Characterizing Subgroups of High-Need, High-Cost Patients Based on Their Clinical Conditions: a Machine Learning-Based Analysis of Medicaid Claims Data

Sudhakar V. Nuti, MSc[1,2], Patrick Doupe, PhD[3], Blanca Villanueva, BS[4], Joseph Scarpa, PhD[2], Emilie Bruzelius, MPH[2,5], and Aaron Baum, PhD[2]

[1]Yale School of Medicine, New Haven, CT, USA; [2]Department of Health System Design and Global Health, and the Arnhold Institute for Global Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [3]Zalando, Inc., Berlin, Germany; [4]CYNGN, Inc., California, USA; [5]Department of Epidemiology, Joseph L Mailman School of Public Health, Columbia University, New York, NY, USA.

KEY WORDS: Medicaid; high-cost patients; high-need patients; machine learning; patient segmentation.

J Gen Intern Med 34(8):1406–8
DOI: 10.1007/s11606-019-04941-8

INTRODUCTION

Health systems are increasingly adopting intensive primary care and care coordination programs to improve outcomes for high-need, high-cost (HNHC) patients, the 5% of patients who account for over 50% of health care costs.[1] However, research on such programs has shown mixed results, improving patient satisfaction but having limited impact on quality of life, illness control, and need for acute care services.[2, 3] As a group, HNHC patients are defined based on their utilization of care, rather than their clinical conditions. Yet, to better manage HNHC patients, clinicians need to match patients to care models tailored to their clinical conditions.[4] Here, we utilized an open-source, machine learning method to describe different subgroups of HNHC patients based on their clinical characteristics for an urban Medicaid population in the Mount Sinai Health System (MSHS).

METHODS

Study Population

We examined administrative claims from 34,764 patients insured by a Medicaid managed care organization that operates in New York and New Jersey who were admitted to at least one hospital contained within MSHS between 1/1/2014 and 12/31/2015. This study was approved by the Icahn School of Medicine at Mount Sinai Institutional Review Board (IRB-16-01066).

Dr. Doupe and Ms. Villanueva were affiliated with the Department of Health System Design and Global Health, Arnhold Institute for Global Health, and Icahn School of Medicine at Mount Sinai during the time the work was conducted.

Published online March 18, 2019

High-Need, High-Cost Criteria

We selected patients ages 18 years and older who fulfilled either of two inclusion criteria: admitted at least three times within any 12-month period between 2014 and 2015 or admitted at least two times within the same time period, with at least one serious mental health condition as a primary diagnosis. We chose this definition based on Johnson et al.[5] A hospitalization was defined as ICD-9 primary diagnosis codes in inpatient hospital claims; secondary and tertiary diagnoses were comorbidities.

Data Preparation

Using Medicaid claims data, we created a dataset of patient features consisting of ICD-9-based clinical condition categories, 31 electronic health record clinical codes, and demographic variables, including age, sex, and neighborhood of residence. We used the Clinical Classification Software scheme to categorize each primary diagnosis ICD-9 code into one of 250 clinical condition categories.[5] For each clinical condition category, we created a variable that was equal to one when a patient's claim line item included a primary diagnosis code fell into that category, and zero otherwise.

Data Analysis

Clustering is an unsupervised machine learning method for exploring non-parametric patterns within data that may not be discernable by parametric multivariate regression methods. We used affinity propagation (AP), a clustering algorithm that does not require the number of clusters in the data set to be known a priori.[6]

We utilized the apcluster package in R (3.3.1) using RStudio (version 0.99.903) for our analysis. For ease of interpretation, we focused on the top 25 clusters by size. The results were interpreted for clinical salience by investigators with clinical expertise (SN and JS).

RESULTS

Cohort Characteristics

There were 2397 patients in our cohort. The average age of patients was 46.5 (standard deviation [SD] 15.0) years and 56% were female (Table 1). The average number of admissions was 79 (SD 45.2) and total cost of care was $50,700 (SD $68,300).

Clinical and Cost Characteristics for Top 25 Clusters

Table 1 presents the main findings. The two largest clusters were characterized by depression and other mood disorders. Twelve of the top 25 clusters were primarily mental health and substance use conditions. Other prominent clusters included pregnancy- and birth-related complications, heart conditions, and diabetes.

Table 1 Characteristics of Top 25 Clusters

Prevalent clinical condition code(s)	Prevalent electronic health record code(s)	Number of patients	Percent of total cohort (%)	Female, % (SD)	Age, years (SD)	Average number of comorbidities (SD)	Total cost of care, '000 $ (SD)
Overall cohort	–	2397	100.0	56.0	46.5 (15.0)	10.8 (9.0)	50.7 (68.3)
Mood disorders	Depression	46	1.9	69.6 (46.5)	31.7 (9.5)	5.4 (4.1)	22.9 (16.6)
Mood disorders	Depression	42	1.8	69.0 (46.8)	38.8 (13.5)	5.3 (3.8)	19.3 (19.5)
Schizophrenia and other psychotic disorders		38	1.6	34.2 (48.0)	40.6 (14.6)	4.7 (4.1)	13.6 (14.1)
Coronary atherosclerosis and other heart disease; nonspecific chest pain	Ischemic heart disease; hypertension; hyperlipidemia; diabetes; depression	34	1.4	61.8 (49.3)	56.7 (5.3)	12.1 (7.1)	76.7 (63.6)
Schizophrenia and other psychotic disorders		34	1.4	35.3 (48.5)	33.8 (11.4)	3.5 (2.8)	25.2 (30.5)
Liveborn; other complications of birth; other pregnancy and delivery including normal		28	1.2	100.0 (0)	33.1 (6.3)	7.4 (4.0)	10.4 (8.2)
Substance-related disorders; alcohol-related disorders	Depression	26	1.1	23.1 (42.7)	40.9 (10.2)	7.0 (3.9)	15.8 (11.3)
Schizophrenia and other psychotic disorders; Mood disorders	Depression	25	1.0	72.0 (45.8)	34.0 (11.8)	6.0 (3.1)	28.2 (20.0)
Substance-related disorders	Hypertension	25	1.0	24.0 (43.5)	46.8 (12.9)	5.2 (3.6)	15.1 (16.1)
Coronary atherosclerosis and other heart disease; nonspecific chest pain	Ischemic heart disease; hypertension; hyperlipidemia	25	1.0	36.0 (49.0)	64.1 (13.4)	12.7 (7.2)	35.4 (33.8)
Polyhydramnios and other problems of amniotic cavity		24	1.0	100.0 (0)	29.4 (5.1)	4.7 (2.2)	7.0 (6.8)
Nonspecific chest pain	Rheumatoid arthritis osteoarthritis; hyperlipidemia	24	1.0	37.5 (49.5)	52.8 (8.3)	8.8 (6.9)	32.4 (21.8)
Other complications of pregnancy	Anemia	24	1.0	95.8 (20.4)	33.1 (7.5)	7.8 (4.1)	19.0 (20.4)
Mood disorders	Depression	23	1.0	73.9 (44.9)	36.7 (14.1)	6.7 (4.7)	47.1 (68.7)
Mood disorders	Depression	22	0.9	68.2 (47.7)	40.5 (14.4)	5.3 (3.4)	43.0 (38.9)
Other complications of birth		22	0.9	90.9 (29.4)	31.7 (10.2)	6.2 (3.4)	11.5 (7.5)
Diabetes mellitus with complications	Diabetes; chronic kidney disease	22	0.9	27.3 (45.6)	42.3 (13.7)	10.1 (7.1)	55.7 (52.3)
Schizophrenia and other psychotic disorders		21	0.9	52.4 (51.2)	37.7 (12.8)	3.2 (2.9)	14.8 (14.1)
Mood disorders	Depression	21	0.9	52.4 (51.2)	34.1 (12.4)	4.6 (3.0)	18.0 (20.9)
Mood disorders	Depression; hyperlipidemia; hypertension	21	0.9	66.7 (48.3)	50.0 (11.7)	8.5 (6.0)	39.7 (25.6)
–*		20	0.8	50.0 (51.3)	47.8 (12.6)	8.6 (6.1)	35.8 (32.6)
Mood disorders; schizophrenia and other psychotic disorders		20	0.8	50.0 (51.3)	37.4 (12.3)	5.3 (6.2)	17.7 (17.9)
Alcohol-related disorders	Depression; diabetes; hyperlipidemia; hypertension	20	0.8	35.0 (49.0)	52.2 (6.4)	5.5 (3.7)	30.7 (28.0)
–*		20	0.8	40.0 (50.3)	45.9 (13.9)	7.6 (5.0)	30.6 (58.0)
Early or threatened labor		20	0.8	95.0 (22.4)	28.0 (4.6)	5.7 (2.4)	29.3 (54.8)

*These clusters did not have a clear association with a clinical condition

There was surprisingly large variation in average costs of care across the top 25 clusters, ranging from $7000 to $76,600 per patient per year.

DISCUSSION

We used an open-source machine learning method to describe different subgroups of HNHC patients based on their clinical characteristics. The largest HNHC patient subgroups were characterized by mental and behavioral health conditions. We found marked heterogeneity in HNHC patient costs across the different subgroups. We also identified an unexpected patient population: patients with pregnancy-related complications.

Corresponding Author: *Aaron Baum, PhD; Department of Health System Design and Global Health, and the Arnhold Institute for Global Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA (e-mail: Aaron.baum@mssm.edu).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. **Blumenthal D**, **Chernof B**, **Fulmer T**, **Lumpkin J**, **Selberg J.** Caring for high-need, high-cost patients—an urgent priority. N Eng J Med 2016;375(10):909–911.
2. **Bleich SN**, **Sherrod C**, **Chiang A**, et al. Systematic Review of Programs Treating High-Need and High-Cost People With Multiple Chronic Diseases or Disabilities in the United States, 2008-2014. Prev Chronic Dis 2015;12:E197.
3. **McCarthy D**, **Ryan J**, **Klein S**. Models of Care for High-Need, High-Cost Patients: An Evidence Synthesis. Issue Brief (Commonw Fund) 2015;31:1–19.
4. **Anderson GF**, **Ballreich J**, **Bleich S**, et al. Attributes common to programs that successfully treat high-need, high-cost individuals. Am J Manag Care 2015;21(11):e597–600.
5. **Johnson TL**, **Rinehart DJ**, **Durfee J**, et al. For many patients who use large amounts of health care services, the need is intense yet temporary. Health Aff (Millwood) 2015;34(8):1312–1319.
6. Agency for Healthcare Research and Quality. Clinical Classifications Software (CCS) for ICD-9-CM. Available at: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed 4 Jan 2019.