Detecting Horse Cardiac Arrhythmia with ECG Data Final Report

Introduction:

A cardiac arrhythmia is a potentially life-threatening condition that causes an irregular rhythm in the heart. Using an electrocardiogram (ECG), clinicians can detect these abnormalities in a non-invasive manner. However, ECG analysis is a time-consuming process, making it a prime candidate for statistical methods to scale the detection of arrhythmia. In this project, we will work with equine ECG data, specifically RR intervals, provided by Dr. Katharyn Mitchell from Cornell University to perform advanced data analysis that will enhance our understanding of cardiac arrhythmias in horses. In our research, we will perform outlier identification and shape analysis as well as use a Gaussian Mixture Model (GMM) for our main classification method. Through this, we aim to find out what RR interval data summary statistics play a role in detecting cardiac arrhythmias in racehorses. Our research would help veterinarians understand how we can better predict race horse health issues before they become a significant problem and guide the conversation surrounding what technologies and investments should be made to keep race horses healthy. Ultimately, allowing race horses and the art of the sport to continue.

Data:

To explore our research question, we are working with ECG data of 13 horses each with 11 windows of time across a 24-hour period. In this data we have 4 variables (dirtyRR, HR, Time, and RR) but for the sake of this project, we are primarily interested in exploring the RR interval data variable. Below, we see a figure which explains what an RR interval is:

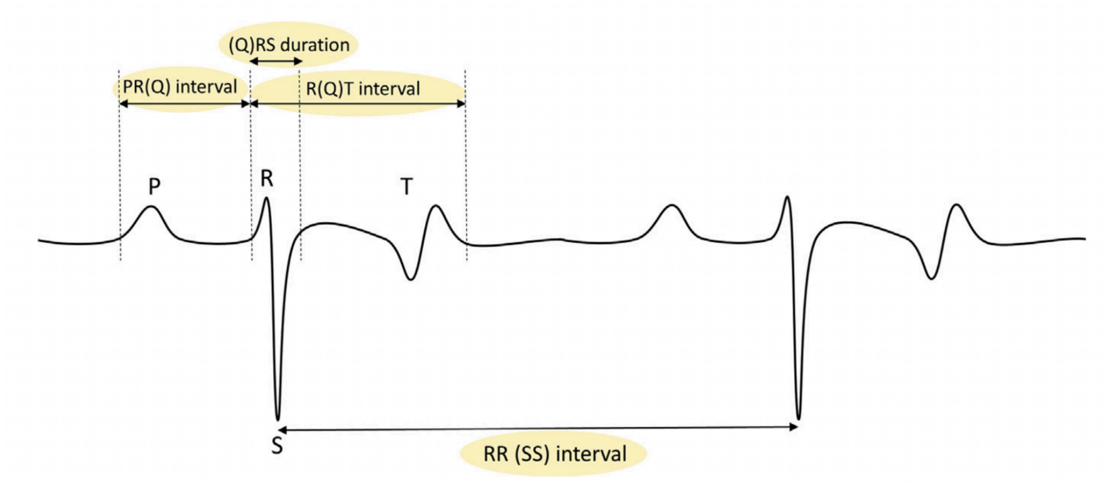


Figure 1: RR Interval Plot

In the experiments conducted by our client, horses are injected with an endotoxin that simulates symptoms of cardiac arrhythmia and they compiled RR interval data for each phase. They collected data starting from 1 hour before administering the injection until up to 20+ hours after the horses received the endotoxin. For our analysis, we decided to use all of these windows but split them into three categories.

Endotoxin Admin	**Severe Symptoms**	**Recovery Stage**
Before Endotoxin Pre-Sedation Post-Sedation	0-1 Hour 1-3 Hours 3-5 Hours	5-7 Hours to the end of the experiment

Table 1: 11 windows of time divided into 3 phases

We chose these categories because we wanted to see if there would be any differences in the RR interval data between when the horses obtain the injection, when they experience the worst symptoms, and when they recover after receiving treatment. More specifically, our window of interest is the severe symptoms window because that is when the horses show the most symptoms in the experiment. At the 5-7 hours window, they receive treatment that helps the horses gradually recover from the symptoms they were experiencing, however, they will still be showing some symptoms of cardiac arrhythmia as they recover.

EDA:

In order to explore our data, we first start with using Poincaré plots to visualize the RR intervals. A Poincaré plot is a recurrence plot that allows us to visualize the correlation between consecutive data points of a time series, commonly used for ECG data. Each point on this plot represents a pair of consecutive heartbeats and these are graphs that show RR(n) on the x-axis and RR(n + 1) (the succeeding RR interval) on the y-axis. We decided to visualize this data by each horse (including all time periods) and each time period (including all horses). Below, we see examples of both of those plots.

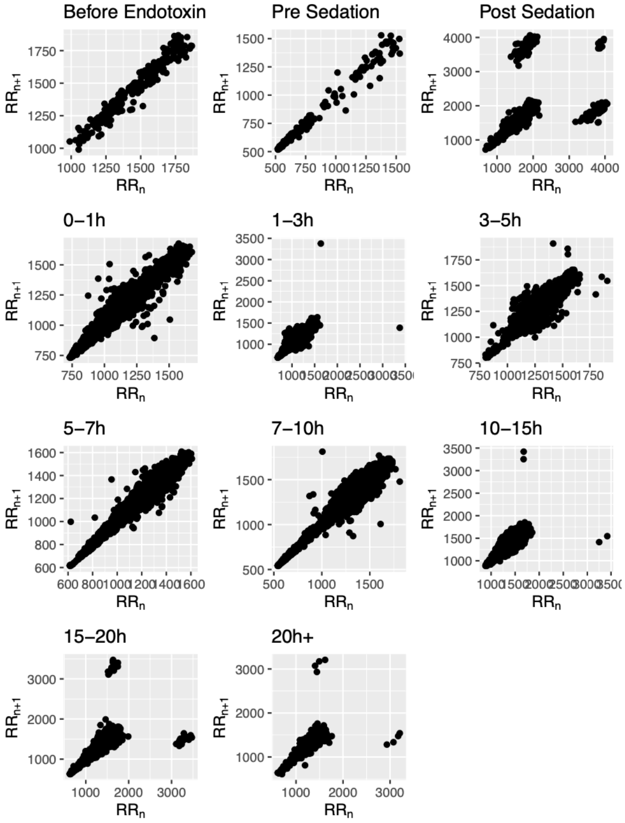


Figure 2: Poincaré plot grid for Horse 2 across all windows of time

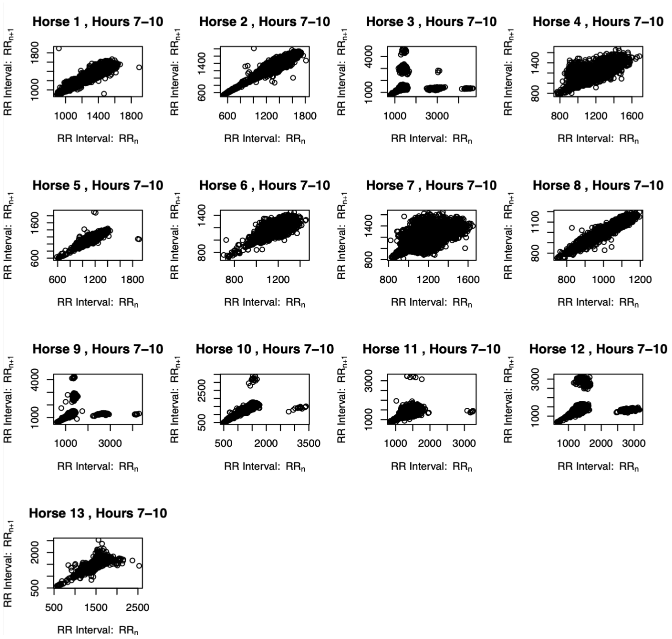


Figure 3: Poincaré plot grid for 7-10 hours for all Horses

By visualizing our data in both of these ways, we were able to observe two things: (1) with each horse, we were able to see how its individual data changed over time and (2) with each window of time, we were able to compare how the horses performed against each other, allowing us to pinpoint any concerning hours in the experiment. Another way we visualized our data was with a line plot displaying all 13 horses and their mean RR interval value at each window of time, separated by the three phases mentioned earlier.

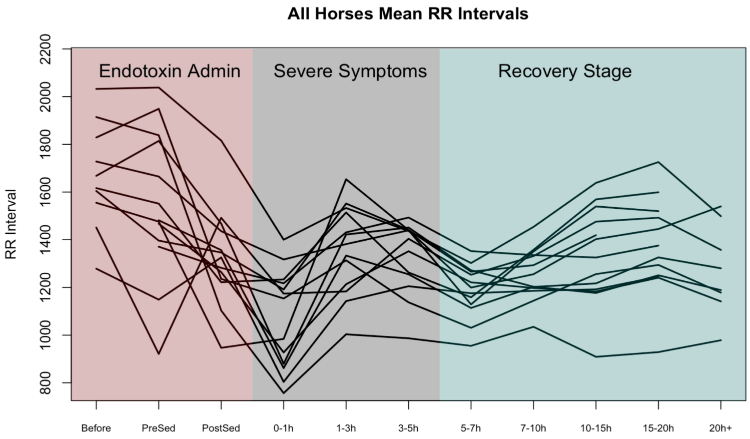


Figure 4: Trends of Mean RR Intervals of all 13 Horses

Methods:

Atrioventricular (AV) blocks take place when “the electrical impulse from the top chamber of the heart (atria) does not conduct to the bottom chamber of the heart (ventricle)” [1]. In this research, AV blocks are treated as outliers because they are common in horses and rarely need treatment in comparison with other kinds of fatal cardiac arrhythmias. On Poincaré plots, AV blocks are represented by large positive outliers right after a normal RR interval. The presence of AV blocks negatively affects the clustering assignments and later the dispersion analysis. Utilizing Horse 12 pre-sedation period as an example in Figure 5 below, we observe three large spikes that are distinguishable from the other RR interval data. They are classified as AV blocks.

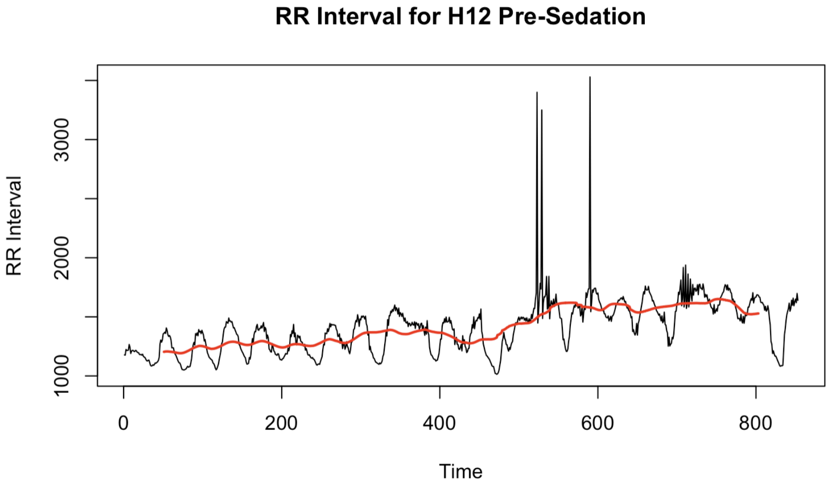


Figure 5: RR interval time series for Horse 12 Pre-Sedation period

Without the removal of those positive outliers, a cluster that is far away from the diagonal line is presented in Figure 6 while constructing a Gaussian Mixture model. If the AV blocks are removed, the cluster assignment will just be a diagonal line and thus is easier for dispersion analysis.

[1] “Cardiology: Cornell University Equine Hospital.” Cornell University College of Veterinary Medicine, September 12, 2022. https://www.vet.cornell.edu/hospitals/services/cardiology-0.

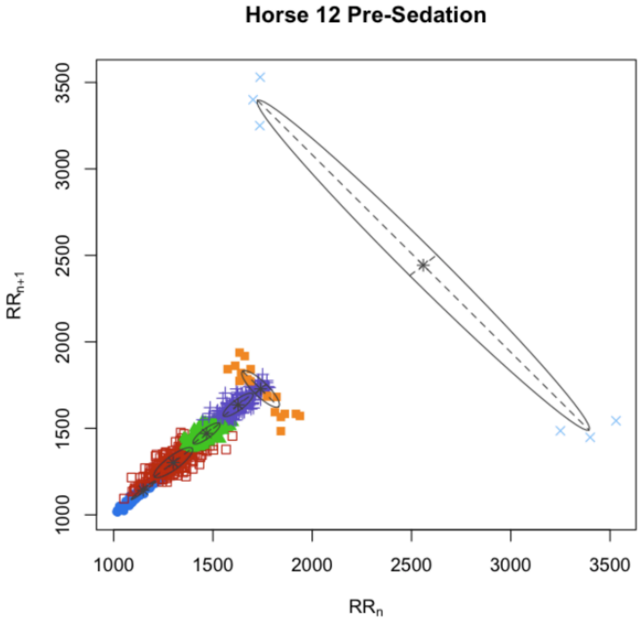


Figure 6: Clustering Assignment for Horse 12 Pre-Sedation period

In this research, we tried three methods for AV block removal. The first method is simply removing the points that are three standard deviations larger than the mean RR interval. The second method includes the usage of the Autoregressive Integrated Moving Average (ARIMA) model. The model is first used to fit the data; then standardized residuals are calculated; and points with standardized residuals that have values three standard deviations larger than the mean residuals are removed. However, neither of the two methods is successful for outlier detection in some windows. Horse 12 stands out the most among all. As Figure 8 demonstrates, for Horse 12 before endotoxin period, a number of outliers away from the diagonal line are still present even if both methods are applied. Note that in all the approaches, a sanity check that confirms whether the beat before and after has normal RR interval is applied to make sure the outlier identified is indeed an AV block rather than premature beats.

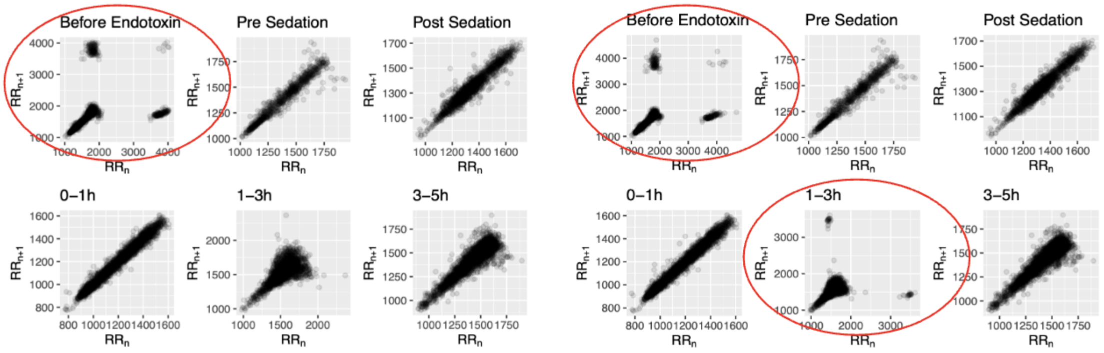


Figure 7: Horse 12 Poincare Plot after 1st & 2nd Approaches

To address the outlier issue above, we introduce the third approach, 1000 threshold removal. We first calculate the change of RR interval (delta RR) for each beat, plot delta RR histograms for each time window, and then determine a threshold which if delta RR exceeds this value, the corresponding RR interval value would be considered an outlier. As Figure 8 demonstrates, we construct the delta RR histogram for each horse and each time period (only horse 12 is included here but all are checked). We determine that 1000 is a reasonable threshold and any point with delta RR greater than 1000 will be considered as outliers and removed. In addition, manual deletion is applied after the delta RR approach to refine our dataset.

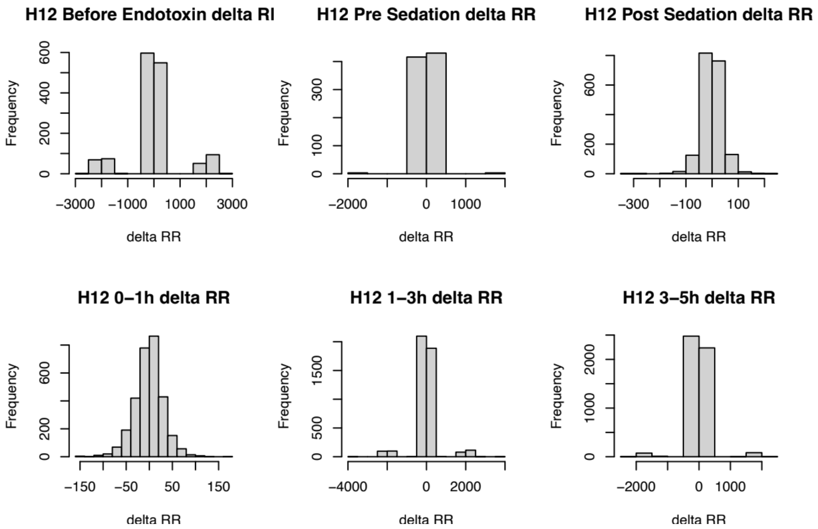
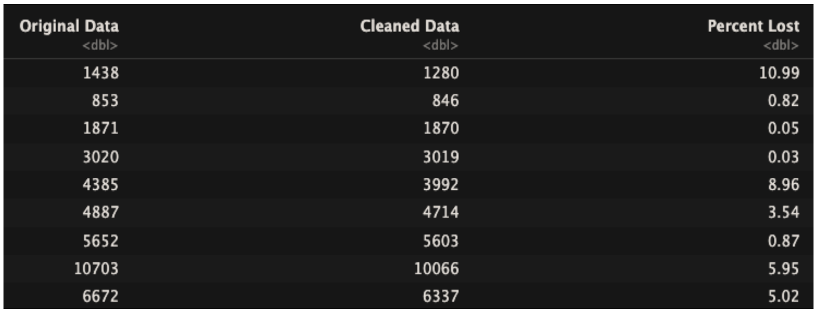


Figure 8: Horse 12 Delta RR Histogram

It is reasonable to propose the concern on whether the above approach is removing too many points. To address this problem, we have calculated percent of data loss after applying the 1000 threshold approach. From Figure 10, we observe that the percent of data loss is within the reasonable range as they are all less than 11%. Percent loss is also confirmed for other horses and all seem to be rational. Besides, we would still prefer type 1 error compared to failing to identify the AV blocks since AV blocks make such a huge difference in statistical analysis but smaller clinically.



Original Data <dbl>	Cleaned Data <dbl>	Percent Lost <dbl>
1438	1280	10.99
853	846	0.82
1871	1870	0.05
3020	3019	0.03
4385	3992	8.96
4887	4714	3.54
5652	5603	0.87
10703	10066	5.95
6672	6337	5.02

Table 2: Horse 12 Percent Data Loss after 1000 Threshold Approach

After removing values based on the threshold, the points now are scattered in a single cluster along the diagonal in all windows of all horses. With only one cluster in each window now, advanced analysis about dispersion and shapes can be applied.

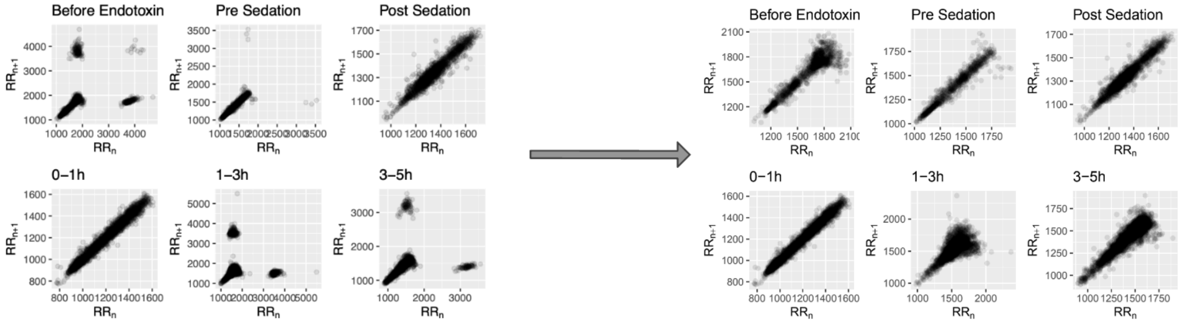


Figure 9: Horse 12 after Data Cleaning

To further quantify the poincare plots for each horse at each period of time, we compute the following summary statistics. Some basic ones are mean, variance, min and max of the RR interval values. We are also using SD1 and SD2. SD1 and SD2 are calculated with the formula below, where SDSD is the standard deviation of the successive difference of RR interval and SDRR is the standard deviation of RR interval.

$$SD_1^2 = \frac{1}{2}SDSD^2$$

$$SD_2^2 = 2SDRR^2 - \frac{1}{2}SDSD^2$$

In terms of ECG data, SD1 would correspond to short-term variability and SD2 would correspond to long-term variability for RR values. We choose SD1 and SD2 because they reflect the shape of the main elliptical cluster on the Poincare plot.

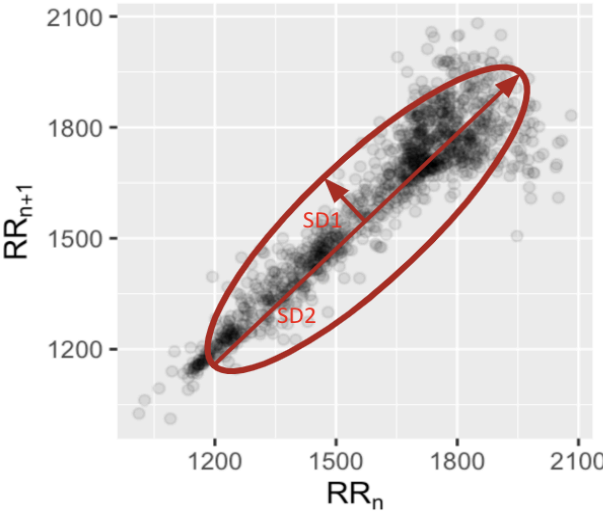


Figure 10: Poincaré plot with SD1 and SD2

If we only have a main cluster along the diagonal and we regard it as an ellipse, then SD1 would reflect the length of the minor axis and SD2 would reflect the length of the major axis.

This brings another reason for AV blocks removal. If the poincare plot contains AV blocks that are not on the diagonal, this would greatly affect the precision of SD1 and SD2 calculation.

The number of Gaussian Components assigned is also used to measure the dispersion of the main clusters. In Gaussian mixture models, we assume the distribution is a mixture of K Gaussian components. If the main diagonal is assigned with more clusters, it means that the distribution is more dispersed and complicated.

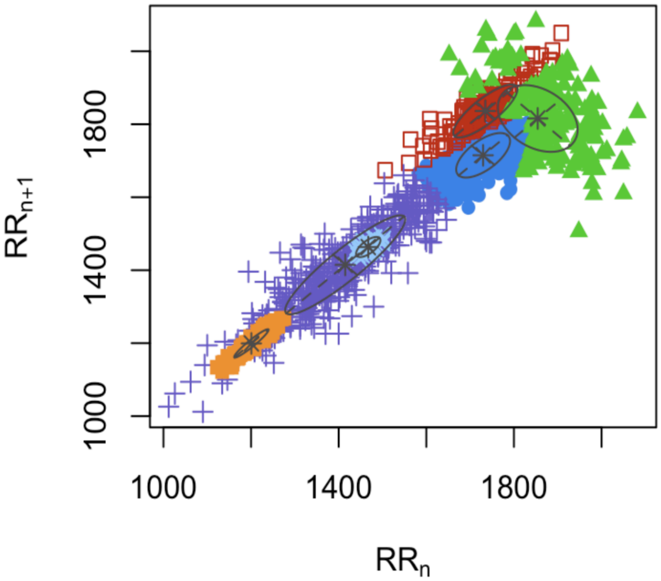
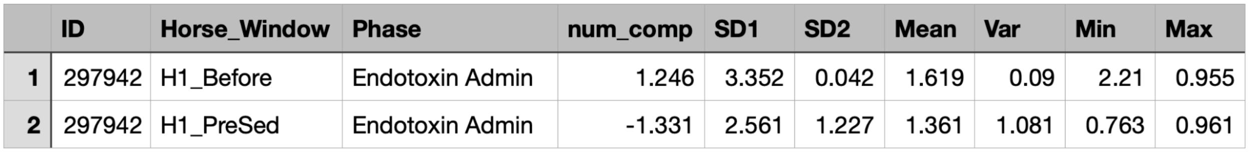


Figure 11: Poincare plot with GMM components (G=6)

After computing these summary statistics for every horse and every time period, we compile them into a table used for further analysis. We also standardize the data to make them centered at 0 with standard deviation of 1. Below is a snapshot of the data we used for PCA, etc.



	ID	Horse_Window	Phase	num_comp	SD1	SD2	Mean	Var	Min	Max
1	297942	H1_Before	Endotoxin Admin	1.246	3.352	0.042	1.619	0.09	2.21	0.955
2	297942	H1_PreSed	Endotoxin Admin	-1.331	2.561	1.227	1.361	1.081	0.763	0.961

Table 3: Snapshot of Summary Statistics used for PCA

Results:

In the summary statistics table of the RR data for each horse at each time window, there are 7 variables we want to take into account to explain the differences across different phases of cardiac arrhythmia. Using PCA, we essentially reduce the number of variables to two, but still preserve the ability to account for 66.6% percent of the total variance in the table.

From the PCA biplot, where the points are colored by the phases of cardiac arrhythmia, we are able to identify some patterns in how the principal components are related to variables in data. The variables that are the most associated with the first principal component (PC1) are the maximum and SD1, as indicated by the two arrows that are almost parallel to the PC1 axis. These two variables are themselves positively correlated, and maximum contributes more to change in PC1 compared to SD1 as it has a longer arrow in the horizontal direction. Increases in maximum and SD1 are associated with lower PC1 values.

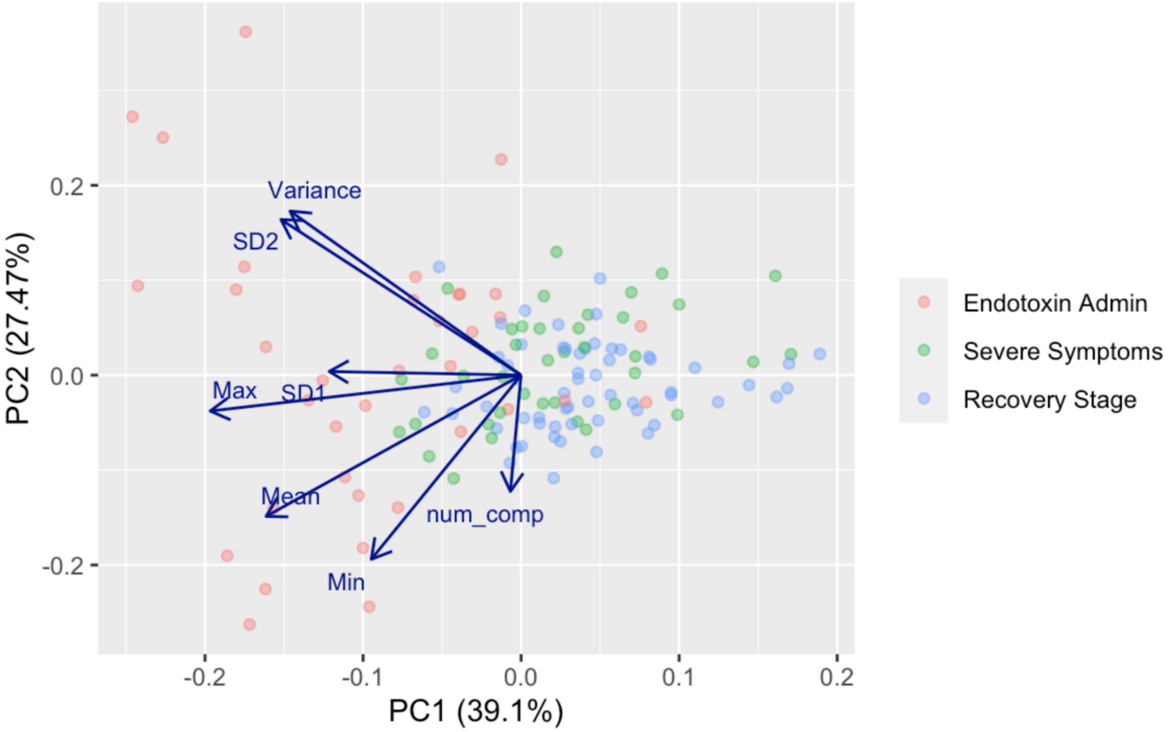


Figure 12: PCA Biplot using RR summary statistics, colored by cardiac arrhythmia phase

One observation from the distribution of the points on the PCA biplot is that the points corresponding to the Endotoxin Admin phase are mostly located to the left of the plot, which at the same time correspond to smaller or negative PC1 values. Combining this with the previous paragraph, we can learn that higher maximum and SD1 values could be a good indicator of a data point in the Endotoxin Admin phase.

The variable with the most contribution to the second principal component (PC2) is the number of components assigned by the Gaussian mixture model. The more number of Gaussian components assigned to the main cluster in a time window, the smaller the PC2 value is. Another method we use to locate trends in the RR summary statistics table is hierarchical clustering with complete linkage. First of all, a distance matrix is computed, which includes the pairwise Euclidean distance between all pairs of data points from the table. Then, a dendrogram is created using the distance matrix. Points that are closer together are more likely to be located on the same branch in the dendrogram. The dendrogram is cut at a distance of 8, which gives three clusters. The hierarchical cluster assignment is used as a categorical variable for each data point in the RR summary statistics table. We then use mosaic plots and chi-squared tests to check if there are associations between cluster assignment and cardiac arrhythmia phases or individual horses.

The method we apply next to visualize if cluster assignment is dependent with the phase or horse is the mosaic plot. The main idea behind a mosaic plot is the conditional probabilities of specific combinations of variables. The null hypothesis is that the two variables are independent. A block being colored blue indicates that this combination of variables occurs more than we would expect under the null hypothesis where they are independent. A red block means that the combination of variables occurs less than expected.

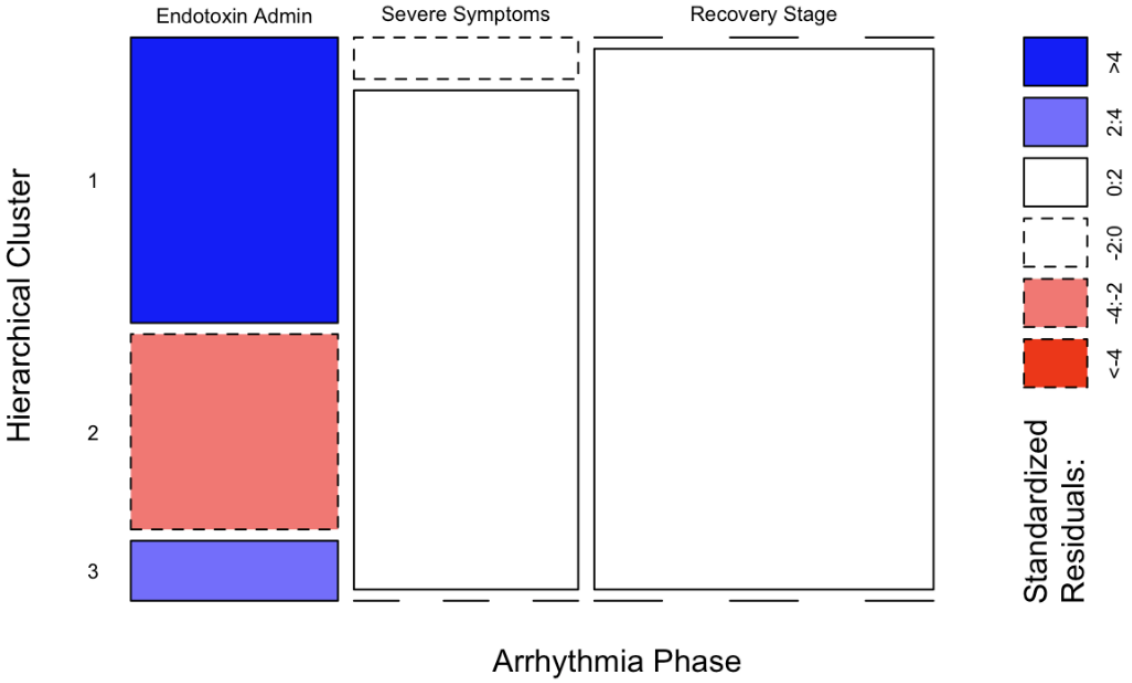


Figure 13: Mosaic Plot of Hierarchical Clustering Assignment vs Cardiac Arrhythmia Phases

The mosaic plot of the hierarchical cluster against phases of cardiac arrhythmia shows that there is dependence between the two variables. The intersections between clusters 1 and 3 and the Endotoxin Admin phase are colored blue. This evidence suggests that we see many more observations in the Endotoxin Admin phase and being assigned to cluster 1 or 3 than what we expected given their individual proportions in the summary statistics table. In other words, if a window is assigned to cluster 1 or 3 using our method of analysis, then it has a much higher probability of being from the Endotoxin Admin phase than the other two phases.

We do not see significant differences between Severe Symptoms and Recovery Stage. This is not surprising, given horses are continuously and gradually feeling better from the 0-1h window on. There is also evidence from the mean RR trend plot above (Figure 4) and the PCA plot point colors (Figure 12) that the differences between the last two phases are much more subtle compared to the Endotoxin Admin phase.

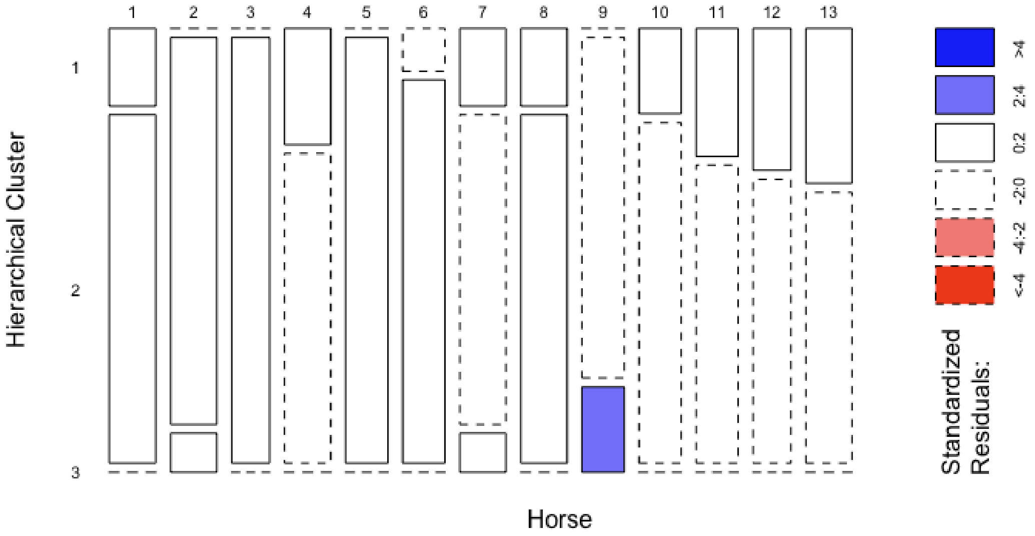


Figure 14: Mosaic plot of hierarchical clustering assignment against horses

The mosaic plot of hierarchical clusters against horses shows that they are almost independent. Most parts of the mosaic plot are not colored, which is evidence that most combinations of the variables are neither more nor less than expected. The intersection of Horse 9 and cluster 3 seems to stand out with more proportions than expected, but it is hard to draw any further conclusions from this observation alone. This mosaic plot also allows us to compare how each horse has different distributions of cluster assignments. There are significant differences between some horses. For example, all windows of horse 5 are assigned to cluster 2, and horses 10 to 13 all have relatively large proportions of windows in cluster 1. Nevertheless, there is not a general trend across horses.

Discussion:

Using our current approach, we can effectively distinguish between the windows of a horse before and after showing severe cardiac arrhythmia symptoms. By computing the seven summary statistics of the window, we can include this window as a row in our table of summary statistics along with all other data that we have. If the hierarchical cluster assigned to the window is 1 or 3, then we know it has a high probability of showing no cardiac arrhythmia symptoms or being injected with endotoxin.

The PCA gives us insight into positive associations between a window's maximum and SD1 to being in the Endotoxin Admin phase. The PCA biplot contains more information on the interactions and relationships between the RR summary statistics. These relationships can be further explored in future research.

In the future, we would want to focus on detecting differences between the severe symptoms phase and the recovery phase. To better explore this idea, we could change our approach to just focus on the experiment data starting from 0-1 hours to the end to see the impact of the time stages on the horse's health after the endotoxin injection. Another approach to consider would be incorporating more horse information, such as age, weight, sex, breed, etc., to further explore variability among the horses and see if these factors play a role in the detection of cardiac arrhythmias. On the other hand, increasing the sample size of horses would allow for this research to be more generalizable.

Acknowledgements:

We express our gratitude to Professor Ron Yurko for his invaluable guidance as our Project Advisor, whose expert insights significantly enriched our research approach and analysis in detecting horse cardiac arrhythmias using ECG data. We are also immensely thankful to Dr. Katharyn Mitchell, our client, who provided us with this dataset and gave us practical perspectives that have been instrumental to our project. Their contributions were vital to the success of our work.

Work Cited:

"Cardiology: Cornell University Equine Hospital." Cornell University College of Veterinary Medicine, September 12, 2022.https://www.vet.cornell.edu/hospitals/services/cardiology-0.

Karmakar, Chandan K, et al. "Complex Correlation Measure: A Novel Descriptor for Poincaré Plot - Biomedical Engineering Online." BioMed
Central, BioMed Central, 13 Aug. 2009.

KJ;, Mitchell. "Equine Electrocardiography." The Veterinary Clinics of North America. Equine Practice, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/30871826/. Accessed 3 May 2024.

Zhang L;Guo T;Xi B;Fan Y;Wang K;Bi J;Wang Y; "Automatic Recognition of Cardiac Arrhythmias Based on the Geometric Patterns of Poincaré Plots." Physiological Measurement, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/25582837/. Accessed 3 May 2024.