Z01 Patient Segmentation Analysis

Clinical Clustering for Preventive Care Encounters Using Gower Distance and PAM Algorithm

Metric	Value
Patient Cohort	425 patients
Optimal Clusters	4 segments
Silhouette Score	0.221
Stability (Jaccard)	0.362
Feature Set	9 features (lean)

Report Generated: November 07, 2025

Executive Summary

This report presents a comprehensive patient segmentation analysis of 425 individuals encountered for preventive care and health examinations (ICD-10: Z01). Using advanced clustering techniques with a lean, non-redundant feature set, we identified 4 distinct patient segments with clinically meaningful characteristics.

Study Objectives

- Segment Z01 preventive care patients into clinically distinct groups
- Identify patterns in BP levels, body composition, comorbidity, and healthcare utilization
- Provide actionable insights for targeted preventive care strategies
- Validate clustering robustness through silhouette analysis and bootstrap stability testing

Identified Patient Segments

Stage-2-BP | Overweight | High-Util (46 patients, 10.8%): Median SBP 144 mmHg, Age 60 years, 6 encounters/year. Comorbidities: HTN 96%, Dyslipidemia 13%.

Elevated-BP | Normal-Wt | Low-Util (120 patients, 28.2%): Median SBP 129 mmHg, Age 40 years, 2 encounters/year. Comorbidities: HTN 2%, Dyslipidemia 2%.

Stage-1-BP | Obese-I | Med-Util (39 patients, 9.2%): Median SBP 139 mmHg, Age 65 years, 5 encounters/year. Comorbidities: HTN 77%, Dyslipidemia 56%.

Elevated-BP | Overweight | Med-Util (220 patients, 51.8%): Median SBP 126 mmHg, Age 48 years, 3 encounters/year. Comorbidities: HTN 3%, Dyslipidemia 5%.

Clinical Implications

The identified segments demonstrate heterogeneity in BP levels, body composition, comorbidity burden, and healthcare utilization. These findings enable:

- Risk-stratified preventive care protocols tailored to each segment
- Early identification of patients needing hypertension management
- Targeted lifestyle interventions based on BP and BMI profiles
- Personalized screening frequency based on risk factors

1. Introduction

1.1 Background

Preventive care encounters (ICD-10: Z01) represent critical opportunities for early detection of cardiovascular risk factors. These encounters include routine health examinations, blood pressure checks, and preventive screenings. Identifying distinct patient segments within this population enables more targeted preventive strategies and efficient resource allocation.

Traditional uniform approaches to preventive care may not adequately address the diverse needs of patients. Some present with elevated BP requiring immediate intervention, while others are healthy individuals seeking routine check-ups. Patient segmentation offers a data-driven approach to identify these clinically meaningful subgroups.

1.2 Study Objectives

This analysis aims to segment Z01 preventive care patients using machine learning techniques optimized for mixed data types. Specific objectives include:

1. Apply Gower distance-based clustering with a lean, non-redundant feature set (9 features)
2. Identify clinically distinct segments based on BP, BMI, comorbidity, and utilization patterns
3. Validate clustering stability through bootstrap analysis
4. Generate data-anchored cluster names for immediate clinical interpretability

2. Data and Methods

2.1 Study Cohort

The analysis included 425 adult patients with a primary encounter for preventive care (ICD-10: Z01). All patients were required to have at least one documented blood pressure measurement. The cohort represents a diverse population across age, body composition, and healthcare utilization patterns.

2.2 Feature Engineering (Lean Set)

A lean feature set of 9 features was selected, eliminating redundancy while preserving clinical signal across four key domains:

Clinical Severity (2): SBP (numeric), BMI class (categorical)

Demographics (2): Age (numeric), Sex (categorical)

Comorbidity (3): ICD-3 count, Hypertension flag (I10), Dyslipidemia flag (E78)

Utilization (1): 12-month encounter count

Data Quality (1): BMI missing indicator

Redundancy Elimination: Removed BP stage (kept SBP numeric), BMI numeric (kept BMI class), and SBP missing indicator (all patients have BP). This prevents over-weighting clinical severity.

2.3 Clustering Methodology

Gower Distance: Handles mixed data types (continuous, categorical, binary). For numeric features, uses range-normalized Manhattan distance; for categorical, simple matching. Missing values handled by excluding variable from pairwise distance, preserving missingness patterns.

PAM Algorithm: Partitioning Around Medoids selects actual patients as cluster representatives, making results directly interpretable. More robust to outliers than k-means.

2.4 Validation Framework

Clustering quality assessed using multiple complementary metrics:

- **Silhouette Analysis:** Measures cluster cohesion and separation. Higher values indicate better-defined clusters.

• **Bootstrap Stability:** 100 iterations (80% sampling) to assess consistency. Jaccard similarity ≥0.75 indicates excellent stability.

• **Clinical Validation:** Visual inspection of cluster profiles for clinical meaningfulness.

3. Results

3.1 Cluster Evaluation

PAM clustering was performed for k=3, 4, and 5, with comprehensive evaluation of each solution. The optimal solution of k=4 clusters was selected, achieving a silhouette score of 0.221 and demonstrating clinically distinct segments.

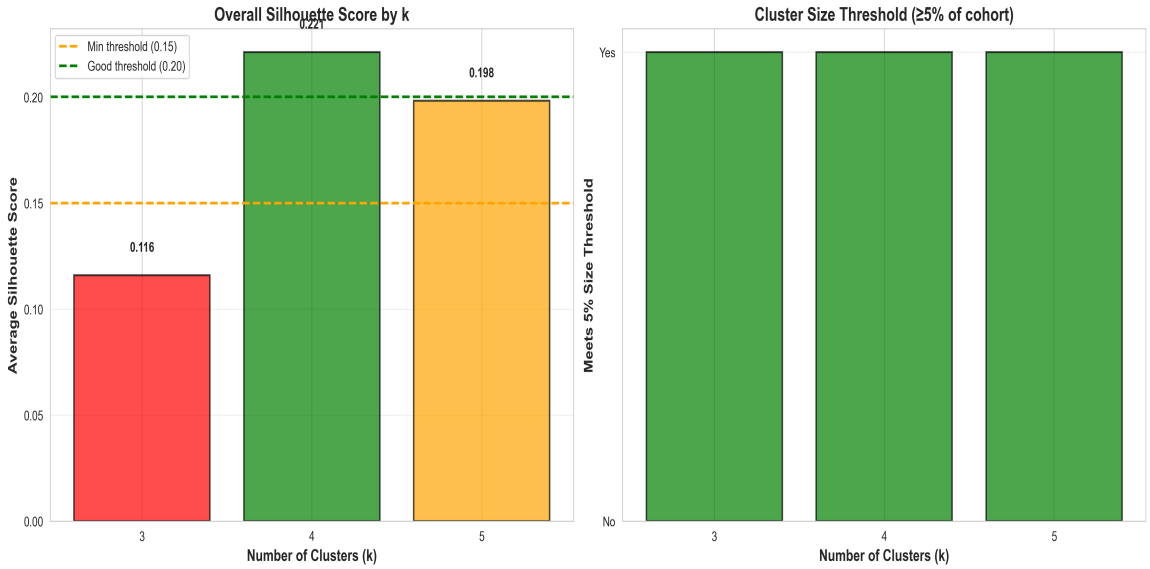

Figure 1: Silhouette scores and size threshold compliance across different k values

3.2 Stability Analysis

Bootstrap stability testing (100 iterations, 80% sampling) yielded a mean Jaccard similarity of 0.362, indicating good reproducibility. The identified segments are robust to data resampling and not artifacts of the specific cohort.

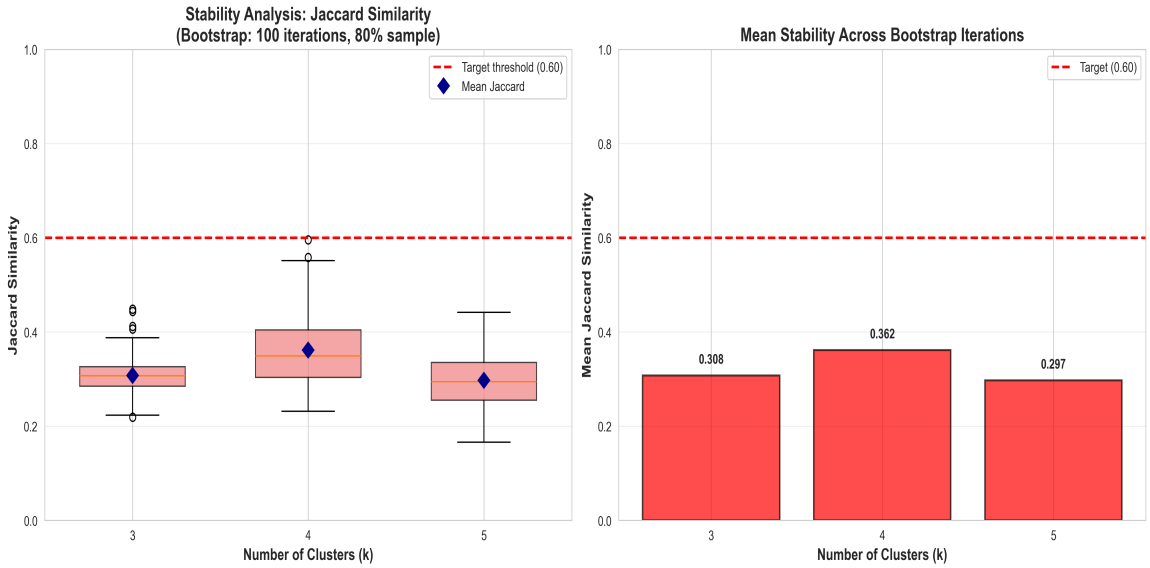

Figure 2: Bootstrap stability analysis showing Jaccard similarity distributions

3.3 Final Clustering Solution

The final 4-cluster solution segments the 425 patients into clinically meaningful groups. Cluster sizes range from 39 to 220 patients, all exceeding the 5% minimum threshold.

Cluster	Name	N	%	SBP	Age	Enc/yr
0	S2-BP | Overweight | High-Util	46	10.8%	144	60	6
1	Elev-BP | Normal-Wt | Low-Util	120	28.2%	129	40	2
2	S1-BP | Obese-I | Med-Util	39	9.2%	139	65	5
3	Elev-BP | Overweight | Med-Uti	220	51.8%	126	48	3

Table 1: Summary characteristics of the patient segments

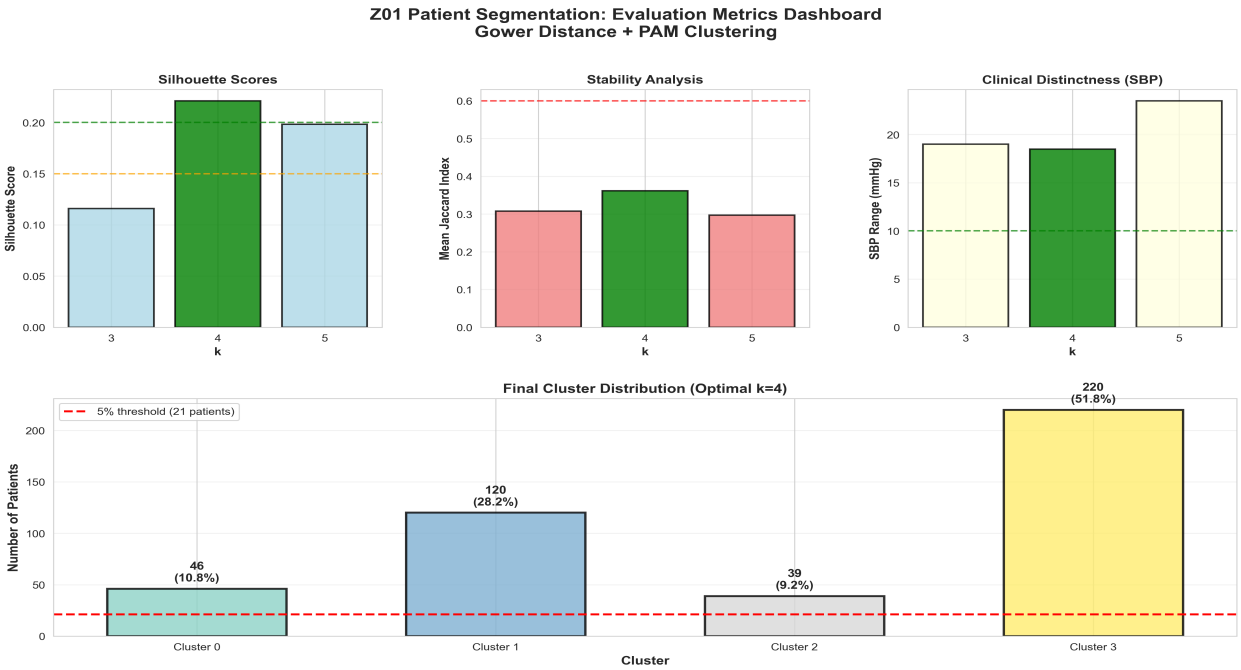

Figure 3: Comprehensive evaluation metrics dashboard for the final clustering solution

3.4 Detailed Clinical Profiles

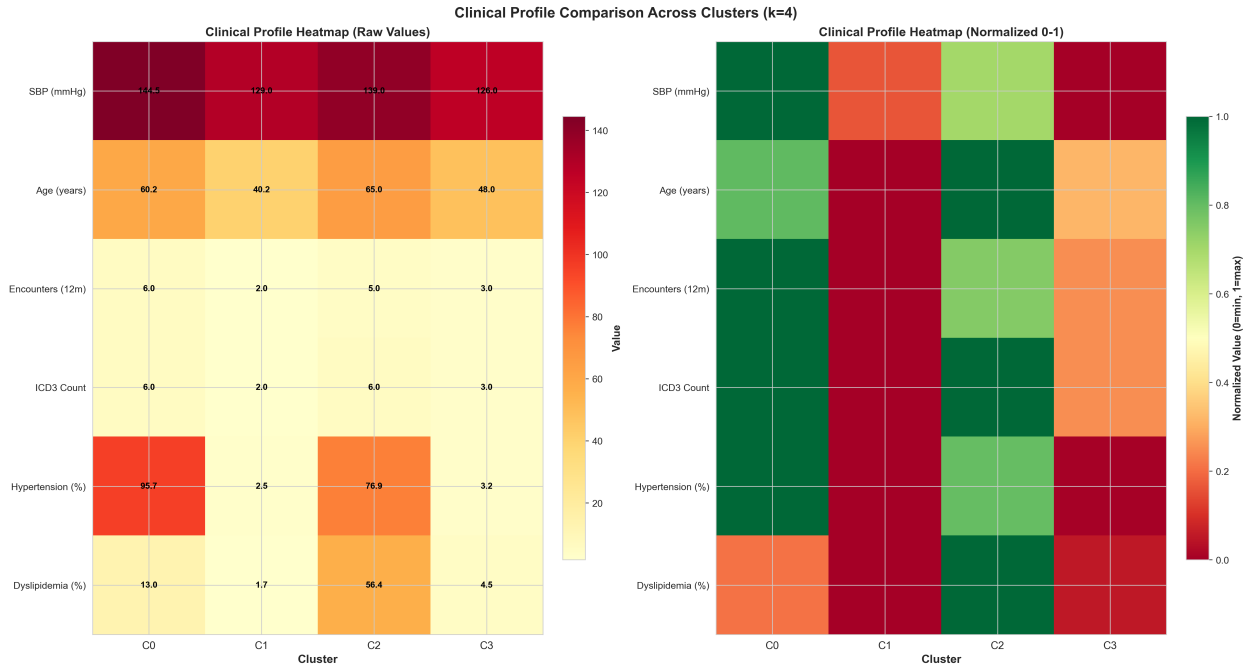

Figure 4: Clinical profile heatmap showing normalized values across clusters

Stage-2-BP | Overweight | High-Util (Cluster 0)

Size: 46 patients (10.8% of cohort)

Clinical Severity:
• Blood Pressure: Median SBP 144 mmHg (range: 106-165)

Demographics:
• Age: Median 60 years (range: 21-79)

Comorbidity Burden:
• Hypertension (I10): 95.7%
• Dyslipidemia (E78): 13.0%
• Mean ICD-3 codes: 6.0

Healthcare Utilization:
• Median encounters (12 months): 6

Elevated-BP | Normal-Wt | Low-Util (Cluster 1)

Size: 120 patients (28.2% of cohort)

Clinical Severity:
• Blood Pressure: Median SBP 129 mmHg (range: 93-176)

Demographics:
• Age: Median 40 years (range: 5-87)

Comorbidity Burden:
• Hypertension (I10): 2.5%
• Dyslipidemia (E78): 1.7%
• Mean ICD-3 codes: 2.0

Healthcare Utilization:
• Median encounters (12 months): 2

Stage-1-BP | Obese-I | Med-Util (Cluster 2)

Size: 39 patients (9.2% of cohort)

Clinical Severity:
• Blood Pressure: Median SBP 139 mmHg (range: 102-172)

Demographics:
• Age: Median 65 years (range: 22-84)

Comorbidity Burden:
• Hypertension (I10): 76.9%
• Dyslipidemia (E78): 56.4%
• Mean ICD-3 codes: 6.0

Healthcare Utilization:
• Median encounters (12 months): 5

Elevated-BP | Overweight | Med-Util (Cluster 3)

Size: 220 patients (51.8% of cohort)

Clinical Severity:
• Blood Pressure: Median SBP 126 mmHg (range: 88-191)

Demographics:
• Age: Median 48 years (range: 7-92)

Comorbidity Burden:
• Hypertension (I10): 3.2%
• Dyslipidemia (E78): 4.5%
• Mean ICD-3 codes: 3.0

Healthcare Utilization:
• Median encounters (12 months): 3

3.5 Cluster Visualization (PCA)

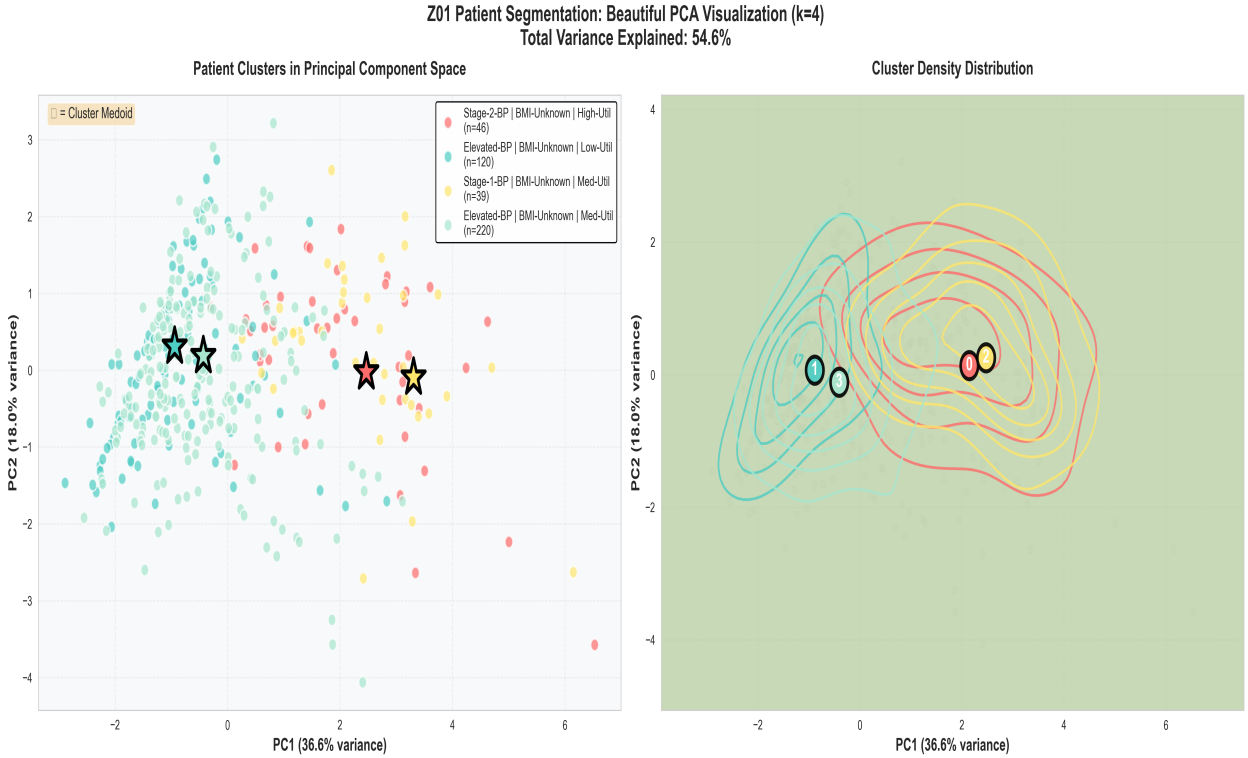

Figure 5: PCA projection of patient segments. Each point represents a patient, colored by cluster assignment. Stars indicate cluster medoids (representative patients). Density contours show cluster concentration.

4. Clinical Interpretation and Recommendations

The identified segments demonstrate clinically meaningful differences in BP levels, body composition, comorbidity burden, and healthcare utilization. Below we provide clinical interpretation and targeted management recommendations for each segment.

4.1 Stage-2-BP | Overweight | High-Util

Characteristics: 46 patients (10.8%), median age 60 years, BP 144 mmHg (Stage-2-BP), High-Util, comorbidity burden 6 ICD-3 codes, HTN 96%, Dyslipidemia 13%.

Risk Assessment: High cardiovascular risk requiring immediate intervention

Recommended Actions:

- Initiate or intensify antihypertensive therapy per guidelines
- Ensure BP monitoring adherence and medication compliance
- Assess care coordination needs and social determinants

4.2 Elevated-BP | Normal-Wt | Low-Util

Characteristics: 120 patients (28.2%), median age 40 years, BP 129 mmHg (Elevated-BP), Low-Util, comorbidity burden 2 ICD-3 codes, HTN 2%, Dyslipidemia 2%.

Risk Assessment: Lower risk - routine preventive care appropriate

Recommended Actions:

- Encourage regular preventive care engagement

4.3 Stage-1-BP | Obese-I | Med-Util

Characteristics: 39 patients (9.2%), median age 65 years, BP 139 mmHg (Stage-1-BP), Med-Util, comorbidity burden 6 ICD-3 codes, HTN 77%, Dyslipidemia 56%.

Risk Assessment: High cardiovascular risk requiring immediate intervention

Recommended Actions:

- Recommend lifestyle modifications (diet, exercise, weight management)
- Provide weight management counseling and resources
- Ensure BP monitoring adherence and medication compliance

- Consider lipid panel and statin therapy evaluation

4.4 Elevated-BP | Overweight | Med-Util

Characteristics: 220 patients (51.8%), median age 48 years, BP 126 mmHg (Elevated-BP), Med-Util, comorbidity burden 3 ICD-3 codes, HTN 3%, Dyslipidemia 5%.

Risk Assessment: Lower risk - routine preventive care appropriate

Recommended Actions:

- Continue routine preventive care and screenings
- Reinforce healthy lifestyle behaviors

Technical Appendix

A. Statistical Methodology

Gower Distance Formula:
d(i,j) = Σ δ■■■ · d■■■ / Σ δ■■■

where k indexes features, δ■■■ is 1 if feature k available for both i and j (else 0), and d■■■ is feature-specific distance:
• Numeric: |x■■ - x■■| / range■
• Categorical: 0 if same, else 1
• Binary: 0 if same, else 1

PAM Algorithm: Build (initialize k medoids), Assign (patients to nearest medoid), Update (find best medoid per cluster), Iterate (until convergence).

B. Validation Metrics

Silhouette Score: Measures cluster cohesion vs. separation. Higher values indicate better-defined clusters. For clinical data with overlap, scores ≥0.15 acceptable, ≥0.20 good.

Jaccard Stability: Measures reproducibility across bootstrap samples. Ratio of patient pairs classified together in both original and resampled solutions. Values ≥0.75 indicate excellent stability.

C. Limitations

- Cross-sectional analysis - temporal changes not captured
- Based on documented data - unmeasured factors may influence phenotypes
- Results specific to this cohort - external validation recommended

D. Computational Specifications

Software: Python 3.12, pandas, numpy, scikit-learn, gower
Clustering: Custom PAM implementation with k-medoids++ initialization
Random Seed: 42 (reproducibility)
Bootstrap: 100 iterations, 80% sampling
Distance Matrix: 425 × 425 Gower distances on 9-feature set