

Long-Term Remission Off Therapy for Anti-Neutrophilic Cytoplasmic Autoantibody Vasculitis

Keliang Gao

Mentors: Professor James Stephen Marron

Nov 12, 2019

Department of Statistics and Operations Research

School of Science and Art

UNC at Chapel Hill

All knowledge is, in final analysis, history.

All sciences are, in the abstract, mathematics.

All judgements are, in their rationale, statistics.

—— Calyampudi Radhakrishna Rao

Contents

Abstract.....	1
1. Introduction.....	1
1.1 Information of disease.....	1
1.2 Remission and active states.....	2
1.3 Introduction of data.....	2
1.3.1 Response variable of interest	2
1.3.2 Introduction of predictors.....	2
2. Data visualization and exploration.....	4
2.1 Biomarkers.....	4
2.1.1 Summary	4
2.1.2 Calprotectin.....	5
2.1.3 Anti-plg-Antibodies;.....	6
2.1.4 PR3THP	6
2.1.5 MPOHL.....	7
2.1.6 RUNX3	8
2.1.7 KIV	8
2.1.8 pr3	9
2.1.9 mpo	10
2.2 Data adjustment	10
2.2.1 Outliers drop-off	10
2.2.2 Data transformation.....	11
3. Comparison of biomarkers from relapse patients group and that from non-relapse group.....	13
3.1 Box Plots of biomarkers of two groups	13
3.2 T-Test.....	18
4. Logistic regression	24
4.1 Response: relapse or not	24
4.2 Predictors	24
4.3 Model developing	24
5. Conclusions and Discussion for Future Work	27
References.....	28
Acknowledgement	31

Abstract

Anti-Neutrophilic Cytoplasmic Autoantibody vasculitis is an autoimmune disease. The vast majority of patients will die within a year if untreated.

In this study, the author investigated data from 412 observations with 10 biomarkers (features) to demonstrate the gap between relapse patients and non-relapse patients as potential candidates in clinical application.

Two biomarkers are dropped from raw dataset due to high extent missing value (missing data > 97%). Among the leftover eight biomarkers, analysis results show that Antipltg_Value ($p=0.002$), PR3THP ($p=0.001$), pr3 ($p=0.012$), and mpo ($p=0.026$) are of significant different between relapse patients group and non-relapse group. Biomarkers of MPOHL, RUNX3 and KIV_Value are of insignificant between the two groups. The biomarker of Calprotectin might be a potential significant predictor while more data is required for validation.

Optimal logistic regression model is built at the end of the report, and can be applied for prediction of relapse.

Keywords: Anti-Neutrophilic Cytoplasmic Autoantibody vasculitis, biomarker, statistics

1. Introduction

1.1 Information of disease

Anti-Neutrophilic Cytoplasmic Autoantibody (ANCA) vasculitis is an autoimmune disease (Jennette et al., 1989). It is an inflammatory disease of small to medium-sized vessels that frequently presents with rapidly progressive glomerulonephritis and renal failure though it can affect any organ system. If untreated, the vast majority of patients will die within a year (Lazarus et al., 2016).

Collectively, these disorders account for a considerable burden of death and disability worldwide and are of great clinical importance owing to the vast improvement in prognosis with treatment (Savage, 2011). As a result, Anti-Neutrophilic Cytoplasmic Autoantibody vasculitis (AAV) has been studied extensively in recent years (Bossuyt, 2017).

There are 2 main kinds of autoantibodies that can be involved in ANCA vasculitis. One is perinuclear ANCA (P-ANCA). This P-ANCA type of autoantibody usually targets and attaches to myeloperoxidase (MPO) inside of neutrophils. The other one is called cytoplasmic ANCA (C-ANCA). This C-ANCA (a different autoantibody) usually targets and attaches to proteinase 3 (PR3), which is also inside of neutrophils (Jennette et al., 2014).

Both MPO and PR3 data are collected and studied in this research.

Research shows that almost half of tuberculosis patients suffered from AAV. Globally in 2014, there were an estimated 9.6 million incident cases of TB: 5.4 million among men, 3.2 million among women and 1.0 million among children. Globally in 2014, there were an estimated 1.5 million deaths from TB: 1.1 million deaths among people who were HIV negative and 390 000 deaths among people who were HIV positive (WHO report, 2015).

Berti et al. reported a prevalence of 42.1 per 100,000 based on the 20-years research in Minnesota USA from 1996 to 2015 (Berti et al., 2017).

1.2 Remission and active states

Patients with ANCA vasculitis can cycle between remission and active disease states (Land et al., 2014). During remission, disease is quiescent and patients exhibit almost no symptoms of disease. However, during active disease (sometimes called a ‘flare’ or ‘relapse’), the patient can have a multitude of disease symptoms that impact any organ system (cutaneous, ENT, chest, renal, cardiovascular, etc.). “True remission” is defined as complete absence of clinical signs and symptoms of AAV for at least 2 years without the use of any immunomodulatory medications, or “Long-Term Remission Off Therapy”.

The current candidate markers are selected based upon their pathogenetic role and association with active ANCA vasculitis. The selected biomarkers for this study are: RNA expression (PR3THP, MPOHL, RUNX3), ANCA Titers (pr3, mpo), complement (C5a, C3a), Anti-plasminogen (Anti-Plg), Calprotectin, KIV, et.al. Detailed information about these biomarkers are introduced in section 1.3.2 (Table 1).

1.3 Introduction of data

Data is detected and collected by Susan Hogan et al from Kidney Center in UNC at Chapel Hill, USA.

1.3.1 Response variable of interest

The response variable of interest is relapse. It is a binary variable with values of Yes or No. Expression level of biomarkers are adopted as response for T-test analysis.

1.3.2 Introduction of predictors

Table 1 shows the definition and explanation of each and every variables in the raw dataset.

Table 1. Information of dataset

Variable	Explanation
PID	Patient Identification (PID, unique to each patient).
SID	Sample Identification (SID, unique to each date a sample was collected for that patient). There are no duplicated SIDs.
Sample Date	The date sample was collected.
LTROT Date	The date the patient entered Long Term Remission Off Therapy.
Last Relapse Date	The date the patient relapsed or had a change in disease state. For many patients this column is blank. However, there are 15 patients who relapsed and had to receive treatment for their disease.
Last FU Date	The last date the patient had chart review performed by Dr. Falk.
CD5	CD5+ B cells, decreased in active disease
Calprotectin	Good marker for stable remission if low or negative.
Anti-Plg Antibodies	Indicator associated with active disease and increased risk (binary).
Anti-plg value	Anti-plasminogen, associated with active disease and increased risk
PR3THP	biomarker
MPOHL	biomarker
RUNX3	biomarker of RNA expression, downregulated in disease
C5a_Value	Complement pathway activation mediators, upregulated in active disease
C3a_Value	Complement pathway activation mediators, upregulated in active disease
KIV-Value	Factors correlated with MPO ANCA disease
pr3	Biomarker of RNA expression, upregulated in active disease
mpo	Biomarker of RNA expression, upregulated in active disease
Duration of LTROT	Duration of LTROT (number of months between each LTROT date and the Sample Date)
Relapse	Relapse Yes or No
Time to Relapse	Time to Relapse (number of months between each LTROT date and the Relapse Date)

2. Data visualization and exploration

2.1 Biomarkers

2.1.1 Summary

The explanatory variables of biomarkers in the dataset are listed in Table 2.

There are 412 observations in the dataset in total. Table 2 shows the statistics of the predictors (mean, SD, min, and max) and the number of missing values. Apparently there are a lot of missing data.

Table 2. Summary of variables

Variable	Number of missing values	N	Mean	Std Dev	Minimum	Maximum
Calprotectin	370	42	1.81	2.22	0.52	14.61
Antiplg_Value	289	123	21.37	12.82	0.00	65.32
PR3THP	263	149	38.36	75.74	0.24	709.38
MPOHL	263	149	131.13	147.61	4.88	959.74
RUNX3	340	72	23.16	7.66	10.09	47.12
C5a_Value	403	9	8.34	2.98	4.46	14.37
C3a_Value	403	9	61.91	22.96	40.00	114.99
KIV_Value	319	93	0.14	0.20	-0.64	0.68
pr3	42	370	15.51	36.41	0.00	200.00
mpo	39	373	21.70	27.70	0.14	142.10

The table shows that there are some variables showing relatively large variation, including PR3THP and pr3. Also there are lots of missing data as shown in Table 2. The missing data is a major problem and common issue in the analysis of clinical research which is caused by patients dropping out of the research before completion. This also include some conditions such as the patients pass away or moved to other city or state. In this study the raw dataset contains 412 observations, while all below biomarkers come with missing data.

The variables of pr3 and mpo both contain around 370 data points; however, most other variables show large amount of missing value. Both complement pathway activation mediators, C5a_Value and C3a_Value, only acquired 9 data points as shown in Table 2.

The extreme small sample size makes it impossible to accomplish data analysis to obtain valuable achievement. For example, the biomarkers of C5a_Value and C3a_Value only show 9 datapoints in the dataset with 403 missing value. It is hard to achieve valuable and reliable statistical analysis results based on those samples. The chance to get accurate and unbiased results is very low, therefore these two predictors are removed and not included in following analysis.

2.1.2 Calprotectin

Distribution of biomarker Calprotectin is shown in Figure 1.

As introduced in section 1.3.2, the variable Calprotectin is a marker for stable remission if it is low or negative. There are 42 data points for Calprotectin among 412 observations.

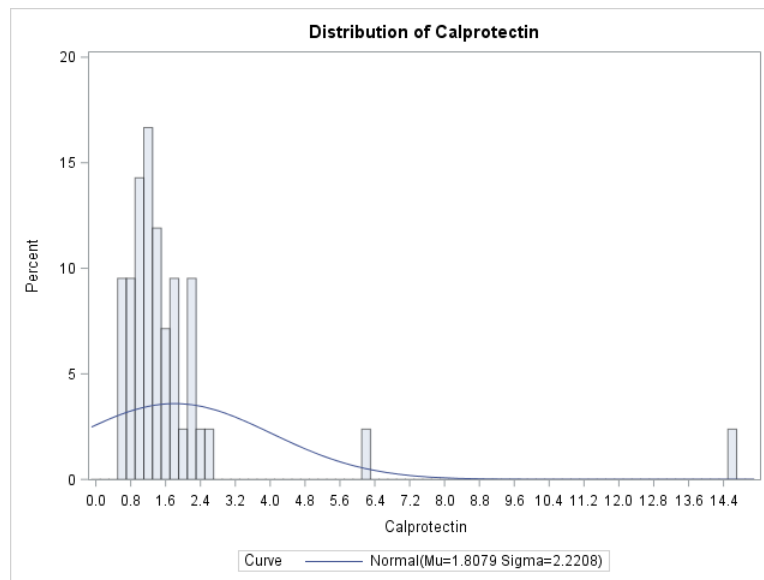


Figure 1. Distribution of biomarker Calprotectin

Figure 1 shows that most values are in the range of 0.4 to 2.8, while there are two data points showing large value. One is 6.27 and the other is 14.61.

These two data points are unreasonably large and are potential outliers or mis-recording. Going back to raw data to double check these values, the point of 6.27 comes from sample ID 5798 of patient ID 23789. There are 12 observations from the identical patient, and there are 5 records of Calprotectin. All other 4 records are between 0.96 and 1.40 from Mar 2010 through Mar 2017, while the data of 6.27 is recorded as Feb

19, 2009. Since it is far away from other data points of the identical patient, I would suggest to remove this point.

As for the point of 14.61, all other records from the same patient vary between 0.94 and 1.24. It is not sure whether the number 14.61 is mis-recorded (e.g, 1.461 instead of 14.61). The huge number was recorded on Jun 18, 2009 and the value on Jul 22, 2010 is 0.94. It seems that the value of 14.61 is not reasonable and very likely a misrecord or mismeasurement, therefore this point is removed.

The distribution of predictor Calprotectin with adjusted data is shown in section 2.2 (see Figure 9).

2.1.3 Anti-plg-Antibodies;

Distribution of biomarker Anti-plg-Antibodies is shown in Figure 2.

As introduced in section 1.3.2, the variable Anti-plg-Antibodies is a marker associated with active disease and increased risk. There are 123 data points for Anti-plg-Antibodies among 412 observations.

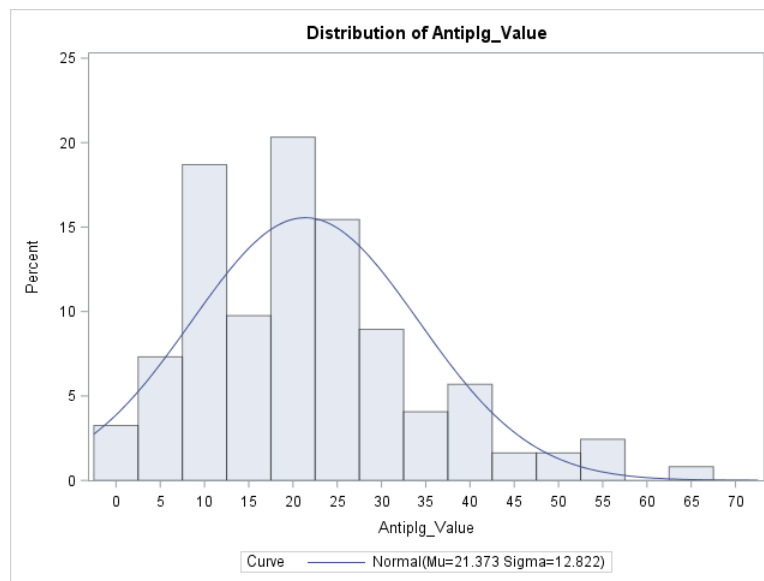


Figure 2. Distribution of biomarker Anti-plg-Antibodies

Figure 2 shows that the distribution of biomarker Anti-plg-Antibodies is close to normal distribution, though it is apparently not a nice normal distribution.

2.1.4 PR3THP

Distribution of biomarker PR3THP is shown in Figure 3.

As shown in Table 2, there are 149 data points for PR3THP among 412 observations.

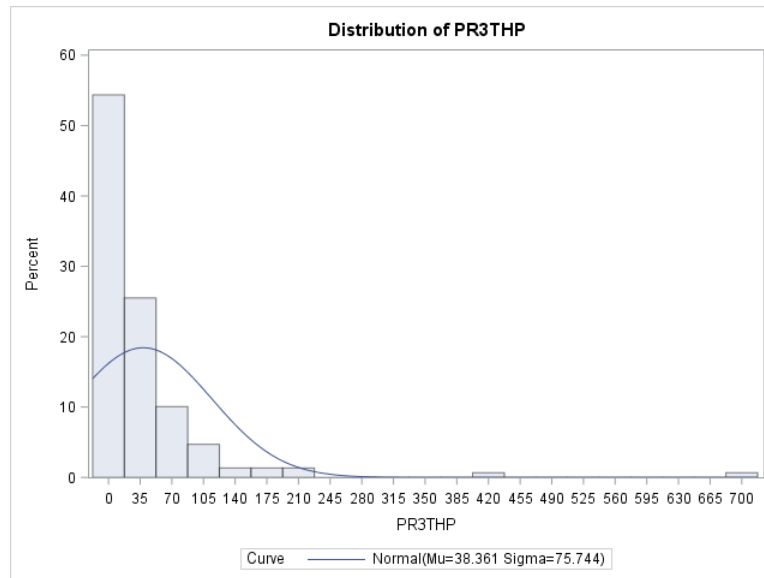


Figure 3. Distribution of biomarker PR3THP

Obviously this is not a normal distribution. It looks more like poisson distribution rather than normal. Here logarithm transformation was applied and the updated figure is shown in section 2.2.

2.1.5 MPOHL

Distribution of biomarker MPOHL is shown in Figure 4.

As shown in Table 2, there are 149 data points for MPOHL among 412 observations.

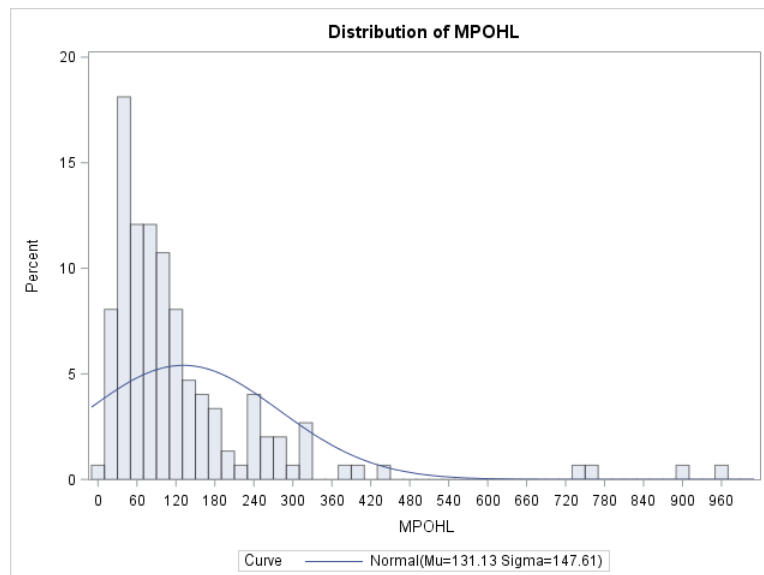


Figure 4. Distribution of biomarker MPOHL

This seems to be right skewed (positive skewed). Therefore logarithm transformation was applied and the updated figure is shown in section 2.2.

2.1.6 RUNX3

Distribution of biomarker RUNX3 is shown in Figure 5.

As introduced in section 1.3.2, the variable RUNX3 is a marker of RNA expression. The level of RUNX3 is lower in patient with active disease compared to those healthy people.

As shown in Table 2, there are 72 data points for RUNX3 among 412 observations.

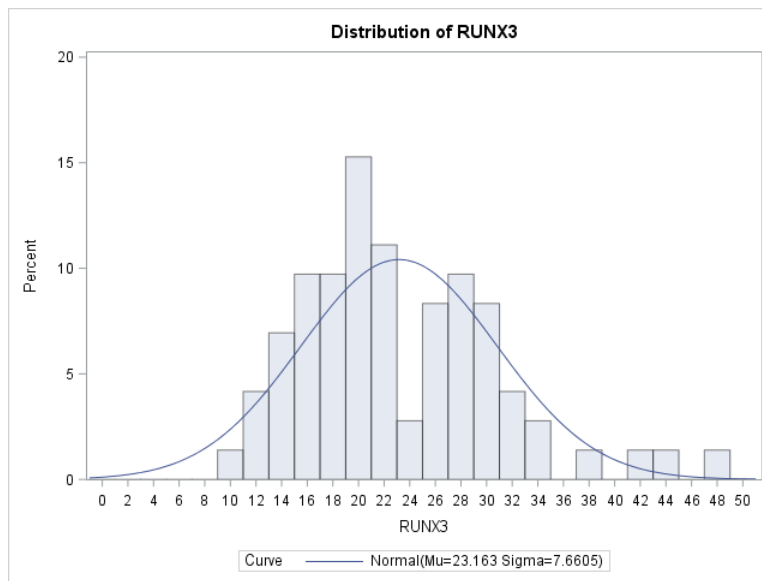


Figure 5. Distribution of biomarker RUNX3

This seems to be similar to normal distribution, even though the histogram is not a perfect normal distribution.

2.1.7 KIV

Distribution of biomarker KIV is shown in Figure 6.

As introduced in section 1.3.2, the variable KIV is a marker correlated with MPO ANCA disease.

As shown in Table 2, there are 93 data points for KIV among 412 observations.

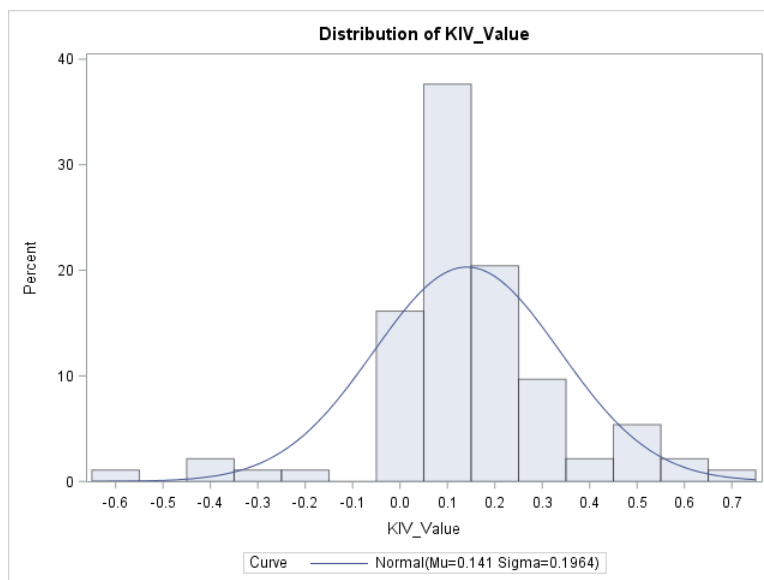


Figure 6. Distribution of biomarker KIV

Figure 6 shows the histogram of biomarker KIV and it is close to normal distribution.

2.1.8 pr3

Distribution of biomarker pr3 is shown in Figure 7.

As introduced in section 1.3.2, the variable pr3 is a marker of RNA expression. The level of pr3 is higher in patient with active disease compared to those healthy people.

As shown in Table 2, there are 370 data points for pr3 among 412 observations.

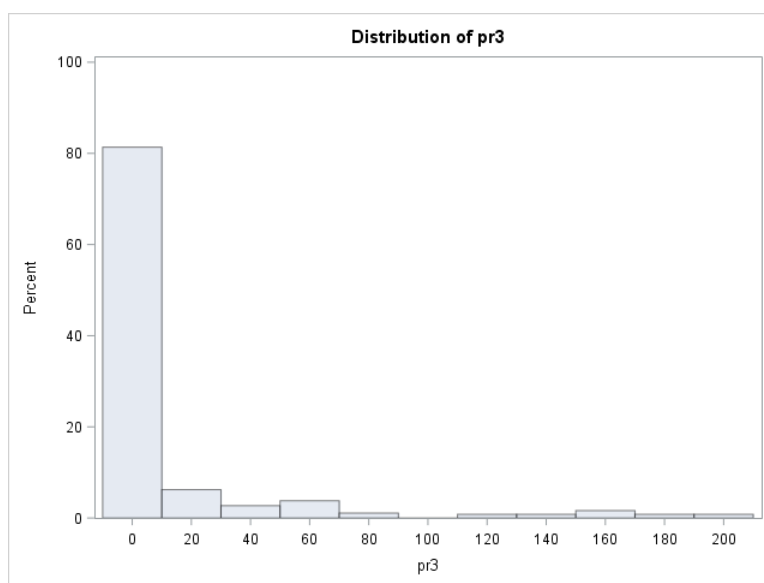


Figure 7. Distribution of biomarker KIV

Figure 7 shows the histogram of biomarker pr3 which is not normal distribution at all. Therefore logarithm transformation was applied and the updated figure is shown in section 2.2. Note that there are lots of zero in the measurement and I added a positive number to each measurement before logarithm transformation to make it valid.

2.1.9 mpo

Distribution of biomarker mpo is shown in Figure 8.

As introduced in section 1.3.2, the variable mpo is a marker of message RNA expression. The level of pr3 is higher in patient with active disease compared to those healthy people.

As shown in Table 2, there are 373 data points for mpo among 412 observations.

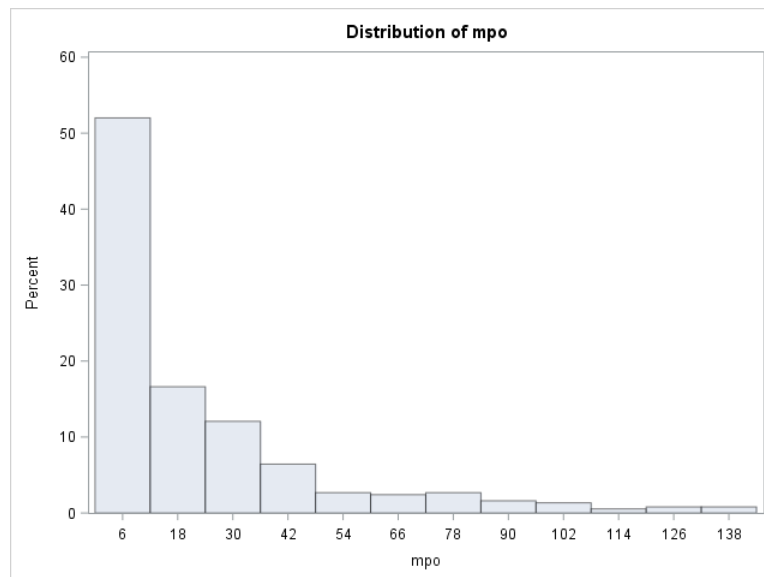


Figure 8. Distribution of biomarker mpo

Figure 8 shows the histogram of biomarker mpo. It looks more like poisson distribution more like normal. Therefore logarithm transformation was applied and the updated figure is shown in section 2.2.

2.2 Data adjustment

2.2.1 Outliers drop-off

As discussed in section 2.1.2, there are two possible outliers in data points for Calprotectin. One is 6.27 while other data from identical patient ranges from 0.96 to 1.40, and the other point is 14.61 while other data from identical patient ranges from 0.94 to

1.24. These two outliers are dropped off and following analysis are based on adjusted dataset.

2.2.2 Data transformation

One of the most common assumptions for statistical analyses is that of normality, with nearly all parametric analyses requiring this assumption in one way or another (Manikandan, 2010). While not all normality assumptions pertain directly to an individual variable's distribution (i.e., the assumption of normality for a regression is that the regression's error is normally distributed, not that all variables in the analysis are normal), it is often easier to meet the assumption if each variable in the analysis *is* normally distributed. We can make that happen by transforming (Stevens, 2009).

Often one of the first steps in assessing normality is to review a histogram of the variable in question.

The log transformation, a widely used method to address skewed data, is one of the most popular transformations used in biomedical and psychosocial research (Templeton, 2011). Due to its ease of use and popularity, the log transformation is included in most major statistical software packages including SAS, Splus and SPSS (Changyong FENG, 2014). Below figures show the distribution of adjusted data and transformed data.

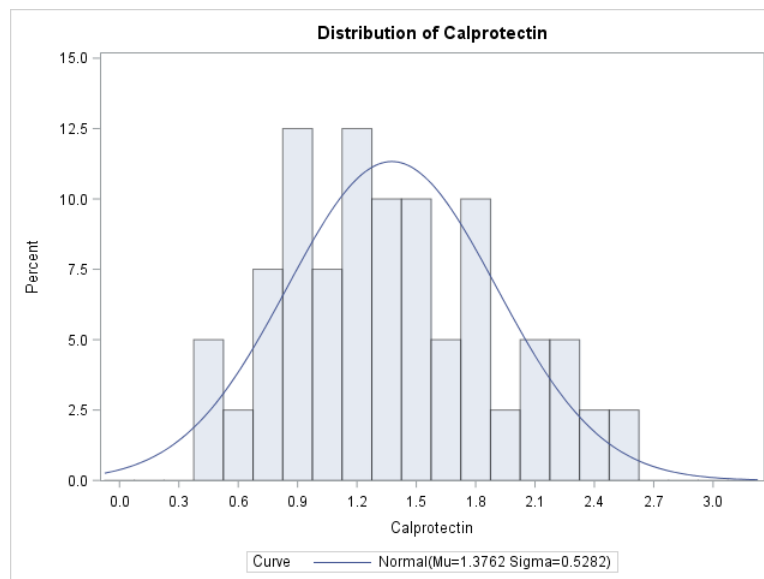


Figure 9. Distribution of biomarker Calprotectin from updated dataset

Figure 9 shows distribution of biomarker Calprotectin with updated dataset in which the two outliers of 6.27 and 14.61 were both removed.

After dropping these two outliers, the data show a close-to-normal distribution.

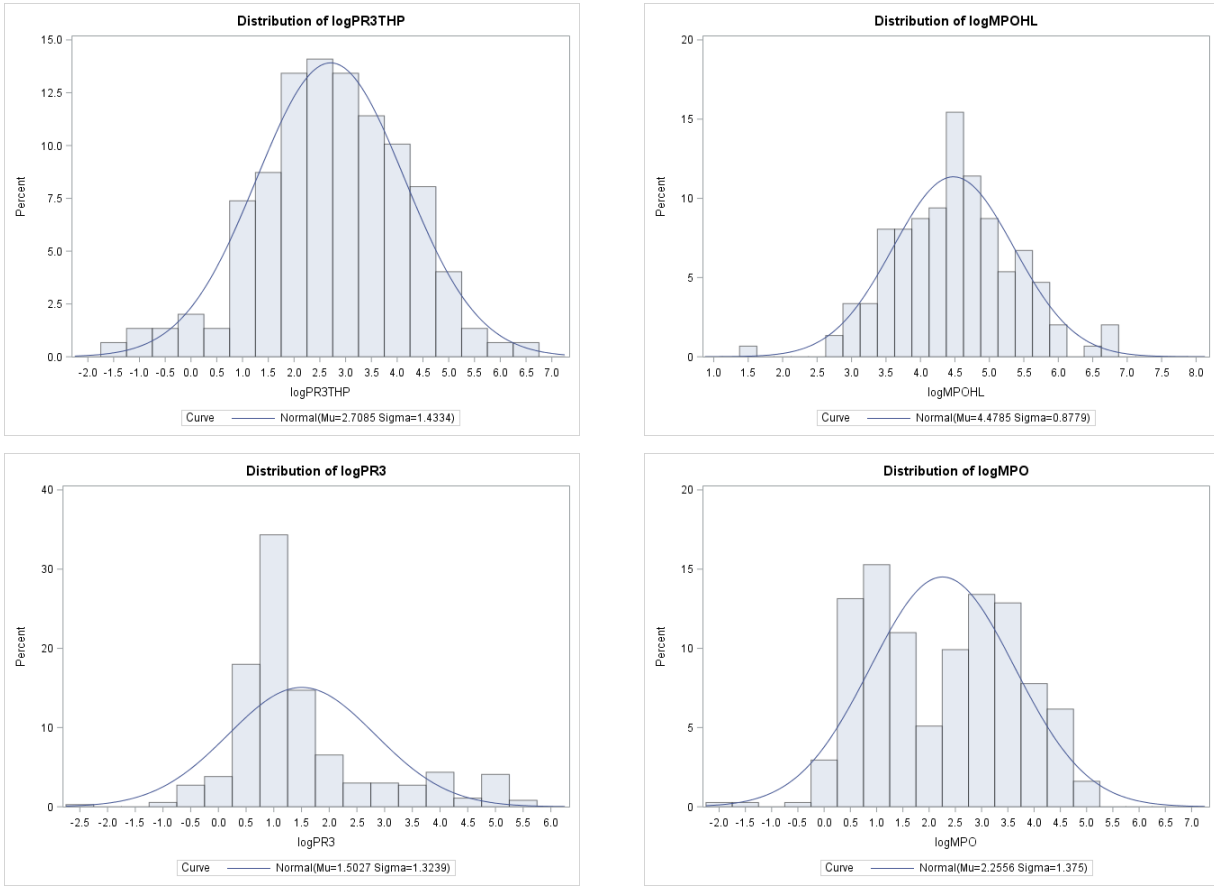


Figure 10. Distribution of transformed biomarkers

Figure 10 shows the distribution of logarithm transformed biomarkers. The results demonstrated that PR3TRP, MPOHL, and pr3 show distribution close to normal, while mpo does not show a satisfied distribution. In this case I would treat it as normal distribution, with further discussion in following paragraphs to investigate more about this biomarker.

While looking deep into the raw data, it is noticed that there are both patients with relapse and patients of non-relapse. There might be difference in expression level of one or more biomarkers, and this might be the possible reason why the distribution demonstrated dual-peak distribution if the difference between those two groups are significant large. Investigation of these difference is carried out and the results of all biomarkers are shown in section 3.1 and section 3.2.

3. Comparison of biomarkers from relapse patients group and that from non-relapse group

3.1 Box Plots of biomarkers of two groups

Before running T-test, ANOVA and setting up a model based on the dataset, results from relapse patients and that from non-relapse subjects are compared as well as validating with biological meaning.

Box plot is applied here to show the difference of distribution of two groups.

Like histograms, box plots show the distribution of continuous data. This type of graph is also called a box-and-whisker plot because of the way it looks. Every part of a box plot tells you something about the distribution of your data.

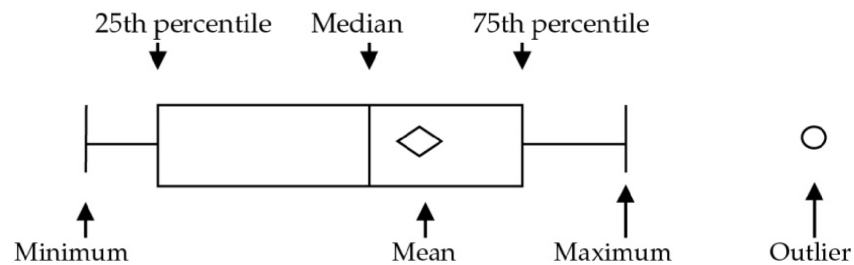


Figure 11. Sketch of box plots

The ends of the box indicate the 25th and 75th percentiles (also called the interquartile range). The line inside the box indicates the 50th percentile (the median), and the marker indicates the mean (see Figure 11). By default, the whiskers cannot be longer than 1.5 times the length of the box. Any points beyond the whiskers are considered outliers and are marked with circles (Delwiche 2012).

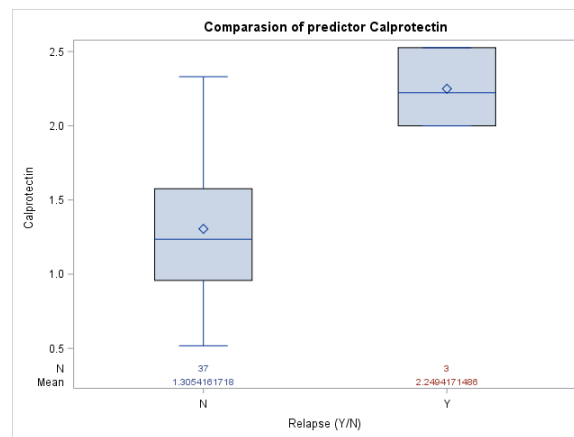


Figure 12. Comparison of biomarker Calprotectin

Figure 12 shows the comparison of biomarker Calprotectin in relapse group and that in non-relapse group. I observed some pattern from the figure that relapse patients show higher level of expression than that of non-relapse patients group.

However, the pattern of relapse group seems unusual. I double checked the raw data and found that there are only three data point of Calprotectin in relapse group. The total amount of observations with Calprotectin is 40, while 37 out of 40 is of non-relapse patient and the data is highly unbalanced. Among the overall 412 observations, there indeed are a lot of patients with relapse. However, most of the relapse patients did not take measurement for Calprotectin. Therefore, the current dataset does not qualify for comparison of Calprotectin in relapse group and non-relapse group. Generally speaking, those will not be included in a statistical report. However, this result is still demonstrated in this report because the analysis results show very strong difference even though there are only 3 samples in the relapse group. Surely this results are not confident or reliable, and it can't validate with only 3 samples in relapse group. However, the strong shift suggests that Calprotectin might be a potential candidate for relapse prediction. It may be of great clinical value. Therefore, if it is possible for the researchers to collect more data from patients of relapse for predictor Calprotectin, there is high chance that we can get solid conclusion about the difference of Calprotectin level between relapse patients and patients of non-relapse.

Next all other biomarkers are tested and the results are shown in Figure 13.

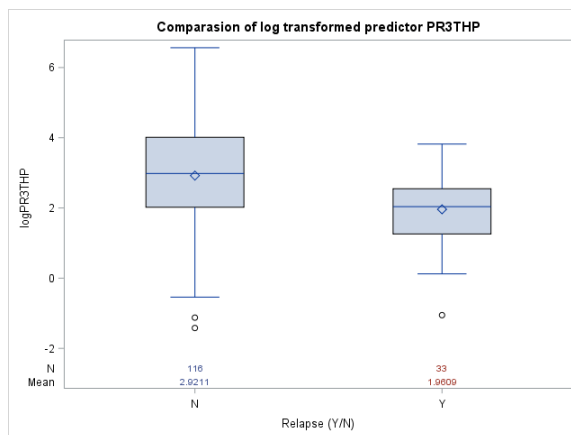
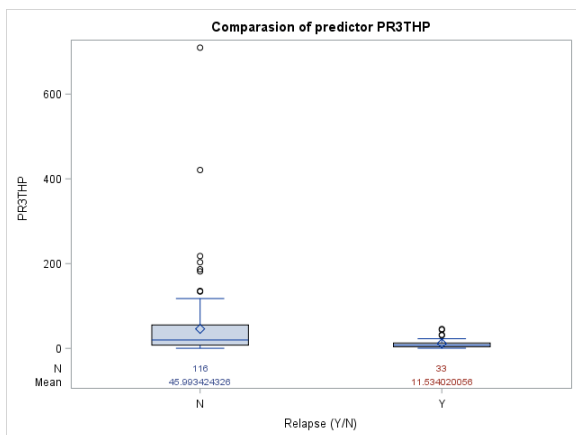
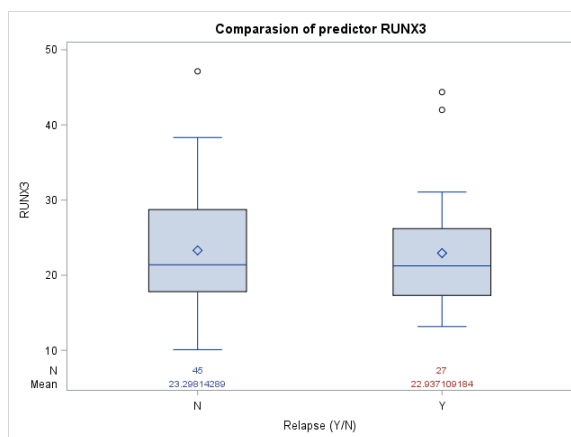
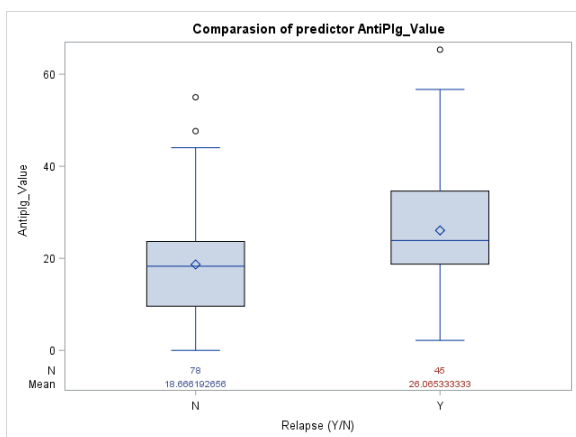
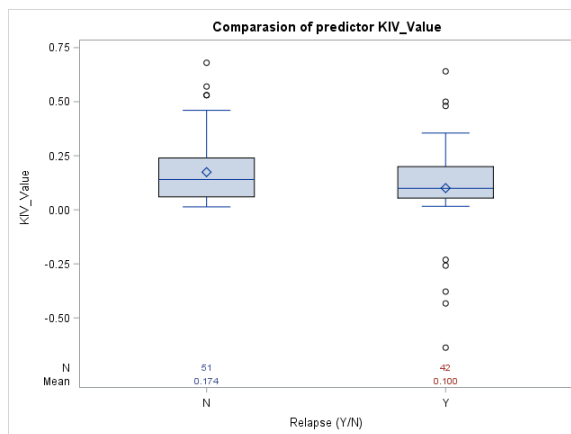
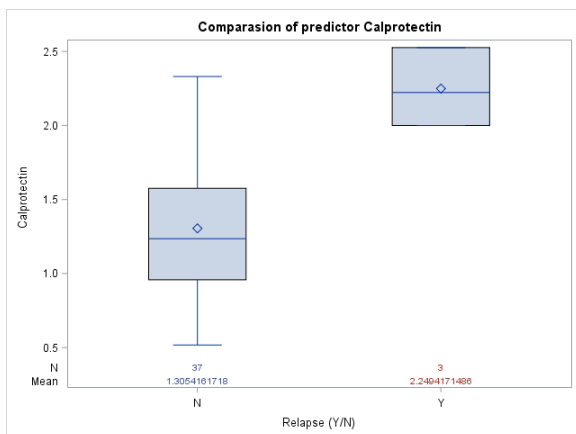
Table 3 below shows the number and mean value of each predictor variable in relapse group and non-relapse group separately.

Table 3. Summary of predictors within each group

Relapse (Y/N)	N Obs	Variable	N	Mean	Std Dev
N	326	Calprotectin	37	1.3054162	0.4792352
		Antiplg_Value	78	18.6661927	11.3579271
		PR3THP	116	45.9934243	84.1457249
		logPR3THP	116	2.9211172	1.4528305
		MPOHL	116	143.1557204	163.0153226
		logMPOHL	116	4.5242975	0.9381007
		RUNX3	45	23.2981429	7.7013286
		KIV_Value	51	0.1743480	0.1564870
		pr3	295	13.5104068	35.5070829
		logPR3	295	1.1341239	2.7493940
		mpo	299	20.8080602	28.0372468
		logMPO	299	2.1769096	1.3844874
Y	86	Calprotectin	3	2.2494171	0.2642659
		Antiplg_Value	45	26.0653333	13.9482141
		PR3THP	33	11.5340201	11.6260366
		logPR3THP	33	1.9609360	1.0849942
		MPOHL	33	88.8732653	53.8691038
		logMPOHL	33	4.3173884	0.6060119
		RUNX3	27	22.9371092	7.7326460
		KIV_Value	42	0.1004876	0.2316550
		pr3	75	23.3993333	39.0375837
		logPR3	75	1.9711439	1.5043614
		mpo	74	25.3152703	26.1863426
		logMPO	74	2.5737339	1.2969474

The results in Table 3 demonstrates the samples of each biomarkers are highly unbalanced between the relapse group and non-relapse group.

Below figures show the comparison of each biomarker between the two groups, along with number of samples and mean of each group.



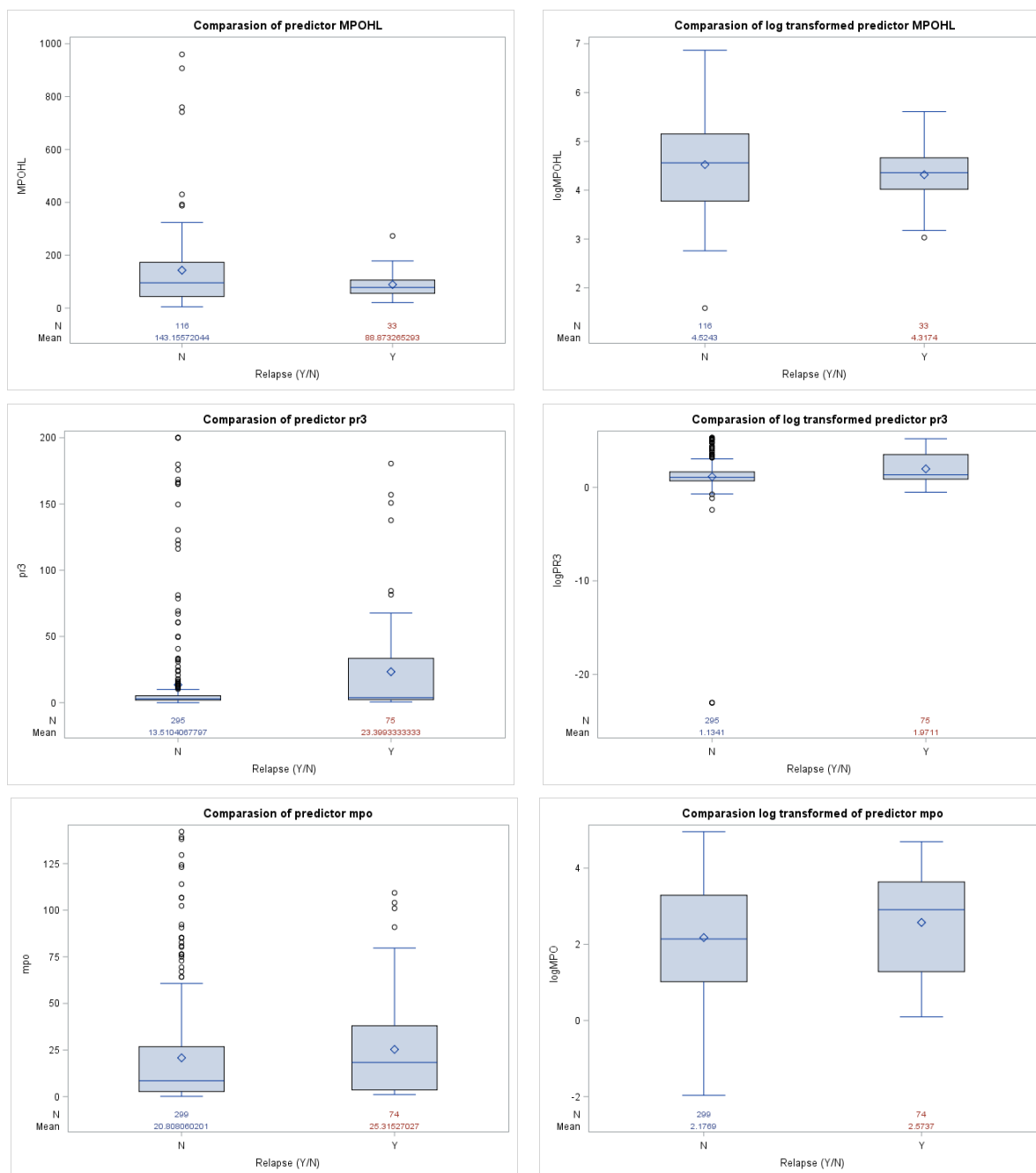


Figure 13. Comparison of biomarkers in relapse group and those in non-relapse group.

Figure 13 shows the comparison of biomarkers, including AntiPlg_Value, PR3THP, MPOHL, RUNX3, C5a_Value, C3a_Value, KIV_Value, pr3 and mpo, as well as the log transformation. Comparing to the raw data, the transformed data show better distribution. While in Figure 13 the mean value of most of the biomarkers seems critical between the relapse patients and non-relapse patients, T-test is carried out to check the statistical significance in following section.

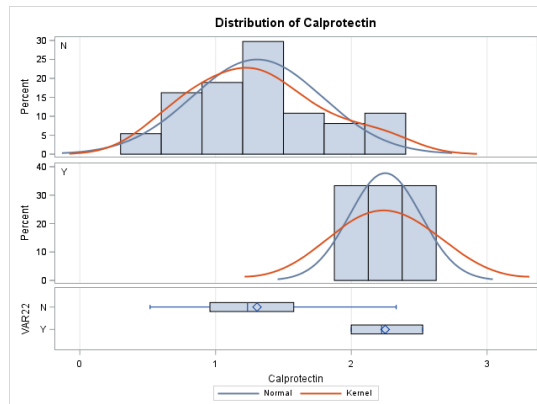
3.2 T-Test

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the probability of difference between two sets of data.

Essentially, a t-test allows us to compare the average values of the two data sets and determine if they came from the same population.

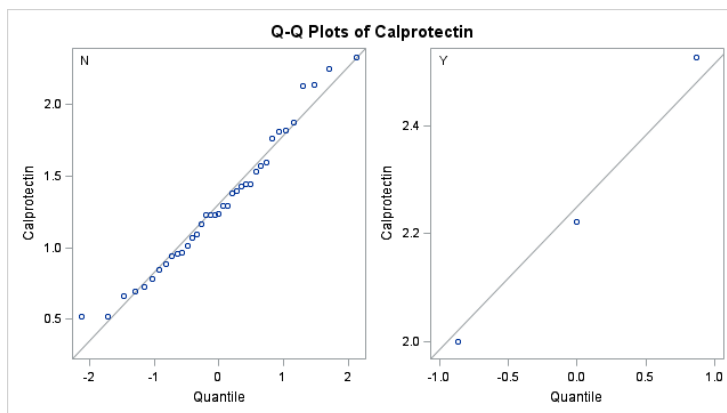
Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement by assuming a null hypothesis that the two means are equal. Based on the applicable formulas, certain values are calculated and compared against the standard values, and the assumed null hypothesis will be rejected (when p-value is smaller than 0.05) or unable to rejected (in this case when p-value is higher than the significant level of 0.05).

If the null hypothesis qualifies to be rejected, it indicates that data readings are strong and are not by chance (Howard J. Seltman, 2018).



T-test shows that the difference is significant, with p-value < 0.05.

Interpretation: this mean that the Calprotectin level in relapse patients is higher than in non-relapse patients.



RELAPSE	N	Mean	Std Dev
N	37	1.3054	0.4792
Y	3	2.2494	0.2643
Diff (1-2)		-0.9440	0.4704

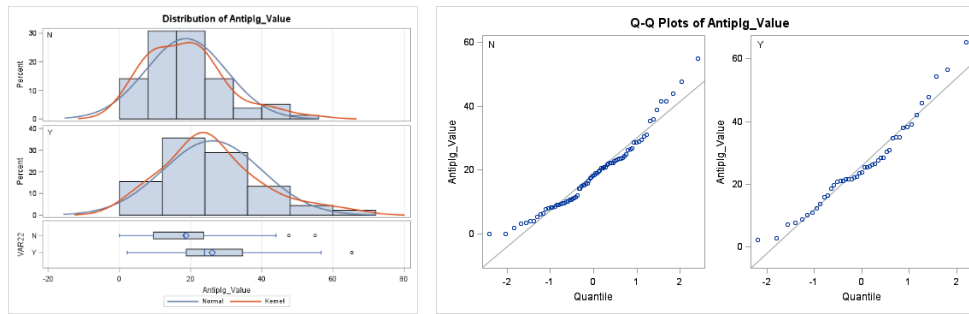
Method	DF	t Value	Pr > t
Pooled	38	-3.34	0.0019
Satterthwaite	3.1962	-5.50	0.0100

Figure 14. T-test of Calprotectin

In Figure 14, the results of T-test show $p\text{-value} = 0.002$ and suggest to reject the null hypothesis. In other words, there is statistically significant difference on biomarker Calprotectin between relapse group and non-relapse group.

As discussed in previous section 3.1, surely this results are not confident or reliable, and it can't validate with only 3 samples in relapse group. However, the strong shift suggests that Calprotectin might be a potential candidate for relapse prediction. It may be of great clinical value. Therefore, if it is possible for the researchers to collect more data from patients of relapse for predictor Calprotectin, there is high chance that we can get solid conclusion about the difference of Calprotectin level between relapse patients and patients of non-relapse.

T-test of other biomarkers are also carried out and the results are shown in Figure 15 through Figure 21. In each of the figure, the above box plot is the results from non-relapse group and the below box plot is from relapse patient group.

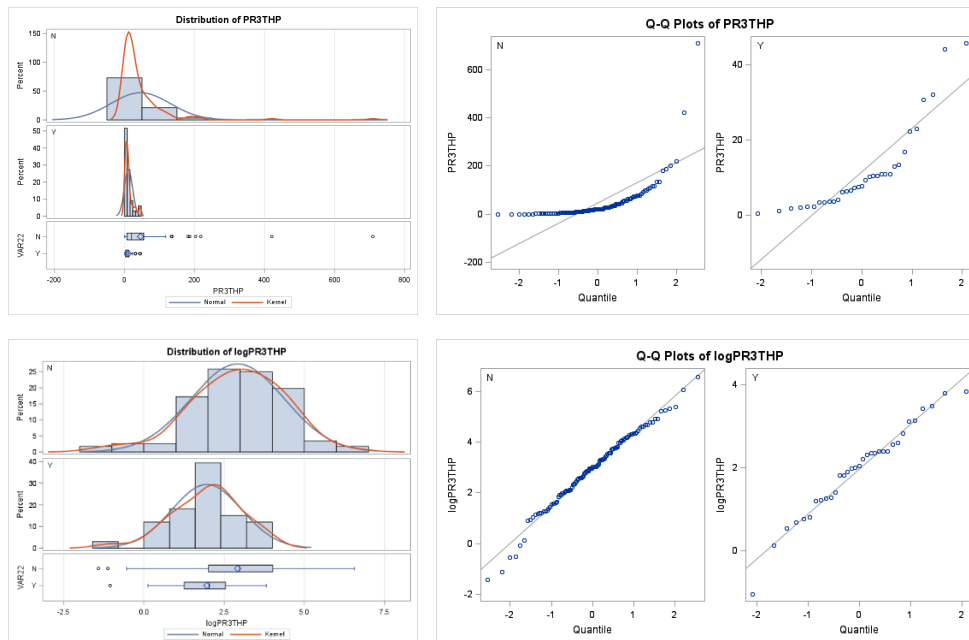


RELAPSE	N	Mean	Std Dev
N	78	18.6662	11.3579
Y	45	26.0653	13.9482
Diff (1-2)		-7.3991	12.3628

Method	Pr > t
Pooled	0.0018
Satterthwaite	0.0034

Figure 15. T-test of AntiPlg_Value

In Figure 15, the results of T-test show $p\text{-value} < 0.05$ and suggest to reject the null hypothesis. In other words, there is statistically significant difference on biomarker AntiPlg_Value between relapse and non-relapse groups.



RELAPSE	N	Mean	Std Dev
N	116	45.9934	84.1457
Y	33	11.5340	11.6260
Diff (1-2)		34.4594	74.6230

Method	Pr > t
Pooled	0.0206
Satterthwaite	<.0001

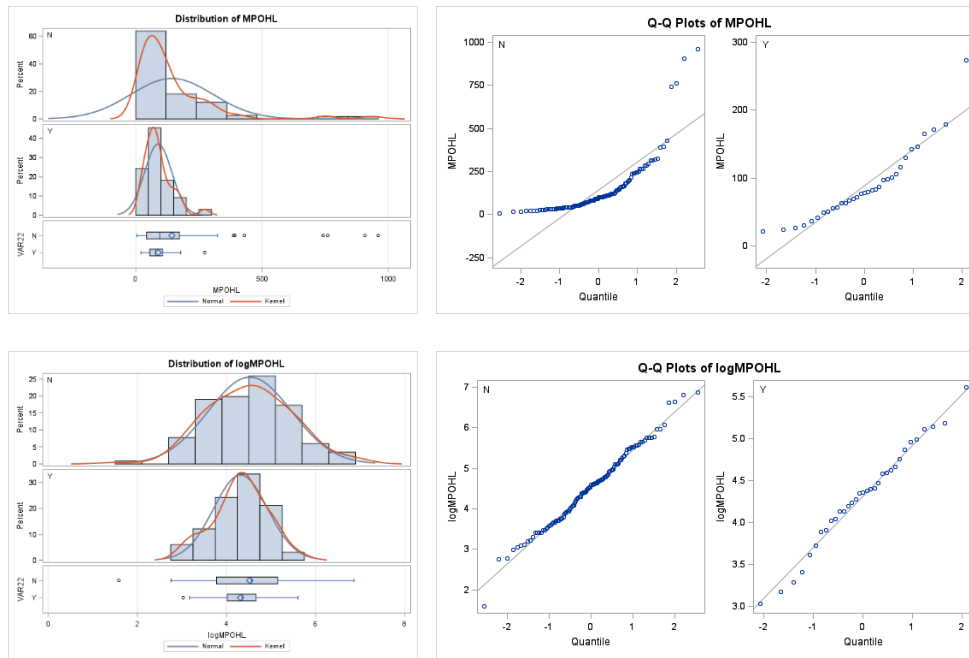
RELAPSE	N	Mean	Std Dev
N	116	2.9211	1.4528
Y	33	1.9609	1.0850
Diff (1-2)		0.9602	1.3811

Method	Pr > t
Pooled	0.0006
Satterthwaite	<.0001

Figure 16. T-test of PR3THP

In Figure 16, the results of T-test show $p\text{-value} < 0.05$ and suggest to reject the null hypothesis. In other words, there is statistically significant difference on biomarker PR3THP between relapse and non-relapse groups.

The top figure is from raw data and the bottom figure is from transformed data. I placed the figures side-by-side to show why transformation is useful and need. The distribution is more close to normal, and the variance is close to ideal situation.



RELAPSE	N	Mean	Std Dev
N	116	143.2	163.0
Y	33	88.8733	53.8691
Diff (1-2)		54.2825	146.4

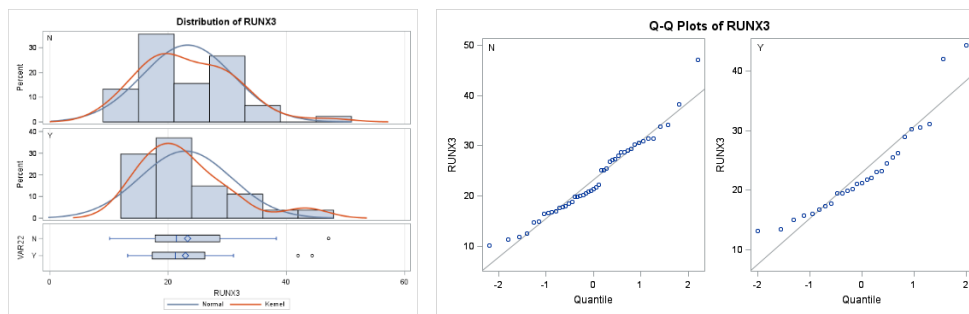
Method	Pr > t
Pooled	0.0621
Satterthwaite	0.0027

RELAPSE	N	Mean	Std Dev
N	116	4.5243	0.9381
Y	33	4.3174	0.6060
Diff (1-2)		0.2069	0.8766

Method	Pr > t
Pooled	0.2335
Satterthwaite	0.1344

Figure 17. T-test of MPOHL

In Figure 17, the results of T-test show $p\text{-value} > 0.05$ and suggest NOT to reject the null hypothesis. In other words, there is NO statistically significant difference on biomarker MPOHL between relapse and non-relapse groups. Again, the log transformation improve the data distribution.

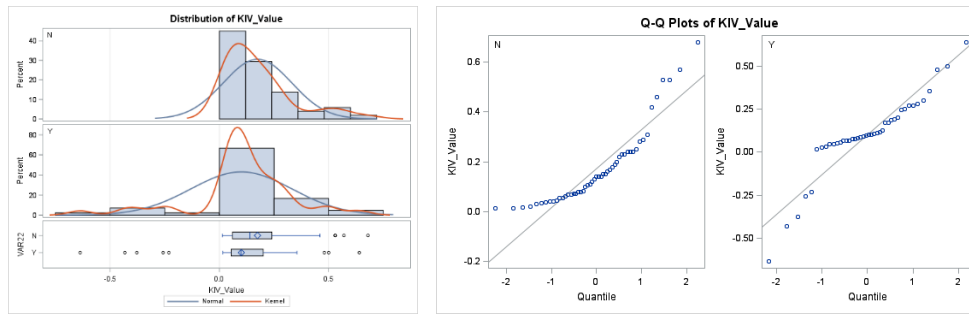


RELAPSE	N	Mean	Std Dev
N	45	23.2981	7.7013
Y	27	22.9371	7.7326
Diff (1-2)		0.3610	7.7130

Method	Pr > t
Pooled	0.8481
Satterthwaite	0.8484

Figure 18. T-test of RUNX3

In Figure 18, the results of T-test show $p\text{-value} > 0.05$ and suggest NOT to reject the null hypothesis. In other words, there is NO statistically significant difference on biomarker RUNX3 between relapse and non-relapse groups.

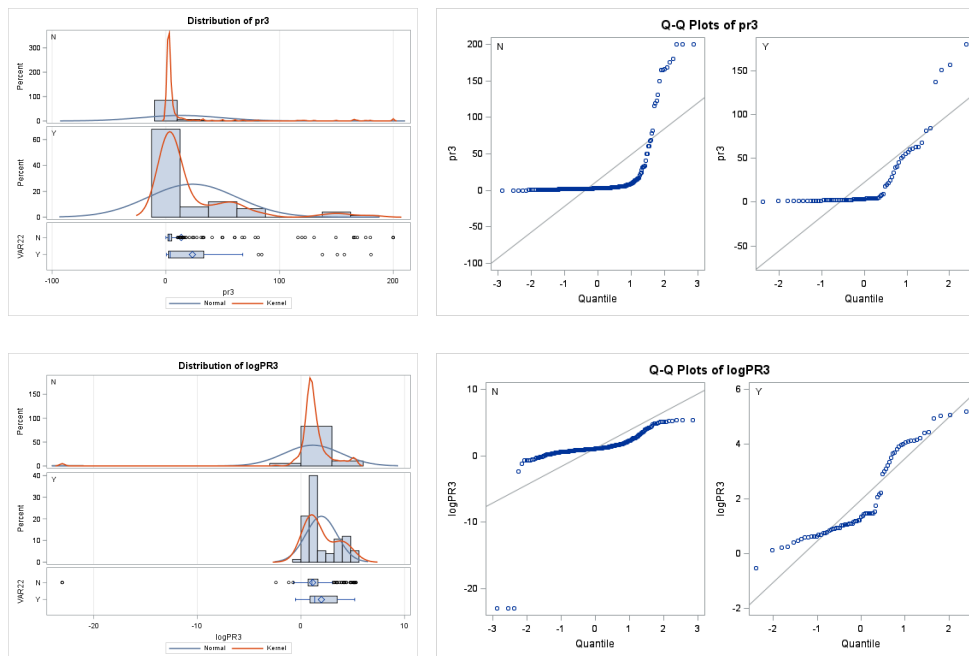


RELAPSE	N	Mean	Std Dev
N	51	0.1743	0.1565
Y	42	0.1005	0.2317
Diff (1-2)		0.0739	0.1940

Method	Pr > t
Pooled	0.0709
Satterthwaite	0.0825

Figure 19. T-test of KIV_Value

In Figure 19, the results of T-test show p-value > 0.05 and suggest NOT to reject the null hypothesis. In other words, there is NO statistically significant difference on biomarker KIV_Value between relapse and non-relapse groups.



RELAPSE	N	Mean	Std Dev
N	295	13.5104	35.5071
Y	75	23.3993	39.0376
Diff (1-2)		-9.8889	36.2447

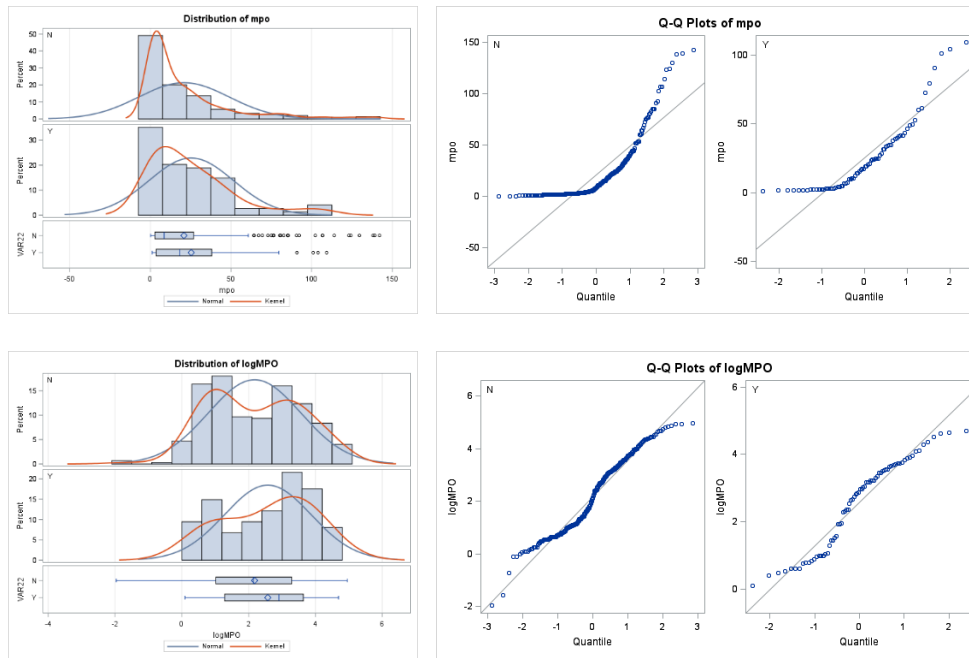
Method	Pr > t
Pooled	0.0355
Satterthwaite	0.0487

RELAPSE	N	Mean	Std Dev
N	295	1.1341	2.7494
Y	75	1.9711	1.5044
Diff (1-2)		-0.8370	2.5484

Method	Pr > t
Pooled	0.0115
Satterthwaite	0.0005

Figure 20. T-test of PR3

In Figure 20, the results of T-test show p-value < 0.05 and suggest to reject the null hypothesis. In other words, there is statistically significant difference on biomarker PR3 between relapse and non-relapse groups. The log transformation data shows distribution close to normal, though the transformed one is not very ideal. Further investigation on this part can be done for future study.



RELAPSE	N	Mean	Std Dev
N	299	20.8081	28.0372
Y	74	25.3153	26.1863
Diff (1-2)		-4.5072	27.6828

Method	Pr > t
Pooled	0.2106
Satterthwaite	0.1938

RELAPSE	N	Mean	Std Dev
N	299	2.1769	1.3845
Y	74	2.5737	1.2969
Diff (1-2)		-0.3968	1.3677

Method	Pr > t
Pooled	0.0260
Satterthwaite	0.0218

Figure 21. T-test of mpo

In Figure 21, the results of T-test show $p\text{-value} < 0.05$ and suggest to reject the null hypothesis. In other words, there is statistically significant difference on biomarker mpo between relapse and non-relapse groups.

Conclusions and discussions:

Above T-test results show that among all biomarkers, Calprotectin, Antipltg_Value, PR3THP, pr3, and mpo are of significant different between relapse patients group and non-relapse group. On the other hand, biomarkers of MPOHL, RUNX3 and KIV_Value are of insignificant between the two groups.

4. Logistic regression

4.1 Response: relapse or not

4.2 Predictors

Calprotectin, AntiPlg_Value, logPR3THP, logMPOHL, RUNX3, KIV_Value, logpr3
logmpo.

4.3 Model developing

SAS programming has been applied for logistic regression analysis. However, the analysis failed due to numerous of missing value. Even if five out of eight biomarkers were dropped off (that is, only three biomarkers are reserved), there are still 319 observations showing missing value. It seems that we may have to pick up only one or two biomarkers to make logistic regression practical. However, it does not make much sense to summarize or predict disease with only one predictor.

Therefore, the feasible suggestion is trying to collect more data so that more investigation and analysis can be carried out in future study.

4.3.1 Full model

First I checked logistic regression with all above biomarkers; however, this did not work. The possible reason is that there are huge amount of missing values as well as unbalanced data. For example, there are 40 observations demonstrating biomarker Calprotectin. However, 37 out of 40 are non-relapse and there are only 3 sample in relapse group. After dropping off Calprotectin, the logistic regression analysis can be carried out. If all current biomarkers are included in the model, the final model is:

$$\text{Probability of relapse} = \frac{\exp(\text{LM})}{1 + \exp(\text{LM})}$$

where $\text{LM} = -79.6 - 28.5 \times (\text{KIV_Value}) + 18.8 \times (\log\text{MPOHL}) + 0.512 \times (\text{RUNX3}) + 0.622 \times (\text{AntiPlg_Value}) - 15.1 \times (\log\text{PR3THP}) + 8.05 \times (\log\text{PR3}) - 3.02 \times (\log\text{MPO})$

However, this model does not fit well. Looking at the p-value of the parameters, all of the p-values are very large suggesting that the factors are not significant (Table 4).

Table 4. Statistical analysis results of logistic regression (full model, n=22)

Parameter	DF	Estimate	Standard Error	Pr > ChiSq
Intercept	1	-79.5856	456.6	0.8616
KIV_Value	1	-28.4621	929.0	0.9756
logMPOHL	1	18.7724	76.2164	0.8054
RUNX3	1	0.5121	4.5330	0.9100
Antiplg_Value	1	0.6215	2.9400	0.8326
logPR3THP	1	-15.1096	38.7668	0.6967
logPR3	1	8.0512	57.0006	0.8877
logMPO	1	-3.0224	39.5198	0.9390

Then I run model selection, and the optimal model is shown below.

4.3.2 Optimal logistic regression model

The optimal logistic regression model is:

$$\text{Probability of relapse} = \frac{\exp(\text{OLM})}{1 + \exp(\text{OLM})}$$

where $\text{OLM} = -2.48 + 0.074 \times (\text{AntiPlg_Value}) - 0.500 \times (\text{logPR3THP}) + 0.609 \times (\text{logPR3})$

Table 5. Statistical analysis of optimal logistic regression model (n=65)

Parameter	DF	Estimate	Standard Error	Pr > ChiSq
Intercept	1	-2.4819	0.9957	0.0127
Antiplg_Value	1	0.0743	0.0304	0.0146
logPR3THP	1	-0.4995	0.2432	0.0400
logPR3	1	0.6090	0.2335	0.0091

Table 5 shows the analysis of optimal logistic regression model, all items in the model are statistically significant (p-value for each item is less than 0.05) at 0.05 significance level.

Comparison should be taken carefully with Table 4 and Table 5. The results show that the optimal logistic regression model comes with all features being significant, while the full model comes with nothing significant. It might be results of model optimization, yet

could also be the results of different sample size. Recall that in section 2.1, Table 2 demonstrated the missing value in this dataset. There are a lot of missing data, and the sample size may decrease drastically along with the increase of number of features adopted in the model. Table 4 shows that the full model contains 7 features, and the sample size is 22. Meanwhile, Table 5 shows that the optimal model contains 3 features, and the sample size is 65. Therefore, the difference of p-value (significance) may be the results of either model selection or sample size change, or both. It will be great helpful is the researcher can collect more data in the future so that further analysis can be carried out with both models by same or similar sample size. In that case, if the optimal model still shows better fitting than full model, we can make solid conclusion of the preferred better model. It is also possible that with increasing sample size, full model also shows features being significant. Based on the current dataset, our best choice is the optimal logistic regression model.

According to this model, given a new patient with known biomarkers of AntiPlg_Value, PR3THP, and PR3, we are able to predict the probability of relapse for this patient.

5. Conclusions and Discussion for Future Work

- 5.1 Analysis shows that among all biomarkers, Calprotectin, Antipltg_Value, PR3THP, pr3, and mpo are of significant different between relapse patients group and non-relapse group. On the other hand, biomarkers of MPOHL, RUNX3 and KIV_Value are of insignificant between the two groups.
- 5.2 Those significant biomarkers are likely potential indicators for disease/relapse predictors. I would suggest the researchers in the lab and in clinical study to pay more attention to these biomarkers because they might be in tight relationship with the disease.
- 5.3 Due to large amount of missing values, further analysis is not practical with current dataset. It is highly suggested to collect more data so that deep investigation can be fulfilled, including but not limited to logistic regression, time series research, survivorship analysis, etc. With more data it is also possible to compare the full model and the current optimal model with same or similar sample size, which will provide solid conclusion.
- 5.4 Logistic regression model is provided in this report for prediction of patient relapse, while it is built based on very limited data. If more data is available in the future, it will be helpful to run validation, ridge regression, and/or LASSO.

References

- Berti A, Cornec D, Crowson CS, Specks U, Matteson EL, The Epidemiology of Antineutrophil Cytoplasmic Autoantibody-Associated Vasculitis in Olmsted County, Minnesota: A Twenty-Year US Population-Based Study, *Arthritis Rheumatol.* 2017 Dec;69(12):2338-2350.
- Jeff Howbert, Machine Learning Dimensionality Reduction, Introduction to Machine Learning, Winter 2014
- G. James, D. Witten, T.Hastie and R. Tibshirani, An Introduction to Statistical Learning, with applications in R" (Springer, 2013)
- Christopher J. C. Burges, Dimension Reduction: A Guided Tour, December 2009, Foundations and Trends® in Machine Learning 2(4). DOI: 10.1561/22000000002
- Lan Huong Nguyen and Susan Holmes, Ten quick tips for effective dimensionality reduction, *PLoS Comput Biol.* June 20, 2019, <https://doi.org/10.1371/journal.pcbi.1006907>
- Sean Simmons, Jian Peng, Jadwiga Bienkowska, and Bonnie Berger, Discovering What Dimensionality Reduction Really Tells Us About RNA-Seq Data, *J Comput Biol.* 2015 Aug 1; 22(8): 715–728.
- Xavier Bossuyt, Jan-Willem Cohen Tervaert, Yoshihiro Arimura, Daniel Blockmans, Luis Felipe Flores-Suárez, Loïc Guillevin, Bernhard Hellmich, David Jayne, J. Charles Jennette, Cees G. M. Kallenberg, Sergey Moiseev, Pavel Novikov, Antonella Radice, Judith Anne Savige, Renato Alberto Sinico, Ulrich Specks, Pieter van Paassen, Ming-hui Zhao, Niels Rasmussen, Jan Damoiseaux & Elena Csernok, Revised 2017 international consensus on testing of ANCAs in granulomatosis with polyangiitis and microscopic polyangiitis, *Nature Reviews Rheumatology* volume 13, pages 683–692 (2017)
- Ian Jolliffe, Principal Component Analysis, 15 October 2005, <https://doi.org/10.1002/0470013192.bsa501>
- Dan Shen, Haipeng Shen, Hongtu Zhu, and J. S. Marron, The Statistics and Mathematics of High Dimension Low Sample Size Asymptotics, *Stat Sin.* 2016 Oct; 26(4): 1747–1770.
- Stevens, J.P., Applied Multivariate Statistics for the Social Sciences, Routledge, 2009, 5th Edition.
- Changyong FENG, Hongyue WANG, Naiji LU, Tian CHEN, Hua HE, Ying LU, and Xin M. TU, Log-transformation and its implications for data analysis, *Shanghai Arch Psychiatry.* 2014 Apr; 26(2): 105–109.

- Ramsay, J.O., Silverman, B.W., Applied Functional Data Analysis, Springer 2002.
- J. Barrientos-Marin , F. Ferraty & P. Vieu, Locally modelled regression and functional data, 18 Journal of Nonparametric Statistics, Nov 2009, 617-632
- Ramsay, James, Silverman, B. W. Functional Data Analysis, Springer 2005.
- B. Lazarus, G. T. John, C. O’Callaghan, and D. Ranganathan, Recent advances in anti-neutrophil cytoplasmic antibody-associated vasculitis, Indian J Nephrol. 2016 Mar-Apr; 26(2): 86–96.
- Manikandan S., Data transformation, 2010, Journal of Pharmacology and Pharmacotherapeutics 1(2):126-7
- J. C. Jennette, A. S. Wilkman, and R. J. Falk, Anti-neutrophil cytoplasmic autoantibody-associated glomerulonephritis and vasculitis, Am J Pathol. 1989 Nov; 135(5): 921–930.
- J. Charles Jennette and Ronald J. Falk, Pathogenesis of antineutrophil cytoplasmic autoantibody-mediated disease, NATURE REVIEWS, VOLUME 10 | AUGUST 2014 | 463-473.
- Judith Land, Abraham Rutgers, Cees G.M. Kallenberg, Anti-neutrophil cytoplasmic autoantibody pathogenicity revisited: pathogenic versus non-pathogenic anti-neutrophil cytoplasmic autoantibody, *Nephrology Dialysis Transplantation*, Volume 29, Issue 4, April 2014, Pages 739–745
- C.O.S Savage, Pathogenesis of anti-neutrophil cytoplasmic autoantibody (ANCA)-associated vasculitis, Clin Exp Immunol. 2011 May; 164(Suppl 1): 23–26.
- Templeton, Gary F. (2011) A Two-Step Approach for Transforming Continuous Variables to Normal: Implications and Recommendations for IS Research, Communications of the Association for Information Systems: Vol. 28, Article 4. DOI: 10.17705/1CAIS.02804
- World Health Organization, Global Tuberculosis Report. 2015
- Howard J. Seltman, 2018, Experimental Design and Analysis, <http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf>
- Armstrong RA1, Slade SV, Eperjesi F., An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry., Ophthalmic Physiol Opt. 2000 May;20(3):235-41.
- Martin G. Larson, SD, Analysis of Variance, Circulation. 2008;117:115-121, DOI: 10.1161/CIRCULATIONAHA.107.654335
- Howard J. Seltman, 2018, Experimental Design and Analysis

David M. Lane, David Scott¹, Mikki Hebl¹, Rudy Guerra¹, Dan Osherson¹, and Heidi Zimmer,
Introduction to Statistics, Rice University

Lora D. Delwiche and Susan J. Slaughter, The Little SAS® Book: A Primer, Fifth Edition, 2012,
SAS Institute Inc., Cary, NC, USA

Acknowledgement

Praise the Lord that I went through the past years and finished the thesis.

It is a mission impossible for me to accomplish this thesis without help and mentorship from Professor James Steven Marron. I appreciate and enjoy his teaching, his patents, and his love to students.

I would like to take this chance to thank those professors and staff, Dr. Chuanshu Ji, Dr. Shankar Bhamidi, Ms. Christine Keat, Ms. Alison Kieber, Dr. Robert Carlstein, Dr. Kai Zhang, Dr. Richard Smith, Dr. Xianming Tan, Dr. James Steven Marron, Dr. Bajhat Qaqish, Dr. Todd Schwartz, Dr. Gary Koch, Dr. Perry Haaland, Dr. Yufeng Liu.

I also thanks my friends and classmates and appreciate the help I received from them, Mr. Yehong Wan, Mr. Tao Bian, Ms. Hongxia Yan, Mr. Weibin Mo, Ms. Yiqing Wei, Ms. Tong Zhu, Ms. Na Lin, Ms. Hang Yu, and many other people.

I also greatly thank my friends who may not be statisticians but support me a lot and pray for me through the hard time: “Silver or gold I do not have, but what I do have I give you.” My great appreciation to Mr. Barry Yang, Mr. Quan Li, Mr. Tianmiao Hu, Ms. Zongping Yu, Ms. Ying Lin, Mr. Xueliang Pan, Mr. Rong Wang, Mr. Xiaochun Sun, Ms. Wei Li, Mr. Jun Li, Ms. Ruihua Zhou, Ms. Zhiping Feng, Mr. Zhidan Xiang, Ms. Li Gou, Mr. Wensheng Cai, Ms. Weihong, Mr. Zuguang Huang, Ms. Yingmiao Liu, Ms. Yahui Li, Mr. Zhimou Wen, Ms. Ping Wu, Pastor Oo, Pastor Chongyao, Pastor Rich, Mr. Bruce Stevenson and Ms. Alice Stevenson. “I once was lost but now I'm found, was blind but now I see.” (John Newton)

Last and foremost I wish to thank my wife, who had stood by me through all my travails, my absences, my fits of pique and impatience. She gave me support and help, discussed ideas and prevented several wrong turns. She also supported the family during much of my graduate studies. Along with her, I want to acknowledge my two kids who loves me a lot more than I love them. My sister and her families, my parents and parents-in-law, and the vast extended family from both my side and my wife's side have all been wonderful – and very patient.

Finally, my sincerely thanks to my committee as well as my department. People here make me feel like at home. I enjoy the time here and appreciate the past years, with tears and smile. This is another milestone in my life, and I will miss Department of Statistics and Operations Research, UNC at Chapel Hill.