Honors Thesis – Spring 2020

A Look Into Systemic Lupus Erythematosus Diagnoses

Brittany Hickerson, MSIS, Stillwater, OK
Dr. Bryan Hammer, Oklahoma State University, Stillwater, OK
Dr. Andy Luse, Oklahoma State University, Stillwater, OK

With little to no professional credentials, it is my personal experiences that qualify me for this research. With multiple members of my family having been diagnosed with a variety of the more common autoimmune disorders over the years, I have a unique desire to uncover the truth.

ABSTRACT

An autoimmune disease or disorder occurs when the individual's immune system becomes confused and attacks/destroys healthy blood tissue by mistake (LFA). Now, in a perfect world, the immune system is supposed to protect us by attacking potentially dangerous foreign bodies but, sometimes the immune system gets it wrong. On a small scale this results in allergies while severe instances can result in a multitude of autoimmune disorders that can frequently lead to death. Currently, the exact cause for this overreaction of the immune system is unknown and, in each instance, it is different but, there are a few running theories on potential triggers. The theory I want to focus on for this project is as follows, "some microorganisms (such as bacteria or viruses) or drugs may trigger changes that confuse the immune system". If this is true, there could be correlations between medicines prescribed, length of surgeries or other similar factors.

Therefore, the goal of this paper is to explore and analyze CHSI data to determine if there are any factors that might lead to the diagnosis of Systemic Lupus Erythematosus (Lupus) in a patient's lifetime. The CHSI data provided is a nationwide dataset that has records pertaining to surgeries and relevant diagnoses in relation to said surgeries. Everything from demographic data on the patient to details about the facility performing the surgery along with medicinal data for after the surgery has been provided. Any correlations identified could have serious ramifications for the medical world as currently, there is no proof that the development of one or more autoimmune diseases is anything other than completely random.

My hypothesis is that the longer a patient is open on an operating table, the greater their odds of developing an autoimmune disorder but, I also intend to explore trends in the length of surgeries, patient suspicions, and any other available category that could be causing the diagnoses.

So, let us begin with a background on Lupus.

INTRODUCTION

Lupus is a systemic autoimmune disease that is the result of your immune system attacking your tissues and organs. This disease attacks many parts of the body including but not limited to the heart, lungs, skin, kidneys, and joints. It is frequently difficult to diagnose as it shares many symptoms with other diseases and like other autoimmune diseases, it impacts women more frequently than men. Furthermore, no two cases of Lupus are the same. Symptoms vary along with the speed at which they develop and at times symptom can disappear completely for extended periods of time but, a few of the most common symptoms are joint pain, fever, and severe fatigue (Mayo). Now, being able to diagnosis a disease is a piece of the puzzle but, what was the cause and what comes next?

For Lupus, the cause is primarily unknown but is believed to be partly due to genetics and partly the environment around you. If you are predisposed potential triggers include excessive exposure to sunlight, infections and occasionally medications but, with medications the symptoms generally go away once the medication is stopped. However, according to past research the following factors, gender (women), age (15-45), and race (African American, Hispanics and Asians), increase your chances of being diagnosed with Lupus (ANRF). So, once you are diagnosed the next steps are fairly complicated. Based upon your symptoms, the treatments differ, and they also differ along with your flare ups, periods where your symptoms are worse. However, the primary options are anti-inflammatory drugs, antimalarial drugs, and immunosuppressants (Mayo). Unfortunately, at this time there is no known cure so no matter how well the symptoms are treated, once diagnosed Lupus is something you will deal with for the rest of your life.

Although, a lot of research has been done on this disease as the Arthritis National Research Foundation alone has funded research for Lupus for over 30 years. Thankfully, more is known with this disease than other autoimmune diseases but, there is still a lot to learn. Currently, research is focusing in one preventative measures because, as mentioned earlier, there currently is not a cure but recently, a specific human chromosome region was tied to an increased risk of developing Lupus. This is a major milestone for predictive attempts and shines hope upon the future. Other studies have included research into the autoantibodies that are the disease, the impact of skin injuries upon kidney inflammation and methods using estrogen to treat Lupus (ANRF). Now, while there have been many other studies, everything I found revolved around preventative measures and then treatment once the disease developed. Therefore, I thought it would be interesting to look into potential cures and the best way I know of to achieve that is to start with what you have. So, next I will outline what we have and what I plan to do with it.

This paper will follow the exploration of a CHSI dataset that contains around one million surgical records and 79 different columns. Only around two thousand of those records contain positive diagnoses for Lupus and therefore, to increase the potential for results, the columns will be broken up into smaller, categorical groups. Before this however, overall ratios for demographic information like gender will be calculated for the entire dataset and then compared to the ratios for the positive cases. Afterwards, using SAS Enterprise Miner there will be an attempt to create a variety of predictive models using the smaller groups which include factors like location or severity of surgery to determine the most at-risk individuals. Lastly, we will discuss what we have learned and the impacts it may have in relation to other research and the search for a cure.

FIRST MAIN TOPIC <DATA PREPARATION>

To start, I was provided with twelve different rpt files covering a wide range of factors revolving around what will be referred to as an encounter for the duration of this report. An encounter is referencing one entry into the dataset and generally comes with demographic information, final diagnosis, and length of stay to name only a few. The following ordered list is my transformation of the files using RStudio.

1. Each of the twelve rpt files was converted into a txt file using the text editor, EmEditor and imported into RStudio.
2. Each of the previously mentioned twelve txt files had a partner file so, I merged them into a more manageable six files, one of which contained forty-five variables and one million rows.
3. To increase usability across platforms I converted them into six separate csv files.
4. I exported these files from RStudio so that I could import them into Tableau to analyze distribution.

It should be noted that each of the final six files had the same unique identifier, Encounter_ID.

The code for the conversion of two txt files in RStudio from start to finish:

```
###Read in txt files from computer
cevnt = read.delim("~/Thesis/Data/control/txt/cevnt.txt")
cevnt2 = read.delim("~/Thesis/Data/case/txt/cevnt.txt")

###Combine files
cevnt_c = rbind(cevnt,cevnt2)

###Export as csv
setwd("file path")
write.csv(cevnt_c,"cevnt_c.csv")
```

After this the next step was to prepare the data for SAS Enterprise Miner but, due to the large number of columns all six files would not efficiently merge together within RStudio. The following list details the steps taken.

1. Each of the six csv files were read back into RStudio after exporting the data for analysis within Tableau.
2. All columns with an excess of null values or redundancies were ignored. For example, if more than 50% of the rows within a column were null, that column was not included in the merged data. Alternatively, a common redundancy occurred with reference to unique identifiers. Only the main one was necessary but, each separate file had its own as well. In this instance every unique identifier except for Encounter_ID was ignored.
3. The new, reduced csv files were then merged on the previously mentioned unique identifier, Encounter_ID. This resulted in a final count of 28 columns.
4. A new column was created and added to the end of the newly merged dataset. This column was set to a 0 if the diagnosis was positive for Lupus and a 1 otherwise.
5. This merged file was then changed into two separate csv files, exported from RStudio, and lastly uploaded into SAS Enterprise Miner.

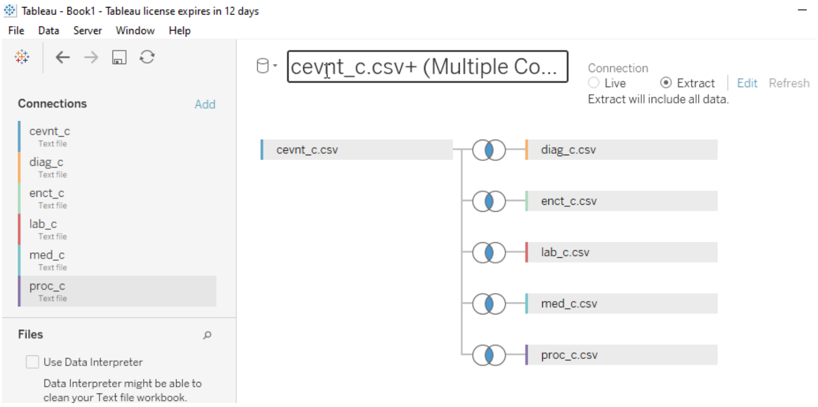
The code detailing this process for two of the csv files can be found below:

```
###Time to add a new column
install.packages("stringr")
library(stringr)
setwd("C:\\Users\\msis5223user\\Documents\\Thesis\\Data\\work\\csv\\")
diag_c = read.csv("diag_c.csv", sep = ',')
diag_c$Target = NA
head(diag_c)
diag_c$Target = ifelse(str_detect(diag_c$DIAGNOSIS_CODE,"M32.9"),0,1)

####Select chosen columns
install.packages("dplyr")
library(dplyr)
rd1 = select(diag_c, ï..ENCOUNTER_ID,….Target)
rd2 = select(enct_c, ï..ENCOUNTER_ID,….Target)
rd3 = select(lab_c, ï..ENCOUNTER_ID,….Target)
rd4 = select(med_c, ï..ENCOUNTER_ID,….Target)
head(rd1)

###merge
enct = merge(rd,rd1,by = 'ï..ENCOUNTER_ID')
write.csv(enct,"enct_final.csv")
.
.
.
med = merge(rd,rd4,by = 'ï..ENCOUNTER_ID')
write.csv(med,"med_final.csv")

###export
write.csv(merged,"sas.csv")
```

SECOND MAIN TOPIC <SUMMARY STATISTICS AND DISTRIBUTIONS>

For this topic, we will be staying within Tableau. As mentioned, at this point we have six individual csv files all with the same unique identifier, Encounter_ID. So, as I imported the files, I connected them on that variable.

Display 1 is a screenshot of that connection.



Connection ○ Live ● Extract

Extract will include all data.

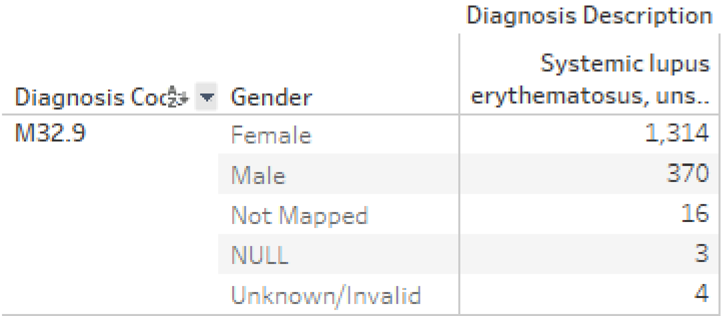
□ Use Data Interpreter

Data Interpreter might be able to clean your Text file workbook.

Display 1. Tableau Data Import and Connection

Throughout this data we have 1,707 positively identified cases of Lupus. The following visuals explore the demographic relationships among these cases, in comparison to the ratios throughout the entire dataset.

Display 2 explores gender.

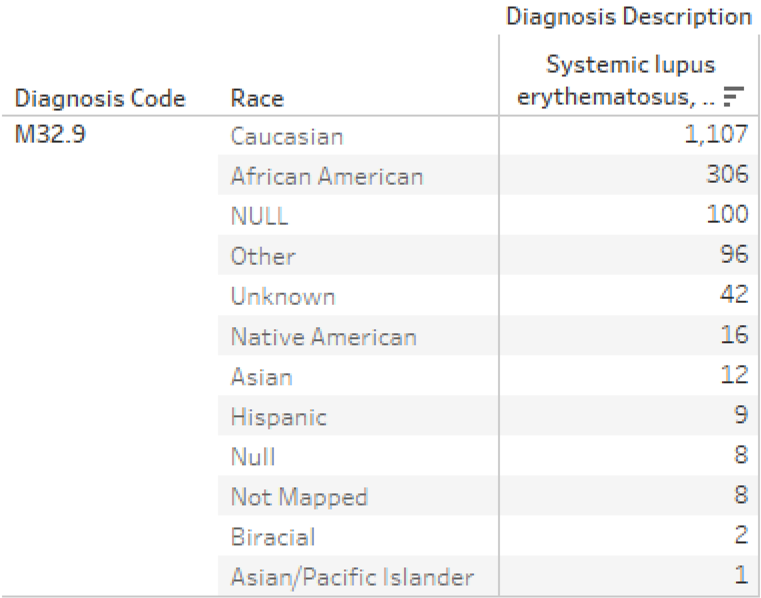


		Diagnosis Description
Diagnosis Coc	Gender	Systemic lupus erythematosus, uns..
M32.9	Female	1,314
	Male	370
	Not Mapped	16
	NULL	3
	Unknown/Invalid	4

Display 2. Gender

In the above display you can see that Females make up 76.98% of the population whereas they only make up 64.89% of the overall population for the entire dataset.

Display 3 explores race.

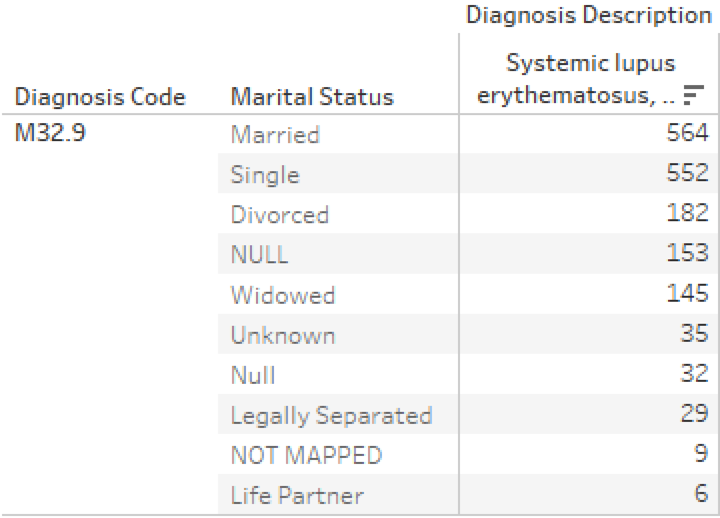


		Diagnosis Description
Diagnosis Code	Race	Systemic lupus erythematosus, ..
M32.9	Caucasian	1,107
	African American	306
	NULL	100
	Other	96
	Unknown	42
	Native American	16
	Asian	12
	Hispanic	9
	Null	8
	Not Mapped	8
	Biracial	2
	Asian/Pacific Islander	1

Display 3. Race

In the above display, 64.85% of the affected are Caucasian while 55.48% of the entire dataset's population is Caucasian.

Display 4 explores marital status.

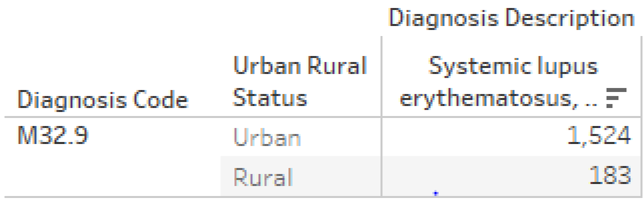


		Diagnosis Description
Diagnosis Code	Marital Status	Systemic lupus erythematosus, ..
M32.9	Married	564
	Single	552
	Divorced	182
	NULL	153
	Widowed	145
	Unknown	35
	Null	32
	Legally Separated	29
	NOT MAPPED	9
	Life Partner	6

Display 4. Marital Status

The display above shows that 33.04% of individuals diagnosed with Lupus are married. In contrast, only 30.17% of individuals are married overall.

Display 5 explores location.

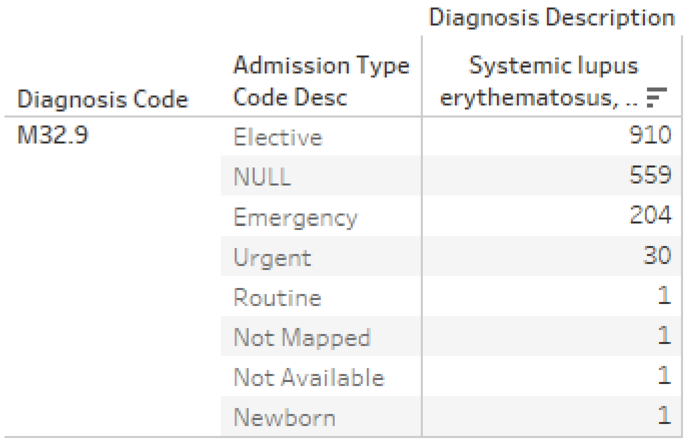


		Diagnosis Description
Diagnosis Code	Urban Rural Status	Systemic lupus erythematosus, ..
M32.9	Urban	1,524
	Rural	183

Display 5. Urban vs. Rural

The above display shows an impressive 89.28% urban population. However, as the overall percentage is 87.31%, this does not appear to be a large difference.

Display 6 explores admission type.



		Diagnosis Description
Diagnosis Code	Admission Type Code Desc	Systemic lupus erythematosus, ..
M32.9	Elective	910
	NULL	559
	Emergency	204
	Urgent	30
	Routine	1
	Not Mapped	1
	Not Available	1
	Newborn	1

Display 6. Admission Type

Unfortunately, this statistic has a lot of null values that throw off the final result. However, the ratio of elective to null to emergency in this smaller group is nearly identical to the overall percentages which indicates a lack of straying from the norm. What we know is that elective procedures top both charts.

SUBHEAD A LEVEL <SUMMARY OF GRAPHICS>

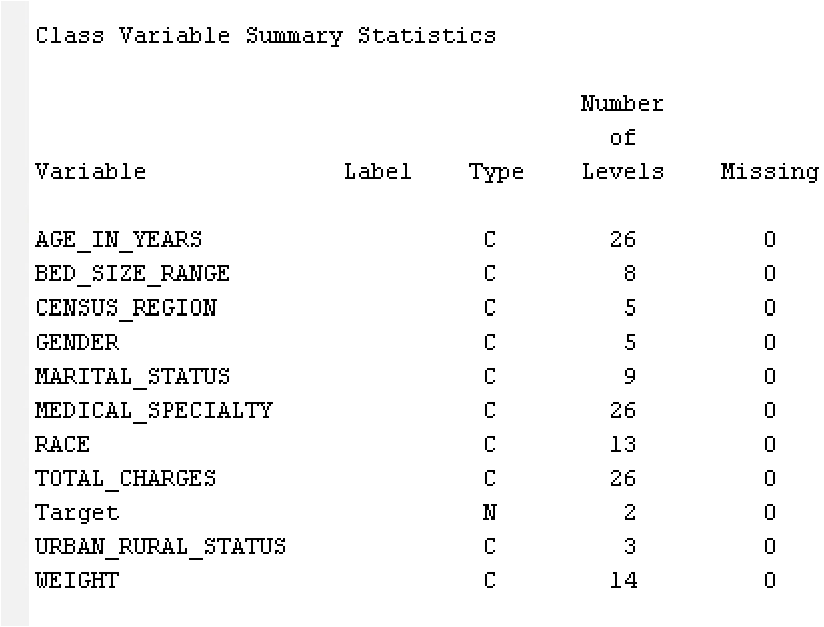
What this shows is interesting. For every demographic that was analyzed the smaller set of Lupus positive patients had a higher percentage of the most common demographic, such as female for gender, than the overall population. However, using the code below in R it was discovered that the difference between the two populations is not significant.

```
prop.test (c(yes_sample1, yes_sample2),c(total_sample1,
total_sample2),alternative="less")
```

THIRD MAIN TOPIC <PREDICTIVE MODEL CREATION>

From this point, I pulled the two datasets with reduced numbers of columns into SAS Enterprise Miner (SAS EM) and began another exploration of the data. However, only one of the datasets resulted in a usable output. This dataset had 17 variables, 1,000,000 rows and the following variables proved to be important to the final model.

Display 7 summary statistics for the variables of importance in the created models.

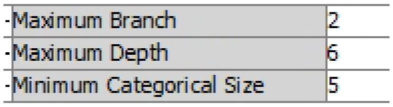


```
Class Variable Summary Statistics

                                         Number
                                           of
Variable              Label     Type     Levels    Missing

AGE_IN_YEARS                     C         26          0
BED_SIZE_RANGE                   C          8          0
CENSUS_REGION                    C          5          0
GENDER                           C          5          0
MARITAL_STATUS                   C          9          0
MEDICAL_SPECIALTY                C         26          0
RACE                             C         13          0
TOTAL_CHARGES                    C         26          0
Target                           N          2          0
URBAN_RURAL_STATUS               C          3          0
WEIGHT                           C         14          0
```

Display 7. Model Variables

Using nodes within SAS EM, I split this dataset into two parts, 70% training and 30% validation. After this I began to run a variety of models as no further data cleaning was necessary. However, only the following four models returned a usable result.

1. Random Forest
    a. I set this node's maximum number of trees to 100 and its significance level to 0.05.
    b. This model resulted in a misclassification rate of 1%.
2. DMINE Regression
    a. I set this node's maximum variable number to 3000 and minimum R-Square value to 0.0005.
    b. This model resulted in a misclassification rate of 0.753%.
3. Decision Tree
    a. The settings for this node can be found below.



Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5

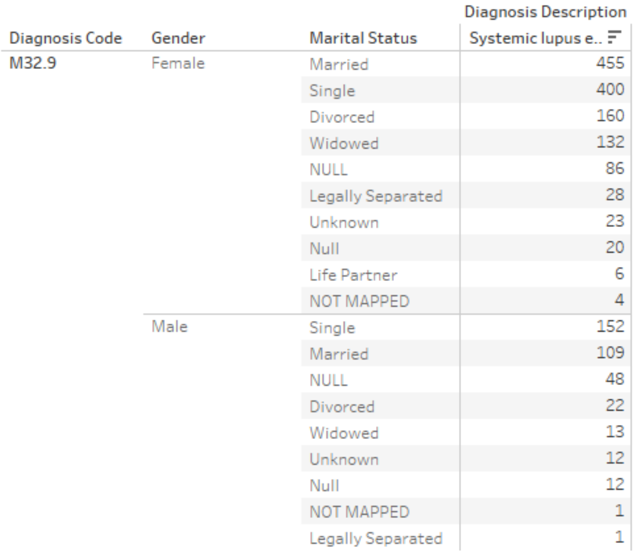
    b. This model resulted in a misclassification rate of 0.731%, which is the best that was achieved.
4. Ensemble
    a. This model is designed to take bits and pieces of previously run models to create the best possible result.
    b. This model resulted in a misclassification rate of 0.753%.

Now, each and every one of these misclassification rates sound wonderful and are well within the acceptable range but, what the misclassification rate does not show is the ratio of positive cases to negative. In fact, this ratio is so small the Ensemble model simply predicted that every single case would be a negative result. Thus, in order to truly test these models a dataset with a greater percentage of positive cases would be necessary.

CONCLUSION

The goal of this paper was to analyze factors that may have an impact on the development of one or more autoimmune diseases and see if anything points towards a potential cure. Throughout the exploration process many factors were analyzed and much was learned but, at this time no usable predictive model was found due to the small percentage of positive cases in the dataset. However, that does not mean the research was not a success. The above findings had me asking some questions, so I created the charts found below. What I found was that, in this dataset, married, white women from the Northeast census region have a higher likelihood to be diagnosed than any other demographic. But, as the numbers were found to not be significant, that can not be applied to other datasets. So, what can we learn from it?

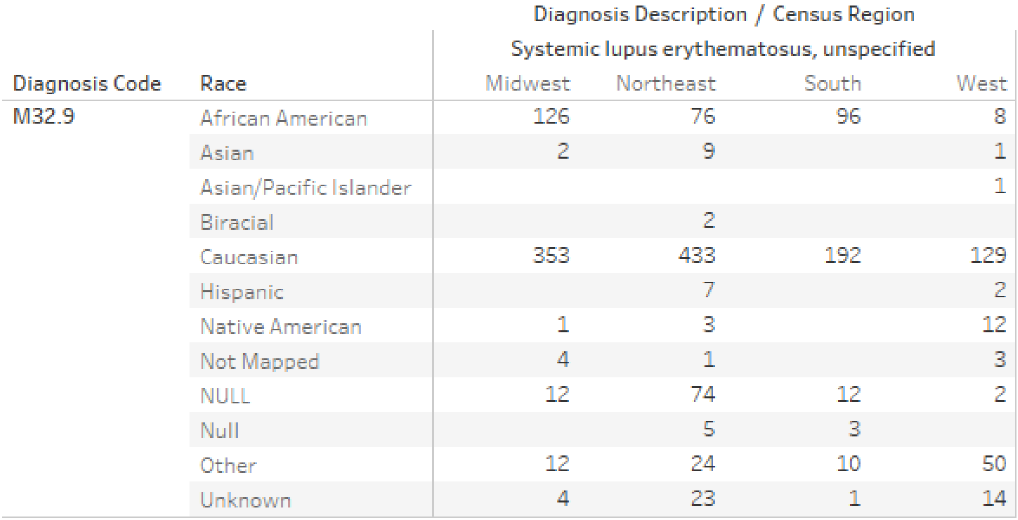
Display 8 is one of two charts that supports the statements made above.



			Diagnosis Description
Diagnosis Code	Gender	Marital Status	Systemic lupus e..
M32.9	Female	Married	455
		Single	400
		Divorced	160
		Widowed	132
		NULL	86
		Legally Separated	28
		Unknown	23
		Null	20
		Life Partner	6
		NOT MAPPED	4
	Male	Single	152
		Married	109
		NULL	48
		Divorced	22
		Widowed	13
		Unknown	12
		Null	12
		NOT MAPPED	1
		Legally Separated	1

Display 8. Gender v Marital Status

Display 9 is the second chart that supports the above statements.



		Diagnosis Description / Census Region			
		Systemic lupus erythematosus, unspecified			
Diagnosis Code	Race	Midwest	Northeast	South	West
M32.9	African American	126	76	96	8
	Asian	2	9		1
	Asian/Pacific Islander				1
	Biracial		2		
	Caucasian	353	433	192	129
	Hispanic		7		2
	Native American	1	3		12
	Not Mapped	4	1		3
	NULL	12	74	12	2
	Null		5	3	
	Other	12	24	10	50
	Unknown	4	23	1	14

Display 9. Race v Census Region

Unfortunately, this does line up with other studies occurring simultaneously. According to those studies, while women are more susceptible, it is typically African Americans, Hispanics, and Asian-Americans that are at risk for Lupus (Mayo). So, why does our dataset say something different? One factor is that this dataset has a higher ratio of Caucasian records than any other race. Another theory could be that perhaps Caucasians are more frequently diagnosed than other races. However, if that were the case, how do the studies determine the races that are at great risk? Ultimately though, I believe the primary issue is sample size. The dataset used for this analysis had a ratio of Lupus positive cases that was not significant in relation to the entire dataset. Now, while this could be resolved with additional cases to study, I would much prefer to leave the cause of Lupus unknown if the alternative is more families having to endure the implications that come from a positive diagnosis. So, acknowledging what we have learned and what studies currently exist, next I will detail what I believe are the next steps in Lupus research.

RECOMMENDATIONS

Something that has been proven is that in some way, shape or form estrogen plays a role in the increased risk for females versus males of developing Lupus (ANRF). This combined with the fact that there are soluble molecules that have been identified to also play a role in the development of Lupus leads me to my conclusion (ANRF). Present day researchers are very close. We as a society are close to not only being able to increase prevention of the disease, but also to stop its progression and/or consistently control its symptoms (ANRF). So, I think the perfect partner to this ongoing research would be to look for the unreported and unconfirmed cases. We know how the symptoms most often present and thanks to current studies we know what genetic markers to look for that can hint at an individual being predisposed. Using this, we should look into hospital records for patients that have no diagnosis. Up to 50% of individuals have non-life-threatening symptoms such as severe fatigue, joint pain, and rash (LFA). Each of these individuals is currently at risk for an incorrect or non-existent diagnosis where a proper one could result in a much better quality of life. This is where Lupus research should head next.

REFERENCES

Lupus Foundation of America (LFA). What is lupus? (2013, July 31). Retrieved from https://www.lupus.org/resources/what-is-lupus#

Mayo Clinic Staff. (2017, October 25). Lupus. Retrieved from https://www.mayoclinic.org/diseases-conditions/lupus/symptoms-causes/syc-20365789

Arthritis National Research Foundation (ANRF). (2019, March 26). Lupus Research: Causes of Lupus. Retrieved from https://curearthritis.org/lupus-research/

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

<Brittany Hickerson>
<brhicke@okstate.edu>