Introducing an Efficient Sampling Method for National Surveys with Limited Sample Sizes: Application to a National Study to Determine Quality and Cost of Healthcare.

Mahboubeh Parsaeian (mahbobehparsaeian@yahoo.com)
https://orcid.org/0000-0003-1928-5424

Mahdi Mahdavi
Tehran University of Medical Sciences, National Institute for Health Research https://orcid.org/0000-0002-7383-3455

Mojdeh Saadati
Iowa State University

Parinaz Mehdipour
Tehran University of Medical Sciences

Ali Sheidaei
Tehran University of Medical Sciences

Shahab Khatibzadeh
Brandeis University

Farshad Farzadfar
Tehran University of Medical Sciences

Saeid Shahraz
Tufts Medical Center

Research article

Keywords: Survey sampling method, small sample size, model-based clustering, validity, sampling efficiency, Iran Quality of Care in Medicine Program (IQCAMP)

Posted Date: April 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-20263/v1

Version of Record: A version of this preprint was published at BMC Public Health on July 17th, 2021. See the published version at https://doi.org/10.1186/s12889-021-11441-0.

Abstract

Background: Due to budget constraints, many national studies are only affordable with small sample sizes. Facing a similar constraint, the national study of “Iran Quality of Care in Medicine Program” (IQCAMP), aims to evaluate the quality and costs of medical care in the country for 8 selected high-cost high-volume diseases by only recruiting 300 participants per disease under study. We proposed and tested a data mining method to efficiently design such national studies with small sample sizes.

Methods: We developed a sampling design based on hierarchical clustering method (HCM) and model-based clustering method (MCM). Based on these methods, we used healthcare structure and outcome data to define the optimum clusters of districts and provinces of Iran. An expert group checked the face validity of the defined clusters. We subsequently compared the internal validity of HCM with MCM by within clusters sum of square, average silhouette width, and Dunn index. The stability validity was also examined by statistical indices. We selected the optimum clusters based on the results of these measures. Per cluster, one province was systematically selected for data collection. The features of selected clusters were delineated by the Decision Tree Learning and the most important distinguishing indicators were identified. The efficiency of the selected sampling method was compared with a simple random sampling through simulation.

Results: MCM and HCM divided the districts into eight and two clusters, respectively. The measures of internal and stability validity showed that clusters created by MCM were more separated, compact, and stable, thus forming our optimum clusters. The probability of death from stroke, chronic obstructive pulmonary disease, and in-hospital mortality rate were the most important indicators that distinguished the eight clusters. Based on the simulation results, MCM increased the efficiency of the sampling design up to 70% compared to simple random sampling.

Conclusions: By using this sampling method we selected 300 participants per health condition across the country only from eight provinces rather than all 31 provinces. Our sampling method benefits studies with small sample sizes by decreasing the level of sampling variability and the overall cost of studies.

Background

There exist a wide variety of sampling methods for designing surveys [1–4]. However, some studies are subject to constraints that make classic sampling methods inappropriate. For instance, multistage sampling is a typical method for the large-scale sampling of a population in national studies [5]. In this method, a region is divided into clusters, and several clusters are randomly selected at different stages. In studies with a large sample size, this sampling method includes samples from the entire spectrum of an outcome. However, the performance of multistage sampling raises major concerns in studies with a small sample size. Firstly, the results of sampling may significantly vary when sampling is iterated. Secondly, selected clusters may not embrace the entire spectrum of the outcome of interest.

Due to budget constraints, many national studies are only affordable with a small sample size. Thus, the efficiency of sampling methods becomes essential. Facing a similar constraint, the national study of “Iran Quality of Care in Medicine Program” (IQCAMP), aims to evaluate the quality and costs of medical care in Iran for selected high-cost high-volume diseases. These diseases are acute myocardial infarction, congestive heart failure, stroke, diabetes mellitus, chronic obstructive pulmonary disease, major depressive disorder, and end-stage renal disease. The study considers measuring several patient-level variables of quality and costs of healthcare at multiple time intervals for three months. IQCAMP is a pilot national-level longitudinal survey with the limitation of sample size (N = 300) patients per condition for the entire nation. We, therefore, propose an innovative sampling design that maximizes the national representativeness.

Method

We adopted a data mining approach and developed a sampling design based on two clustering methods; hierarchical clustering method (HCM) and model-based clustering method (MCM). The details of these two clustering strategies can be found elsewhere [6–8]. The sampling design used the clustering algorithms that divided the country into a minimum number of clusters for the maximum representativeness of health care quality and costs (HCQC) indicators. A cluster represents a number of homogenous districts or provinces.

The input data of the method consisted of prior information of HCQC indicators that were available from the national surveys and registries [9–12]. The input data was used by clustering methods to divide 413 districts of the country into homogeneous clusters. A key criterion in cluster selection was to minimize the within-cluster differences and at the same time to maximize the between-cluster differences in the indicators of HCQC. We used district level data to cluster the provinces. In the first stage of validity assessment, an expert group checked the face validity of the selected clusters. We subsequently compared the internal validity and stability of selected clusters by statistical indices. We used a Decision Tree Learning to explore the distinct features of the clusters. Subsequently, we conducted a simulation to compare the efficiency of the clustering method with a simple random sampling (SRS).

Clustering Methods

Hierarchical clustering method decomposes data hierarchically. The decomposition is undertaken by an agglomerative approach [13]. There are different methods for agglomeration of similar objects, i.e. input data. We chose the Complete Method that computes the distance between all objects and merges objects with the least distance. MCM is a finite mixture model, which assumes that data originate from heterogeneous clusters. Using the Bayesian Information Criteria (BIC), we compared models obtained from MCM that differ in the number of clusters and/or cluster distribution [8]. We considered a model with a larger BIC value as an optimum model. We provided the mathematical formulation of this model in the Additional file 1 Part A. We used R programming language version 3.5.1 and its “mclust” package version 5.2 (Additional file 2) [8, 14].

Input Data

The input data consisted of micro-level data of HCQC indicators. We relied on the Donabedian's Structure-Process-Outcome (SPO) model for selecting the indicators of HCQC [15]. Furthermore, we considered indicators of healthcare delivery context in order to comply with the contemporary healthcare evaluation methods [16]. The selected indicators consisted of patient demands (as a part of context), service structures, and health outcomes. Patient demands described the characteristics of the population seeking health services, including the type of health needs. Service structures described insurance arrangements and health care resources that were used to provide services. Health outcomes described the clinical health states of populations (Table 1).

We used the data from the national surveys and registries including 'Iran 2016 STEPwise approach to Surveillance of Non-Communicable Diseases (STEPS) study' [9], the 'Death Registration System' (DRS) in 2015 [12], 'Iran 2011 Hospital Data' [11], and 'Iran 2014 Healthcare Utilization study' [10]. We developed a combined metric to discriminate between different patterns of demand, structure, and outcome of the diseases under study. This metric used the aggregate information at district or province level, whichever feasible, to create homogeneous clusters. The metric in principle used district level data. However, in some of the data sources, district level data were unavailable, thus we relied on province level data (See Table 1).

Assessment of Clustering Tendency and Optimality

To achieve meaningful and interpretable clusters, we checked the clustering tendency of input indicators using Hopkins' statistics [14, 17]. The values of Hopkins' statistic higher than 0.5 were considered clusterable data. The optimum number of clusters were automatically determined through MCM. However, HCM could not directly estimate this number. We used R package NbClust to estimate an optimum number of clusters and used this number for HCM [18]. The package used 30 indices to estimate the optimum number of clusters, i.e. the number recommended by the majority of indices.

Validation

Internal Validation

Internal validation examines compactness and separation of clusters. Compactness measures within-cluster variations. Separation uses the information of between-cluster variations. Three measures were applied for internal validation: Silhouette width, Dunn index, and within-cluster sum of square [13, 14, 19]. Silhouette width compares "the average distance between a district and other districts within a same cluster" with "average distance between a district and other districts in other clusters. The values of Silhouette width ranges between − 1 (observation placed in the wrong cluster) and + 1 (observations are well matched to its own cluster). The greater the values of this index, the higher the compactness and separation of the clusters.

Dunn index is calculated as the minimum distance of objects between clusters to the maximum distance of objects in the same cluster. It ranges from zero to infinite. The larger the value of this index, the better the performance of the clustering method. Furthermore, the within-cluster sum of square indicates how

closely objects were related in the same clusters. Smaller values of this measure indicated a higher homogeneity of clusters.

Cluster Stability Validation

We used four indices to measure cluster stability. The measures were the average proportion of non-overlap (APN), the average distance (AD), the average distance between means (ADM), and the figure of merit (FOM) [19]. These measures compare clustering results from the original data with those of data from which one column is removed in a step-wise fashion. For all stability indices, smaller values indicate a better stability.

Identification of Clusters' Features

We used a Decision Tree Learning (DTL) to describe the clusters' features in terms of HCQC indicators [20, 21]. DTL uses partitioning rules to classify districts into several homogeneous sub-groups based on most important differentiating HCQC indicators. Partitioning rule was defined as conditions to assign districts into clusters based on the value of HCQC indicators. The algorithm continues the recursive partitioning of data to accurately predict cluster labels.

Comparison and Simulation of Sampling Methods

We used simulation technique to compare the efficiency of sampling between the clustering method and the simple random sampling. Based on the clustering methods, we selected one district per cluster and based on a simple random sampling, we selected the same number of district randomly out of all 413 districts of the country. Next, we estimated the weighted mean of the HCQC indicators for samples selected using two methods. We simulated these estimates 1000 times and calculated the mean and variance of these estimates. Sampling efficiency was defined by the ratio of the variance of simulated estimates in SRS ($\bar{x}_{SRS}$) to the variance of simulated estimates in the clustering method ($\bar{x}_{cluster}$).

Results

The Hopkins' statistic of input measures was estimated as 0.67, indicating a good clustering tendency. Figure 1 demonstrates the number of clusters recommended by different statistical indices. The X-axis shows the number of recommended clusters (k) and the Y-axis shows the number of indices proposed k. As shown in Fig. 1, most indices recommended two clusters, which was considered as the optimum number of clusters in HCM. Whereas, MCM recommended eight clusters based on the BIC criteria (Additional file 1 Part B).

The Validity of Clustering Methods

We compared the internal validity of results from MCM and HCM. We added another scenario to test the performance of HCM with eight clusters (HCM-8) (Table 2). The within-cluster sum of square in MCM with eight clusters (MCM-8) was lower than HCM with two clusters (HCM-2). The Dunn index of MCM-8 was higher than that of HCM-2. These results indicate that MCM-8 clusters are more compact and separated

than HCM-2. However, the average silhouette width of HCM-2 is larger than MCM-8. Comparing the clustering methods with the same number of cluster, the average silhouette width of MCM-8 is larger than HCM-8. Thus, the model-based method outweighs the hierarchical method with a same number of cluster.

The results of four stability measures are given in Table 2. AD, ADM and FOD selected MCM-8 as a more stable model; whereas, APN identified HCM-2 as a more stable model. Based on internal and stability validity, we selected MCM with eight clusters as a final classification of districts in this study. The geographic distribution of clusters and the districts of each cluster in MCM-8 is depicted in Fig. 2.

The number of districts in clusters varies from 31 to 86. Cluster 1 has the least number of districts and cluster 2 has the largest. Since the input data is at the district level, MCM assigns districts into clusters. To generalize the clustering result to province level, we assigned a province to a cluster that the majority of districts and the largest weighted population of that province fall into that cluster. To select one province per cluster, we calculated the distance of each province from other provinces in the same cluster and selected the province with minimum distance from other provinces (Additional file 1 Part C).

Features of Identified Clusters

The features of MCM-8 clusters are shown in Fig. 3. The most significant HCQC indicators that make distinctions between clusters were the probability of death from stroke, the probability of death from Chronic Obstructive Pulmonary Disease (COPD), in-hospital mortality rate, patient's exchange rate, the mortality rate caused by the adverse events of medical treatment, the probability of death from Chronic Kidney Disease (CKD), and all-cause mortality ratio (Fig. 3).

The decision tree identified 10 partitioning rules. Except for clusters six and eight, other clusters had unique features and were identified by only one partitioning rule. For instance, the distinct features of cluster 1 were as follows: all 31 districts had the probability of death from stroke < 0.008, the probability of death from COPD < 0.006, the mortality by adverse events of medical treatment < 33, the probability of death from CKD ≥ 0.021, and the all-cause mortality ratio < 1. These values were considered as cut-off points for partitioning. DTL accurately placed all 31 districts in this cluster.

Per cluster eight and six, DTL identified two rules. In cluster eight, out of 61 districts, 52 were identified by one rule and nine districts by an another. These rules were similar in the probability of death from stroke and the probability of death from COPD while they were different in the mortality rate caused by the adverse events of medical treatment, the probability of death from CKD, the all-cause mortality ratio, and the patient exchange rate. Similarly, among 42 districts in cluster six, 28 districts had one partitioning rule and 14 districts were identified by the other partitioning rule (see cluster's features in Fig. 3).

The Efficiency of Clustering Methods

Table 3 illustrates the sampling efficiency of key features of MCM-8 clusters detected by the DTL. The simulation results showed that the clustering method decreased the sampling variance of all these features compared to SRS. The highest reduction in a sampling variance, by 1.7 times, was related to the probability of death from stroke. The next higher reduction, 1.5 times, was for the probability of death from

COPD and the patient exchange rate. The lowest reduction was related to the mortality rate attributed to the adverse events of medical treatments with sampling efficiency 1.2.

Discussion

We used a data mining method to satisfy the sampling design requirements of the IQCAMP, a national pilot survey with a limited budget and sample size. The model-based clustering method divided districts into eight clusters; whereas, the hierarchical clustering method divided districts into two clusters. Before conducting the validity assessment through statistical analysis an expert group approved the face validity of the methods. The internal validity as measured by the within-cluster sum of square and Dunn index showed that the clusters of districts in MCM-8 had higher compactness and separation in comparison with HCM-2. Moreover, the majority of stability indices recognized that MCM with eight clusters is more stable than HCM with two clusters. Therefore, we selected MCM with eight clusters as the final model for sampling design. These clusters were mainly characterized by the probability of death from stroke, COPD, and CKD, in-hospital mortality rate, patient's exchange rate, the mortality rate attributed to adverse events of medical treatment, and all-cause mortality ratio. The simulation results showed the MCM-8 improved sampling efficiency up to 1.7 times compared with the simple random sampling. However, sampling efficiency may vary depending on the most important distinguishing indicators of clusters' features.

In the use of clustering methods, we built on earlier studies [22, 23]. Though there exists a large number of clustering methods in the literature, we used the MCM as it has several advantages over other clustering methods. It relies on statistical models and requires no pre-specified number of clusters [24]. HCM has also been extensively used in the literature [25, 26].

Relying on the contemporary theories of quality of healthcare [15, 16], an array of input indicators from demands, structures, and health outcomes were used by clustering methods to define optimum homogenous strata, which cover the spectrum of quality and cost in national studies as aimed by IQCAMP. The more homogeneous the strata, the more efficient the stratified sampling design [23]. This innovative way to define strata is an efficient alternative to conventional stratified sampling which defines strata based on, for instance, geographical units. This property is particularly desired in surveys with small sample size, which is potent to a larger variability of sampling results. Besides, by using this sampling method, we select participants only from eight provinces (rather than 31 provinces), which makes the sampling more feasible and affordable. Therefore, our sampling method benefits studies with small sample sizes by decreasing the level of sampling variability and the overall cost and feasibility of the study.

However, the efficiency of our final sampling method is measured by the quality and costs indicators of targeted health conditions. These indicators only relatively specify the aspects of quality and cost. Therefore, steps should be taken to include as much as inclusive, relevant, and precise prior information of quality and costs of health conditions for sampling.

Disease-specific surveys such as IQCAMP require large registries and health information systems that are barely available in developing countries. Usually, the information on the resource use (cost and utilization) and quality of services of different health conditions are limited to small samples collected by non-representative sampling methods such as convenient sampling [27–29]. Thus, the proposed clustering method is very appealing for developing countries where healthcare data are limited. This strategy helps policymakers to conduct small sample size surveys even with a limited budget.

The present study is subject to limitations. The first limitation regards the availability of district level data for some of the input measures, that were only available at provincial level. Thus, we used provincial level data for all districts of a province. The second limitation refers to representativeness of sampling results. The proposed method lies in the middle of a spectrum of sampling methods with convenient sampling methods at one extreme and simple random sampling at the other. Though the sampling method is far away from convenient sampling, an extent to which it comes closer to a representative sampling is unclear and needs to be evaluated in future studies. Worth to note that, the validity of the method depends on the appropriateness of the selected prior information on the quality and costs of the health condition. To maintain the validity, we selected HCQC indicators based on the accepted Donabedian SPO model [16] and the Design Science paradigm [30].

For the simplicity of sampling design, we used a common definition of strata for all eight health conditions in this research. This was motivated by the fact that access to prior information for each condition was limited. Furthermore, this common definition facilitated the administrative arrangement for data collection. However, with sufficient information per health condition, the definition of strata based on condition-specific outcomes could increase the sampling efficiency. We therefore call future research to address efficiency gain, cost, and feasibility of using condition-specific health outcomes to define strata for health conditions that are studied in the present research.

Conclusions

The present study provides a framework to efficiently design rigorous and affordable national studies with small sample size in countries with limited resources. The sampling method can approximately determine the optimum strata of survey variable, which leads to substantial gains in the precision of the estimates in the stratified sampling. Availability of inclusive, relevant, and precise prior information is critical for the application of this sampling method in other similar settings.

Abbreviations

APN: Average Proportion of Non-overlap; AD: Average Distance; ADM: Average Distance between Means; BIC: Bayesian Information Criteria; CKD: Chronic Kidney Disease; COPD: Chronic Obstructive Pulmonary Disease; DM: Diabetes Mellitus; DRS: Death Registration System; DTL: Decision Tree Learning; FOM: Figure of Merit; HCM: Hierarchical Clustering Method; IHD: Ischemic Heart Disease; IQCAMP: Iran Quality of Care in Medicine Program; MCM: Model-based Clustering Method; SBP: Systolic Blood Pressure; SPO:

Donabedian Structure-Process-Outcome; SRS: Simple Random Sampling; STEPS: STEPwise approach to Surveillance Non-Communicable Diseases study.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and material:

The data that support the findings of this study are available from Non-Communicable Disease Research Center (NCDRC), Tehran, Iran. Data are available from the authors upon reasonable request and with permission of NCDRC.

Competing interests:

SSH is a member of the editorial board of the BMC Public Health. Other co-authors declare that they have no competing interests.

Funding:

IQCAMP project was supported by the IR Iran National Institute for Medical Research Development [NIMAD 942936]. The funding body had no role in the design of the study, analysis, and interpretation of data or in writing the manuscript.

Authors' contributions:

MP, SSH, FF, MS contributed in study planning and design. MS, ASH contributed in data collection, data preparation. MP, MS, PM conducted the statistical analyses. SSH, FF, and SHKH commented on the results. MM and MP drafted the first manuscript and revised the manuscript after input from the other authors. All authors read and approved the final manuscript.

Acknowledgements:

We would like to express our appreciation to Non-Communicable Disease Research Center (NCDRC), Tehran, Iran that provided us the input data to conduct this research.

References

1. Paul S, Levy, Lemeshow S. Sampling of Populations: Methods and Applications. 4th ed. Hoboken: John Wiley & Sons; 2008.
2. Pfeffermann D, Rao CR. Sample Surveys: Design, Methods and Applications. 1st ed. Amsterdam: Elsevier; 2009.

3. Lehtonen R, Pahkinen E. Practical Methods for Design and Analysis of Complex Surveys. 2nd ed. England: John Wiley & Sons; 2004.
4. Valliant R, Dever JA, Kreuter F. Practical Tools for Designing and Weighting Survey Samples. 1st ed. New York: Springer; 2018.
5. Donald H, McBurney, White TL: **Research Methods**, 8th edn. United States: Wadsworth, Cengage Learning; 2010.
6. Murtagh F. A Survey of Recent Advances in Hierarchical Clustering Algorithms. Comput J. 1983;26(4):354–9.
7. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540–2.
8. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. The R journal. 2016;8(1):289–317.
9. Djalalinia S, Modirian M, Sheidaei A, Yoosefi M, Zokaiee H, Damirchilu B, et al. Protocol Design for Large-Scale Cross-Sectional Studies of Surveillance of Risk Factors of Non-Communicable Diseases in Iran: STEPs 2016. Arch Iran Med. 2017;20(9):608–16.
10. Etemad K, Yavari P, Mehrabi Y, Haghdoost A, Motlagh ME, Kabir MJ, et al. Inequality in Utilization of In-patients Health Services in Iran. International journal of preventive medicine. 2015;6:45.
11. Niakan Kalhori SR, Tayefi B, Noori A, Mearaji M, Rahimzade S, Zandian E, et al. Inpatient data, inevitable need for policy making at national and sub-national levels: a lesson learned from NASBOD. Arch Iran Med. 2014;17(1):16–21.
12. Sheidaei A, Gohari K, Kasaeian A, Rezaei N, Mansouri A, Khosravi A, et al. National and Subnational Patterns of Cause of Death in Iran 1990–2015: Applied Methods. Arch Iran Med. 2017;20(1):2–11.
13. Mooi E, Sarstedt M. A concise guide to market research: The process, data, and methods using IBM SPSS Statistics. New York: Springer; 2011.
14. Kassambara A: **Practical guide to cluster analysis in R**: **Unsupervised machine learning**, vol. 1: STHDA; 2017.
15. Donabedian A. Evaluating the quality of medical care. 1966. Milbank Q. 2005;83(4):691–729.
16. Mahdavi M, Vissers J, Elkhuizen S, van Dijk M, Vanhala A, Karampli E, et al. The relationship between context, structure, and processes with outcomes of 6 regional diabetes networks in Europe. PloS one. 2018;13(2):e0192599.
17. Ackerman M, Ben-David S. **Clusterability: A theoretical study**. In: *Artificial Intelligence and Statistics: 2009*; 2009: 1–8.
18. Charrad M, Ghazzali N, Boiteau V, Niknafs A. **NbClust Package: finding the relevant number of clusters in a dataset**. *UseR! 2012* 2012.
19. Brock G, Pihur V, Datta S, Datta S: **clValid**, **an R package for cluster validation**. *Journal of Statistical Software (Brock et al, March 2008)* 2011.

20. Kassambara A: **Machine Learning Essentials**: **Practical Guide in R**: CreateSpace Independent Publishing Platform; 2018.
21. Torgo L: **Data mining with R**: **learning with case studies**: Chapman and Hall/CRC; 2011.
22. Kim YJ, Oh Y, Park S, Cho S, Park H. Stratified Sampling Design Based on Data Mining. Healthcare Informatics Research. 2013;19(3):186–95.
23. Miller TJ, Skalski JR, Ianelli JN. Optimizing a stratified sampling design when faced with multiple objectives. ICES J Mar Sci. 2007;64(1):97–109.
24. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001;17(10):977–87.
25. Rana S, Garg R: **Application of Hierarchical Clustering Algorithm to Evaluate Students Performance of an Institute**. In: *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT): 12–13 Feb. 2016 2016*; 2016: 692–697.
26. Hirano S, Sun X, Tsumoto S. Comparison of clustering methods for clinical databases. Inf Sci. 2004;159(3):155–65.
27. Chang Ch, Chen SY, Lan YT. Service quality, trust, and patient satisfaction in interpersonal-based medical service encounters. BMC Health Serv Res. 2013;13(1):22.
28. Aliman NK, Mohamad WN. Linking service quality, patients' satisfaction and behavioral intentions: an investigation on private healthcare in Malaysia. Procedia-Social Behavioral Sciences. 2016;224:141–8.
29. Sumaedi S, Yarmen M, Yuda Bakti IGM. Healthcare service quality model: a multi-level approach with empirical evidence from a developing country. International Journal of Productivity Performance Management. 2016;65(8):1007–24.
30. Denyer D, Tranfield D, Van Aken JE. Developing design propositions through research synthesis. Organization studies. 2008;29(3):393–413.

Tables

Table 1: A list of input HCQC indicators included in the clustering methods

Factor	Variable name	Variable definition	Geographical unit	Data Source	Year
Demand/ Disease patterns	Inpatient	Annual average number of inpatient	District	Utilization	2014
	Outpatients	Annual average number of outpatient	District	Utilization	2014
	Hospitalization rate	Hospitalization rate per 1000 population	Province	Hospital Data	2011
	Patient exchange rate	Ratio of sending referrals to receiving referrals (Patient exchange rate)	Province	Hospital Data	2011
	SBP	Mean SBP among hypertensive patients	District	STEPs	2016
	Glucose	Mean of glucose among patients with DM	District	STEPs	2016
	Cholesterol	Mean cholesterol among patients with hyperlipidemia	District	STEPs	2016
Structure	Basic insurance coverage	Basic insurance coverage (% of the population with basic insurance)	District	Utilization	2014
	Complementary insurance coverage	Complementary insurance coverage (% of the population with complementary insurance)	District	Utilization	2014
	Bed density	Number of beds per 1000 population	Province	Hospital Data	2011
	Physician density	Number of physicians per 1000 population	Province	Hospital Data	2011
Outcome	Probability of dying IHD	Probability of dying from IHD* among adults (age≥30)	Province	DRS	2015
	Probability of dying Stroke	Probability of dying from Stroke among adults (age≥30)	Province	DRS	2015
	Probability of dying COPD	Probability of dying from COPD* among adults (age≥30)	Province	DRS	2015
	Probability of dying Diabetes	Probability of dying from Diabetes mellitus among adults (age≥30)	Province	DRS	2015
	Probability of dying CKD	Probability of dying from CKD* among adults (age≥30)	Province	DRS	2015
	Neonatal mortality rate	Neonatal mortality per 1000 live births	District	DRS	2015
	Adverse effect mortality	Mortality rate due to the adverse effect of medical treatment	Province	DRS	2015
	All-cause mortality ratio	Expected mortality rate to observed mortality rate	Province	DRS	2015
	Mortality rate in hospital	Mortality rate among 1000 hospitalized patients	Province	Hospital Data	2011

SBP: Systolic Blood Pressure, DM: Diabetes Mellitus, IHD: Ischemic Heart Disease, COPD: Chronic Obstructive Pulmonary Disease, CKD: Chroni Kidney Disease.

District is defined as a geographical region with administrative boundaries and an independent network of healthcare provisioning. Provinc comprises a set of districts and has few managerial authorities in planning and organization of health services. Provincial level is the first level c country subdivisions. Input data consist of data from 31 provinces and 413 districts of Iran.

Utilization study measures the use of inpatient and outpatient health services by individuals using a representative sample of population.

STEPs is a national survey based on the WHO stepwise approach to study non-communicable disease risk factors.

DRS abbreviates Death Registration System.

Hospital Data is a research project that studies 0.5% of all inpatient cases in hospitals owned by Ministry of Health and Medical Education in 2011 i Iran.

The adverse effect of medical treatment refers to unintended consequences of any types of medical interventions including prevention, diagnosis treatment, and rehabilitation.

Table 2: Comparison of internal and stability validity by clustering method

Validity Indexes	MCM-8 Model-based with eight clusters	HCM-2 Hierarchical with two clusters	HCM-8 Additional scenario; HCM with eight clusters
Internal Validity Indexes			
Within-clusters Sum of Squares*	**255.87**	384.55	292.65
Average silhouette width**	0.14	**0.17**	0.09
Dunn index**	**0.27**	0.20	0.19
Stability Validity Indexes			
Average Proportion of Non-overlap (APN)***	0.13	**0.09**	0.19
Average Distance AD***	**1.11**	1.37	1.24
Average distance between means (ADM)***	**0.19**	0.19	0.28
Figure of Merit (FOM)***	**0.21**	0.24	0.22

* The lower the value of the within-cluster sum of square, the higher the extent of compactness.

** The higher the value of Dunn index and average silhouette width, the higher the extent of compactness and separation.

*** For all stability indices, smaller values indicate better stability validity.

Table 3: Comparison of efficiency of clustering-based sampling to SRS based on distinct cluster features

Variables	Sampling efficiency
The ratio of sending referrals to receiving referrals (Patient exchange rate)	1.5
The probability of dying from Stroke (age≥30)	1.7
The probability of dying from COPD* (age≥30)	1.5
The probability of dying from CKD** (age≥30)	1.4
The mortality rate attributed to the adverse effect of medical treatment	1.2
Expected mortality rate to observed mortality rate	1.3
The mortality rate among 1000 hospitalized patients	1.4

* COPD: Chronic Obstructive Pulmonary Disease

** CKD: Chronic Kidney Disease

Figures

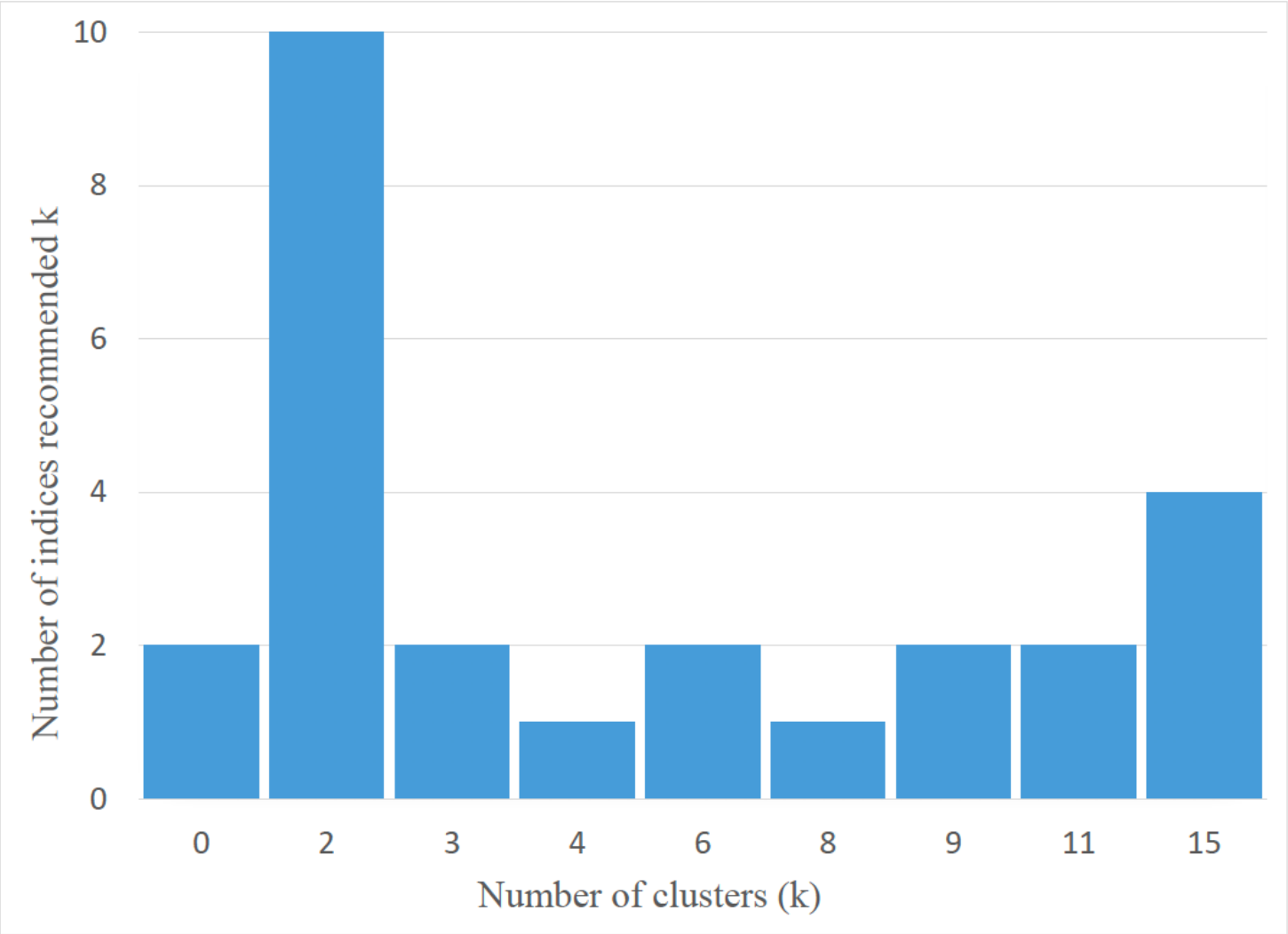

Figure 1

Proposed number of clusters by NbClust package

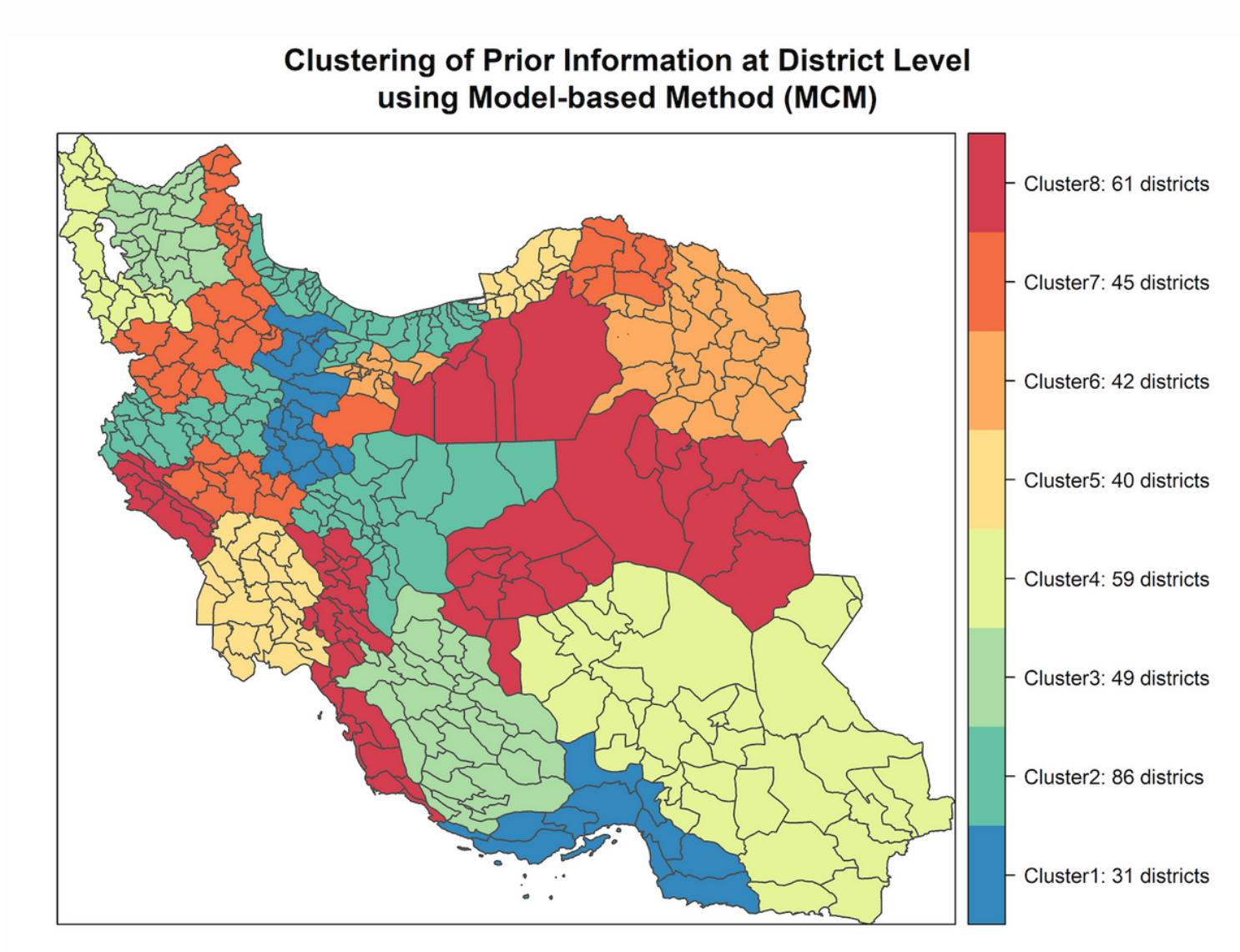

Figure 2

Geographic distribution of 8 clusters identified by the Model-based Clustering Method

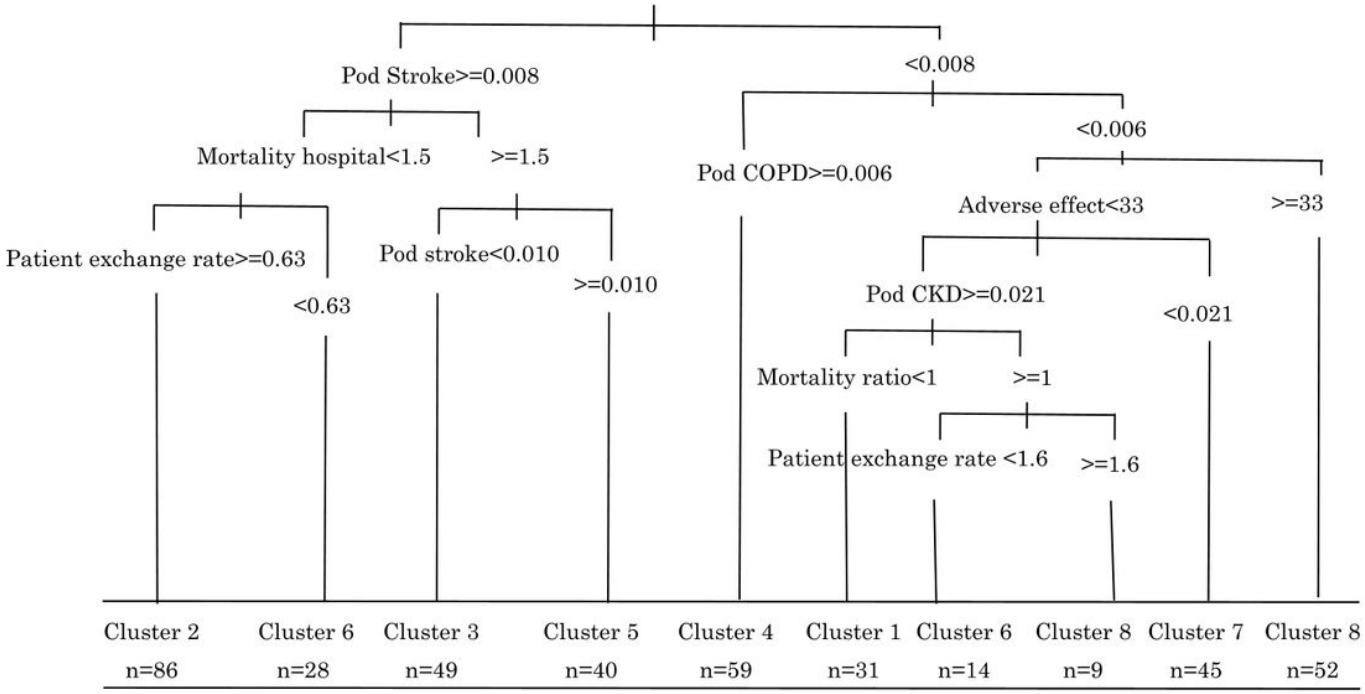

Figure 3

Using decision tree learning to describe distinctive features of 8 clusters identified by the Model-based Clustering Method

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Figure1additionalfile1.tiff
- Additionalfile1.docx
- Additionalfile2RCode.pdf