ICDSOS
The 3rd International Conference on Data Science and Official Statistics
November 27 - 28, 2025

"From Data to Impact: Leveraging Innovation in Data Science and Official Statistics for Data-Driven Policy Making and Global Competitiveness"

Identifying Stratifications of Cancer Patient Visits: Approach of Clustering Using PCA of Mixed Data

K Yunitaningtyas[1,*] and Herianti[2]

[1] Center for Health Financing, Ministry of Health Indonesia
[2] Secretariat of Health Development Policy Agency, Ministry of Health Indonesia

*Corresponding author's email: kristiana.yunitaningtyas@kemkes.go.id

Abstract. Cancer is a significant contributor to the burden of non-communicable diseases and one of the diseases with the highest costs in Indonesia's health insurance system. Understanding key factors influencing cancer patient visits and risk groups under national health insurance supports evidence-based and sustainable cancer care financing. The aim is to identify key factors influencing inpatient visits among cancer survivors and map risk patterns to improve cancer health service policies, using a 1% sample of claim data from the national health insurance (JKN) program. The PCA of mixed data analysis revealed that cost-severity level and contribution-ward classes shared influence of the visits. After PCA, K-Means was applied and 4 clusters were obtained. K-Means can give better understanding of the patient visits, especially the need for distinct strategies to be implemented for the groups so that the burden of cancer disease financing under the national health insurance program can be reduced.

Keyword: Cancer Patients, Claim Data, Clustering, JKN, Principal Component Analysis

1. Introduction

National social security in Indonesia is administered according to Law of the Republic of Indonesia Number 40 of 2014 concerning the National Social Security System. Social security is a government-mandated system designed to provide social protection so that residents can meet their basic needs for a decent life [1]. One part of the national social security system is the National Health Insurance, abbreviated as JKN, in the form of health protection so that participants receive health care benefits and protection in meeting basic health needs which is provided to everyone who has paid contributions or whose contributions are paid by the central or regional government.

In general, there are two types of payments to healthcare providers in the health insurance system: retrospective and prospective payments. Retrospective payments are made or agreed upon after the provision of services, such as fee-for-service, payment per itemized bill, and payment per diem. Prospective payments are made or agreed upon in advance before the provision of services, including case-based/casemix payments, capitation payments, and global budgets. In Indonesia's national health insurance system, payments to healthcare facilities use prospective payments: capitation at primary healthcare facilities and case-based payments at advanced healthcare facilities. These payments are administered by a legal entity called the Health Social Security Administering Body (BPJS Kesehatan) to healthcare facilities used to provide individual healthcare services, including promotive, preventive, curative, and rehabilitative services.

The case-based group system currently used for hospital payments is the Indonesian Case Based Groups (INA-CBG) [2]. There are approximately 22,000 diagnosis and procedure codes grouped into 1,075 INA-CBG group codes consisting of 786 inpatient codes and 289 outpatient codes. Case grouping is done using a software algorithm-based technology called a grouper [3]. Grouping by grouper is applied to all types of diseases including catastrophic diseases.

Catastrophic illness is an illness that requires special expertise because its therapy uses sophisticated medical equipment and/or requires lifelong health services [4]. Catastrophic illness is an illness that requires special expertise because its therapy uses sophisticated medical equipment and/or requires lifelong health services. Based on data from the World Health Organization (WHO), globally in 2021, noncommunicable diseases were the major leading causes of death, representing 68% of the top ten causes. Cancer remains one of the leading causes of death globally. In 2022, nearly 20 million new cases and 9.7 million deaths were recorded worldwide. By 2050, annual new cases are projected to reach 33 million, with cancer-related deaths expected to rise to 18.2 million [5][6]. The cancer burden continues to rise with heavy physical, emotional, and financial pressure on individuals, families, community, and health systems. Many health systems in low-middle income countries are least equipped to cope with this burden and large numbers of cancer patients do not have access to quality diagnosis and treatment. In contrast, countries with strong health systems, survival rates of many types of cancers are improving due to better early detection access, effective treatment and survivorship care [5].

Based on Basic Health Survey (Riset Kesehatan Dasar) 2018 [7], the prevalence of cancer in all ages based on doctor diagnosis is 1.79% (per thousand) and more recently, according to Indonesian Health Survey data (2023) national prevalence (per thousand) of cancer in all ages based on doctor diagnosis is 1.2% with the highest prevalence is in the age range between 55-64 years old (3.2%). Women (2.0%) have higher prevalence than men (0.5%) [8]. Although there is a notable decrease, cancer prevalence typically changes slowly over time, and differences in survey design, measurement methods, or reporting practices may partly explain the variation. In 2023 BPJS Kesehatan reported top 10 number of cases and costs of catastrophic diseases in national health insurance participants, where cancer is in second place in terms of the number of cases and the highest costs under heart disease, namely around 3.8 million cases with a total cost reaching 5.9 trillion rupiah [9]. This indicates that cancer is a significant contributor to the burden of non-communicable diseases in Indonesia and is one of the diseases with the highest costs in the national health insurance system, due to its increasing prevalence, the need for long-term therapy, and the high cost of treatment. This situation can impact the sustainability of JKN financing and place a significant social and economic burden on patients' families. Unlike the other diseases, cancer care involves complex, long term, and often high-cost treatments such as chemotherapy, radiotherapy, and surgery. These variations in treatment intensity, cost, and outcomes make cancer a major driver of health expenditures and resources.

This study applies the PCA Mixed Data method to identify key factors influencing cancer patient visits and uses K-Means to classify risk groups under the national health insurance program, providing evidence-based insights for sustainable cancer care financing.

2. Research Method

2.1. Data source and variables

The data used in this study is secondary data, namely 1% sample data from the National Health Insurance program (JKN) issued by BPJS Kesehatan in 2023. This sample data represents all membership and health service data, consisting of membership data and claims from the utilization of JKN services by participants at health service facilities collaborating with BPJS Kesehatan. The 1% JKN sample data has gone through a standardization and extraction process consisting of service activities claimed by health facilities to BPJS Kesehatan. The data used is individual data on patients recovered from hospitalization based on visits which are then combined with membership data with a total of 2,935 rows of data in 2023. Individual data is filtered only with the Casemix Main Group code in INA-CBG, the first digit of which is code C, which is myeloproliferative system and neoplasm.

The variables contained in the data in this study are:

a. Patient age is the patient's age when visiting JKN healthcare services at the hospital;
b. Length of hospitalization is the number of days the patient was hospitalized, calculated from the time the patient was admitted to the hospital until the patient was discharged;
c. Claim cost is the total cost of one episode of care for each patient visit after verification by BPJS Kesehatan;
d. Gender is the gender of the patient accessing JKN healthcare services at the hospital;
e. Ward class is the level of inpatient care privileges a patient receives during hospitalization. These are: Class I (the best inpatient facility), Class II (medium inpatient facility), and Class III (basic inpatient facility). Inpatient class usually aligns with the premium class, but patients have the right to upgrade it;
f. Patient contribution class is the monthly premium payment level for JKN participants, consisting of Class 1 (the highest premium), Class 2 (medium premium), and Class 3 (the lowest premium); and
g. Case severity level, is the third digit of the INA-CBG code, indicates the severity of the case, influenced by the presence of comorbidities or complications during treatment that consist of mild (severity level 1), moderate (severity level 2), and severe (severity level 3).

These variables age, length of stay, claim cost, gender, patient class, premium class, and case severity level are linked with disease risk (likelihood and severity of illness), resource utilization (hospital services, length of stay, medications), and service level (class of care and insurance coverage). These factors together are the main drivers of health insurance claim frequency and cost. The software used to explore the data is Microsoft Excel and to analyze using RStudio.

2.2. Data analysis methods

2.2.1. Data preprocessing.

Data normalization is used as preprocessing technique in this study. This technique adjusts or scales variables so they contribute equally during analysis. It is a crucial step in preparing data to ensure better classification performance before applying machine learning algorithms. By transforming features into a common range, normalization prevents larger numerical values from overshadowing smaller ones, to have a mean of zero and standard deviation of one. The primary goal is to reduce bias from features with higher numerical influence when distinguishing between pattern classes [10]. Z-Score is one of data normalization methods or generally called as standardization. The standard deviation of a dataset indicates how widely the scores are spread out from their mean. In the data used in this study, the numerical variables have striking differences in scale where age and length of stay have different units with the cost (millions of rupiah). Normalization of the dataset with respect to its standard deviation is achieved by dividing each observation by the standard deviation of the distribution. The mean of the dataset is subtracted from each observation then divided by standard deviation with formula (1).

$$Z_n = \frac{Y_n - M}{\hat{S}} \tag{1}$$

where Z_n is the Z-score of N data, Y_n is a set of scores of N data, and $M$is mean of N data [11]. Through basic algebra, it can be shown that a Z-score has a mean equal of zero and a standard deviation of one, making it suitable for comparing observations with different units present in the variables in the research data.

2.2.2. Principal Component Analysis (PCA).

The PCA was discovered by Pearson in 1901 and then due to its versatility in applications, PCA has been extended in many directions and independently developed by many including Hotelling in 1933, by Karhunen in 1947 and by Loève in 1948 [12]. Principal Component Analysis (PCA) is a widely used technique for reducing the number of variables in a dataset while maintaining most of its original information. It is particularly useful for large or complex datasets that contain numerous variables. PCA is considered one of the most well-known and effective methods for reducing dimensionality of data [13].

The main objection of PCA is to check whether the first few components explain most of the variation in the data. If they do, it means the problem can be described with fewer dimensions than the original number of variables. This happens because some variables may be highly correlated, essentially providing the same information. In such cases, the first few components should ideally be meaningful, make the data easier to understand, and allow future analyses to be done with fewer variables. Therefore, it is often the case that an examination of the reduced dimension data set will allow the user to spot trends, patterns, and outliers in the data that would be difficult to perform in original data [14].

A data matrix *X* is the input for PCA in which the columns represent different variables and the rows correspond to values measured on the variables. The maximal number of components in PCA is equal to the number of variables or the number of observations, whichever is smaller. The principal components are linear combinations of the original variables, as a weighted sum of the input variables in equation (2).

$$\begin{aligned} PC_1 &= w_{1,1}Var1 + w_{1,2}Var2 \\ PC_2 &= w_{2,1}Var1 + w_{2,2}Var2 \end{aligned} \tag{2}$$

where *wi,j* is a weighting coefficient. Importantly, the first principal component (PC1) points in the direction of the largest variance in the data set. PC2 is orthogonal to PC1 and shows the direction of the second greatest variance, and so forth, until the maximal number of components is reached. The first principal component (PC1) often explains most of the variance and the original high-dimensional data can be represented using only a few keys of principal components. The remaining higher dimensions can be omitted with little loss of information [15].

Standard PCA is generally used for data with numerical variables, while multiple correspondence analysis is used to handle categorical variables. Multiple correspondence analysis (MCA) is a statistical method that extends simple correspondence analysis to explore relationships and visualize data with more than two categorical variables. MCA is a qualitative version of PCA, which is used for quantitative data. MCA helps to understand how different categories within the variables are associated with one another [16]. When data consists of both numeric and categorical variables, another approach to reducing the dimensionality of the data variables must be used, namely PCA on mixed data. Several techniques are available in statistical software, specifically in R through the Factor Analysis of Mixed Data (FAMD) function [17]. FAMD is applied to the data used in this study because it consists of numerical and categorical variables which are important information from each patient visit.

2.2.3. K-Means clustering.

K-Means clustering is a classification method often used to assign data into predefined groups or labels that is included in an unsupervised learning method. In the unsupervised approach, new data is grouped based on similarity to existing data patterns, after the algorithm identifies clusters within the training data. This method generates distinguishing rules or boundaries between groups [18]. In general, the stages in cluster analysis start with obtaining data sources with specific characteristics and ending with evaluating the clustering results. These stages include three main parts: feature/transformation selection, algorithm selection or clustering technique selection, and finally, evaluation of the clustering results [19]. In completing the grouping, there are three main things that must be done, namely finding the measure of similarity between observations, how to form groups or clusters, and determining the number of groups or clusters to be created.

K-Means is a technique for classifying object characteristics. It is one of the simplest forms of clustering analysis and is widely used on unlabeled data. The concept of the K-Means algorithm separates data into several separate parts. K-Means clustering begins randomly and is repeated until groups with similar characteristics are found. Observations with the same characteristics, or in this case, similarity, are grouped together. Other observations with different characteristics are grouped together so that the data within a group has a low level of variation [20].

K-Means uses distance to divide numerical data into a number of groups. The output of K-Means analysis is to classify data by maximizing the similarity of data within a cluster and minimizing the similarity of data between clusters. The similarity measure used in clustering is a distance function. K-Means Clustering will classify the dataset into as many as k clusters, with each cluster having certain characteristics based on the cluster center point (centroid) closest to the data. Therefore, maximizing data similarity is obtained based on the shortest distance between the data and the centroid point.

The Elbow method is a practical approach for determining the optimal number of clusters (k) in the K-Means algorithm by calculating the within-cluster sum of squares (WCSS) for each value of k [22]. The resulting data is visualized in a plot, with the x-axis representing the number of clusters and the y-axis displaying the WCSS values. The point where the decrease in WCSS slows, forming an elbow-like shape, indicates the most optimal number of clusters.

Selecting appropriate evaluation metrics is crucial for assessing clustering performance. Davies-Bouldin Index (DBI), introduced by David L. Davies and Donald W. Bouldin in 1979, is a popular method for choosing the optimal number of clusters [22]. The DBI evaluates clustering quality by specifically calculating the average similarity between each cluster and its most similar neighboring cluster, where a lower overall DBI value indicates the best number of clustering [23].

3. Result and Discussion

3.1. Data exploration

Data exploration is conducted to understand the types, patterns, and characteristics of each variable available in the dataset. The dataset consists of both numerical and categorical variables. Numerical variables are explored to examine measures of central tendency, dispersion, and distribution, while categorical variables are explored to identify frequency, proportion, and the distribution of each category.

Table 1. Descriptive statistics of numeric variables.

Variable	Minimum	Median	Mean	Maximum
Age of patients per visit (years old)	0	49	46	96
Length of Stay (LOS) of patients per visit (day)	1	3	4	56
Cost of patients per visit (Rupiah)	1,518,900	3,158,000	4,742,280	51,130,600

3.1.1 Age of patients per visit.

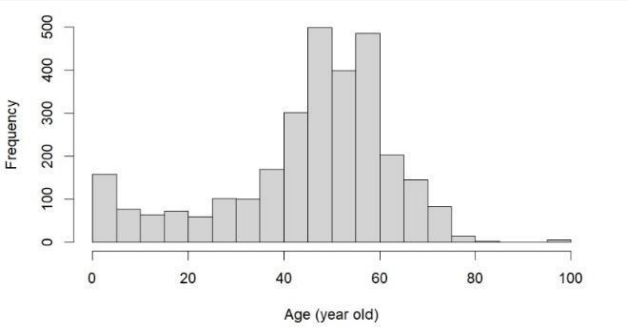

Figure 1. Histogram of Patient Age while visiting inpatient care.

Table 1 shows the median age of cancer patients accessing the hospitals, which is 49 years old, and the mean age is 46 years old, indicating that most patients are middle-aged. The histogram shows the distribution of patient ages that ranges from 0 to 96 years old, which 0 means that there are patients less than 1 year old, namely infants or neonates. Most cancer patients who visit hospitals are between 40 and

65 years old, with the highest frequency around ages 50–60. Fewer patients are observed at younger ages (below 20) and older ages (above 80). The histogram shows the largest concentration of patients between 40 and 65 years old, with fewer patients in younger (below 20) and older (above 80) age groups.

3.1.2 Length of stay (LOS) of patients per visit.

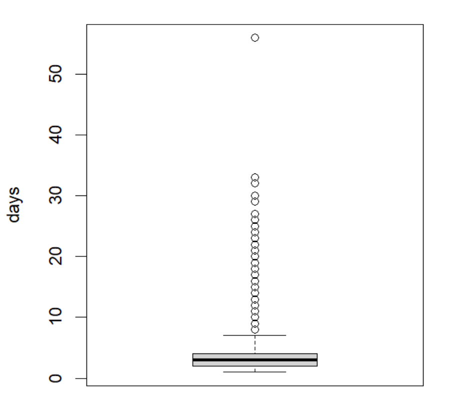

Figure 2. Boxplot of Length of Stay (LOS) patient visits.

From Table 1, it can be observed that the average length of stay for cancer patients in the hospital is 4 days, with a median of 3 days. Most patients were hospitalized for only a few days, as illustrated by the boxplot in figure 2. However, there are numerous outliers, with some patients staying more than 10 days, and an extreme case of 56 days, as also indicated in table 1.

3.1.3 Cost of patients per visit.

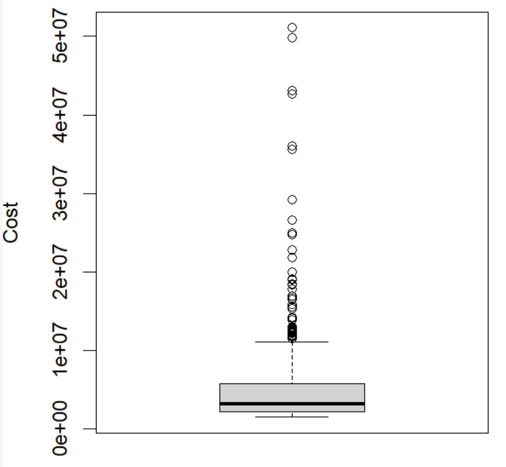

Figure 3. Boxplot of cost of patient visits.

The average cost of a patient for a visit in the hospitals is Rp4,742,280 and the median is Rp3,158,000 according to table 1. The range of patient costs per visit is quite large, as shown in figure 3. The data distribution is heavily right-skewed. The minimum cost of a cancer patient visit is Rp1,518,900 and the maximum is Rp51,130,600. While most patients incurred relatively low costs, there are many outliers with significantly higher expenses. Several numbers of patients had extremely high costs, contributing to the wide variation observed in the data, with many patients having unusually high costs.

3.1.4 Gender of patients per visit.

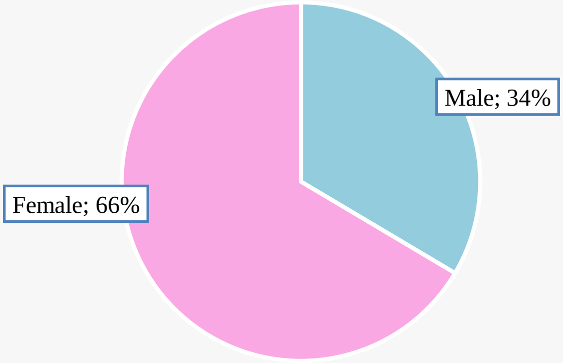

Figure 4. Pie chart of patient gender based on visits.

Patients gender was identified from membership data that has been merged with the individual visit data. Of all patient visits, 984 visits or 34% were made by men, while 1,951 visits or 66% were made by women.

3.1.5 Ward class of patients per visit.

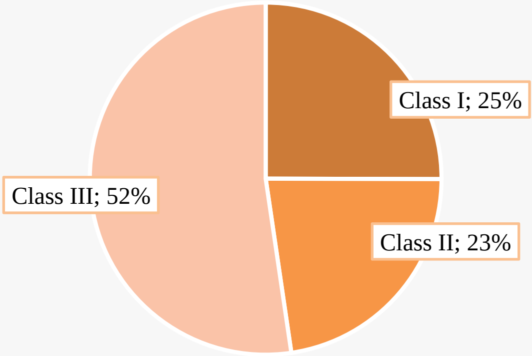

Figure 5. Pie chart of patient ward class based on visits.

This variable is obtained from membership data that identifies inpatient visits according to the type of ward. A large share of cancer patient hospitalizations per visit occurred in the third-class ward (1,536 visits, 52%) followed by first-class ward (735 visits, 25%), and then the least is the second-ward class (664 visits, 23%) according to figure 5.

3.1.6 Contribution class of patient per visit.

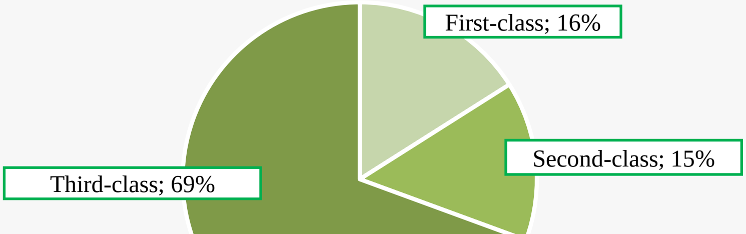

Figure 6. Pie chart of patient contribution class based on visits.

The variable is constructed from membership records indicating the contribution type for each inpatient visit. A substantial proportion of cancer inpatient visits were third-class (2,037 visits, 69%), then to first-class (470 visits, 16%), with the second-class showing the lowest number of visits (428 visits, 15%).

3.1.7 Severity level of case group per visit.

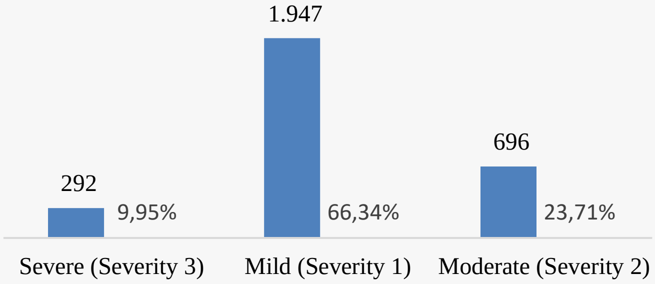

Figure 7. Bar chart of severity levels.

Figure 7 presents the distribution of severity levels among cancer inpatient visits. The majority of visits were classified as mild or severity level 1 (1,947 visits, 66%), followed by moderate or severity level 2 (696 visits, 24%), and severe or severity level 3 (292 visits, 10%). These findings indicate that most cancer patient hospitalizations involve fewer complications and comorbidities.

3.2. Data standardization

Standardization was applied to the numerical variables: age, LOS, and cost using R's scale () function, which performs z-score normalization which centering by the mean and scaling by the standard deviation.

Table 2. The results of standardization.

	Mean	Median	Standard Deviation
Age	0	0.193	1
LOS	0	-0.246	1
Cost	0	-0.403	1

3.3. Results of PCA and formed factors

Table 3. The results of PCA by FAMD technique.

Eigenvalues	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5
Variance	2.322	1.709	1.660	1.609	0.938

Percent (%) of variance	23.216	17.088	16.598	10.690	9.383
Cumulative Percent (%) of var	23.216	40.304	56.902	67.592	76.975

The results of Principal Component Analysis by Factorial Analysis of Mixed Data (FAMD) show that the first five dimensions capture about 76.98% of the total variance, which indicates that the dimensionality reduction still preserves most of the original information. Dimension 1 explains 23.2% of the variance, followed by Dimension 2 with 17.1%, and Dimension 3 with 16.6%, so that the first three dimensions together already account for more than half of the variance (56.9%). Dimensions 4 and 5 contribute smaller shares, at 10.7% and 9.4% respectively, but they still add valuable information to the overall structure of the data. In practice, the first two dimensions can be used to visualize the distribution of individuals and the separation between clusters, while the contribution of variables to these dimensions helps to identify which characteristics drive the grouping. Meanwhile, examining cluster profiles with respect to the original variables, such as cost, severity, and length of stay, allows a more meaningful interpretation of each cluster's distinctive features.

Table 4. The results of PCA by FAMD technique.

Variables	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Continous Variables									
Age	-0.505	10.997	0.255	0.428	10.719	0.183	-0.073	0.32	0.005
LOS	0.514	11.372	0.264	-0.367	7.871	0.134	-0.168	1.718	0.029
Cost	0.805	27.891	0.648	-0.071	0.293	0.005	-0.075	0.338	0.006
Categorical Variables									
Contribution Class 1	1.539	6.667	0.269	2.074	22.334	0.488	-1.319	9.586	0.198
Contribution Class 2	-0.625	6.262	0.174	-0.508	7.987	0.194	-2.861	37.462	-0.603
Contribution Class 3	-0.507	3.454	0.04	-0.554	7.598	0.481	-0.221	3.283	-0.287
Mild (Inpatient Severity 1)	0.915	3.829	0.063	-0.151	0.31	0.007	-0.162	0.205	-0.006
Moderate (Inpatient Severity 2)	0.915	3.329	0.081	-0.018	0.023	0.001	-0.16	2.095	0.002
Severe (Inpatient Severity 3)	-1.637	17.803	0.681	0.339	0.533	0.029	-0.163	0.205	0.002
Male	0.282	0.936	0.107	-0.517	2.185	0.168	-0.089	0.096	0.034
Female	-0.282	0.936	0.107	0.517	2.185	0.168	-0.049	0.049	0.074
Inpatient Class I	0.464	0.645	0.029	-0.692	3.821	0.108	-0.089	0.089	-0.16
Inpatient Class II	0.464	0.645	0.029	0.446	1.103	0.027	2.585	39.235	0.896
Inpatient Class III	-0.507	2.834	0.236	-0.823	13.767	0.622	-0.344	2.397	0.102

From the result summary in table 4, it is divided into two types of variables in forming the principal factors namely continuous and categorical variables. From table 3, the values in Dim.1, Dim.2, Dim.3 show the loading factor for each variable. The loading factor shows the strength of the relationship. A high absolute loading, close to +1 or –1, means the variable is strongly correlated with that principal dimension. A value near 0 means weak or no relation. It also shows the direction of the relationship. Positive loadings mean the variable increases as the dimension increases and negative loadings mean the variable decreases as the dimension increases. Ctr means contribution to the definition of a principal component (%) and cos2 means quality of representation on a given dimension, if cos2 is close to 1 then the variable/category is well represented by that axis.

3.3.1. Continuous variables.
Age has the main contribution in Dimension 1 (loading -0.50) with the value of contribution 10.997% and value of cos2 is 0.25. This means Age is one of the important variables shaping this axis and has moderate representation (not too weak, not too strong). LOS also has quite significant loading in Dimension 1 (0.51) with value of contribution is 11.37% and value of cos2 is 0.26. This means LOS is one of the contribution variables and has moderate representation (not too weak, not too strong). Cost has a very significant loading factor in Dimension 1 (0.80) with the value of contribution in this dimension is 27.89% and value of cos2 is 0.65. Cost is the dominant variable shaping Dimension 1 and Dimension 1 is strongly aligned with Cost. This variable is both important in defining the axis (high ctr) and also well explained by the axis (high cos2).

3.3.2. Categorical variables.
Contribution Class, according to table 2, generally has a higher loading value in Dimension 2. Class 1, which has the highest contribution, has a loading of 2.07, a contribution value of 22.33%, and a cos2 (quality of representation) of 0.49. Class 2 in Dimension 2 has a loading value of -0.51, a contribution value of 7.99%, and a quality of representation of 0.19. Class 3, which has the lowest contribution, has a loading value of -0.554, a contribution value of 7.60%, and a quality of representation of 0.481. Severity level, which generally has a significant contribution to Dimension 1, is shown in table 2, having the largest loading in Dimension 1 compared to the other dimensions. Mild severity has a loading of 0.91, a contribution value of 3.83%, and a quality of representation of 0.06. Meanwhile, moderate severity also has a loading on Dimension 1 of 0.91, contributing 3.33% and a low representation of 0.08. Severe, or the highest severity, contributes significantly to Dimension 1 with a negative loading of -1.64, contributing 17.80%, and is well explained by the axis with a high cos2 value.

Gender, male and female, have small contributions and a low cos2 value in both Dimensions 1, 2, and 3, indicating that gender does not play a significant role in the primary dimension. According to table 2, Ward Class has a fairly significant loading factor on Dimension 2. Class I has a negative loading factor of -0.69, contributing 3.82%, and quality representation, indicated by a cos2 value of 0.11. Class II has a loading factor of 0.45, contributing 1.10% to the principal component, and a representation value of 0.03. Class III also has a negative loading factor, namely -0.82 with a fairly large contribution of 13.77% and is well explained by the axis with cos2 0.62.

3.3.3. Main dimensions.

Table 5. Contribution of each variable to PCA dimensions.

Variable	Dim.1	Dim.2
Age	10.997	10.718
LOS	11.372	7.871
Cost	27.891	0.293
Contribution Class	12.748	31.946
Severity	27.069	4.073
Gender	2.573	6.006
Ward Class	7.350	39.093

Table 5 shows the significant contribution of each variable to the main dimensions. In Dimension 1, the most dominant contributing variables are Cost (27.9%), Severity (27.1%), and those with moderate contributions are Age (11.0%), and LOS (11.4%). Dimension 1 primarily separates patient visits based on treatment costs and disease severity, with additional influences from age and length of stay. Meanwhile, Dimension 2 is dominated by the variables Ward Class (39.1%) and Contribution Class (31.9%), with Age (10.7%) still having a significant influence. Dimension 2 is more related to the separation of patient visits based on inpatient room segmentation and participant contributions. The contribution of each variable can also be seen through the plots in figure 8 below.

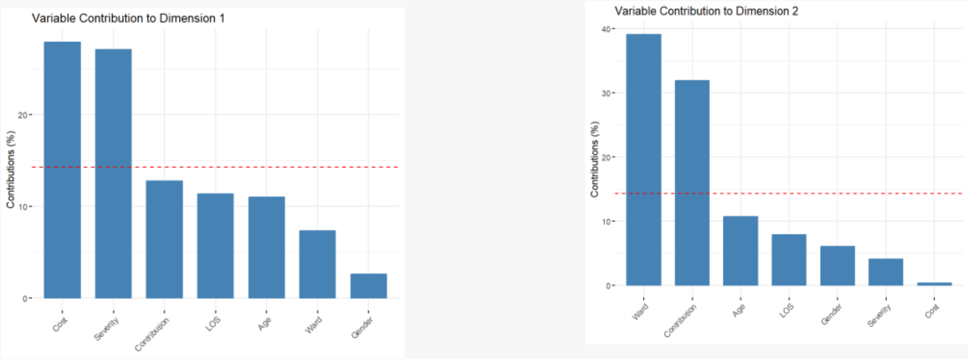

Figure 8. Plots of variable contribution to Dimension 1 and 2.

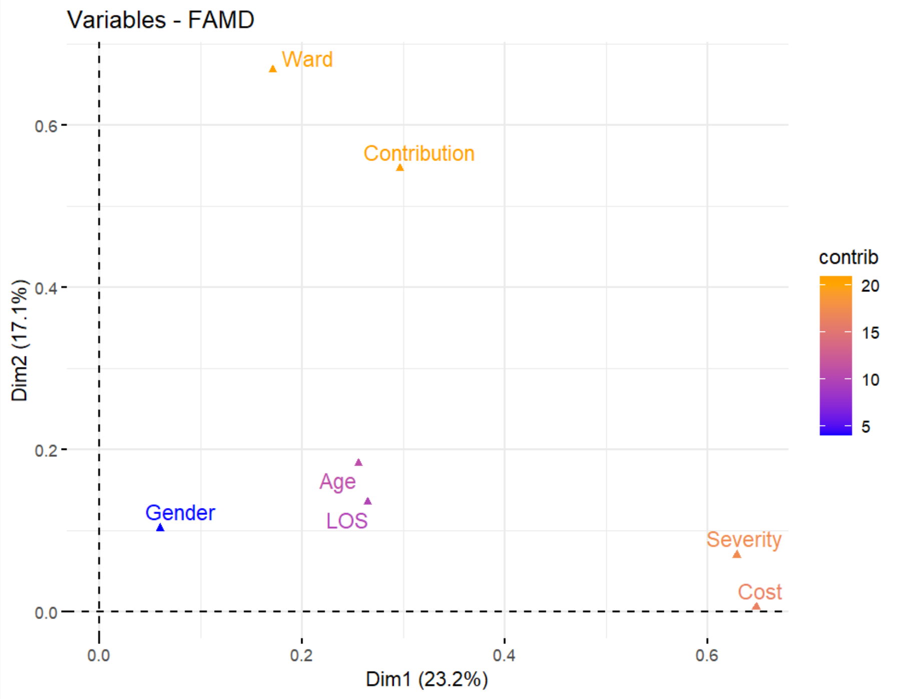

Figure 9. Two-dimensional contribution variable plot.

A two-dimension contribution plot in figure 9 visualizes the contribution of variables, measuring the extent to which each variable influences the formation of a specific principal component axis. A higher contribution value indicates a greater impact on the formation of the principal component. Dimension 1, which explains 23.2% of the variance, distinguishes patients based on treatment costs and the Severity category, with additional influence from age and length of stay. The two dominant variables in Dimension 1 are Cost (27.9%) and Severity (27.1%), while age and length of stay per patient visit contribute moderately.The variance of Dimension 2, which is 17.1%, is associated with the Contribution category (31.9%) and the Ward variable (39.1%).

3.4. K-Means clustering analysis

The results of PCA using the FAMD technique show that four variables are significant and dominant contributors in two dimensions. These are collectively used for K-Means analysis. Determining the optimal number of clusters is a key component of the K-Means algorithm. Two methods are presented for determining the optimum k value: the Elbow Method and the Davies-Bouldin Index (DBI).

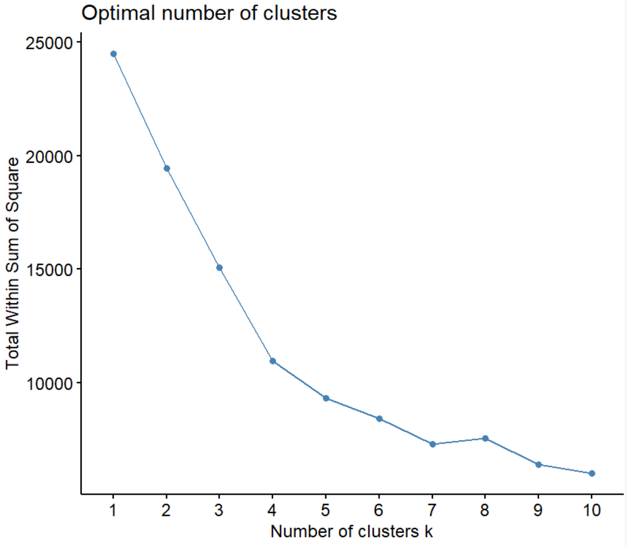

Figure 10. Elbow Plot of optimal number of clusters.

Figure 10 shows the Elbow Plot results from determining the k value, with the X-axis representing the optimal number of clusters and the Y-axis representing the total within-cluster sum of squares (WCSS). Optimal clustering is determined when there is an elbow point where the WCSS decline begins to slow. The graph shows that from k=1 to k=4 the WCSS value decreases sharply, and after k=4 the decline begins to slow. Therefore, it can be said that the optimal cluster is at k=4.

Table 6. Davies-Bouldin Index for optimal cluster.

Number of Cluster	DBI
2	0.933
3	0.991
4	0.880
5	1.109
6	1.252
7	1.414
8	1.467

In addition to the Elbow Plot, the optimal number of clusters is also seen through the Davies-Bouldin Index (DBI) criteria for each number of clusters. The lower the DBI value, the more optimal the cluster formed. Table 4 shows that the cluster with a number of 4 has the lowest DBI value compared to the other clusters, namely 0.880. This DBI value strengthens the evidence that clusters or k = 4 is the most optimal number of clusters, so a K-Means analysis was conducted for visits to cancer patients in inpatients with a number of clusters of 4.

3.5. Visit segmentations of cancer patients based on K-Means results

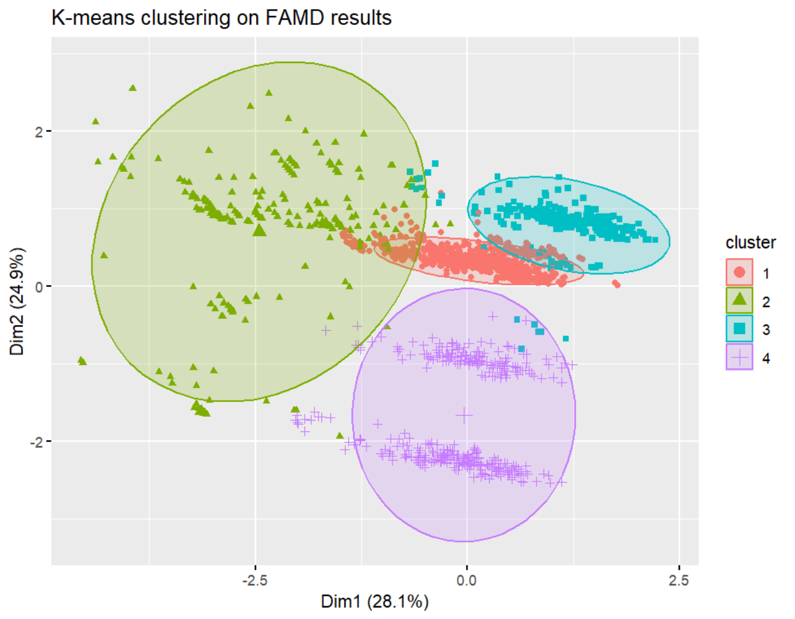

Figure 11. K-Means Clustering result of cancer patient visits.

Figure 11 shows the results of K-means clustering analysis applied to data that has been reduced using FAMD. The plot displays the data points on a two-dimensional graph, with Dimension 1 (28.1% of total variance) captures variations in treatment cost and disease severity and Dimension 2 (24.9% of total variance) represents differences in ward class and contribution level. From the clustering results, 4 clusters were obtained with each cluster representing different characteristics. The higher Dimension 1 means the higher cost patient visits and severity level of cancer, meanwhile the higher value of Dimension 2 means the lower class of the patient segmentation on ward and contribution (Class III of ward and Class 3 of contribution are the basic).

The characteristics of each cluster can be defined as:

a. Cluster 1 (Red Circles): This cluster has 1,553 visits, is located on the right of the Dimension 1. It is indicating that in this group the visits have moderate cost and disease severity. Therefore, these patients are positioned at the positive value and top-middle of the Dimension 2, they fall into a segment of the mid-level contribution amount and ward class. This could represent a group of patients with standard ward class and basic contribution that incurs moderate costs and conditions;

b. Cluster 2 (Green Triangles): Situated at the top of the plot, this cluster has 293 visits. It shows low to moderate cost and severity, but is distinctly high on patient segmentation. This group likely represents a specific segment, potentially a different type of contribution or ward class, but the majority located in the top of the plot. This group indicates the visits that do not have the highest disease severity and have higher ward and contribution level but still require a certain level of care;

c. Cluster 3 (Cyan Squares): This cluster has 449 visits, located in the far right of the plot, with high cost and severe severity and a top-middle position on Dimension 2 that shows the patients have high contribution and ward class. Cluster 3 describes patients with severe/complex cases requiring costly care but treated in higher-class wards with possibly more affluent or high-contribution patients;

d. Cluster 4 (Purple Plus Signs): This cluster has 640 visits, represents moderate cost and severity (values on Dimension 1 around 0). The wide range of Dimension 2 suggests this group is diverse in terms of their contribution and ward class but are unified by their relatively lower values and the patients mostly are from lower class ward and contribution. This could represent financially vulnerable patients facing severe illness.

Combining PCA with K-Means clustering has significant benefits. Firstly, it is used to segment the insured population into meaningful groups. This analysis can identify distinct groups with similar characteristics so it can provide an overview of the services provided to each group instead of treating all cancer patients the same. Secondly, once the distinct segments have been identified, the health insurance provider can tailor policies and services to meet the specific needs of each group, for instance:

a. For patients in cluster 1, the preventive and early detection programs should be introduced so they do not progress to severe cases with higher cost;
b. For patients in cluster 2, ensure the resource used and promote day care or outpatient treatment models so the burden of inpatient can be decreased;
c. For patients in cluster 3, need to be introduced to co-payment or risk sharing for those who have high-income to be an effort to reduce the burden of national health insurance program;
d. For patients in cluster 4, they need to be prioritized to get the social subsidies and strengthen the access to the national health insurance program so they will not be burdened by the financial problems.

Furthermore, the clustering logic also can be used to detect unusual or fraudulent activity related to cancer treatment in health national insurance. After clustering the data of all service providers, any entity that falls outside a cluster, or identified as an outlier, could be flagged for further investigation. The insights gained from clustering can also inform financial models and help set contribution rates or forecast costs more accurately specifically in maintaining the sustainability of health insurance financing. Finally, the clusters formed can be used to predict future healthcare costs for each group of cancer patient visits.

4. Conclusion

Two main dimensions were derived from seven variables describing cancer patient visits in the national health insurance inpatient data, using PCA for mixed data (FAMD). Dimension 1 reflects cost and disease severity, while Dimension 2 captures participant segmentation based on contribution and ward classes. Applying K-Means clustering to these dimensions produced four distinct groups.

The integration of PCA for mixed data and K-Means enables the identification of patient segments with similar characteristics, detection of anomalous or irregular patterns, and estimation of healthcare costs per group. These insights provide valuable evidence for optimizing contribution rates, improving cost forecasting, and enhancing the financial sustainability of the national health insurance program.

Acknowledgement

The authors would like to thank the Center for Health Financing and Health Policy Agency, Ministry of Health Indonesia for supporting this paper.

References

[1] A. Wijaya, *Hukum Jaminan Sosial Indonesia*. Jakarta: Sinar Grafika, 2017.

[2] Ministry of Health Indonesia, *Regulation of the Minister of Health Number 3 of 2023 concerning Standard Tariffs for Health Services in the Implementation of the Health Insurance Program*, Jakarta: Ministry of Health, 2023.

[3] Ministry of Health Indonesia, *Minister of Health Regulation Number 26 of 2021 concerning Guidelines for Indonesian Case Base Groups (INA-CBG) in the Implementation of Health Insurance*, Jakarta: Ministry of Health, 2021.

[4] H. Heniwati and H. Thabrany, “Perbandingan Penyakit Katastropik Peserta Jaminan Kesehatan Nasional di Provinsi DKI Jakarta dan Nusa Tenggara Timur,” *Jurnal Ekonomi Kesehatan Indonesia*, vol. 1, no. 2, December 2016.

[5] World Health Organization, “Health Topics: Cancer”, [Online]. Address: https://www.who.int/health-topics/cancer [access on 15 August 2025].

[6] National Cancer Institute, “Cancer Statistics”, [Online]. Address: https://www.cancer.gov/about-cancer/understanding/statistics [access on 12 October 2025].

[7] Ministry of Health Indonesia, *Basic Health Research in Numbers*, Jakarta: Penerbit Litbangkes, 2018.

[8] Ministry of Health Indonesia, *Indonesia Health Survey in Numbers*, Jakarta: BKPK, 2023.

[9] BPJS Kesehatan, *Health Insurance Program Implementation Report*, Jakarta: BPJS Kesehatan, 2023.

[10] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance”, *Applied Soft Computing*, volume 97, part B, December 2020.

[11] H. Abdi, B. Edelman, D. Valentin, W.J. Dowling, *Experimental Design and Analysis for Psychology*, Oxford: Oxford Press, 2009.

[12] A. Torokhti, S. Friedland, “Towards theory of generic Principal Component Analysis”, *Journal of Multivariate Analysis*, volume 100, April 2009.

[13] I. Jolliffe, “Principal Component Analysis”, *Lovric, M. (eds) International Encyclopedia of Statistical Science. Springer*, January 2014.

[14] M. Richardson, “Principal Component Analysis”, May 2009, [Online]. Address: https://people.maths.ox.ac.uk/richardsonm/SignalProcPCA.pdf [access on 15 July 2025].

[15] H. Todorov, D. Fournier, G. Susanne, “Principal components analysis: theory and application to gene expression data analysis”. *Genomics and Computational Biology*, 4. 100041, January 2018

[16] ZI Kalantan, NA. Alqahtani, “A study of principal components analysis for mixed data”, *International Journal of Advanced and Applied Sciences*, 6(12), October 2019.

[17] M. Chavent, V. Kuentz, A. Labenne, J. Saracco. “Multivariate analysis of mixed data: The R Package PCA mix data”, *Electronic Journal of Applied Statistical Analysis*, *15*(3), 606-645, 2022.

[18] K. Yunitaningtyas, A.M. Yolanda, “Klasifikasi Kabupaten/Kota di Provinsi Nusa Tenggara Timur Berdasarkan Indikator Status Kesehatan Masyarakat”, 2022, JSTAR 2(1), 1-19, July 2022.

[19] Z. Liu, C. Ren, W. Cai, “Overview of Clustering Analysis Algorithms in Unknown Protocol Recognition”, *MATEC Web of Conference*, CSCNS2019, 309 (03008), 2020.

[20] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms 2nd Edition*, New Jersey: John Wiley & Sons Ltd, 2011.

[21] R. Singh, R. Reddy, V. Kapoor, P. Churi, “K-means clustering analysis of crimes on Indian women”, *Journal of Cybersecurity and Information Management (JCIM)*, *4*(1), 5-25, 2020.

[22] S. Nawrin, Md. Rahman, S. Akhter, “Exploring K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System”, *International Journal of Advanced Computer Science and Applications,* IJACSA 2017.080337, 2017.

[23] A.A. Vergani and E. Binaghi, “A soft davies-bouldin separation measure”, *IEEE International Conference on Fuzzy Systems (FUZZIEEE)*, pp. 1-8, 2018