new statistical methodology to conduct our research study.

2.3 METHODOLOGY

2.3.1 Univariate Analysis

To have a better illustration of the data in hand we have started to examine each of the variables in the data set and delve through the descriptive statistic to have an idea of the different categorical variables. So, in this type of classical problems we had to start with the classical approach to each of the variables starting with the severity levels of the individuals. From Table 1, we have the summary information of this severity level.

Level of severity	**Total Count**
Mild	146
Moderate	116
Severe	415
Total	677

Table 1: Frequency of Severity level

To have a clear picture on the proportions of the severity level Figure 3, we can refer back to the Figure 1 Pie chart with corresponding proportions.

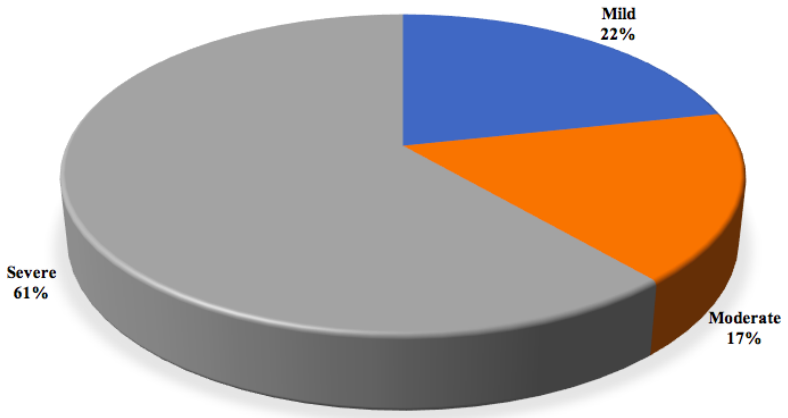

Figure 3: Pie Chart for Severity Level

From the same data set, if we take a look at the overall proportion of the individuals with respect to the different races then, we have the following illustration in Figure 4. It indicates that about 88% of the individuals are white according to the data we have.

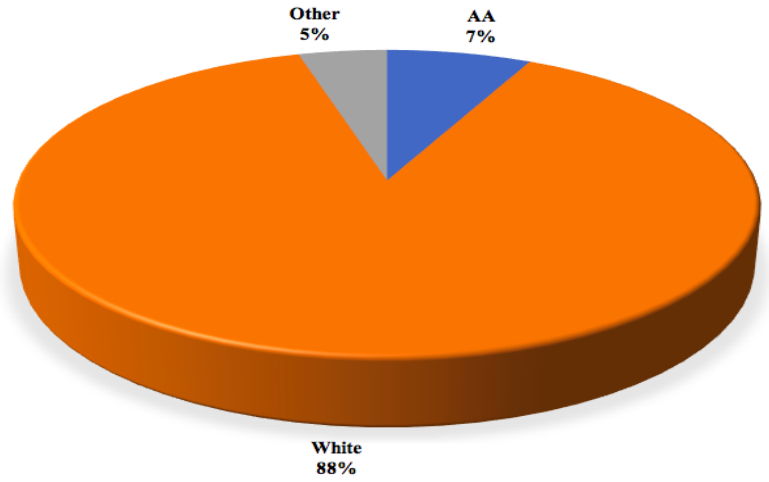

Figure 4: Pie Chart for Race

Similarly, we have visualized the other variables which is History of Inhibitors [3] presence in the body for individuals that plays a very significant role for the hemophilia. From the clinical research it was found that approximately 15%-20% of people with hemophilia will develop an antibodycalled an inhibitorto the product used to treat or prevent bleeding episodes. Developing an inhibitor is one of the most serious and costly complications of hemophilia [3]. So, from this fact we intend to look at the overall proportions in this very study data and as per our data set at hand there are approximately

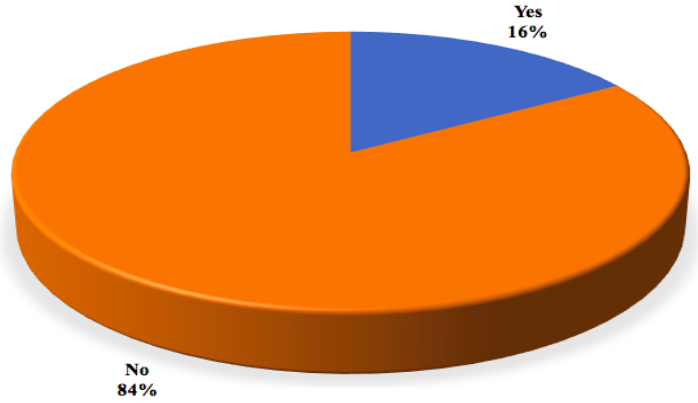

Figure 5: Pie Chart for Inhibitor History

84% of the individuals reported for hemophilia for **F8** have a history of inhibitors. This inhibitor was present in the individuals blood either due to inheritance from the parents or from the family.

2.3.2 Bi-variate Analysis

Now, after having a preliminary idea about each of the variables, we wanted to have a better picture of the cross tabulation for contingency and independence relationship. From this point of analysis [10], we will have an in-depth idea of the relationship among and between the variables and from there we will be able to do further analysis and get a clearer idea of the variables. To do that, we started with the contingency table of the the variables level of Severity and Races of Individuals.

	Severity			**Total**
	Mild	**Moderate**	**Severe**	
African American	4	9	36	49
White	137	101	360	598
Others	5	6	19	30
Total	146	116	415	677

Table 2: Contingency Table for Severity Levels and Races

From Table 2, we would like to answer the question “Does severity level depend on the races of individuals?”To answer this question, first of all we intend to look at the proportions of the severity levels for the three separate races. From Figure 6,

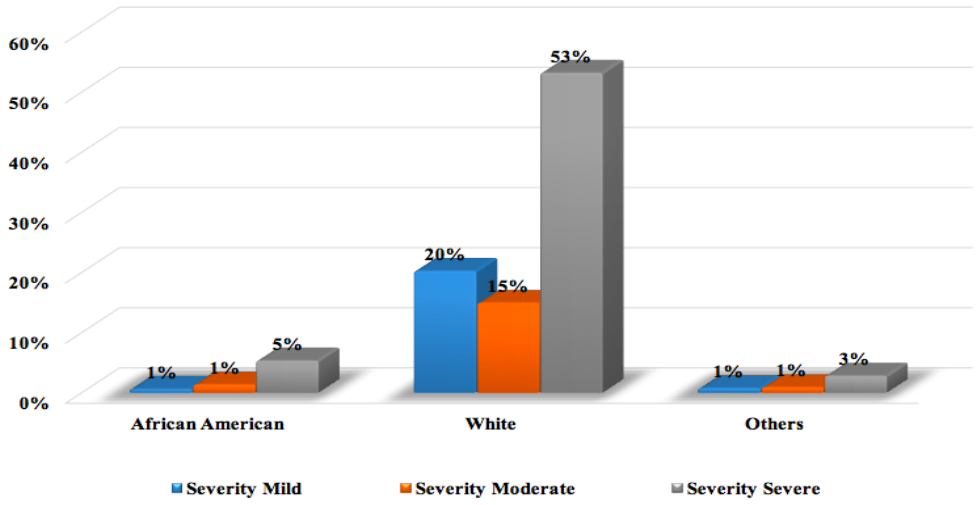

Figure 6: Pie Chart for Inhibitor History

it clearly indicates that the proportions of the races are significantly different from each other and might be affecting the severity levels of the individuals, but, without any statistical validation, it is only the hypothesis and we need to test this hypothesis to prove or disprove the claim that,

H_0 : *Race and Severity levels are Independent*
H_1 : *Race and Severity levels are not independent*

After having done the χ^2 test of independence from parametric perspective (Appendix A Table 4), we have found that the races and the level of severity for the reported Hemophilia A are not significantly dependent for the this US population at 5% level of significance with a p-value= 0.168. Alternatively, if we do the Non-parametric hypothesis test with the Kendall's tau [5] *b* and *c* (τ_b and τ_c) we have found negative correlations (Table 4 and 5) between these two categorical variables. Moreover, to be more comfortable with our findings, we have also examined the symmetric measures of association [6] of ordinal categorical response variable named *Phi* and *Cramer's V* (Appendix A Table 6) and also came to the same conclusions. We continued to look at the association [7] between severity level and the history of inhibitors among the individuals and the numerical summary is in Table 2 B as following,

Inhibitor	**Severity**			**Total**
	Mild	**Moderate**	**Severe**	
Yes	7	9	94	110
No	139	107	321	567
Total	146	116	415	677

Table 3: Contingency Table for Severity Levels and Inhibitor

To examine the association between inhibitor history and the level of severity of the individuals with Hemophilia A, we were interested to see if there is any relationship between these two categorical variables. So, we started with the hypothesizes as follows:

H_0 : *Inhibitor and Severity levels are Independent*
H_1 : *Inhibitor and Severity levels are not independent*

After conducting χ^2 test (Appendix, Table 8), as we did for level of severity and races, interestingly we have

found that there are very strong dependencies between the two categorical variables Level of Severity and History of Inhibitors presence in the family history of the individuals with Hemophilia A at 5% level of significance with p-value < 0.05. Also, it indicates a rejection of the hypothesis that all the population proportions of severity level are the same when there is a presence of inhibitor in the individuals. Moreover, to make sure the dependencies are really statistically significant, we have done the symmetric measures of association named Phi, Cramer's V and Kendall's τ_c (Appendix Table 9) and all of them proved significant relationship between Inhibitor and severity levels of the individuals. Now, we also need to test the independence between the attributable variables called Races and History of inhibitors, and from Table 10 (Appendix A) , we see that, there is one cell with less than 5% expected frequency which violates the assumption to proceed to do χ^2 test. So, we did the Fisher's Exact test and from Table 11, we see that, at 5% level of significance there is not enough evidence to reject the Hypothesis (p-value >0.05) that the races of individuals and history of inhibitors are independent according the data set we are dealing with. This conclusion made us to proceed with treating these two categorical variables (Race and History of inhibitor) are independent of each other. Since, we have already found that the history of inhibitor plays a significant role on the level of severity for the individuals of Hemophilia A patient with F8, we decided to run further analysis of the pairwise proportions of the variable Severity level, and to conduct this analysis we have used the method called "The Marascuilo Procedure"[9].

Rejecting the null hypothesis in a χ^2 test of equality of proportions in a $2 \times c$ table (Table 2 B) only allows us to reach the conclusion that not all c population proportions are equal, but which of the proportions differ? Because the result of χ^2 test for equality of proportions does not specifically answer this question, we need to use a multiple comparisons procedure such as the Marascuilo procedure. This procedure enables us to make comparisons between all pairs of groups. First, we have computed the differences $p_k - p_{\acute{k}}$ (where $k \neq k'$) among all $c \times (c-1) \div 2$ (where, *c= total number of categories or columns in the* $2 \times c$ *contingency table*) pairs. Then using Equation (1):

$$Critical\ Range = \sqrt{\chi^2_U}\sqrt{\left[\left(\frac{p_k\,(1-p_k)}{n_k}\right) + \left(\frac{p_{k'}\,(1-p_{k'})}{n_{k'}}\right)\right]} \tag{1}$$

here,
U= $(a,\ d.f.)\ and\ d.\ f.\ = c-1$;
c= *number of categories in column variable*

we have computed a different critical range for each pairwise comparison of sample proportions. In the final step, we compared each of the $c \times (c-1) \div 2$ pairs of sample proportions against its corresponding critical range. Then we decided a specific pair is significantly different if the absolute difference in the sample proportions $|p_k - p_{\acute{k}}|$ is greater than its critical range. In accordance with the Table 8, we have decided that there was evidence of a significant difference among the population proportions of severity level in the presence of Inhibitor for individuals. And since we have three levels of severity, then according to the Marascuilo Procedure we needed to do $3 \times (3-1) \div 2 = 3$ pairwise comparisons. After calculating three sample proportions for three different severity levels mild, moderate and severe when inhibitor are present in the blood are $p_1 = 0.048$, $p_2 = 0.078$, $p_3 = 0.227$ respectively.

Now, since the overall level of significance is 0.05, the upper tail critical value of the χ^2 test statistic for a chi-square distribution having $(c-1) = 2$ degrees of freedom is 5.991, thus, $\sqrt{\chi^2_U}$= 2.448. Next, we computed the three pairs of absolute differences in sample proportions and their corresponding critical ranges. If the absolute difference is greater than it's critical range, the proportions are significantly different:

Absolute difference in Proportions	**Critical Range**
$\|p_1 - p_2\| = \|0.048 - 0.078\| = 0.03$	0.07476462
$\|p_1 - p_3\| = \|0.048 - 0.227\| = 0.179$	0.06639704
$\|p_2 - p_3\| = \|0.078 - 0.227\| = 0.149$	0.07904317

Table 4: Comparison of severity Proportions in presence of inhibitor in the blood

From the above table, we see that the absolute difference of the sample proportions between mild and severe is greater than their critical range and it also the same situation between moderate and severe when history of inhibitors is positive for the individuals. So, the result indicates that, there is no significant difference between Mild and Moderate level of severity. However, both the Mild and Moderate levels are shown to differ significantly from the Severe level of Severity. When presence of inhibitor is "Yes "the proportions for the severe level of the disease is highest; therefore, if anyone decides to take any actions to address this problem, then he/she should start with the Severe category of the disease.

3 LOCAL and CUMULATIVE ODDS RATIOS COMPARISON

3.1 LOCAL ODDS RATIOS

For $r \times c$ tables, odds ratios can use each pair of rows in combination with each pair of columns. For rows *a* and *b* and columns *c* and *d*, the odds ratio $n_{ac}n_{bd}/n_{bc}n_{ad}$ uses four cells falling in a rectangular pattern. All such odds ratios of this type are determined by a basic set of $(r-1) \times (c-1)$ odds ratios [10]. One basic set consists of the odds ratios:

$$\hat{\theta} = \frac{n_{ij}n_{rc}}{n_{rj}n_{ic}} \tag{2}$$
$$i = 1, ..., r-1, \, j = 1, ..., c-1$$

which use cell in the last row and last column as a baseline. Each odds ratio is formed using the rectangular array of cells determined by rows i and r and columns j and c. In literature it is defined as the "Local Odds Ratios".

3.2 CUMULATIVE ODDS RATIOS

Another family of odds ratios that distinguishes between rows and columns is "Cumulative Odds Ratios "[10], these odds ratios are local in the row variable but global in the column variable and is defined as:

$$\hat{\theta}_{ij}^{C} = \frac{\left(\sum_{b \le j} n_{ib}\right)\left(\sum_{b > j} n_{i+1,b}\right)}{\left(\sum_{b > j} n_{ib}\right)\left(\sum_{b \le j} n_{i+1,b}\right)} \tag{3}$$

An equivalent definition for these odds ratios uses the sample conditional cumulative distribution functions of Y given X, since, they provide a comparison of pairs of levels of X with respect to their entire conditional distribution on Y. Moreover, an equivalent definition for these odds ratios uses the sample conditional cumulative distribution functions of Y given X:

$$\hat{\theta}_{ij}^{C} = \frac{\left(\hat{F}_{j|i}\right) / \left(1 - \hat{F}_{j|i}\right)}{\left(\hat{F}_{j|i+1}\right) / \left(1 - \hat{F}_{j|i+1}\right)} \tag{4}$$

We have used these odds ratios in order to answer the questions we have, "For those individuals having disease Hemophilia A from African American race, what is the odds being in the Mild stage of disease rather than in the Moderate or severe stage? ". and "what is the odds of being in the mild stage of the disease rather than in the Moderate or in the Sever Stages of the disease for Whites?"and so on. So the following table will give us the idea about addressing these type of questions. After calculating the Local and Cumulative odds rations from Table 2A, we have found the following:

Odds Ratios	Local	Cumulative
θ_{11}	0.328	0.30
θ_{12}	0.891	0.55
θ_{21}	1.63	1.49
θ_{22}	0.89	1.14

Table 5: Local and Global Odds ratios of Severity Level and Race

These values mean that for those individuals with Hemophilia for *F8* of African American, the estimated odds being in Mild stage of the disease rather than Moderate stage of the disease are $\theta_{11}\,(local) = 0.328$ times the corresponding estimated odds for those of Whites. On the other hand, for the African American individuals, the estimated odds of being in the Mild level of severity of the disease rather than in Moderate or Severe stage of the disease are $\theta_{12}\,(Cumulative) = 0.30$ times the corresponding odds for those of Whites and Other races. Now , if we do the similar estimates from Table 2 B, were we have 2×3 contingency table for Severity level in column and History of inhibitors in the row, and from this sort of contingency table we should have $(r-1) \times (c-1) = (2-1) \times (3-1) = 2$ basic set of odds ratios in terms of Local and Cumulative. After applying the equations (2) and (3), we have the following table:

Odds Ratios	Local	Cumulative
θ_{11}	0.60	0.21
θ_{12}	0.30	0.22

Table 6: Local and cumulative Odds ratios for severity level and Inhibitor History

From above table, the estimated value means for those of Individuals having positive History of Inhibitors the estimated odds being in the severity level named Mild rather than the Moderate level of the disease are 0.6 times the corresponding estimated odds of those with the negative or no history of inhibitors. And from the estimated cumulative odds ratios column we see that for those with the presence of inhibitors the estimated odds being in Mild stage of the disease rather

than Moderate or Severe level of the disease are 0.30 times the corresponding estimated odds of the those without the inhibitor presence.

			Severity Level of Hemophilia A			Total
			Mild	Moderate	Severe	
Races of Individuals	African American	Count	4.00	9.00	36.00	49.00
		Expected Count	10.60	8.40	30.00	49.00
		%Within Race of Individuals	8.20%	18.40%	73.50%	100.00%
	Whites	Count	137.00	101.00	360.00	598.00
		Expected Count	129.00	102.50	366.60	598.00
		%Within Race of Individuals	22.90%	16.90%	60.20%	100.00%
	Others	Count	5.00	6.00	19.00	30.00
		Expected Count	6.50	5.10	18.40	30.00
		%Within Race of Individuals	16.70%	20.00%	63.30%	100.00%
	Total	Count	146.00	116.00	415.00	677.00
		Expected Count	146.00	116.00	415.00	677.00
		%Within Race of Individuals	21.60%	17.10%	61.30%	100.00%

Table 7: Cross Tabulation of Race and Severity Levels of Hemophilia A

χ^2 Test Results			
	Value	d. f.	Asymp. Sig. (2- Sided)
Pearson χ^2	6.45	4	0.168
No. of valid Cases	677		

Table 8: Chi- Squared test of Association between Race and Severity level

			Severity Level of Hemophilia A	Race of Individuals
Kendall's τ_b	Severity Level of Hemophilia A	Correlation Co- efficient	1.00	-0.051
		Sig. (2- tailed)	.	0.159
	Race of Individuals	Correlation Co- efficient	-0.051	1.00
		Sig. (2- tailed)	0.159	.
		N	677	677

Table 9: Non- parametric Correlation (Kendall's τ_b) between Race and Severity Level

Symmetric Measures					
		Value	Asymp. Std. Error	Approx. τ_b	Approx. Sig.
Nominal vs. Nominal	Phi	0.098			0.168
	Cramer's V	0.069			0.168
Ordinal vs. Ordinal	Kendall's τ_c	-0.026	0.017	-1.57	0.117
No. of Valid Cases		677			

Table 10: Symmetric Measures of Association between Races and Severity Level

History of Inhibitor in Individuals vs. Level of Severity of Individuals Cross tabulation						
			Severity Level of Hemophilia A			Total
			Mild	Moderate	Severe	
History of Inhibitor	NO	Count	139.00	107.00	321.00	567.00
		Expected Count	122.30	97.20	347.00	567.00
		%Within History of Inhibitor	24.50%	18.90%	56.60%	100.00%
	YES	Count	7.00	9.00	94.00	110.00
		Expected Count	23.70	18.80	67.40	110.00
		%Within Race of Individuals	6.40%	8.20%	85.50%	100.00%
	Total	Count	146.00	116.00	415.00	677.00
		Expected Count	146.00	116.00	415.00	677.00
		%Within Race of Individuals	21.60%	17.10%	61.30%	100.00%

Table 11: Cross Tabulation of Inhibitor History and Severity Levels of Hemophilia A

χ^2 Test Results			
	Value	d. f.	Asymp. Sig. (2- Sided)
Pearson χ^2	32.719	2	0.00
No. of valid Cases	677		

Table 12: Chi-squared test of independence between Inhibitor and Severity level

Symmetric Measures					
		Value	Asymp. Std. Error	Approx. τ_b	Approx. Sig.
Nominal vs. Nominal	Phi	0.22			0.00
	Cramer's V	0.22			0.00
Ordinal vs. Ordinal	Kendall's τ_c	0.161	0.024	6.7	0.00
No. of Valid Cases		677			

Table 13: Symmetric Measures of Association between Races and Severity Level

Presence of Inhibitor vs. Race of Individuals having Hemophilia A Cross tabulation						
			Races of Individuals Having Hemophilia A			Total
			African American	Whites	Others	
Presence of Inhibitor	NO	Count	36.00	504.00	27.00	567.00
		Expected Count	41.00	500.80	25.10	567.00
		%Within History of Inhibitor	6.30%	88.90%	4.80%	100.00%
	YES	Count	13.00	94.00	3.00	110.00
		Expected Count	8.00	97.20	4.90	110.00
		%Within Race of Individuals	11.80%	85.50%	2.70%	100.00%
	Total	Count	49.00	598.00	30.00	677.00
		Expected Count	49.00	598.00	30.00	677.00
		%Within Race of Individuals	7.20%	88.30%	4.40%	100.00%

Table 14: Cross Tabulation of Inhibitor History and Race

χ^2 Test & Fisher's Exact Test Results					
	Value	d. f.	Asymp. Sig. (2- Sided)	Exact Sig. (2- sided)	Exact Sig. (1- sided)
Pearson χ^2	4.79	2	0.091	0.086	
Fisher's Exact Test	4.404			0.109	
No. of valid Cases	677				

Table 15: Chi-squared test of independence between Inhibitor and Severity level

4 DISCUSSION and FURTHER SCOPE OF STUDY

From the above study, it is pretty conclusive that the severity level of hemophilia A is not much related to the races of the individuals of US population as per our working data set. Also, we have found that race and inhibitor history are

independent of each other. But, the severity level is highly dependent on the History of Individuals for US populations as we have proved it statistically significant. Also, from the local and cumulative odds ratios indicates that the Odds of Whites being in the Mild level of the disease is higher than those of the African Americans and other race categories. The Non- paramteric and Bayesian analysis [8] can be done in this study to have a better insights, but that could be another scope to extend this study further.

REFERENCES

[1] Soucie, J M and Evatt, B and Jackson, D, *Occurrence of hemophilia in the United States. The Hemophilia Surveillance System Project Investigators*, Am J Hematol, 59- 4 (1998) 288- 94.

[2] Ullman, M and Hoots, W K, *Assessing the costs for clinical care of patients with high-responding factor VIII and IX inhibitors* , Haemophilia, 12 Suppl 6 (2006) 74-9; discussion 79-80.

[3] Soucie, J M and Symons, 4th, J and Evatt, B and Brettler, D and Huszti, H and Linden, J, *Home-based factor infusion therapy and hospitalization for bleeding complications among males with haemophilia* , Haemophilia, 7- 2 (2001) 198- 206.

[4] Britton, Beverly, *Myths & facts...about hemophilia*, Nursing, 33 (2003) 78.

[5] M. G. Kendall, *A New Measure of Rank Correlation*, Biometrika, 30 (1938) 81- 93.

[6] William H. Kruskal, *Ordinal Measures of Association*, Journal of the American Statistical Association, 53, 284, (1958) 814- 861.

[7] R. C. Lewontin and J. Felsenstein, *The Robustness of Homogeneity Tests in* $2 \times N$ *Tables*, 21- 1 (1965) 19-33.

[8] Elaheh Rabiei and Enrique Lopez Droguett and Mohammad Modarres, *A prognostics approach based on the evolution of damage precursors using dynamic Bayesian networks*, Advances in Mechanical Engineering, 8- 9 (2016) 1687814016666747.

[9] Marascuilo, Leonard A. And Slaughter, Robert E., *Statistical Procedures For Identifying Possible Sources Of Item Bias Based On X2 Statistics*, 18, 4, (1981) 229- 248.

[10] Agresti, Alan, *Analysis of Ordinal Categorical Data; 2nd ed.*, (2012).