MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190203

Characterizing Frequent Flyers of an Emergency Department Using Cluster Analysis

Emile Ramez Shehada[a], Lu He[a], Elizabeth V. Eikey, PhD[a], Maxwell Jen, MD[b], Andrew Wong, MD[c], Sean D. Young, PhD[d,e], Kai Zheng, PhD[a]

[a] *Department of Informatics, University of California, Irvine, Irvine, CA, USA*
[b] *Department of Emergency Medicine University of California, Irvine, Irvine, CA, USA*
[c] *Emergency Department, University of California, Davis, Davis, CA, USA*
[d] *Department of Family Medicine, University of California, Los Angeles, Los Angeles, CA, USA*
[e] *University of California Institute for Prediction Technology, Los Angeles, CA, USA*

Abstract

Emergency department (ED) overcrowding has been a pain point in hospitals across the globe. "Frequent flyers," who visited the ED at a much higher rate than average, account for almost one third of ED visits even though they represent only a small proportion of all ED patients. In this study, we used data-mining methods to cluster ED frequent flyers at a large academic medical center in the US. The objective was to identify distinct types of frequent flyers, and the common characteristics associated with each type. The results show that the frequent flyers at the ED have three subgroups each exhibiting distinct characteristics: (1) the elderly with chronic health conditions, (2) middle-aged males with unhealthy behavior, and (3) adult females who are generally healthy. These findings may inform targeted interventional strategies for patients of each subgroup, who likely have distinct reasons for visiting the ED frequently, to reduce ED overcrowding.

Keywords:

Data Mining; Hospital Emergency Service; Cluster Analysis

Introduction

ED overcrowding has long been an issue in hospitals across the world [1–3]. Frequent flyers, which are generally defined as patients who visit the ED four or more times per year [4,5], represent as little as 4.5% to 8% of all ED patients but account for as much as 28% of all ED visits [6]. Accordingly, hospitals have a vested interest in characterizing their populations of ED in order to determine how to reduce their impact on ED operations.

Researchers in the field of emergency medicine have written extensively about the typical characteristics of ED frequent flyers. However, there is still too much variation; frequent flyers simply are not a homogenous group. Within the literature, there is a general consensus that three factors are associated with the number of times a frequent flyer visits the ED per year: mental illness, substance abuse and dependence, and alcohol abuse [7–9]. Additionally, employment status and government insurance are often correlated with high ED use [6,10,11]. For instance, younger age groups were also implicated in ED overuse due to the characteristics of the neighborhood surrounding the hospital or the type of hospital from which the data was acquired [6,7]. In some cases, researchers observed a bimodal age distribution with increased risks in frequents users who were either 25 to 44 years old or over 64 years old [7]. Similarly, research has generally found that males and minorities tend to use the ED more frequently, but findings are sometimes inconsistent. For example, a 2005 study conducted by Blank et al. concluded that the sex and race composition of their ED frequent flyer population did not differ substantially from that of their general ED population, which was 51% female and 57% white [10]. Another study conducted by Milbrett and Halm indicated that the frequent flyer population at their "large Midwestern urban hospital" was "commonly" female, middle-aged and white [11].

Thus, there is no "one-size-fits-all" list of characteristics that apply to all hospitals. In order to support hospital-level decision making, a different approach to uncovering these subgroups is necessary. There is a need to identify hospital-specific subgroups and characteristics of frequent flyers. Clustering, or the process of organizing objects or measurements into groups, is a promising method. It differs from mere classification in that the groups used in classification are pre-defined groups, whereas the groups exposed by clustering do not exist beforehand [12]. As an automated unsupervised method, it is also less labor-intensive than manually identifying subgroups in a population and therefore is realistic to be implemented in real-life settings. In previous work, clustering techniques have been used on ED data to both predict patient outcomes or characterize ED usage. In 2007, Huang et al. used the K-means algorithm to cluster patients according to medical utilization, discovering that their population of frequent flyers also tended to utilize other medical services more often [13]. In a similar vein as Huang, Hastings et al. used latent cluster analysis (LCA) to reveal medical utilization clusters among elderly ED patients [14].

In this paper, we performed cluster analysis on 1748 unique patients who have visited the ED more than 4 times in a year at an UC medical center to identify characteristics of subgroups. We applied two commonly used cluster methods that are able to handle mixed data types: the K-prototypes algorithm and Partition-Around-Medoid (PAM). Finally, we discuss the three subgroups of frequent flyers that emerged from our analysis and the implications for using subgroup level interventions to reduce ED misuse.

Methods

Data description and processing

The data received from UCI Medical Center consisted of records of 1748 unique patients. Some data was derived from other data elements; for instance, blood pressure and height were used to yield a computed hypertension, which was cross-referenced against patients' medical history. Computed hypertension is used in the final analysis in lieu of both blood pressure and height. Patient medical history was also used to determine whether the patient had any history of substance abuse, mental illness, or cancer. The data contains both categorical attributes and continuous attributes. Categorical attributes include: hypertension, Diabetes, Binge drinking, tobacco, gender, employment, insurance, substance abuse, other mental illness, and cancer. Continuous attributes include: age, Body Mass Index (BMI), average emergency severity index (ESI), heart rate (HR), blood pressure (RR), temperature (TempC) and oxygen saturation (SpO2). Table 1 provides descriptive statistics of our dataset. Decimals in the table indicate percentages for categorical variables.

Table 1-Descriptive statistics

Variables	Mean (SD)
Demographic	
Gender (Male=1)	.52
Age	46 (18)
Race (Hispanic)	.47
Race (White)	.36
Race (Black)	.05
Race (Asian)	.08
Race (Others)	.04
Medical history	
Cancer (Yes=1)	.18
Diabetes (Yes=1)	.16
Other mental illness (Yes=1)	.33
Health indicators	
BMI	28 (7.3)
HR	91 (16)
RR	18 (2.2)
TempC	37 (.27)
SpO2	98 (1.9)
Hypertension (Normal)	.21
Hypertension (Prehypertension)	.44
Hypertension (Stage1)	.26
Hypertension (Stage2)	.07
Hypertension (ZHT Crisis)	.02
Binge Drinking (Daily or almost daily)	.21
Binge Drinking (Weekly)	.09
Binge Drinking (Monthly)	.07
Binge Drinking (Less than monthly)	.17
Binge Drinking (Never)	.57
Substance Abuse (Yes=1)	.12

Cluster methods

While K-means is often regarded as the most widely used cluster method, it can only handle numerical data [15]. Though categorical data can be transformed through techniques such as one-hot encoding, important information may be lost in the process. Another commonly used cluster method, K-modes, can only handle categorical data, and similarly, while we can condense numerical attributes into categorical ones, how the intervals are defined is also in question. Therefore, to handle the mixed data type we have, we choose the K-prototypes algorithm. To validate the resulting clusters produced by the K-prototypes algorithm, we used PAM on the same dataset. For each attribute, we performed the Kruskal-Wallace test to see if there exists any statistically significant difference across the three clusters. To sketch an overview of the subgroup characteristics, we combined several sub-categories, for instance, for tobacco use, "never" and "former" are combined to be "no". Since cluster analysis is unsupervised, we do not know how much natural clusters exist in the dataset. Therefore, we computed the Silhouette coefficient as well as the gap statistics to decide how many clusters we should expect from our dataset.

K-prototypes

K-prototype was first proposed by Huang in 1998 [16]. It differs from k-modes in that k-prototypes allow clustering of datasets with mixed data attributes, utilizing a combined dissimilarity measure. Another feature of K-prototypes is that for each resulting cluster, the algorithm returns a real data point from the original dataset as the "prototype" or the representation of the cluster. Cluster methods such as K-means usually return an averaged centroid instead of a real data point.

PAM

PAM, also known as the k-medoids algorithm, clusters data according to medoids [17]. The user inputs the desired number of clusters and the algorithm selects a random set of k items to be the set of starting medoids. Then, the algorithm constructs clusters by iterating through the remaining observations and assigning them to the "closest" medoids. "Closeness" in a PAM analysis is determined by a distance function. Euclidean distance is frequently used in k-means and similar clustering techniques where the data being clustered is entirely continuous. However, because our dataset contains several categorical variables, we used Gower distance, which is capable of handling dichotomous, continuous, and categorical data [18].

Optimal number of clusters: The Silhouette Coefficient & The Gap Statistic

To select the optimal number of clusters to feed into PAM, we used silhouette width. Silhouettes allow users to visualize the quality of a specified number of clusters [19]. When the data has been partitioned in a manner that closely tracks the "natural" cluster, silhouette width is generally high. As Figure 1 shows, when there are 3 clusters, the Silhouette coefficient reaches the peak, which indicates that more natural clusters can be produced. The gap statistic is another way to decide the optimal number of clusters, which is illustrated in much detail in Tibshirani et al.'s paper [20]. Simply put, the higher the gap statistic is, the more natural the resulting clusters will be. Figure 2 shows the gap statistic where there are local peaks at k = 3 and k = 5. However, the Silhouette coefficient hits a local minima at k = 5 but peak at k = 3. Balancing the Silhouette coefficient and the Gap statistic, we decided to use k = 3 as the number of clusters for the following cluster analysis.

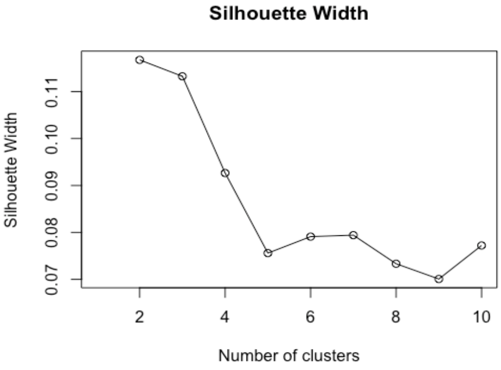

Figure 1 – The Silhouette Coefficient

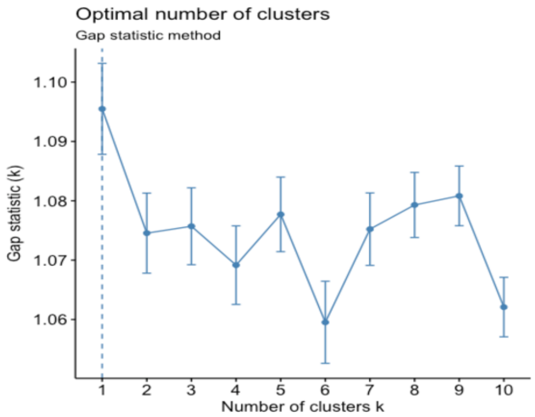

Figure 2 – The Gap statistic

Results

In Table 2, we present the demographical characteristics of the three clusters produced by K-prototypes and PAM. In Table 3, we present the medical history features of the subgroups. In Table 4, we present the health indicator features of the subgroups. The p-values of the Kruskal-Wallace test are marked next to each attribute. The results of K-prototypes are listed on the left of each column, and the results of PAM are on the right. For attributes that contain multiple categories, we only report the categories that are of the majority within the clusters.

Table 2 – Demographical features of subgroups (: p< 0.001)*

Attributes	Cluster 1		Cluster 2		Cluster 3	
Size	379	399	685	520	684	829
Gender* (Male=1)	.44	.64	.62	.73	.36	.24
Age*	72.9	67.1	49.7	43.8	30.3	39.6
Race* (Hispanic)					.55	.72
Race (White)	.48	.55	.33	.58		
Employment* (Retired)	.69	.61				
Employment* (Unemployed)			.76	.76	.71	.72

Table 3 – Medical history features of subgroups

Attributes	Cluster 1		Cluster 2		Cluster 3	
Size	379	399	685	520	684	829
Cancer* (Yes=1)	.38	.37	.16	.09	.09	.14
Diabetes* (Yes=1)	.28	.24	.18	.13	.07	.14
Mental illness* (Yes=1)	.39	.37	.31	.37	.32	.29

Table 4 – Health indicators of subgroups

Attributes	Cluster 1		Cluster 2		Cluster 3	
Size	379	399	685	520	684	829
BMI	26.1	26.0	28.4	27.8	28.0	28.6
HR*	84.3	87.8	87.4	92.7	98.9	92.0
RR*	17.9	17.9	17.8	17.8	17.9	17.9
TempC*	36.8	36.8	36.8	36.8	36.9	36.9
SpO2*	97.3	97.4	98.0	98.0	98.4	98.3
Binge Drinking (Yes=1)*	.27	.23	.49	.83	.30	.28
Tobacco Use (Yes=1)*	.10	.07	.37	.70	.32	.12
Substance Abuse (Yes=1)*	.06	.07	.15	.25	.13	.07

Next, we summarize and describe the distinguishable characteristics of the three subgroups that have emerged from the cluster analysis. The two algorithms seem to diverge in deciding the Binge Drinking and Tobacco Use categories for cluster 2, but if we compare them across different subgroups, they still remain meaningful. Overall we observed differences that are statistically significant across the three groups through the Kruskal-Wallace test.

The Elderly Subgroup

The two cluster algorithms all clearly identified the elderly subgroup, whose members have an averaged age of 67.1~72.9. More than 62% of this subgroups are retired and primarily white. Members in the elderly subgroup also have relatively healthy habits and normal BMI, with the highest proportion (73%~77%) that has never engaging in binge drinking, the highest proportion (90%~93%) of currently non-smokers, and the highest proportion (93%~94%) without substance abuse problems. However, the elderly subgroup has a larger proportion that have cancer (38%) and diabetes (28%), compared to the middle-aged subgroup where the proportion of having cancer is 16% and diabetes is 18%, and the adult subgroup (where the proportion of having cancer is 9% and diabetes 7%).

The Middle-aged Subgroup

The largest proportion of the middle-aged subgroup are unemployed (76%) and male (62%~73%) with an averaged age of 44~49 years old. Compared to the elderly subgroup, the middle-aged subgroup has a lower proportion with cancer (16%) and diabetes (18%). However, the middle-aged subgroup has less healthy habits in general – 30%~66% are current tobacco smokers and 15%~25% have substance abuse issues, which are higher than the elderly subgroup (6%~7%). The middle-aged subgroup also has higher BMI are considered as overweight concerning their age groups.

The Adult Subgroup

This subgroup consists of primarily unemployed (72%) females (64%~76%) that are generally below 40 years old (younger than the middle-aged subgroup). They have the least proportion of having cancer (9%) and diabetes (7%). Compared to the middle-aged subgroup, they have healthier habits – less proportion are current smokers (12%~32%) and only 7%~13% have substance abuse issues. This subgroup also has higher than average BMI and are considered obese concerning their age group.

Discussion

While the two algorithms have some slight discrepancies in deciding certain characteristics, many of their results are consistent. Comparing across the three subgroups, our findings are consistent with previous literature that older population, unemployed population and patients with mental illnesses are more likely to become frequent flyers. However, one significant contribution of our study is that the frequent flyer population consists of subgroups that are substantially different from each other. If only looking at a macro level, these subgroup characteristics may be easily obfuscated. For instance, the general characteristics such as chronic health conditions simply do not apply to all subgroups – the middle-aged subgroup and the adult subgroup seem to have other concerns that lead to their overuse of the ED resource. In addition, the adult subgroup that has the lowest proportion of having chronic health conditions and unhealthy behaviors also takes up a large part of the frequent flyer population, and reasons behind this are rarely explored in existing literature. More investigation needs to be done for the adult female subgroup in order to understand their concerns. Unemployment and burdens of child raising could be factors that contribute to this subgroup's overuse of ED resource. The elderly group who need to cope with chronic health conditions may have difficulty commuting to the right medical facility, and ED may be the most feasible choice in such situations. For the middle-aged male subgroup, less healthy behaviors such as binge drinking and substance abuse may play a bigger role in overusing the ED resource. Thus, simply implementing interventions that are built upon the general frequent flyer characteristics is not able to address the concerns faced by the those different subgroups. Hospitals may consider first investigate what subgroups of frequent flyers constitute the patient populations at their EDs, why certain subgroups are more likely to make repeated visits to EDs, and devise separate intervention programs that target at those subgroups to address their varying needs.

The frequent flyer population at different locations may differ significantly, and our study has suggested cluster analysis as a feasible and promising exploratory stage to understand which subgroups constitute their frequent flyer population at a particular site. Simply borrowing intervention programs from hospitals at different locations or of different settings may fail to suit the needs of the particular frequent flyer subgroup at one hospital.

Limitations and Future Work

Our analysis was based on the structured Electronic Health Records of patients at a single academic medical center, and therefore the results may not be generalizable to hospitals at other locations or of other settings. In addition, while our analysis was able to reveal the subgroups of the frequent flyer population of the study site, the data we used lacks the explanatory power to uncover why these three groups become frequent flyers of ED. It is our future work to extend the analysis to clinical notes that contain richer information of patients' visits.

Conclusions

We performed cluster analysis on 1748 frequent flyers at an academic Medical Center. Three subgroups with substantially different characteristics have emerged from our analysis: the elderly subgroup with chronic health conditions, the middle-aged male subgroup with unhealthy behaviors and the adult female subgroup that are generally healthy. Our findings suggested that the commonly used general characteristics may not apply to all subgroups and different subgroups may face varying challenges in reaching out the right medical resource. Our study sets the stage for more tailored subgroup-level interventions to reduce ED overuse in hospitals across the world.

Acknowledgements

The project described was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institute of Health, USA, through Grant UL1 TR001414. The content is solely the responsibility of the authors and does not necessarily represnet the official views of the NIH.

References

[1] R.W. Schafermeyer and B.R. Asplin, Hospital and emergency department crowding in the United States, *Emerg Med* **15** (2003), 22–27.

[2] R.A. Barish, P.L. Mcgauly, and T.C. Arnold, Emergency room crowding: a marker of hospital health, *Trans A. Clin Climatol Assoc* **123** (2012), 304–311.

[3] P. Bodenmann, V.-S. Velonaki, J.L. Griffin, S. Baggio, K. Iglesias, K. Moschetti, O. Ruggeri, B. Burnand, J.-B. Wasserfallen, F. Vu, J. Schupbach, O. Hugli, and J.-B. Daeppen, Case management may reduce emergency department frequent use in a universal health coverage system: a randomized controlled trial, *J Gen Intern Med* **32** (2017), 508–515.

[4] K.A. Hunt, E.J. Weber, J.A. Showstack, D.C. Colby, and M.L. Callaham, Characteristics of frequent users of emergency departments, *Ann Emerg Med* **48** (2006), 1–8.

[5] C.A. Grover and R.J. Close, Frequent users of the emergency department: risky business, *West J Emer. Med* **10** (2009), 193–194.

[6] E. LaCalle and E. Rabin, Frequent users of emergency departments: the myths, the data, and the policy implications, *Ann Emerg Med* **56** (2010), 42–48.

[7] Y. Ledoux and P. Minner, Occasional and frequent repeaters in a psychiatric emergency room, *Soc Psychiatry Psychiatr Epidemiol* **41** (2006), 115–121.

[8] L.E. Mehl-Madrona, Prevalence of psychiatric diagnoses among frequent users of rural emergency medical services, *Can J Rural Med* **13** (2008), 9.

[9] J. Billings and M.C. Raven, Dispelling an urban legend: frequent emergency department users have substantial

burden of disease, *Health Aff Proj Hope* **32** (2013), 2099–2108.

[10] F.S.J. Blank, H. Li, P.L. Henneman, H.A. Smithline, J.S. Santoro, D. Provost, and A.M. Maynard, A descriptive study of heavy emergency department users at an academic emergency department reveals heavy ED users have better access to care than average users, *J Emerg Nur.* **31** (2005), 139–144.

[11] P. Milbrett and M. Halm, Characteristics and predictors of frequent utilization of emergency services, *J Emerg Nurs* **35** (2009), 191–198.

[12] R.A. Meyers, ed., *Encyclopedia of Analytical Chemistry: Applications, Theory, and Instrumentation*, Wiley, Chichester, NY, 2000.

[13] J.-A. Huang, R.-H. Weng, C.-S. Lai, and J.-S. Hu, Exploring medical utilization patterns of emergency department users, *J Formos Med Assoc* **107** (2008), 119–128.

[14] S.N. Hastings, C. Horney, L.R. Landerman, L.L. Sanders, M.B. Hocker, and K.E. Schmader, Exploring patterns of health service use in older emergency department patients, *Acad Emerg Med* **17** (2010), 1086–1092.

[15] R.O. Duda, *Pattern Classification and Scene Analysis*, Wiley, Chichester, NY, 1973.

[16] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Min Knowl Discov* **2** (1998), 283–304.

[17] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data, An Introduction to Cluster Analysis*, Wiley, Chichester, NY, 1990.

[18] J.C. Gower, A general coefficient of similarity and some of its properties, *Biometrics* **27** (1971), 857–871.

[19] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J Comput Appl Math* **20** (1987), 53–65.

[20] R. Tibshirani, G. Walther, and T. Hastie, Estimating the number of clusters in a data set via the gap statistic, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** (2001) 411–423.

Address for correspondence

Lu He
Department of Informatics,
Donald Bren School of Information and Computer Sciences,
University of California, Irvine
E-mail: lhe11@uci.edu

Kai Zheng, PhD
Department of Informatics,
Donald Bren School of Information and Computer Sciences,
University of California, Irvine
E-mail: kai.zheng@uci.edu