

Clustering Outpatient and Inpatient Medical Records Based on Patient Complaints using K-Means

Nery Adelia Br Purba^{1*}, Magdalena Simanjuntak², Marto Sihombing³

^{1,2,3}*Sekolah Tinggi Manajemen Informatika dan Komputer Kaputama, Binjai, Indonesia*
neryadelia373@gmail.com^{1*}, magdalena.simanjuntak84@gmail.com², martosihombing45@gmail.com³

Abstract

Medical record archives contain patterns that can support hospital service planning, but large datasets are difficult to analyze manually. This study developed a clustering application to group outpatient and inpatient medical records at Delia Hospital based on age, patient complaint, and type of service. A total of 600 medical records from 2025 were selected, cleaned, transformed into numerical codes, and processed in MATLAB using the K-means algorithm with squared Euclidean distance. Three scenarios were evaluated, namely K = 3, K = 4, and K = 5, using cluster variance and the distribution of members as comparison measures. The cluster variance decreased from 4.7356 at K = 3 to 2.8841 at K = 4 and 2.0662 at K = 5. The five-cluster solution produced 129, 95, 122, 127, and 127 members, respectively, with dominant complaints of asthma, vertigo, diabetes, fractures, and gastritis. The results indicate that K = 5 provides the most compact and detailed grouping. The resulting information can support the evaluation of outpatient services, inpatient capacity, and hospital resource planning.

Keywords: Clustering; K-means; Medical records; Patient complaints; Service type

1. Introduction

Medical records are essential documents in health-service delivery because they contain patient identity, examination history, diagnoses, treatment, medical actions, and other services received by patients. Beyond their administrative function, medical records provide clinical information that supports continuity of care and communication among health professionals [1]. When the number of records increases, however, the information stored in archives becomes difficult to examine manually and potentially useful patterns may remain unidentified.

Delia Hospital in Langkat Regency stores outpatient and inpatient medical records containing diverse patient ages, complaints, and service types. Outpatient and inpatient services have different operational requirements related to registration, monitoring, facilities, and continuity of care [2][3]. The records had mainly functioned as administrative archives, while systematic grouping had not been applied to reveal patterns among patients. As a result, hospital staff faced limitations in obtaining concise information about dominant complaints, the characteristics of patient groups, and the tendency of each group to use outpatient or inpatient services.

Clustering is an unsupervised data-mining technique that separates objects into groups so that objects within the same group are more similar than objects in different groups. Data-mining methods are used to extract previously hidden patterns from accumulated datasets [4]. The K-means algorithm is widely used because its process is relatively simple, computationally efficient, and suitable for numerical datasets [5][6][7][8]. Previous studies have applied K-means to health-related datasets, including regional stunting indicators [5], web-based medical records [9], patient-service grouping at Delia Hospital [10][11], medical-record management in Binjai [12], and severity-based hospital patient grouping [13]. These studies show that clustering can transform accumulated health data into structured information.

The present study focuses on clustering medical records according to three attributes: age, complaint, and type of service. Unlike an implementation that reports only one predetermined number of clusters, this study compares K = 3, K = 4, and K = 5 using cluster variance, minimum variance, and maximum variance. The contribution is an integrated MATLAB application that reads transformed records from Microsoft Excel, performs clustering, presents the centroid and dominant characteristics of each group, visualizes the clusters, and saves the results. The study aims to identify the most compact grouping and to produce information that can support hospital service planning without replacing clinical decision-making.

2. Research method

2.1. Research data and variables

The research used 600 medical records from Delia Hospital for the 2025 service period. The records represented both outpatient and inpatient services. The clustering process did not use patient names as analytical attributes; only the three research variables required for grouping were processed. This approach allowed the analysis to focus on service patterns while reporting the results in aggregate form.

The source data were categorical and therefore had to be transformed into numerical values before being processed by K-means. Age was represented by five ordinal category codes, complaint by twenty category codes, and service type by two codes. Outpatient service was coded as 1 and inpatient service as 2. The variables used in the analysis are summarized in Table 1.

Table 1: Variables used in the clustering process

Variable	Symbol	Code range	Description
Age	X	1-5	Ordinal age category recorded in the medical record
Complaint	Y	1-20	Patient complaint category, including asthma, diabetes, vertigo, fracture, and gastritis
Service type	Z	1-2	Outpatient = 1 and inpatient = 2

2.2. Data preparation

The research stages followed the Knowledge Discovery in Database process. Data selection was performed to retain the attributes relevant to the research objective. Data cleaning was then applied to identify incomplete or inconsistent records. During transformation, each category was converted into its corresponding numerical code. The resulting matrix consisted of 600 rows and three numerical columns representing X, Y, and Z.

The transformed dataset was stored in a Microsoft Excel file and imported into the MATLAB application. The application validated that at least three numerical columns were available and excluded rows containing empty numerical values. The same validated dataset was used in all clustering scenarios to ensure that the comparison among K values was consistent.

2.3. K-means clustering procedure

The K-means procedure began by selecting the number of clusters, initializing the centroid, assigning every record to the nearest centroid, recalculating the centroid from the members of each cluster, and repeating the assignment process until no meaningful membership change remained. Squared Euclidean distance was used as the proximity measure. To reduce the effect of random initialization, the MATLAB implementation used a fixed random seed and ten replications for each K value. The clustering scenarios evaluated in this study were K = 3, K = 4, and K = 5. The procedure is illustrated in Fig. 1.



Fig. 1: K-means clustering procedure used in the study

2.4. Cluster evaluation

The resulting clusters were evaluated by examining the number of members, centroid positions, dominant categories, variance, minimum variance (V_{min}), maximum variance (V_{max}), and overall cluster variance. The dominant age, complaint, and service type were determined using the mode among the members of each cluster. This distinction was important because complaint and service-type codes are categorical; therefore, the numerical centroid was interpreted as a center of distribution rather than directly as a category label.

Cluster variance was used as the primary comparison measure among the three K scenarios. A lower value indicates that records are, on average, more closely distributed around their respective centroids. The selected solution was the scenario that achieved the smallest cluster variance while still producing interpretable groups.

3. Results and discussion

3.1. Application implementation

The clustering application was implemented in MATLAB and organized into four main modules: the main menu, clustering process, medical-record information, and help. The clustering module imports the Excel dataset, displays the number of rows and columns, allows the number of clusters to be selected, performs K-means, and presents the centroid, dominant characteristics, number of members, and evaluation values. It also provides functions to visualize the result and export the clustered records to Microsoft Excel.

Fig. 2 shows the main processing interface. The interface separates data input, K-means configuration, centroid information, cluster-member information, and detailed cluster characteristics. This design makes the analytical output easier to interpret without requiring the operator to read program code or conduct manual calculations.

Fig. 2: MATLAB interface for processing and displaying the clustering results

3.2. Comparison of cluster scenarios

All three scenarios successfully divided the 600 records into groups. Increasing K produced more specific partitions and reduced the overall dispersion of records around their centroids. The comparison is presented in Table 2.

Table 2: Comparison of the evaluated cluster scenarios

Number of clusters	Members in each cluster	Cluster variance	Interpretation
3	189; 186; 225	4.7356	Largest dispersion
4	123; 147; 171; 159	2.8841	More compact than K = 3
5	129; 95; 122; 127; 127	2.0662	Smallest dispersion

At K = 3, the records formed groups of 189, 186, and 225 members. The solution had a cluster variance of 4.7356, indicating that the three broad groups still contained relatively high internal variation. The dominant complaints were fracture, diabetes, and gastritis. The first group tended toward inpatient service, whereas the second and third groups tended toward outpatient service.

At $K = 4$, the cluster variance decreased to 2.8841. The additional cluster divided the complaint-code distribution into narrower intervals. The member counts were 123, 147, 171, and 159. Asthma and headache emerged as dominant complaints in separate groups, while diabetes and fracture dominated the other groups. This solution was more compact than $K = 3$ but still had higher dispersion than $K = 5$.

At $K = 5$, the cluster variance reached 2.0662, the lowest value among the evaluated scenarios. The distribution of 129, 95, 122, 127, and 127 members also produced relatively balanced groups, except for Cluster 2, which was smaller. The decrease in variance indicates that adding the fifth cluster separated records with different complaint and service characteristics more effectively.

3.3. Characteristics of the selected five-cluster solution

The detailed characteristics of the selected $K = 5$ solution are shown in Table 3. Adult-category records (code 4) were dominant in every cluster, while complaint and service patterns differentiated the groups.

Table 3: Centroid and dominant characteristics of the $K = 5$ solution

Cluster	Centroid (X; Y; Z)	Members	Dominant complaint	Dominant service
1	4.0388; 2.6124; 1.5039	129	Asthma: 37 (28.68%)	Inpatient: 65 (50.39%)
2	4.1263; 18.6421; 1.6526	95	Vertigo: 27 (28.42%)	Inpatient: 62 (65.26%)
3	4.0492; 6.6311; 1.4098	122	Diabetes: 39 (31.97%)	Outpatient: 72 (59.02%)
4	4.1417; 14.3701; 1.3622	127	Fracture: 42 (33.07%)	Outpatient: 81 (63.78%)
5	3.9449; 10.3386; 1.2047	127	Gastritis: 40 (31.50%)	Outpatient: 101 (79.53%)

Cluster 1 contained 129 records and was dominated by asthma. Its service distribution was almost balanced, with 65 inpatient and 64 outpatient records. Cluster 2 contained 95 records, was dominated by vertigo, and had the strongest inpatient tendency: 62 records, or 65.26%, used inpatient services. This group can therefore be considered the most relevant for evaluating inpatient readiness among the five clusters.

Cluster 3 contained 122 records and was dominated by diabetes, with 59.02% of its members receiving outpatient services. Cluster 4 contained 127 records and was dominated by fracture, while 63.78% of its members received outpatient services. Cluster 5 also contained 127 records, was dominated by gastritis, and showed the strongest outpatient tendency, reaching 79.53%. This group is particularly relevant for evaluating outpatient capacity, follow-up services, and supporting examinations.

The three-dimensional visualization of $K = 5$ is shown in Fig. 3. The axes represent age code, complaint code, and service-type code. Because the variables were transformed from categories, the visualization should be interpreted as the separation of coded service patterns rather than as a clinical scale. The centroid markers indicate the numerical center of each group.

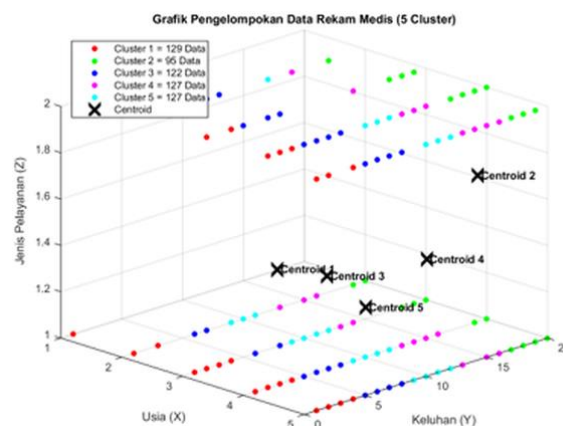


Fig. 3: Three-dimensional visualization of the $K = 5$ clustering result

3.4. Implications and limitations

The resulting groups provide an initial analytical view of how patient complaints relate to outpatient and inpatient services. Groups with a stronger inpatient tendency can support the review of room availability, nursing capacity, and monitoring requirements. Groups dominated by outpatient services can support the evaluation of polyclinic workload, medicine availability, control schedules, and diagnostic support. The results are intended as management information and must not be used as a substitute for diagnosis or clinical decisions.

The study has several limitations. First, the clustering used only three variables, so other factors such as sex, diagnosis, length of stay, severity, comorbidity, and treatment outcome were not represented. Second, numerical coding of nominal complaints creates an ordered numerical space even though the clinical categories do not necessarily have a natural distance. The results are therefore most appropriate for exploratory grouping within the coding scheme used by the hospital. Future research should compare alternative encodings, distance measures for mixed data, and additional validation indices to strengthen the interpretation.

4. Conclusion

A MATLAB-based system was successfully developed to cluster 600 outpatient and inpatient medical records from Delia Hospital using age, complaint, and service type. The system transformed the variables into numerical form, applied K-means, and displayed the centroid, number of members, dominant characteristics, evaluation values, and graphical visualization.

The comparison of $K = 3$, $K = 4$, and $K = 5$ showed that the five-cluster solution was the most compact, with a cluster variance of 2.0662. The five groups contained 129, 95, 122, 127, and 127 records and were dominated by asthma, vertigo, diabetes, fracture, and gastritis, respectively. The information can support the review of outpatient and inpatient service needs, resource allocation, and hospital planning. Continuous data standardization and the use of additional variables and validation methods are recommended for further development.

References

- [1] S. Apriliyani, "Penggunaan rekam medis elektronik guna menunjang efektivitas pendaftaran pasien rawat jalan di Klinik Dr. Ranny," *Cerdika: Jurnal Ilmiah Indonesia*, vol. 1, no. 10, pp. 1399–1410, 2021, doi: 10.36418/cerdika.v1i10.209.
- [2] T. F. H. Farida and S. Setiatin, "Analisis lama waktu tunggu pelayanan pasien rawat jalan di bagian pendaftaran Rumah Sakit X Bandung," *Akrab Juara: Jurnal Ilmu-Ilmu Sosial*, vol. 6, no. 5, pp. 119–126, 2021, doi: 10.58487/akrabjuara.v6i5.1692.
- [3] M. E. Putra, N. Desreza, and R. A. Atika, "Hubungan kualitas pelayanan terhadap kepuasan pasien rawat inap di Rumah Sakit Umum Teungku Fakinah Banda Aceh tahun 2025," *Future Academia: The Journal of Multidisciplinary Research on Scientific and Advanced*, vol. 3, no. 4, pp. 1547–1558, 2025, doi: 10.61579/future.v3i4.575.
- [4] R. K. Barus, "Perlindungan hak pasien sebagai konsumen dalam rekam medis," *Jurnal Juristic*, vol. 4, no. 2, p. 171, 2024.
- [5] M. F. Amalia and D. B. Arianto, "Implementasi algoritma K-means clustering dalam klasterisasi kabupaten/kota Provinsi Jawa Barat berdasarkan faktor pemicu stunting pada balita," *Simkom*, vol. 9, no. 1, pp. 36–46, 2024, doi: 10.51717/simkom.v9i1.356.
- [6] E. Prasetyo, *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta, Indonesia: Andi, 2012.
- [7] Kusri and E. T. Luthfi, *Algoritma Data Mining*. Yogyakarta, Indonesia: Andi, 2009.
- [8] F. Marisa, "Educational data mining: Konsep dan penerapan," *Jurnal Teknologi Informasi*, vol. 4, no. 2, pp. 91–93, 2013.
- [9] F. H. Nasution, R. Kurniawan, and R. A. Putri, "Sistem informasi rekam medis pada puskesmas dengan menerapkan algoritma K-means berbasis web," *Jurnal Ilmiah Sains dan Teknologi*, vol. 9, no. 2, 2025.
- [10] S. Monika, "Pengelompokan pelayanan pasien rawat inap dan pasien rawat jalan berdasarkan jenis penyakit pasien menggunakan metode clustering (Studi Kasus: Rumah Sakit Delia)," in *Prosiding Seminar Nasional Informatika (SENATIKA)*, 2021, pp. 144–153.
- [11] P. R. S. Br. Ginting, R. Buaton, and S. Ramadani, "Pengelompokan pasien BPJS dan pasien umum berdasarkan pelayanan rumah sakit dengan menggunakan metode clustering (Studi Kasus: Rumah Sakit Delia)," *Jurnal Ilmiah Kaputama (JIKA)*, vol. 6, no. 2, pp. 89–98, 2022.
- [12] D. N. Br. Purba, M. Sihombing, and I. Ambarita, "Pengelompokan data rekam medis pada pasien penyakit dalam untuk meningkatkan manajemen informasi kesehatan berdasarkan wilayah Kota Binjai menggunakan algoritma clustering K-means (Studi Kasus: Artha Medika Binjai)," *Jurnal Penelitian Teknologi Informasi dan Sains*, vol. 2, no. 3, pp. 103–117, 2024, doi: 10.54066/jptis.v2i3.2378.
- [13] G. Alasi and R. A. Putri, "Penerapan K-means clustering untuk pengelompokan pasien rumah sakit berdasarkan tingkat keparahan penyakit," *Jurnal Multimedia dan Teknologi Informasi (Jatilima)*, vol. 7, no. 3, pp. 475–486, 2025, doi: 10.54209/jatilima.v7i03.1598.