Algorithmic Prediction of Health-Care Cost

Samuel Keiling

CONTENTS

01
Introduction
Global Challenge of Rising Health Care Costs

02
Problem Statement
Limitations of Traditional Methods & Benefits of Data Mining

03
Data Mining Process
Review of Research Objective and Variables

04
Baseline Method
Using the Traditional methods

05
***Classification Trees**
Data Mining Method 1

06
***Clustering**
Data Mining Method 2

07
Key findings
Results and comparison

08
Conclusion
Implication of the study

01

Introduction

Global Challenge of Rising Health Care Costs

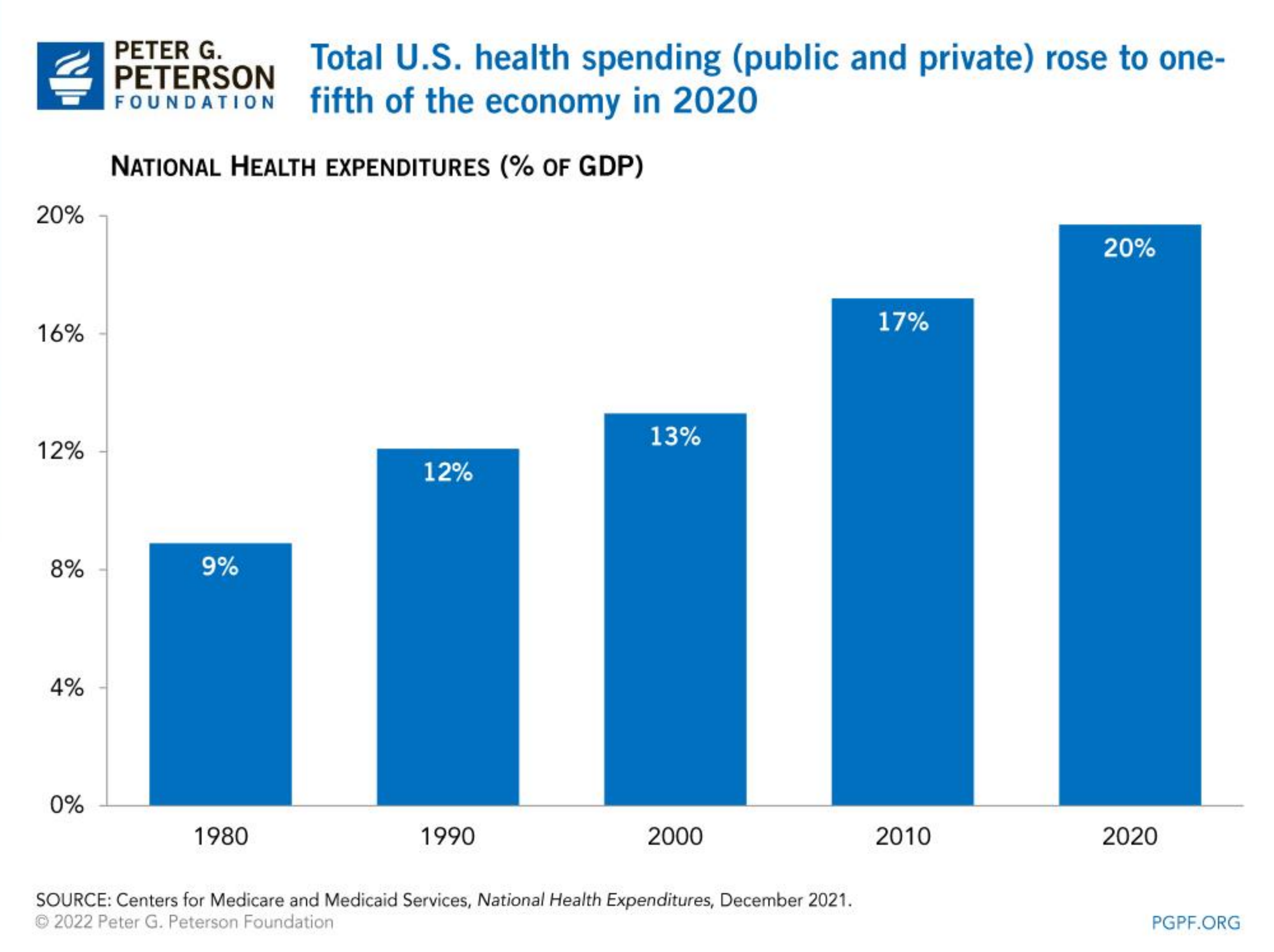
Fundamental Problem

The rising cost of health care is one of the world’s most important problems

Accordingly, predicting such costs with accuracy is a significant first step in addressing this problem.

01 Insights.

Rising health care costs are outpacing economic growth in many countries



PETER G. PETERSON FOUNDATION

Total U.S. health spending (public and private) rose to one-fifth of the economy in 2020

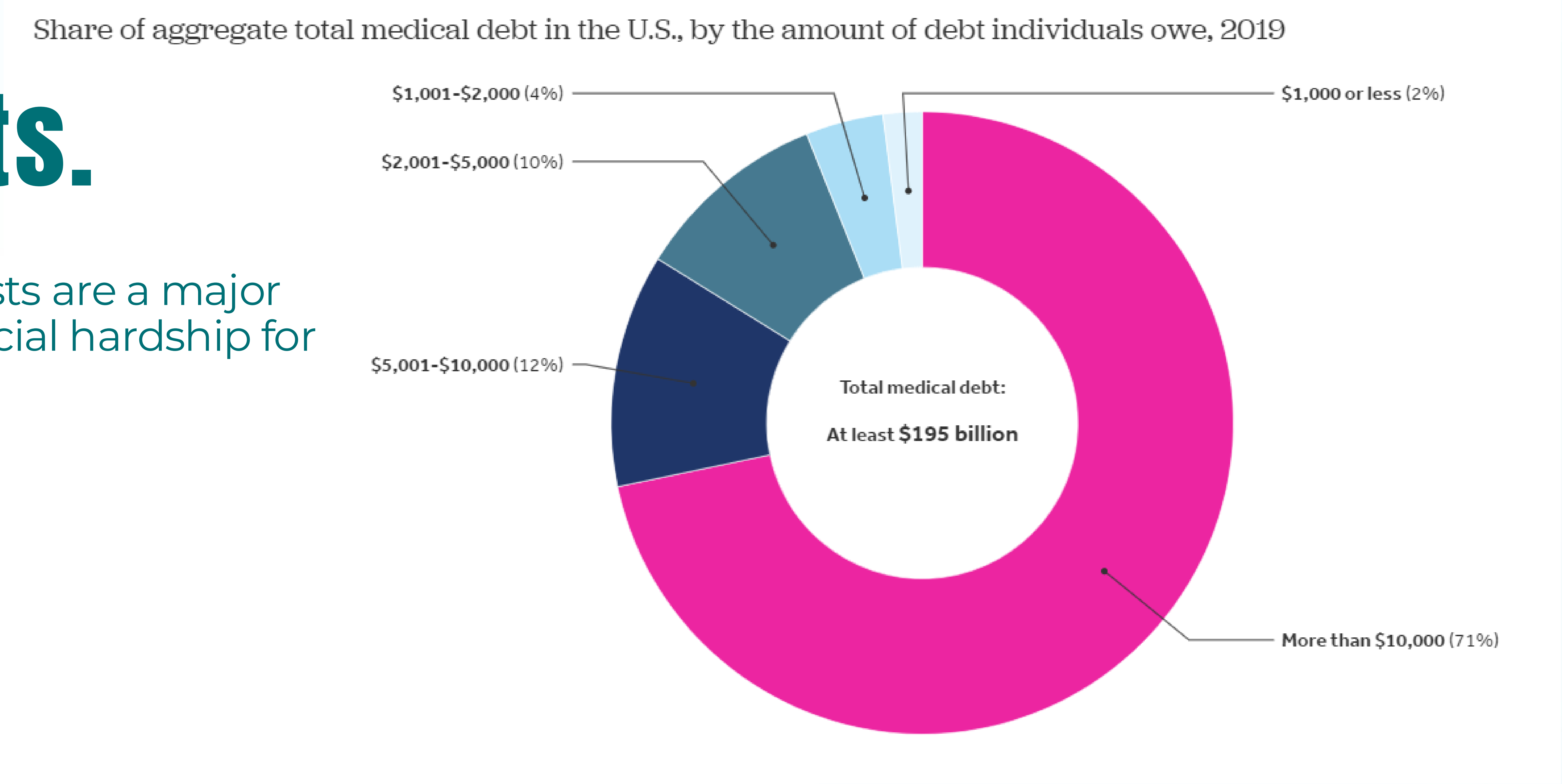
NATIONAL HEALTH EXPENDITURES (% OF GDP)

SOURCE: Centers for Medicare and Medicaid Services, *National Health Expenditures*, December 2021.

PGPF.ORG

01 Insights.

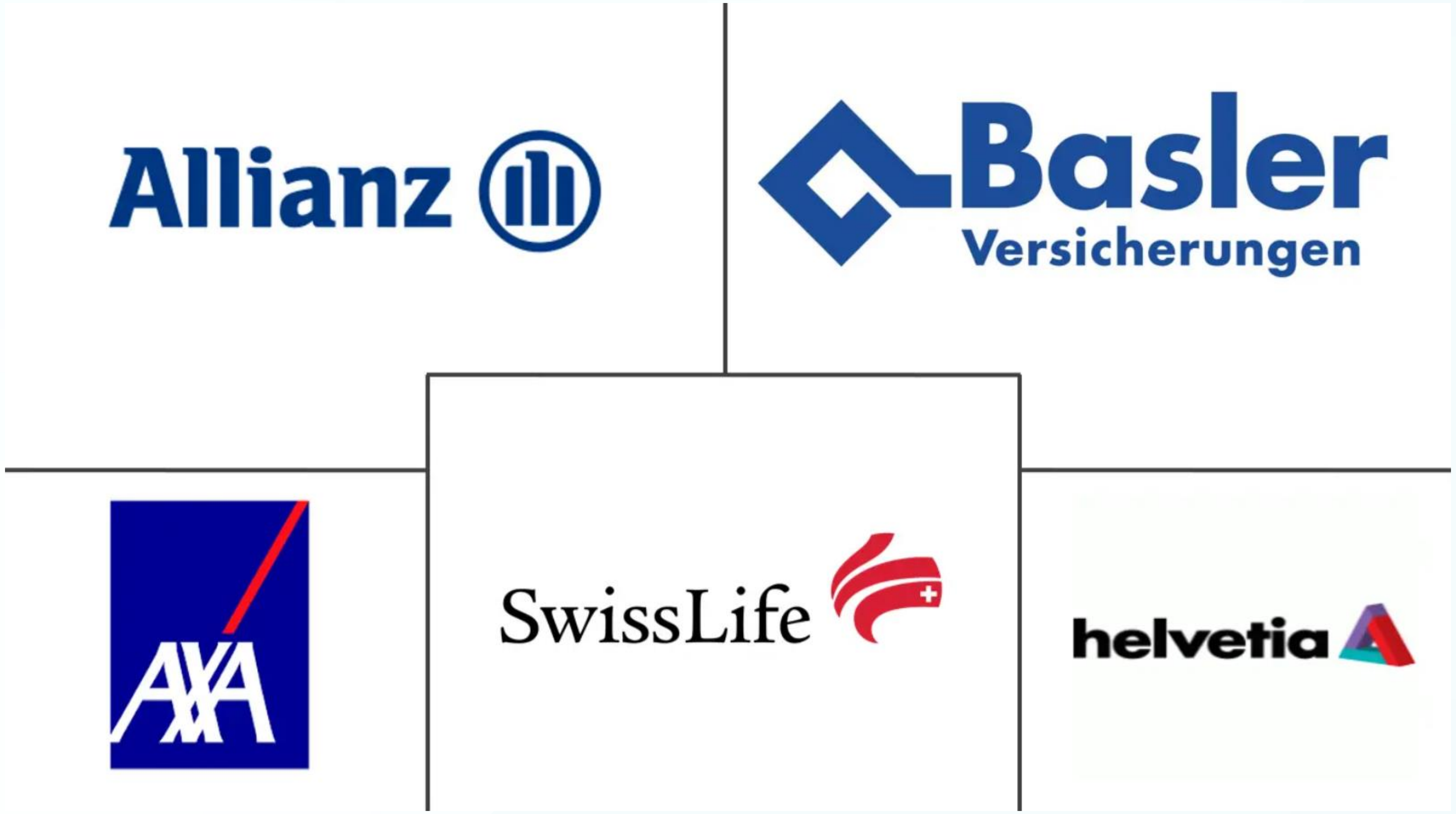
Health care costs are a major source of financial hardship for many people



Share of aggregate total medical debt in the U.S., by the amount of debt individuals owe, 2019

01 Insights.

Ability to Predict Health Costs is key for fair pricing within the Insurance Industry & better management of National Health Systems



Fundamental Problem

The rising cost of health care is one of the world’s most important problems

Traditional methods for predicting medical costs have not been appropriately validated using populations that the methods have not seen.

02

Problem Statement

Limitations of Traditional Methods & Benefits of Data Mining

Methodological Problem

Since the 1980s, there has been research on the predictive modeling of medical costs based on (health insurance) claims data using heuristic rules and regression methods. These methods, however, have not been appropriately validated using populations that the methods have not seen.

Objective of the Case

Predicting Health Costs using modern data-mining methods

General.

Traditional Methods

Regression Analysis
Hypothesis Testing
Time Series Analysis
ANOVA (Analysis of Variance)
Chi-Square Test
Correlation Analysis
Descriptive Statistics
t-Test
Factor Analysis
Decision Trees

✓ Data Mining Method

Clustering
Classification (e.g., Decision Trees, Naive Bayes, Support Vector Machines)
Association Rule Mining
Regression Trees
Neural Networks
K-Nearest Neighbors (K-NN)
Principal Component Analysis (PCA)
Self-organizing Maps (SOM)
Apriori Algorithm
Random Forest

Case Specific.

Traditional Method

Simple Regression Models

Heuristic-Based Approaches

Generalized Linear Models (GLMs)

High-Cost Case Identification

✓ Data Mining Methods

Classification Trees

Clustering

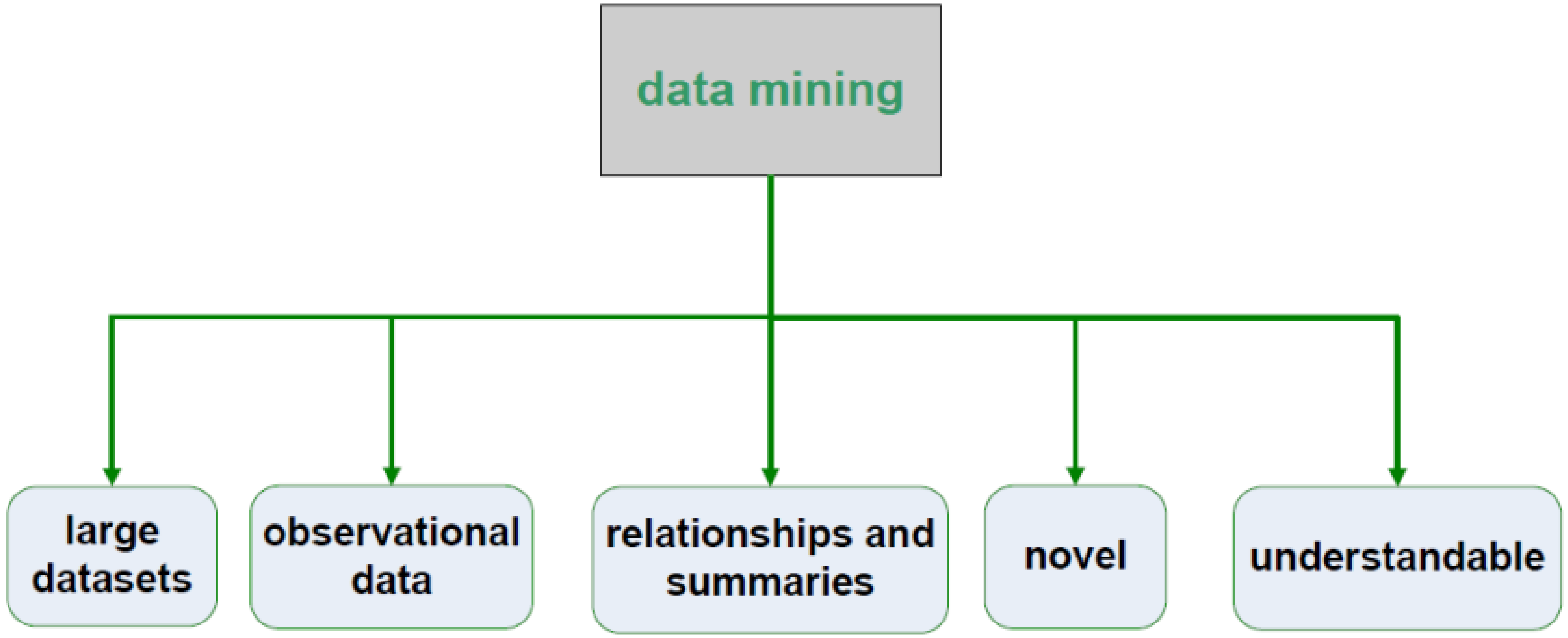
03

Data Mining Process

Review of Research Objective and Variables

Features of Data Mining

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. [Hand, Mannila, Smyth]



data mining

large datasets

observational data

relationships and summaries

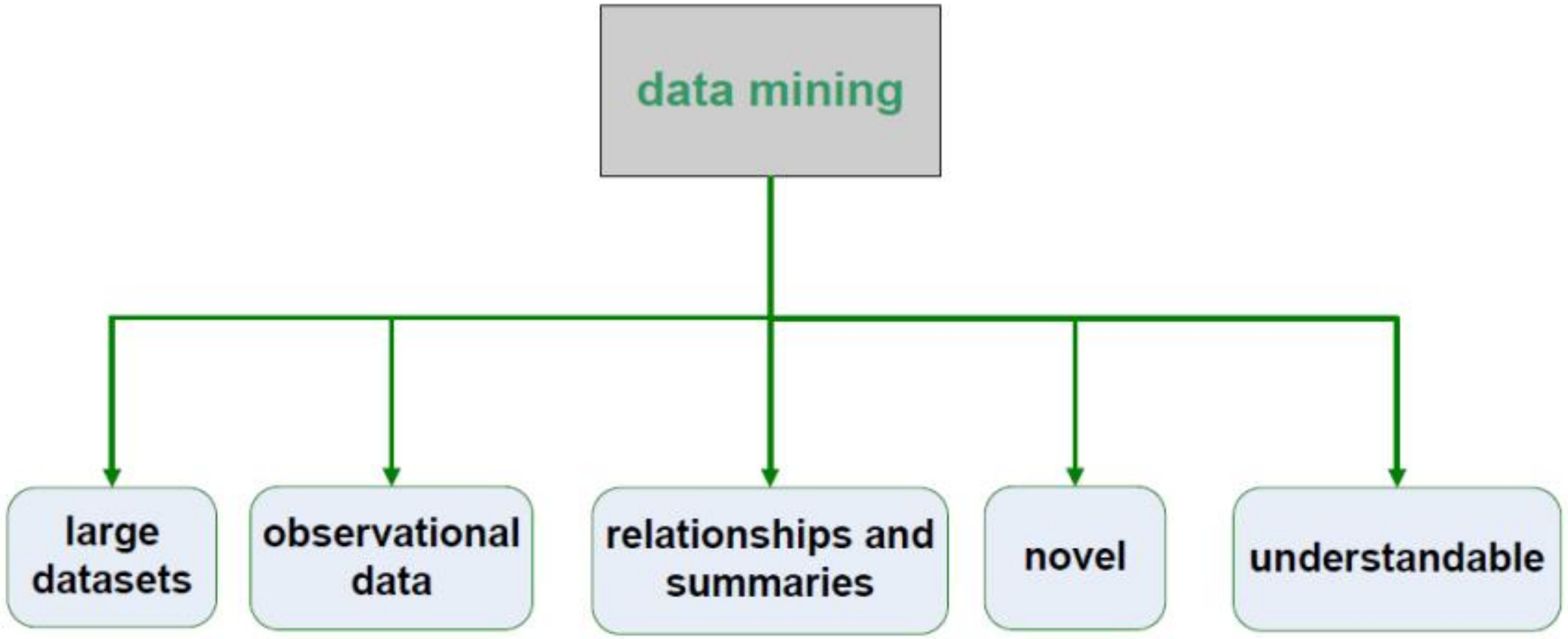
novel

understandable

Objective of the Case - details

Predicting Health Costs using modern **data-mining methods,** specifically classification trees and clustering algorithms, along with claims data from over 800,000 insured individuals over three years, to provide rigorously validated predictions of health-care costs in the third year, based on medical and cost data from the first two years.

Case Data Mining Process



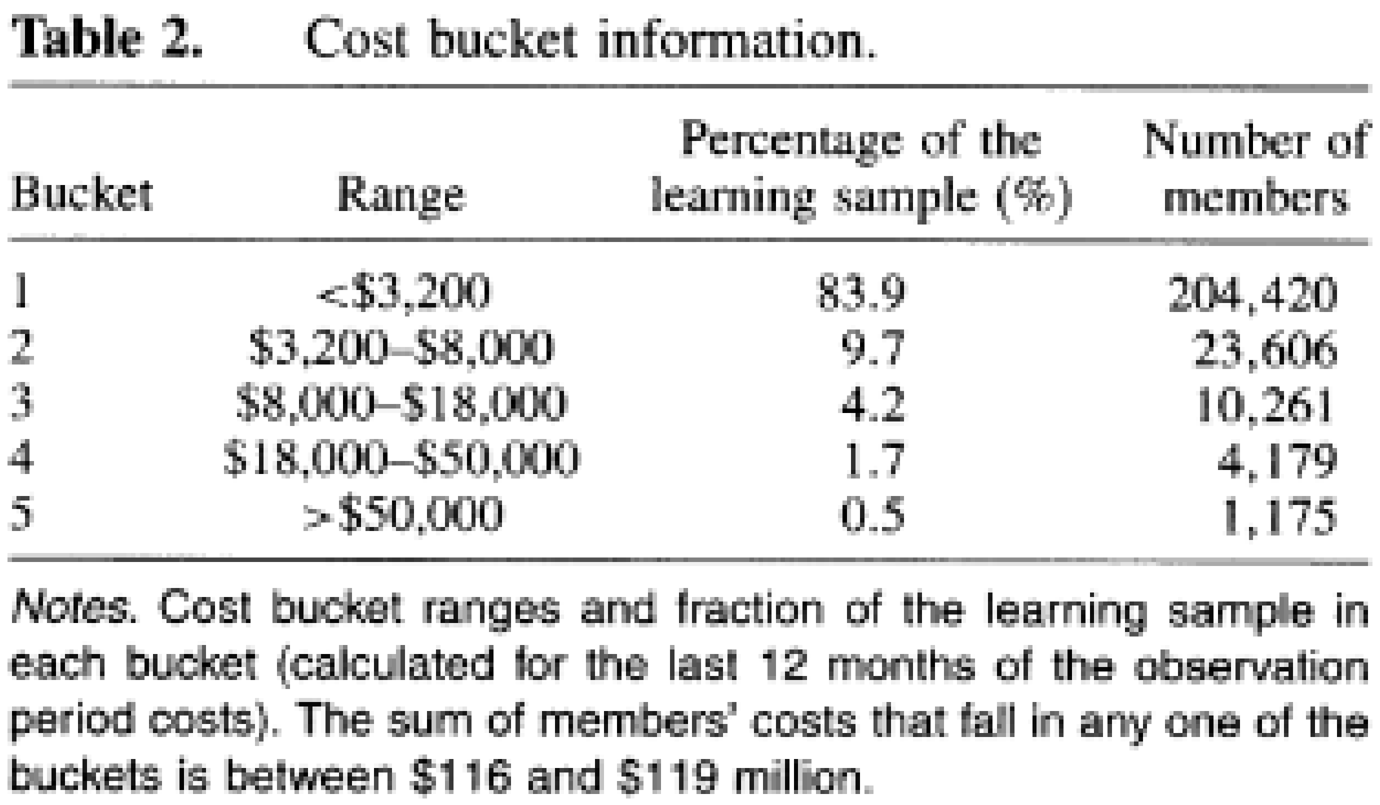


Table 2. Cost bucket information.

Bucket	Range	Percentage of the learning sample (%)	Number of members
1	<$3,200	83.9	204,420
2	$3,200–$8,000	9.7	23,606
3	$8,000–$18,000	4.2	10,261
4	$18,000–$50,000	1.7	4,179
5	>$50,000	0.5	1,175

Notes. Cost bucket ranges and fraction of the learning sample in each bucket (calculated for the last 12 months of the observation period costs). The sum of members' costs that fall in any one of the buckets is between $116 and $119 million.

Large datasets: Costs of 800,000 individuals over three years

Observational data: Insurance claims

Relationships and summaries: Medical variables as a predictor of costs

Novel: A model that can be accurately applied to other samples

Understandable: Intuitive as Medical conditions are a likely key determinant of healthcare costs

Useful: Model that allows Insurance Companies and Health Care providers be more efficient

Predicting Health Costs - case

“X” variables

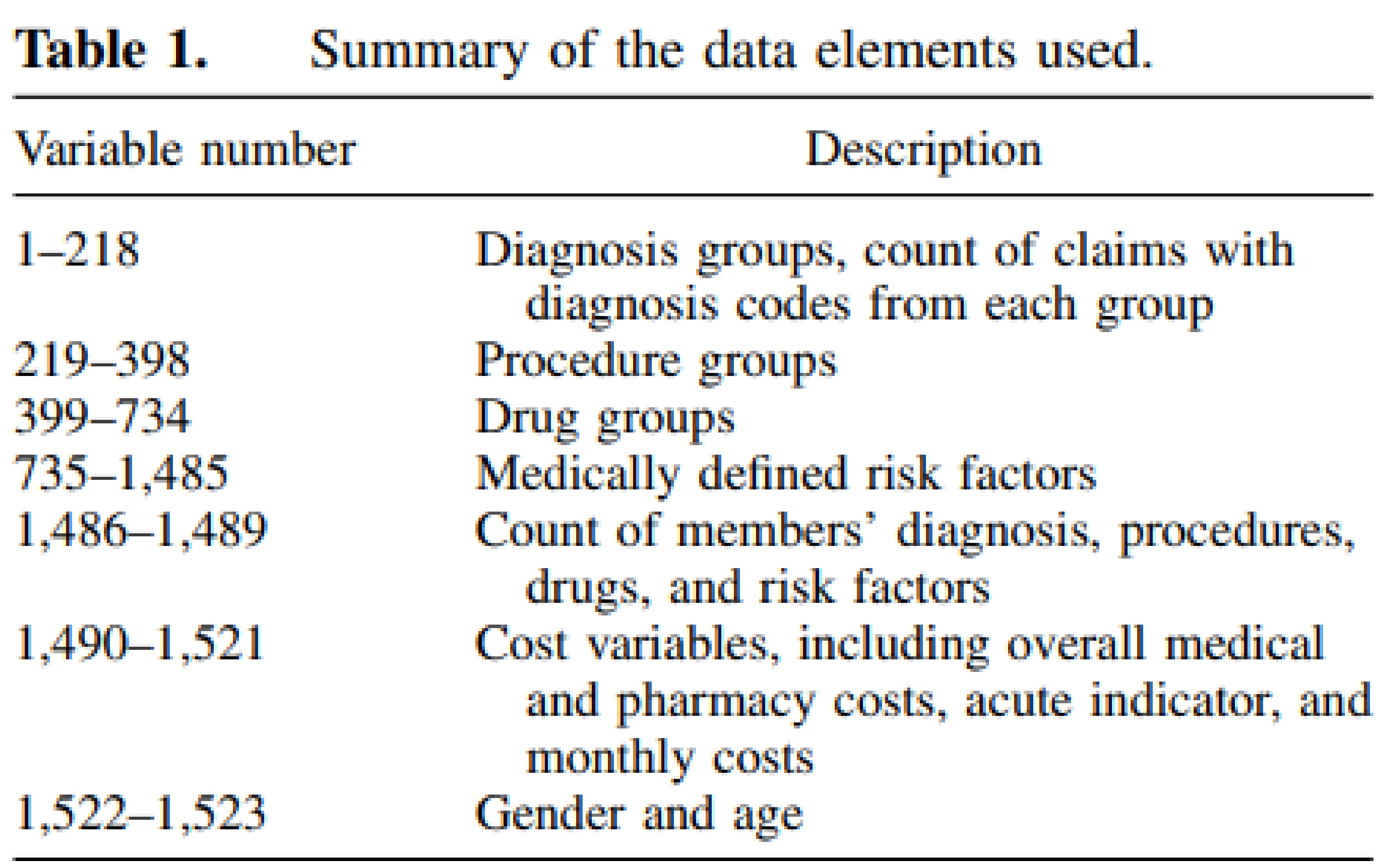


Table 1. Summary of the data elements used.

Variable number	Description
1–218	Diagnosis groups, count of claims with diagnosis codes from each group
219–398	Procedure groups
399–734	Drug groups
735–1,485	Medically defined risk factors
1,486–1,489	Count of members’ diagnosis, procedures, drugs, and risk factors
1,490–1,521	Cost variables, including overall medical and pharmacy costs, acute indicator, and monthly costs
1,522–1,523	Gender and age

Y

Predicted Patient’s Health Costs

*predictions of health-care costs in the third year, based on medical and cost data from the first two years.

04

Baseline Method

Using the Traditional methods

Baseline Method - Forecast

The baseline method in the case involves using the healthcare cost incurred during the last 12 months of the observation period as the prediction for an individual's overall healthcare cost in the result period.

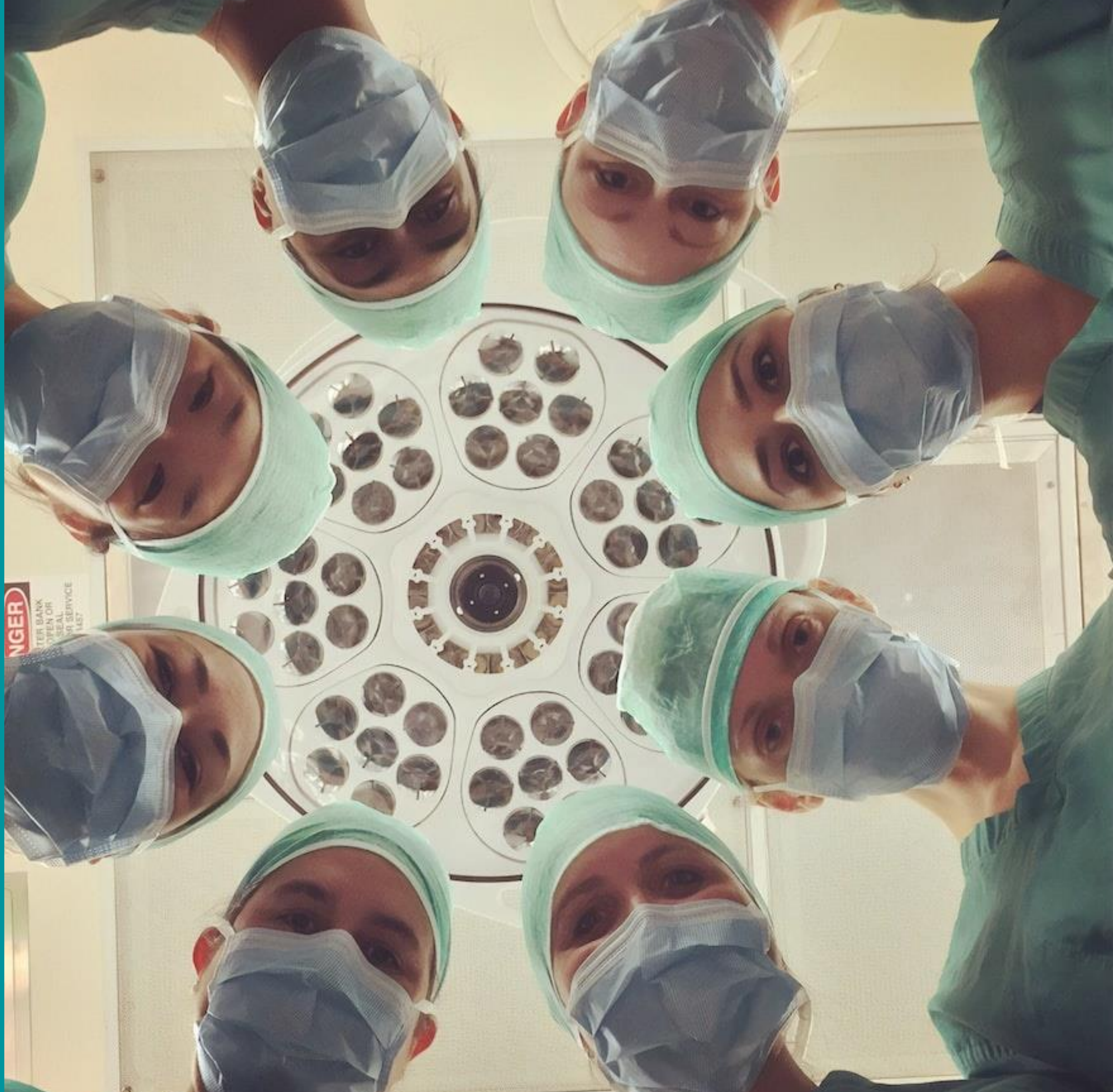
This approach relies on the strong correlation between an individual's current healthcare expenses and their overall health status.

The baseline method provides a reference point for evaluating the performance of more advanced prediction models in forecasting healthcare costs, and its accuracy varies depending on individuals' cost buckets from the observation period.

05

*Classification Trees

Data Mining Method 1

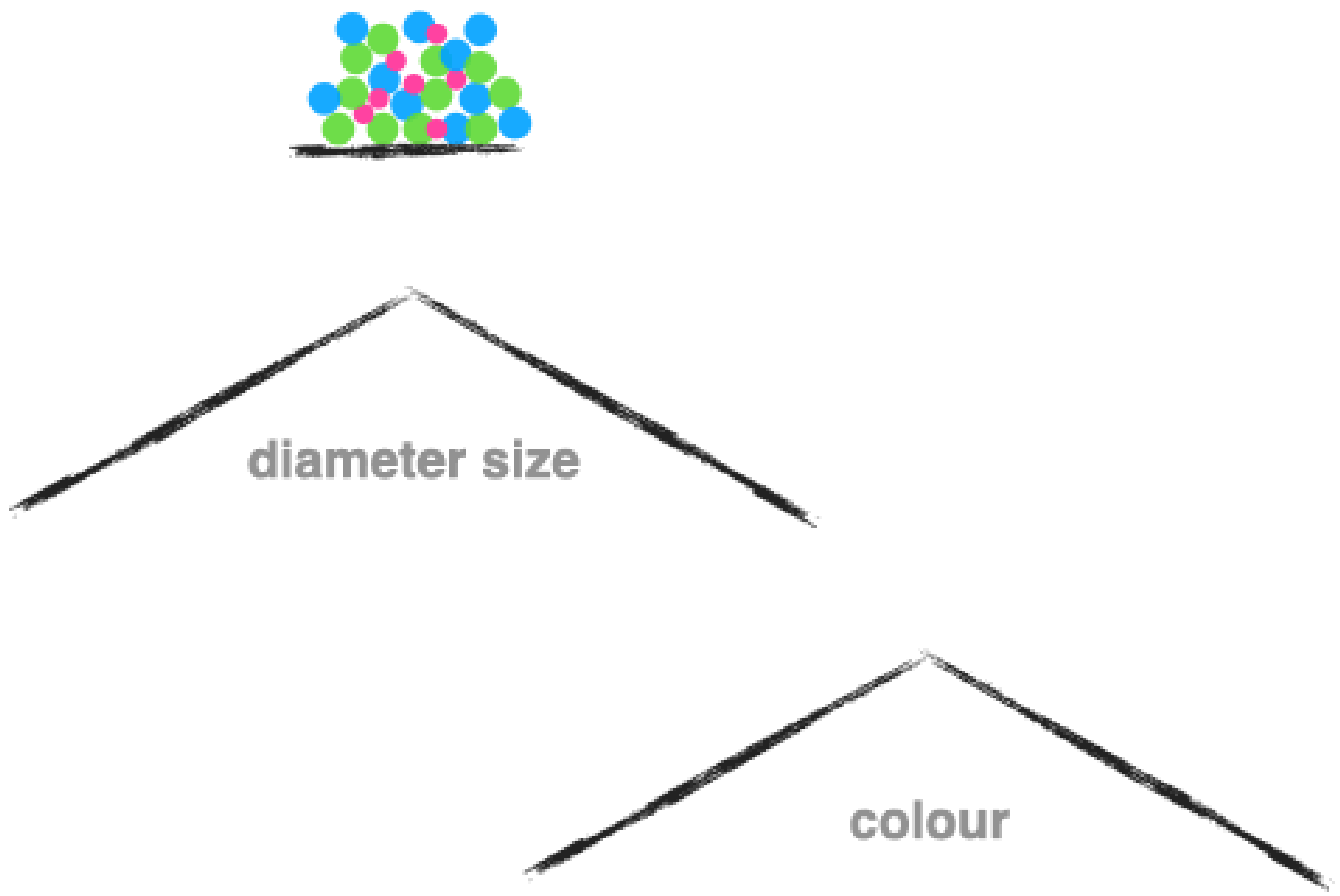


What are Classification Trees?

Classification trees are a data mining technique that recursively divides a dataset into smaller, more homogenous groups based on certain criteria.

They create a tree-like structure where each branch represents a decision based on attributes, leading to the classification of data into different categories or classes.

Classification trees are interpretable and widely used for tasks like predicting outcomes or classifying data points into predefined categories, making them valuable in fields such as medicine, finance, and more.



Classification Trees - in the case

Classification trees are a widely applied technique in various fields, including medicine, to develop criteria for classifying different conditions.

These trees recursively partition the member population into smaller, more homogenous groups based on known result period costs, creating a graphical and interpretable representation.

An example classification tree is provided, demonstrating how it can predict result period health-care costs based on diagnoses such as coronary artery disease (CAD) and diabetes.

The authors note that more complex classifiers can be derived from the full dataset, and these classifiers consider both cost and medical information, along with age, to identify high-risk members of the population.

Classification Tree

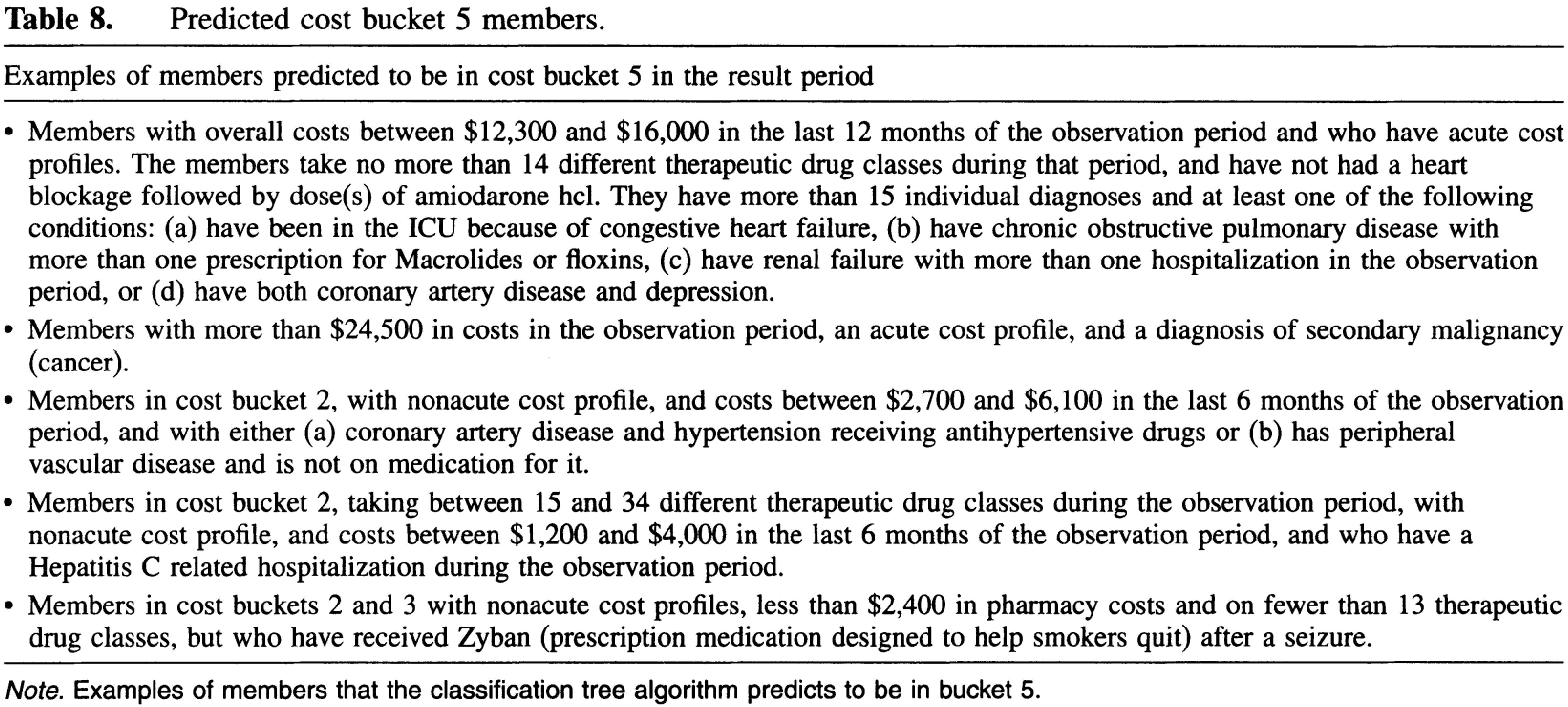


Table 8. Predicted cost bucket 5 members.

Examples of members predicted to be in cost bucket 5 in the result period

- Members with overall costs between $12,300 and $16,000 in the last 12 months of the observation period and who have acute cost profiles. The members take no more than 14 different therapeutic drug classes during that period, and have not had a heart blockage followed by dose(s) of amiodarone hcl. They have more than 15 individual diagnoses and at least one of the following conditions: (a) have been in the ICU because of congestive heart failure, (b) have chronic obstructive pulmonary disease with more than one prescription for Macrolides or floxins, (c) have renal failure with more than one hospitalization in the observation period, or (d) have both coronary artery disease and depression.
- Members with more than $24,500 in costs in the observation period, an acute cost profile, and a diagnosis of secondary malignancy (cancer).
- Members in cost bucket 2, with nonacute cost profile, and costs between $2,700 and $6,100 in the last 6 months of the observation period, and with either (a) coronary artery disease and hypertension receiving antihypertensive drugs or (b) has peripheral vascular disease and is not on medication for it.
- Members in cost bucket 2, taking between 15 and 34 different therapeutic drug classes during the observation period, with nonacute cost profile, and costs between $1,200 and $4,000 in the last 6 months of the observation period, and who have a Hepatitis C related hospitalization during the observation period.
- Members in cost buckets 2 and 3 with nonacute cost profiles, less than $2,400 in pharmacy costs and on fewer than 13 therapeutic drug classes, but who have received Zyban (prescription medication designed to help smokers quit) after a seizure.

Note. Examples of members that the classification tree algorithm predicts to be in bucket 5.

06

*Clustering

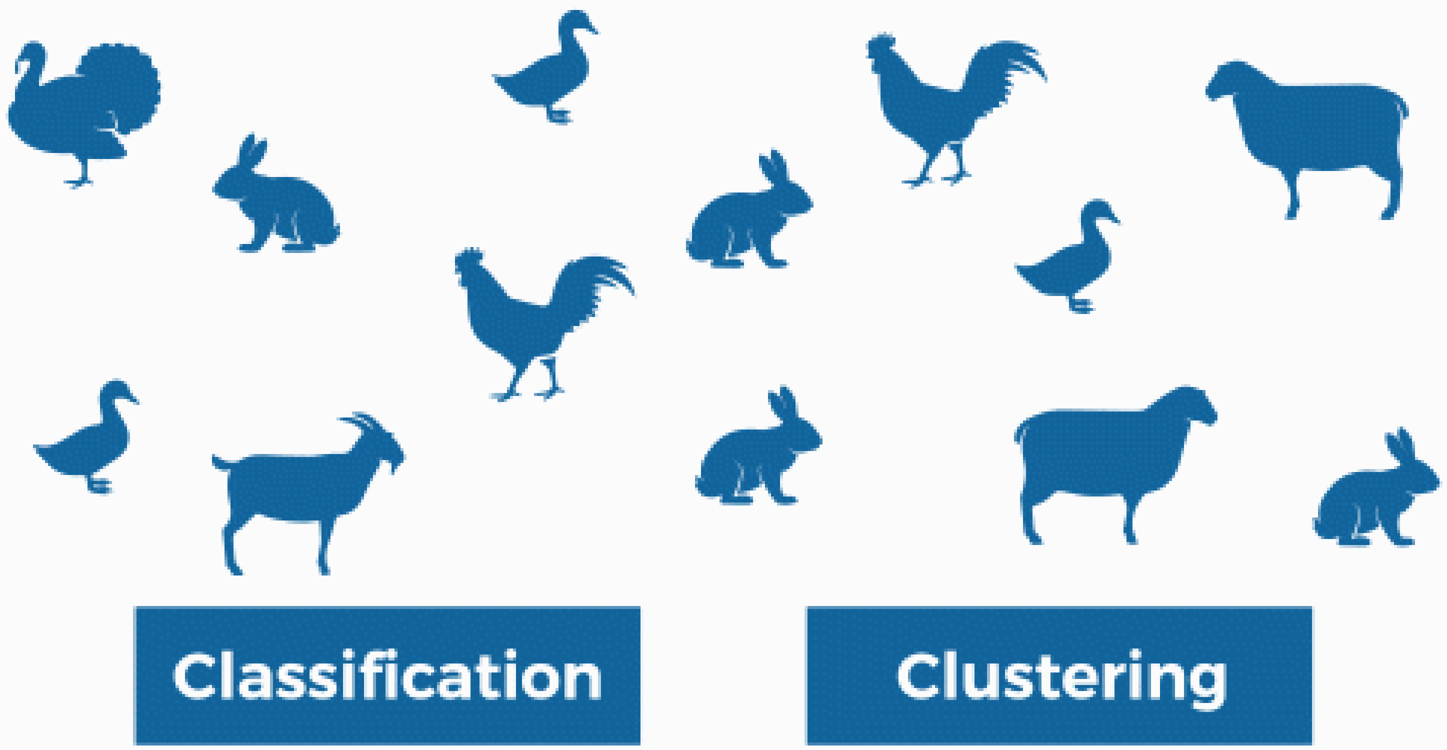
Data Mining Method 2

What is Clustering?

Clustering in data mining is a technique used to group similar data points together while separating dissimilar ones.

It involves finding patterns or structures in data, categorizing them into clusters, and organizing them based on their inherent similarities or dissimilarities.

Clustering helps identify natural groupings within datasets, enabling data analysts and machine learning algorithms to uncover insights, detect anomalies, and make data-driven decisions in various applications, including customer segmentation, image analysis, and pattern recognition.



Clustering - in the case

This case uses a prediction clustering method based on the Eigen Cluster algorithm.

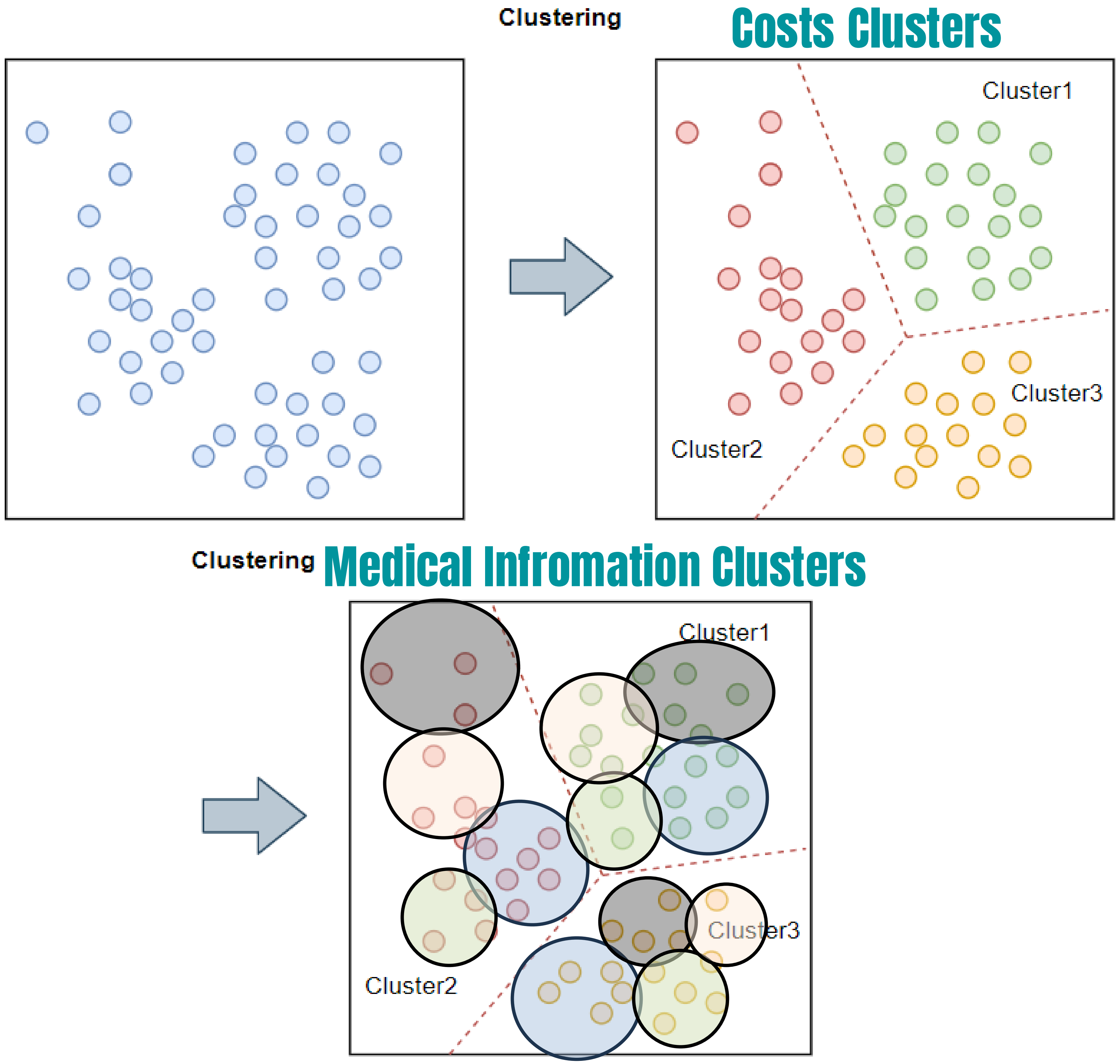
Initially, members are clustered based on monthly healthcare cost data, with an emphasis on recent months. The clusters represent individuals with similar cost profiles.

Then, the algorithm incorporates medical data to create clusters with both cost and medical condition similarities.

Forecasts are made based on known costs. For instance, one cluster contains individuals with cancer indicators, indicating higher future healthcare costs, while another comprises those in physical therapy or with musculoskeletal characteristics, suggesting lower future costs due to expected recovery.

Clustering improves health profile differentiation and healthcare cost prediction.

Two Step Clustering



Clustering

Costs Clusters

Clustering

Medical Infromation Clusters

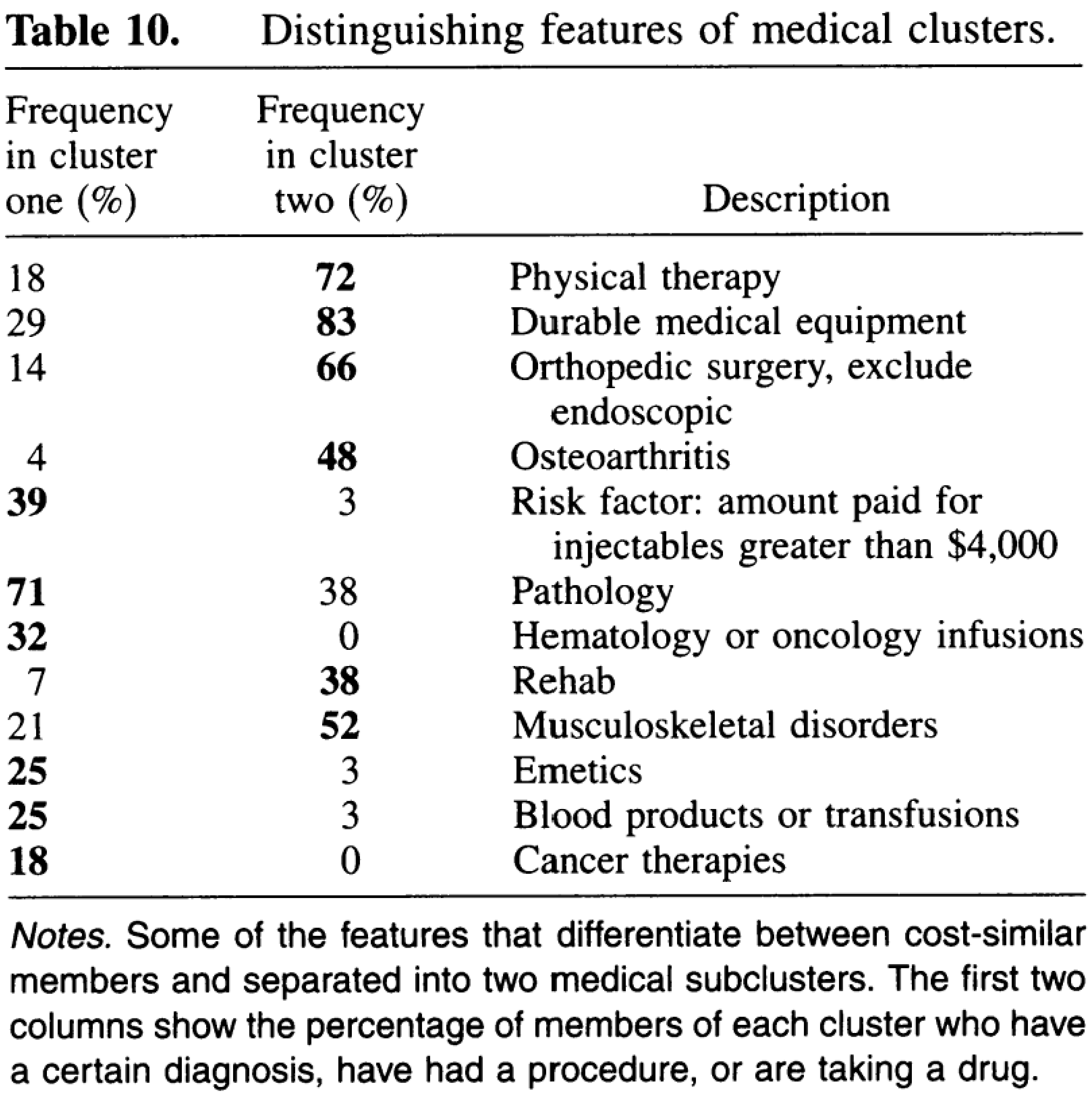


Table 10. Distinguishing features of medical clusters.

Frequency in cluster one (%)	Frequency in cluster two (%)	Description
18	**72**	Physical therapy
29	**83**	Durable medical equipment
14	**66**	Orthopedic surgery, exclude endoscopic
4	**48**	Osteoarthritis
39	3	Risk factor: amount paid for injectables greater than $4,000
71	38	Pathology
32	0	Hematology or oncology infusions
7	**38**	Rehab
21	**52**	Musculoskeletal disorders
25	3	Emetics
25	3	Blood products or transfusions
18	0	Cancer therapies

Notes. Some of the features that differentiate between cost-similar members and separated into two medical subclusters. The first two columns show the percentage of members of each cluster who have a certain diagnosis, have had a procedure, or are taking a drug.

07

Key findings

Results and comparison

Comparing Results

Baseline

Lowest Performance

The baseline method was used to establish a benchmark for comparison. It correctly predicted the right cost bucket for 80% of the population. The average penalty error for the baseline was 0.431, and the absolute prediction error was $2,677.

Classification Trees

Classification trees achieved promising results, correctly predicting the right cost bucket for over 84% of the population. They reduced the average penalty by 10.5% and the absolute prediction error by over 16%, indicating a significant improvement compared to the baseline method.

Clustering

The clustering algorithm also showed substantial improvements over the baseline method.

It performed comparably with classification trees and had an edge in terms of the absolute prediction error.

Comparing Results

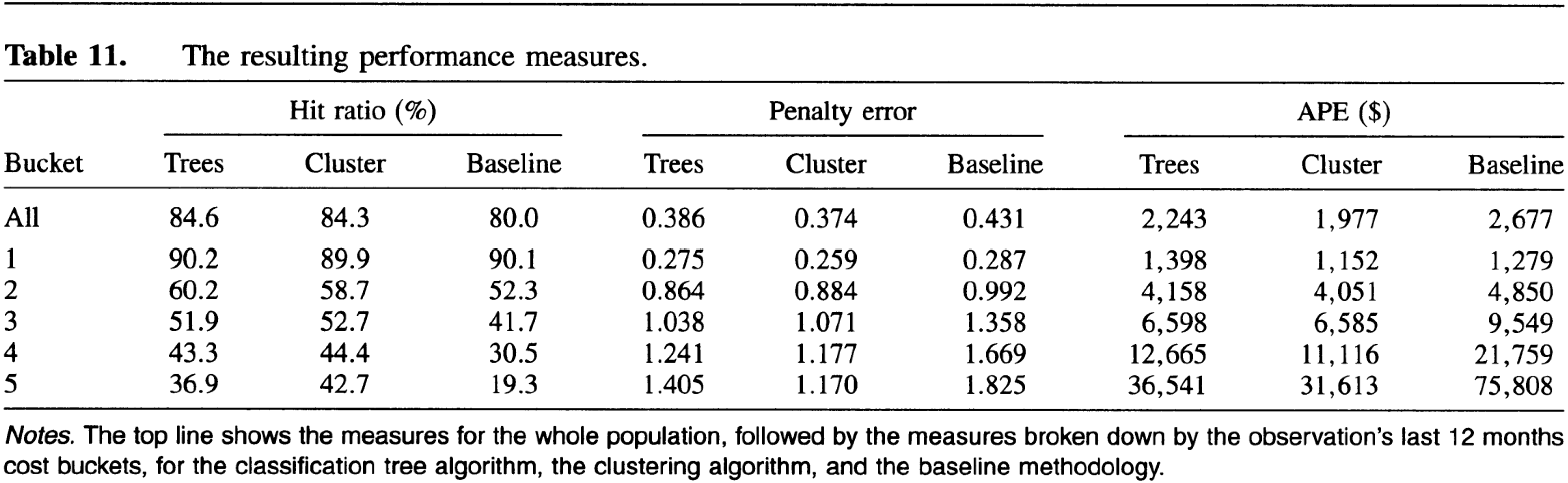


Table 11. The resulting performance measures.

	Hit ratio (%)			Penalty error			APE ($)		
Bucket	Trees	Cluster	Baseline	Trees	Cluster	Baseline	Trees	Cluster	Baseline
All	84.6	84.3	80.0	0.386	0.374	0.431	2,243	1,977	2,677
1	90.2	89.9	90.1	0.275	0.259	0.287	1,398	1,152	1,279
2	60.2	58.7	52.3	0.864	0.884	0.992	4,158	4,051	4,850
3	51.9	52.7	41.7	1.038	1.071	1.358	6,598	6,585	9,549
4	43.3	44.4	30.5	1.241	1.177	1.669	12,665	11,116	21,759
5	36.9	42.7	19.3	1.405	1.170	1.825	36,541	31,613	75,808

Notes. The top line shows the measures for the whole population, followed by the measures broken down by the observation's last 12 months cost buckets, for the classification tree algorithm, the clustering algorithm, and the baseline methodology.

Hit Ratio: % of correct cost bucket forecasted

Penalty error: Average cost bucket misclassification

APE: Absolute Prediction Error, Average difference between predicted cost and real cost

08

Conclusion

Implication of the study

Conclusion

Overall Improvement:

The key takeaway is that both data-mining methods outperformed the baseline methodology, with considerable enhancements in accuracy, particularly for the most expensive healthcare cost scenarios.

Classification Trees

Clustering

Better Models:

Hence, per the data-mining methods we attained a better way of pricing insurance and assisting National Health Systems manage costs than traditional regression model.

Appendix: R2 of Testing Sample

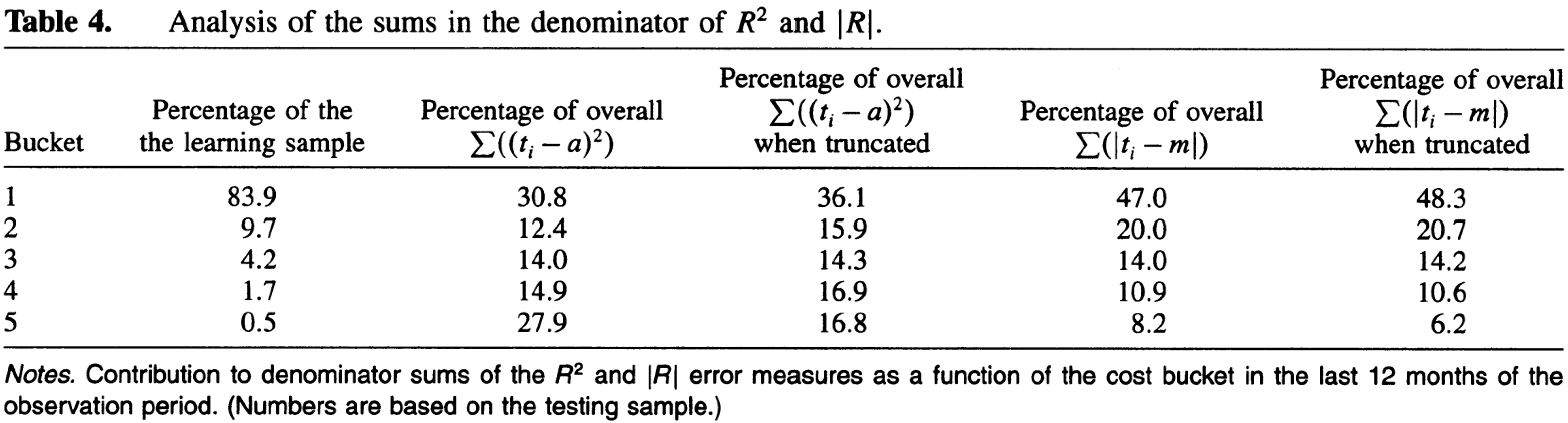


Table 4. Analysis of the sums in the denominator of R^2 and $|R|$.

Bucket	Percentage of the the learning sample	Percentage of overall $\sum((t_i - a)^2)$	Percentage of overall $\sum((t_i - a)^2)$ when truncated	Percentage of overall $\sum(\|t_i - m\|)$	Percentage of overall $\sum(\|t_i - m\|)$ when truncated
1	83.9	30.8	36.1	47.0	48.3
2	9.7	12.4	15.9	20.0	20.7
3	4.2	14.0	14.3	14.0	14.2
4	1.7	14.9	16.9	10.9	10.6
5	0.5	27.9	16.8	8.2	6.2

Notes. Contribution to denominator sums of the R^2 and $|R|$ error measures as a function of the cost bucket in the last 12 months of the observation period. (Numbers are based on the testing sample.)

$$R^2 = 1 - \frac{\sum_i (t_i - f_i)^2}{\sum_i (t_i - a)^2},$$

$$R| = 1 - \frac{\sum |t_i - f_i|}{\sum |t_i - m|},$$

Appendix: Cost only results

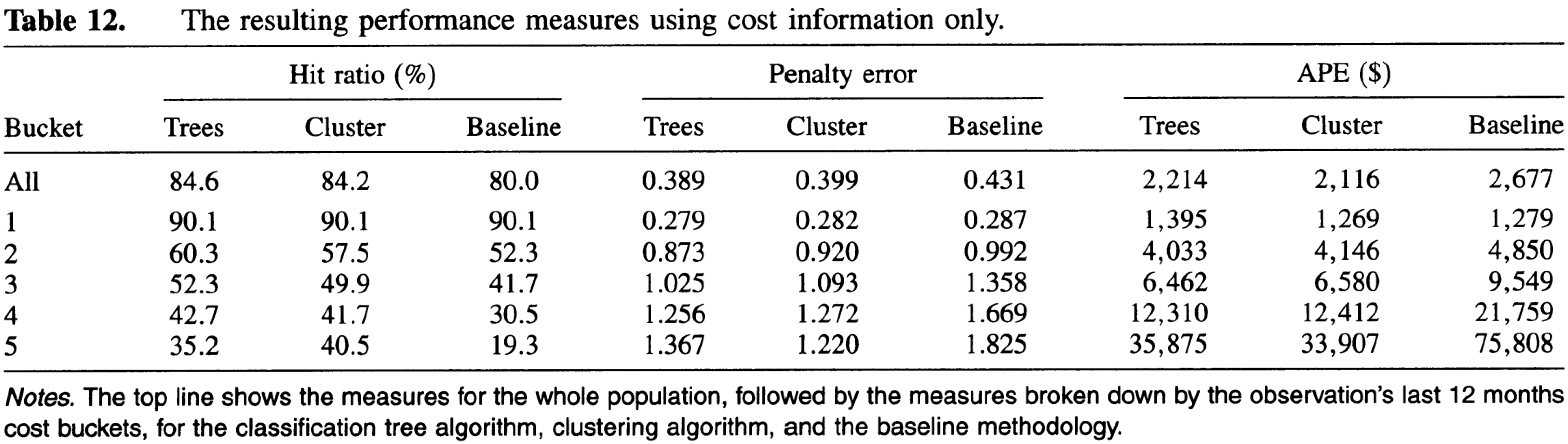


Table 12. The resulting performance measures using cost information only.

	Hit ratio (%)			Penalty error			APE ($)		
Bucket	Trees	Cluster	Baseline	Trees	Cluster	Baseline	Trees	Cluster	Baseline
All	84.6	84.2	80.0	0.389	0.399	0.431	2,214	2,116	2,677
1	90.1	90.1	90.1	0.279	0.282	0.287	1,395	1,269	1,279
2	60.3	57.5	52.3	0.873	0.920	0.992	4,033	4,146	4,850
3	52.3	49.9	41.7	1.025	1.093	1.358	6,462	6,580	9,549
4	42.7	41.7	30.5	1.256	1.272	1.669	12,310	12,412	21,759
5	35.2	40.5	19.3	1.367	1.220	1.825	35,875	33,907	75,808

Notes. The top line shows the measures for the whole population, followed by the measures broken down by the observation's last 12 months cost buckets, for the classification tree algorithm, clustering algorithm, and the baseline methodology.