Original Article

Healthc Inform Res. 2013 September;19(3):186-195.
http://dx.doi.org/10.4258/hir.2013.19.3.186
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Stratified Sampling Design Based on Data Mining

Yeonkook J. Kim, MS, MA[1], Yoonhwan Oh, MS[1], Sunghoon Park, MS[2], Sungzoon Cho, PhD[2], Hayoung Park, PhD[1]
[1]Technology Management, Economics and Policy Graduate Program, Seoul National University, Seoul; [2]Department of Industrial Engineering, Seoul National University, Seoul, Korea

Objectives: To explore classification rules based on data mining methodologies which are to be used in defining strata in stratified sampling of healthcare providers with improved sampling efficiency. **Methods:** We performed k-means clustering to group providers with similar characteristics, then, constructed decision trees on cluster labels to generate stratification rules. We assessed the variance explained by the stratification proposed in this study and by conventional stratification to evaluate the performance of the sampling design. We constructed a study database from health insurance claims data and providers' profile data made available to this study by the Health Insurance Review and Assessment Service of South Korea, and population data from Statistics Korea. From our database, we used the data for single specialty clinics or hospitals in two specialties, general surgery and ophthalmology, for the year 2011 in this study. **Results:** Data mining resulted in five strata in general surgery with two stratification variables, the number of inpatients per specialist and population density of provider location, and five strata in ophthalmology with two stratification variables, the number of inpatients per specialist and number of beds. The percentages of variance in annual changes in the productivity of specialists explained by the stratification in general surgery and ophthalmology were 22% and 8%, respectively, whereas conventional stratification by the type of provider location and number of beds explained 2% and 0.2% of variance, respectively. **Conclusions:** This study demonstrated that data mining methods can be used in designing efficient stratified sampling with variables readily available to the insurer and government; it offers an alternative to the existing stratification method that is widely used in healthcare provider surveys in South Korea.

Keywords: Sampling Studies, Decision Trees, Data Mining

Submitted: August 16, 2013
Revised: September 28, 2013
Accepted: September 28, 2013

Corresponding Author
Hayoung Park, PhD
Technology Management, Economics and Policy Graduate Program, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea. Tel: +82-2-880-2575, Fax: +82-2-873-7229, E-mail: hayoungpark@snu.ac.kr

I. Introduction

Stratified random sampling or stratified sampling, as opposed to simple random sampling, is often used in the field of healthcare management and policy [1]. A stratified sample is defined as one resulting from classification of population into mutually exclusive groups, called strata, and choosing a simple random sample from each stratum. The main reason for using stratified sampling instead of simple random sampling is improved efficiency of sampling [2,3]. Sampling efficiency is the amount of information obtained for a given sampling cost, and the efficiency of stratified sampling is usually better than the efficiency of simple random sampling because the classification of population into strata can reduce variability of measurements within a stratum and result in smaller bounds on estimation errors [2,3]. Therefore, stratification should reduce variability within strata and increase variability between strata to achieve the intended improvement of sampling efficiency. Well-designed stratified sampling should be able to define strata where measurements within strata are homogeneous.

Stratified sampling is widely used in studies regarding healthcare policy formulation to sample healthcare providers in South Korea. The variables often used to define strata in these studies include the type of medical facilities defined by the medical law—tertiary hospitals, general hospitals, hospitals, and clinics, the type of location defined by the local autonomy law—district of special and metropolitan city, county of special and metropolitan city, city of province, and county of province, the size of providers defined by the number of beds a provider operates, and the type of ownership. However, the appropriateness of the variables in defining strata within which measurements should be homogeneous has been rarely studied, although managerial and clinical characteristics of providers vary widely within categories of the variables used to define strata. Conflicting interests of stakeholders are common in healthcare policy scenes, and the appropriateness and generalizability of samples has been a source of controversy as to the validity of study findings. Scheaffer et al. [3] suggested 3 criteria to be considered in deciding the design of stratified sampling: 1) measurements within strata should be homogeneous, 2) the survey cost should be reduced by stratification, and 3) estimates of population parameters for each stratum should be useful.

The objective of this study was to demonstrate how data mining methodologies can be used to design stratification rules that improve sampling efficiency compared to designs that are generally used in policy studies in South Korea. We introduce the method by presenting the study that we conducted as a part of a project supported by a grant from the Health Insurance Review and Assessment Service (HIRA) of South Korea; in the study, a framework was developed for updating payment rates in the Diagnosis Related Groups (DRGs) based prospective payment system as well as work processes, including sampling healthcare providers and collecting data needed to analyze updates from the sample [4]. The project concerned factors that affect changes in the cost of treating patients that should be factored into payment updates. Examples of those factors are changes in prices of input resources, such as labor and materials, and changes in productivity.

In this study, we explored variables and classification rules based on data mining methodologies for defining strata in surveys of healthcare providers. We performed cluster analysis to segment medical facilities with similar characteristics. Then, decision trees were created to generate rules to clearly identify and analyze each cluster, which is used to define sampling strata. We compared the efficiency of the proposed sampling design with that of a conventional design of stratified sampling to assess the effectiveness of this approach.

In data mining or the process of Knowledge Discovery in Databases, interesting patterns and knowledge are discovered from large amounts of data. A pattern is interesting if it is valid on test data with some degree of certainty, novel, useful, and easy to understand [5]. Data mining involves a number of common classes of tasks. One of them is cluster analysis, or simply clustering. It is the process of partitioning a set of data observations into subsets. Each subset is a cluster, such that observations in a cluster are similar to one another, yet dissimilar to observations in other clusters. Another task is classification. Classification is a form of data analysis that extracts models or rules describing important data classes or labels. Such models, called classifiers, predict categorical class labels, which are discrete and unordered.

One natural application of cluster analysis is consumer segmentation in consumer relationship management (CRM). For example, Hung et al. [6] used k-means clustering methods to segment the customers in the Taiwanese telecommunication market into five clusters of roughly equal size according to their billing amounts, tenure months, and payment behaviors. Then, they created a decision tree model in each cluster to see if the churn behaviors, or attrition rates, differed for various segments. There was also a study that applied k-means clustering and decision tree algorithms to detect healthcare providers with abusive billing patterns in South Korea [7]. Ngai et al. [8], in their review article, showed that decision tree and k-means clustering were highly popular data mining tools in CRM research.

II. Methods

1. Analysis

This research selected k-means clustering and decision tree induction as data mining techniques to segment and classify healthcare providers. The k-means algorithm is a method of cluster analysis which aims to partition n observations into a predetermined number of k clusters in which each observation belongs to the cluster with the nearest mean. The algorithm proceeds as follows. First, it randomly selects k of the observations in the data set, each of which initially represents a cluster mean or center. For each of the remaining observations, an observation is assigned to the cluster to which it is the most similar, based on the Euclidean distance between the observation and the cluster mean. The k-means algorithm then iteratively improves the within-cluster variation, and the iterations continue until the assignment becomes stable. On the other hand, the construction of decision tree classifiers does not require any domain knowledge or parameter setting. Decision trees can easily be converted to classification rules that are intuitive and generally easy to understand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology [5].

We studied a sampling design for single specialty clinics or hospitals in the two specialties, general surgery and ophthalmology. We chose the nine most relevant and representative variables to be used in clustering among 94 highly correlated variables we had in the study database of healthcare providers. They included type of provider location (REG), population density of the region (POPD), number of specialists the provider had (MD_SPEC), number of beds (FAC_SICK), number of inpatients per specialist (NPAT_SPEC), lengthiness index (LI), costliness index (CI), case-mix index (CMI), and rate of annual change in number of inpatients per specialist (CHG_NPAT_SPEC). Details of variable definition are given in Table 1. The number of clusters for each specialty of clinics and hospitals was decided based on discussions with experts in the project team and Silhouette coefficients, which measure cohesion and separation of clusters. Clustering was performed using clustering functions in the Matlab ver. 7.11 program (MathWorks Inc., Natic, MA, USA). Then, Spotfire ver. 4.5 (TIBCO Spotfire, Somerville, MA, USA) was used to

Table 1. Definition and data sources of the study variables

Variable	Description/definition	Data source
Type of provider location (REG)	1: District of special and metropolitan city, 2: County of special and metropolitan city, 3: City of province, 4: County of province,	HIRA provider profile data
Population density (POPD)	Population per square kilometer of the administrative region the provider is located at the level of city/county/district	KOSTAT data
Type of ownership (ESTA_CD)	1: National and public, 2: Foundation except for medical foundation, 3: Medical foundation, 4: Private, 5: Military, 6: Other	HIRA provider profile data
Number of specialists (MD_SPEC)	Total numberof specialists	HIRA provider profile data
Number of beds (FAC_SICK)	Total number of beds in operation	HIRA provider profile data
Number of inpatients per specialist (NPAT_SPEC)	(Total number of inpatients in DRGs paid by the DRG-based prospective payment system and treated by the specialty)/(total number of specialists in the specialty)	HIRA provider profile data and claims data
Lengthiness index (LI)	(Mean length of stay)/(case-mix adjusted expected mean length of stay)	HIRA claims data
Costliness index (CI)	(Mean charges)/(case-mix adjusted expected mean charges)	HIRA claims data
Case-mix index (CMI)	(case-mix adjusted expected mean charges)/(mean charges of inpatients discharged from all providers in the type the provider is assigned)	HIRA claims data
Rate of annual change in number of inpatients per specialist (CHG_NPAT_SPEC)	{(2-year moving average of the number of inpatients per specialist at the year t)–(2-year moving average of the number of inpatients per specialist at the year t-1)} / (2-year moving average of the number of inpatients per specialist at the year t-1)	HIRA provider profile data and claims data

HIRA: Health Insurance Review and Assessment Service, KOSTAT: Statistics Korea, DRG: Diagnosis Related Group.

visualize the clustering results. We used SAS E-miner ver. 4.3 (SAS Institute Inc., Cary, NC, USA) to create decision trees.

We evaluated the performance of the proposed sampling design by assessing the degree of variance reduced or explained by the classification of clinics and hospitals with a single specialty into strata as a measure of sampling efficiency following the suggestion by Scheaffer et al. [3]. We compared the efficiency of the stratification proposed in this study with the efficiency of conventional stratification based on the type of administrative region of provider location and the size of inpatient bed. The type of administrative region had four categories, namely, district of special and metropolitan city, county of special and metropolitan city, city of province, and county of province; size had two categories, that is, with 30 or more beds and with fewer than 30 beds. We conducted an analysis of variance (ANOVA) to compute the degree of variance reduction achieved by stratifications. The dependent variable used in the analysis was the annual rate of change in the number of inpatients per specialist which is a measure of changes in productivity, and the categorical independent variables were strata. SAS ver. 9.3 (SAS Institute Inc.) was used for statistical computations and tests.

2. Data

We constructed a study database of providers with data from the HIRA and Statistics Korea (KOSTAT). The major sources of data from the HIRA were inpatient insurance claims data and provider profile data. The inpatient insurance claims data comprised claims submitted to the HIRA by providers for insurance payments from January 2006 to December 2011, and there was a total of 50,769,933 records. Each record included claims information as to a hospitalization for inpatient care,

Table 2. General characteristics of clinics and hospitals in each cluster

Cluster	Variable	General surgery		Ophthalmology	
		No. of providers/mean	Percent/SD	No. of providers/mean	Percent/SD
1	Hospital	1	2	-	-
	Clinic	54	98	96	100
	Private	55	100	96	100
	District, special and metropolitan city	22	40	51	53
	County, special and metropolitan city	1	2	1	2
	City, province	21	38	41	43
	County, province	11	20	3	3
	Population density[a]	4,569	5,898	8,263	7,105
	Number of specialists	1	0	1	1
	Number of beds	12	16	1	2
	Number of inpatients per specialist[b]	22	27	26	17
	Change rate of the number of inpatients per specialist[b]	0.23	0.34	−0.13	0.20
2	Hospital	2	2	4	4
	Clinic	98	98	89	96
	Foundation except for medical foundation	1	1	-	-
	Medical foundation	1	1	1	1
	Private	98	98	92	99
	District, special and metropolitan city	41	41	42	45
	County, special and metropolitan city	2	2	-	-
	City, province	42	42	50	54
	County, province	15	15	1	1
	Population density[a]	6,446	7,648	6,226	5,475
	Number of specialists	1	0	3	2
	Number of beds	15	12	8	9
	Number of inpatients per specialist[b]	28	45	410	229
	Change rate of the number of inpatients per specialist[b]	−0.33	0.15	0.07	0.10

Table 2. Continued

Cluster	Variable	General surgery		Ophthalmology	
		No. of providers/mean	Percent/SD	No. of providers/mean	Percent/SD
3	Hospital	6	2	-	-
	Clinic	301	98	263	100
	Private	307	100	263	100
	District, special and metropolitan city	153	50	115	44
	County, special and metropolitan city	-	-	3	1
	City, province	152	50	112	43
	County, province	2	1	33	13
	Population density[a]	7,313	6,596	6,292	6,806
	Number of specialists	1	1	1	1
	Number of beds	13	8	1	2
	Number of inpatients per specialist[b]	381	224	222	150
	Change rate of the number of inpatients per specialist[b]	−0.01	0.13	−0.07	0.10
4	Hospital	-	-	-	-
	Clinic	-	-	291	100
	Private	-	-	291	100
	District, special and metropolitan city	-	-	135	46
	County, special and metropolitan city	-	-	2	1
	City, province	-	-	129	44
	County, province	-	-	25	9
	Population density[a]	-	-	7,228	7,361
	Number of specialists	-	-	1	0
	Number of beds	-	-	1	1
	Number of inpatients per specialist[b]	-	-	285	219
	Change rate of the number of inpatients per specialist[b]	-	-	0.24	0.21

SD: standard deviation.

[a]Number of persons per kilometer. [b]Inpatients include only those classified to the Diagnosis Related Groups (DRGs) paid by the DRG-based prospective payment system and treated by the specialty concerned.

and the information items in each record were patient ID, provider ID, type of provider, Korean DRG code, gender, age, length of stay, and charges. Provider profile data comprised provider information as of the end of each calendar year from 2006 to 2011, and there was a total of 174,081 records. Each record included provider ID, type of medical facilities, type of ownership, establishment date, numbers of general doctors, specialists in each of 26 specialties, nurses, and inpatient beds. We attached the population density statistics obtained from KOSTAT to the profile data by the provider location and year [9]. We computed yearly statistics using claims data for each provider, such as total numbers of inpatients and inpatients in the four specialties of—general surgery, obstetrics and gynecology, ophthalmology, and otolaryngology—who were paid by the DRG-based prospective payment system, LI, CI, and CMI, and merged them with provider profile data by provider ID and year. We further computed measures of physician productivity in the four specialties in terms of the number of inpatients per specialist with merged data. We used the two-year moving average to compute annual change in physician productivity which was used as the dependent variable in the ANOVA. Table 1 shows the detailed definition of study variables and data sources.

We excluded providers which were established in the year 2009 or after for the completeness of data and excluded those with an absolute change rate value greater than 0.5 as outliers from the study. Finally, we used data for the year 2011 in this study, and 442 clinics and hospitals with the specialty of general surgery and 715 clinics and hospitals with the specialty of ophthalmology were included in the analyses.

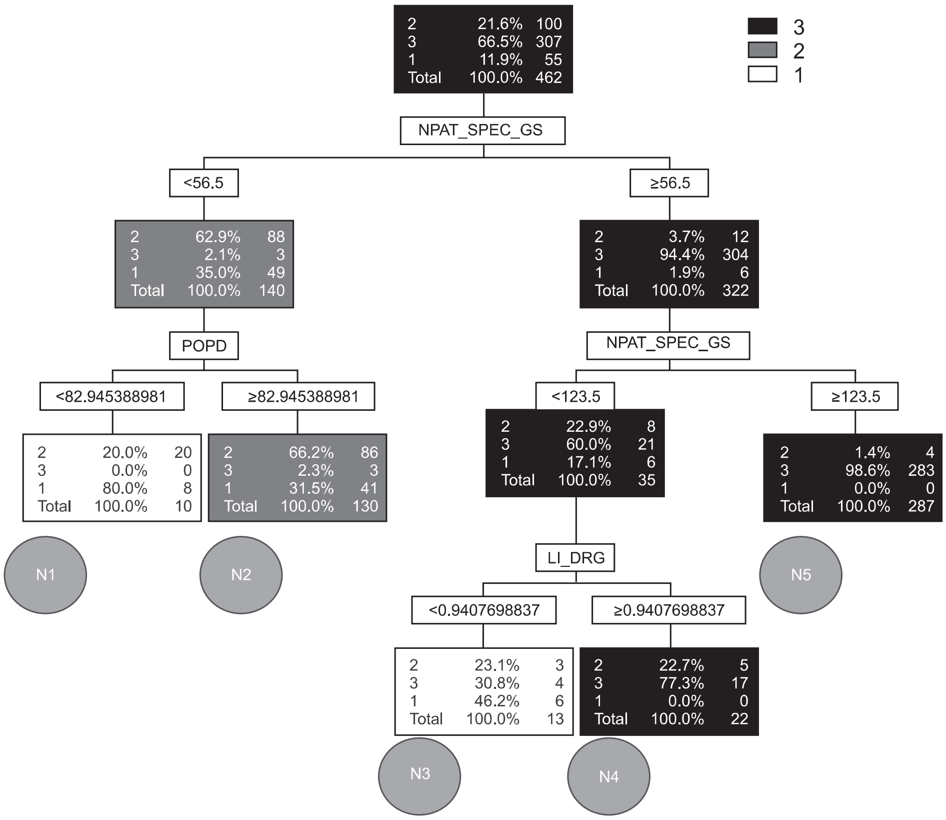

Figure 1. Decision tree inducted to stratify general surgery (GS) clinics and hospitals. NPAT_SPEC: number of inpatients per specialist, POPD: population density of the region, LI: lengthiness index, DRG: Diagnosis Related Group.

III. Results

1. Clustering

We identified three clusters of general surgery clinics and hospitals, and their general characteristics are presented in Table 2. Providers in cluster 1 were mostly private clinics in rural areas with a low number of inpatients per specialist and a decreasing number of inpatients in the DRGs paid by the DRG-based prospective payment system. Those in cluster 2 were relatively large private clinics and hospitals with a decreasing number of inpatients, and those in cluster 3 were mostly urban clinics with a large number of inpatients.

We identified four clusters of ophthalmology clinics and hospitals, and their general characteristics are presented in Table 2. Providers in the cluster 1 were relatively small private clinics located in metropolitan areas with a low number of inpatients in the DRGs paid by the DRG-based prospective payment system. Those in cluster 2 were large clinics with multiple specialists and beds for inpatient care, and those in cluster 3 were private clinics in rural areas with a decreasing number of inpatients in the DRGs paid by the DRG-based prospective payment system. Providers in cluster 4 were large private clinics with an increasing number of inpatients.

2. Decision Tree and Rules to be Used in Stratification

The decision tree we constructed to identify three clusters of general surgery clinics and hospitals is shown in Figure 1. Three variables were used to classify clinics and hospitals into homogeneous groups or strata, namely, the number of inpatients per specialist who were in the DRGs paid by the DRG-based prospective payment system (NPAT_SPEC_GS), population density (POPD), and lengthiness index (LI_DRG). The suggested classification rule would result in five strata:

- Stratum 1: clinics and hospitals with fewer than 57 inpatients per specialist and located in a city/county/district with a population density lower than 83,
- Stratum 2: clinics and hospitals with fewer than 57 inpatients per specialist and located in a city/county/district with a population density higher than 82,
- Stratum 3: clinics and hospitals with 57 to 123 inpatients per specialist whose lengthiness index was less than 0.95,
- Stratum 4: clinics and hospitals with 57 to 123 inpatients per specialist whose lengthiness index was larger than 0.94,
- Stratum 5: clinics and hospitals with more than 123 inpatients per specialist.

The general characteristics of clinics and hospitals in each stratum are presented in Table 3. The results show that clin-

Table 3. General characteristics of clinics and hospitals in each stratum

Stratum	Variable	General surgery		Ophthalmology	
		No. of providers/mean	Percent/SD	No. of providers/mean	Percent/SD
1	Hospital	-	-	-	-
	Clinic	11	100	155	100
	District, special and metropolitan city	-	-	88	57
	County, special and metropolitan city	-	-	2	1
	City, province	3	27	62	40
	County, province	8	73	3	2
	Number of inpatients[a]	58	83	39	35
	Number of specialists	1	0	1	1
	Number of beds	9	9	1	2
2	Hospital	5	3	-	-
	Clinic	156	97	520	10
	District, special and metropolitan city	61	38	244	47
	County, special and metropolitan city	4	2	4	1
	City, province	78	48	220	42
	County, province	18	11	52	10
	Number of inpatients[a]	145	183	256	203
	Number of specialists	1	0	1	1
	Number of beds	17	17	0	1
3	Hospital	8	2	-	-
	Clinic	318	98	98	100
	District, special and metropolitan city	167	51	25	26
	County, special and metropolitan city	-	48	2	2
	City, province	156	-	45	46
	County, province	3	1	26	27
	Number of inpatients[a]	618	432	716	297
	Number of specialists	1	1	1	0
	Number of beds	13	9	1	1
4	Hospital	-	-	1	6
	Clinic	22		17	94
	District, special and metropolitan city	13	59	13	72
	County, special and metropolitan city	-	-	-	-
	City, province	7	32	5	28
	County, province	2	9	-	-
	Number of inpatients[a]	170	155	291	207
	Number of specialists	1	0	2	1
	Number of beds	9	7	7	6
5	Hospital	-	-	5	6
	Clinic	19	100	80	94
	District, special and metropolitan city	7	37	37	44
	County, special and metropolitan city	-	-	-	-
	City, province	11	58	47	55
	County, province	1	5	1	1
	Number of inpatients[a]	305	196	1,157	815
	Number of specialists	1	0	3	2
	Number of beds	19	7	10	10

SD: standard deviation.

[a] Inpatients include only those classified to the Diagnosis Related Groups (DRGs) paid by the DRG-based prospective payment system and treated by the specialty concerned.

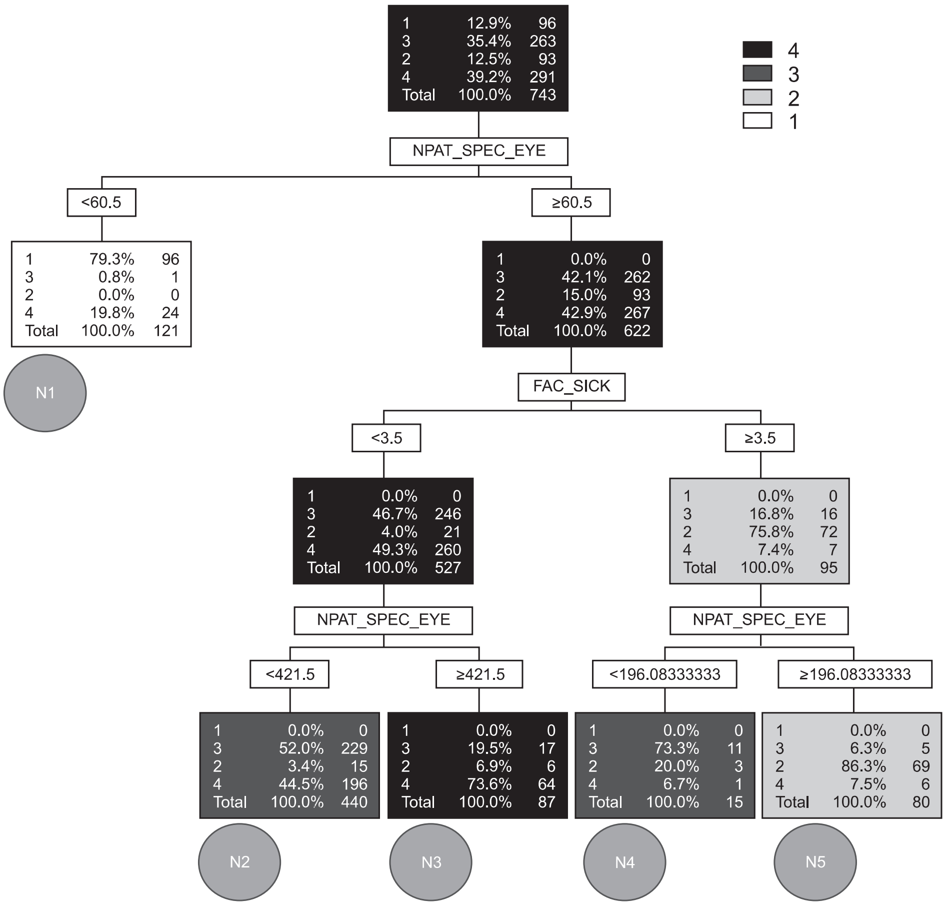

Figure 2. Decision tree inducted to stratify ophthalmology clinics and hospitals. NPAT_SPEC: number of inpatients per specialist, FAC_SICK: number of beds.

Table 4. Evaluation results from the analyses of variance

	No. of observations	Stratification based on data mining		Conventional stratification	
		No. of strata	Variance reduction (%)	No. of strata	Variance reduction (%)
General surgery	442	5	22.24	8	1.77
Ophthalmology	715	5	8.29	8	0.19

ics and hospitals in strata 2 and 3 had significantly different inpatient volumes in the DRGs paid by the DRG-based prospective payment system in general surgery even though the characteristics of the type of location and the number of beds were not very different, which means they may have been classified into a stratum in a conventional stratified sampling that uses the type of location and the number of beds as stratification variables. Also, clinics in strata 4 and 5 appear to be similar in terms of the type of location and the number of specialists even though the productivities of specialist measured by the number of inpatients per specialist are not.

The decision tree we constructed to identify four clusters of ophthalmology clinics and hospitals is shown in Figure 2. Two variables were used to classify clinics and hospitals into homogeneous groups or strata, namely, the number of patients per specialist who were in the DRGs paid by the DRG-based prospective payment system (NPAT_SPEC_EYE) and the number of beds (FAC_SICK). The suggested classification rule would result in five strata:

- Stratum 1: clinics and hospitals with fewer than 60 inpatients per specialist,
- Stratum 2: clinics and hospitals with fewer than 4 beds and with 60 to 421 inpatients per specialist,
- Stratum 3: clinics and hospitals with fewer than 4 beds and with more than 421 inpatients per specialist,
- Stratum 4: clinics and hospitals with more than 3 beds and with 60 to 196 inpatients per specialist,
- Stratum 5: clinics and hospitals with more than 3 beds and with more than 196 inpatients per specialist.

The general characteristics of clinics and hospitals in each stratum are presented in Table 3. The patient volumes of clinics in strata 1 and 2 were different, although they were mostly located in urban areas with a single specialist in ophthalmology, which implies that clinics in stratum 1 may focus in ophthalmology care not covered by the national health insurance. Stratum 3 includes clinics with a single specialist and a single bed in all types of location whose patient volume in the DRGs paid by the DRG-based prospective payment system is heavy; therefore, the physician productivity in ophthalmology care in those DRGs is high.

3. Evaluation

The performance of the stratification using the variables and rules found through data mining methods were far better than the performance of conventional stratification based on the type of location and bed size in both specialties (Table 4). The percentages of variance explained by the stratification based on data mining methods for general surgery and ophthalmology clinics and hospitals were 22% and 8%, respectively, whereas the percentages explained by conventional stratification were 2% and 0.2%; although fewer strata were created by data mining methods than by conventional stratification. These findings imply that the homogeneity of measurements, changes in physician productivity in this evaluation, within strata is better with stratification by data mining methods than that of conventional stratification; thus, sampling efficiency is improved.

IV. Discussion

This study attempted to find a stratified sampling design based on data mining methods that achieves improved sampling efficiency over designs conventionally used in studies of healthcare providers for management and policy decisions in South Korea. Utilizing widely used data mining methods, we wanted to provide an explanatory study that would offer an efficient alternative that can be easily adopted by health care professionals. Specifically, cluster analysis groups a data set according to perceived intrinsic characteristics or similarity and is known to find structure in data which, in turn, can be identified as stratification rules by the application of decision tree induction on the cluster labels [5,10]. The stratification rules found in this study defined strata with far better homogeneity within strata in measurements with several classification variables and fewer strata than conventional stratified sampling did. In our evaluation study of clinics and hospitals in the specialties of general surgery and ophthalmology, the performance of stratification by the type of provider location and bed size which is widely used in policy studies in South Korea was so inadequate that the stratification did not add any value in sampling efficiency over simple random sampling. Clinics and hospitals with single specialty were quite heterogeneous within categories of the type of provider location and bed size, the variables routinely used to classify providers to homogenous groups, and they were not good predictors at all.

Fueled by conflicting interests of stakeholders, inappropriate sampling and generalizability of samples have been major sources of controversy over findings of policy studies, particularly studies concerning payment rates which are affected by providers' managerial, financial, and environmental characteristics. This study demonstrated that data mining methods can be applied to find an intelligent sampling design with data that are routinely collected and readily available to the insurer, the government, and researchers. Stratification rules may need to be updated in three to five years as the economic, social, and policy environment changes. The sensitivity of sampling results and performance to changes in classification thresholds should be further investigated to achieve robust sampling designs.

This study is not free from limitations, and further studies are needed. First, we did not examine the applicability of the sampling designs found in this study in a variety of study settings. Although we can assume that the classification rules could hold in studies concerning providers' managerial and financial performance, we did not explicitly look into the applicability of study findings in this regard. Second, the study findings suggest different classification variables and rules for providers in different specialties and this may be obvious to achieve maximum efficiency. However, we did not explore whether it would be optimal to have different rules for different groups of providers in terms of sampling efficiency, overall costs, and the simplicity of sampling schemes. Third, we did not perform out-of-sample analysis in building our decision trees. However, overfitting would not be a major problem in this study due to pruning procedures we performed. In a future study, we may want to perform out-of-sample analysis and compare the predictability of our model with other popular data mining methods, such as support vector machines. Lastly, during the variable selection process in the clustering analysis, due to time constraints, we automatically excluded variables with 80% or more missing values from the analyses. However, in a future study, we may want use an appropriate algorithm to impute missing values for some key variables and test if the use of these variables would result in better clustering.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Graubard BI, Korn EL. Modelling the sampling design in the analysis of health surveys. Stat Methods Med Res 1996;5(3):263-81.
2. Lohr SL. Sampling: design and analysis. Belmont (CA): Duxbury Press; 1999.
3. Scheaffer RL, Mendenhall W 3rd, Lyman Ott R. Elementary survey sampling. 6th ed. Belmont (CA): Duxbury Press; 2006.
4. Park H, Kang G, Shin K, Oh Y, Lee C, Lee E, et al. A study on updating payment rates in the DRG based prospective payment system. Seoul, Korea: Seoul National University R&D Foundation; 2013.
5. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. San Francisco (CA): Morgan Kaufmann Publishers; 2011.
6. Hung SY, Yen DC, Wang HY. Applying data mining to telecom churn management. Expert Syst Appl 2006;31(3):515-24.
7. Shin H, Park H, Lee J, Jhee WC. A scoring model to detect abusive billing patterns in health insurance claims. Expert Syst Appl 2012;39(8):7441-50.
8. Ngai EW, Xiu L, Chau DC. Application of data mining techniques in customer relationship management: a literature review and classification. Expert Syst Appl 2009;36(2):2592-602.
9. Statistics Korea. Korean Statistical Information Service [Internet]. Daejeon, Korea: Statistics Korea; c2013 [cited at 2013 Sep 1]. Available from: http://kosis.kr/eng/database/database_001000.jsp?listid=A&subtitle=Population/Household#jsClick.
10. Jain AK. Data clustering: 50 years beyond k-means. Pattern Recognit Lett 2010;31(8):651-66.