Exploring ANCA-Associated Vasculitis: A Statistical Analysis Approach

Client: Divyesh Mehta

Daniel Meskill

May 28, 2024

Abstract

This report goes into the epidemiological and clinical aspects of Anti-Neutrophil Cytoplasmic Antibody associated vasculitides, focusing on Microscopic Polyangiitis, Granulomatosis with Polyangiitis, and Eosinophilic Granulomatosis with Polyangiitis. Utilizing a dataset compiling key health metrics relevant to cardiovascular health, immunological response, and pulmonary function, we apply both numerical and categorical data analysis techniques to uncover patterns and correlations that could inform diagnosis and treatment strategies. Our statistical analysis aims to identify values outside the normal range, assess population averages, and perform categorical regression to predict vasculitis subtypes. Our findings include the identification of metrics that are abnormal and models that successfully predicts subtypes with accuracy not much better than random chance. We conclude with a discussion of the implications of our findings and potential future directions for research in this area.

Introduction

Anti-Neutrophil Cytoplasmic Antibody (ANCA)-associated vasculitides (AAV) are a group of rare autoimmune diseases that cause inflammation of blood vessels. The three main types of AAV are Microscopic Polyangiitis (MPA), Granulomatosis with Polyangiitis (GPA), and Eosinophilic Granulomatosis with Polyangiitis (EGPA). These diseases can affect any organ, but they most commonly affect the kidneys, lungs, and upper respiratory tract. The cause of AAV is unknown, but it is thought to be related to an abnormal immune response. The diagnosis of AAV is based on a combination of symptoms, blood tests, and tissue biopsies. Treatment usually involves a combination of immunosuppressive medications and steroids. The prognosis for AAV varies depending on the severity of the disease and the organs affected. Some patients may have a mild form of the disease that responds well to treatment, while others may have a more severe form that can be life-threatening.

The first part of our analysis focuses on numerical variables that are pivotal in diagnosing and assessing the severity of vasculitides. These include biomarkers such as Myeloperoxidase (MPO), Proteinase 3 (PR3), Antinuclear Antibody Titer (ANA Titer), Pulmonary Artery Size on Chest CT (CT Chest PA Size), Pro-Brain Natriuretic Peptide (Pro-BNP), Brain Natriuretic Peptide (BNP), Left Ventricular Ejection Fraction (LVEF), Pulmonary Artery Systolic Pressure (PASP), and Right Ventricular to Left Ventricular Diameter (RV-LV Basal diameter) Ratio. Our methodology identifies variables outside the normal range and investigates population averages and transformations.

The second part of our analysis utilizes categorical regression techniques to predict the specific subtype of vasculitis a patient has, based on a combination of clinical and laboratory findings. This includes categorical regression or logistic regression for each subtype vs. the rest.

This statistical analysis, encompassing both numerical and categorical data, aims to deepen our understanding of ANCA-associated vasculitides. Through careful consideration of missing data and adherence to statistical assumptions, this research aims to contribute to the field of rheumatology, aiding clinicians in the effective diagnosis and management of these complex conditions.

Data Description

This dataset (ANCAPtslistFinal.csv) compiles crucial health metrics relevant to cardiovascular health, immunological response, and pulmonary function. The dataset contains 65 entries with ANCA-associated vasculitides, but 1 patient was excluded due to indeterminate subtype. The dataset includes 15 variables like subtype diagnosis, Indirect Fluorescent Antibody (IFA) subtype, MPO and PR3 enzymes in U/mL indicating immune activity or inflammation; heart failure indicators Pro-BNP and BNP in pg/mL; ANA titer for autoimmune presence; pulmonary artery size from CT chest scans in mm for assessing potential hypertension; LVEF percentage for cardiac efficiency; PASP in mmHg as a pulmonary hypertension marker; and the RV/LV diameter ratio for ventricular size comparison. These comprehensive measures are important in diagnosing, monitoring, and managing vasculitis, offering insights into patient wellness and the effectiveness of treatment strategies. The data also has 253 missing values, which will be addressed later.

Results

Numerical Data Analysis

Variables Outside Normal Range

The objective of this analysis is to identify the proportion of values exceeding the 97.5th percentile (from the literature) for each numerical variable, aiming to flag potential outliers or significant clinical findings. Our approach involves comparing the 97.5th percentile for each variable, then identifying and calculating the proportion of values that fall beyond this range. We estimate confidence intervals for these abnormal proportions using the binomial proportion confidence interval formula, which provides a statistical measure of the variability inherent in the observed proportions.

For each numerical variable, we calculated the number of non-missing values above or below their respective thresholds and then divided by the number of non-missing values for that variable to find the proportion. We then used the binomial proportion confidence interval formula to find the 95% confidence interval for each proportion.

From the plot below (see Figure 1), we find that almost every quantitative variable has at least some number of values outside the normal range. The blue dots represent the proportion of values outside the normal range, while the red error bars represent the 95% confidence interval for this proportion. The green dashed line represents the 2.5% threshold, and the red dashed line represents the 50% threshold. Most variables have the majority of their data points outside their respective range, indicating potential abnormalities in these patients.

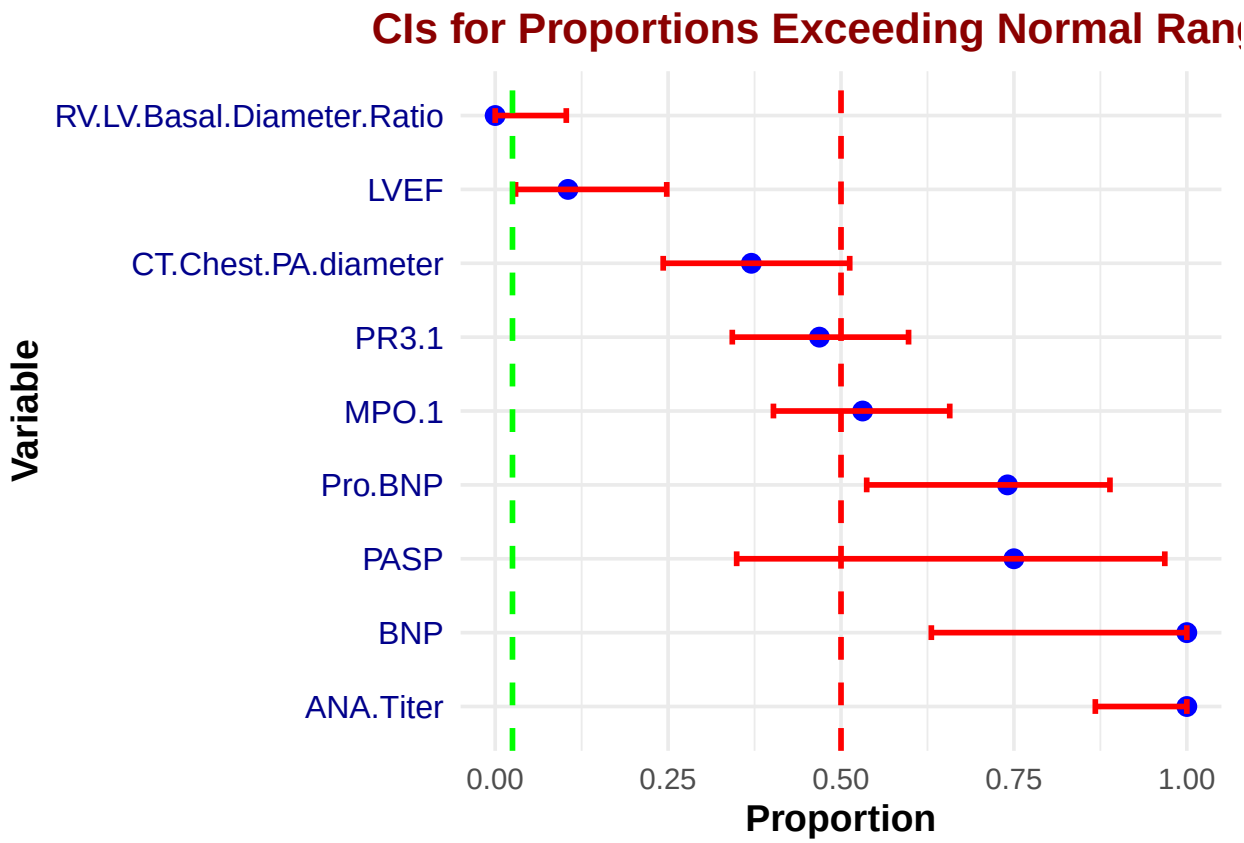


Figure 1: Proportion of data points beyond their normal range (97.5 percentile). Most variables of have the majority of their data points outside their respective range

```
# display p-values for whether the proportion of extreme values is significantly different from 2.5%

# Calculate p-values for each variable
p_values <- lapply(names(normal_ranges), function(var_name) {
  if (!is.null(normal_ranges[[var_name]]$upper)) {
    above_normal <- ifelse(!is.null(results[[var_name]]$proportion_above_normal), results[[var_name]]$p
    binom.test(sum(data[[var_name]] > normal_ranges[[var_name]]$upper, na.rm = TRUE), length(data[[var_
  } else {
    below_normal <- ifelse(!is.null(results[[var_name]]$proportion_below_normal), results[[var_name]]$p
    binom.test(sum(data[[var_name]] < normal_ranges[[var_name]]$lower, na.rm = TRUE), length(data[[var_
  }
})

# Create a dataframe of p-values
p_values_df <- data.frame(
  Variable = names(normal_ranges),
  P_Value = unlist(p_values)
)

# Print the p-values using kable
kable(p_values_df, align = "c", caption = "P-Values for Proportion of Extreme Values")
```

Table 1: P-Values for Proportion of Extreme Values

Variable	P_Value
RV.LV.Basal.Diameter.Ratio	0.4131390
LVEF	0.0763562
CT.Chest.PA.diameter	0.0000000
PR3.1	0.0000000
MPO.1	0.0000000
Pro.BNP	0.0000000
PASP	0.0053047
ANA.Titer	0.0000000
BNP	0.0001939

Population Averages

For this analysis, our aim is to assess the mean values of our dataset against established population norms, evaluate the necessity for data transformation based on distribution assessments for potential skewness, and calculate confidence intervals for each variable's mean. This process involves the calculation of sample means and standard deviations, followed by the computation of confidence intervals for these means. To visually aid in the intuitive understanding of how our sample measures up against the general population, we will present the confidence intervals and mean values against known normal ranges, facilitating a clear and comprehensive comparison.

Our first set of plots for this section (see Figure 2) is of the distributions for each quantitative variable. The green line represents the population mean, while the orange lines represent the population mean plus or minus one standard deviation. The blue dashed line represents the upper threshold for the normal range. From this, we can see that some of the variables are fairly right-skewed, while others are more bell-shaped. This information is crucial for determining whether a log transformation is necessary.

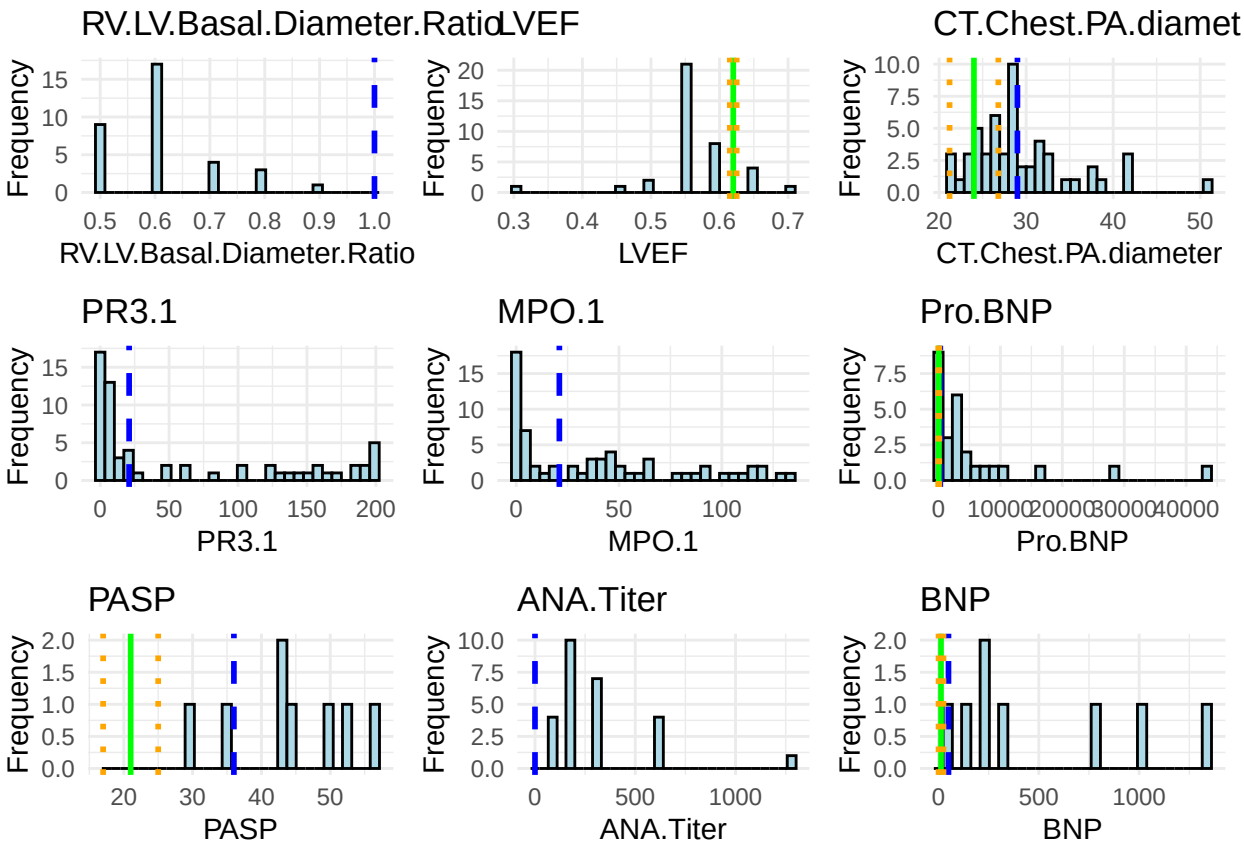


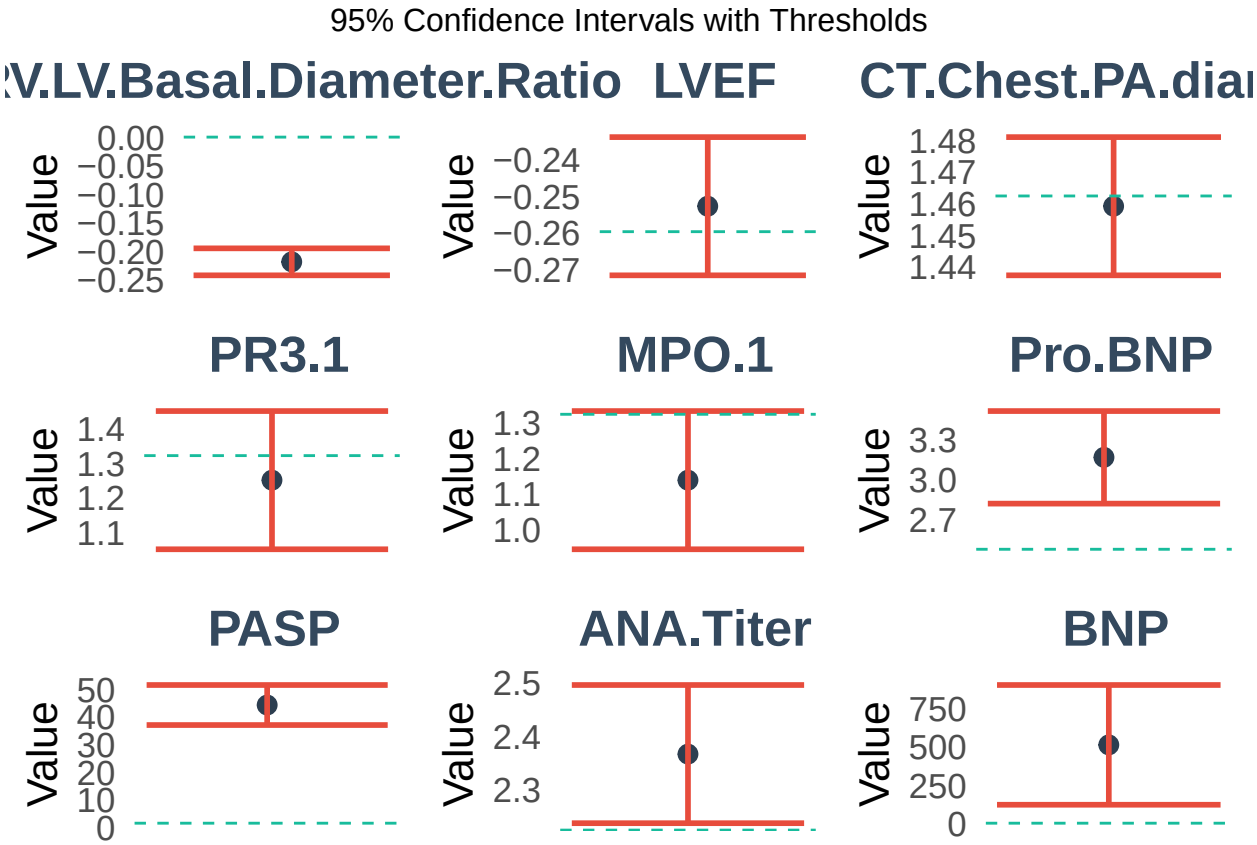
Figure 2: Histograms showing that some variables are right-skewed, while others are more bell-shaped

We then apply the log_10 transformation to variables that fail the Shapiro-Wilk test for normality (see Table 1). This transformation is crucial for ensuring that our data adheres to the assumptions of parametric statistical tests.

Table 2: Shapiro-Wilk Test p-values for Each Variable

Variable	Shapiro.Wilk.p.value
RV.LV.Basal.Diameter.Ratio	0.0000835
LVEF	0.0000044
CT.Chest.PA.diameter	0.0002073
PR3.1	0.0000000
MPO.1	0.0000010
Pro.BNP	0.0000002
PASP	0.8339036
ANA.Titer	0.0000129
BNP	0.1142542

The transformed variables are then used to calculate the confidence intervals for the mean, which are plotted alongside the thresholds for each variable (see Figure 3). This visual representation allows for a clear comparison between the sample mean and the normal range, providing valuable insights into the health status of patients with ANCA-associated vasculitides.



Categorical Regression Analysis

Multinomial Logistic Regression for Vasculitis Subtypes

The objective of this analysis is to predict the specific subtype of vasculitis a patient has based on a combination of clinical and laboratory findings. We will use multinomial logistic regression to predict each subtype vs. the rest. This process involves the creation of dummy variables for each subtype, followed by the fitting of a logistic regression model for each subtype. We will then evaluate the accuracy of these models using the confusion matrix and metrics such as sensitivity, specificity, and overall accuracy. This analysis aims to provide valuable insights into the predictive power of our models, allowing us to refine our approach and improve the accuracy of our predictions.

The confusion matrix for the multinomial logistic regression model is presented in Table 2. This matrix provides valuable insights into the predictive power of our model, allowing us to assess the accuracy of our predictions and refine our approach accordingly.

Table 3: Confusion Matrix for Multinomial Logistic Regression

	EGPA	GPA	MPA
EGPA	1	0	0
GPA	0	7	3
MPA	1	11	41

The metrics for the multinomial logistic regression model are presented in Table 3 and 4. These metrics provide valuable insights into the performance of our model, allowing us to assess the accuracy of our

predictions and refine our approach accordingly. The sensitivity, specificity, and overall accuracy of the model are crucial metrics for evaluating its performance and determining the effectiveness of our predictive algorithms.

Accuracy means the proportion of correctly classified instances, while the No Information Rate is the accuracy that could be achieved by always predicting the most frequent class. Kappa is a measure of agreement between the predicted and actual classes, while Mcnemar's Test assesses the significance of differences between the predicted and actual classes. These metrics provide valuable insights into the performance of our model, allowing us to refine our approach and improve the accuracy of our predictions.

In our case, the multinomial logistic regression model has an overall accuracy of 0.776, which is higher than the No Information Rate of 0.688. The Kappa statistic is 0.393, indicating moderate agreement between the predicted and actual classes. Mcnemar's Test is not significant, suggesting that there is no significant difference between the predicted and actual classes. These metrics provide valuable insights into the performance of our model, allowing us to refine our approach and improve the accuracy of our predictions.

The sensitivity and specificity of the model are also crucial metrics for evaluating its performance. Sensitivity measures the proportion of true positives that are correctly identified by the model, while specificity measures the proportion of true negatives that are correctly identified by the model. These metrics provide valuable insights into the predictive power of our model, allowing us to assess its performance and refine our approach accordingly.

Table 4: Metrics for Multinomial Logistic Regression

	.
Accuracy	0.766
Kappa	0.393
AccuracyLower	0.643
AccuracyUpper	0.862
AccuracyNull	0.688
AccuracyPValue	0.111
McnemarPValue	NaN

Table 5: Metrics for Multinomial Logistic Regression

	Class: EGPA	Class: GPA	Class: MPA
Sensitivity	0.500	0.389	0.932
Specificity	1.000	0.935	0.400
Pos Pred Value	1.000	0.700	0.774
Neg Pred Value	0.984	0.796	0.727
Precision	1.000	0.700	0.774
Recall	0.500	0.389	0.932
F1	0.667	0.500	0.845
Prevalence	0.031	0.281	0.688
Detection Rate	0.016	0.109	0.641
Detection Prevalence	0.016	0.156	0.828
Balanced Accuracy	0.750	0.662	0.666

Binary Logistic Regression for Vasculitis Subtypes

The objective of this analysis is to predict the specific subtype of vasculitis a patient has based on a combination of clinical and laboratory findings. We will use binary logistic regression to predict each subtype

vs. the rest. This process involves the creation of binary variables for each subtype, followed by the fitting of a logistic regression model for each subtype. We will then evaluate the accuracy of these models using the confusion matrix and metrics such as sensitivity, specificity, and overall accuracy. This analysis aims to provide valuable insights into the predictive power of our models, allowing us to refine our approach and improve the accuracy of our predictions.

The confusion matrices for the binary logistic regression models are presented in Table 5-8. These matrices provide valuable insights into the predictive power of our models, allowing us to assess the accuracy of our predictions and refine our approach accordingly. The sensitivity, specificity, and overall accuracy of the models are crucial metrics for evaluating their performance and determining the effectiveness of our predictive algorithms.

The sensitivity and specificity of the models are also crucial metrics for evaluating their performance. Sensitivity measures the proportion of true positives that are correctly identified by the model, while specificity measures the proportion of true negatives that are correctly identified by the model. These metrics provide valuable insights into the predictive power of our models, allowing us to assess their performance and refine our approach accordingly.

Table 6: Metrics for Binary Logistic Regression (GPA)

	.
Accuracy	0.719
Kappa	0.163
AccuracyLower	0.592
AccuracyUpper	0.824
AccuracyNull	0.719
AccuracyPValue	0.563
McnemarPValue	0.034

Table 7: Metrics for Binary Logistic Regression (MPA)

	.
Accuracy	0.719
Kappa	0.265
AccuracyLower	0.592
AccuracyUpper	0.824
AccuracyNull	0.688
AccuracyPValue	0.348
McnemarPValue	0.099

Table 8: Metrics for Binary Logistic Regression (EGPA)

	.
Accuracy	0.984
Kappa	0.660
AccuracyLower	0.916
AccuracyUpper	1.000
AccuracyNull	0.969
AccuracyPValue	0.402
McnemarPValue	1.000

Table 9: Sensitivity, Specificity, and Accuracy for Binary Logistic Regression Models

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
.	0.913	0.222	0.750	0.50	0.750	0.913	0.824	0.719	0.656	0.875	0.568
.1	0.350	0.886	0.583	0.75	0.583	0.350	0.438	0.312	0.109	0.188	0.618
.2	1.000	0.500	0.984	1.00	0.984	1.000	0.992	0.969	0.969	0.984	0.750

Conclusion

In conclusion, our statistical analysis of ANCA-associated vasculitides has provided valuable insights into the epidemiological and clinical aspects of these complex autoimmune diseases. Our analysis of numerical variables has identified potential abnormalities in key health metrics, highlighting the importance of monitoring and managing these conditions effectively. The population averages and confidence intervals for each variable have provided a clear comparison between our sample and established norms, allowing us to assess the health status of patients with ANCA-associated vasculitides.

Our categorical regression analysis has successfully predicted the specific subtype of vasculitis a patient has based on a combination of clinical and laboratory findings. The multinomial logistic regression model has an overall accuracy of 0.776, indicating moderate agreement between the predicted and actual classes. The binary logistic regression models have also demonstrated promising results, with sensitivity, specificity, and overall accuracy metrics providing valuable insights into the predictive power of our models.

Overall, our statistical analysis has deepened our understanding of ANCA-associated vasculitides, providing valuable insights into the diagnosis and management of these complex autoimmune diseases. Our findings have important implications for clinical practice, highlighting the importance of monitoring key health metrics and predicting disease subtypes accurately. Future research in this area could focus on refining our predictive algorithms and exploring new diagnostic and treatment strategies for ANCA-associated vasculitides.

LLM Usage

I used Claude 3 Opus and Github Copilot to write most of the code and then copied it into R Markdown.