Combining Contrast Mining with Logistic Regression To Predict Healthcare Utilization in a Managed Care Population

Lincoln Sheets[1, 2]; Gregory F. Petroski[2]; Yan Zhuang[3]; Michael A. Phinney[3]; Bin Ge[2]; Jerry C. Parker[2]; Chi-Ren Shyu[1]

[1]University of Missouri, MU Informatics Institute, Columbia, Missouri, USA;
[2]University of Missouri, School of Medicine, Columbia, Missouri, USA;
[3]University of Missouri, College of Engineering, Columbia, Missouri, USA;

Keywords
Data mining, prediction models, clinical decision support, data reuse, practice management

Summary

Background: Because 5% of patients incur 50% of healthcare expenses, population health managers need to be able to focus preventive and longitudinal care on those patients who are at highest risk of increased utilization. Predictive analytics can be used to identify these patients and to better manage their care. Data mining permits the development of models that surpass the size restrictions of traditional statistical methods and take advantage of the rich data available in the electronic health record (EHR), without limiting predictions to specific chronic conditions.

Objective: The objective was to demonstrate the usefulness of unrestricted EHR data for predictive analytics in managed healthcare.

Methods: In a population of 9,568 Medicare and Medicaid beneficiaries, patients in the highest 5% of charges were compared to equal numbers of patients with the lowest charges. Contrast mining was used to discover the combinations of clinical attributes frequently associated with high utilization and infrequently associated with low utilization. The attributes found in these combinations were then tested by multiple logistic regression, and the discrimination of the model was evaluated by the c-statistic.

Results: Of 19,014 potential EHR patient attributes, 67 were found in combinations frequently associated with high utilization, but not with low utilization (support>20%). Eleven of these attributes were significantly associated with high utilization ($p<0.05$). A prediction model composed of these eleven attributes had a discrimination of 84%.

Conclusions: EHR mining reduced an unusably high number of patient attributes to a manageable set of potential healthcare utilization predictors, without conjecturing on which attributes would be useful. Treating these results as hypotheses to be tested by conventional methods yielded a highly accurate predictive model. This novel, two-step methodology can assist population health managers to focus preventive and longitudinal care on those patients who are at highest risk for increased utilization.

Correspondence to
Lincoln Sheets, MD, PhD
University of Missouri, Columbia, Missouri
telephone: 417–860–1197
fax: 573–884–4808
Email: SheetsLR@health.missouri.edu.

Appl Clin Inform 2017; 8: 430–446
https://doi.org/10.4338/ACI-2016-05-RA-0078
received: May 26, 2016
accepted: February 21, 2017
published: May 3, 2017
Citation: Sheets L, Petroski GF, Zhuang Y, Phinney MA, Ge B, Parker JC, Shyu C-R. Combining contrast mining with logistic regression to predict healthcare Appl Clin Inform 2017; 8: 430–446
https://doi.org/10.4338/ACI-2016-05-RA-0078
Funding
LS and JCP: This publication was made possible by Grant Number 1C1CMS331001–01–00 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services or any of its agencies. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. MAP is supported by the US Department of Education Graduate Assistance in Areas of National Need (GAANN) Fellowship under grant number P200A100053, and YZ and CRS are supported by the Shumaker Endowment for biomedical informatics. The high performance computing infrastructure used in this research is currently supported by the National Science Foundation under grant number CNS-1429294.

1. Background and Significance

1.1 Scope of Problem

To achieve the "Triple Aim" of (a) better health outcomes, (b) better healthcare delivery, and (c) lower costs [1], managed care programs seek to improve interactions between informed, activated patients and prepared, proactive providers [2], including preventive care [3]. Unfortunately, healthcare often fails to provide effective coordination of care across a target population [4, 5]. When care coordinators do not know which of their patients are most "at risk" for increased healthcare needs, they typically allocate their time by responding to the patient in front of them at the moment [6].

Predictive analytics can be used to rapidly spot opportunities to improve care management [7]. Because 5% of patients incur 50% of healthcare expenses [8], population health managers need to focus preventive and longitudinal care on those patients who are at highest risk of increased utilization. This approach can facilitate the transition from traditional "reactive" models of medical care [6] to one of maintaining health and avoiding preventable conditions. Focusing proactive and preventive care on these high-risk patients directly addresses the Triple Aim by lowering costs and improving health outcomes, and indirectly may also improve healthcare delivery [9].

1.2 Limitations of Current Methods

Current models that predict health risks for community-dwelling older adults achieve discrimination measures up to about 70%, as measured by c-statistic [10]. Because regression analysis and other traditional statistical methods are constrained by the limited number of attributes that can be used [11], most predictive algorithms have focused on specific conditions such as diabetes [12] or hypertension [13]. However, population health managers need predictive analytics that identify patients at increased risk for all-cause healthcare utilization.

Higher accuracies have been achieved by more specialized prediction models, such as one for imaging utilization [14]. Other investigators [15–17] have built successful models on the basis of demographic and utilization characteristics using a limited subset of clinical data. However, these strategies may not fully exploit the highly detailed clinical history available in electronic health records (EHR). Other studies [18] have used rich clinical data to identify practice patterns without explicitly predicting outcomes. Data mining algorithms permit the development of models that use the rich data available in the EHR [19], without limiting predictions to specific chronic conditions or high-level summaries (such as restricted EHR data).

2. Objective

The objective was to demonstrate the usefulness of unrestricted EHR data for predictive analytics in managed healthcare.

3. Methods

3.1 Population

LIGHT^2 (**L**everaging **I**nformation Technology to **G**uide **H**i-Tech and **H**i-Touch Care) was a Health Care Innovation Award from the Centers for Medicare and Medicaid Services to examine the use of advanced health information technology and care coordination in a managed population. The LIGHT^2 program recruited primary care patients at the University of Missouri Health System who were already enrolled in Medicare or Medicaid. The cohort comprised of 9,568 patients who were enrolled in LIGHT^2 on or before July 1, 2013.

3.2 Data Source

We retrieved all patient diagnoses, prescriptions, and other clinical attributes from the EHR of the University of Missouri Health System as maintained by clinicians during the fiscal years ending in 2012 and 2013.

3.3 Data Selection

We selected hospital and clinic charges as the outcome of interest for this study because they are easily measured, continuously distributed, and can be compared comprehensibly between diverse patients or populations. We selected the 5% of patients (n=479) with the highest health system charges during FY2013 (the fiscal year ending on March 31, 2013). The FY2013 charges for this top 5% ranged from $94,896 to $3,029,833; and the top 5% accounted for 49.7% of charges incurred by the entire LIGHT[2] cohort for that year (▶Figure 1). The FY2012 charges for the top 5% of patients in that fiscal year ranged from $63,967 to $4,288,603, which we used to define the independent variable of high prior-year cost.

Mining data to contrast two or more conditions, or contrast mining [20], requires comparison groups from comparable populations. Other data reduction techniques such as principal components are less than ideal for several reasons: they do not make explicit use of the known-groups nature of the problem, are not well suited to binary data, and would be computationally impractical with the large number of characteristics considered here. Furthermore, both principal component analysis and factor analysis aim at finding linear combinations of features as opposed to identifying individual features that best discriminate between groups. For this application of contrast mining, we used multiple comparison groups in order to test the flexibility and robustness of the methodology under varying input conditions. We first excluded patients with zero healthcare system charges on the grounds that individuals with no recent hospital or outpatient visits may not have current medical histories in the healthcare system EHR. Therefore, the comparison groups comprised each of the lowest non-zero 5%, 10%, 20%, 30%, 40%, and 50% of FY2013 charges (▶Table 1).

3.4 Data Projection

The EHR records at the time of data collection contained a mixture of diagnosis codes from the International Classification of Diseases, 9th Revision (ICD-9) nomenclature and the Systematized Nomenclature of Medicine (SNOMED). The patient records selected for contrast mining contained 3,998 unique SNOMED codes and 3,615 unique ICD-9 codes. These records also contained 10,725 unique medication prescriptions and nine demographic attributes (i.e., age, gender, race/ethnicity, marital status, English fluency, Medicaid coverage, high prior-year (FY2012) costs, body mass index (BMI), and history of adherence to prescription instructions.

We also categorized the 3,615 ICD-9 codes in the dataset into 612 diagnosis-related groups (DRG), and the 10,725 prescriptions into 55 higher-level therapeutic classes. All 19,014 attributes were collected for the selected patients at the end of FY2012, prior to the FY2013 outcome of interest (▶Figure 2).

3.5 Data Mining

In order to process contrast mining algorithms, we built a distributed association-rule mining (ARM) tool suite on Apache Spark in HDFS (Hadoop Distributed File System) [21]. Because ARM requires binary values, we transformed all variables (i.e., attributes) to true-or-false flags using a PHP script. Because ARM analyses identify the presence of attributes in each combination, but cannot identify the absence of any attribute or combination of attributes, flags must be coded for all possible categorical values in association rule mining (even when the categories are mutually exclusive), rather than the n-1 categories used in traditional regression. For example, we transformed each categorical variable (i.e., race/ethnicity and marital-status) to a set of binary values: (a) "race/ethnicity=white-non-Hispanic or not, =Hispanic or not, =African-American or not, =Asian or not, =Native-American or not, =other or not, =unknown or not," and (b) "marital-status=single or not,

=married or not, =divorced or not, =widowed or not." We transformed the two continuous variables (i.e., age and BMI) to binary flags after transformation to standard [22] categories: (a) "age=18–24 or not, =25–44 or not, =45–64 or not, =65-84 or not, =85-or-older or not," and (b) "BMI=less-than-18.5 or not, =18.5–24.9 or not, =25–29.9 or not, =30-or-higher or not." For each of these sets of binary values created from categorical variables, only one is true for any given patient. For example, if "marital-status=married" is true, then "marital-status=single," "=divorced," and "=widowed" are false.

We then discovered frequent attribute combinations using an "Apriori" algorithm [23] with a minimum support of 0.2 (i.e., excluding attribute combinations found in less than 20% of transactions or fewer than 192 out of 958 patients). We chose this parameter, which should identify 20% of 5% of the population or 1% overall, in order to strike a balance between the recognition of rare conditions in an intrinsically sparse dataset and the elimination of outliers that could misrepresent typical clinical histories. We limited results to attribute combinations that included the outcome of interest (i.e., FY2013 charges over $94,895 or not).

3.6 Statistical Confirmation

In the second step, we dissected the attribute combinations found frequently (20% or more) in patients with high utilization and infrequently in patients with low utilization into individual attributes. Because some age categories were found infrequently in some comparison groups but not in others, "Age" was restored to a continuous integer variable; and because all patients were marked as either "female" or "male", the "male" flag was dropped and all patients were marked as female or not. We then treated these contrasting attributes as hypotheses to be tested with multiple regression, using the entire population as the validation set.

We used forward selection with $p < 0.05$ as the entry criterion to add attributes to a simplified regression model for each comparison group. Interaction terms were not included. Because the dependent variable was expressed as a binary classifier (high vs. low utilization), we used logistic regression [24] to construct the risk prediction model. For each candidate predictor we calculated the Variance Inflation Factor (VIF) resulting from the regression of that variable on the other candidate predictors. None of the VIF values exceed 3.8, substantially less than the standard rule of thumb that a VIF of 10 or greater signals instability in the regression coefficients [25]. In addition, we examined influence plots from the final model to see if individual cases exerted extreme influence on the regression coefficients, identifying no remarkable observations.

The discrimination of the resulting prediction was evaluated by testing the predicted outcome against the actual outcome (FY2013 charges over $94,895 or not) for the entire study population of 9,581 patients. Discrimination was defined as the c-statistic, or the area under the receiver operating characteristic curve of sensitivity versus one-minus-specificity [26]. Each comparison group (lowest non-zero 5%, 10%, 15%, 20%, 30%, 40%, and 50%) was contrast-mined independently against the 5% of patients with highest FY2013 charges, and the resulting models were tested independently. The attributes common to all these models also were used to derive a combined model using all FY2013 observations, which was also tested independently.

4. Results

Contrast mining of 19,014 clinical attributes from the first year of EHR data for 479 high-utilization patients and comparison groups with low-utilization patients (ranging from the lowest 5% to the lowest 50%) identified 5,188 attribute combinations frequently found (support of 20% or more) in patients with high utilization in the second year, but infrequently in other patients (▶Table 2). Not all combinations were infrequent in all comparison groups, but at least 5,178 of the 5,188 were found in all seven contrast mining analyses. These 5,188 contrasting combinations were made up of 67 unique attributes (▶Table 3). Logistic regression of the 67 attributes found eleven attributes to be significantly ($p<0.05$) associated with high utilization (▶Table 4). The elven attributes comprised four diagnoses (i.e., depressive disorder, essential hypertension, ischemic heart disease, and osteoarthrosis), one demographic attribute (i.e., obesity), and six prescription types (i.e., anti-infectives,

benzodiazepines, beta-adrenergic blocking agents, quinolones, respiratory agents, and selective serotonin reuptake inhibitor antidepressants).

The c-statistic of the resulting model was 0.8436, with a 95% confidence interval of (0.8227, 0.8645). By assuming sensitivity and specificity errors to be equally important, an optimal threshold for the model was calculated to minimize the distance to the upper left corner of the operator receiving characteristic graph (▶Figure 3). This distance was calculated as

$$\sqrt{(1 - sensitivity)^2 + (1 - specificity)^2}$$ [27]

and tuning the model to this threshold produced a sensitivity of 0.770, a specificity of 0.812, a positive predictive value of 0.202, and a negative predictive value of 0.983. Please refer to the second paragraph of the "Primary Findings" (Section 5.1, below) for an interpretation of these measures.

5. Discussion

5.1 Primary Findings

A novel, two-step combination of EHR data mining with multiple logistic regression yielded a manageably small number of clinical attributes, which accurately predicted the 5% of patients who incurred nearly 50% of healthcare expenses. The model presented here has the virtue of simplicity and interpretability while still achieving an area under the ROC curve of 0.84, markedly higher than ROC value of 0.7 reported in comparable models [10]. Although adding interaction effects and nonlinear effects of continuous variables (e.g., age) to the logistic model might slightly improve this already reasonably high accuracy, it would come at the cost of a more complex model that might impede clinical interpretation. We felt that this model performed adequately without the added complexity, and demonstrated the methodology using unrestricted EHR data.

While the positive predictive value of 20% and negative predictive value of 98% appear low and high, respectively, they are reasonably useful given a population in which only 5% of patients are truly positive for high cost, and 95% of patients are negative. For example, a positive predictive value of 20% would result in five patients receiving the intervention of care management for every patient actually destined to incur high costs without intervention. This over-treatment penalty may be reasonable because care management is both extremely safe and relatively inexpensive, and because the 98% negative predictive value of the model would direct population health managers away from nearly all patients who will not incur the highest 5% of costs without the intervention.

These examples demonstrate the utility of mining the rich data available in the EHR to predict the small number of patients who will incur the majority of healthcare expenses, which support population health managers in focusing preventive and longitudinal care more effectively. This could support the Triple Aim [1] by improving health outcomes (for example, improving blood sugar control or blood pressure control in high-risk patients), improving healthcare delivery (for example, proactively reaching out to patients with unmet health management needs), and reducing costs (for example, using earlier lower-cost interventions such as frequent outpatient visits to reduce expensive inpatient stays).

All of the four diagnoses found to be associated with high utilization are among the ten most expensive medical conditions in the U.S. in 2013 [28]: (a) ischemic heart disease (second most expensive), (b) depression (third), (c) osteoarthrosis (fifth), and (d) hypertension (eighth). Of the prescription types found to be associated with high utilization, beta-adrenergic blocking agents may be indicative of ischemic heart disease (second most expensive); benzodiazepines may be indicative of depression (third), and respiratory agents may be indicative of chronic obstructive pulmonary disease (sixth). The partial congruence of the sample model with the medical conditions known to be most expensive validates the generalizability of these findings, while demonstrating the potential for other, novel discoveries (i.e., a nearly ten-fold increase in the odds of high costs associated with obesity, increased risks associated with anti-infectives in general and quinolones specifically, and risk reduction associated with SSRI antidepressants).

This sample prediction model for high healthcare utilization, or similar models derived using the same methodology, may be more suitable for secondary prevention than primary prevention since

many of the associated attributes are chronic conditions or therapeutics. For example, identification of hypertension and obesity as risk factors for high utilization should alert population health managers to monitor blood pressure and body weights more closely in high-risk patients, or review their medications more often. This method would also be applicable to disease-specific models or to other outcomes of interest, such as inpatient, emergency, or outpatient charges considered separately. Multiple models could be created from the same algorithm by limiting the population sample (for example, to patients with diabetes or those with hyperlipidemia) or by excluding some attributes which may not be interesting or may not be actionable (for example, excluding patients with high prior-year costs, or testing demographics and diagnoses but ignoring prescriptions).

The coefficients of the final regression model can be used to calculate a relative score [29] for all patients in a population (▶Table 4). This score gives an approximate relative risk of high utilization in the upcoming year, and patient interventions could be prioritized by ranking these scores. Alternatively, clinical alerts could be triggered for patients with scores exceeding a given threshold. By adjusting the threshold of the scoring system, the sensitivity and specificity of the model could be tuned to identify only as many high-risk patients as can be managed. However, because population health management is a low-risk and relatively low-cost intervention, clinical applications may benefit from greater sensitivity even at the price of lower specificity.

Some common attributes (e.g., gender=female, gender=male, race= white-non-Hispanic, or age=65-84 in this population) were found in attribute combinations associated with high utilization, but they clearly were not independent predictors of high utilization since they also were found in attribute combinations associated with low utilization or not predictive of utilization. This may explain why no demographic attributes other than obesity were identified in the final model. It is surprising that age and high prior-year costs were not significant predictors of high utilization, and these attributes may be found to be predictive in other populations.

Dissecting the associated combinations into separate attributes yields more robust predictors by generalizing the specific combinations of attributes found in a given population, reducing the number of rules (from thousands to tens, in this case), and testing the combined effects of the attributes by traditional statistical methods to identify the significant predictors

5.2 Limitations

While data mining techniques other than contrast mining can be used to discover associations with continuous outcomes, the focus of this demonstration was on a policy-relevant binary outcome: "high cost" and "not high cost," based on the well-supported contrast between patients in the higher 5% and lower 95% of costs [8]. Multivariate regression is not limited to binary outcomes, however, and linear regression on actual charges could have been used to describe or predict the central portion of the cost distribution.

Because this was a single-system study, the generalizability of these results to other populations is not clear. Predicting high hospital and clinic utilization reflects an important outcome of interest, but may exclude some patients who died in the second year before incurring charges high enough to exceed the measurement threshold. Furthermore, at the time these data were gathered, the University of Missouri EHR was undergoing a transition from ICD-9 to SNOMED coding. Since the same disease may have been recorded with an ICD-9 code in some patient records and a SNOMED code in others, the predictive power of some diseases may have been split between two diagnosis codes that were unrecognized synonyms. Lastly, hospital charges were used as a proxy for healthcare costs, but claims data would be a more accurate source of cost information.

5.3 Future Research

The implementation of these predictors as clinical alerts would allow quantitative and qualitative measurement of their clinical impact, in order to test the hypothesis that this predictive methodology can facilitate more efficient deployment of preventive and longitudinal care. Comparing these results to prior literature would help determine their clinical utility, and future studies might also survey expert clinical opinion as to the utility of these predictors of high utilization in population management. In addition, it would be useful to duplicate this method with other patient populations,

with higher and lower support values for "frequent" associations, and with expanded data sources including geospatial and socioeconomic attributes.

Further studies are also needed to incorporate Medicare and Medicaid claims data for the LIGHT[2] enrollees during the measurement period and to expand the attribute set with socio-economic status attributes, second-order attributes such as number of co-morbidities and poly-pharmacy, and intervention data such as nursing contacts and disease-management training.

6. Conclusions

A novel, two-step analysis of the electronic health records of 9,581 Medicare and Medicaid patients generated hypotheses with contrast mining and tested them with multiple logistic regression. This method yielded multiple similar models, each comprising a manageably small number of attributes that accurately predicted which patients would be in the 5% of patients with the highest healthcare utilization in the following year. The similarity of the models derived from varying comparison groups illustrate the flexibility and robustness of this approach. Because this method is not hypothesis driven, but draws predictors from the broader set of inputs available in a clinical EHR, it has the potential of discovering novel predictors, which may make it particularly useful in improving predictive discrimination over existing hypothesis-driven models. The method identified both expected and novel predictors including four diagnosis codes (i.e., depressive disorder, essential hypertension, ischemic heart disease, and osteoarthrosis), one demographic attribute (i.e., obesity), and six prescription types (i.e., anti-infectives, benzodiazepines, beta-adrenergic blocking agents, quinolones, respiratory agents, and SSRI antidepressants).

By predicting the small number of patients who will incur the majority of healthcare expenses, this method can support population health managers in focusing preventive and longitudinal care more effectively. This model, and similar models developed by combining contrast mining with logistic regression on readily available EHR data, could be used by population health managers to further the "Triple Aim" of better health outcomes, better healthcare delivery, and lower costs [1].

Questions

1. Your organization's CMIO (Chief Medical Information Officer) has asked you to propose informatics-based strategies for a new population health management program. How can population health informatics be employed to improve healthcare outcomes and costs?

A While informatics can support clinical decision support for individual patients, it does not apply to population health.

B Predictive analytics can support the transition from the traditional "reactive" model of medical care to one of avoiding preventable conditions.

C Web-based computerized diagnostic systems can be used to replace physicians for most health care delivery.

D The field of informatics is not mature enough to contribute to these organizational goals.

ANSWER: B. The Chronic Care Model [2] proposed improving the effectiveness of interactions between patients and providers as a way of promoting the "Triple Aim" of healthcare: better health, better care, and lower costs [1]. By bridging the implementation gaps in the Chronic Care Model, well-designed predictive analytics support the transition from the traditional "reactive" model of medical care [6] to one of maintaining health and avoiding preventable conditions [3]. Predictive analytics are potentially powerful tools for predicting population health outcomes [7].

2. You have been asked to choose between data analytic approaches for discovering actionable clinical predictors in electronic health records. Which of these strategies is likely to be useful?

A Logistic regression against the tens of thousands of fields in an electronic health record will reliably identify the few important predictors of outcomes and costs.

B Quantitative methods aren't needed, because qualitative methods such as surveys and focus groups will discover any important medical evidence.
C Combining contrast mining with multiple regression can produce a manageably small number of understandable and actionable rules.
D The field of informatics is not mature enough to contribute to these organizational goals.

ANSWER: C. Predictive analytics can be used to rapidly spot opportunities to improve care management [7], but regression analysis and other traditional statistical methods are constrained by the limited number of attributes that can be used [11]. However, a two-step process of data mining to reduce the number of candidate predictors followed by multiple regression to test the remaining candidates will permit the development of models that surpass the size restrictions of traditional statistical methods.

Clinical Relevance Statement
Accurate prediction of the 5% of patients who incur 50% of healthcare expenses is needed to permit population health managers to focus preventive and longitudinal care effectively. Combining contrast mining, which permits the use of the rich data available in the EHR, with testing by traditional statistical methods created flexible and highly accurate healthcare predictive analytics which can support population health management.

Conflicts of Interest
The authors have no conflict of interest to declare.

Protection of Human and Animal Subjects
This project was funded by the Center for Medicare and Medicaid Services (CMS) to expand the scope of services to a population of CMS beneficiaries, so the Health Sciences Institutional Review Board deemed the project to be a quality improvement initiative that did not require a formal patient consent process since the explicit purpose of data use was to improve patient care; the IRB number is 2001677-QI.

Acknowledgements
The authors would like to thank Hongfei Cao, PhD, for computational advice.

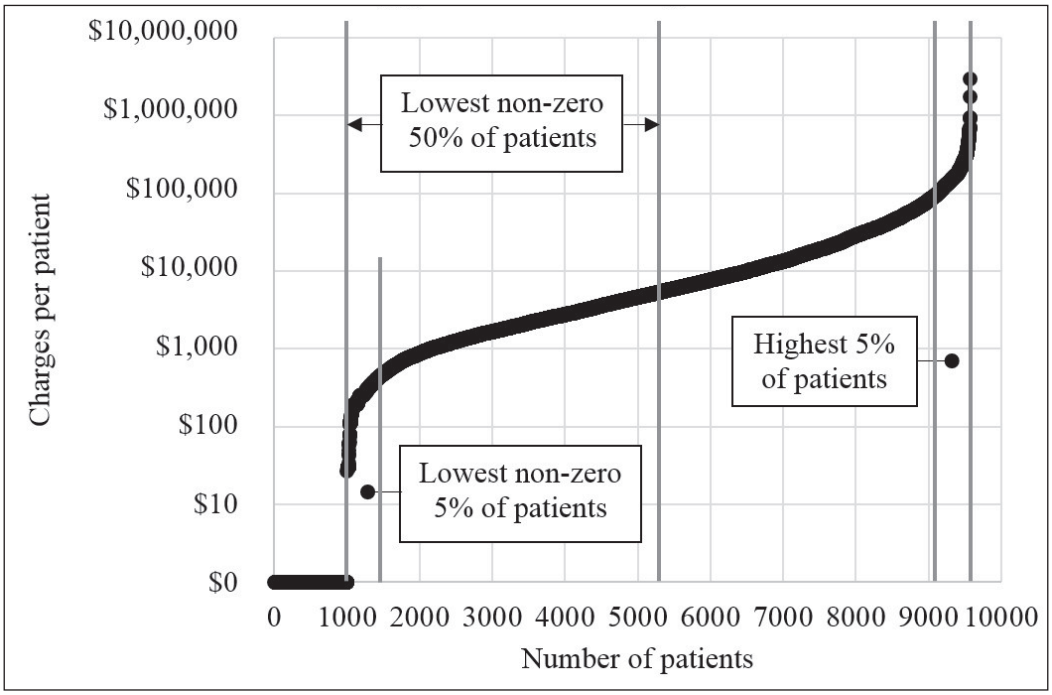

Fig. 1 Logarithmic distribution of FY2013 charges by patient

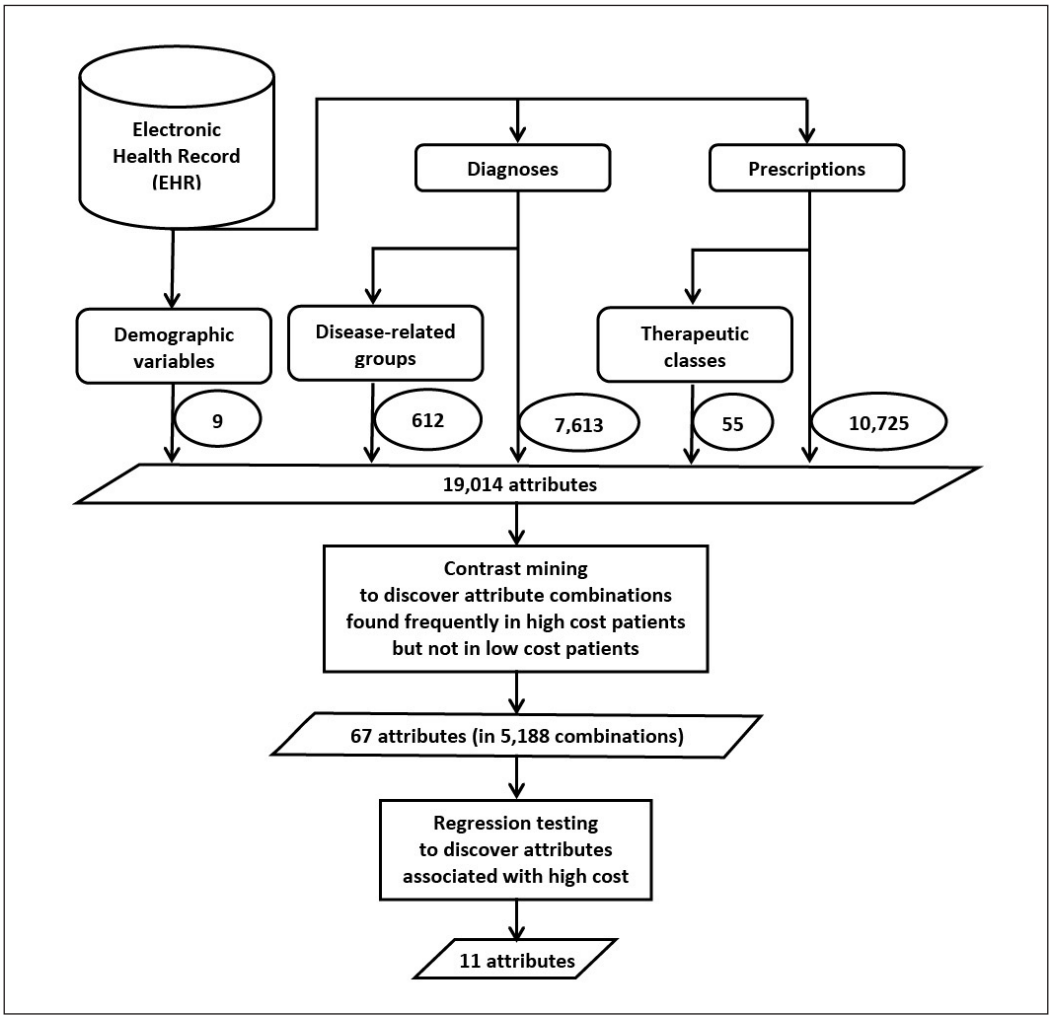

Fig. 2 Data selection, projection, and mining

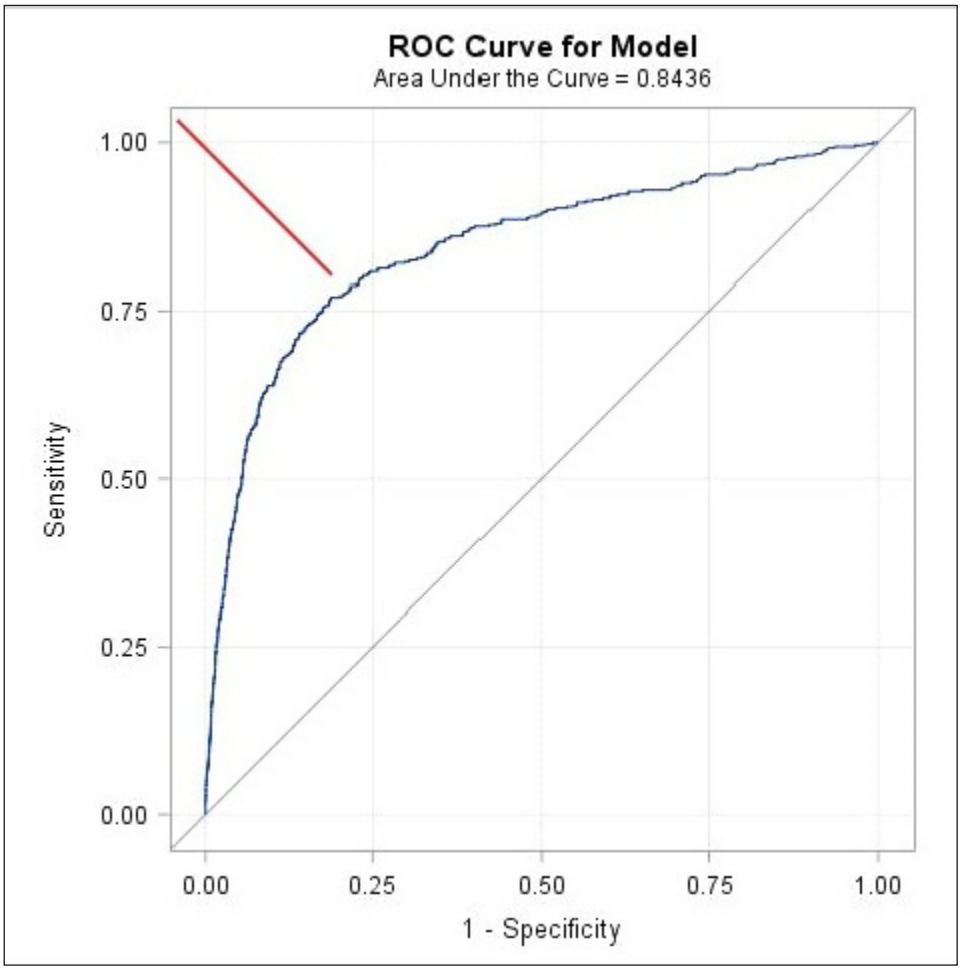

Fig. 3 Receiver operating characteristic (ROC) curve for the final model

Table 1 Comparison groups from patients with lowest non-zero charges in FY2013

	5%	10%	15%	20%	30%	40%	50%
Lowest charge in range				$27			
Highest charge in range	$470	$853	$1,221	$1,621	$2,646	$4,300	$6,963
Percentage of all charges	<0.1%	0.2%	0.5%	0.8%	1.8%	3.5%	6.1%

Table 2 Ten (out of 5,188) combinations frequently associated with high utilization

Attribute Combination	Support
Narcotic analgesics, Analgesics, Platelet aggregation inhibitors	0.21
Antihyperlipidemic agents, Analgesics, HMG CoA reductase inhibitors	0.39
Antidepressants, ICD9=311 (Depressive disorder), Antihistamines	0.20
Beta-adrenergic blocking agents, Cardioselective beta blockers, Nutritional products	0.29
Narcotic analgesics, Respiratory agents, Nutritional products	0.20
Race=White, Salicylates, Antiplatelet agents, Platelet aggregation inhibitors, Age=65to84	0.25
Antiplatelet agents, Analgesics, Beta-adrenergic blocking agents, Platelet aggregation inhibitors	0.33
Vitamins, Gastrointestinal agents, Salicylates, Nutritional products, Antiplatelet agents	0.20
Narcotic analgesics, Anxiolytics/sedatives/hypnotics	0.25
Narcotic/analgesic combinations, Gastrointestinal agents, Laxatives	0.23

Table 3 Individual attributes found in combinations associated with high utilization

Size of low-cost comparison group:	5%	10%	15%	20%	30%	40%	50%
Number of contrasting combinations:	**5178**	**5180**	**5188**	**5179**	**5179**	**5179**	**5179**
Age=25to44	X	X	-	-	-	-	-
Age=45to64	X	X	X	X	X	X	X
Age=65to84	X	X	X	X	X	X	X
Race/ethnicity=White/non-Hispanic	X	X	X	X	X	X	X
Female	X	X	X	X	X	X	X
Male	X	X	X	X	X	X	X
Obesity	X	X	X	X	X	X	X
Taking Rx as prescribed	X	X	X	X	X	X	X
Taking Rx not as prescribed	X	X	X	X	X	X	X
Medicaid	X	X	X	X	X	X	X
Prior High Cost	X	X	X	X	X	X	X
ICD9=250 (Diabetes mellitus)	X	X	X	X	X	X	X
ICD9=272.4 (Hyperlipidemia)	X	X	X	X	X	X	X
ICD9=311 (Depressive disorder)	X	X	X	X	X	X	X

Table 3 Continued

Size of low-cost comparison group:	5%	10%	15%	20%	30%	40%	50%
ICD9=401.1 (Benign essential hypertension)	X	X	X	X	X	X	X
ICD9=401.9 (Unspecified essential hypertension)	X	X	X	X	X	X	X
ICD9=414 (Ischemic heart disease)	X	X	X	X	X	X	X
ICD9=715 (Osteoarthrosis)	X	X	X	X	X	X	X
Adrenergic bronchodilators	X	X	X	X	X	X	X
Alternative medicines	X	X	X	X	X	X	X
Analgesics	X	X	X	X	X	X	X
Angiotensin converting enzyme inhibitors	X	X	X	X	X	X	X
Antiarrhythmic agents	X	X	X	X	X	X	X
Anticonvulsants	X	X	X	X	X	X	X
Antidepressants	X	X	X	X	X	X	X
Antidiabetic agents	X	X	X	X	X	X	X
Antiemetic antivertigo agents	X	X	X	X	X	X	X
Antihistamines	X	X	X	X	X	X	X
Antihyperlipidemic agents	X	X	X	X	X	X	X
Anti-infectives	X	X	X	X	X	X	X
Antiplatelet agents	X	X	X	X	X	X	X
Antipsychotics	X	X	X	X	X	X	X
Anxiolytics, sedatives and hypnotics	X	X	X	X	X	X	X
Benzodiazepine anticonvulsants	X	X	X	X	X	X	X
Benzodiazepines	X	X	X	X	X	X	X
Beta-adrenergic blocking agents	X	X	X	X	X	X	X
Bronchodilators	X	X	X	X	X	X	X
Calcium channel blocking agents	X	X	X	X	X	X	X
Cardioselective beta blockers	X	X	X	X	X	X	X
Cardiovascular agents	X	X	X	X	X	X	X
Dermatological agents	X	X	X	X	X	X	X
Diuretics	X	X	X	X	X	X	X
Gamma-aminobutyric acid analogs	X	X	X	X	X	X	X
Gastrointestinal agents	X	X	X	X	X	X	X
HMG CoA reductase inhibitors	X	X	X	X	X	X	X
Hormones/hormone modifiers	X	X	X	X	X	X	X
Iron products	X	X	X	X	X	X	X
Laxatives	X	X	X	X	X	X	X
Minerals and electrolytes	X	X	X	X	X	X	X
Miscellaneous analgesics	X	X	X	X	X	X	X
Miscellaneous anxiolytics, sedatives and hypnotics	X	X	X	X	X	X	X

Table 3 Continued

Size of low-cost comparison group:	5%	10%	15%	20%	30%	40%	50%
Muscle relaxants	X	X	X	X	X	X	X
Narcotic/analgesic combinations	X	X	X	X	X	X	X
Narcotic analgesics	X	X	X	X	X	X	X
Nonsteroidal anti-inflammatory agents	X	X	X	X	X	X	X
Nutraceutical products	X	X	X	X	X	X	X
Nutritional products	X	X	X	X	X	X	X
Platelet aggregation inhibitors	X	X	X	X	X	X	X
Proton pump inhibitors	X	X	X	X	X	X	X
Quinolones	X	X	X	X	X	X	X
Respiratory agents	X	X	X	X	X	X	X
Salicylates	X	X	X	X	X	X	X
Skeletal muscle relaxants	X	X	X	X	X	X	X
SSRI antidepressants	X	X	X	X	X	X	X
Thiazide and thiazide like diuretics	X	X	X	X	X	X	X
Vitamin and mineral combinations	X	X	X	X	X	X	X
Vitamins	X	X	X	X	X	X	X

Table 4 Regression model of attributes significantly ($p < 0.05$) associated with high utilization

Attribute	Coefficient	p-value	Odds Ratio	95% Confidence Limits	
Diagnoses					
ICD9=311 depressive disorder	0.5568	<0.0001	1.707	1.343	2.168
ICD9=401.9 unspecified essential hypertension	0.3967	0.0007	1.423	1.128	1.795
ICD9=414 ischemic heart disease	0.5939	<0.0001	1.828	1.386	2.411
ICD9=715 osteoarthrosis	1.0479	<0.0001	2.769	2.192	3.499
Demographic Attribute					
Obesity (BMI ≥ 30)	2.3520	<0.0001	9.496	7.530	11.976
Prescription Types					
Anti-infectives	0.4136	0.0060	1.504	1.117	2.025
Benzodiazepines	0.2975	0.0139	1.307	1.026	1.665
Beta-adrenergic blocking agents	0.2832	0.0148	1.314	1.047	1.649
Quinolones	0.4916	0.0087	1.674	1.158	2.421
Respiratory agents	0.3030	0.0063	1.340	1.076	1.668
Selective serotonin reuptake inhibitor (SSRI) antidepressants	-0.4062	0.0019	0.655	0.506	0.847

* Intercept = –4.2585 with $p < 0.0001$

References

1. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008; 27(3): 759-769.
2. Wagner EH, Glasgow RE, Davis C, Bonomi AE, Provost L, McCulloch D, Carver P, Sixta C. Quality improvement in chronic illness care: a collaborative approach. Jt Comm J Qual Improv 2001; 27(2): 63–80.
3. Glasgow RE, Orleans CT, Wagner EH. Does the chronic care model serve also as a template for improving prevention? Milbank Q 2001; 79(4): 579–612, iv-v.
4. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness: the chronic care model, Part 2. JAMA 2002; 288(15): 1909-1914.
5. Coleman K, Austin BT, Brach C, Wagner EH. Evidence on the Chronic Care Model in the new millennium. Health Aff (Millwood) 2009; 28(1): 75–85.
6. Snyderman R, Williams RS. Prospective medicine: the next health care transformation. Acad Med 2003; 78(11): 1079-1084.
7. Bradley P. Predictive analytics can support the ACO model. Healthc Financ Manage 2012; 66(4): 102-106.
8. Cohen SB, Uberoi N, United States Agency for Healthcare Research and Quality. Differentials in the concentration in the level of health expenditures across population subgroups in the US, 2010. Rockville: Agency for Healthcare Research and Quality; 2013.
9. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Aff (Millwood) 2014; 33(7): 1148-1154.
10. O'Caoimha R, Cornallya N, Weathersa E, O'Sullivana R, Fitzgeralda C, Orfilad F, Clarnettee R, Paúlf C, Molloya DW. Risk prediction in the community: A systematic review of case-finding instruments that predict adverse healthcare outcomes in community-dwelling older adults. Maturitas 2015; 82(1): 3–21.
11. Kantardzic M. Data mining: Concepts, models, methods, and algorithms. 2nd ed. Hoboken: John Wiley & Sons; 2011.
12. Khalid JM, Raluy-Callado M, Curtis BH, Boye KS, Maguire A, Reaney M. Rates and risk of hospitalisation among patients with type 2 diabetes: retrospective cohort study using the UK General Practice Research Database linked to English Hospital Episode Statistics. Int J Clin Pract 2014; 68(1): 40-48.
13. Sun J, McNaughton CD, Zhang P, Perer A, Gkoulalas-Divanis A, Denny JC, Kirby J, Lasko T, Saip A, Malin BA. Predicting changes in hypertension control using electronic health records from a chronic disease management program. J Am Med Inform Assoc 2014; 21(2): 337-344.
14. Hassanpour S, Langlotz CP. Predicting High Imaging Utilization Based on Initial Radiology Reports: A Feasibility Study of Machine Learning. Acad Radiol 2016; 23(1): 84-89.
15. Chechulin Y, Nazerian A, Rais S, Malikov K. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). Healthc Policy 2014; 9(3): 68–79.
16. Dove HG, Duncan I, Robb A. A prediction model for targeting low-cost, high-risk members of managed care organizations. Am J Manag Care 2003; 9(5): 381-389.
17. Gildersleeve R, Cooper P. Development of an automated, real time surveillance tool for predicting readmissions at a community hospital. Appl Clin Inform 2013; 4(2): 153-169.
18. Wright A, McCoy A, Henkin S, Flaherty M, Sittig D. Validation of an Association Rule Mining-Based Method to Infer Associations Between Medications and Problems. Appl Clin Inform 2013; 4(1): 100-109.
19. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques, 2nd Ed. San Francisco: Morgan Kaufmann; 2005.
20. Dong G. Preliminaries. In: Dong G, Bailey J, editors. Contrast data mining: concepts, algorithms, and applications. Boca Raton: CRC Press; 2013. p. 8.
21. Shvachko K, Kuang H, Radia S, Chansler R. The Hadoop distributed file system. Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. New York: Institute of Electrical and Electronics Engineers; 2010.
22. Health Data Interactive. Atlanta: Centers for Disease Control and Prevention; c2016 [updated 2016 May 16, cited 2016 May 26]. Available from: http://www.cdc.gov/nchs/hdi.htm.
23. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases 1994; 487-499.
24. Cox, DR. The regression analysis of binary sequences (with discussion). J Roy Stat Soc B 1958; 20: 215–242.
25. Myers RH. Classical and Modern Regression with Applications, Second Edition. Boston: PWK Kent; 1990.
26. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143(1): 29–36.
27. Zhou Xh, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. New York: John Wiley and Sons; 2002.

28. Total Expenses and Percent Distribution for Selected Conditions by Type of Service: United States, 2013. Rockville: Agency for Healthcare Research and Quality; c2016 [updated 2016 May 26, cited 2016 May 26]. Available from: http://meps.ahrq.gov/mepsweb/data_stats/tables_compendia_hh_interactive.jsp?_SERVICE=MEPSSocket0&_PROGRAM=MEPSPGM.TC.SAS&File=HCFY2013&Table=HCFY2013_CNDXP_C&_Debug=
29. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York: Cambridge University Press; 2007.