

Introduction

Cancer has proven to be an ever-growing, omnipresent disease which touches the lives of countless people around the world. In the United States, there are around 15 million cancer survivors, 1.65 million new cancer diagnoses, and around 590,000 cancer related deaths every year. Despite this massive number of impacted lives, up until now it has been impossible to retrieve patients' data to actually help influence treatment options. Our Cancer Medical Record Time-Slicer uses patients' data to come up with innovative treatment approaches based on their past history, and on the history of similar patients.

Motivation

All4Cure seeks to aggregate knowledge across patients, their caregivers, clinicians, and researchers. We hope that our Capstone Project can help realize a day where every cancer patients' experience adds to an ever growing body of knowledge which progressively improves our understanding and treatment of cancer.

Technical Approach

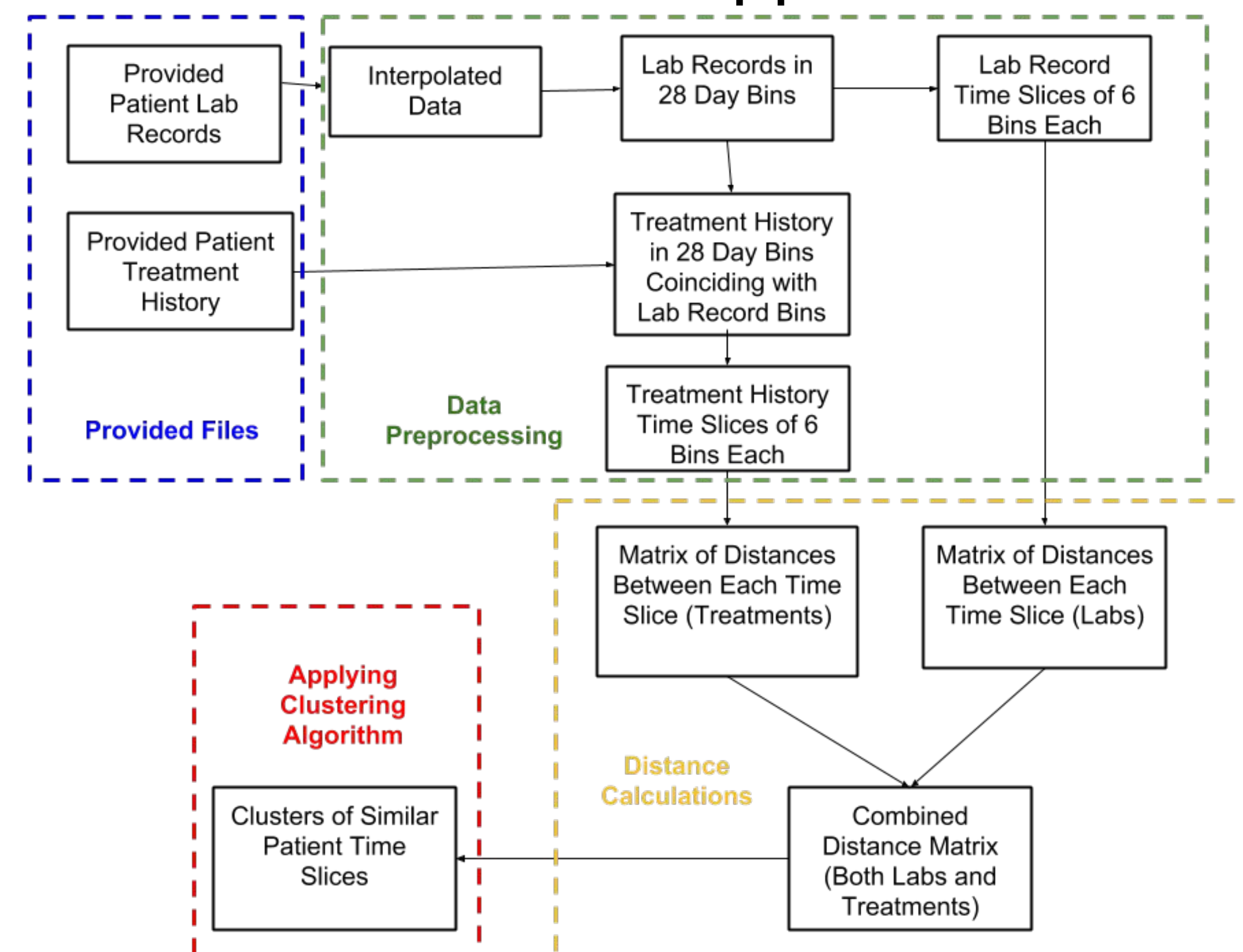


Figure 1. Block Diagram for the entirety of the project

- Interpolate patient data to ensure each time-slice has a matching number of evenly spaced entries
- Normalize the data in one of several ways: Min-Max, Z-Score, and Unit-Norm
- Calculate the correlation based on either: Pearson, Spearman, or Kendall-Tau, then normalize this matrix
- Use one-hot binary encoding for treatments and calculate the distance matrix using the Jaccard distance calculation
- Linearly combine both distance matrices into a combined normalized distance matrix
- Run clustering algorithm on the combined distance matrix

Presentation of Results

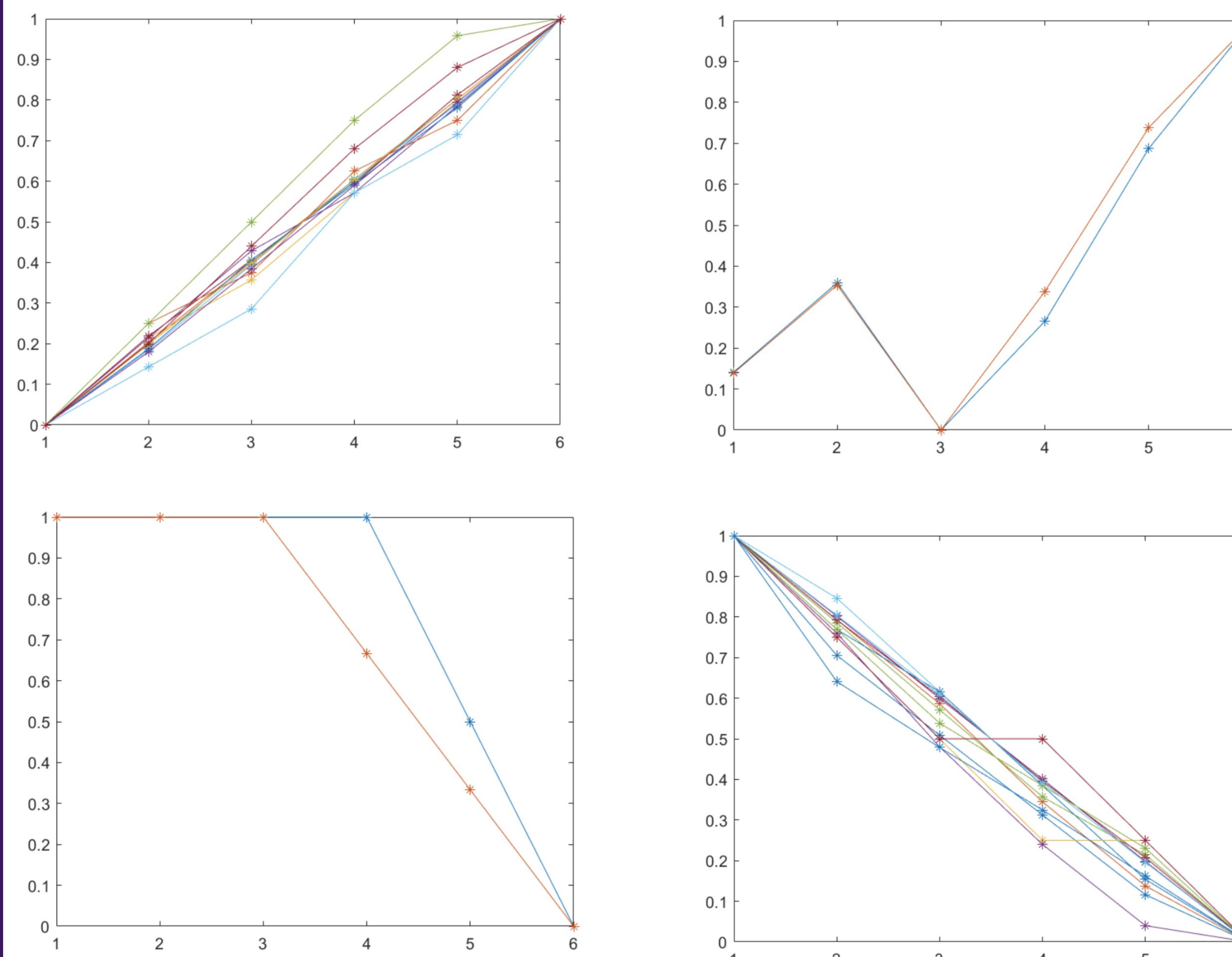


Figure 2. Four of the normalized clusters returned from the FLC clustering method. In total, on a dataset of 50 time-slices, 9 non-noise clusters were made.

ID	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
MM-2 65.0	0	0	2, 4, 7	2, 4, 7	2, 4, 7	2, 4, 7
MM-4 0.0	0	2, 4, 7	2, 4, -1, 7	2, 4, 7	2, 4, 7	2, 4, 7
MM-2 67.0	0	-1	1, 2, 4, 7	1, 2, 4, 7	2, 4, -1, 7	1, 2, 4, -1
MM-3. 0	0	1, 2, 4	1, 2, 4	2, 4	3, 4, 5, -1	3, 4, 5, -1

ID	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
MM-4 0.7	5	1, 3, 4	3, 4, -1	3, 4, -1	3, 4, -1	3, 4, -1
MM-4 0.8	3, 4	3, 4	3, 4, -1	3, 4, -1	3, 4	3, 4, -1
MM-4 0.9	3, 4	3, 4, -1	9, 4, -1, 15	9, 4, -1, 15	9, 4, 15	9, 4, -1, 15
MM-4 0.10	9, 4, 15	9, 4, 15	9, 4, -1, 15	9, 4, 15	9, 4, 15	9, 4, 15
MM-4 0.11	9, 4, 15	9, 2, 4, 15	2, 4, 9, 15, -1	9, 2, 4, 15	9, 2, 4, 15	2, 4, 9, 15, -1
MM-4 0.12	9, 2, 4, 15	9, 2, 4, 15	9, 2, 4, 15	2, 4, 9, 15, -1	9, 2, 4, 15	9, 2, 4, 15

Figure 3. Two of the clusters returned from the treatment clustering method. Each treatment is encoded to an integer. On a dataset of 50 time-slices, 5 non-noise clusters were made.

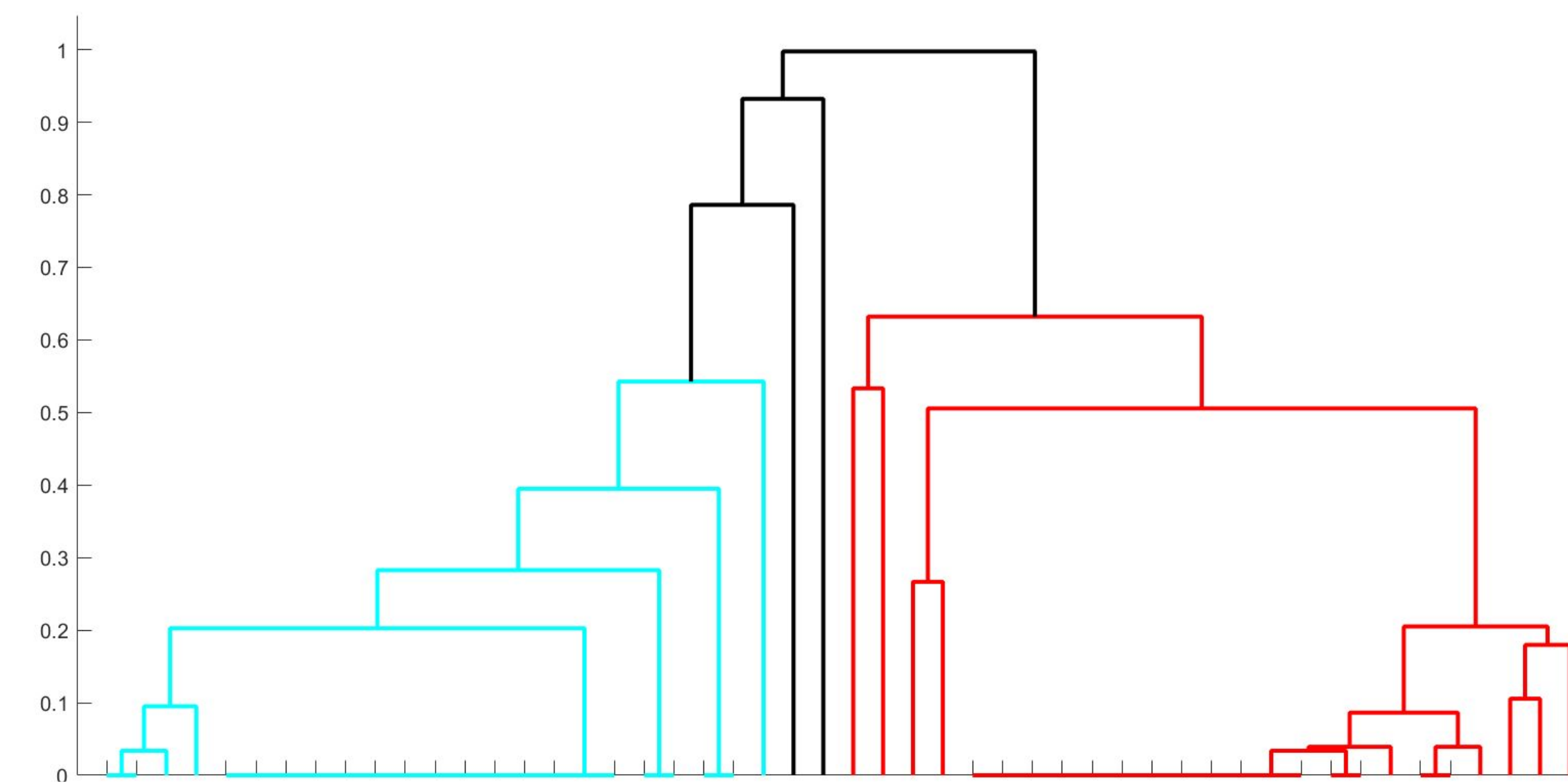


Figure 4. Visual representation of clustering called a dendrogram. Made through another clustering method where each time-slice starts as its own cluster, and as the algorithm progresses up the graph, distances are spanned between each cluster to connect it with its next nearest cluster. Used for visualizing the distances between clusters.

Discussion of Results

Visually, the lab clusters have been verified to be correct. After manually sorting each of the patients into clusters, we found very few differences between our visual clustering and our programs' clustering. Our observations are the following:

- A few dominant trends in the time-slices: There are two clusters that the majority of time-slices fall into.
 1. A straight upward trend
 2. A straight downward trend

The treatment clustering seems equally promising. When reviewing the results with help from those more knowledgeable in myeloma, there were very few abnormally placed time-slices. Our observations are the following:

- A few different common myeloma drug combinations formed dominant clusters. The use of common drug combinations is a phenomenon we had discussed beforehand and hoped to capture in our clustering algorithm, so we were pleased to see these clusters form.
- A large cluster forms from all the time-slices for which no treatment is taken. This is expected as well.

Future Steps

One improvement that we are working on implementing is the ability to fit a prospective patient into pre-existing clusters. Using this ability in conjunction with the current dashboard will allow doctors to easily recommend treatments based on observing similar time-slices to the patient's history up to this point.

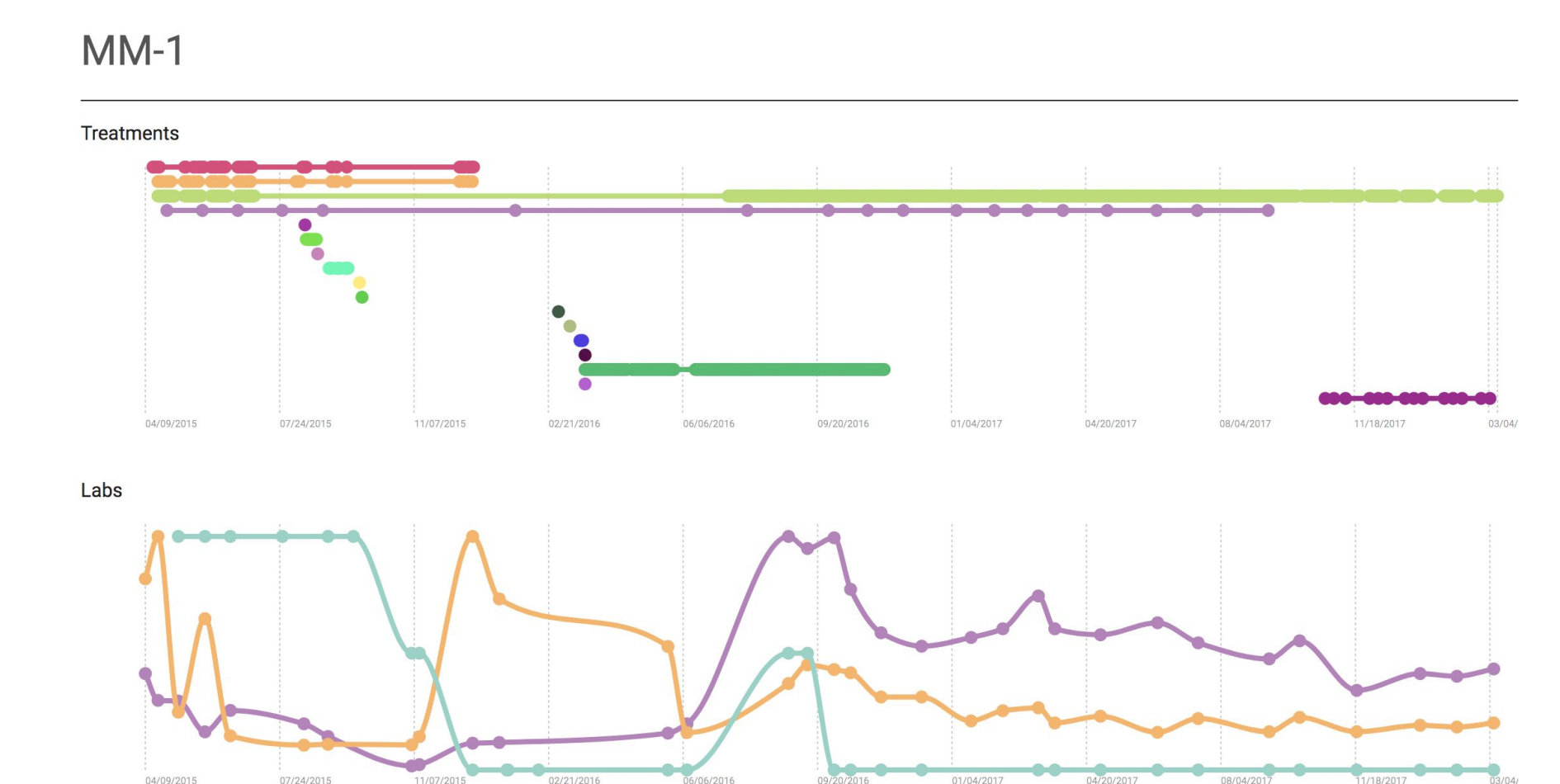


Figure 5. The current All4Cure dashboard

Conclusion

Our research has successfully proven that we can draw meaningful results from clustering patients' data based on their FLC value and treatment history. Our program has paved the way for All4Cure to implement these findings in a variety of different ways. Depending on the circumstances, the end-user may decide to substitute in a different clustering algorithm, or normalize the data in a different way. Not only that, but in the future, our program can also help medical professionals decide on the best course of action for the treatment of a patient, based on their clustering and the treatment history of successful patients with a similar background.