CHARACTERISTICS OF SUPER-UTILIZERS OF CARE AT THE UNIVERSITY HOSPITALS OF GENEVA USING LATENT CLASS ANALYSIS

Gilles Cohen[1], Pascal Briot[2] and Pierre Chopard[2]

[1]Finance Directorate Geneva University Hospitals Geneva, Switzerland
[2]Quality and Patient Safety Division Geneva
University Hospitals Geneva, Switzerland

ABSTRACT

In hospitalized populations, there is significant heterogeneity in patient characteristics, disease severity, and treatment responses, which generally translates into very different related outcomes and costs. A better understanding of this heterogeneity can lead to better management, more effective and efficient treatments by personalizing care to better meet patients' profiles. Thus, identifying distinct clinical profiles among patients can lead to more homogenous subgroups of patients. Super-utilizers (SUs) are such a group, who contribute a substantial proportion of health care costs and utilize a disproportionately high share of health care resources. This study uses cost, utilization metrics and clinical information to segment the population of patients (N=32,759) admitted to the University Hospital of Geneva in 2019 and thus identifies the characteristics of its SUs group using Latent Class Analysis. This study shows how cluster analysis might be valuable to hospitals for identifying super-utilizers within their patient population and understanding their characteristics.

KEYWORDS

Latent Class Analysis, Clustering, Super-Utilizers, Inpatient Segmentation, Hospital Efficiency, Quality Improvement.

1. INTRODUCTION

The ongoing increase in healthcare expenditures [1] [2] and the introduction of new payment incentives which favor reductions in avoidable admissions and reoperations [3] [4][5] are forcing hospitals to develop new quality improvement strategies and improve their efficiencies. Since the greater share of hospital expenditure is often directed toward a limited number of patients commonly referred in the literature as super-utilizers (SUs)[6] [7] [8] [9], identifying these patients and designing better targeted interventions for them have the potential to increase appropriateness of care, improve outcomes and reduce costs. This study aims to stratify the population of patients admitted for more than 24 hours to the University Hospitals of Geneva and discharged between January 1, 2019 and December 31, 2019 applying cluster analysis on utilization data using demographics, admission and medical data.

The proposed approach uses Latent Class Analysis (LCA)to identify distinct patient clusters within our inpatient data. LCA is a model-based method that determines clusters of patients by common underlying unobserved characteristics. It is an iterative, maximum likelihood method that estimates how patterns in patient characteristics can be summarized into a finite number of

David C. Wyld et al. (Eds): MLDS, NECO, SEMIT, IBCOM, SPPR, SCAI, CSIA, ICCSEA - 2021
pp. 01-15, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111801

groups, or latent classes, by producing a probability distribution over the cluster allocation for each patient.LCA is convenient for analysis of categorical variables that are commonly found in clinical settings.

Clustering has been used to identify new disease subgroups in a diverse range of conditions, such as asthma, chronic lung disease (COPD), chronic heart failure (CHF) and neurological disorders. Nevertheless, the application of clustering to health care delivery is still emerging.

2. METHODS

2.1. Data and Variables

The Hôpitaux Universitaires de Genève (HUG) in Geneva is the largest academic medical center in Switzerland with approximately 2,000 acute care beds and 47,000 admissions per year. Located on 8 different sites, the hospital offers acute, intensive and long-term inpatient care, including pediatric and psychiatric care as well as rehabilitation and ambulatory care. All data for the study were collected from the HUG Enterprise Data Warehouse (EDW). The EDW contains information from several information systems including the patient administrative file (DPA - Dossier Patient Administratif), the clinical data repository which includes data from the HUG electronic medical record system (DPI - Dossier Patient Integré), the accounting costing system, and other operation tracking systems. Case costing at the HUG is determined using the standardized cost accounting model known by the acronym REKOLE developed by the Swiss Hospital Association (H+) [10]. It is based on real and normative costs which provide detailed information on the direct and indirect costs associated to each patient hospital stay. All costs are quoted in Swiss francs (CHF). From the EDW we used patient hospital utilization data. Detailed admission data were gathered from hospital discharge summaries comprising admission and discharge dates, admission and discharge disposition, length of stay (LOS), level of care provided (standard care or intensive care), category of services provided including surgical interventions, medications, tests, imaging and both primary and secondary diagnoses. The Elixhauser comorbidity index was calculated for every admission using the International Classification of Diseases, Version 10 [11] using a coding algorithm. DiagnosesICD-10 codes were matched with chapter headings. These data are gathered and coded systematically for each admission by coding service. Patients with missing data were excluded from the analysis.

Since we focused on high-cost patients according to the costs charged, we examined the distribution of health care costs in our data set representing all patients with a non-psychiatric inpatient admission discharged between January 1, 2019, and December 31, 2019. This rapid analysis confirms that the distribution of health care costs is highly concentrated on a small number of patients. In Figure 1, the population on the horizontal axis is segmented into deciles, starting from the decile of patients with the lowest costs consumption on the left to the decile of patients with the highest costs consumption on the right. The vertical axis shows the cumulative costs consumption for all patients. Thus, the 58.5% indicated above the 90% on the horizontal axis signifies that 90% of individuals (the least costly) accounted for only 58.5% of the total costs of the population. While the other 10% of the population generated 41.5% of the total costs. Therefore, the high-costs group was defined as the top 10 percentile of patients incurring the largest total (direct and indirect) admission costs.

The primary objective of this study was to characterize the high-costs users compared to the remaining 90% of patients according to patient characteristics, primary diagnoses, as well as their admission (emergency department) and discharge dispositions (e.g., home, acute care transfer, and long-term care transfer)and selected comorbidity score.

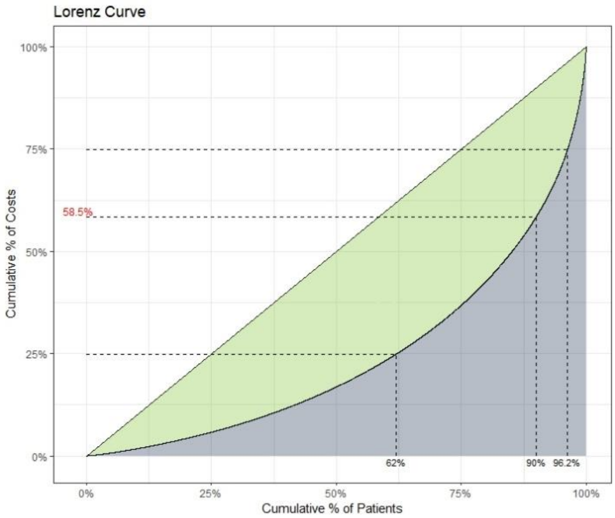

Figure 1. Lorenz curve for the total costs distribution. The diagonal line is the line of equality (for a perfectly equal distribution of costs per patient). Greater distance from the equality line indicates greater disparity in the distribution of total HUG costs.

2.2. Methodology

2.2.1. Latent Class Analysis

Let $\mathbf{X}$be the $N{\times}M$ data matrix, where each row $\mathbf{X}_n$is the realization of an M-dimensional vector of random dichotomous or polytomous variables $\mathbf{X}=(X_{n1},\ldots,X_{nM})$. Model based clustering assumes that each $\mathbf{X}_n$ comes from a finite mixture of G probability distribution in the exponential family, such as Bernoulli or Multinomial, each representing a different cluster, class or group. The general form of finite joint distribution of observed variables is as follow:

$$p(\mathbf{X}_n) = \sum_{g=1}^{G} \tau_g p(\mathbf{X}_n|\theta_g) \tag{1}$$

where the τ_g are the mixing probabilities and θ_g is the parameter set corresponding to component g. The component densities completely describe the cluster structure of the data and each observation belongs to the respective cluster in accordance with a set of unobserved cluster membership indicators $\mathbf{z}_n=(z_{n1},z_{n2,}\ldots,z_{nG})$ such that $z_{ng}=1$if $\mathbf{X}_n$ arises from the gth subpopulation.

When grouping multivariate categorical data, a prevalent model-based approach is the latent class analysis (LCA) model. In this setting, within each class, each variable $\mathbf{X}_m$ is modelled using a multinomial distribution, thus

$$p(X_m|\theta_g) = \prod_{c=1}^{C_m} \theta_{gmc}^{\mathbb{1}\{X_m=c\}}$$

where c=1,…,C_m are the possible categories values for variable m, θ_{gmc} is the probability of the variable taking value c given class g, and $\mathbb{1}\{X_m = c\}$ is the indicator function that is 1 if the variable takes value c, and 0 if not. In LCA, the variables are considered to be statistically independent given the class value of an observation. This is a primary assumption referred to as the local independence assumption [12].Transgressions of this assumption often cause the incompatibility of latent class models. The variables are then modelled for each variable within each group with a multinomial density giving the following factorization of the joint component density:

$$p(\mathbf{X}_n|\theta_g) = \prod_{m=1}^{M}\prod_{c=1}^{C_m}\theta_{gmc}^{\mathbb{1}\{X_{nm}=c\}}$$

accordingly the overall density in (1) turn into

$$p(\mathbf{X}_n) = \sum_{g=1}^{G}\tau_g\prod_{m=1}^{M}\prod_{c=1}^{C_m}\theta_{gmc}^{\mathbb{1}\{X_{nm}=c\}}$$

For a specified value G, the set of LCA model parameters is typically estimated by maximum likelihood by means of the Expectation-Maximization (EM) [13]. The algorithm is initialized with a random set of starting values. Therefore, it is usually recommended to run the procedure a bunch of times and then to pick the best solution [14]. More information about the model and parameter estimation can be found in [15][16] [17] and [14]. Concerning parameters interpretation, in the LCA model the parameter θ_{gmc} represents the probability of occurrence of attribute c for variable $\mathbf{X}_m$ in class g. Hence for the variables in the HUG dataset, θ_{gmc} will stand for the probability of having a certain criterion for each patient who belongs to the class g.

2.2.2. Model Selection

Various LCA models are being specified by the assignment of different values to G. For the purpose of selecting the optimal clustering model, various measures have been considered [18] and their performance were compared [19][20].Selecting the number of classes usually requires estimating models with incremental numbers of latent classes, and picking the number of classes based on the model that best fit the observed data. However, statistical criteria must always be assessed in combination with interpretability[21]. A class solution with better statistics is not of any use if it does not make any sense theoretically. Most current ways to decide the number of classes can be broken down into three categories: information-theoretic methods, likelihood ratio statistical test methods, and entropy-based criterion. Information criteria (ICs) are fitted indices that are frequently considered in a broad variety of statistical models and are used to make comparisons between a set of models. ICs consider model complexity into account and are also used to assess statistical fit. These indices comprise the Akaike Information Criterion (AIC) [22], the Consistent Akaike Information Criterion (cAIC)[23], the Bayesian Information Criterion (BIC) [24] and the adjusted Bayesian Information Criterion (aBIC)[25], where lower values denote a better fitting model. The AIC can be defined as:

$$AIC = -2LL + 2p,$$

where p is the number of free model parameters and LL the log-likelihood. The cAIC is a variant of the AIC but also punishes the value of -2 times the log-likelihood of the model for the number of free model parameters and sample size (Bozdogan, 1987). The cAIC is described as:

$$cAIC = -2LL + p[log(n) + 1],$$

where p is the number of free parameters and n is the sample size. The BIC also incorporates an adaptation for sample size and is given as follows:

$$BIC = -2LL + 2p\log(n),$$

where p is the number of free parameters and n is the sample size. Finally, aBIC is a by-product of BIC that decreases the penalty related to the sample size. The aBIC is defined as:

$$aBIC = -2LL + 2p\log[(n+2)/24],$$

where again p is the number of free parameters and n is the sample size.

The second type of methods for assessing model fit in latent class models involves likelihood ratio (LR) statistic tests. These tests compare the relative fit of two models that disagree in a set of parameter restrictions. For example, it compares a nested (n-1)-class solution to an (n)-class solution. The final category of the fit tests used to evaluate latent class models is the measure of entropy. The entropy index is based on the uncertainty of classification [26] [27]. Basically, the uncertainty of classification is evaluated at the individual level using the posterior probability; thus, entropy is a measure of the aggregated classification uncertainty. The uncertainty of classification is raised when the posterior probabilities are very close across classes. The normalized version of entropy, which scales to the interval [0, 1], is commonly used as a model selection criteria indicating the level of separation between classes. A higher value of normalized entropy represents a better fit; values > 0.80 indicate that the latent classes are highly discriminating [28].

3. RESULTS

3.1. LCA Results

A sequence of models was fitted to the data with the number of classes ranging from 1 to 12. The number of classes was determined by the evaluation of model fit indices (Table 1). Smaller values indicate better latent class separation except for entropy where values near 1 indicate better latent class separation. Regarding the relative goodness-of-fit indices, the value of BIC, aBIC, and cAIC continued to decline for the estimated models from the single-class model to the twelve-class model, while they reached a flattening from the five-class model onwards. However, there was no substantial improvement in either BIC, aBIC or cAIC fit beyond models with nine to twelve classes indicated by the elbow-shaped curve in Figure 2.Moreover, upon examination, the eight-class model appeared to have a meaningful interpretation. Therefore, based on a trade-off between several fitting indices, parsimony, and interpretability of the model, the eight-class model was retained as the final model.

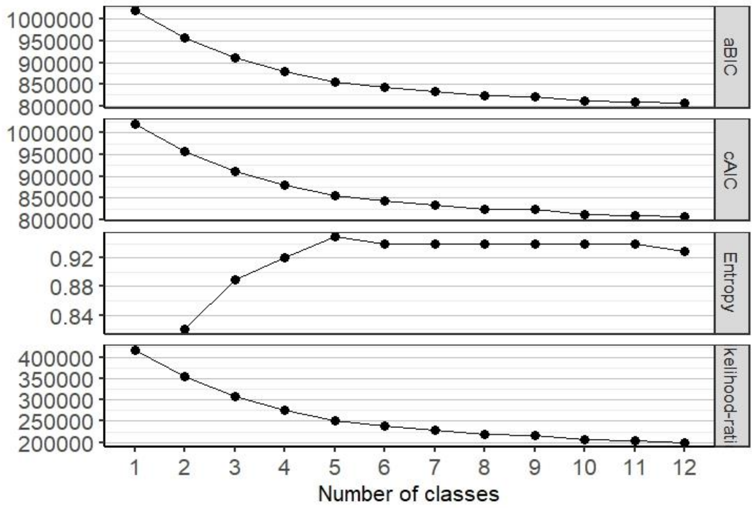

Figure 2. Plots showing goodness of fit with varying number of classes

Table 1. Fit Statistics for Latent Class Analyses

Number of latent class	**Number of parameters estimated**	**BIC**	**aBIC**	**cAIC**	**Entropy**	**likelihood-ratio**
1	32701	1020364.3	1020179.9	1020422.3	-	419028.1
2	32642	957342.3	956970.4	957459.3	0.82	355392.7
3	32583	910366.9	909807.6	910542.9	0.89	307804.0
4	32524	880006.5	879259.7	880241.5	0.92	276830.1
5	32465	855091.5	854157.2	855385.5	0.95	251301.7
6	32406	842763.4	841641.6	843116.4	0.94	238360.2
7	32347	834598.9	833289.6	835010.9	0.94	229582.2
8	32288	824737.8	823240.9	825208.8	0.94	219107.7
9	32229	818230.5	816546.2	818760.5	0.94	211987.0
10	32170	813011.7	811139.9	813600.7	0.94	206154.8
11	32111	810219.2	808159.9	810867.2	0.94	202748.9
12	32052	808030.0	805783.2	808737.0	0.93	199946.2

BIC: Bayesian information criterion; aBIC: adjusted Bayesian information criterion; cAIC: consistant Akaike information criterion

3.2. Results for the groups

3.2.1. Results for demographics and mode of admission and discharge from hospital

32,759 unique patients across 8 groups were identified by the clustering method. The number of patients per group ranges from 2,735 (8.4%) to 5,711 (17.4%) with an average of 4,095. Groups 6 and 8 have only single patients (N = 5,927; 18.1%) and group 3 has only women patients (N = 3,981; 12.2%) as described in table 2 below.

Table 2. Gender and status distribution of patients per groups

	Group 1 (N = 5,711)	Group 2 (N = 3,909)	Group 3 (N = 3,981)	Group 4 (N = 4,054)	Group 5 (N = 3,476)	Group 6 (N = 3,192)	Group 7 (N = 5,701)	Group 8 (N = 2,735)	All Groups (N = 32,759)
Men	2622 (45.9%)	1694 (43.3%)	0 (0%)	2236 (55.2%)	2084 (60.0%)	1586 (49.7%)	2774 (48.7%)	1589 (58.1%)	14585 (44.5%)
Women	3089 (54.1%)	2215 (56.7%)	3981 (100%)	1818 (44.8%)	1392 (40.0%)	1606 (50.3%)	2927 (51.3%)	1146 (41.9%)	18174 (55.5%)
Single	2706 (47.4%)	2504 (64.1%)	1570 (39.4%)	2495 (61.5%)	1889 (54.3%)	3192 (100%)	3214 (56.4%)	2735 (100%)	20305 (62.0%)
Couple	3005 (52.6%)	1405 (35.9%)	2411 (60.6%)	1559 (38.5%)	1587 (45.7%)	0 (0%)	2487 (43.6%)	0 (0%)	12454 (38.0%)

The patients' age showed a bimodal distribution with a first mode in the 0 to 18 age range (N = 6781; 20.7%) and the second mode in the 75 and above age range (7107; 21.7%). Groups 6 and 8 include mostly young patients less than 19 years of age (99.8% and 81.7% respectively). Group 3 includes nearly only young adult patients from age 19 to 44 (99.2%) while group 7 has a majority of older adults from age 75 and above (57.5%) and no young patients (less than 18 years old) as described in table 3 below.

Table 3. Age bracket distribution of patients per groups

Age bracket	Group 1 (N = 5,711)	Group 2 (N = 3,909)	Group 3 (N = 3,981)	Group 4 (N = 4,054)	Group 5 (N = 3,476)	Group 6 (N = 3,192)	Group 7 (N = 5,701)	Group 8 (N = 2,735)	All Groups (N = 32,759)
[0,18]	161 (2.8%)	373 (9.5%)	13 (0.3%)	566 (14.0%)	248 (7.1%)	3185 (99.8%)	0 (0%)	2235 (81.7%)	6781 (20.7%)
(18,34]	858 (15.0%)	540 (13.8%)	2687 (67.5%)	699 (17.2%)	109 (3.1%)	0 (0%)	182 (3.2%)	347 (12.7%)	5422 (16.6%)
(34,44]	1008 (17.7%)	333 (8.5%)	1260 (31.7%)	503 (12.4%)	145 (4.2%)	4 (0.1%)	285 (5.0%)	78 (2.9%)	3616 (11.0%)
(44,54]	1152 (20.2%)	371 (9.5%)	21 (0.5%)	663 (16.4%)	373 (10.7%)	3 (0.1%)	451 (7.9%)	67 (2.4%)	3101 (9.5%)
(54,64]	1107 (19.4%)	407 (10.4%)	0 (0%)	704 (17.4%)	600 (17.3%)	0 (0%)	648 (11.4%)	2 (0.1%)	3468 (10.6%)
(64,74]	766 (13.4%)	409 (10.5%)	0 (0%)	476 (11.7%)	750 (21.6%)	0 (0%)	857 (15.0%)	6 (0.2%)	3264 (10.0%)
(74,150]	659 (11.5%)	1476 (37.8%)	0 (0%)	443 (10.9%)	1251 (36.0%)	0 (0%)	3278 (57.5%)	0 (0%)	7107 (21.7%)

Admissions to the HUG were done majorly via the emergency department (ED) for all the groups (55.7%) with groups 2 and 7 at 92.5% and 93.3% respectively. Group 6 was the exceptionwith only 50 patients out of 3,142 (1.6%) admitted via the ED. On the average 78.3% of all patients (N = 25,654) were discharged to home with the exception of group 5 with only 49.5% discharged to home (N = 1719). Groups 5 and 7 had the most patients transferred to rehabilitation with 32.1% and 23.4% respectively; while groups 1, 3,6 and 8 had the least with 0.4%, 0.1%, 0.0% and 0.8% respectively. These results are tabulated in table 4 below.

Table 4. Mode of admission and discharge from hospital for patients per groups

	Group 1 (N = 5,711)	Group 2 (N = 3,909)	Group 3 (N = 3,981)	Group 4 (N = 4,054)	Group 5 (N = 3,476)	Group 6 (N = 3,192)	Group 7 (N = 5,701)	Group 8 (N = 2,735)	All Groups (N = 32,759)
ED	1766 (30.9%)	3617 (92.5%)	2686 (67.5%)	1313 (32.4%)	2212 (63.6%)	50 (1.6%)	5319 (93.3%)	1299 (47.5%)	18262 (55.7%)
Not ED	3945 (69.1%)	292 (7.5%)	1295 (32.5%)	2741 (67.6%)	1264 (36.4%)	3142 (98.4%)	382 (6.7%)	1436 (52.5%)	14497 (44.3%)
Home	5547 (97.1%)	2185 (55.9%)	3947 (99.1%)	3510 (86.6%)	1719 (49.5%)	3160 (99.0%)	3037 (53.3%)	2549 (93.2%)	25654 (78.3%)
Rehab	21 (0.4%)	669 (17.1%)	2 (0.1%)	414 (10.2%)	1116 (32.1%)	1 (0.0%)	1336 (23.4%)	22 (0.8%)	3581 (10.9%)
Others	143 (2.5%)	1055 (27.0%)	32 (0.8%)	130 (3.2%)	641 (18.4%)	31 (1.0%)	1328 (23.3%)	164 (6.0%)	3524 (10.8%)

3.2.2. Results for diagnoses, procedures and Elixhauser index

Groups 1 and 4 show a range of precisely targeted procedures (such as obstetric technics and operations on musculoskeletal system) and primary diagnoses (such as diseases of the digestive system) while groups 2 and 6 show no procedures done in 2019. In addition, group 6 shows a majority (61.1%) of diagnoses related to factors influencing the health status and reasons to access the health system.

35.4% of group 1 patients received operations of the digestive system with 30.9% of patients diagnosed with a disease of the digestive system. Of all patients with operations of the digestive systems (N = 3003), group 1 includes 67.4% patients (N = 2024) and of all patients with a primary diagnosis of disease of the digestive system (N = 2774), group 1 includes 63.6% patients (N = 1763).

90.3% of group 4 patients received operations of the musculoskeletal system with 49.5% of patients diagnosed with a disease of the musculoskeletal system and 47.1% with traumatic lesions. Of all patients with operations of the musculoskeletal system (N = 4473) group 4 includes 81.8% patients (N = 3660) and of all patients with a primary diagnosis of disease of the musculoskeletal system or traumatic lesions (N = 6407), group 4 includes 61.1% patients (N = 3917).94.0% of the patients (N = 3,742) in group 3 (women only patients) received obstetric procedures.

These results are summarized in tables 5 and 6 below.

Table 5. Distribution of procedure categories for patients by groups

Procedure categories	Group 1 (N = 5,711)	Group 2 (N = 3,909)	Group 3 (N = 3,981)	Group 4 (N = 4,054)	Group 5 (N = 3,476)	Group 6 (N = 3,192)	Group 7 (N = 5,701)	Group 8 (N = 2,735)	All Groups (N = 32,759)
Operations on the nervous system	196 (3.4%)	0 (0%)	10 (0.3%)	111 (2.7%)	177 (5.1%)	0 (0%)	116 (2.0%)	83 (3.0%)	693 (2.1%)
Operations on the urinary system	440 (7.7%)	0 (0%)	0 (0%)	1 (0.0%)	129 (3.7%)	0 (0%)	195 (3.4%)	37 (1.4%)	802 (2.4%)
Operations on male genital organs	314 (5.5%)	0 (0%)	0 (0%)	0 (0%)	15 (0.4%)	0 (0%)	0 (0%)	47 (1.7%)	376 (1.1%)
Operations on female genital organs	688 (12.0%)	0 (0%)	198 (5.0%)	0 (0%)	25 (0.7%)	0 (0%)	0 (0%)	2 (0.1%)	913 (2.8%)
Obstetric techniques	0 (0%)	0 (0%)	3742 (94.0%)	0 (0%)	1 (0.0%)	0 (0%)	0 (0%)	0 (0%)	3743 (11.4%)
Operations on musculoskeletal system	0 (0%)	0 (0%)	0 (0%)	3660 (90.3%)	504 (14.5%)	0 (0%)	294 (5.2%)	15 (0.5%)	4473 (13.7%)
Operations on integumentary system	424 (7.4%)	0 (0%)	1 (0.0%)	204 (5.0%)	73 (2.1%)	0 (0%)	107 (1.9%)	29 (1.1%)	838 (2.6%)
Diagnostic and therapeutic techniques	572 (10.0%)	0 (0%)	24 (0.6%)	22 (0.5%)	828 (23.8%)	0 (0%)	4158 (72.9%)	1841 (67.3%)	7445 (22.7%)
Operations of the nose, mouth and pharynx	317 (5.6%)	0 (0%)	0 (0%)	0 (0%)	23 (0.7%)	0 (0%)	8 (0.1%)	370 (13.5%)	718 (2.2%)
Operations of respiratory system	147 (2.6%)	0 (0%)	0 (0%)	10 (0.2%)	127 (3.7%)	0 (0%)	107 (1.9%)	49 (1.8%)	440 (1.3%)
Operations of cardiovascular system	195 (3.4%)	0 (0%)	3 (0.1%)	41 (1.0%)	672 (19.3%)	0 (0%)	228 (4.0%)	103 (3.8%)	1242 (3.8%)
Operations of digestive system	2024 (35.4%)	0 (0%)	2 (0.1%)	0 (0%)	736 (21.2%)	0 (0%)	142 (2.5%)	99 (3.6%)	3003 (9.2%)
Other classified procedures	377 (6.6%)	0 (0%)	0 (0%)	5 (0.1%)	103 (3.0%)	0 (0%)	0 (0%)	57 (2.1%)	542 (1.7%)
Procedures non classified elsewhere	17 (0.3%)	0 (0%)	1 (0.0%)	0 (0%)	62 (1.8%)	0 (0%)	346 (6.1%)	3 (0.1%)	429 (1.3%)
No procedure	0 (0%)	3909 (100%)	0 (0%)	0 (0%)	1 (0.0%)	3192 (100%)	0 (0%)	0 (0%)	7102 (21.7%)

Table 6. Distribution of diagnosis categories for patients by groups

	Group 1 (N = 5,711)	Group 2 (N = 3,909)	Group 3 (N = 3,981)	Group 4 (N = 4,054)	Group 5 (N = 3,476)	Group 6 (N = 3,192)	Group 7 (N = 5,701)	Group 8 (N = 2,735)	All Groups (N = 32,759)
Certain infectious and parasitic diseases	34 (0.6%)	117 (3.0%)	0 (0%)	0 (0%)	202 (5.8%)	2 (0.1%)	255 (4.5%)	77 (2.8%)	687 (2.1%)
Tumors	1426 (25.0%)	42 (1.1%)	0 (0%)	29 (0.7%)	775 (22.3%)	2 (0.1%)	195 (3.4%)	0 (0%)	2469 (7.5%)
Diseases of the blood, hematopoietic organs, immunity system	22 (0.4%)	23 (0.6%)	0 (0%)	0 (0%)	29 (0.8%)	0 (0%)	91 (1.6%)	27 (1.0%)	192 (0.6%)
Endocrinien, metabolic and nutritionel diseases	274 (4.8%)	103 (2.6%)	0 (0%)	0 (0%)	98 (2.8%)	6 (0.2%)	154 (2.7%)	35 (1.3%)	670 (2.0%)
Diseases of the circulatory system	208 (3.6%)	410 (10.5%)	0 (0%)	0 (0%)	771 (22.2%)	0 (0%)	1820 (31.9%)	32 (1.2%)	3241 (9.9%)
Mental and behavior diseases	4 (0.1%)	409 (10.5%)	0 (0%)	0 (0%)	15 (0.4%)	0 (0%)	177 (3.1%)	41 (1.5%)	646 (2.0%)
Diseases of the nervous system	231 (4.0%)	122 (3.1%)	0 (0%)	9 (0.2%)	85 (2.4%)	0 (0%)	345 (6.1%)	119 (4.4%)	911 (2.8%)
Diseases of the eyes	95 (1.7%)	69 (1.8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	20 (0.4%)	50 (1.8%)	234 (0.7%)
Diseases of the respiratory system	235 (4.1%)	495 (12.7%)	0 (0%)	0 (0%)	130 (3.7%)	2 (0.1%)	909 (15.9%)	808 (29.5%)	2579 (7.9%)
Diseases of the digestive system	1763 (30.9%)	286 (7.3%)	0 (0%)	20 (0.5%)	421 (12.1%)	4 (0.1%)	193 (3.4%)	87 (3.2%)	2774 (8.5%)
Diseases of the skin and subcutaneous tissue	117 (2.0%)	84 (2.1%)	0 (0%)	28 (0.7%)	36 (1.0%)	1 (0.0%)	68 (1.2%)	36 (1.3%)	370 (1.1%)
Diseases of the musculoskeletal system	0 (0%)	214 (5.5%)	0 (0%)	2006 (49.5%)	157 (4.5%)	1 (0.0%)	170 (3.0%)	38 (1.4%)	2586 (7.9%)
Diseases of the urinary track system	970 (17.0%)	237 (6.1%)	0 (0%)	1 (0.0%)	107 (3.1%)	6 (0.2%)	214 (3.8%)	66 (2.4%)	1601 (4.9%)
Traumatic lesions, poisoning and other external cause of illness	81 (1.4%)	609 (15.6%)	0 (0%)	1911 (47.1%)	461 (13.3%)	1 (0.0%)	619 (10.9%)	139 (5.1%)	3821 (11.7%)
Pregnancy and delivery	0 (0%)	101 (2.6%)	3981 (100%)	0 (0%)	4 (0.1%)	1 (0.0%)	1 (0.0%)	0 (0%)	4088 (12.5%)
Perinatal related illness	0 (0%)	1 (0.0%)	0 (0%)	0 (0%)	12 (0.3%)	1037 (32.5%)	0 (0%)	782 (28.6%)	1832 (5.6%)
Genetic malformations and chromosomic abnormalities	26 (0.5%)	1 (0.0%)	0 (0%)	49 (1.2%)	110 (3.2%)	154 (4.8%)	0 (0%)	212 (7.8%)	552 (1.7%)
Abnormal results from exams and labs non classified elsewhere	92 (1.6%)	526 (13.5%)	0 (0%)	1 (0.0%)	53 (1.5%)	26 (0.8%)	467 (8.2%)	118 (4.3%)	1283 (3.9%)
Factors influencing health status and reasons to access health system	133 (2.3%)	60 (1.5%)	0 (0%)	0 (0%)	10 (0.3%)	1949 (61.1%)	3 (0.1%)	68 (2.5%)	2223 (6.8%)

The Elixhauser comorbidity index was calculated for each patient based on their diagnosis codes. The distribution per group for chronic heart failure (CHF), cardiovascular disease (CARIT), chronic obstructive pulmonary disease (COP), and diabetes (DIABC) do not show any significance difference across the groups. The proportion of patients across the groups exhibiting each conditions are very homogeneous as described in table 7 below.

Table 7. Distribution Elixhauser comorbidity index for selected conditions for patients by groups

	Group 1 (N = 5,711)	Group 2 (N = 3,909)	Group 3 (N = 3,981)	Group 4 (N = 4,054)	Group 5 (N = 3,476)	Group 6 (N = 3,192)	Group 7 (N = 5,701)	Group 8 (N = 2,735)	All Groups (N = 32,759)
CHF									
0	5306 (92.9%)	3632 (92.9%)	3744 (94.0%)	3794 (93.6%)	3260 (93.8%)	2982 (93.4%)	5340 (93.7%)	2574 (94.1%)	30632 (93.5%)
1	405 (7.1%)	277 (7.1%)	237 (6.0%)	260 (6.4%)	216 (6.2%)	210 (6.6%)	361 (6.3%)	161 (5.9%)	2127 (6.5%)
CARIT									
0	5062 (88.6%)	3432 (87.8%)	3545 (89.0%)	3583 (88.4%)	3080 (88.6%)	2827 (88.6%)	5059 (88.7%)	2421 (88.5%)	29009 (88.6%)
1	649 (11.4%)	477 (12.2%)	436 (11.0%)	471 (11.6%)	396 (11.4%)	365 (11.4%)	642 (11.3%)	314 (11.5%)	3750 (11.4%)
COPD									
0	5412 (94.8%)	3716 (95.1%)	3801 (95.5%)	3866 (95.4%)	3313 (95.3%)	3049 (95.5%)	5435 (95.3%)	2589 (94.7%)	31181 (95.2%)
1	299 (5.2%)	193 (4.9%)	180 (4.5%)	188 (4.6%)	163 (4.7%)	143 (4.5%)	266 (4.7%)	146 (5.3%)	1578 (4.8%)
DIABC									
0	5143 (90.1%)	3549 (90.8%)	3589 (90.2%)	3663 (90.4%)	3192 (91.8%)	2898 (90.8%)	5182 (90.9%)	2497 (91.3%)	29713 (90.7%)
1	568 (9.9%)	360 (9.2%)	392 (9.8%)	391 (9.6%)	284 (8.2%)	294 (9.2%)	519 (9.1%)	238 (8.7%)	3046 (9.3%)

3.2.3. Results for top 10 percentile of cost and clinical outcomes

Group 5 (N = 3,476) had 80.5% of its patients in the top 10 percentile for total costs compared to all the other groups combined with 3.0% of their patients (N = 883).

Group 5 patients had the most number of patients with more than 10 ambulatory visits (42.9%), more than 10 different diagnoses (69.9%), more than 3 procedures (90.5%), more than 10 lab tests (80.2%), more than 10 medications (96.3%), and more than 2 hospitalizations (23.9%).

Group 5 had also the most number of patients who were discharged to rehabilitation facilities after their hospital stay (32.1%).

More group 5 patients were 65 years and older (N = 2,001; 57.6%) than any other groups except group 7 (N = 4,135; 72.5%). While group 7 had more patients 65 years and older than group 5, it also had no patient less than 19 years of age while group 5 had 248 patients (7.1%).

Group 7 provides some other results which are noteworthy. After group 5, it has the most number of patients (N = 371; 6.5%) in the top 10 percentile of costs; with more than 10 diagnoses (N = 2,028; 35.6%); with more than 10 tests (N = 2,649; 46.5%); and with more than 10 medications (N = 4,628; 81.2%).

These results are tabulated in table 8 below.

Table 8. Distribution of costs percentile and clinical outcomes for patients per groups

	Group 1 (N = 5,711)	Group 2 (N = 3,909)	Group 3 (N = 3,981)	Group 4 (N = 4,054)	Group 5 (N = 3,476)	Group 6 (N = 3,192)	Group 7 (N = 5,701)	Group 8 (N = 2,735)	All Groups (N = 32,759)
Percentile distribution of costs									
Top 10th percentile	133 (2.3%)	32 (0.8%)	13 (0.3%)	157 (3.9%)	2797 (80.5%)	11 (0.3%)	371 (6.5%)	166 (6.1%)	3680 (11.2%)
Bottom 90th percentile	5578 (97.7%)	3877 (99.2%)	3968 (997%)	3897 (96.1%)	679 (19.5%)	3181 (99.7%)	5330 (93.5%)	2569 (93.9%)	29079 (88.8%)
Ambulatory visits									
0 - 4	2320 (40.6%)	2660 (68.0%)	1998 (50.2%)	1073 (26.5%)	1075 (30.9%)	3076 (96.4%)	3828 (67.1%)	2115 (77.3%)	18145 (55.4%)
5 - 10	1758 (30.8%)	647 (16.6%)	1294 (32.5%)	1907 (47.0%)	910 (26.2%)	95 (3.0%)	1085 (19.0%)	407 (14.9%)	8103 (24.7%)
> 10	1633 (28.6%)	602 (15.4%)	689 (17.3%)	1074 (26.5%)	1491 (42.9%)	21 (0.7%)	788 (13.8%)	213 (7.8%)	6511 (19.9%)
Hospital admissions									
1	4919 (86.1%)	3359 (85.9%)	3663 (92.0%)	3695 (91.1%)	1623 (46.7%)	3117 (97.7%)	4495 (78.8%)	2362 (86.4%)	27233 (83.1%)
2	619 (10.8%)	420 (10.7%)	273 (6.9%)	311 (7.7%)	1021 (29.4%)	71 (2.2%)	860 (15.1%)	333 (12.2%)	3908 (11.9%)
> 2	173 (3.0%)	130 (3.3%)	45 (1.1%)	48 (1.2%)	832 (23.9%)	4 (0.1%)	346 (6.1%)	40 (1.5%)	1618 (4.9%)
Number of diagnoses									
1	1433 (25.1%)	600 (15.3%)	10 (0.3%)	553 (13.6%)	3 (0.1%)	1680 (52.6%)	38 (0.7%)	547 (20.0%)	4864 (14.8%)
2 - 10	4210 (73.7%)	2773 (70.9%)	3804 (95.6%)	3489 (86.1%)	1043 (30.0%)	1512 (47.4%)	3635 (63.8%)	2124 (77.7%)	22590 (69.0%)
> 10	68 (1.2%)	536 (13.7%)	167 (4.2%)	12 (0.3%)	2430 (69.9%)	0 (0%)	2028 (35.6%)	64 (2.3%)	5305 (16.2%)
Number of treatments									
0	0 (0%)	3908 (100.0%)	0 (0%)	0 (0%)	0 (0%)	3192 (100%)	0 (0%)	0 (0%)	7100 (21.7%)
1 - 2	4059 (71.1%)	0 (0%)	2665 (66.9%)	3194 (78.8%)	329 (9.5%)	0 (0%)	4499 (78.9%)	2151 (78.6%)	16897 (51.6%)
> 3	1652 (28.9%)	1 (0.0%)	1316 (33.1%)	860 (21.2%)	3147 (90.5%)	0 (0%)	1202 (21.1%)	584 (21.4%)	8762 (26.7%)
Number of labs (Tests)									
1 - 10	5241 (91.8%)	3275 (83.8%)	3884 (97.6%)	3948 (97.4%)	689 (19.8%)	3181 (99.7%)	3052 (53.5%)	2639 (96.5%)	25909 (79.1%)
> 10	470 (8.2%)	634 (16.2%)	97 (2.4%)	106 (2.6%)	2787 (80.2%)	11 (0.3%)	2649 (46.5%)	96 (3.5%)	6850 (20.9%)
Number of medications									
1 - 10	2526 (44.2%)	2100 (53.7%)	2859 (71.8%)	2294 (56.6%)	130 (3.7%)	3192 (100%)	1073 (18.8%)	2315 (84.6%)	16489 (50.3%)
> 10	3185 (55.8%)	1809 (46.3%)	1122 (28.2%)	1760 (43.4%)	3346 (96.3%)	0 (0%)	4628 (81.2%)	420 (15.4%)	16270 (49.7%)

4. DISCUSSION

This study was conducted to determine how cluster analysis using the SU criterion used in the literature and by Grafe et al [29]might applied to the inpatient population of the Hôpitaux Universitaires de Genève. The results show that the LCA clustering model is able to generate 8 groups with distinctive characteristics. In particular, the algorithm was able to identify a group with mostly patients less than 19 years of age who use the hospital for health related factors but not serious illness as well as a group with only women who use the hospital for only women related procedures and diagnoses and two other groups whose patients are greater utilizers of digestive and musculoskeletal procedures with consistent related diagnoses. Across and among the groups the results for the variables studied appear highly coherent and as would be expected demonstrating that the clustering algorithm appears robust in stratifying the population of patients admitted to the Hôpitaux Universitaires de Genève in 2019.

Most important among the 8 groups was the group of patients whose costs are in the top 10^{th} percentile (Group 5) and for whom the use of ambulatory and inpatient services is the greatest as well as the use of treatments, test (labs) and medications. Given the consistency of the results for these patients and the coherence we observed across the other groups (described above), we are confident that group 5 represent the super-utilizers of care for the HUG in 2019.

In this investigation, we demonstrated the use of cluster analysis to identify distinct subgroups of patients with specific combinations of co-occurring conditions in a large academic medical center.

The model revealed the expected segmentation by age brackets and gender such as with group 6 (patients less than 19 years of age) and group 3 (women only patients) along with the expected utilisation of care services such as pregnancy and delivery for group 3. The identification of these expected groups in our analysis provide assurance of the validity of our data mining method.

The cluster analysis provided also a data driven approach to identifying at least 3very distinctive clinically relevant groups of patients with patterns of care utilization that could be targeted with new, enhanced care management strategies. The super-utilizers (group 5, N = 3476, 10.6% of all patients) including mostly SUs (N = 2797, 8.5% of all patients). The patients who consistently (93.3% of patients) are admitted via the ED (group 7, N = 5701, 17.4% of all patients) including at least some SUs (N = 371, 1.0% of all patients). The musculoskeletal patients (group 4, N = 4054, 12.4% of all patients) whose care and costs are mostly related to problems associated with the musculoskeletal system including at least some SUs (N = 157, 0.5% of all patients).Together these 3 groups (N = 12947, 39.5% of all patients) which alone contain the majority of SUs (N = 3325, 10% of all patients), and considering only the SUs, account for all costs above the 90% percentile which means that targeted intervention to improve the care of these patients will have the most impact on total costs for the HUG.

While the model appears coherent and robust to further assess the stability of these clusters over time, analyses should be conducted on cohorts from different years. A larger population of patients (over multiple years) might also provide more power to detect significant difference in the Elixhauser comorbidity index across groups.

Like any investigation, the characteristics of our clusters are constrained to our data and setting. Reproducing these analyses in different settings and different patient populations may potentially yield different clusters. However, these differences would and should nevertheless inform on different management strategies specific to populations in those settings.

In this study we showed how cluster analysis can be used to identify homogeneous groups of complex patients from a large heterogeneous population. Such data science methods demonstrate that it is possible to use the conceptual findings of this investigation to raise awareness of the need for a more personalized approach of care management services for patients with high levels of healthcare utilization (super utilizers). However, further understanding of the care management needs of clusters of patients with similar comorbidities and care utilization is warranted before designing specific tailored interventions.

5. CONCLUSIONS

This study identified SU criterion that have commonly been used in the literature and applied these criterion to the inpatient population of a large academic medical center. The procedures and results reported illustrate how cluster analysis can be helpful in differentiating homogeneous groups of complex patients from a large heterogeneous population. These results should help in the application of more targeted interventions per subgroups to improve appropriateness of care, improve outcomes and reduce costs.

ACKNOWLEDGEMENTS

The medical directorate, the service for quality and patient safety, and the department of finance of the Hôpitaux Universitaires de Genève provided useful insights into the broad focus and goals of this study. We appreciate their inputs, especially from our colleagues in the medico-economic unit, which were most helpful and constructive for the analysis and writing of this paper. This study will not have been possible without the assistance and collaboration of all these people.

REFERENCES

[1] Spending on health: Latest trends. Health Spending Latest Trend Brief.2018. [En ligne]. Available: https://www.oecd.org/health/health-systems/Health-Spending-Latest-Trends-Brief.pdf.
[2] R. W. Raghupathi. Healthcare Expenditure and Economic Performance: Insights From the United States Data. Front Public Health, vol. 8, 2020.
[3] R. Axon, M. Williams. Hospital readmission as an accountability measure. JAMA, vol. 305, pp. 504-5, 2011.
[4] H. A. Purdey. Predicting and preventing avoidable hospital admissions: a review. The journal of the Royal College of Physicians of Edinburgh, vol. 43, pp. 340-4, 2013.
[5] C. Schwierz. Cost-Containment Policies in Hospital Expenditure in the European Union. 2016.
[6] M. Berk, A. Monheit. The concentration of health care expenditures, revisited. Health Affairs (Project Hope)., vol. 20, pp. 9-18, 2001.
[7] S. Rais, A. Nazerian, S. Ardal, Y. Chechulin, N. Bains, K. Malikov, A. Nazerian. High-cost users of Ontario's healthcare services.Healthc Policy., vol. 9, pp. 44-51, 2013.
[8] W. P. Wodchis, P. C. Austin, D. A. Henry. A 3-year study of high-cost users of health care.CMAJ : Canadian Medical Association journal., vol. 188, pp. 182-188, 2016.
[9] S. B. Cohen. The Concentration of Health Care Expenditures and Related Expenses for Costly Medical Conditions, 2012.Statistical Brief (Medical Expenditure Panel Survey (US)), Agency for Healthcare Research and Quality (US), 2014.
[10] P. Besson. Manuel Rekole® Comptabilité Analytique A L'hôpital. Les Hôpitaux de Suisse, 2013.
[11] H. Organization. ICD-10 Classification of Mental and Behavioral Disorders (The): Diagnostic Criteria for Research. World Health Organization, 1993.
[12] C. C. Clogg. Latent Class Models for Measuring. Latent Trait and Latent Class Models, pp. 173-205, 1988.
[13] A. Dempster, N. Laird, D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society, series B, vol. 39, pp. 1-38, 1977.
[14] D. J. Bartholomew, K. Martin, I. Moustaki. Latent Variable Models and Factor Analysis: A Unified Approach. Wiley Series in Probability and Statistics, 2011.
[15] P. F. Lazarsfeld, N. W. Henry. Latent Structure Analysis, Houghton, Mifflin, 1968.
[16] C. C. Clogg. Latent class models, chez Handbook of statistical modeling for the social and behavioral sciences, 1995.
[17] A. Agresti. Categorical data analysis, New York: Wiley, 2002.
[18] G. McLachlan, D. Peel. Finite mixture models, New York: Wiley, 2000.
[19] J. G. Dias. Latent Class Analysis and Model Selection, chez From Data and Information Analysis to Knowledge Engineering, 2006.
[20] C. Biernacki. Model selection theory and considerations in large scale scenarios, chez Research Summer School on Statistics for Data Science - S4D, Caen France, 2018.

[21] B. O. Muthén, L. K. Muthén. Integrating person-centered and variablecentered analyses: Growth mixture modeling with latent trajectory classes.Alcoholism: Clinical & Experimental Research, vol. 24, pp. 882-891, 2000.

[22] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, vol. 19, pp. 716-723, 1974.

[23] H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika, vol. 52, pp. 345-370, 1987.

[24] G. Schwarz. Estimating the dimension of a model. Annals of Statistics, vol. 6, pp. 461-464, 1978.

[25] S. L. Sclove. Application of model-selection criteria to some problems in multivariate. Psychometrika, vol. 52, pp. 333-343, 1987.

[26] M.-C. Wang, Q. Deng, X. Bi, H. Ye, W. Yang. Performance of the entropy as an index of classification accuracy in latent profile analysis: A Monte Carlo simulation study. Acta Psychologica Sinica, vol. 49, pp. 1473-1482, 2017.

[27] G. Celeux, G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification, vol. 13, pp. 195-212, 1996.

[28] C. Larose, O. Harel, K. Kordas, D. Dey. Latent Class Analysis of Incomplete Data via an Entropy-Based Criterion. Stat Methodology, vol. 32, pp. 107-121, 2016.

[29] Grafe, Carl J et al. "How to Classify Super-Utilizers: A Methodological Review ofSuper-Utilizer Criteria Applied to the Utah Medicaid Population, 2016-2017.Population health management vol. 23,2 (2020): 165-173.