Vifada Management and Social Sciences, E-ISSN 2987-1999
This work is Licensed under a Creative Commons Attribution 4.0 International License

Strategic Patient Segmentation Based on Age, Length of Stay, and Hospital Revenue Using K-Means Clustering: a Hospital Management Perspective

Almerya Indriastuti [1*] *Arief Wibowo* [2]

[1,2] *Faculty of Economics and Business, Budi Luhur University, Jakarta, Indonesia*

Received: 2026, 07, 18 **Accepted:** 2026, 09,04
Available online: 2026, 09,28

*Corresponding author.
E-mail addresses: 2531600506@student.budiluhur.ac.id

KEYWORDS

Keywords:

patient segmentation; K-Means clustering; length of stay; hospital revenue; hospital management; healthcare analytics; data-driven decision-making

Conflict of Interest Statement:

The author(s) declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2026 **Vifada Management and Social Sciences.** All rights reserved.

ABSTRACT

Purpose: This study aims to identify strategic patient segments based on age, length of stay, and hospital revenue using K-Means clustering and to interpret the resulting segments from a hospital management perspective.

Research Method: A quantitative descriptive-exploratory design was employed using anonymized secondary inpatient data from a private international-standard hospital in South Tangerang, Banten, Indonesia. The variables included patient age, length of stay, and hospital revenue. Data were cleaned, transformed, normalized, and analyzed using the K-Means algorithm in RapidMiner Studio, resulting in three clusters.

Results and Discussion: Three patient segments with distinct service-utilization and financial profiles were identified. Cluster 0 generated average revenue of IDR 163,171,150, Cluster 1 generated IDR 330,793,288, and Cluster 2 recorded the highest average revenue of IDR 994,300,700. Despite relatively similar patient ages and lengths of stay, substantial revenue differences suggest variations in case complexity, medical procedures, and resource intensity.

Implications: The segmentation model can support differentiated service strategies, resource allocation, capacity planning, and data-driven revenue management. Future studies should incorporate clinical severity, payment methods, treatment types, and alternative clustering techniques.

Originality: This study demonstrates the managerial application of K-Means clustering by integrating demographic, service-utilization, and financial variables to develop actionable patient segmentation for hospital decision-making

1) Introduction

Hospitals are increasingly required to balance clinical quality, patient experience, operational efficiency, and financial sustainability in an environment characterized by rising service complexity and limited organizational resources. The growing availability of administrative and financial healthcare data has created new opportunities for hospitals to support strategic decision-making through healthcare analytics. By transforming routinely collected inpatient data into actionable information, hospitals can better understand patient heterogeneity, optimize resource allocation, improve capacity planning, and support evidence-based management decisions. However, hospital populations remain highly heterogeneous in their demographic characteristics, hospitalization patterns, service utilization, and financial contribution. Consequently, identifying meaningful patient segments has become an important analytical approach for supporting differentiated management strategies and more efficient use of hospital resources.

Patient heterogeneity creates a practical management problem because individuals treated within the same institution may require substantially different levels of care, medical procedures, technologies, and hospital resources. High-need patients cannot be identified adequately through a single indicator because their service requirements are usually shaped by combinations of clinical complexity, previous utilization, and organizational context (Hewner et al., 2023). Research involving high-need and high-cost hospitalized children also demonstrates that expenditure patterns are associated with multiple demographic, clinical, and utilization characteristics rather than one isolated factor (P. Zhang & al., 2025). When hospitals manage such heterogeneous patients as a homogeneous population, they risk producing undifferentiated service strategies, inefficient capacity planning, and inappropriate allocation of limited resources.

Length of stay (LOS) is an important operational indicator because it reflects the duration of inpatient resource utilization and directly influences bed occupancy, workforce allocation, discharge planning, and hospital capacity management. Although LOS has frequently been investigated as an outcome variable in predictive modeling, it also provides valuable operational information for understanding differences in patient service utilization (Jain et al., 2024). Within the context of patient segmentation, LOS represents the intensity of hospital resource consumption rather than a variable to be predicted. Patients with similar demographic characteristics may experience different hospitalization durations, reflecting variations in care processes and operational demands (Lee et al., 2024). When analyzed together with demographic attributes such as age and financial indicators such as hospital revenue, LOS contributes to a multidimensional representation of patient profiles that supports strategic segmentation and managerial decision-making. Therefore, in this study, LOS is incorporated as an operational dimension that complements demographic and financial characteristics for identifying relatively homogeneous patient groups through K-Means clustering, rather than as the target of a predictive model (Bopche et al., 2024).

Hospital revenue is similarly important because it reflects the financial contribution generated through diagnostic examinations, medication, medical procedures, accommodation, specialist services, and other forms of resource utilization. Patients with comparable ages and lengths of stay may generate substantially different revenues when they receive services with different levels of complexity and intensity. Machine-learning research on cardiac patients indicates that clinical and demographic variables jointly influence hospitalization patterns that affect operational planning and resource requirements (AlMuhaideb et al., 2024). Explainable artificial intelligence research also demonstrates that prolonged hospitalization is influenced by multidimensional patient and treatment characteristics that cannot be represented by duration alone (Yasin et al., 2024). Therefore, integrating demographic, operational, and financial attributes is necessary to obtain a more complete understanding of patient profiles from a hospital management perspective.

Recent healthcare studies have increasingly used data-driven population segmentation to identify relatively homogeneous groups within heterogeneous patient populations. A segmentation study using German claims data demonstrated that healthcare utilization information could distinguish population groups with different service needs and patterns of care consumption(Pioch et al., 2023). A scoping review found that machine-learning-based population segmentation has been applied across various clinical settings, with K-Means representing the most frequently used unsupervised algorithm (Liu et al., 2023). These developments demonstrate that clustering is useful when predefined class labels are unavailable and the purpose of the analysis is to discover naturally occurring structures within healthcare data.

K-Means clustering is particularly relevant because it is computationally efficient, comparatively easy to implement, and capable of producing clusters that can be interpreted by hospital managers. Research using more than 21 million outpatient prescription claims showed that K-Means could identify insured groups with different utilization and risk characteristics (Momahhed et al., 2023). An Indonesian hospital study also demonstrated that K-Means could segment patient visit patterns and provide information for service planning (Marlina et al., 2025). These studies confirm the applicability of K-Means to large healthcare datasets, although the managerial usefulness of the results depends on the variables selected and the quality of cluster validation.

The use of longitudinal and electronic health record data has further expanded the sophistication of patient clustering. Unsupervised clustering of longitudinal clinical measurements can identify distinct patterns of disease development and healthcare trajectories from routinely collected records (Mariam et al., 2024). Deep representation learning can also generate clinically meaningful patient subgroups from complex longitudinal survival data contained in electronic health records (Qiu et al., 2025). Clustering of electronic health records among patients with atrial fibrillation has revealed subgroups with different prognoses and patient trajectories (Y. Zhang & al., 2025). These studies illustrate the analytical potential of patient clustering, but their primary orientation remains clinical rather than managerial or financial.

Other recent studies have applied clustering to identify disease-specific patient profiles and variations in healthcare utilization. K-Means clustering among patients with chronic ulcers distinguished groups with different ages, comorbidities, hospitalization burdens, and medical resource consumption. Density-based clustering of open medical records also identified clinically interpretable patient profiles while emphasizing the importance of selecting algorithms and validation measures appropriate to the structure of the data. A literature-based study of hospital length-of-stay prediction concluded that prediction models are highly dependent on patient characteristics, hospital context, and the availability of relevant variables (Deschepper et al., 2025). These findings suggest that patient profiles should be interpreted carefully and translated into management decisions only after considering the scope and limitations of the variables used.

Despite the growing body of healthcare clustering research, several gaps remain. First, previous patient-segmentation studies have predominantly emphasized disease characteristics, clinical risk, healthcare trajectories, prescription behavior, or service utilization, whereas comparatively less attention has been given to integrating financial indicators such as hospital revenue into patient segmentation models. Second, existing studies have generally interpreted patient clusters from clinical or epidemiological perspectives, with more limited discussion of their implications for service differentiation, hospital capacity planning, patient portfolio management, and financial sustainability (Kim et al., 2025). Third, although demographic, operational, and financial variables have each been examined in healthcare analytics, studies integrating age, length of stay, and hospital revenue within a single unsupervised clustering framework for strategic hospital management remain limited.

A further theoretical gap concerns the limited connection between patient clustering and management concepts. Robust patient grouping is essential for using electronic health records to study heterogeneous populations and to design decisions that are relevant to distinct patient categories. From a strategic management perspective, patient segments can support more focused decisions concerning service design, capacity utilization, relationship management, and allocation of organizational resources. From a segmentation perspective, the clusters can function as the foundation for identifying groups that require different service propositions and managerial responses. Nevertheless, the clusters should not be interpreted as direct representations of clinical severity when the model contains only demographic, operational, and financial variables.

Based on these gaps, this study addresses the following research questions: (1) Can the integration of demographic characteristics (age), operational service utilization (length of stay), and financial characteristics (hospital revenue) reveal strategically meaningful inpatient segments with distinct financial profiles? (2) How can these patient segments be interpreted from a hospital management perspective to support managerial decision-making?

Accordingly, this study aims to examine whether the integration of demographic, operational, and financial dimensions can identify financially distinct inpatient segments and to interpret the resulting clusters in terms of their managerial relevance for service differentiation, capacity planning, resource allocation, and revenue management.

The novelty of this study does not lie in proposing a new clustering algorithm or merely combining multiple variables. Instead, it lies in examining the analytical value of integrating age, length of stay, and hospital revenue to generate a more comprehensive understanding of patient financial heterogeneity. In particular, the study explores whether financially distinct patient profiles can emerge that may not be sufficiently captured when relying primarily on conventional operational indicators such as length of stay. The resulting clusters are therefore interpreted as exploratory strategic patient profiles that may support data-driven hospital management decisions.

In addition, this study provides empirical evidence from a private international-standard hospital in South Tangerang, Indonesia, thereby extending the application of patient segmentation into the context of strategic hospital analytics. Given its exploratory nature, the study does not claim definitive superiority of the proposed segmentation model over alternative specifications.

Comparative clustering models and additional statistical validation are therefore recommended for future research.

2) Literature Review and Research Propositions

Patient Segmentation in Hospital Management

Patient segmentation refers to the systematic classification of patients into relatively homogeneous groups based on demographic, clinical, behavioral, operational, or financial characteristics. In hospital management, segmentation enables decision-makers to move beyond aggregate patient statistics and recognize that different patient groups may impose different demands on services, capacity, resources, and financial management. Accordingly, segmentation can function not merely as a statistical classification technique but as a strategic analytical approach for understanding patient heterogeneity and supporting differentiated managerial responses.

Unlike conventional market segmentation, however, patient segmentation in healthcare cannot be based exclusively on commercial attractiveness or financial contribution. Hospitals must simultaneously consider clinical needs, ethical responsibilities, accessibility of care, resource requirements, and organizational sustainability. Therefore, the managerial relevance of patient segmentation lies in identifying meaningful differences among patient groups while maintaining a balance between service quality, operational efficiency, and financial sustainability.

Previous studies have demonstrated the usefulness of clustering in organizing complex healthcare data. Frangky et al. (2025), for example, analyzed approximately 58 million patient visits and showed that clustering techniques could reveal diagnostic patterns and relationships that would be difficult to identify through conventional aggregate analysis. Such findings demonstrate the ability of clustering to reduce the complexity of heterogeneous medical records and transform large datasets into more interpretable patient profiles.

Nevertheless, much of the existing patient-segmentation literature remains oriented toward clinical conditions, diagnostic patterns, risk classifications, or patterns of healthcare utilization. Although these dimensions are important for clinical decision-making, comparatively less attention has been directed toward combining demographic, operational, and financial information for strategic hospital-management purposes. This limitation creates an opportunity to examine whether patient segmentation can provide additional managerial value when patient characteristics are analyzed jointly with utilization and financial outcomes.

Demographic, Operational, and Financial Dimensions of Patient Segmentation

The present study conceptualizes patient segmentation through three complementary dimensions: demographic characteristics represented by age, operational utilization represented by length of stay, and financial characteristics represented by hospital revenue. The integration of these dimensions provides the analytical framework underlying the study.

Age represents an important demographic characteristic because healthcare needs and service-utilization patterns may vary across stages of the life cycle. Different age groups may experience different disease patterns, treatment requirements, and levels of healthcare consumption. However, age should not be interpreted as a direct indicator of clinical severity. Patients of similar ages may have substantially different diagnoses, medical procedures, and resource requirements. Therefore, age is treated in this study as a demographic segmentation attribute rather than as a determinant of clinical complexity.

Length of stay represents an operational dimension reflecting the duration of inpatient service utilization. Longer hospitalization generally requires greater use of hospital beds, nursing services, medication, diagnostic facilities, and supporting resources. Consequently, length of stay is frequently used as an indicator of operational pressure and hospital-resource consumption. Studies related to patient flow have demonstrated that differences in hospitalization patterns can substantially influence bed utilization, patient waiting times, and operational efficiency (Zamani et al., 2025; Khedr et al., 2025).

However, length of stay does not necessarily correspond proportionally with financial outcomes. Two patients with comparable hospitalization durations may generate substantially different levels of revenue depending on their diagnoses, medical procedures, specialist services, medications, diagnostic examinations, and intensity of resource utilization. This limitation suggests that reliance on operational measures alone may not sufficiently explain financial heterogeneity among hospital patients.

Hospital revenue therefore provides an additional financial dimension. Revenue reflects the financial value generated through services such as accommodation, medical procedures, specialist consultations, pharmaceuticals, laboratory examinations, imaging, medical devices, and other inpatient services. Integrating revenue into patient segmentation makes it possible to investigate whether patients with relatively similar demographic and operational characteristics exhibit substantially different financial profiles.

This perspective is supported by previous studies showing that financial patterns among healthcare users are heterogeneous. Ng et al. (2025), for instance, applied longitudinal K-Means clustering to identify patient groups with distinct healthcare-expenditure trajectories over five years. Their findings indicate that financial utilization does not develop uniformly across patient populations and that such heterogeneity can provide important information for resource planning. Similarly, research integrating patient grouping with resource-allocation models has demonstrated that patient segments can support differentiated allocation of healthcare resources and service priorities (Akbari-Moghaddam et al., 2025).

The literature therefore suggests that demographic, operational, and financial indicators capture different but complementary aspects of patient characteristics. Considering these dimensions simultaneously may provide a richer understanding of inpatient heterogeneity than relying on a single operational or financial indicator.

K-Means Clustering as a Strategic Hospital Analytics Approach

K-Means is an unsupervised machine-learning technique designed to partition observations into a predetermined number of relatively homogeneous clusters. Each observation is assigned to the nearest centroid, after which the centroids are iteratively recalculated until cluster membership becomes relatively stable. Unlike supervised learning, K-Means does not require predefined outcome categories or class labels, making it appropriate for exploratory studies seeking to discover naturally occurring patterns within patient data.

The use of K-Means in the present study is supported by the numerical characteristics of age, length of stay, and hospital revenue. However, these variables are measured on substantially different scales. Hospital revenue, in particular, has a much larger numerical magnitude than either age or length of stay. Without appropriate preprocessing, variables with larger scales may disproportionately influence distance calculations and consequently affect cluster formation. Data normalization is therefore an important methodological prerequisite.

Sapitri et al. (2025) applied Z-score normalization before K-Means clustering of BPJS Health patient records and reported stable cluster formation after multiple iterations. Their study illustrates the importance of standardizing healthcare variables with different numerical ranges. Siahaan et al. (2025) similarly demonstrated that clustering inpatient data could distinguish patient groups and provide information relevant to operational efficiency, resource allocation, and service differentiation.

Other studies extend the usefulness of clustering beyond operational classification. Patient segmentation at a private hospital in Jambi, for example, identified groups based on differences in perceived service quality, demonstrating that clustering results can be interpreted from a service-management perspective. Kim et al. (2025) found that clustering emergency department patients according to demographic and physiological characteristics revealed heterogeneity that conventional triage classifications did not fully capture. Likewise, Ang et al. (2025) showed that unsupervised learning could identify clinically heterogeneous septic-shock subgroups with different lengths of stay, treatment requirements, and patterns of resource consumption.

Collectively, these studies demonstrate that clustering can reveal multidimensional differences among patients that may be obscured by conventional classifications. Nevertheless, most existing applications emphasize clinical risk, diagnostic categories, expenditure patterns, service perceptions, or resource consumption individually. Fewer studies explicitly examine whether demographic, operational, and hospital-revenue information can be combined to generate strategically interpretable inpatient profiles.

Managerial Interpretation of Patient Clusters

The value of patient clustering does not end with the statistical identification of groups. From a hospital-management perspective, clusters become useful when their characteristics can be translated into implications for service differentiation, capacity planning, resource allocation, and financial management.

Patient-flow research has demonstrated that differences in patient characteristics and hospitalization patterns influence bed utilization, service delays, and operational costs. Zamani et al. (2025) developed a strategic and tactical patient-flow framework showing that improved coordination of patient transitions can enhance capacity utilization and generate financial savings. Similarly, Khedr et al. (2025), based on 179,066 emergency department visits, found that patient-transfer strategies could reduce unnecessary patient days and improve bed utilization without compromising the intended level of care.

These findings illustrate why patient segments should not be interpreted solely according to revenue levels. A high-revenue cluster, for example, may also reflect patients receiving complex procedures, intensive medical services, expensive medications, or other resource-intensive treatments. Conversely, a lower-revenue segment may consume substantial bed capacity despite generating relatively lower financial returns. Therefore, managerial interpretation requires simultaneous consideration of operational utilization and financial contribution rather than assuming that higher revenue automatically represents greater organizational value.

This integrated perspective is particularly important for hospitals seeking to balance financial sustainability with service quality and equitable access to care. Patient segmentation can therefore function as a decision-support mechanism by helping hospital managers identify groups with different service patterns, resource demands, and financial characteristics.

Synthesis of Previous Studies and Research Gap

The reviewed literature demonstrates that patient clustering has been applied to diagnostic-pattern discovery, clinical-risk identification, healthcare-expenditure analysis, patient-flow management, service-quality evaluation, and resource-allocation decisions. These studies consistently show that unsupervised clustering techniques can reveal patient heterogeneity that conventional aggregate statistics or predefined classifications may overlook.

However, several limitations remain. First, most patient-segmentation studies predominantly emphasize clinical characteristics, disease severity, physiological indicators, healthcare expenditure, or service utilization. Second, operational indicators such as length of stay are frequently examined independently from direct hospital-revenue characteristics. Third, relatively few studies integrate demographic, operational, and financial information within a single segmentation framework and subsequently translate the resulting clusters into strategic managerial implications.

Accordingly, the contribution of the present study does not lie in proposing a new clustering algorithm. Rather, it lies in examining the analytical value of integrating age, length of stay, and hospital revenue to determine whether these dimensions can reveal financially and operationally distinct inpatient profiles. The study further extends previous research by interpreting the resulting clusters as strategic inputs for service differentiation, capacity planning, resource allocation, and revenue management in a private international-standard hospital in South Tangerang, Indonesia.

Research Propositions

Because this study employs an exploratory unsupervised clustering design rather than a causal or hypothesis-testing framework, formal directional hypotheses are not proposed. Instead, the literature supports the following research propositions:

RP1: The integration of patient age, length of stay, and hospital revenue is expected to reveal distinct inpatient segments with different demographic, operational, and financial profiles.

DOI: 10.70184/v1kq1v81

RP2: Patients with relatively similar age and length-of-stay characteristics may exhibit different hospital-revenue profiles, indicating financial heterogeneity that cannot be explained by operational utilization alone.

RP3: The identified patient segments are expected to provide managerially relevant information for service differentiation, capacity planning, resource allocation, and hospital revenue management.

3) Research Design and Methodology

This study employed a quantitative descriptive-exploratory design to identify strategic patient segments based on demographic, operational, and financial characteristics. The quantitative approach was appropriate because the analysis relied on numerical inpatient data, while the descriptive component was used to characterize the profiles of the resulting patient segments. The exploratory component was intended to identify naturally occurring patterns without predefined class labels or the testing of causal relationships. Accordingly, K-Means clustering was selected as an unsupervised analytical technique to group inpatient episodes according to similarities in age, length of stay, and hospital revenue.

The research design was aligned with the study's research questions concerning whether the integration of demographic, operational, and financial dimensions could reveal strategically meaningful patient segments and how these segments could be interpreted from a hospital management perspective. Thus, the analysis focused not only on identifying statistical clusters but also on translating the resulting profiles into managerial insights related to service differentiation, capacity planning, resource allocation, and revenue management.

Data Source and Unit of Analysis

The study used anonymized secondary data obtained from the inpatient database of a private international-standard hospital located in South Tangerang, Banten, Indonesia. The data were derived from routine administrative and financial records generated during inpatient services. The unit of analysis was an inpatient episode, defined as one hospitalization period from admission to discharge.

The study applied a total-data approach to all eligible inpatient records available in the dataset. This approach was appropriate because the objective was to examine patterns across all eligible inpatient episodes rather than estimate population parameters from a probability sample. Records were included when complete and valid information was available for all variables used in the clustering analysis. Records containing missing values, duplicate observations, invalid age values, inconsistent admission or discharge information, nonpositive hospitalization duration, or incomplete revenue information were excluded during data preparation.

All directly identifiable patient information, including names, addresses, identification numbers, and contact details, was excluded before analysis. The data were analyzed only in anonymized and aggregate form to maintain patient confidentiality.

Variables

Three numerical variables were used as clustering attributes. **Age** represented the patient's age at the time of hospitalization and was measured in completed years. It represented the demographic dimension of patient segmentation and was not interpreted as a direct proxy for clinical severity.

Length of stay (LOS) represented the duration of inpatient treatment from admission to discharge and was measured in days. LOS was used as an operational indicator because hospitalization duration is associated with bed utilization and the consumption of nursing and supporting hospital resources.

Hospital revenue represented the total financial value recorded for services delivered during a single inpatient episode and was measured in Indonesian rupiah. Depending on the hospital records, the value incorporated inpatient accommodation, medical procedures, diagnostic examinations, pharmaceutical services, specialist services, and other recorded hospital services.

The three variables therefore represented complementary dimensions of hospital management: age captured demographic characteristics, length of stay represented operational utilization, and hospital revenue represented financial characteristics. Their joint use was intended to investigate whether patient groups with relatively similar demographic and hospitalization profiles could nevertheless exhibit substantially different financial characteristics.

Data Preparation

Data preparation involved completeness checking, duplicate detection, numerical-format verification, and logical-validity assessment. Missing or incomplete records were excluded because K-Means requires complete numerical observations, while duplicate records were removed to prevent the same inpatient episode from influencing cluster formation more than once.

Potential extreme observations, particularly in hospital revenue, were examined during data cleaning. Extreme values were not automatically removed solely because of their magnitude. Observations were retained when they represented valid inpatient episodes rather than data-entry or recording errors, because legitimate high-revenue episodes may contain important information regarding patient heterogeneity and hospital resource utilization.

Because age, length of stay, and hospital revenue were measured on substantially different numerical scales, the variables were normalized using **Min-Max scaling**, transforming each attribute to a range between 0 and 1. This procedure prevented hospital revenue, which had a considerably larger numerical magnitude than age and length of stay, from disproportionately dominating the Euclidean distance calculations. Normalized values were used for cluster formation, whereas the original measurement units were retained for descriptive reporting and managerial interpretation.

K-Means Clustering Procedure

The clustering analysis was conducted using RapidMiner Studio. K-Means with Euclidean distance was employed to partition inpatient episodes into relatively homogeneous groups. The algorithm initially established cluster centroids, assigned each observation to the nearest centroid, recalculated the centroid positions based on cluster membership, and iteratively repeated the process until convergence.

The number of clusters was not determined solely on the basis of managerial interpretability. Candidate cluster solutions were first evaluated using the **Elbow Method**, which examined changes in within-cluster variation across alternative values of k. The resulting solutions were further assessed using the **Silhouette coefficient** to evaluate the degree of cohesion within clusters and separation between clusters. Based on these evaluations, a three-cluster solution (k = 3) was selected as an appropriate representation of the underlying patient-segmentation structure.

The three-cluster solution was subsequently profiled using centroid values and the distribution of inpatient episodes within each cluster. The normalized centroids were interpreted alongside the original values for age, length of stay, and hospital revenue to facilitate practical interpretation.

Cluster Profiling and Managerial Interpretation

Cluster labels were assigned after the clustering procedure based on the relative characteristics of their centroid profiles and were not predefined before the analysis. Cluster 0 was characterized as a moderate-revenue segment, Cluster 1 as a high-revenue segment, and Cluster 2 as a premium-revenue segment. These labels were used solely to facilitate managerial interpretation and did not represent judgments regarding patient importance, quality of care, or social value.

Particular attention was given to whether clusters with relatively similar average ages and lengths of stay exhibited substantially different hospital-revenue profiles. Such patterns were interpreted as evidence of financial heterogeneity that might not be apparent from operational indicators alone.

The managerial interpretation focused on the potential relevance of the identified segments for service differentiation, inpatient capacity planning, resource allocation, and revenue management. However, the clusters were not interpreted as direct measures of disease severity, clinical risk, treatment quality, or profitability. Revenue differences may also reflect diagnosis, medical procedures, treatment intensity, technology utilization, class of care, and payment method, none of which were directly included in the present clustering model.

Accordingly, the results should be interpreted within the boundaries of the three variables analyzed. Although the Elbow Method and Silhouette coefficient were used to support the selection of the cluster solution, future studies could further evaluate robustness using additional validation indices, such as the Davies–Bouldin and Calinski–Harabasz indices, and compare the resulting segmentation with alternative algorithms such as hierarchical clustering, K-medoids, or Gaussian mixture models.

4) Findings and Discussion

a) Findings

Descriptive Statistics

Descriptive analysis was conducted to summarize the demographic, operational, and financial characteristics of the inpatient episodes included in the clustering analysis. The three variables analyzed were patient age, length of stay, and hospital revenue, representing the demographic, operational, and financial dimensions of inpatient management.

Patient age ranged from 0 to 88 years, with a mean of 39.81 years. Length of stay ranged from 1 to 30 days, with an average hospitalization duration of 6.85 days. Hospital revenue showed considerably greater dispersion, ranging from IDR 2,306 to IDR 1,218,100,750, with a mean of IDR 36,099,806.27 per inpatient episode.

Table 1. Descriptive Statistics of the Research Variables

Variable	Minimum	Maximum	Mean
Age (years)	0	88	39.81
Length of Stay (days)	1	30	6.85
Hospital Revenue (IDR)	2,306	1,218,100,750	36,099,806.27

Source: Authors' calculation based on the processed inpatient dataset.

The descriptive statistics indicate substantial heterogeneity across the inpatient records. The wide age range shows that the hospital served patients from infancy to older adulthood, while variation in length of stay reflects differences in hospitalization duration. Hospital revenue displayed the greatest numerical dispersion. The substantial difference between the minimum and maximum values suggests that the revenue distribution may be strongly right-skewed; therefore, the mean should be interpreted cautiously as a measure of central tendency. This heterogeneity provides an empirical basis for applying clustering to identify more homogeneous patient groups that may not be visible through aggregate statistics alone.

Cluster Validation and Selection

Candidate K-Means solutions were evaluated using the Elbow Method and Silhouette coefficient. These procedures were used to assess within-cluster compactness and separation between clusters rather than selecting the number of clusters solely on the basis of managerial interpretability. The evaluation supported a three-cluster solution, and therefore k = 3 was retained for the final segmentation analysis.

K-Means Clustering Results

K-Means clustering produced three patient segments based on similarities in normalized age, length of stay, and hospital revenue. Normalized values were used during cluster formation, while the original measurement units were used to profile and interpret the resulting clusters.

Table 2. Patient Cluster Profiles

Cluster	Average Age	Average Length of Stay	Average Hospital Revenue
Cluster 0	39.44 years	6.91 days	IDR 163,171,150
Cluster 1	43.40 years	8.10 days	IDR 330,793,288
Cluster 2	43.00 years	8.00 days	IDR 994,300,700

Source: Authors' calculation using RapidMiner Studio.

The three clusters displayed relatively modest differences in average age and length of stay but substantially larger differences in hospital revenue. Average age ranged from 39.44 to 43.40 years, while average length of stay ranged from 6.91 to 8.10 days. In contrast, average hospital revenue differed markedly across the three groups.

The contrast between Cluster 1 and Cluster 2 was particularly notable. Their average ages differed by only 0.40 years and their average lengths of stay by only 0.10 days. Nevertheless, the average hospital revenue of Cluster 2 exceeded that of Cluster 1 by approximately IDR 663.51 million. This pattern indicates that similar demographic and hospitalization-duration profiles may be associated with substantially different financial outcomes.

Cluster Profiles

Cluster 0: Moderate-Revenue Segment.
Cluster 0 had an average age of 39.44 years, an average length of stay of 6.91 days, and average hospital revenue of IDR 163,171,150. Relative to the other clusters, this group had the shortest average hospitalization duration and the lowest financial profile. It was therefore labeled the moderate-revenue segment.

Cluster 1: High-Revenue Segment.
Cluster 1 had an average age of 43.40 years, an average length of stay of 8.10 days, and average hospital revenue of IDR 330,793,288. It recorded the highest average age and longest average hospitalization duration. Based on its relative financial position, it was categorized as the high-revenue segment.

Cluster 2: Premium-Revenue Segment.
Cluster 2 had an average age of 43.00 years, an average length of stay of 8.00 days, and average hospital revenue of IDR 994,300,700. It recorded the highest financial profile despite having almost the same average age and length of stay as Cluster 1. It was therefore categorized as the premium-revenue segment.

The labels moderate-revenue, high-revenue, and premium-revenue were assigned only to facilitate managerial interpretation. They represent relative financial profiles and should not be interpreted as indicators of clinical severity, treatment priority, patient importance, or eligibility for hospital services.

Comparative Cluster Characteristics

Table 3. Comparative Characteristics of Patient Segments

Indicator	Cluster 0	Cluster 1	Cluster 2	Main Pattern
Average Age	39.44	43.40	43.00	Relatively similar
Average Length of Stay	6.91	8.10	8.00	Clusters 1 and 2 almost identical

DOI: 10.70184/v1kq1v81

Average Hospital Revenue	IDR 163.17 million	IDR 330.79 million	IDR 994.30 million	Largest observed difference
Segment Classification	Moderate revenue	High revenue	Premium revenue	Relative financial profile

Source: Authors' calculation based on the cluster results.

The comparative results show that hospital revenue exhibited the largest observed variation across the three cluster profiles. By contrast, average age and length of stay varied only modestly, especially between Clusters 1 and 2. This pattern suggests that demographic characteristics and hospitalization duration alone are insufficient to explain the observed financial heterogeneity among inpatient episodes.

Visualization of Cluster Distribution

The clustering visualization showed that Cluster 0 was concentrated in the lower-revenue region and generally displayed a shorter average length of stay. Cluster 1 and Cluster 2 occupied similar positions in terms of age and hospitalization duration but were separated by substantial differences in their revenue profiles.

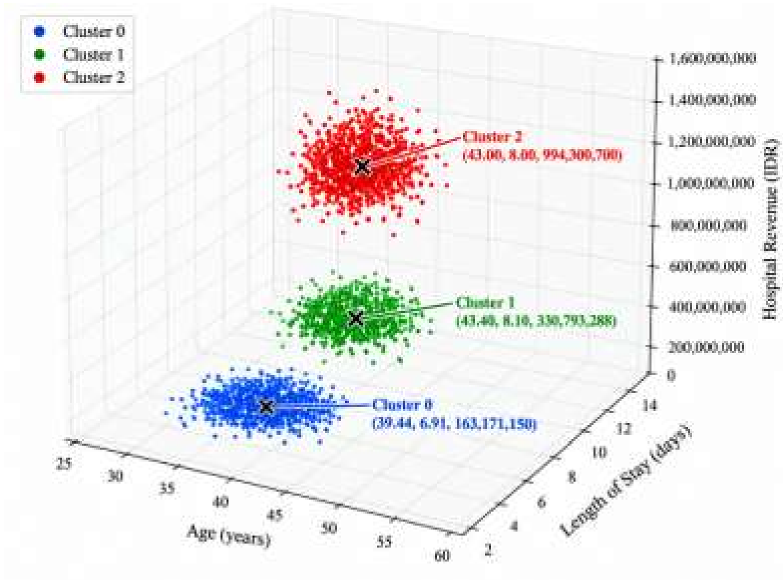

Figure 1. Distribution of Patient Clusters Based on Age, Length of Stay, and Hospital Revenue

The cluster visualization was consistent with the centroid profiles. Cluster 0 was concentrated in the lower-revenue region, whereas Clusters 1 and 2 showed relatively similar positions in terms of average age and hospitalization duration but differed substantially in their financial profiles. The visual results therefore reinforce the finding that financial heterogeneity may persist even among patients with broadly similar demographic and operational characteristics.

Overall, four principal findings emerged. First, the inpatient dataset exhibited substantial heterogeneity in age, length of stay, and hospital revenue. Second, K-Means clustering identified three distinguishable inpatient segments. Third, hospital revenue showed considerably greater differences across the cluster profiles than age or length of stay. Fourth, Clusters 1 and 2 demonstrated that similar demographic and hospitalization profiles can coexist with markedly different financial outcomes.

b) **Discussion**

The findings demonstrate the potential of K-Means clustering to transform routinely collected inpatient data into strategically interpretable patient segments. The three identified clusters differed most visibly in their financial profiles, while average age and length of stay remained relatively similar. This finding supports the proposition that integrating demographic, operational, and financial dimensions can reveal forms of inpatient heterogeneity that may not be visible when relying on hospitalization duration alone.

Cluster 0 represented the moderate-revenue segment and had the shortest average length of stay. From a managerial perspective, this segment may be associated with opportunities to emphasize efficient care pathways, discharge coordination, and bed-turnover management. However, its lower financial profile should not be interpreted as indicating lower clinical importance or lower priority of care.

Cluster 1 occupied an intermediate financial position and had the longest average length of stay. This combination suggests that managers may need to consider both hospitalization duration and service-resource requirements when planning bed capacity and coordinating inpatient services. Nevertheless, the present study cannot determine the underlying causes of this profile because diagnosis, procedure type, treatment intensity, and other clinical variables were not included in the clustering model.

Cluster 2 represented the premium-revenue segment. Its most notable characteristic was that its average age and length of stay were almost identical to those of Cluster 1, while its average hospital revenue was substantially higher. This finding is managerially important because it indicates that hospitalization duration alone cannot adequately represent the financial characteristics of an inpatient episode. Differences in medical procedures, specialist involvement, technology use, pharmaceuticals, class of care, or payment mechanisms may help explain this pattern, although these factors were not directly examined in this study.

These results strengthen the managerial argument for integrating operational and financial information in hospital planning. Bed utilization, for example, should not be evaluated solely through length of stay. Two patient groups with comparable hospitalization durations may require very different combinations of procedures, specialist services, technology, medications, and supporting resources. Therefore, linking patient-flow indicators with financial information may provide a more comprehensive basis for resource-allocation and capacity-planning decisions.

The findings also support the use of differentiated managerial strategies across patient segments. For the moderate-revenue segment, hospitals may emphasize process efficiency and coordinated discharge planning. For the high-revenue segment, management may focus on integrated service coordination and resource scheduling. For the premium-revenue segment, attention may be directed toward ensuring the availability of specialized services and high-intensity resources when clinically required. These implications should be interpreted as management-oriented possibilities rather than deterministic prescriptions derived directly from the clustering algorithm.

From a theoretical and analytical perspective, the study extends patient-segmentation research by integrating demographic, operational, and financial dimensions within a single exploratory framework. Its contribution does not lie in proposing a new clustering algorithm but in demonstrating how age, length of stay, and hospital revenue can be jointly analyzed and translated into strategically meaningful patient profiles. In particular, the marked financial difference between Clusters 1 and 2 illustrates the analytical value of incorporating financial information into patient segmentation rather than relying exclusively on conventional operational indicators.

However, the interpretation of these clusters remains subject to important limitations. The model includes only age, length of stay, and hospital revenue and therefore cannot identify the specific clinical or administrative mechanisms underlying financial differences. Variables such as diagnosis, procedure type, disease severity, class of care, insurance category, payment method, treatment intensity, and patient outcomes should be incorporated in future studies. Further research should also assess the robustness of the segmentation using additional validation indices and alternative clustering approaches such as hierarchical clustering, K-medoids, or Gaussian mixture models

5) Conclusion

This study identified three strategic inpatient segments based on age, length of stay, and hospital revenue using K-Means clustering. Cluster 0 represented the moderate-revenue segment, characterized by the lowest average age, shortest length of stay, and lowest financial profile. Cluster 1 represented the high-revenue segment, with the highest average age, longest length of stay, and an intermediate financial profile. Cluster 2 represented the premium-revenue segment and exhibited an average age and length of stay that were nearly identical to those of Cluster 1, while generating a substantially higher financial profile. These findings demonstrate that inpatient groups with relatively similar demographic and operational characteristics may exhibit markedly different financial profiles.

The study contributes to strategic hospital analytics by demonstrating the analytical value of integrating demographic, operational, and financial dimensions within a single exploratory patient-segmentation framework. Rather than proposing a new clustering algorithm, its contribution lies in translating the resulting clusters into managerially relevant patient profiles. The findings suggest that patient segmentation can complement conventional operational indicators by providing additional information for service differentiation, inpatient capacity planning, resource allocation, and revenue management. However, these financial profiles should not be interpreted as indicators of clinical severity, treatment priority, or patient importance, and managerial decisions should remain consistent with clinical needs and ethical principles.

This study is subject to several limitations. The clustering model incorporated only age, length of stay, and hospital revenue and was based on data from a single private hospital, thereby limiting the broader generalizability of the findings. Important clinical and administrative factors, including diagnosis, disease severity, medical procedures, treatment intensity, class of care, insurance category, payment method, and patient outcomes, were not incorporated into the analysis. Although the Elbow Method and Silhouette coefficient were used to support the selection of the three-cluster solution, future studies should further examine cluster robustness using additional validation measures, such as the Davies-Bouldin and Calinski-Harabasz indices. Future research should also include data from multiple hospitals, incorporate broader clinical and administrative variables, and compare K-Means with alternative clustering approaches such as hierarchical clustering, K-medoids, DBSCAN, or Fuzzy C-Means to assess the consistency and managerial usefulness of the resulting patient segments.

Statement on the Use of Generative Artificial Intelligence (AI)

During the preparation of this manuscript, the authors used **ChatGPT (OpenAI) and Grammarly** solely to support language editing, grammar correction, clarity, and readability. All AI-assisted content was critically reviewed, revised, and validated by the authors. These tools were not used as authors or as substitutes for scientific judgment, data analysis, interpretation of findings, or the formulation of conclusions. The authors take full responsibility for the accuracy, originality, integrity, and final content of the manuscript. This disclosure complies with the journal's Policy on the Use of Generative Artificial Intelligence (GenAI).

References

Akbari-Moghaddam, M., Li, N., Down, D. G., & Hands, K. (2025). Patient segmentation and resource allocation for tailored healthcare delivery. *Journal of the Operational Research Society*. https://doi.org/10.1080/01605682.2025.2546058

AlMuhaideb, S., Bin Shawyah, A., Alhamid, M. F., et al. (2024). Machine learning-guided length of stay prediction for cardiac patients. *Healthcare, 12*(11). https://doi.org/10.3390/healthcare12111110

Deschepper, M., De Smedt, C., & Colpaert, K. (2025). A literature-based approach to predict continuous hospital length of stay in adult acute care patients using admission variables. *International Journal of Medical Informatics, 193*, 105678. https://doi.org/10.1016/j.ijmedinf.2024.105678

Frangky, F., Sinaga, R., & Raihansyah, M. (2025). Analisis segmentasi pasien berdasarkan persepsi kualitas pelayanan dengan algoritma clustering. *Explorer, 5*(1). https://doi.org/10.47065/explorer.v5i1.1818

Hewner, S., Sullivan, S. S., Yu, G., et al. (2023). Identifying high-need primary care patients using nursing knowledge and machine learning methods. *Journal of Nursing Scholarship, 55*(5). https://doi.org/10.1111/jnu.12875

Jain, R., Singh, M., Rao, A. R., & Garg, R. (2024). Predicting hospital length of stay using machine learning on a large open health dataset. *BMC Health Services Research, 24*. https://doi.org/10.1186/s12913-024-11238-y

Khalil, H., et al. (2025). Implementing value-based healthcare: A scoping review of models, enabling factors, and barriers. *BMC Health Services Research*. https://doi.org/10.1186/s12913-025-12698-6

Khedr, A., Hassan, E., Asim, R., et al. (2025). The impact of a novel transfer process on patient bed days and length of stay: A five-year comparative study. *International Journal of Environmental Research and Public Health, 22*(6). https://doi.org/10.3390/ijerph22060871

Kim, J., Jang, E., Kwon, S., & Song, M. (2025). Unsupervised clustering of 41,728 emergency department visits: Insights into patient profiles and KTAS reliability. *Healthcare, 13*(23). https://doi.org/10.3390/healthcare13233073

Lee, H., Kim, Y., Lee, J., et al. (2024). Hospital length of stay prediction for planned admissions using standardized electronic health record data. *Journal of Medical Internet Research, 26*. https://doi.org/10.2196/59260

Liu, P., Wang, Z., Liu, N., & Peres, M. A. (2023). A scoping review of the clinical application of machine learning in data-driven population segmentation analysis. *Journal of the American Medical Informatics Association, 30*(9). https://doi.org/10.1093/jamia/ocad111

Mariam, A., Javidi, H., Zabor, E. C., Zhao, R., Radivoyevitch, T., & Rotroff, D. M. (2024). Unsupervised clustering of longitudinal clinical measurements in electronic health records. *PLOS Digital Health, 3*(10). https://doi.org/10.1371/journal.pdig.0000628

Marlina, M., Hindrawan, D., Kornela, A., & Heikal, J. (2025). Implementation of K-Means clustering algorithm for segmentation of patient visit patterns in public hospitals. *Jurnal Pendidikan Tambusai, 9*(1). https://doi.org/10.31004/jptam.v9i1.26592

Momahhed, S. S., Emamgholipour Sefiddashti, S., Minaei, B., & Shahali, Z. (2023). K-means clustering of outpatient prescription claims for health insureds in Iran. *BMC Public Health, 23*. https://doi.org/10.1186/s12889-023-15753-1

Ng, S. H.-X., Kaur, P., Tan, L. L. C., et al. (2025). Identifying clusters of healthcare expenditure trajectories in end-stage organ disease. *BMC Health Services Research, 25*. https://doi.org/10.1186/s12913-025-13590-z

Pioch, C., Henschke, C., Lantzsch, H., Busse, R., & Vogt, V. (2023). Applying a data-driven population segmentation approach in German claims data. *BMC Health Services Research, 23*. https://doi.org/10.1186/s12913-023-09620-3

Qiu, J., Hu, Y., Li, L., et al. (2025). Deep representation learning for clustering longitudinal survival data from electronic health records. *Nature Communications, 16*. https://doi.org/10.1038/s41467-025-56625-z

Sapitri, A., Nurdin, N., & Afrilia, Y. (2025). Implementation of clustering method using K-Means algorithm for grouping BPJS Health patient medical record data. *Journal of Applied Informatics and Computing, 9*(5). https://doi.org/10.30871/jaic.v9i5.10046

Yasin, P., Yimit, Y., Cai, X., et al. (2024). Machine learning-enabled prediction of prolonged length of stay in hospital. *European Journal of Medical Research, 29*. https://doi.org/10.1186/s40001-024-01988-0

Zamani, H., Parvaresh, F., & Nasr Isfahani, M. (2025). Optimizing patient flow logistics: Strategic challenges, tactical solutions, and operational excellence. *BMC Health Services Research, 25*. https://doi.org/10.1186/s12913-025-13516-9

Zhang, P., et al. (2025). Predicting high-need high-cost pediatric hospitalized patients using machine learning. *Scientific Reports, 15*. https://doi.org/10.1038/s41598-025-99546-z

Zhang, Y., et al. (2025). Clustering electronic health record data to identify distinct clinical trajectories and prognostic subgroups among patients with atrial fibrillation. *Scientific Reports*. https://doi.org/10.1038/s41598-025-91287-3