Clustering Literature Review and Research

Tamunoebite Clement

December 7, 2025

1 Introduction

Clustering is an unsupervised machine learning technique used to group similar data points based on patterns in their attributes. It identifies natural structures within the data such that points within the same cluster exhibit greater similarity to one another than to those in different clusters. In this paper, we examine foundational concepts in clustering as well as recent applied research from leading journals, with a focus on studies relevant to CMS healthcare data and patient subgroup analysis.

2 Distances

One of the most important elements in clustering is the choice of the distance metric. Distance measures are the backbone of clustering algorithms, and understanding them is critical—they can significantly influence the outcome of the analysis. Selecting the appropriate distance metric can dramatically improve both the accuracy and the insights produced by the algorithm. In this literature review, we explore the most relevant literature on distance metrics, examine how similar studies have applied them, and consider how best to use them in the context of our own CMS Medicaid healthcare data.

2.1 Euclidean Distance

Euclidean distance is the most commonly used metric for clustering, especially in **continuous-scaled data**. It measures the straight-line distance between two points in an n-dimensional space—essentially the hypotenuse in the Pythagorean theorem:

$$\text{distance} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Since Euclidean distance is sensitive to scale, it is essential to standardize all features before applying it. Without standardization, features with larger ranges can dominate the distance calculation, skewing the clustering results.

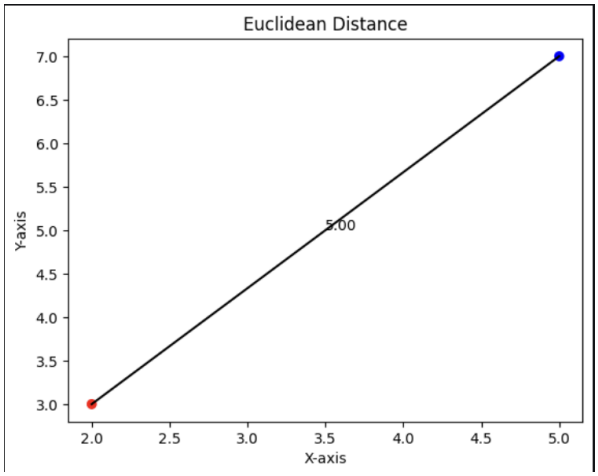

Figure 1: Euclidean Distance Graph

2.2 Manhattan Distance

Manhattan distance is the sum of the absolute differences between the Cartesian coordinates of two points. It is often visualized as the distance that one would travel along a city grid, moving only horizontally and vertically. This metric is well-suited for data with discrete variables, like the number of clothes in your closet. Moreover, Manhattan distance is more robust to outliers because it does not square outliers like Euclidean distance does.

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

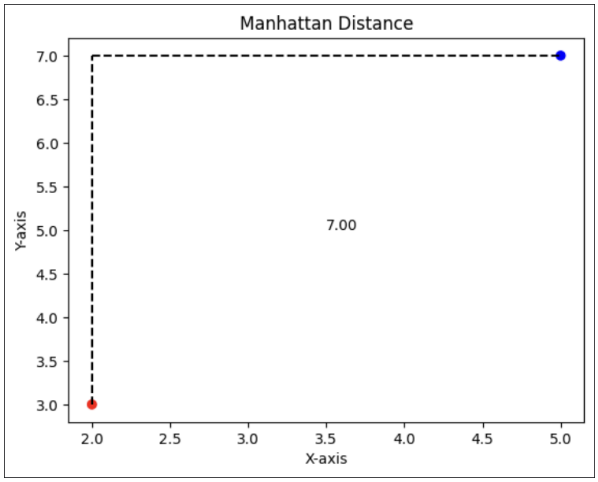

Figure 2: Manhattan Distance Graph

2.3 Minkowski Distance

Minkowski distance is useful when you want to generalize both Euclidean and Manhattan distances. It introduces a parameter p that allows you to control the form of the distance. When $p = 1$, it becomes **Manhattan distance**; when $p = 2$, it becomes **Euclidean distance**. Choosing values beyond $p = 2$ is typically done intentionally, often to emphasize larger differences between features or to fit specific domain needs. Minkowski distance is especially useful when you want flexibility in how similarity is measured across multiple dimensions.

$$D(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

2.4 Cosine Similarity

Cosine similarity measures the cosine of the angle between two data points, with higher values indicating greater similarity. It focuses on the **shape** or **direction** of the feature vectors, rather than their **magnitude**. This makes it ideal when you're interested in comparing patterns rather than scale. Unlike Euclidean distance, cosine similarity doesn't give much weight to how large the features are—only how they vary in relation to each other.

$$\text{Cosine Similarity} = \cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

For example, imagine you and your friend are using a dating app. You swipe right on many profiles, and your friend swipes right on only a few. Cosine similarity won't be influenced by how many people each of you swiped on. Instead, it focuses on which specific profiles you both liked—capturing *similarity in preference*, not quantity.

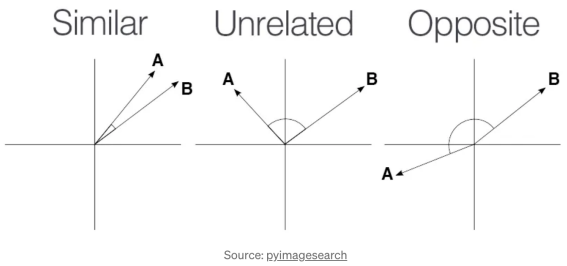

Figure 3: Cosine Similarity Graph

2.5 Jaccard similarity

Jaccard Similarity measures the similarity between two sets by dividing the size of their intersection by the size of their union:

$$\text{Jaccard Similarity} = \frac{|A \cap B|}{|A \cup B|}$$

It is primarily used for **binary features** (1 or 0), where 1 indicates presence and 0 indicates absence.

For example, suppose you want to find which Spotify users are most similar based on the songs they've listened to. You could construct a similarity index by counting how many songs *both users have listened to* (1's in common). Importantly, mutual 0's are ignored—songs that neither user has listened to—since Spotify contains millions of tracks, and most users haven't listened to most of them.

Jaccard similarity focuses only on the *positive overlap* in behavior, making it especially useful for sparse datasets.

2.6 Gower Distance

Gower distance is a similarity metric designed to measure dissimilarity between data points that include a mix of numerical and categorical variables—specifically continuous, ordinal, nominal, and binary features. Unlike traditional metrics such as Euclidean distance, which are limited to continuous data, Gower distance is well-suited for datasets with mixed data types. It standardizes each feature to a [0, 1] range before computing the overall distance, allowing all variable types to contribute meaningfully

to the comparison. Gower distance is especially useful in clustering applications where preserving the heterogeneity of features is essential. The general formula is:

$$D_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} \cdot d_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

where: D_{ij} is the Gower distance between observations i and j
p is the total number of features (variables)
d_{ijk} is the partial distance for feature k between i and j

$$w_{ijk} = \begin{cases} 1 & \text{if both } x_{ik} \text{ and } x_{jk} \text{ are non-missing} \\ 0 & \text{otherwise} \end{cases}$$

Click here to learn more about Gower distance.

3 Overview of our Dataset

When it comes to clustering and machine learning in general, knowing the data types of your features is critical when choosing a distance method. Below, our a features for our healthcare dataset and the common preprocessing techniques that we can use for them:

Feature Type	**Examples**	**Preprocessing**
Continuous	Total Medicaid Paid Amount IP, Total Medicaid Paid Amount ER, Age	Standardization (e.g., z-score)
Discrete (Counts)	Number of Conditions, Number of ER visits, Total IP Claim, Enrollment Months	Can standardize or leave raw (depends on range)
Ordinal Categorical	Age Group, Year	Label encoding or map to ordered integers
Nominal Categorical	Race, Sex, Eligibility Group	One-hot encoding
Binary/Flags	POA, Tech, Transplant	No transformation or treat as 0/1
Identifiers	Patient ID, FIPS	Drop (not used in clustering)

Table 1: Feature Types, Examples, and Pre-processing Approaches

Demographic variables like race and gender are not meant to drive clustering or prediction. Their role is interpretive—tools for stratification, adjustment, and context—not mechanisms for segmentation. Used as direct inputs, they risk masking more meaningful clinical signals, introducing noise instead of clarity. The focus should remain on the variables that truly reflect severity and utilization: the number of ER visits, inpatient stays, the sequence of claims over time, and the distinction between chronic and acute conditions. These features speak directly to need, burden, and system engagement. They form the backbone of any attempt to define subgroups based on medical complexity.

Clustering methods must match the structure of the data. Whether using K-means, hierarchical models, or other unsupervised approaches, feature types can't be treated carelessly. Continuous measures like cost or frequency must be standardized; categorical features like eligibility or age group must be encoded appropriately. Distances must be chosen with precision—Euclidean or Manhattan for continuous data, Gower for mixed types. Mixing variable types without transformation doesn't just reduce rigor—it introduces distortion. If the goal is to surface meaningful clusters, the process must be deliberate. Every method used should be interrogated, compared, and validated—not just for technical performance, but for how well it reveals the actual patterns in the lives and care of these children.

Cost distributions in the dataset reveal a pronounced imbalance driven by extreme cases. Inpatient and ER Medicaid costs show clear signs of distortion—distributions pulled sharply by extreme values. The median inpatient cost sits at $24,000, while the mean climbs past $52,000. For ER, the median

holds at just over $2,000, but the mean drifts higher to $3,000. The outliers push those means upward with force: inpatient costs reach as high as $2.4 million, and ER costs peak around $39,000. These are not statistical flukes—they signal a distinct subgroup, a cluster of high-cost, high-need "superusers" who reshape the entire landscape.

Statistic	Inpatient	Emergency Room
Count	1,589	1,455
Sum	$83,913,080	$4,430,904
Mean	$52,808	$3,045.30
Median	$24,415.38	$2,012.10
Max	$2,375,675	$38,810.94
Min (Non-Zero)	$17.89	$2.63

Figure 4: Descriptive Statistics Table

This structural imbalance has real implications for clustering. Algorithms like hierarchical clustering with Euclidean distance are sensitive to such extremes, risking distortion unless carefully standardized. In contrast, methods like K-medoids or distance metrics like Manhattan or Gower offer stronger resistance to skewed data. Recognizing the presence of this superuser group isn't optional—it's essential. Without accounting for it, any grouping risks missing the very patterns that define the problem.

4 Literature Review

In this project, the goal is to identify clinically meaningful subgroups among high-cost, high-need children with complex conditions enrolled in Medicaid. We are not clustering for abstraction—we are searching for patterns tied to real-world differences in care utilization, clinical severity, and potential intervention pathways. As such, our literature review focuses on studies that apply unsupervised learning methods, particularly clustering, to identify subpopulations within heterogeneous patient cohorts. Priority is given to work that uses real clinical or claims data, handles mixed variable types, and validates clusters through differences in outcomes, risk, or resource needs. We are especially interested in studies published in high-impact journals such as JAMA or Lancet, as well as work in the machine learning and health analytics space where methodological rigor is paired with clinical relevance. These studies serve as both methodological and conceptual benchmarks for the direction of our own analysis.

Relevant Literature Links

- **4.1 Clinical Sepsis Phenotypes in Critically Ill Patients**
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10538192/
- **4.2 Machine Learning Consensus Clustering in Black Kidney Transplant Recipients**
  https://jamanetwork.com/journals/jamasurgery/fullarticle/2791955
- **4.3 Data-Driven Subtypes of Diabetes (Alternative Reference)**
  https://www.nature.com/articles/s41598-025-93961-y
- **4.4 Reporting of Subgroup Analyses in Clinical Trials (Statistics in Medicine)**
  https://onlinelibrary.wiley.com/journal/10970258
- **4.5 Latent Class Analysis and k-Means Clustering for Complex Patient Profiles**
  https://pubmed.ncbi.nlm.nih.gov/
- **4.6 Prevalence of Children With Medical Complexity and Health Care Utilization**
  https://pubmed.ncbi.nlm.nih.gov/

- **4.7 Heterogeneous Treatment Effects of Therapeutic-Dose Heparin in COVID-19**
https://www.nejm.org/doi/full/10.1056/NEJMoa2105911
- **4.8 Multimorbidity Patterns in High-Need, High-Cost Elderly Patients**
https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2764783
- **4.9 Identification and Prediction of Chronic Diseases Using Machine Learning Approach**
https://onlinelibrary.wiley.com/doi/epdf/10.1155/2022/2826127
- **5.0 Clustering Approaches for Mixed-Type Data: A Comparative Study**
https://medium.com/analytics-vidhya/the-ultimate-guide-for-clustering-mixed-data-1eefa0b4743b
- **5.1 Recommendations for validating hierarchical clustering in consumer sensory projects**
https://www.sciencedirect.com/science/article/pii/S2665927123000904

4.1 Use of Machine Learning Consensus Clustering to Identify Distinct Subtypes of Black Kidney Transplant Recipients and Associated Outcomes

Go in depth with this paper.

4.2 Lancet Diabetes Endocrinol. (2018) – Data-Driven Subtypes of Diabetes

This study looks very promising but I am not able to access it.

4.3 Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials

Speaks more to the current challenges when it comes to statistics in medicine. It's more of a conversational piece

4.4 Use of Latent Class Analysis and k-Means Clustering to Identify Complex Patient Profiles

When it comes to hospitalization costs, a small number of medically complex patients account for a disproportionate share of medical spending. It is not only paramount that we identify these high-need individuals, but also develop innovative solutions to reduce the costs associated with their care. To achieve this goal, we must answer the question: *How can we identify patients with medically complex conditions?* Equally important is the question: *How can we classify these patients into distinct subgroups for targeted care?* This essential research is the primary focus of the paper *"Use of Latent Class Analysis and k-Means Clustering to Identify Complex Patient Profiles,"* a cohort study that analyzed a dataset of the most medically complex patients within Kaiser Permanente Northern California. The primary objective of this study runs parallel to our intent to identify the severity and utilization patterns of medically complex children. It is crucial to critically examine the methodologies presented in this paper—including their approach to feature selection and clustering—to inform our own subgroup analysis design, identify gaps in current knowledge, and uncover promising directions for future research.

This multidimensional dataset used in this study contained over 5,000 clinical variables. It is known that in data preprocessing, a large number of features can be redundant or highly correlated. The authors of this paper resolved this issue by using a multitude of feature selection techniques to reduce the number of features to 97, significantly simplifying the model by only using relevant features. Feature selection techniques ranged from objective statistical methods, such as the Jaccard similarity metric and VARCLUS procedure for reducing multicollinearity, to clinician-guided choices based on known indicators of health status. They identified high-need patients by comorbidity, measured by the

Comorbidity Point Score, version 2 (COPS2), a weighted comorbidity index that quantifies a patient's overall disease burden based on diagnoses recorded within the past 12 months. In addition, they used high healthcare utilization defined by a LOH score > 0.25 (Likelihood of Hospitalization) and/or ≥ 2 emergency room visits. They used two unsupervised grouping methods in order to classify distinct patient profiles: Latent Class Analysis and K-Means Clustering. Moreover, they used Generalized Low-Rank Models (GLRM), a preprocessing technique similar to PCA but applicable to mixed data types, to transform binary variables into continuous latent variables. Both methods contributed to robust cluster formation; more specifically, the two algorithms were used to assess the consistency of patient groupings across models.

4.5 Prevalence of Children With Medical Complexity and Associations With Health Care Utilization and In-Hospital Mortality

Guantao reviewed this paper already. It's about classification of children with complex conditions not nesscarrily subgroup analysis.

4.6 Multimorbidity patterns in high-need, high-cost elderly patients

4.7 Identification and Prediction of Chronic Diseases Using Machine Learning Approach

4.8 Heterogeneous Treatment Effects of Therapeutic-Dose Heparin in Patients Hospitalized for COVID-19

4.9 Recommendations for validating hierarchical clustering in consumer sensory projects

Rather than estimating a single average treatment effect, the study titled *Heterogeneous Treatment Effects of Therapeutic-Dose Heparin in Patients Hospitalized for COVID-19* investigates how the effectiveness of therapeutic-dose heparin varies between different subgroups of patients. This work acknowledges the limitations of population-wide inference and instead focuses on uncovering treatment effect heterogeneity through a combination of statistical modeling and machine learning.

The first approach relies on conventional **subgroup analysis**, stratifying patients by baseline characteristics such as illness severity, sex, and body mass index. The treatment effect within each subgroup is estimated using **Bayesian cumulative logistic regression** with interaction terms, allowing for probabilistic interpretation through posterior distributions. Odds ratios (ORs) are used as the primary measure of treatment effect: values greater than 1 indicate benefit, while values less than 1 indicate harm. Results reveal that patients with lower baseline severity are more likely to benefit from therapeutic-dose heparin, while severely ill patients show little to no benefit, suggesting meaningful heterogeneity in response to treatment.

The second approach, known as **risk-based modeling**, involves developing a multivariable prediction model to estimate each patient's baseline risk. Patients are sorted into deciles of predicted risk, and the treatment effect is reassessed using Bayesian methods within each stratum. The results reveal a striking contrast: patients in the lowest-risk decile have a 92% posterior probability of benefit, while those in the highest-risk decile have an 87% probability of harm. This approach emphasizes that treatment effects are not uniform but instead depend heavily on individual patient context.

The third and most flexible method utilizes a **causal forest**, a machine learning algorithm capable of estimating **individual-level treatment effects**. Unlike the prior two approaches, the causal forest does not require explicit grouping. Instead, it learns treatment heterogeneity directly from the data, producing personalized effect estimates. This model identified a subgroup of patients—specifically, those who were obese and on mechanical support—who were predicted to be harmed by treatment. Such precision enables a more nuanced understanding of treatment allocation and risk.

Together, these three approaches highlight the critical importance of moving beyond average effects in clinical research. The study demonstrates that treatment benefits and harms are not evenly distributed and that methods capable of capturing this variation—such as Bayesian interaction modeling, risk stratification, and causal forests—offer a clearer picture of who stands to benefit or be harmed. These insights underscore the need for more individualized approaches in therapeutic decision-making, especially in heterogeneous high-risk populations.

5 Clustering Implementation

This section describes the implementation of four clustering algorithms to identify subgroups within the study population. For each method, select distance metrics will be tested, and the number of resulting clusters will be reported. Evaluation will focus on cluster quality and consistency using metrics such as the Silhouette Score and others as appropriate.

Each model will be briefly summarized with a rationale for its use. Key hyperparameters, including the number of clusters (k) and distance measures, will be tuned and assessed for robustness. Cluster outputs will then be evaluated based on their ability to distinguish meaningful patterns in downstream variables such as LP cost and number of chronic conditions.

For feature selection, we used number of conditions, total inpatient claims, number of er visits, present on admission, Tech, Transplant, total medicaid paid amount inpatient, total medicaid paid amount emergency room.

Put this in a table with the data types of the features

5.1 Hierarchical Clustering

Hierarchical clustering is an unsupervised machine learning algorithm that builds a hierarchy of clusters, typically visualized using a dendrogram. There are two main approaches: agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering, the more commonly used method and the one applied in this study, begins by treating each data point as an individual cluster. Clusters are then iteratively merged based on similarity, determined by a chosen distance metric (e.g., Euclidean distance) and a linkage criterion (e.g., complete linkage). When it comes to hierarchical clustering, it is essential to understand the common linkage methods used, as they directly influence how clusters are formed and interpreted.

- **Single Linkage (Nearest Neighbor):** Defines the distance between two clusters as the shortest distance between any two points, one from each cluster. This method tends to produce long, chain-like clusters and is highly sensitive to noise and outliers.
- **Complete Linkage (Farthest Neighbor):** Measures the distance between two clusters as the greatest distance between any two points in the respective clusters. It typically produces more compact, spherical clusters and is less sensitive to outliers.
- **Average Linkage:** Computes the distance between two clusters as the average of all pairwise distances between points in the two clusters. It offers a balance between single and complete linkage, often resulting in moderately elongated clusters.
- **Ward's Method:** Merges clusters based on minimizing the increase in total within-cluster variance. It is the most similar to k-means clustering, as both aim to minimize within-cluster sum of squares. While k-means directly optimizes this objective function, Ward's method applies it in a hierarchical framework, often leading to more balanced and interpretable dendrograms.

5.1.1 Data Preprocessing

As part of our preprocessing pipeline for hierarchical clustering, we began by removing 134 rows with missing values, reducing the dataset from 1,589 to 1,455 complete records. We then applied the IQR method to filter out extreme outliers in cost and utilization features. These variables were highly right-skewed, making z-score–based detection unreliable, as it would have flagged clinically relevant high-cost cases as statistical anomalies. Instead, we used a non-parametric IQR approach with 5th and 95th percentile bounds to retain valid extremes while removing only the most outlying

values—mitigating distortion in downstream clustering. For dissimilarity measurement, we used Gower distance, which natively handles mixed data types and internally standardizes numeric features, making it especially appropriate for heterogeneous health datasets. After preprocessing, we swept $k = 2$ to 7, evaluating cluster validity using Silhouette Score and Dunn Index. We tested Euclidean and Manhattan distances in combination with various linkage methods and preprocessing strategies, including scalers such as StandardScaler and PowerTransformer. However, these configurations consistently produced impractically small or fragmented clusters that lacked clinical relevance. In contrast, the combination of Gower distance with complete linkage, without the need for external scaling, yielded the best results at k = 3 and k = 5 offering a strong balance between objective performance and interpretability.

- **Cluster 1 (n=51):** High-complexity super-utilizers with the highest IP claims and costs; also elevated ER use—likely high-need, high-cost patients.
- **Cluster 2 (n=64):** Moderate utilizers with disproportionately high inpatient costs despite lower condition burden—possibly tech- or transplant-dependent cases.
- **Cluster 3 (n=492):** ER-dominant, high-complexity group with the highest ER use and elevated overall costs—ideal candidates for care coordination.
- **Cluster 4 (n=4):** Tiny, homogeneous group with low utilization and low costs—likely an edge case or noise cluster.
- **Cluster 5 (n=830):** Low-complexity baseline group with the lowest costs and utilization—represents the majority of stable Medicaid patients.

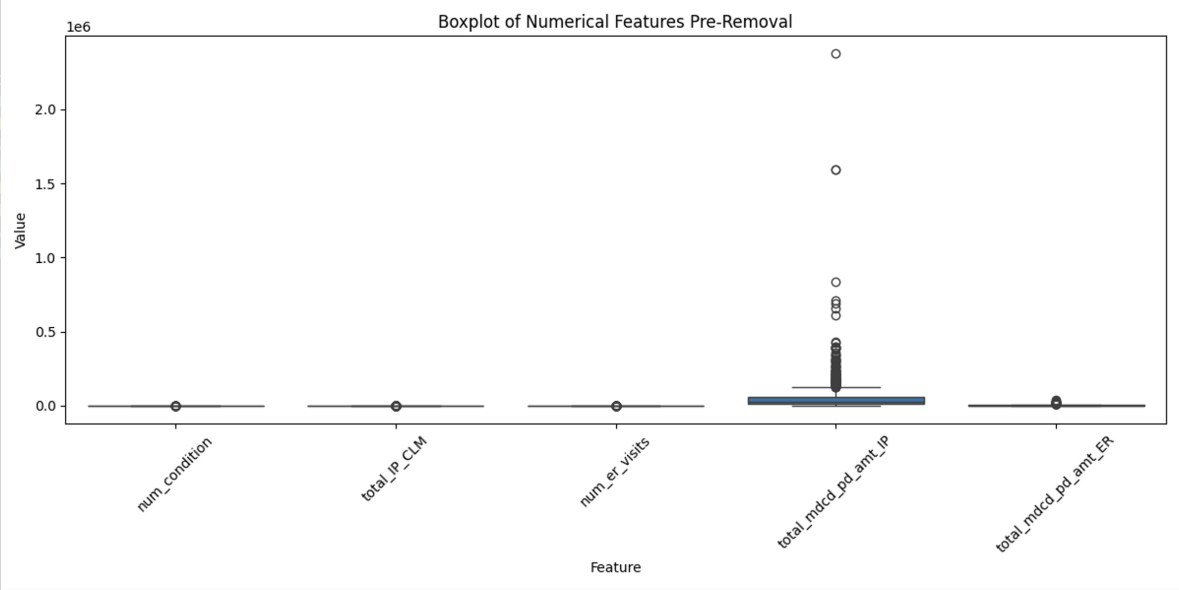

Figure 5: Box Plot of Numerical Features Pre-Removal

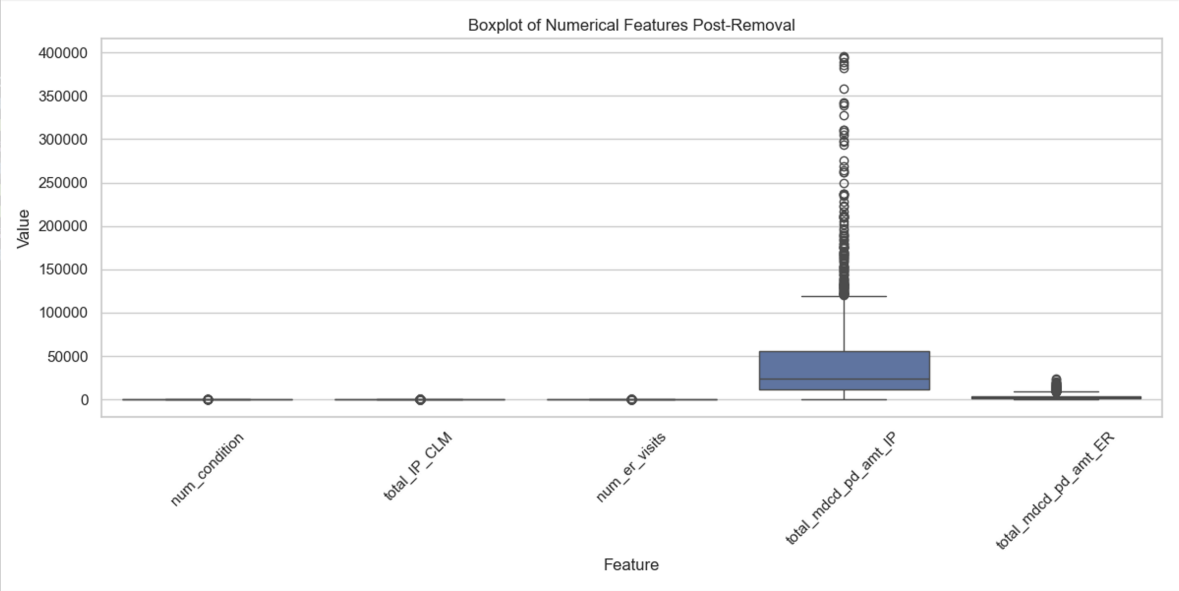

Figure 6: Box Plot of Numerical Features Post-Removal

These plots illustrate the impact of IQR-based outlier removal on numerical features. Extreme values—particularly in `total_mdcd_pd_amt_IP`—were compressing the visual scale and distorting the overall distribution. After outlier removal, the cost-related features appear better centered and more interpretable.

Table 2: Hierarchical Clustering with Gower Distance: k Sweep Results

k	**Silhouette Score**	**Dunn Index**
2	0.4896	0.2089
3	**0.6122**	**0.2467**
4	0.6161	0.2534
5	**0.6266**	**0.3104**
6	0.4631	0.0130
7	0.4013	0.0131

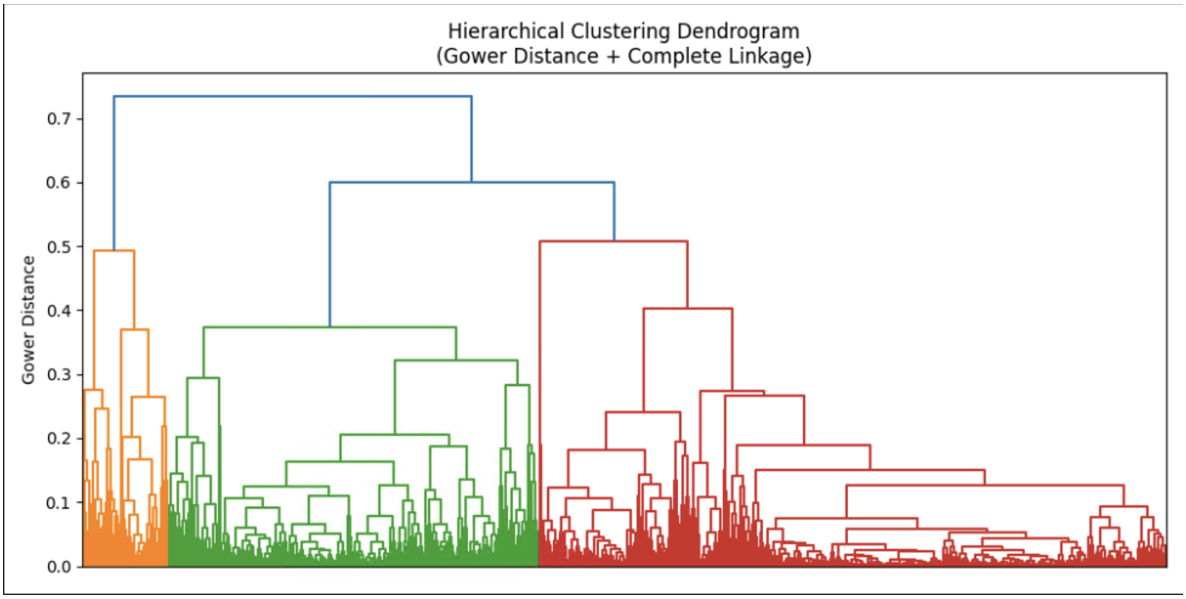

Figure 7: Hierarchical Clustering Dendrogram for Gower Distance + Complete Linkage

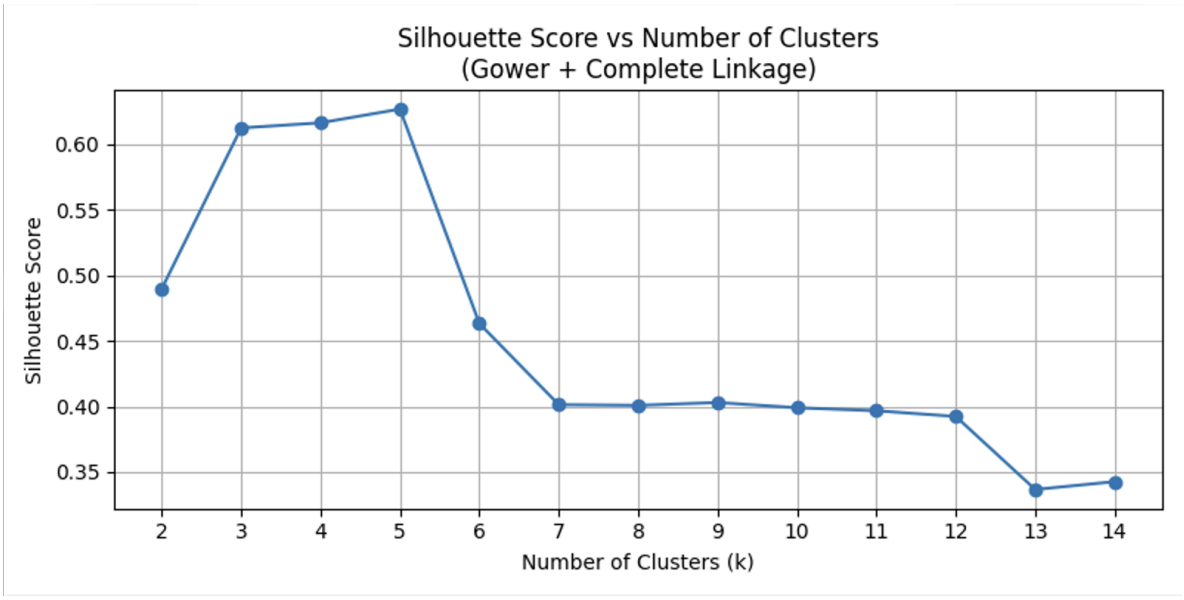

Figure 8: Box Plot of Numerical Features Pre-Removal

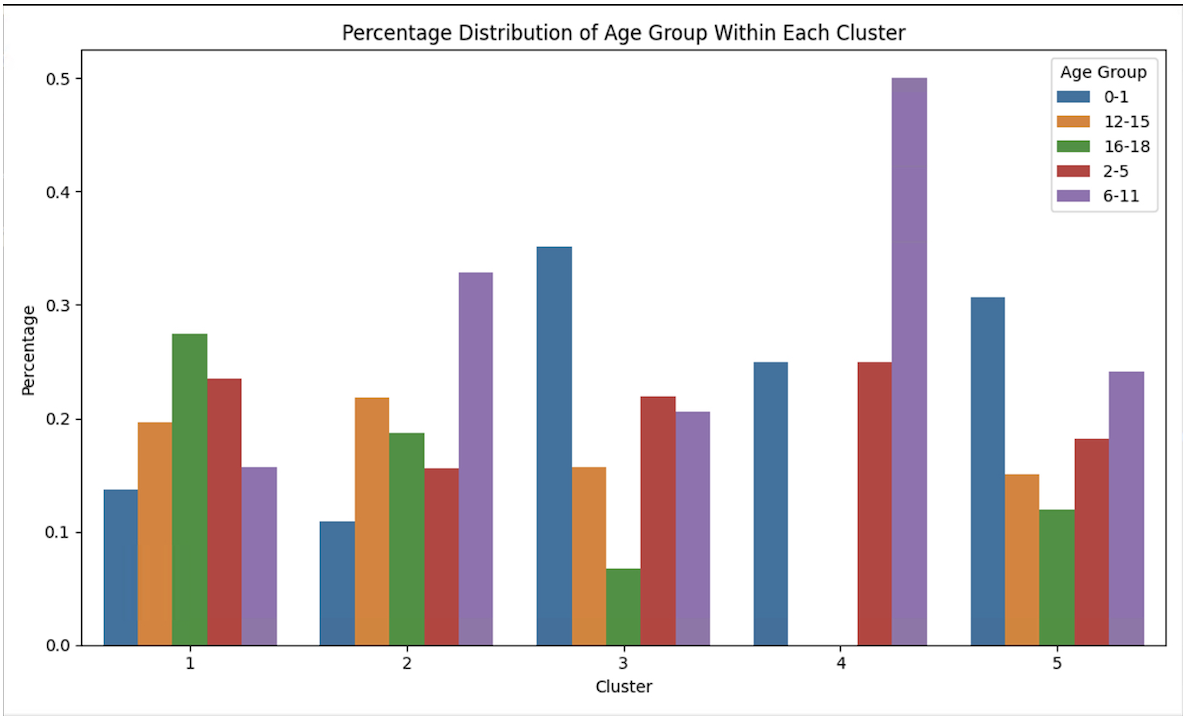

Figure 9: Age Group

Cluster Composition by Demographics

Table 3: Cluster Summary Statistics

Cluster	**num_condition**				**total_IP_CLM**				**num_er_visits**			
	size	median	mean	std	size	median	mean	std	size	median	mean	
1	51	2	1.980392	1.029373	51	6	6.529412	3.754213	51	10	14.01961	10.
2	64	1	1.625	0.863731	64	4	4.59375	3.308089	64	9	12.8125	11.
3	492	2	2.162602	0.862186	492	4	5.178862	3.230538	492	14	16.74797	13.
4	4	1	1.000000	0.000000	4	2	2.25	0.5	4	14	14	10.
5	830	1	1.26988	0.51462	830	3	3.610843	2.417593	830	10	13.85663	12

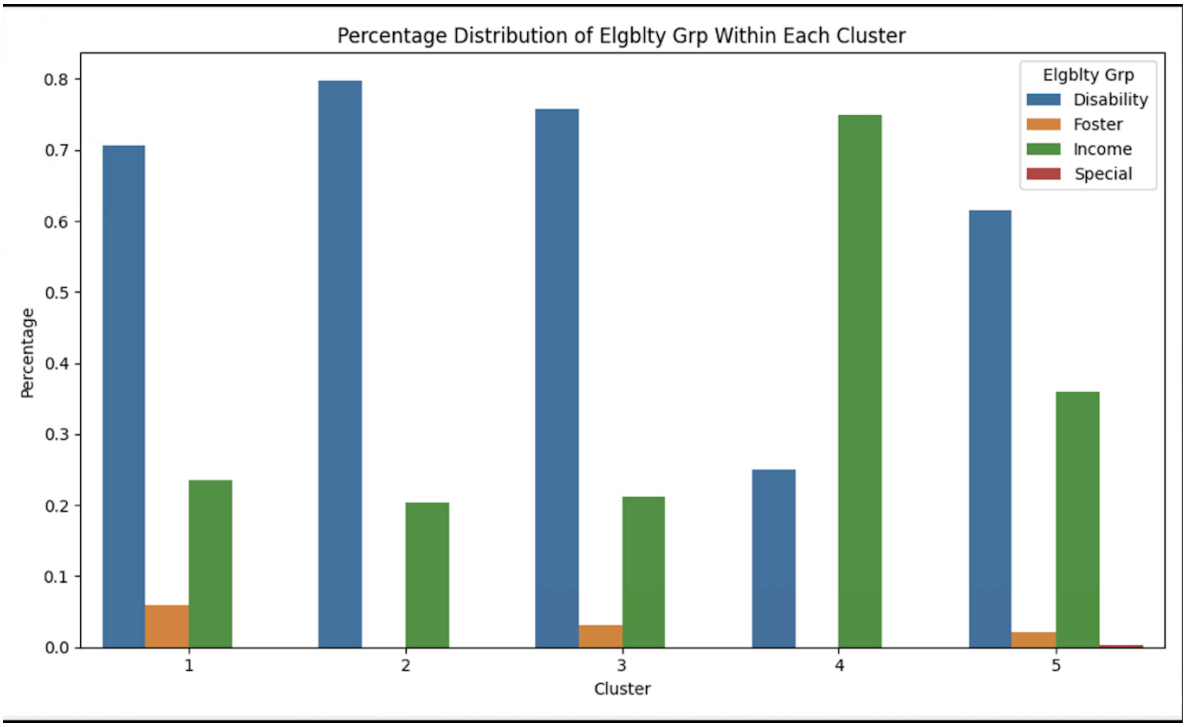

Figure 10: Eligibility Group

Age Group. Cluster 3 contains a disproportionately high number of infants (0–1 years), suggesting that this group may reflect early childhood or neonatal care needs.

Clusters 1 and 2 show a more even age distribution, with slight over-representation in older age groups (12–18). Cluster 5 exhibits a relatively balanced composition, but with noticeable weight toward the youngest and 6–11 groups.

Eligibility Group. Clusters 1, 2, and 3 are composed primarily of individuals in the *Disability* eligibility group, each exceeding 70%.

Cluster 5 presents a more mixed profile, with both *Disability* and *Income* groups present, and small proportions from *Foster* and *Special* categories.

These patterns provide interpretability to the clustering results, revealing demographic distinctions across patient subgroups that may inform downstream analysis or intervention strategies.

5.2 K-prototypes

K-Prototypes failed to meet the benchmarks to be included apart of this analysis

6 K-Medoids

K-medoids is a clustering algorithm that, unlike k-means, selects actual data points (called medoids) as the centers of clusters. It assigns each observation to the nearest medoid based on a chosen distance metric. Because it works with arbitrary distance matrices and does not rely on means, k-medoids is well-suited for our dataset, which contains mixed data types and heavily right-skewed cost and utilization features. Its robustness to outliers and flexibility with distance metrics (e.g., Gower) make it a strong choice for clustering complex healthcare data.

We applied k-medoids clustering using a Gower distance matrix and tested values of k from 2 to 6 to identify the optimal number of clusters. For each value of k, we evaluated clustering quality using Silhouette Score and Dunn Index to determine how cohesive and well-separated the resulting

groups were. This approach allowed us to select a clustering structure that was both interpretable and statistically robust.

Table 4: PAM Clustering with Gower Distance: k Sweep Results

k	**Silhouette Score**	**Dunn Index**
2	0.6741	0.1994
3	0.6932	0.0590
4	**0.7198**	**0.2501**
5	0.5746	0.0501
6	0.5025	0.0501

7 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that groups together points that are closely packed (i.e., high local density) while identifying points in low-density regions as outliers. Unlike k-means or k-medoids, DBSCAN does not require specifying the number of clusters in advance and can discover clusters of arbitrary shape. This makes it particularly well-suited for our dataset, which includes mixed data types, heavily skewed cost and utilization variables, and rare but clinically important outliers. By operating on a precomputed distance matrix (e.g., Gower), DBSCAN can effectively isolate both meaningful patient subgroups and extreme high-need cases without manual outlier removal.

Table 5: DBSCAN Parameter Sweep Results (ε and `min_samples`)

ε	**min_samples**	**n_clusters**	**n_noise**	**Silhouette**	**Dunn**
0.03	5	4	58	0.6835	0.0272
0.03	9	4	83	0.6677	0.0152
0.05	5	4	17	0.7115	0.0875
0.05	10	4	21	0.7089	0.0450
0.07	5	4	13	0.7147	0.1112
0.09	10	4	9	0.7170	0.1579
0.11	10	4	6	**0.7204**	**0.1856**

Table 6: Clustering Similarity Between PAM and DBSCAN

Metric	**Score**
Adjusted Rand Index (ARI)	0.712933
Normalized Mutual Information (NMI)	0.775827
Jaccard Score	0.048019

Patient Clustering Results

We applied k-prototypes clustering to identify patient subgroups with similar patterns of healthcare utilization and cost, incorporating both numerical features (e.g., number of inpatient claims, ER visits, total Medicaid paid amounts) and categorical covariates. This method is well-suited for mixed-type healthcare data and has been previously applied in chronic and complex patient population studies [**?**, **?**, **?**].

The optimal clustering solution yielded five distinct groups:

- **Cluster 0 (High-need, High-cost):** Comprised 95 patients with the highest average number of chronic conditions (mean: 3.03), the highest inpatient claim count (mean: 12.1), and total

inpatient costs exceeding $96,800 on average. These patients represent a multi-morbid, high-acuity group with complex care needs, similar to the "frail-elderly" or "high-acuity" phenotypes found in other unsupervised clustering analyses [**?**, **?**].

- **Cluster 1 (ER Super-utilizers):** A group of 190 patients with relatively low comorbidity (mean: 1.58 conditions), yet exceptionally high emergency department use (mean: 38 visits) and the highest mean ER-related Medicaid cost ($8,803). These patients may resemble the "psychiatric illness" or "less-engaged" profiles reported in latent class analyses [**?**].

- **Cluster 2 (Inpatient Cost Outliers):** Although only moderately complex clinically (mean: 1.88 conditions), this cluster showed extreme total inpatient costs (mean: $221,107), possibly reflecting rare catastrophic episodes or billing artifacts. This pattern is consistent with prior findings in chronic urticaria and rare disease cost clusters [**?**].

- **Cluster 3 (Moderate, Chronic Users):** Representing 415 patients, this cluster showed intermediate values across all utilization metrics, suggesting a stable chronic care group with manageable inpatient and ER costs. Such groups are commonly observed in population health segmentation for care coordination [**?**].

- **Cluster 4 (Low-need, Low-cost):** The largest group (n = 665), characterized by the lowest comorbidity burden, inpatient and ER claims, and overall cost. These patients likely represent a relatively healthy or low-risk segment, commonly observed in unsupervised segmentation studies [**?**, **?**].

Collectively, these clusters demonstrate the value of unsupervised learning for uncovering heterogeneous healthcare utilization patterns in chronic disease populations. The approach also aligns with prior literature emphasizing interpretable cluster profiles as a basis for targeted interventions [**?**, **?**].

We don't have ground truth labels for our data. So we can't use rand index. Silhouette score is our best bet.