Clustering Methods for Finding Insights in PRO-CTCAE Data

Anya Greenberg, Thomas Hanson, Ethan Otto, Austin Varela
Group 3, DSC 383W
Goergen Institute for Data Science, University of Rochester

1 Introduction

As medical treatment has gotten more advanced, life expectancies have continued to rise resulting in a larger aged population needing medical care. With age comes both a greater risk of developing and dying from cancer (Ershler, 2003; Yancik & Ries, 2004) with roughly 60% of all cancers being reported by the older adult population. Compounding this issue is the disparity in outcomes between younger and older patients partially due to the older population's lower representation in research concerning cancer treatment and care standards. Due to this discrepancy, close to 60% of geriatric patients develop significant toxicities after starting treatment, additionally various age-related conditions were significantly associated with this toxicity (Mohile). These older patients also experienced a high interference with daily living and quality of life, which they generally consider a higher priority than survival rate (Mohile). These geriatric patients are both at a higher risk, and have so far been less prioritized in research causing difficulty in creation of treatment plans and higher prevalence of negative outcomes.
By better understanding this unique collection of patients and how their existing conditions interact with both their advanced stage cancer and the treatments, we can help advise those treatment plans to hopefully reduce the patient's toxicities and negative outcomes.

The Patient Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) is a measure of patient symptoms as experienced by the patient themselves rather than by a clinician report. This criteria is being evaluated in a UR NCORP network (Mohile R01CA177592funds URCC 13059/"GAP") study attempting to reduce treatment toxicities by using a geriatric assessment model (Mohile). In addition to the PRO-CTCAE, the clinician-reported version of the CTCAE and the Karnofsky Performance Status (KPS) are also being evaluated. The KPS allows physicians to classify the functional impairment of a patient. The NCORP study is comprised of roughly 700 patients over 70 years of age, with age related conditions, with advanced stage cancers and hopes to further explore the relatively novel insights that can be gained by evaluating the patient reported outcomes. Specifically, the study's aim is to determine whether there is a strong link between the PRO-CTCAE data and the tolerability of treatments in these geriatric patients. The goal of our work was mainly exploratory in nature, aiming to begin understanding and gaining information from the PRO-CTCAE data using unsupervised learning methods to cluster patients and their symptoms.

Unsupervised learning is a subset of machine learning, defined as methods that seek to discover novel patterns in data without existing labels or direct supervision. This is in contrast to many common machine learning methods which train a model using labeled training data to predict a class label (i.e. supervised learning for classification). Some of the most common unsupervised learning methods are clustering methods, which aim to group data points by commonality between shared attributes. How this is done is largely dependent on the data, type of algorithm, and distance metric used. Due to the imprecision of defining a "cluster", there are a wide variety of clustering algorithms which can have significantly different results based on their assumptions. However there is no one best algorithm, and generally the different algorithms will behave differently on different datasets. For example: connectivity-based, or hierarchical, clustering works by linking nearby data points together into clusters. It returns a dendrogram of the data showing each cluster merge, which can be useful itself but it is very slow and vulnerable to outliers. There are also centroid based methods which seek to define a cluster by a central point (which may be generated rather than actually existing in the data), which tend to run very quickly but can fail due to assumptions of cluster shape (circular as a data point will be assigned to it's nearest centroid), and focuses on optimizing cluster centers rather than cluster borders. Use of these methods on the PRO-CTCAE data can help generate groupings of patients and symptoms to better understand the symptoms and toxicities experienced by these patients, as well as to help inform care plans.

2 Data Set Description

The PRO-CTCAE data gathered from the control arm of the GAP70+ study has been digitized and pre-processed before being received by our team. Specifically, the digitized data has been converted to a CSV file with missing fields in the original data marked with the "." symbol. The data contains 1,210 rows; each row indicating a completed PRO-CTCAE assessment for a given patient at a particular time point. Each patient can have up to four separate entries, where each of these entries corresponds to a scheduled assessment which may be given at one of four time points: 1) initial screening, 2) 4-6 week follow-up, 3) 3 month follow-up, and 4) 6 month follow-up. These time points are consecutive, such that no patient can have a given time point without all previous time points being present (ex. all patients with an entry for the 6 month follow-up will have one for their initial screening, 4-6 week follow-up, and 3-month follow-up). The PRO-CTCAE data also contains 52 columns. These columns consist of a randomly assigned, de-identified patient identifier, the assessment time point, the "KPS" score, and the general category of the patient's cancer. The remaining 48 columns represent the symptoms' properties (severity, frequency, and degree of interference with daily living). It should be noted that while cancer type is a data field, it may have less relevance in the studied population than in most other oncology patient populations because all patients in the study are 70 years of age or older and have advanced-stage cancer (which may tend to have more homogenized symptoms). Yet, the strong bias towards gastro-intestinal (GI) and lung cancer types (fig 1) mean that the potential minor differences between cancer types may lead to an over-representation of GI and lung cancer specific adverse events in the data set. Additionally, all symptomatic fields and KPS are ordinal, with set ranges of 0-4 and 0-100 (by 10), respectively. The KPS field can be converted into fewer ordinal categories based on the accepted classification of scores (0-40 indicating inability to care for one's self; 50-70 indicating inability to work; 80-100 indicating normal ability to carry on daily activities). Since KPS is created based on a physician's evaluation of patient symptoms, it is not an independent data field - it is instead highly correlated to symptomatic data fields and may be considered more of an outcome.

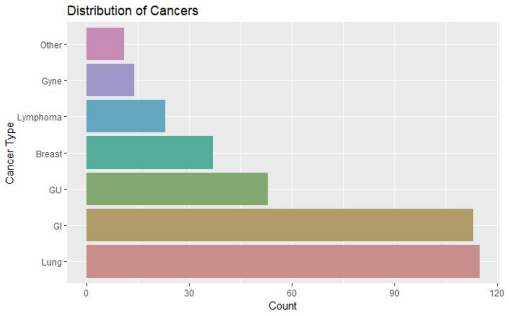

Figure 1: *The distribution of cancer types among individual patients.*

Since the PRO-CTCAE data is part of a larger study, each patient also has other attributes associated with them (dubbed as "outcome" variables). These attributes provide a useful set of data to help contextualize findings derived from the PRO-CTCAE data. One set of outcome data is based on clinician completed evaluations of each patient at all assessments the patient attended and these data are ordinal in nature (with the exception of the binary outcome of CalcMinicog). Such evaluations rate the patients' ability to: perform daily activities (ADL and IADL), move and walk (SPPBTotalSumCalc), and perform cognitive tasks (CalcMinicog and CalcMinicogScore). The other set of outcome data provided to us relates to binary variables indicating: whether the patient had a high level (grade 3-5) of toxicity measured at some point and whether the patient had to be hospitalized over the course of their treatments.

2.1 Missing Data

As with many other real-world data sets, there are a few cases in the PRO-CTCAE data set where missing data is present. The cases where missing data is present, however, is non-random. Instead, specific clinical context determines the reason behind why the data is missing. There are two specific clinical reasons for data points being missing. Firstly, the PRO-CTCAE form is contextual - meaning that certain questions are only asked if there is clinical relevance for doing so. Additionally, since PRO-CTCAE is a patient self-reporting form, it's possible that a patient may refuse to answer a question. So for any given patient, there may be a non-random feature that does not contain a value. Fortunately, only six rows of the data contain missing values. Of which, only one of the rows took place on a screening assessment. This is particularly relevant when considering the other type of missing data present. While ideally, all patients have four individual assessments present in the data, there are numerous cases where patients do not have all assessments present. Overall, there is a seemingly linear trend downward with respect to the total number of patients assessed at each of the four time points (fig 2). The loss of a significant portion of data over time makes it difficult to produce longitudinal findings across all time points. Therefore, as the screening assessment contains the largest number of patients and only a single missing data point, most analyses were performed on just screening data. As a result, analyses did not focus on longitudinal results across all time points. Screening findings, however, do have the benefit of acting as a potential diagnostic tool for physicians.

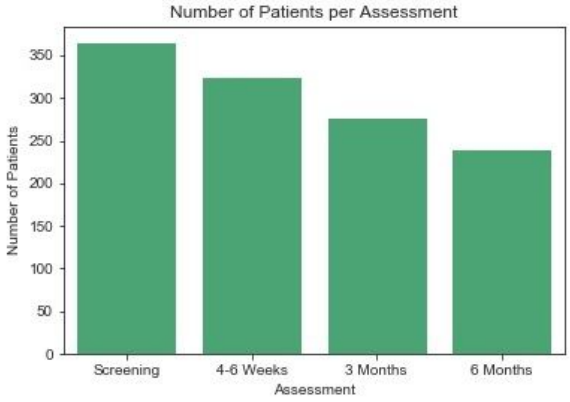

Figure 2: *There is a consistent decrease in patients over time. With 366 individuals present at the screening, 326 at the 4-6 week follow-up, 277 at the 3 month follow-up, and 241 at the 6 month follow-up.*

2.2 Imputation

Since there is only one case with missing data at the screening assessment, imputation would likely not affect results very strongly, so for most exploratory analyses in the early stages of the project, rows containing missing data were simply dropped. However, as there is a large number of features, with a relatively few number of data points--imputation was used in cluster-based analyses. The instance where missing data occurred was a GI cancer patient without a hand-foot syndrome severity score. Since the median and mode value of hand-foot syndrome severity among all other patients at the screening time point is 0, the missing value was replaced with a severity score of 0 (little to no instances of hand-foot syndrome). Additionally, over-sampling methods were considered to increase the occurrence of minority types of cancer. Specifically, the SMOTE method of minority over sampling was evaluated (Chawla et al, 2002), however it was found to overfit models severely so oversampled or bootstrapped data sets were not used.

3 Exploratory Analysis

3.1 Occurrence of Symptoms Across Cancer Type

Since there is a large number of features, it is important to understand the relationships across features to uncover potential data biases and to just simply get a better "feel" for the data to be able to make informed decisions regarding handling of the data. While it is expected that the fact that the patients are all older individuals with advanced stage cancer would make differences between cancer types relatively non-apparent, it is still important to consider the slight differences among cancer populations. An effective method to do this is to simply generate a heatmap of relative occurrences of variables. "Occurrence" is defined as the number of times a specific feature had a non-zero data point among the data. This transformation is utilized as it is difficult to compare patient-reported symptom data across different individuals and perception of a specific symptom's effects can differ from person-to-person. Additionally, this analysis was only performed on variables indicating either a symptom's frequency or severity as the interference of daily living appeared in no non-redundant symptom classes. Since there is an imbalance in cancer types, occurrences were normalized such that they represented the proportion of patients with a given symptom present relative to the cancer type of the patient. This allows for occurrences to be able to be compared across cancer types more easily. Generally, the occurrence of symptoms appear to be fairly constant across cancer types on the screening as expected (figs 3a-b). However there are a handful of clear outliers. Generally, the "other" cancer tends to be the most different cancer type - but no meaningful generalizations can be derived as it both has the fewest number of patients of all cancers (fig 1), but it also is a combination of several minority cancer types. Regardless, in the heat map from frequency variables (fig 3a) it can be seen that there is a spike in occurence of urination control issues in gynecological cancer compared to other cancers. And there appears to be a relative lack of nausea among genitourinary cancer patients compared to other cancers. There are a few other outliers of note among the severity variables (fig 3b). Specifically, lung cancer patients are more likely to have shortness of breath issues and lymphoma appears to be the least likely to cause fatigue issues.

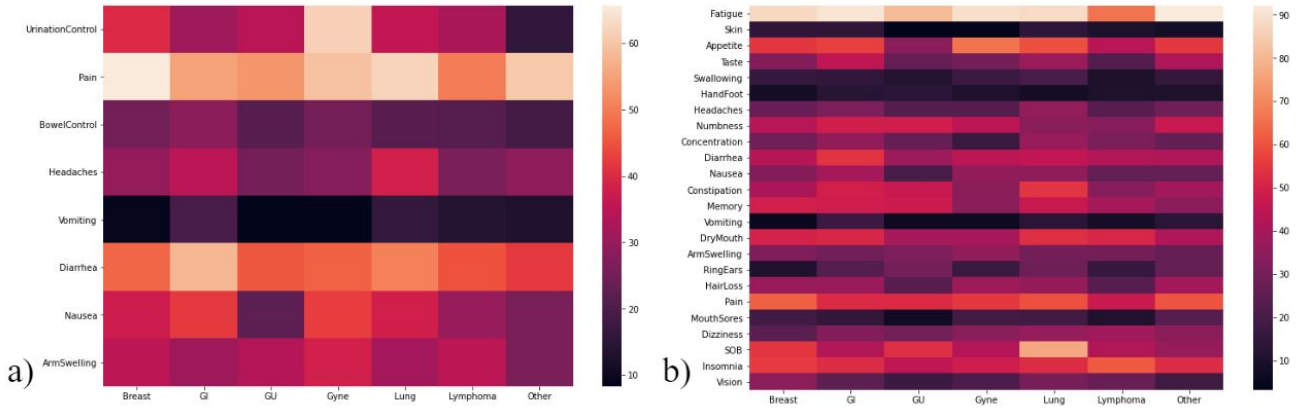

Figure 3: *Brighter boxes indicate that the symptom (indicated by the y-axis) was present in a large number of patients with the cancer type indicated by the x-axis. (a) The relative percentage of occurrences of each symptom at the screening assessment according to "frequency" variables across cancer types. (b) The relative percentage of occurrences of each symptom at the screening assessment according to "severity" variables across cancer types.*

A similar exploratory analysis can be performed to determine the change in occurrence of these variables across cancer types. By comparing the difference in occurrence from the screening assessment and the final, 6 month follow-up, it is possible to potentially understand the relative rate of development

of these symptoms over 6 months. Similar to the analyses done on the screening assessment, it is possible to visualize the changes as a heat map (figs 4a-b). There are a few more hot spots able to be seen than in the heat maps of screening symptoms. For example, among the frequency variables (fig 4a):

- There is a large increase in urination control issues over time among lymphoma patients, with a large decrease in urination control issues over time among breast cancer patients.
- There is a decrease in pain among gastro-intestinal cancer patients.
- There is a decrease in nausea and arm swelling in gynecological cancer patients.

And among the severity variables (fig 4b):

- There is an increase in numbness among gastro-intestinal cancer patients.
- There is a decrease in appetite issues among gastro-intestinal, genitourinary, and gynecological cancers.
- There is a decrease in ringing ears among breast and gynecological cancer patients.
- And there is a decrease in pain issues in breast and gastro-intestinal cancer patients.

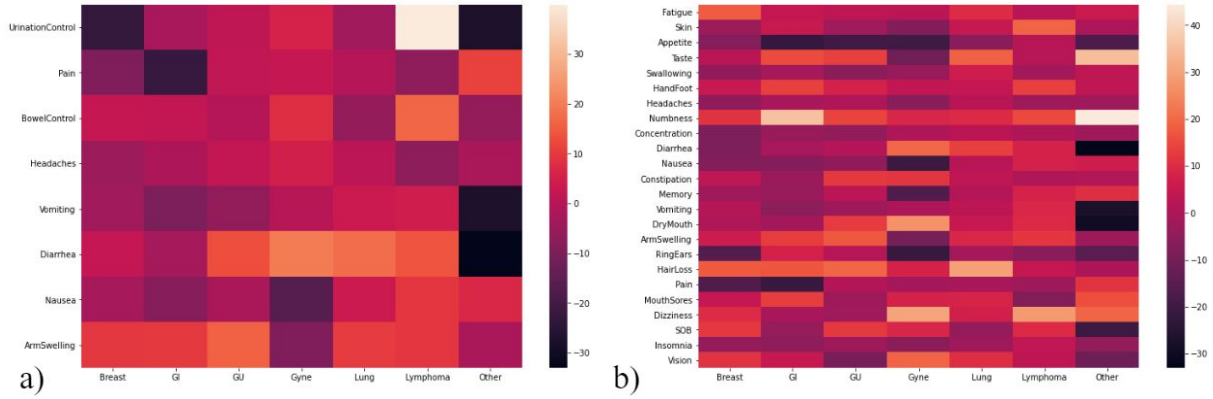

Figure 4: *Brighter boxes indicate that the symptom (indicated by the y-axis) was present in a large number of patients with the cancer type indicated by the x-axis. (a) The relative change percentage of occurrences of each symptom from the screening assessment to the 6 month follow-up assessment according to "frequency" variables across cancer types. (b) The relative change percentage of occurrences of each symptom from the screening assessment to the 6 month follow-up assessment according to "severity" variables across cancer types.*

3.2 Correlation

While it is important to understand how cancer types affect the distribution of symptomatic variables, it is also clear that while most symptomatic features are uniformly distributed among cancer types, these features are likely not independent of each other. For example, having high frequency and severity of a symptom like pain would logically correlate to having a high interference with dailing living from pain. In hopes to verify such a claim, pairwise spearman rank correlations were performed on all symptom features to all other symptom features (fig 5). Frequency, severity, and interference features of symptoms were highly correlated to each other (something that is expected). But there are a few cases of unrelated symptom features tending to correlate to each other. Most of these also make sense logically; several examples include correlations between: concentration and memory, taste and appetite, and nausea and vomiting. This could prove useful in potential future feature selection as these sets of features may have redundancy. The major unexpected correlation, or rather lack thereof, is between KPS and all other individual symptoms. This may indicate that KPS cannot be well predicted from patient reported symptom data.

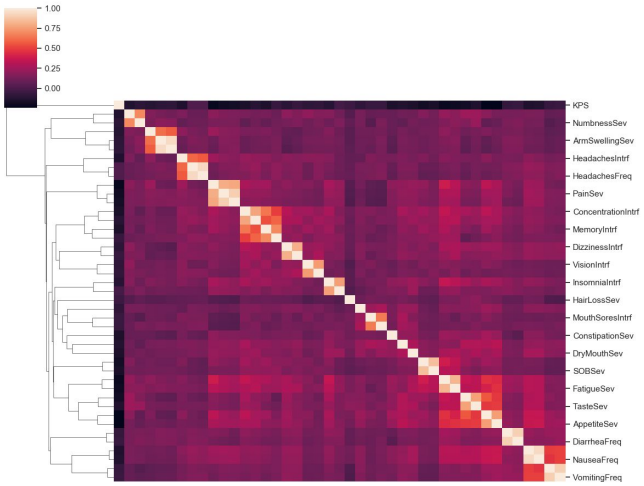

Figure 5: *Spearman's rank correlation map of all symptom variables against all other variables on the screening assessment.*

3.3 Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction method which transforms high dimensional data to a lower dimension through the use of the covariance matrix and its eigenvalues and eigenvectors. The bases of these eigenvectors are called principal components and contain individual dimensions of the data which are uncorrelated. This procedure seeks to represent the greatest variance in the data with fewer dimensions. By reducing the dimensions of a dataset, natural clusters may become more apparent in visualizations of the transformed data.

As such, we applied PCA to the PRO-CTCAE data and plotted the second principal component against the first, coloring each data point according to cancer type. As shown in figure 6, no natural clusters in the data were realized and the different cancer types are scattered throughout the plot when performing PCA on the PRO-CTCAE screening data. A similar pattern was seen at all time points and thus it can be confirmed that the symptoms do not cluster around cancer type.

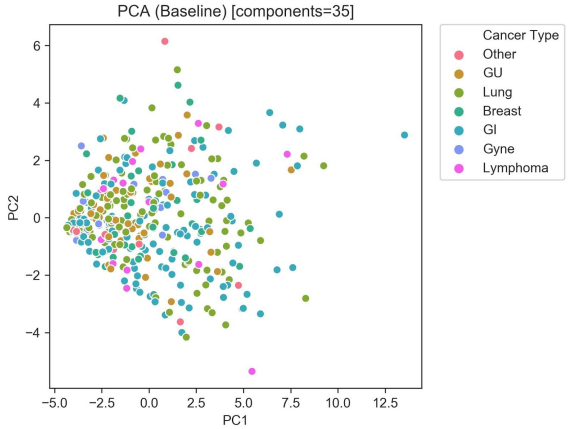

Figure 6: *Scatter plot of second principal component against first principal component of PCA on screening (baseline) PRO-CTCAE data using 35 components. Each color represents a different canner type.*

3.3 Association Rules

In order to understand the relations between symptoms and cancer types, association rules were developed using the Frequent Pattern Growth (FPGrowth) algorithm, the methods of which are explained below (sec.4.3). More than 4000 rules were generated with a minimum support of 10 and a confidence of 80%. Most notably, we found that high appetite severity and low concentration severity when coupled with low insomnia severity or low dizziness severity were strong indicators of GI cancer.

Association Rules Sample: Support = 10 | Confidence > 0.8		
Symptom Set	**Implies**	**Confidence**
('DizzinessSev_L', 'SkinSev_L')	VisionSev_L	0.83
('AppetiteSev_H', 'InsomniaSev_L', 'ConcentrationSev_L')	cancertype_GI	0.85
('AppetiteSev_H', 'DizzinessSev_L', 'ConcentrationSev_L')	cancertype_GI	0.86

Table 1: *Table showing selected results from FPGrowth association analysis on data.*

3.4 Summary

In this section, we did an exploratory analysis of the PRO-CTCAE data and obtained some key findings. When analyzing the occurrence of symptoms across cancer type, we found that there were some relations between the prevalence of a symptom and a specific cancer type. For example, a higher proportion of lung cancer patients have some degree of shortness of breath than patients with other cancers. From these heatmaps (fig 3-4), it can also be seen that some symptoms such as pain and fatigue were prevalent across all cancer types while others such as vomiting and hand-foot syndrome had very few occurrences in the dataset as a whole. Additionally, we found some insights into how these symptoms may change over the course of 6 months. Through Spearman's rank correlation, we found that the severity, frequency, and interference with daily living of a symptom were all highly correlated (as expected) and discovered that some symptoms were correlated, as well, such as concentration and memory, taste and appetite, and nausea and vomiting. Lastly, the PCA plot (fig 6) supports our assumption that symptoms do not naturally cluster around cancer type likely due to the properties of the study participant population.

4 Model development

4.1 Clustering

As different clustering algorithms perform better on different data and are solved by different methods, we thought it best to construct a collection of clusterings using different algorithms. The algorithms used include KMeans, Affinity Propagation, Spectral Clustering, Ward Hierarchical Clustering, and Mean Shift. All of these produced different clusterings as expected, however the mean shift algorithm's clusters were extremely weighted such that almost all patients were within one cluster, with the others having as little as 1 patient. This likely meant that the algorithm found outliers more than clusters, which was not applicable to our current work so it was not explored further. Most of these methods required us to supply a number of clusters to generate, which based on suggestions from our sponsor team was generally set to be four or less. All clustering methods implemented with the scikit learn package (Scikit-Learn).

KMeans is a centroid based clustering method, it attempts to form clusters by first selecting points to be starting centroids then repeatedly assigning all data points to their closest centroid and updating the centroid by taking the mean of each point in the cluster. Formally, the KMeans algorithm seeks to solve the following problem: Given observations x_n and sets S_k,

$\arg\min_s \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 = \arg\min_s \sum_{i=1}^{k} |S_i| var(S_i)$ where μ_i is the mean of points in cluster S_i. This is equivalent to minimizing the squared deviations of points within a cluster: $arg_s\, min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x,y \in S_i} ||x - y||^2$. The result of this method can be heavily influenced by the start position, so it is common practice to run it multiple times with randomized start locations. Another key limitation is the assumptions made in its cluster model, namely it assumes spherical clusters of equal size. The algorithm is good for general purpose use, and best with few clusters.

Affinity Propagation is another centroid based algorithm, whose goal is to find "exemplars", or data points that are good representatives of their clusters. Unlike KMeans, it does not require a number of clusters to be specified beforehand, and makes few assumptions about internal cluster structure. Assignment of exemplars is done using a responsibility matrix **R**, where *r(i, k)* is how well suited x_k is to be the exemplar for x_i relative to other candidate exemplars, and an availability matrix **A** where $a(i,k)$ is how appropriate it would be for x_i to select x_k as it's exemplar. The algorithm first fills the **R** matrix, using a similarity function $s(i,k)$: $r(i,k) = s(i,k) - \max_{k' \neq k}[a(i,k') + s(i,k')]$. The **A** matrix is then updated by $a(i,k) = \min\{0, r(k,k) + \sum_{i' \in \{i,k\}} \max(0, r(i',k))\}$ for $i \neq k$ and $a(k,k) = \sum_{i' \neq k} \max(0, r(i',k))$. The algorithm then iterates until cluster boundaries remain unchanged across iterations or a maximum number of iterations. The implementation used from the sklearn library also utilizes a damping factor to prevent large oscillations during the update phase. With this damping factor λ $r_{t+1}(i,k) = \lambda \cdot r_t(i,k) + (1-\lambda) \cdot r_{t+1}(i,k)$; $a_{t+1}(i,k) = \lambda \cdot a_t(i,k) + (1-\lambda) \cdot a_{t+1}(i,k)$, where t is the current iteration number. As the algorithm does not take as input a number of clusters, the tuning of hyperparameters was done to create the desired number of clusters. The only main drawback of affinity propagation is it's runtime, which is on the order of $O(N^2T)$. The algorithm is applicable to situations with a variable number of clusters, uneven cluster size and a non-flat geometry.

Spectral Clustering is distinct from the first two methods as it is not centroid based, but instead uses the eigenvalues of a similarity matrix to project the data into a lower dimensional space to cluster more efficiently. The similarity matrix **A**, where a(i,k) is a measure of similarity between points i and k, is the input for the model along with a number of clusters to create. First the algorithm calculates the laplacian of **A**, which is defined as $L = D - A$, where D is the diagonal matrix $D_{ii} = \sum_j A_{ij}$. It then calculates the eigenvalues of the laplacian. An eigenvalue being the scalar value λ which satisfies the following equation for a corresponding non 0 eigenvector x of A: $Ax = \lambda x$. These eigenvalues then relate to a connected graphical representation of the data, and by evaluating the first non-zero eigenvalues the algorithm can find connections to 'cut', thus separating the data into the desired number of clusters. An eigenvalue of 0 represents unconnected components, and thus a value close to zero indicates a connection that is almost not there. The algorithm is mostly used for situations with a low number of even sized clusters on a non-flat geometry.

Ward Hierarchical Clustering was the final method used, which is an agglomerative hierarchical clustering using Ward's method as the distance metric. Agglomerative hierarchical clustering starts by assuming each datapoint is it's own cluster, then iteratively computes the distances between each cluster and merges two clusters. In our case, we used Ward's method for our distance metric, where for the distance between points i and j: $d(i,j) = ||X_i - X_j||^2$, which is equal to the squared euclidean distance. The Ward method also relates to the linking criteria, or how the clusters are merged. In standard single linkage, the closest two clusters are joined, for Ward instead we minimize the sum of squared differences

within all clusters and to join seek to minimize the variance in the generated clusters. Using Ward's method helps one of the common drawbacks of agglomerative clustering where single linkage will tend to create very uneven cluster sizes. Other drawbacks include the runtime, which tends to be $O(n^3)$, and the algorithm in general is not guaranteed to give the optimal clustering.

4.2 Subspace Clustering

Subspace clustering is a field of clustering algorithms that build reduced dimension representations of the data before applying a clustering method to the data. A popular subcategory of subspace clustering is sparse subspace clustering (SSC) algorithms which we applied to the sparse PRO-CTCAE dataset.

The full description of SSC can be found here (Vidal), but we'll provide a brief explanation of what it does. A SSC algorithm assumes that the data was generated from linear subspaces of the features, then clusters the points based on how close points are within the subspaces. This has the benefit of not clustering the data on features that a datapoint doesn't use in a subspace. Therefore, features that aren't likely part of a subspace for a datapoint, aren't involved in the cluster assignment of that point. Now that we have the subspace representations, the algorithm calculates an affinity matrix and clusters the data. The clustering algorithm itself is very similar to spectral clustering, mentioned previously.

We implemented variants of SSC: Orthogonal Matching Pursuit (OMP) and Elastic Net Subspace Clustering (EnSC). Detailed descriptions of both can be found here (Vidal). The main difference is that OMP is faster than typical SSCs and EnSC is faster and can be adjusted to handle noise better than OMP.

Overall, the SSC methods didn't appear to cluster much better than Ward Hierarchical Clustering or KMeans. This likely means the assumption of linear subspaces was fundamentally incorrect. This might also relate back to our previous dimensionality reduction attempts during the EDA where we didn't notice any clear clustering along the PCA components. This provides further evidence that the data doesn't cluster well, even when we try to find subspaces of features that cluster rather than clustering the whole dataset. Unfortunately, it's very difficult to interpret the subspaces and the python packages we deployed don't offer us any analytical capabilities.

4.3 Association Rules Analysis

An alternative lens to view relationships between data points is through association rules. Association rules are clear "if X then Y" conditionals where X and Y are certain feature values. For instance, X could be having abdominal pain and scored 4 on a GI test and Y might have GI cancer. Given a dataset we can construct association rules for the given likelihood of the rule being true given X (the confidence), and the number of times we see an example of this rule holding (the support). The idea behind rules is that we may discover rules that could have diagnostic or evaluative purposes. For instance, symptoms → cancer type, and cancer type → symptoms.

Association rule building is roughly n^2, where n is the number of distinct values of all features. Therefore, most datasets contain far too many column values to generate association. However, for the PRO-CTCAE dataset, by labeling severity symptoms as either high (>2) or low (1<=2), we were able to build rules that were constructed in a few minutes. We generated several excel sheets of rules for a confidence of 80% for various supports.

5 Outcome Evaluation and Performance

The main mission of this project was to discover novel clusters in the dataset and relate them to some explainable phenomena. However, due to the multidimensionality of the data, and the lack of clear separable clusters, it's not entirely clear what many of the clusters represent. We've speculated some of

the clusters might capture an indicator of a patient's overall wellbeing at the end of the study. Conveniently, at the end of the study a number of outcome metrics were recorded measuring a patient's overall mental and physical health on certain criteria. Hence, we wanted to develop models to measure the association and compare the predictive power between the clusters and the outcome features. These outcomes were: Activities of Daily Living (ADL), Instrumental Activities of Daily Living (IADL), Short Physical Performance Battery (SPPBTotalSumcalc), Mini-Cog Impairment (CalcMinicog), Mini-Cog Score (CalcminicogScore), High_Toxcity, Hospitalization, stayed_until_end, and KPS. A description of each outcome can be found in the "Dictionary.xlsx" document. We should also note that KPS is not technically an outcome variable as it was recorded during the baseline survey, but might yield some insights regardless.

The structure we've established for evaluating these associations is a four-step process. First perform an association test between the clusters and outcomes. Then, fit outcome cluster pairs that pass to a linear model and record metrics. Afterward, perform an association test between outcomes and the symptoms. Finally, we regress each outcome on its significant symptoms and record metrics.

We could have stopped after the first two steps since we now have metrics measuring how well clusters predicted outcomes. However, the metrics would be even more insightful if we had an idea of what the best possible predictive model was given the data set. As mentioned before, clustering is an unsupervised learning model. Hence, the clustering algorithms were not told to find correlations with the outcomes, just to find sections of the symptom data that appeared similar. On the other hand, supervised learning algorithms would build models by training on outcomes and symptoms to predict outcomes using symptoms. This can give us a ceiling on how good our clusters could be, since the classification accuracy of a supervised learning model should always beat that of an unsupervised model. Furthermore, directly training models on the outcomes and symptoms allows us to see how much outcome predictive power exists in the PRO-CTCAE dataset.

The third step, finding associations between the symptoms and the outcomes, has the additional benefit of establishing which symptoms are related to which outcomes. It's possible that we have outcomes that don't correspond to any of the symptoms or outcomes that only correspond to a few. This information could be used to launch future exploratory research in these associations.

5.1 Binary Outcomes

We settled on using a chi-square association test to look for associations for each pair of binary outcomes and set of clusters. This involves multiple repeated tests so we adjusted the p-values using the Benjamini-Hochberg procedure to avoid p-hacking. We have the following binary outcomes: CalcMinicog, High_Toxcity, Hospitalization, and , stayed_until_end. Only the Hospitalization outcome had any clusters with a p-value less than 0.1. The table with those cluster hospitalization pairs are displayed in Table 2. We selected the two pairs Hospitalization-KMeans and Hospitalization-Ward to build logistic regression models.

The metric area-under-curve (AUC) measures how well the model is able to distinguish between the classes. It uses the true positive rate, and false positive rate rather than the raw predictions of the model. This is important since most of these models will always predict the majority class, making accuracy-based metrics impractical.

5.2 Ordinal Outcomes

For ordinal outcomes we chose an F test as our association test. The F test assumes the outcomes are normally distributed so we applied a boxcox transform to the outcomes. None of the outcomes pass any normality tests with most scoring a p-value around 0.01. We then corrected the F test p-values with

Benjamini-Hochberg. The only pairs that passed were SPPBTotalSumcalc with Spectral and Ward clustering as seen in Table 2. The SPPBTotalSumcalc went through a boxcox transform to make the outcome distribution as normal as possible. The regression R^2 for both pairs are both under 0.035 and can be found in Table 4. This suggests that the clusters can't explain 4% of the variance present in the outcomes. Therefore, the clusters are associated with SPPBTotalSumcalc but don't represent it.

Outcome Clusters Trend Test Results										
	Clusters									
Outcomes	af	sc	km	ms	wd	SSC-OMP	LSSC	EnSC-50	EnSC-1	EnSC-0.1
ADL	0.000	0.000	0.000	0.094	0.000	0.708	0.011	0.223	0.920	0.742
IADL	0.000	0.000	0.000	0.285	0.000	0.690	0.362	0.181	0.935	0.483
SPPBTotalSumcalc	0.118	0.052	0.040	0.381	0.035	0.394	0.385	0.442	0.383	0.661
CalcMinicogScore	0.882	0.661	0.468	0.468	0.764	0.495	0.708	0.661	0.639	0.468
CalcMinicog	0.797	0.385	0.681	0.442	0.753	0.495	0.661	0.383	0.701	0.681
High_Toxcity	0.383	0.699	0.494	0.825	0.708	0.448	0.699	0.112	0.495	0.708
Hospitalization	0.112	0.163	0.001	0.591	0.001	0.925	0.120	0.183	0.218	0.825
stayed_until_end	0.044	0.385	0.001	0.764	0.094	0.920	0.472	0.102	0.825	0.472

Table 2: *The p values of the trend association tests on each pair of clusters and outcomes. The green highlighted values passed a false discovery rate of 10%.*

5.3 Outcome on Symptom Regression

5.3.1 Ordinal Association Tests

We needed to gather the symptoms that are associated with each outcome. Recall that the symptoms are on an ordinal scale as well as the outcomes. We also expect that when one symptom increases that an associated outcome would not decrease. Given this expected relationship, we chose two deploy two kinds of trend tests as our association tests. For binary outcomes we used a Cochran-Armitage test, and for ordinal outcomes we used Linear By Linear tests. Both have a null hypothesis that there is no association and an alternative hypothesis saying the outcome increases with the symptom. The main caveat is that it assumes each symptom's values are equally distributed, but this is not the case with our data.

We continued even though the data did not meet the equal interval assumption. Then we implemented the test for each pair and made the appropriate Benjamini-Hochberg correction. The p-value and 0.1 false discovery rate results are shown in Tables 3 and 4. The most notable observations from these tests is that toxicity is not associated with any symptoms, ADL and IADL are associated with almost all the symptoms, and both MiniCog outcomes are associated with the mental symptoms.

Outcome Symptom Trend Test Results								
	Outcomes							
Symptoms	ADL	IADL	SPPB-calc	Calc-Score	Calc-cog	Toxcity	Hospital	stayed-end
ArmSwellingSev	0.000	0.000	0.016	0.554	0.256	0.691	0.580	0.163
PainSev	0.000	0.000	0.003	0.029	0.019	0.115	0.011	0.012
HeadachesSev	0.001	0.000	0.444	0.078	0.087	0.622	0.451	0.831
NauseaSev	0.000	0.000	0.243	0.935	0.747	0.941	0.435	0.262
VomitingSev	0.000	0.003	0.831	0.935	0.729	0.946	0.811	0.302
DiarrheaSev	0.012	0.104	0.729	0.533	0.859	0.578	0.323	0.694
FatigueSev	0.000	0.000	0.000	0.369	0.477	0.632	0.026	0.114
AppetiteSev	0.000	0.000	0.072	0.445	0.743	0.605	0.298	0.036
NumbnessSev	0.000	0.000	0.184	0.788	0.666	0.961	0.666	0.883
VisionSev	0.000	0.000	0.089	0.182	0.379	0.617	0.558	0.243
SOBSev	0.012	0.005	0.217	0.666	0.636	0.569	0.014	0.691
InsomniaSev	0.000	0.001	0.823	0.998	0.734	0.666	0.927	0.905
TasteSev	0.000	0.000	0.147	0.541	0.465	0.823	0.166	0.009
DizzinessSev	0.000	0.000	0.490	0.167	0.339	0.823	0.169	0.849
MouthSoresSev	0.000	0.000	0.823	0.946	0.888	0.743	0.551	0.883
ConcentrationSev	0.000	0.000	0.578	0.024	0.061	0.743	0.167	0.533
MemorySev	0.000	0.000	0.184	0.001	0.005	0.157	0.022	0.682
ConstipationSev	0.043	0.029	0.642	0.927	0.666	0.750	0.416	0.027
SwallowingSev	0.000	0.000	0.927	0.849	0.946	0.323	0.026	0.215
DryMouthSev	0.000	0.000	0.193	0.732	0.402	0.134	0.031	0.005
HandFootSev	0.000	0.000	0.971	0.551	0.888	0.818	0.166	0.927
RingEarsSev	0.033	0.081	0.927	0.939	0.939	0.946	0.946	0.323
SkinSev	0.000	0.000	0.533	0.435	0.971	0.553	0.102	0.925
HairLossSev	0.006	0.004	0.961	0.691	0.969	0.927	0.734	0.849
KPS	0.000	0.000	0.000	0.146	0.025	0.722	0.018	0.000

Table 3: *The p values of the trend association tests on each pair of symptoms and outcomes. The green highlighted values passed a false discovery rate of 10%.*

5.3.2 Outcome Symptom Regression

Now that we have a list of symptoms associated with each outcome, we can regress each outcome on those symptoms. We decided to continue using logistic and linear regression models, but we experimented with more advanced gradient boosted decision tree models which yielded similar results. The R^2 and AUCs of each model were recorded in Table 5. To directly compare these results to the cluster results, the cluster column contains the score from the best performing cluster model. From this we can draw two conclusions. None of the supervised linear regression models are very predictive ([AUC ~ 0.6, R^2~0.3] suggest a little predictive power, but not much) . Secondly, that most direct symptom models exceed the cluster models with exception to the Hospitalization outcome models. This implies that whatever little information in the dataset can be used to predict hospitalization is being picked up on by our clusters.

Outcome Cluster Regression Results			
Outcome	Cluster	Metric	Value
Hospitalization	km	AUC	0.603
Hospitalization	wd	AUC	0.623
stayed_until_end	af	AUC	0.601
stayed_until_end	km	AUC	0.599
stayed_until_end	wd	AUC	0.578
ADL	af	R2	0.067
ADL	sc	R2	0.076
ADL	km	R2	0.075
ADL	ms	R2	0.042
ADL	wd	R2	0.066
ADL	LSSC	R2	0.029
IADL	af	R2	0.081
IADL	sc	R2	0.088
IADL	km	R2	0.097
IADL	wd	R2	0.098
SPPBTotalSumcalc	sc	R2	0.031
SPPBTotalSumcalc	km	R2	0.021
SPPBTotalSumcalc	wd	R2	0.034

Table 4: *The results of the linear/logistic regression on each of the outcomes on the clusters. The value is either an AUC metric for logistic regression or R^2 metric for linear regression*

Outcome Symptom Regression Results		
Outcome	Metric	Value
CalcMinicog	AUC	0.630
Hospitalization	AUC	0.671
stayed_until_end	AUC	0.651
ADL	R2	0.239
IADL	R2	0.239
SPPBTotalSumcalc	R2	0.115
CalcMinicogScore	R2	0.057

Table 5: *The results of the linear/logistic regression on each of the outcomes on their significant symptoms. The value is either an AUC metric for logistic regression or R^2 metric for linear regression*

Outcome Cluster Regression Comparison				
Outcome	Cluster	Metric	Best Cluster Value	Symptom Value
Hospitalization	wd	AUC	0.623	0.671
stayed_until_end	af	AUC	0.601	0.651
ADL	sc	R2	0.076	0.239
IADL	wd	R2	0.098	0.239
SPPBTotalSumcalc	wd	R2	0.034	0.115

Table 6: *The results of table 4 and 5 compared to each other. Clearly, the symptom regression outperforms cluster regression. However, the gap for Hospitalization and stayed_until_end is narrow. Impling that the cluster is capturing much of the symptoms' effect on outcome.*

5.3.3 Hospitalization Symptoms Regression

One of the benefits of using linear models is that the coefficients are typically interpretable. We thought these coefficients could provide some insight into what symptoms indicate future hospitalization. The coefficients for each symptom are listed in Table 7. The larger the value, the more that value contributes to the model's decision to predict that the patient will need hospitalization. According to the coefficients, the most important features are swallowing severity, skin severity, bowel control frequency

and shortness of breath severity. Also, an interesting observation is that fatigue severity contributes almost nothing to a hospitalization decision.

Hospitalization Coefs	
Symptom	Coef
PainSev	0.225
SOBSev	0.215
SwallowingSev	0.145
MemorySev	0.133
DryMouthSev	0.099
FatigueSev	0.049

Table 7: *The regression coefficients from the logistic regression of Hospitalization on the significant symptoms*

6 Conclusion and Next steps

The goal of this project was to discover whether the information collected from the PRO-CTCAE can inform treatment plans for geriatric oncology patients. In order to do this, we generated novel clusters in the dataset and evaluated them on outcome variables to see if there were any explainable relations between the two.

After receiving the pre-processed PRO-CTCAE data, we first conducted an exploratory data analysis, and found that the symptoms did not contain any natural clusters and did not cluster around cancer type. Then we developed 4 different clustering models (KMeans, affinity propagation, spectral clustering, and ward hierarchical clustering) and 2 variants of sparse subspace clustering (OMP and EnSC). After retrieving cluster labels from these methods, we received datasets on outcome variables encompassing the patients' physical and mental capabilities as well as whether they were hospitalized and whether they had any grade 3-5 toxicities. With this data, we evaluated our clusters based on the outcome data. Through this analysis, we found that the clusters generated by KMeans and Ward Hierarchical algorithms were picking up on signals in symptoms predicting hospitalization, but other outcome variables were not represented.

As this project was exploratory in nature, there are many avenues for future work. One of the challenges in our analysis was the ratio of dimensions to number of patients. For a dataset of ~50 features, we only had ~350 samples to analyze and cluster. We were also dealing with class imbalance in terms of cancer type and the majority of the symptom scores were non-normally distributed. These problems could be handled in two ways: 1) obtain more samples or 2) dimensionality reduction. We attempted SMOTE and bootstrapping which caused overfitting in our models, so obtaining more data points in new studies would be beneficial. While we also attempted PCA and subspace clustering with minimal insightful results, it would be interesting to explore both in more depth. We only made a slight foray into subspace clustering and did so without analytical capabilities. Perhaps more research into different subspace clustering methods and the time to develop code from scratch would provide new insights. Additionally, our sponsor team brought our attention to known subspaces within the data. Other studies (Reeve) have shown associations between certain clusters of symptoms; using these subsets may also provide new insights into how interpretations of PRO-CTCAE data may be applied in treatment plans.

7 References

Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

Ershler, W.B. (2003). Cancer: A disease of the elderly. *Journal of Supportive Oncology*, 1, 5–10

Hurria A, Dale W, Mooney M, Rowland JH, Ballman KV, Cohen HJ, Muss HB, Schilsky RL, Ferrell B, Extermann M, Schmader KE, Mohile SG, Cancer and Aging Research Group. Designing therapeutic clinical trials for older and frail adults with cancer: U13 conference recommendations. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. Aug 20 2014;32(24):2587-2594. PMCID:PMC4129504

Mohile, Supriya et al.. Proposal to evaluate association between PRO-CTCAE and toxicity in geriatric oncology treatments. Unpublished.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Vidal, René. "Sparse Subspace Clustering." *Vision Lab*, John Hopkins University,

Reeve, Bryce B et al. “Recommended patient-reported core set of symptoms to measure in adult cancer treatment trials.” *Journal of the National Cancer Institute* vol. 106,7 dju129. 8 Jul. 2014, doi:10.1093/jnci/dju129

Yancik, R., & Ries, L. (2004). Cancer in older persons: An international issue in an aging world. *Seminars in Oncology*, 31, 128–136.