A Healthcare Utilization Analysis Framework for Hot Spotting and Contextual Anomaly Detection

Jianying Hu, Ph.D., Fei Wang, Ph.D., Jimeng Sun, Ph.D.
Robert Sorrentino, M.D., and Shahram Ebadollahi, Ph.D.
IBM T. J. Watson Research Center, New York, U.S.

Abstract

Patient medical records today contain vast amount of information regarding patient conditions along with treatment and procedure records. Systematic healthcare resource utilization analysis leveraging such observational data can provide critical insights to guide resource planning and improve the quality of care delivery while reducing cost. Of particular interest to providers are *hot spotting*: the ability to identify in a timely manner heavy users of the systems and their patterns of utilization so that targeted intervention programs can be instituted, and *anomaly detection*: the ability to identify anomalous utilization cases where the patients incurred levels of utilization that are unexpected given their clinical characteristics which may require corrective actions. Past work on medical utilization pattern analysis has focused on disease specific studies. We present a framework for utilization analysis that can be easily applied to any patient population. The framework includes two main components: utilization profiling and hot spotting, where we use a vector space model to represent patient utilization profiles, and apply clustering techniques to identify utilization groups within a given population and isolate high utilizers of different types; and contextual anomaly detection for utilization, where models that map patient's clinical characteristics to the utilization level are built in order to quantify the deviation between the expected and actual utilization levels and identify anomalies. We demonstrate the effectiveness of the framework using claims data collected from a population of 7667 diabetes patients. Our analysis demonstrates the usefulness of the proposed approaches in identifying clinically meaningful instances for both hot spotting and anomaly detection. In future work we plan to incorporate additional sources of observational data including EMRs and disease registries, and develop analytics models to leverage temporal relationships among medical encounters to provide more in-depth insights.

1 Introduction

The purposes of analysis of patient healthcare resource utilization patterns include resource planning, allocation and the evaluation of the appropriateness, medical needs and efficiency of health care service and procedures. Such analysis is of increasing importance for health care institutions to ensure effective and efficient patient care delivery. Patient medical records today include a large number of entries related to patient conditions along with treatments and procedures received. Utilization analysis based on such observational data collected through normal course of care delivery and carried out in a systematic manner can be leveraged to improve care delivery in many ways. Two areas in particular have attracted significant attention recently. The first is the notion of *hot spotting*, which is the ability to identity in a timely manner patients who are heavy users of the system and their patterns of use, so that targeted intense intervention and follow up programs can be put in place to address their needs and change the existing, potentially ineffective, utilization pattern [9]. The second is anomaly detection, where the goal is to identify utilization patterns that are unusual given patients' clinical characteristics, including both underutilization and overutilization. The former may indicate a gap in medical service that if left unaddressed could result in further deterioration of patient's condition leading to situations requiring more costly and less effective interventions. The latter incurs unnecessary cost and waste of precious healthcare resources that could have been directed towards cases in real need. Estimates has put the waste caused by overutilization at more than 30% of the total medical cost and this has been confirmed by real world medical management experiences [1].

This paper describes a new framework for utilization analysis designed to address these needs. The framework includes two main components. The first component is Utilization Profiling and Hot Spotting, where we use a vector space model to represent patient utilization profiles, and apply clustering techniques to identify dominant utilization groups within a given population. Hot spotting can then be performed by analyzing small and isolated high utilization groups. The second component is Contextual Anomaly Detection for Utilization. Typical anomaly detection methods focus on identifying data instances that deviate from the majority of the samples [6]. However for healthcare utilization anomaly detection the context provided by the patient's clinical characteristics is extremely important. A given utilization instance may be perfectly normal for one patient, but unexpected for another patient with different clinical conditions such as comorbidities. We propose novel methods for contextual anomaly detection designed to detect utilization anomalies in such settings. Our method is based on building models trained from observational data to compute the expected utilization levels for each patient given his/her clinical and demographic characteristics, and then examining the difference between the expected and actually levels based on well established statistical methods.

The proposed approaches have been evaluated using out-patient data for a population of 7667 diabetes patients collected over a one year period. The main contribution of this paper is the adaptation and integration of advanced machine learning techniques into an important care management application that can be used to perform systematic utilization analysis on any given patient population, to identify clinically meaningful cases of heavy utilization as well as anomalous utilization. Such insights can potentially assist care providers achieve better resource allocation and better management of gaps and opportunities in care, leading to improved patient outcomes at reduced cost in the long run. It's worth noting that utilization anomalies could also be indicators of potential medical fraud that require further investigation, however this aspect will not be the focus of this paper.

2 Background

Existing work on medical utilization pattern analysis has focused on disease specific studies and has not directly proposed a general framework for addressing the issues of hot spotting or anomaly detection. For example, Barsky et al. introduced a clustering method to detect medical care utilization patterns for somatizing patients [2]. Nicholson et al. conducted research on patterns of ambulatory care use for gynecologic conditions [18]. Eisele et al. studied the ambulatory medical care utilization patterns before and after the diagnosis of dementia in Germany [7]. Ruchlin et al. learned the resident medical care utilization patterns in continuing care retirement communities [20]. Bushche et. al. analyzed ambulatory medical care utilization by elderly patients in relation to patient conditions in Germany [5]. While these past studies each shed valuable light on the factors affecting the pattern of utilization in a specific disease condition, they were not designed to provide systematic approaches that can be adopted for routine utilization analysis on any given patient population.

Anomaly detection as a general topic has been studied for wide ranging domains including financial fraud detection, industrial damage detection, social media analysis, and medical and public health anomaly detection [6]. Of particular relevance to medical utilization analysis are two main types of anomaly detection, namely Point Anomalies and Contextual Anomalies.

Point anomalies refer to cases where an individual data instance (e.g., number of visits of different types) can be considered as anomalous to the rest of the data. This is the simplest form of anomaly and is the focus of majority of the research on anomaly detection in general. Most of the prior work in medical domain falls into this category [19, 25, 15, 22]. The utilization profiling and hot spotting component of our proposed framework can be considered an instance of this type of anomaly detection using a clustering based approach (other common methods include classification based, statistical techniques, information theory based, etc.) [6]. In the clustering based approach, singleton clusters as well as small clusters are considered point anomalies, based on the operating assumption that normal data instances belong to large, dense clusters, while anomalies belong to small, isolated ones .

Contextual anomalies are more complex cases where a data instance is anomalous in a specific context (e.g., given the particular characteristics for a patient), but not otherwise, hence cannot be detected using methods designed for point anomalies. In order to detect contextual anomalies, each data instance has to be defined using two sets of attributes:

(1) Contextual attributes. The contextual attributes are used to determine the context for that instance, for example, patient characteristics such as comorbidities.

(2) Behavioral attributes. The behavioral attributes are used to determine the non-contextual characteristics of an instance, for example, number of clinical visits of various types (primary care physician, specialists, emergency room etc.)

The goal of contextual anomaly detection is to determine whether a particular value for the behavioral attributes is unusual within a specific context. Methods for contextual anomaly detection are particularly valuable in medical utilization analysis as they provide more comprehensive indicators by evaluating the utilization profile of each patient in the context of what is expected for patients with similar characteristics. In doing so they can uncover important anomalies for further investigation that would remain undetected using point anomaly detection methods.

We introduce a new approach to contextual anomaly detection in the context of medical utilization analysis based on expectation modeling, using regression models. Similar type of regression model based approach has been applied before to anomaly detection in time series [6], but to our knowledge has never been applied to medical utilization analysis. The most closely related work in medical domain is by Hauskrecht and colleagues [12, 11]. They studied methods for conditional anomaly detection to flag potential medical error by identifying medical actions that are unusual with respect to past patients and their conditions. However their methods are classification based and designed to assess the appropriateness of specific medical actions as binary variables. Such methods cannot be easily extended to the study of general muti-dimensional utilization patterns with arbitrary values.

3 Methods

The proposed utilization analysis framework contains two main components. In the first component a multidimensional clustering method is applied to segment a given patient population into groups with similar utilization profiles as characterized by numbers of clinical visits of different types. The purpose of the clustering analysis is two fold: 1) to identify the dominant utilization patterns within the population, and 2) to identify groups of high utilizers and understand their general characteristics. In contrast to patient stratification based on total cost or a particular type of utilization (e.g. emergency visits), such multidimensional clustering analysis provides a more comprehensive understanding of the characteristics of different types of utilization.

In the second component we apply the concept of contextual anomaly detection to this domain and develop an expectation modeling based approach to identify patients with anomalous utilization records. The basic idea is as follows. First a regression model is trained from the observational data of recorded patients characteristics and corresponding utilization profiles. This model is then used to calculate the expected utilization level (behavior) with respect to any given patient profile (context). Then a comparison is made between the observed and expected behaviors and the Grubb's test which is widely used for anomaly detection [10, 6] is deployed to determine whether there is an anomaly. In the following we describe each component in detail.

3.1 Utilization Profiling and Hot Spotting

We use p_i to indicate the i-th patient. The whole patient population set is denoted by $\mathcal{P} = \{p_1, p_2, \cdots, p_n\}$, where n is the number of patients. The patient's utilization is characterized by the number of different types of visits (e.g., visit to Primary Care Physician (PCP), visit to specialist, lab visit, etc.) incurred by this patient during a certain time period. We use v_i to denote the visit of type i, and $\mathcal{V} = \{v_1, v_2, \cdots, v_d\}$ to denote the set of different types of visits (suppose there are totally d types of visits). Then the utilization of patient p_i is represented as a d-dimensional vector $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{id}]^\top$, where x_{ij} is the number of v_j visits incurred on patient p_i. We call $\mathbf{x}_i$ the *utilization profile* of patient p_i. For example, if the utilization profile of a patient is $[0, 1, 3, 0, 0]^\top$, this means that there are totally $d = 5$ types of visits, and the patient had 1 v_2 type visit as well as 3 v_3 type visits.

Once each patient is represented as a vector in a multi-dimensional space, a variety of clustering algorithms can be applied to segment the patient population into cohorts of patients with similar utilization profiles. To choose the most appropriate algorithm the following factors need to be taken into consideration:

- *Interpretability*. We not only want to identify the patient cohorts, but also want to understand how those patients are grouped together.
- *Stability*. We want the algorithm to be stable such that clustering assignments do not change much against small parameter and/or data perturbation.
- *Scalability*. As our goal is to provide a general tool for patient utilization analysis, where we may encounter a large patient population, it is important for the approach to be capable of handling large scale data.

In general, data clustering algorithms can be classified into two categories: *partitioning methods* and *hierarchical methods* [13]. Partitioning methods formulate clustering as an optimization problem, which makes it performance-driven. Here performance measure can be cluster compactness as used in K-means [16], normalized graph cuts [21] or the margins between different clusters [26]. However, these methods are usually not stable. Furthermore, as they focus on algorithm performance rather than interpretability, it is often difficult to interpret the cluster procedure and results. Based on these considerations, we decided to adopt one of the most representative hierarchical methods: *Hierarchical Agglomerative Clustering* (HAC) [13], which merges the data vectors one by one in a bottom-up manner according to a distance metric. As a by-product, HAC generates an easy to explore dendrogram explaining the whole clustering procedure, which makes it very convenient for users to control and investigate the clustering results.

One potential problem of HAC is its scalability, as it relies on pairwise data distances. This results in an at least $O(n^2)$ computational complexity. To alleviate this problem, we borrow an idea from image segmentation and develop a hybrid two-stage HAC algorithm. In image segmentation, a common approach for handling large number of pixels is to first aggregate them into a set of homogeneous "superpixels", and then perform clustering on those superpixels [23]. Similarly, in the first stage of our approach, we segment the patient population into a large number of micro clusters, then in the second stage, we perform HAC on these micro clusters. The algorithm flow is illustrated in Fig.1, where each blue dot corresponds to a patient, the outside green circle represents the utilization profile vector space. We use red shaded areas to indicate patient clusters, and red dots to denote cluster representatives (e.g., cluster mean) .

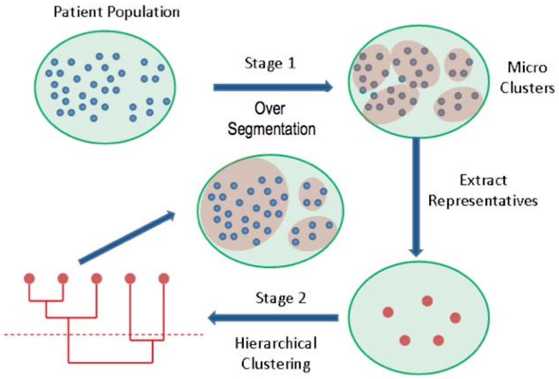


Figure 1: Two-stage clustering for large patient population. The whole population is first over segmented to a set of micro-clusters, then the mean of each micro-cluster is extracted as representatives for further clustering using HAC.

Many efficient partition based methods could be used for the over-segmentation stage. We chose to use *Classification And Regression Tree* (*CART*) [3] algorithm to take advantage of the fact that cost information is typically available in utilization data, and patients with very similar utilization profiles should also have very similar cost. The vectors representing patients' utilization profiles are treated as input features and used to predict cost as the target variable. The CART algorithm constructs a rule based decision tree to segment the patient set by recursively partitioning the feature space until the patients within each partition satisfy certain purity constraint (based on cost). The final partitions correspond to the leaf nodes of the tree.

At the end of the first stage, the mean utilization profile from each micro-cluster is extracted and treated as the cluster representative. Then in the second stage we cluster these representatives with HAC. As stated above, HAC starts with every representative as a cluster, and merge them step by step. At each step, two nearest clusters are merged. Here the distance of two clusters $\mathcal{P}_i$ and $\mathcal{P}_j$ is measured by

$$d(\mathcal{P}_i, \mathcal{P}_j) = \frac{1}{n_i n_j} \sum_{p_k \in \mathcal{P}_i} \sum_{p_l \in \mathcal{P}_j} \|\mathbf{x}_k - \mathbf{x}_l\| \tag{1}$$

where n_i, n_j are the sizes of $\mathcal{P}_i, \mathcal{P}_j$, and $\|\mathbf{x}_k - \mathbf{x}_l\| = \sqrt{(\mathbf{x}_k - \mathbf{x}_l)^\top (\mathbf{x}_k - \mathbf{x}_l)}$ is the *Euclidean distance* between $\mathbf{x}_i$ and $\mathbf{x}_j$. $\mathbf{x}_k, \mathbf{x}_l$ are utilization profiles of p_k, p_l. It can be easily seen that this is in fact the average distance between all pairs of data points with one in $\mathcal{P}_i$ and one in $\mathcal{P}_j$.

3.2 Contextual Anomaly Detection for Utilization

Our contextual anomaly detection approach consists of the following three steps. First, we learn functions that map clinical characteristics (contextual attributes) to utilization characteristics. These regression models are then used to estimate the expected number of visits of each type. Finally, a statistical test (Grubb's test [10]) is applied to check if a significant difference exists between expected and actual utilization levels.

In the first step, the contextual attributes include patient demographics (age and gender) and clinical features characterized by ICD-9 codes. One potential shortcoming of using ICD-9 codes alone is that they do not adequately reflect clinical relations and risk groupings among different diagnoses. Much past work on Health Risk Assessment (HRA) has deployed various methods of generating risk groups, and reported improved accuracy of healthcare cost prediction using such groupings [24]. We adopted the Hierarchical Condition Categories (HCC) used in Medicare Risk Adjustment provided by CMS (Centers for Medicare and Medicaid Services) in addition to the ICD-9 codes in the contextual attributes. For both ICD-9 codes and HCC codes, the clinical feature is defined as the percent of times that specific diagnosis was given in the utilization analysis period, which provides a measure of dominance of the corresponding condition for a patient. The target variables for the expectation models are the behavioral attributes, which in this case are the numbers of visits for the different utilization types. A separate expectation model is built independently for each utilization type.

For the regression model, we explored several advanced function learners:

- Classification And Regression Trees (CART) [3] is similar to a decision tree except at the leaf level a regression model is constructed in order to map to a continuous target variable, instead of doing a majority vote as in a decision tree classifier.
- Random Forest (RF) [4] is an ensemble version of CART where multiple CART trees are built on the bootstrapping samples of the entire patient set. Here the bootstrapping is uniform such that each data point has an equal opportunity to be sampled with replacement.
- Multivariate Adaptive Regression Splines [8] (MARS) a non-parametric regression technique and can be seen as an extension of linear models that automatically models non-linearities and interactions between variables.

We used 10-fold cross-validation to evaluate these different methods. In our experiments, Random Forest consistently outperform all other methods in all cases, and was thus adopted in our system.

Once the regression models have been trained, they are used to compute the expected level of utilization of each specific type for each patient given the contextual attributes of the patient. The difference between the expected and actual utilization (the residual error) can then be used to determine whether there is an anomaly. Intuitively, an actual utilization level should be declared anomalous if it deviates too much from the expected level. The key question to answer is: how much is too much? Certain utilization types may naturally have a wider range of variability associated with them than others and thus should be allowed larger deviation. We deploy Grubb's test which had been widely used in the anomaly detection literature [10, 6] to take into consideration this inherent variability.

The test statistic for the i-th patient of j-th type of utilization as

$$z_i^j = |r_i^j - \bar{r}^j|/s^j \tag{2}$$

where r_j^i is the squared error between the expected and the actual value of the j-th type of utilization$j$ on the patient i, $\bar{r}^j$ and s^j are the mean and standard deviation of r_j^i. Then the patient i is declared as anomaly on the j-th type of utilization if

$$z_j^i > \frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2_{\alpha/2n,n-2}}{n-2+t^2_{\alpha/2n,n-2}}} \tag{3}$$

Here n is the number of patients and $t_{\alpha/2n,n-2}$ is a threshold used to declare an instance to be anomalous or normal. This threshold is the value taken by a t-distribution at a significance level of $\alpha/2n$. The significance level measures the confidence associated with the threshold and indirectly controls the number of instances declared as anomalous. A patient is declared an anomaly if it is identified as an anomaly by the Grubb's test for at least one type of visits.

4 Results and Analysis: Diabetes Patient Management

4.1 Data description

The proposed framework has been tested using claims data collected from a network of physicians over a one year period. While the framework is very general and can be applied to any patient population, it is useful to focus on a specific use case to investigate whether the results are clinically meaningful. We thus constrained our experiments to the diabetes patient population which contains a total of 7,667 patients. For this population 98% of the total visits belong to one of the top six visit types as given in Table 1, which also provides statistics for these top visit categories. From the table it can be clearly observed that the majority of the patients had relatively low level of utilization. For example, the 50 percentile of the total number of visits is only 12, i.e., half of the patients only made up to 12 visits to medical facilities during the year.

Table 1: Descriptions and statistics of different types of visits

Visit Type	Description	#Visits	Summary Statistics			Percentiles		
			median	mean	std	50%	80%	95%
1	PCP visit	61,253	6	7.99	6.90	6	12	20
2	Specialist visit	77,255	6	10.08	15.88	6	15	32
3	Emergency visits	5,731	0	0.75	3.06	0	0	4
4	Outpatient hospital visits	34,047	0	4.44	13.46	0	6	18
5	Inpatient hospital visits	20,826	0	2.72	12.68	0	0	14
6	Patient's home	15,389	0	2.00	5.15	0	4	9

4.2 Results for Utilization Segmentation and Hot Spotting

The modified Hierarchical Agglomerative Clustering approach described in Section 3.1 was applied to this patient population. The resulting dendrogram can then by explored interactively by a domain expert. For each cluster, the user can examine the representative utilization profile of the cluster (computed as the cluster mean), average cost, and patient characteristics such as mean age, sex ratio and dominant diagnoses. For this diabetes population, a close examination by the MD in our group revealed that a total of 10 clusters provides a meaningful level of segmentation.

Figure 2 shows the representative utilization profile for each cluster (cluster mean). Table 2 shows for each cluster the cluster size, average cost, average age and a clinical description of the cluster derived by the MD based on information provided by the system as explained above. It can be clearly seen from this analysis that clusters 1-4 represent well managed patients with varying but stable conditions, leading to relatively low level of utilization and cost. Cluster 5, 6 and 8 represent patients with more advanced disease state and advancing complications, thus requiring increased utilization. Finally, clusters 7, 9 and 10 are the ”hot spot” patients with advanced conditions requiring intense utilization of different types. These are patients who will likely benefit from an intensive disease management program.

4.3 Results for Contextual Anomaly Detection

As described in Section 3.2, a separate expectation model was trained for each one of the top six utilization types using Random Forest regression model, using diagnoses, age and sex as the contextual attributes to predict the expected level of utilization. Table 3 shows the prediction results for each each one of the six utilization types using standard 10 fold cross validation. As can be seen in the table, a positive R^2 measure was achieved for all utilization types, including even Emergency visit, which is particularly difficult to predict because of the sparsity of the event, and large degree of randomness (e.g., accidents). For visit type involving less degree of randomness, the performance improves as expected. Particularly, for Specialist and Inpatient hospital visits we achieved R^2 values greater than 0.3. These results indicate that the proposed expectation model can indeed lead to better prediction of expected utilization level than using population mean, which should lead to more personalized and clinically meaningful anomaly detection.

For each patient the difference between the expected level and actual level of utilization is compared against the mean residual error and the Grubb's test is used to determine if this different is anomalous. A patient is considered anomalous if he/she is signaled as such for at least one of the utilizaton types. Using a significance level of 0.05 in Grubb's test, a total of 51 anomalies were detected. These anomalies can then be explored in the system by examining the actual vs. expected utilization levels, and contextual attributes including age, sex and dominant diagnosis to determine the next step of investigation. Here we provide sample investigations of three of the patients with anomalous utilizations. Figure 3 shows the expected vs. actual utilization for each patient, and Table 4 provides the characteristics

Table 2: Detailed Analysis of the 10 Utilization Clusters (see Fig. 2 for the mean utilization profiles of the clusters)

ID	Size	Ave. Cost/Patient	Ave. Age	Cluster Description
1	377	5,758	69	Cohort consists primarily of well-managed Diabetics with Hypertension, Hyperlipidemia and cardiac arrhythmias, with cost-effective use of Primary Care, Specialty Care and Outpatient Hospital Clinics, avoiding Hospitalizations and ER Visits
2	959	4,720	70	Cohort consists primarily of well-managed Diabetics with Hypertension, Hyperlipidemia and some cardiac disease with cost-effective use of Primary Care, requiring some Specialist visits while avoiding Outpatient Clinics, Hospitalizations and ER Visits.
3	807	2,580	66	Cohort of Diabetics with complications of Hypertension, Hyperlipidemia and some with cardiac arrhythmias. They make cost-effective use of Primary Care, while avoiding Specialist, Outpatient Clinics, Hospitalizations and ER Visits.
4	5013	1,573	63	Cohort of younger patients with uncomplicated Diabetics with Hypertension, Hyperlipidemia, making minimal use of services. This cohort is a target for interventions with preventative services
5	239	10,150	69	Cohort of Diabetics with Hypertension, Hyperlipidemia, cardiac arrhythmias, and arthritis, making extensive use of Specialists, while avoiding Hospitalizations and ER Visits.
6	127	11,738	69	Cohort of more advanced diabetics with increasing comorbities, and complications requiring periodic hospitalization for exacerbations of heart failure, pulmonary disease and renal failure.
7	3	66,480	63	Cohort of high-utilizing diabetics with advancing renal failure, cardiac and peripheral vascular disease. Leukemia is a significant coexisting comorbidity. These patients require frequent expensive hospital outpatient procedures and use of specialists.
8	112	16,375	66	Cohort of diabetics with advancing complications thereof, being managed by specialists. Higher representation of women. Outpatient hospital visits likely include treatment of peripheral vascular disease, vascular ulcers and bone infections.
9	14	42,559	71	Cohort of older diabetics likely to have a high percentage of smokers with costly complications including malignancies, COPD and complex outpatient treatment thereof.
10	16	41,980	57	Cohort of end-stage diabetics with advanced complications including renal failure, heart failure, poorly controlled hypertension, requiring frequent hospitalizations, outpatient hospital procedures, and home health visits.

of these patients along with investigation notes and recommendations from the MD on our team who examined the cases.

5 Conclusions

We present a novel framework for utilization analysis that can be used to perform systematic and timely identifications of heavy users of different types as well as contextual anomalies, i.e., utilization instances that are unexpected given patients' clinical characteristics. In order to assess the general applicability of the framework, in this initial exploration we restricted our experiments and analysis to the most widely available type of data, i.e., claims data including diagnosis, demographics, and medical utilization records. Our evaluations and case studies demonstrate the usefulness of the proposed approaches in identifying clinically meaningful instances for both hot spotting and anomaly detection, using the most basic observational data as described above. Clearly many other data sources such as EMRs and patient and disease registries could provide additional information relevant to utilization analysis. In our future work we plan to expand our framework to leverage these additional data sources to provide enhanced performance and additional actionable insight. Another limitation of the proposed methods is that we currently do not consider temporal relationships among different medical events or encounters. Exploration of such relationships could provide deeper context and more fine-grained utilization patterns as well as contextual attributes. Temporal event analysis in medical domain

Table 3: Prediction Performance of The Utilization Regression Model (Random Forest Regression)

	PCP	Specialist	Emergency	Inpatient Hospital	Outpatient Hospital	Patients' Home
RMSE	6.11	13.26	2.93	10.30	12.15	4.81
R^2	0.22	0.30	0.085	0.34	0.18	0.13

Table 4: Example Analysis of Detected Contextual Anomalies (see Fig. 3 for expected vs. actual utilization levels)

ID	Age	Top Cormobidities	Anomaly Analysis
6071	48	other symptoms involving abdomen and pelvis; heart failure; essential hypertension; obesity and other hyperalimentation; nondependent abuse of drugs; ill-defined descriptions and complications of heart disease; acute bronchitis and bronchiolitis	Diabetic with advancing complications including heart failure, thus requiring large number of specialist visits. Patient is underutilizing PCP and Specialists, and appears non-compliant with poor dietary control and is likely a smoker. This person over-utilizes the ER, possibly for both diabetic and respiratory complications, as well as drug-seeking behavior.
1311	67	acquired hypothyroidism; disorders of lipoid metabolism; asthma;nontoxic nodular goiter; essential hypertension;other disorders of urethra and urinary tract; other and unspecified anemias	Older diabetic, with prior thyroidectomy and resultant hypothyroidism, plus possible urinary stress incontinence, as well as poorly controlled asthma. The latter may be due to suboptimal medication regime and/or non-compliance, resulting in unexpectedly frequent PCP and Specialist visits, and hospital admissions from the doctor's office. Home health visits are likely for home respiratory therapy. Patient likely has been trained to contact her doctors instead of using the ER.
2815	59	other disorders of cervical region; symptoms involving head and neck; disorders of lipoid metabolism; other disorders of soft tissues; intervertebral disc disorders; other forms of chronic ischemic heart disease; gastrointestinal hemorrhage; cellulitis and abscess of finger and toe	likely a non-compliant diabetic with hyperlipidemia, cardiac disease and vascular disease leading to skin ulcers and infections. Cervical disc disease is a comorbidity. This patient over-utilizes specialists and the ER, most likely due to diabetic complications, and chronic pain related to cervical disc disease. Unexpected hospitalizations are likely due to complications related to non-compliance, and outpatient hospital visits may be related to antibiotic treatment of skin infections secondary to vascular disease and poor self-care. Alcohol abuse should also be investigated in light of the history of gastrointestinal hemorrhage

has been widely studied in the literature for discovery of medical knowledge and decision support [17, 14]. The incorporation and expansion of these methodologies for medical utilization analysis is another importance direction of our future work.

References

[1] http://leanmedicalcare.org/?p=85.

[2] A. J. Barsky, E. J. Orav, and D.W. Bates. Distinctive patterns of medical care utilization in patients who somatize. *Med Care.*, 44(9):803–811, 2006.

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* CRC Press, Boca Raton, 1984.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] H. V. D. Bussche, G. Schon, T. Kolondo, H. Hansen, K. Wagscheider, G. Glaeske, and D. Koller. Patterns of ambolatory medical care utilization in elderly patients with special reference to chronic diseases and multimorbidity - results from a clams data based observational study in germanuy. *BMC Geriatrics*, 11(1), 2011.

[6] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *Technical Report TR 07-017, Dept. of Computer Engineering, Univ. Minnesota*, 2007.

[7] M. Eisele, H. van den Bussche, D. Koller, B. Wiese, H. Kaduszkiewicz, W. Maier, G. Glaeske, S. Steinmann, K. Wegscheider, and G. Schön. Utilization patterns of ambulatory medical care before and after the diagnosis of dementia in germany–results of a case-control study. *Dement. Geriatr. Cogn. Disord.*, 29(6):475–483, Jul. 2010.

[8] J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.

[9] A. Gawande. The hot spotters. *New Yorker*, January 2011.

[10] F. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[11] M. Hauskrecht, M. Valko, I.Batal, G. Clermont, S. Visweswaran, and G. Cooper. Conditional outlier detection for clinical alerting. In *Proceedings of the Annual American Medical Informatics Association (AMIA) Symposium*, pages 286–290, 2010.

[12] M. Hauskrecht, M. Valko, B. Kveton, S. Visweswaram, and G. Cooper. Evidence-based anomaly detection in clinical domains. In *Proceedings of the Annual American Medical Informatics Association (AMIA) Symposium*, pages 319–323, 2007.

[13] A. K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Upper Saddle River, NJ, USA, 1988.

[14] N. Lee, A. Laine, J. Hu, F. Wang, J. Sun, and S. Ebadollahi. Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients. In *First IEEE International Conf. on Health Informatics, Imaging and Systems Biology*, 2011.

[15] Jessica Lin, Eamonn Keogh, Ada Fu, and Helga Van Herle. Approximations to magic: Finding unusual medical time series. In *In 18th IEEE Symp. on Computer-Based Medical Systems (CBMS)*, pages 23–24, 2005.

[16] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[17] M.R. Moskovitch and Y. Shahar. Medical temporal knowledge discovery vis temporal abstraction. In *AMIA Annual Symposium Proceedings*, pages 452–456, 2009.

[18] W. K. Nicholson, S. A. Ellison, H. Grason, and N. R. Powe. Patterns of ambulatory care use for gynecologic conditions: A national study. *Am. J. Obstet. Gynecol.*, 184(4):523–530, Mar. 2001.

[19] K. I. Penny and I. T. Jolliffe. A comparison of multivariate outlier detection methods for clinical laboratory safety data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(3):295–307, sep 2001.

[20] H. S. Ruchlin, S. Morris, and J. N. Morris. Resident medical care utilization patterns in continuing care retirement communities. *Health Care Financ Rev.*, 14(4):151–168, Summer 1993.

[21] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 888–905, 2000.

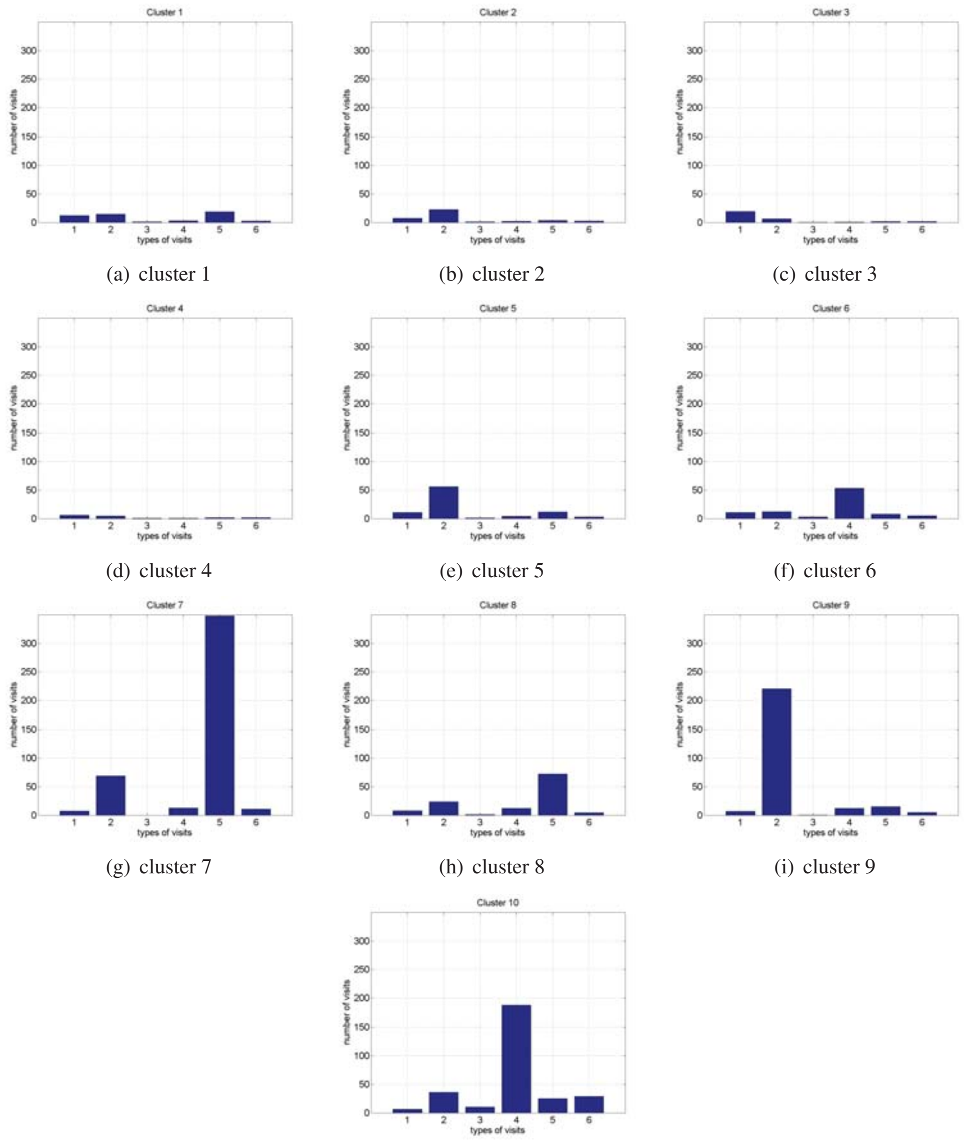
[22] Z. Syed, M. Saeed, and I. Rubinfeld. Identifying high-risk patients without labeled training data: Anomaly detection methodologies to predict adverse outcomes. In *Proceedings of the Annual American Medical Informatics Association (AMIA) Symposium*, pages 772–776, 2010.

[23] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *Proceedings of the 2010 European Conf. on Computer Vision (ECCV)*, pages 319–323, 2007.

[24] R. Winkelman and S. Mehmod. A comparative analysis of claims-based tools for health risk assessment. *Society of Actuaries Report*, 2007.

[25] W. K. Wong, A. Moor, G. Cooper, and M. Wagner. Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the 20th International Conference on Machine Learning*, pages 808–815, 2003.

[26] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *Advances in Neural Information Processing Systems 17*, 2004.



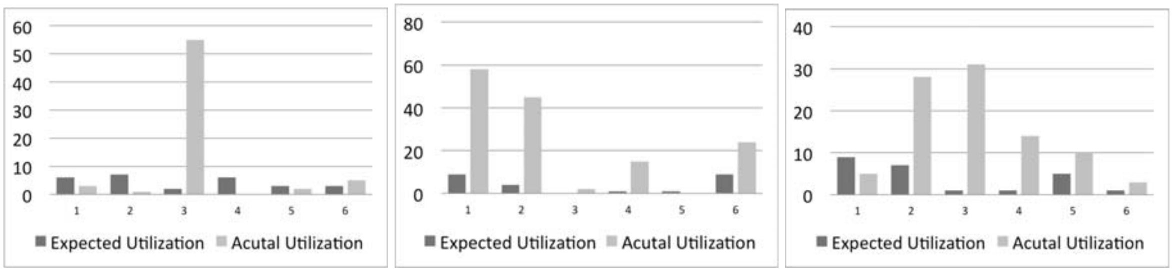
(a) cluster 1 (b) cluster 2 (c) cluster 3

(d) cluster 4 (e) cluster 5 (f) cluster 6

(g) cluster 7 (h) cluster 8 (i) cluster 9

(j) cluster 10

Figure 2: Mean utilization profiles of the identified clusters. Visit types: 1-PCP visit; 2-Specialist visit; 3-Emergency visit; 4-Inpatient hospital; 5-Outpatient hospital; 6-Patient's home.



(a) patient 6071 (b) patient 1311 (c) patient 2815

Figure 3: Actual vs. expected utilization for three patients detected as contextual anomalies. Visit types: 1-PCP visit; 2-Specialist visit; 3-Emergency visit; 4-Inpatient hospital; 5-Outpatient hospital; 6-Patient's home.