Patient Phenotypes to Identify Resource Allocation and Usage in Primary Care

Stephen Ford, Rehan Merchant, Avinaash Pavuloori, Ryan Williams, Cait Dreisbach, Angela Saunders, Christian Wernz, Jonathan Michel

Abstract— **Resource allocation, including decisions about clinical and administrative staffing, language interpreter requirements, and billing procedures, is challenging in a complex medical system. In the setting of limited resources and high patient need, identification of patients who require a high amount of medical, nursing, and clinical services need to be identified for optimal care. The purpose of this paper is to identify the factors that predict patient phenotypes, a set of observable characteristics of an individual, that reflect their primary care resource usage. The data used in this study are de-identified, patient level data (n=34,957) between January 2019 to December 2021. We used k-means clustering to identify patient phenotypes based on the frequency of primary care and emergency department visits. Using multinomial regression, we then identified *insurance type, comorbidity score, age, race, language, gender, hypertension, chronic opioid, obesity, prediabetes, tobacco usage, congestive heart failure,* and *chronic obstructive pulmonary disease* as significant predictors for the primary care usage phenotypes. Having a more complete, holistic understanding of patient resource phenotypes can help leaders to make important decisions regarding optimal hospital resource allocations. Future work using our methods could be used to prospectively identify patients in high-need resource phenotypes compared to individuals with average annual usage.**

I. Introduction

Predicting patient phenotypes is important, as it could potentially help health systems and clinics to identify high resource needs based on patient characteristics. A phenotype is a set of observable characteristics of an individual. By determining the important predictors for primary care and emergency department usage, we can identify patient populations that could benefit from proactive intervention in the healthcare system and assist hospital leadership with making informed, data-driven decisions regarding resource allocation. Findings from a study in Ontario, Canada showed that 5% of healthcare users consumed 61% of hospital and home care spending (Checulin et al., 2014). Likewise, a study in British Columbia found that 5% of healthcare users consumed 30% of spending on physician services (Checulin et al., 2014). Due to the large financial and resource impact from a small portion of the patient population, it is important to develop interventions directed at these patients that could improve patient outcomes and thus reduce healthcare spending (Checulin et al., 2014).

Previous research has used clustering methods for discovering meaningful patient phenotypes. A K-Means clustering analysis by Nnoaham et al. (2020) was able to identify ten population segments which had distinct profiles of healthcare use, morbidity, demographic characteristics, and risk attributes. Their analysis used seven resource utilization variables including elective and non-elective inpatient admissions, outpatient encounters, Emergency Department (ED) visits, primary care visits and prescriptions (Nnoaham & Cann, 2020). They then analyzed segments *post hoc* to understand morbidity, risk and demographic profiles (Nnoaham & Cann, 2020). Further, K-Means clustering has also been used in many medical applications such as Alzheimer's disease detection among patients and the diagnosis of breast cancer (Alashwal et al., 2019).

The primary objective of this study is to better understand the patient population at four primary care clinics across a major academic health system. We aimed to 1) describe the patient sample by relevant demographics (i.e., age, gender, diagnosis, comorbidity score), 2) identify patient phenotypes based on frequency of ED visit, primary care visits, and related encounters, and 3) determine the factors that predict high resource usage phenotypes.

II. Methods

A. Data Description

This study was approved by the University of Virginia Institutional Review Board (#23444). The data used in this study represented encounter-level patient information from January 2019 to December 2021. In order to aggregate by a unique patient, we grouped by *patient ID* so that each record would be a distinct patient. Once we grouped by unique *patient ID* and selected columns of interest, the data set had n=34,957 patients and 35 columns. The dataset included demographic characteristics (age, gender, race, ethnicity, insurance status), health status (Elixhauser comorbidity index, various diseases and health status registries), resource needs (ED visits, primary care visits , primary care encounters). In order to address missing data, we converted empty cells to NA values. We then changed NA to zero for the registry columns in our data set to make each disease or health status registry a binary variable. NA or NULL values in our Elixhauser comorbidity score column were changed to zero because those values represented insufficient diagnostic data to determine the score for that patient. A total of 27,424 patients (78.5%) had a complete comorbidity score.

B. *Feature Engineering*

Feature engineering tasks for this study were to collapse categorical variable levels to reduce noise in the data including insurance type (Medicare, Medicaid, Self-Pay, and Private), medical registries (Tobacco Registry, Opioid Registry, and Wound Registry), and language preferences (English, Spanish, and Other). Specific to resource and clinic usage, we also calculated important variables such as *Total Primary Care Visits*, *Emergency Department Visits,* and *Primary Care Encounters*. We define a primary care visit as an interaction where a patient visits the same clinic as their primary care provider. All other interactions are grouped together as *Primary Care Encounters* and include, but are not limited to, a MyChart message, a telephone call, a prescription request, billing encounters, home care visits, nursing only visits, and pharmacy only visits. We built our usage phenotypes on the three features discussed above because they provide important resource allocation insights and show which patients are high primary care or emergency department utilizers.

C. *Statistical Analysis*

We completed an exploratory data analysis (EDA) using a subset of the data (n=25,994 patient records, 27 column variables). Our EDA results helped us to inform our feature selection in the larger dataset. Our final analysis was conducted on n=34,957 patient records. After calculating descriptive statistics, we used K-Medoid clustering to identify distinct patient phenotypes. These phenotypes were built on three data features: *Primary Care Visits*, *Primary Care Encounters*, and *Emergency Department Visits*. These features were standardized by subtracting the features mean value and dividing by the feature's mean absolute value before calculating clusters. After identifying our phenotypes using the medoid representative of that phenotype, we used the six phenotypes as the response variable for a multinomial regression model.

The first step in our multinomial regression was determining what subset of predictors to include in our model. We identified significant predictors in our data using univariate analysis and clinical domain expertise. To reduce the amount of registries in the model we started with a baseline model using the following predictors: i*nsurance type, comorbidity score, age, race, language, and gender*. We then tested the nine registries through a manual forward selection process and determined the following registries to be significant: *hypertension, chronic opioid, obesity, prediabetes, tobacco, congestive heart failure, and copd*. We then built our final model using a manual forward selection process to determine which registries to include in our model. We then used a machine learning approach to identify the accuracy of predicting a patient's resource phenotype. We split our data into a test train split of 70-15-15 for test, train and validation sets. Coefficients were analyzed at a 0.05 significance level. Our analysis was completed using the R packages pam, nnet, and tidyverse in RStudio (Version 4.1.3).

III. RESULTS

A. *Identification of Patient Phenotypes*

Using *Primary Care Visits*, *Primary Care Encounters*, and *Emergency Department Visits* (ED_Visits), the K-medoid clustering indicated that the data is best represented by six clusters (the phenotypes) as confirmed by the ELBO plot. Figure 1 below, shows the ELBO plot and the total sum of squares reaching a minimum at six clusters.

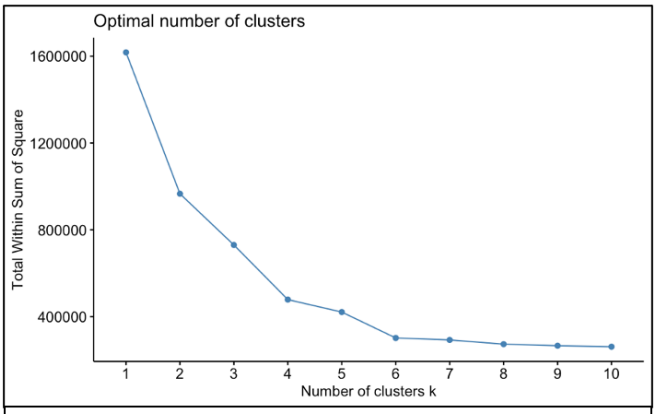

Figure 1: This image is an ELBO plot used for our clustering analysis. As you can see from this plot, the total sum of squares reaches a minimum at 6 clusters. Clusters 7-10 show slight improvement in sum of squares, but the change is minimal. Since the ELBO plot supports the use of six clusters, we opted to use K = 6 clusters for our medoid clustering.

There was improvement with having phenotypes range from 7-10, but the reduction was minimal. After the identification of six phenotypes, the output resulted in n = 9,544 for cluster 1, n = 908 for cluster 2, n = 3,682 for cluster 3, n = 2,741 for cluster 4, n = 11,458 for cluster 5, and n = 6,624 for cluster 6. We labeled each cluster by the relevant usage phenotype. For example, cluster 2 had an average value of 10 for ED_Visits, so we labeled this cluster as *High ED Usage* patients. We labeled the remaining clusters using a similar method, and labeled patient profiles based on primary care and ED usage. Refer to Table 1 below, for a summary of our phenotype labels and results.

Cluster represented by best Medoid	Cluster Phenotype Label	Average Primary Care Visits	Average Primary Care Encounters	Average Emergency Department Visits
Cluster 1	*Annual PC*	3	5	0
Cluster 2	*High ED Usage*	8	15	10
Cluster 3	*ED Usage*	2	5	3
Cluster 4	*Highest PC*	17	26	1
Cluster 5	*Minimal PC*	0	1	0
Cluster 6	*High PC*	8	12	0

Table 1: This table represents the patient ID's for the medoids that best represent each cluster. Patient ID's have been replaced with the cluster that is represented by these medoids**.** Clusters 1-6 have been labelled as: "Annual PC", "High ED Usage", "ED Usage", "Highest PC", "Minimal PC", and "High PC." These labels were chosen based on the average values in the table above. "Highest PC", "High PC", "Annual PC, and "Minimal PC" are referring to the Primary Care utilization based on average primary care visits and average primary care encounters for these groups. "High ED Usage" and "ED usage" labels were determined based on Average Emergency Department Visits for these clusters.

Variable	Level(s) if applicable	High ED Usage (n = 908)	ED Usage (n = 3,682)	Highest PC (n = 2,741)	High PC (n = 6,624)	Annual PC (n = 9,544)	Minimal PC (n = 11,458)
Insurance type	Private	215 (23.7)	1,465 (39.8)	800 (29.2)	2,949 (44.5)	5,828 (61.1)	6,336 (55.3)
	Medicaid	258 (28.4)	789 (21.4)	449 (16.4)	678 (10.2)	791 (8.3)	1,164 (10.2)
	Medicare	361 (39.8)	986 (26.8)	1,379 (50.3)	2,591 (39.1)	2,039 (21.4)	1,728 (15.1)
	Self Pay	74 (8.1)	442 (12.0)	113 (4.1)	406 (6.1)	886 (9.3)	2,230 (19.5)
Race	White	426 (46.9)	1,736 (47.1)	1,643 (59.9)	4,165 (62.9)	5,313 (55.7)	4,922(43.0)
	Asian	3 (.3)	22 (0.6)	20 (0.7)	76 (1.1)	196 (2.1)	173 (1.5)
	Black	388 (42.7)	1,034 (28.1)	881 (32.1)	1,450 (21.9)	1,351 (14.2)	1,304 (11.4)
	Other	24 (2.6)	177 (4.8)	56 (2.0)	234 (3.5)	459 (4.8)	599 (5.2)
	Unknown	67 (7.4)	713 (19.4)	141 (5.1)	699 (10.6)	2,225 (23.3)	4,460 (38.9)
Gender	Male	435 (47.9)	1,842 (50.0)	1,031 (37.6)	2,690 (40.6)	4,411 (46.2)	5,371 (46.9)
	Female	473 (52.1)	1,840 (50.0)	1,710 (62.4)	3,933 (59.4)	5,130 (53.8)	6,084 (53.1)
	Unknown	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.0)	3 (0.0)	3 (0.0)
Elixhauser	N/A	7.86 (4.3)	4.23 (3.5)	7.64 (3.2)	5.18 (2.9)	3.03 (2.5)	1.73 (2.4)
Age	N/A	56.39 (16.4)	53.3 (18.4)	63.67 (15.2)	61.24 (16.3)	53.69 (17.3)	48.26 (17.6)

Table 2: This is a table of patient demographic information broken down by the clusters labeled from our results

B. *Sample Characteristics*

Table 2 above, shows important patient demographic information broken down by the clusters labeled from our results. The average age of our population is 54 years old. Further describing the demographic characteristics, 45% (n= 15,780) of the population is male and 55% (n=19,170) are females. Our population consists of 52.1% (n=18,205) patients who identify as White, 1.4% (n=490) patients who identify as Asian, 18.3% (n=6,408) patients who identify as Black, and 28.2% (n=9,854) who identify as Other. There are 50.3% (n=17,593) patients with Private Insurance, 11.8% (n= 4,129) with Medicaid, 26.0% (n=9,084) with Medicare, and 11.9% (n=4,151) with Self Pay. More specifically for the primary care phenotypes, the average age of a patient in the *Highest PC* column was 63.7 years of age. Females comprise 62% (n=1710) and 59% (n=3933) of the *Highest PC* and *High PC* phenotypes respectively, with males making up only 38% (n=1031) and 41% (n=2690) respectively.

C. *Factors that predict patient resource phenotype*

Our multinomial regression identified Medicaid (*OR*=2.28, *95CI=[1.95 ,2.67]*, $p<0.00$), and Medicare insurance (*OR*=1.69, *95CI=[*1.48,1.93], $p<0.00$), Elixhauser (*OR*=1.40, *95CI*=[1.37,1.42], $p<0.00$), age (*OR*=1.01, *95CI=[*1.01,1.01], $p<0.00$), patients who identify their race as Black (*OR*=1.77, *95CI=[*1.58,1.99], $p<0.00$) or unknown (*OR*=0.72, *95CI=[*0.59, 0.88], $p<0.00$), primarily Spanish language (*OR*=1.38, *95CI=[*1.06, 1.80], *p=0.02*), identifying as female (*OR*=1.29, *95CI=[*1.16, 1.42], $p<0.00$) or unknown (*OR*=0.57, *95CI*=[0.54, 0.61], $p<0.00$), hypertension registry (*OR*=1.30, *95CI*=[1.15, 1.46], $p<0.00$), chronic opioid use registry (*OR*=5.89, *95CI*=[4.71, 7.37], $p<0.00$), obesity registry (*OR*=1.69, *95CI*=[1.52, 1.88], $p<0.00$), prediabetes registry (*OR*=0.82, *95CI*=[0.72, 0.95], $p=0.01$), congestive heart failure registry (*OR*=1.45, *95CI*=[1.20, 1.75], $p<0.00$) and COPD registry (*OR*=1.52, *95CI*=[1.26, 1.84], $p<0.00$) as significant predictors in the *High PC* phenotype compared to the *Annual PC*.

In the *Highest PC* compared to the *Annual PC* phenotype the significant predictors and their significant level are Medicaid (*OR*=1.33, *95CI*=[1.18, 1.50], $p<0.00$) and Medicare (*OR*=1.32, *95CI*=[1.20, 1.45], $p<0.00$) insurance, Elixhauser (*OR*=1.18, *95CI*=[1.16, 1.19], $p<0.00$) , age (*OR*=1.01, *95CI*=[1.01, 1.01], $p<0.00$), Black (*OR*=1.21, *95CI*=[1.11, 1.33], $p<0.00$) and unknown (*OR*=0.77, *95CI*=[0.70, 0.86], $p<0.00$) in race, female (*OR*=1.22, *95CI*=[1.138, 1.30], $p<0.00$) and unknown (*OR*=1.95, *95CI*=[0.20, 18.95], $p<0.00$) in sex, hypertension registry (*OR*=1.34, *95CI*=[1.24, 1.45], $p<0.00$) , chronic opoid use registry (*OR*=2.35, *95CI*=[1.89, 2.92], $p<0.00$) , and obesity registry (*OR*=1.50, *95CI*=[1.39, 1.61], $p<0.00$). Our final model predicting all phenotypes had a validation accuracy of 44% and test accuracy of 45%.

IV. Discussion

Our clustering output was able to provide meaningful results. Upon reviewing the medoids that best represent each cluster, we were able to name six groups based on the medoid from each of the clusters. We labeled our patient profiles as *Annual PC, High ED Usage, ED Usage, Highest PC, Minimal PC,* and *High PC*. Our final model had a total of 13 variables. The final variables in our best model are: *insurance type, comorbidity score, age, race, language, gender, hypertension, chronic opioid usage, obesity, prediabetes, tobacco, congestive heart failure,* and *chronic obstructive pulmonary disease*. The best model had a validation and test accuracy of 44% and 45%, respectively.

We were able to predict which cluster a patient would belong to 2.5 times better than random and extract interesting insights upon evaluating the odds ratios from our results. We saw that Medicaid patients, compared to those with private insurance, were 2.16 times more likely to be in the *Highest PC* cluster when compared to patients in the *Annual PC* cluster. We saw that Medicare patients were 1.65 times more likely to be in the *Highest PC* cluster when compared to patients in the annual PC cluster and using private insurance. We also found that patients who identify their race a*s Black, c*ompared to those who identify as white, have a 77% higher odds of belonging to the *High PC* and a 21% higher odds of belonging to the *Highest PC* phenotypes

compared to *Annual Usage.* Finally, we observed that patients who are included on the *Chronic Opioid Registry*, compared to those who are not, are 6 times more likely to be in the *Highest PC* phenotype when compared to patients in the *Annual PC* phenotype who are not on the *Chronic Opioid Registry*. This information can be beneficial to the health system in our study when planning resource allocation and developing patient engagement strategies.

According to research, patients that are uninsured primarily use the ED three times more likely than their uninsured counterparts thus making them the most resource intensive patient group (Barish et al,. 2012). Since uninsured patients are most likely to utilize the ED, they present the highest costs since hospitals must provide care to all patients seeking emergency care (Barish et al,. 2012). Additionally, in our dataset approximately 68% of *High ED* phenotype patients were on either Medicaid or Medicare, and these patients pay the lowest premiums, thus resulting in lower profits for the hospital. Though assessing ED usage was not the primary goal of this study, ED crowding is a prevalent issue across health systems, especially during the novel Coronavirus pandemic. ED crowding is important to solve as transferring between hospitals provides higher risks the longer a patient is in transit between health systems (Kolenich, et al., 2022). By determining *High ED* utilizing patients in two different usage phenotypes, the health system in our study can potentially create proactive educational measures to prevent these patients from depending on the ED for medical attention.

Engaging patients is important, as research indicates that engaged patients have a better experience of care and health outcomes. Past studies have found numerous reasons for patients lacking active engagement in the receipt of their care. Some of these reasons include patients not understanding medical jargon, lack of access to outside resources, clinicians ignoring or not hearing patient information clearly, and patients not receiving necessary information from clinicians (Mishra et al., 2016). A health survey showed that while 85% of patients asked staff medical questions, only 17% managed to take an active role in their care (Mishra et al., 2016). This suggests that patients view their care in a passive manner and do not want to interfere with their doctor's care routine. A major limitation of past research on clinical engagement is that it did not explore the role of clinicians in supporting patient needs. Rather than exploring the reasons for reduced engagement, research needs to explore how provider support can help increase patient engagement (Mishra et al., 2016). Our usage phenotype information will be useful for health systems since it can identify groups of patients that could benefit from increased patient engagement, therefore reducing the frequency of health care usage and offsetting resource constraints.

The strengths of our study include, 1) analysis of a high volume of patients with diverse clinical needs, 2) patients representing four different clinics within a single health system, and 3) our use of K-medoid clustering for identification of the relevant phenotypes. It has been found that K-Medoids clustering is more robust to noise and outliers when compared to K-Means (Gorban, 2011). The reason for this is because K-Medoids minimize a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances (Gorban, 2011). However, this paper also has several limitations that could be improved upon in future work. First, our accuracy only achieves approximately 50%, indicating that there are likely latent or unmeasured patient-centered or clinical variables that contribute to resource usage that are not included in the analysis. We expect that variables such as zip code, diagnosis codes, mental health, and more sociodemographic variables may help to increase accuracy of our model.

V. Conclusion

Identifying patient profiles based on cost utilization can be beneficial for health systems when planning resource allocation. The phenotype groups that were labeled in our study can help to determine which groups should be prioritized when planning for primary care resource allocation. Studying the patient populations through our multinomial regression model and their respective resource needs in primary care can help clinical leaders identify patients that could benefit from increased engagement opportunities.

REFERENCES

[1] Chechulin, Y., Nazerian, A., Rais, S., & Malikov, K. (2014, February). *Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada)*. Healthcare policy = Politiques de sante. Retrieved April 1, 2022

[2] Nnoaham, K. E., & Cann, K. F. (2020, May 27). *Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? - BMC public health*. BioMed Central. Retrieved April 1, 2022.

[3] Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019, May 24). *The application of unsupervised clustering methods to Alzheimer's disease*. Frontiers in computational neuroscience. Retrieved April 1, 2022

[4] Barish, R. A., McGauly, P. L., & Arnold, T. C. (2012). *Emergency room crowding: A marker of hospital health*. Transactions of the American Clinical and Climatological Association. Retrieved March 22, 2022

[5] Kolenich, E., Brown, B., & Gentry, M. (2022, January 15). *Crowded hospitals in Virginia have few open beds to offer the rush of patients*. Richmond Times-Dispatch. Retrieved March 22, 2022.

[6] Mishra, S. R., Haldar, S., Pollack, A. H., Kendall, L., Miller, A. D., Khelifi, M., & Pratt, W. (2016, May 7). *"not just a receiver": Understanding patient behavior in the hospital environment*. Retrieved March 21, 2022

[7] Gorban, A. (2011). *Means and K-Medoids*. University of Leicester Mathematics. Retrieved April 3, 2022