

Examining Healthcare Utilization Patterns of Elderly Middle-Aged Adults in the United States

Cilia E. Zayas,¹ Zhe He,² Jiawei Yuan,³ Mildred Maldonado-Molina,¹ William Hogan,¹
François Modave,¹ Yi Guo,¹ and Jiang Bian¹

¹Health Outcomes and Policy, University of Florida, Gainesville, FL

{cilia, mmmm, hoganwr, modavefp, yiguo, bianjiang}@ufl.edu

²School of Information, Florida State University, Tallahassee, FL

Zhe.He@cci.fsu.edu

³Department of ECSSE, Embry-Riddle Aeronautical University, Daytona Beach, FL

yuanj@erau.edu

Abstract

Elderly patients, aged 65 or older, make up 13.5% of the U.S. population, but represent 45.2% of the top 10% of healthcare utilizers, in terms of expenditures. Middle-aged Americans, aged 45 to 64 make up another 37.0% of that category. Given the high demand for healthcare services by the aforementioned population, it is important to identify high-cost users of healthcare systems and, more importantly, ineffective utilization patterns to highlight where targeted interventions could be placed to improve care delivery. In this work, we present a novel multi-level framework applying machine learning (ML) methods (i.e., random forest regression and hierarchical clustering) to group patients with similar utilization profiles into clusters. We use a vector space model to characterize a patient's utilization profile as the number of visits to different care providers and prescribed medications. We applied the proposed methods using the 2013 Medical Expenditures Panel Survey (MEPS) dataset. We identified clusters of healthcare utilization patterns of elderly and middle-aged adults in the United States, and assessed the general and clinical characteristics associated with these utilization patterns. Our results demonstrate the effectiveness of the proposed framework to model healthcare utilization patterns. Understanding of these patterns can be used to guide healthcare policy-making and practice.

Introduction

The 2012 Institute of Medicine (IOM) report on *"Best care at Lower Costs: The Path to Continuously Learning Health Care in America"* emphasizes that the growing complexity and fragmented nature of the US healthcare delivery system, resulting in areas of inefficiencies and uncoordinated care

delivery, causes harm to not only patients' financial life but also their health outcomes (Smith et al. 2012). Healthcare utilization patterns are often complex. For example, fragmented care often leads to duplicative, however, avoidable services. Further, an increasing number of patients with complex comorbidities exhibit higher use patterns (Burns et al. 2014). And, wide variations in the utilization of healthcare services, unrelated to patient health outcomes, have been observed across healthcare organizations (HCOs), geographic areas, providers, and payers (Newhouse et al. 2013).

Existing work on analyzing healthcare utilization patterns has been mostly one-dimensional and primarily focused on specific diseases or conditions that may not be generalizable to other areas (Eisele et al. 2010; Jin et al. 2014; Ilinca and Calciolari 2015). Other related research has examined healthcare utilization patterns by segmenting population groups in a number of ways, including but not limited to, income, the communities they live in, proximity to primary care providers, insurance, age, race, and or gender to examine healthcare utilization patterns exhibited by each. While these studies underscore specific aspects of healthcare utilization patterns, it is important to examine the problem by a patient's collective healthcare profile rather than one aspect alone.

In this work, we address this gap through a novel multi-level framework for healthcare utilization analysis. In particular, we use a vector space model to characterize a patient's utilization profile as the numbers of utilizations of different healthcare services. We then apply a random forest (RF) regression model to predict patients' total expenditures based on their utilization profiles. RF models often outperform other prediction methods, and are robust against over-fitting (Breiman, 2001). Additionally, the RF predictor

provides a dissimilarity measure (Shi and Horvath, 2006) between two patients considering both their utilization profiles and total expenditures. By leveraging the RF dissimilarity measure, we can cluster patients into groups with similar utilization profiles using hierarchical clustering approaches. We applied the proposed methods on the Medical Expenditures Panel Survey (MEPS) datasets (meps.ahrq.gov), from which we identified dominant utilization patterns, and assessed the general and clinical characteristics of healthcare utilization patterns of elderly and middle-aged adults in the United States. The results of our experiments demonstrate the effectiveness of the proposed framework and yield valuable insights on healthcare utilization patterns. Furthermore, our approach of modeling healthcare utilizations makes no assumptions about the underlying patient population, thus, is generalizable to patient groups other than middle-aged and elderly adults. In short, the proposed framework can leverage the vast amounts of readily accessible public datasets and uncover meaningful utilization patterns that can be used to inform policy-making.

Background

Variations in healthcare utilization of the older patient population and their adverse effects

Middle-aged and elderly adults are particularly vulnerable to variability in healthcare use, including both over- and under-utilization of healthcare services, which has the potential to cause unnecessary personal and financial harm (Farrow 2010; Nicholas and Hall 2012; Lipitz-Snyderman and Bach 2013; Kale et al. 2013). Overuse is often defined as services that are not supported by evidence, duplicative of other tests or procedures already received, potentially harmful, or not truly necessary (Burns, Dyer and Bailit 2014). On the other hand, underuse represents care that is not sufficient or appropriate in type, location, intensity, or timeliness to meet the patient's medical needs (Congressional Budget Office 2008). The elderly, more than any other population group, underutilize necessary preventive health services, which are known to improve their quality of care (Nicholas and Hall 2012). For example, despite being common in older adults, late life mood and anxiety disorders are highly undertreated. Approximately 70% of older adults with mood and anxiety disorders were found to either underuse or not use mental health services.

Analysis of healthcare utilization

Existing literature on utilization pattern analysis of older patient population groups predominately applied multi-variable and multinomial logistic regression models, reported prevalence rates, or compared proportions of classified groups. Very few studies applied machine learning approaches to analyze utilization patterns. Moreover, the larger body of research focused on medical utilization pattern analysis has been primarily conducted from a disease or condition-specific perspective, where the goal was to

segment the patient population by one dimension—the disease or condition of interest, then examine if patterns of healthcare utilizations varied after diagnosis.

For example, Eisele et al. (2010) designed a case-control study to examine changes in utilization of ambulatory medical care services before and after the diagnosis of dementia in Germany. Jin et al. (2014) used claims data to compare utilization patterns of disease-modifying anti-rheumatic drugs in elderly Korean Rheumatoid Arthritis patients by medical service, age-group, gender and geographic areas, separately. Ilinca and Calciolari (2015) used survey data to examine the impact that frailty had on utilization of primary and hospital care services among frail/elderly Europeans. Rosemann et al. (2007) administered questionnaires and used hierarchical stepwise multiple linear regression models to examine health care services utilization patterns of primary care patients with osteoarthritis. Oymoen Pottegard and Almarsdottir (2015) studied prescription drug utilization patterns and characteristics of high users of prescription drugs among the elderly Danish population. They applied multivariable logistic binary regression to study the top 1 percentile of the population that made up the largest share of prescription drugs dispensed at pharmacies.

Data Set and Sample Population

The MEPS is a representative survey used to collect comprehensive data on healthcare utilization and expenditures in the United States. The MEPS currently has two major components: the Household Component (HC) and the Insurance Component (IC). The HC collects data from a sample of families and individuals. During the household interview the MEPS collects detailed data related to respondents' demographic characteristics, health status, health conditions, use of medical services, charges and sources of payment, access to care, satisfaction with care, health insurance coverage, income and employment. The HC is supplemented by data from their medical providers. We used the 2013 Full-Year Consolidated Data File to select our patient utilization, expenditure, demographic and clinical characteristic features (i.e., 32 variables). The analytic data set was limited to adults 45 or greater.

Table 1. Summary statistics of healthcare utilization.

Description	Occurrence	mean (sd)	Percentile	
			50th	75th
Prescription medications (including refills)	343	19.39 (27.59)	9	27
Office-based visits to physicians	169	4.55 (7.63)	2	6
Office-based visits to non-physicians	158	2.78 (8.22)	0	2
Hospital outpatient visits to physicians	55	0.29 (1.70)	0	0
Hospital outpatient visits to non-physicians	166	0.39 (2.62)	0	0
Home health provider days	524	4.66 (32.10)	0	0
Emergency room visits	10	0.23 (0.65)	0	0
Hospital inpatient days	184	0.59 (4.27)	0	0

After the preprocessing, our study sample included 12,310 elderly (65 and older) and middle-aged (45–64) respondents in the 2013 MEPS dataset. We used a vector space model to represent individual's utilization profile based on the number of times each care service was used. Table 1 provides summary statistics of the utilization profiles of all patients in

our sample. The majority of the MEPS respondents had a relatively low level of utilization.

Methods

Our overall goal is to cluster patient populations with similar utilization profiles and expenditures into logical groups. Conceptually, a patient's demographic (e.g., age, gender, and social-economic status) and clinical characteristics (i.e., health status, and diagnoses) affect the degree of healthcare services used, which in turn affects the patient's total healthcare expenditure. Thus, we propose a hybrid learning system that combines a Random Forest (RF) regression model with Hierarchical Agglomerative Clustering (HAC). The overall process, as depicted in Figure 1, can be separated into three main components: (1) derive a sense of dis-/similarity based on a RF regression model between two patients' utilize profiles characterized by the number of office-based, outpatient, emergency room, and inpatient visits, the number of home care days, and the number of prescription medications; (2) identify clusters of patients with similar utilization patterns and expenditures using HAC; and (3) identify dominant utilization patterns by the patient population, identify distinct groups of high, low and median level-utilizers, and examine the clinical and demographic characteristics of the respective cluster groups. In the following sections, we describe each step and the basic procedures in further detail.

Step 1: Define a similarity measure between patient utilization profiles using random forest regression.

A patient's utilization profile (i.e., X_i) is characterized by the number of different types of healthcare service utilization (e.g., outpatient visits, prescribed drugs, etc.) incurred by this patient during a 1 year period. Each patient incurs a specific amount of healthcare expenditures (i.e., y_i) during the same period. Individuals with similar utilization profiles would incur similar expenditures (but not vice versa). Thus, training a RF regression model to predict total expenditures based on patients' utilization profiles would result in a logical similarity measure between observations. Patients with both similar utilization profiles and expenditures will more frequently end up in the same terminal node of each decision tree. After a tree is grown, put all the data, both training and out-of-bag (oob) samples, down the tree. If cases i and j are in the same terminal node, increase their similarity by one. At the end, normalize the similarities by dividing by the number of trees in the RF model. We denote the similarity between patient i and j as S_{ij} .

Step 2: Identify clusters of patients with similar utilization profiles using hierarchical clustering.

Based on these similarities, we then apply a hierarchical clustering approach to build a hierarchy of clusters. In particular, we use a bottom-up clustering approach, i.e., agglomerative, where each patient starts in its own cluster, and pairs of patients are merged as each moves up the

hierarchy. To decide which patients/clusters should be combined, a dissimilarity measure is required. We derive the dissimilarity between patient i and j as $D_{ij} = 1 - S_{ij}$. Further, we need a linkage criterion that determines the distance between *sets of observations* as a function of the pairwise distances between observations. In our study, we use the mean linkage clustering as our linkage criterion.

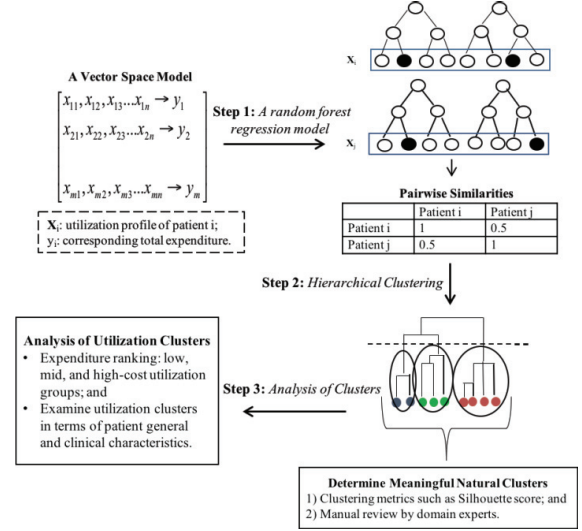


Figure 1. The overall process flow of the proposed healthcare utilization analysis framework.

The next step is to choose the best cutoff line on the dendrogram where natural clusters form. This concept is similar to determining the number of clusters (k) in k-mean clustering. A typical method is to choose the best clusters based on a clustering metric such as the Silhouette coefficient score. A Silhouette score measures the internal consistency of the learned clusters as it gives a sense of how well each object relates within its cluster (Rousseeuw 1987). Nevertheless, the existing literature indicates clustering metrics (i.e., internal consistency measures like the Silhouette Score) do not reliably identify natural clusters (Almedia et al. 2011, 2012). Thus, these methods should be supplemented with a manual review of the clusters formed.

Step 3: Analysis of the learned clusters.

We first rank the learned clusters according to the mean expenditure, and categorize them as low-, mid-, and high-utilization groups. We can then examine and compare patients' general and clinical characteristics among these utilization groups.

For patient characteristics, we selected variables based on the Anderson healthcare utilization model (Andersen and Aday 1974, 1978), where usage of healthcare services is determined by three dynamics: predisposing, enabling, and need. Predisposing factors include characteristics such as age, race, sex, religion, and values concerning health and illness, which are used to describe a patient's tendency to utilize healthcare services. Enabling factors are described as the 'means' by which individuals utilize healthcare services.

Table 2. Prediction performance of the random forest regression model on the 2013 MEPS dataset.

	Overall Model	Inpatient hospital days	Office-based physician visits	Office-based non-physician visits	Outpatient hospital physician visits	Outpatient hospital non-physician visits	Emergency room visits	Prescription medications	Home health days
NRMSE	1.68	1.89	2.11	2.23	2.23	2.25	2.17	2.16	2.22
R-squared	0.46	0.32	0.15	0.06	0.05	0.04	0.11	0.12	0.07

For example, variables such as family income, insurance coverage, geographic location and community attributes are considered enabling factors. Need factors represent severity of illness and are considered the most influential cause of healthcare utilization. The Anderson model also makes a distinction between the need perceived by an individual or evaluated by the healthcare system.

Based on the Anderson model, we selected: 1) gender, marital status, race/ethnicity, and education as variables to represent predisposing factors; 2) family income (as a percentage of the poverty line), as a proxy for socio-economic status, insurance coverage, employment status, and geographic region to represent enabling factors; and 3) patient self-report perceived personal health and mental health status plus 12 priority clinical conditions to represent need factors. The selected clinical conditions are specifically identified as priority conditions in the MEPS Household survey procedure due to their prevalence, expense, or relevance to health policy. Conditions, such as cancer, diabetes, emphysema, high cholesterol, hypertension, heart disease, and stroke can be both life-threatening and challenging to manage. Other conditions, such as arthritis and asthma are also flagged as priority clinical conditions, but are defined as chronic manageable conditions.

Results and Analysis

Performance of the random forest regression model

We first trained a RF regression model, following machine learning best practice (e.g., using cross-validation for hyper-parameter fitting—to choose the number of estimators in the RF model), to predict total healthcare expenditures based on patients' healthcare utilization profiles. As shown in Table 2, the overall RF regression model, considering all utilization features, exhibits reasonable performance ($r^2 = 0.46$, $\text{nmse} = 1.68$). When we examined individual feature's predictive power (i.e., using one feature at a time to train the RF model), the number of hospital inpatient days ($r^2 = 0.32$, $\text{nmse} = 1.89$), visits to office-based physicians ($r^2 = 0.15$, $\text{nmse} = 2.11$), prescription medications ($r^2 = 0.12$, $\text{nmse} = 2.16$), and visits to the emergency room ($r^2 = 0.11$, $\text{nmse} = 2.17$) are relatively more important than other features in predicting a patient's total healthcare expenditure.

Results for hierarchical clustering

From the learned RF regression model, we derived a pairwise dissimilarity (distance) matrix of all patients in the dataset. We then applied the HAC based on this distance matrix using the average linkage criterion. To find the natural clusters, we first used the Silhouette method, i.e., iteratively increase the number of clusters (k), measure the average Silhouette score with configuration k , and aim to choose the best k (best cutoff on the dendrogram) with the

highest Silhouette coefficient. By definition, the Silhouette score (s) is between -1 and 1 (i.e., $s \in [-1, 1]$), where the higher the Silhouette score, the better the segmentation of the learned clusters. However, in our case, the Silhouette score continues to increase as k increases. The Silhouette scores for $k=20$, 100, and 150 are -0.592, 0.007, and 0.065, respectively. This is understandable as the higher the k , the more the model overfits the training data, and the better the Silhouette metric becomes. Thus, we manually examined different configurations of k , and chose the ones that exhibit meaningful results with a reasonable Silhouette metric.

What we mean by meaningful is that the clustering result based on the use of care services and expenditures should correspond to logical segmentations of patient demographics and clinical profiles as well. For example, patients with a higher number of healthcare visits should also represent patients with low health status, and vice versa. Upon manual review of the clusters under each configuration, we found that patients are well segmented when $k=20$, 100, and 150, and the clustering results are meaningful. Due to space limitations, we only present results for $k=150$.

Analysis of the clusters

After deriving the clusters, we first ranked them by mean healthcare expenditures of all the patients within each cluster, and categorized the clusters into low-, mid-, and high-utilization groups accordingly.

We then analyzed predisposing, enabling, and need characteristics such as gender, marital status, race/ethnicity, family income (as a percentage of the federal poverty line) and perceived health status of the learned clusters. Table 3 provides the summary statistics of four cluster groups we selected for presentation. The patients' general and clinical characteristics as well as their utilization patterns are clearly different between low-, mid-, and high-utilization groups demonstrating the effectiveness of the proposed clustering approach. Detailed analysis is presented in Table 4 discussing the patients' general and clinical characteristics associated with their utilization profiles within each cluster.

Conclusion

The application of machine learning approaches in healthcare settings is promising. This study presented a simple but novel vector space model of patients' utilization profiles. Our evaluations, using the 2013 MEPS dataset, demonstrate the usefulness of the proposed approaches in identifying meaningful utilization patterns of elderly and middle-aged adults in the United States.

Future work will consist of validating the results on other datasets and studying whether and how utilization patterns would change across time.

Table 3. Patients characteristics of four learned utilization groups (k=150).

	Low-Utilization Cohort (96) (n = 1,837)	Mid-Utilization Cohort (34) (n = 140)	Mid-Utilization Cohort (54) (n = 171)	High-Utilization Cohort (63) (n = 150)
Mean Age (SD)	54.65 (8.14)	62.44 (10.59)	66.15 (10.61)	67.30 (13.03)
Age group				
45 - 64	88.19%	62.14%	40.35%	44.00%
65 - 85	11.81%	37.86%	59.65%	56.00%
Employment status				
Employed	70.99%	40.00%	39.77%	15.33%
Unemployed	29.01%	60.00%	60.23%	84.67%
Insurance status				
Private	44.2%	49.29%	70.18%	40.67%
Public	14.32%	45.71%	26.32%	52.67%
Uninsured	41.48%	5.00%	3.51%	6.67%
Health status				
Excellent / Very good	58.46%	25.71%	45.62%	12.67%
Good	30.92%	37.86%	35.67%	25.33%
Fair / Poor	10.62%	36.43%	18.71%	62.00%
Mental health status				
Excellent / Very good	64.83%	45.00%	54.38%	33.34%
Good	28.52%	34.29%	35.67%	30.67%
Fair / Poor	6.64%	20.72%	9.94%	36.00%
Reported clinical conditions				
Diabetes	2.99%	25.71%	23.98%	36.00%
Cancer	3.48%	18.57%	29.82%	22.67%
Coronary heart disease	1.58%	17.86%	9.94%	29.33%
Angina	0.71%	8.57%	4.09%	9.33%
Heart attack	1.36%	13.57%	5.85%	25.33%
Other heart disease	4.03%	28.57%	21.05%	40.67%
Stroke	1.52%	15.00%	9.36%	29.33%
Use of healthcare services (SD)				
Office-based (physicians)	0.00 (0.00)	5.29 (3.01)	12.30 (2.92)	9.35 (14.62)
Office-based (non-physicians)	0.00 (0.00)	0.50 (0.78)	2.01 (2.40)	5.66 (15.49)
Outpatient (physicians)	0.00 (0.00)	1.45 (0.83)	0.00 (0.00)	0.78 (3.44)
Outpatient (non-physicians)	0.00 (0.00)	0.65 (1.78)	0.62 (1.39)	1.61 (10.27)
Home health (#days)	0.12 (2.83)	0.00 (0.00)	0.01 (0.08)	23.32 (51.40)
Emergency room visits	0.00 (0.00)	0.28 (0.79)	0.19 (0.50)	1.33 (1.41)
Hospital stays (#days)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	26.23 (26.61)
Prescription medications	0.00 (0.00)	38.82 (11.12)	16.54 (3.52)	48.81 (42.75)
Outcome (SD)				
Total healthcare expenditures	\$127.57 (\$836.76)	\$6,3050.84 (\$6,089.32)	\$6,649.37 (\$10,029.70)	\$57,894 (\$54,690.20)

Table 4. Detail analysis of four learned utilization groups (k=150).

#	Category	#Patients	Mean Exp. /Patient	Mean Age	Description of the clusters
96	Low	1,837	\$128	55	Average use of all healthcare services was reported at 0. This cohort consists primarily of middle-aged, employed patients. Over 50% of the patients perceived their personal and mental health status as excellent or very good. Less than 4% reported having one or more priority clinical conditions.
34	Mid	140	\$6,351	62	Average use of most services was reported close 0, with three exceptions: office and outpatient visits to physicians and Rx medications are used at 5, 1, and 39/patient, respectively. This cohort consists of 62% middle-aged adults, and 40% of this cohort is employed. In comparison to the low and other mid use cohort, the percentage who perceived their personal and mental health status as fair or poor was much higher. Between 9% and 29% reported having a priority clinical condition.
54	Mid	171	\$6,649	66	Average use of most services was reported close to 0, with three exceptions: office visits to physicians and non-physicians and Rx medications at 12, 2, and 16/patient, respectively. 60% of this cohort is elderly. In comparison to the low and other mid use cohort, the percentage who perceived their personal and mental health status as fair or poor was between that of the other two. Between 4% and 30% reported having a priority clinical condition.
63	High	150	\$57,894	67	In comparison to the other 3 cohorts, average use of all services is the highest. Particularly, average patient uses 49 Rx medications. The percentage of patients who were elderly and perceived their personal and mental health status as fair or poor was also higher. Between 9% and 41 % reported having a priority clinical condition.

Acknowledgements

This work was supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1TR001427.

References

- Andersen, R., Aday, LA. 1978. Access to medical care in the U.S.: realized and potential. *Med Care* 16(7):533–546.
- Aday, LA., Andersen, R. 1974. A framework for the study of access to medical care. *Health Serv Res* 9(3):208–220.
- Almedia, H., Guedes, D., Meira, W., Zaki, M. 2011. Is there a best quality metric for graph clusters? In proceedings of the European Conference on Machine Learning and knowledge discovery (1) 44-59.
- Almedia, H., Neto, D., Meria, W., Zaki, M. 2012. Towards a better quality metric for graph cluster evaluation. *Journal of Information and Data Management* 3(3): 378 – 393.
- Baxter, J., Bryant, S., Scarbro, S., Shetterly, S. 2001. Patterns of rural Hispanic and Non-Hispanic White health care use. *Research on Aging* 23(1): 37-60.
- Breiman, L. 2001. Random Forests. *Machine Learning* 45(1): 5-32.
- Burns, M., Dyer, M., Bailit, M. 2014. Reducing overuse and misuse: State strategies to improve quality and cost of healthcare. Robert Wood Johnson Foundation, Princeton, NJ.
- Cameron, K., Song, J., Manheim, L., Dunlop, D. 2010. Gender Disparities in Health and Healthcare Use Among Older Adults. *Journal of Women's Health*. 19(9):1643-1650.
- Congressional Budget Office. 2008. The Overuse, underuse, and misuse of healthcare. Congressional Report. Statement of Peter R. Orszag Director of Congressional Budget Office before the committee on finance. United States Senate. https://www.cbo.gov/sites/default/files/110th-congress-2007-2008/reports/07-17-healthcare_testimony.pdf - accessed on November 11, 2015.
- Cooper, R., Cooper, M., McGinley, E., Fan, X., Rosenthal, J. 2012. Poverty, Wealth, and Health Care Utilization: A Geographic Assessment. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*. 89(5):828-847.
- Eisele M, van den Bussche H, Koller D, Wiese B, Kaduszkiewicz H, Maier W, et al. 2010. 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.
- Farrow, F. 2010. Overutilization and underutilization of preventive services in elderly populations, A conundrum. *Marquette Elder's Advisor* 12 (1):103-122.
- Hill, J., Fillit, H., Shah, S., Del Valle, M., Futterman, R. 2005. Patterns of healthcare utilization and costs for vascular dementia in a community-dwelling population. *Journal of Alzheimer's disease* 8(5): 43-50.
- HealthyPeople2020. Older adults. <http://www.healthypeople.gov/2020/topics-objectives/topic/older-adults> accessed on December 1, 2015.
- Ilinca S, Calciolari S. 2015. The Patterns of Health Care Utilization by Elderly Europeans: Frailty and Its Implications for Health Systems. *Health Serv Res* 50(1):305-320.
- Jin, X., Lee, J., Choi, N., Seong, J., Shin, J., Kim, Y., et al. 2014. Utilization Patterns of Disease-Modifying Antirheumatic Drugs in Elderly Rheumatoid Arthritis Patients. *J Korean Med Sci* 29(2):210-216.
- Kale, M., Bishop, T., Federman, A., Kehani, S. 2013. Trends in the overuse of ambulatory healthcare services in the United States. *JAMA Intern Med*. 173(2):142-148.
- Lipitz-Snyderman, A., Bach, P. 2013. Overuse: when less is more, more or less. *JAMA* 173 (14): 1277–1278.
- Mandsager, P., Lebrun-Harris, L., Sripipatana, A. 2015. Health center patients' insurance status and healthcare use prior to implementation of the Affordable Care Act. *American Journal of Preventive Medicine* 49 (4): 545-552.
- Medical Expenditure Panel Survey Data Set. 2013. U.S. Department of Health and Human Services: Agency for Healthcare Research and Quality (AHRQ). <http://meps.ahrq.gov/mepsweb/>
- Newhouse, J., and Committee on geographic variation in healthcare spending an promotion of high-value care. 2013. Variation in healthcare spending: Target decision making, not geography. Institute of Medicine, Washington, D.C.
- Nicholas, J., Hall, W. 2012. Screening and preventive services for older adults. *The Mount Sinai Journal of Medicine*, 78(4): 498–508.
- Oymoen A, Pottegard A, Almarsdottir AB. 2015. Characteristics and drug utilization patterns for heavy users of prescription drugs among the elderly: a Danish register-based drug utilization study. *Eur J Clin Pharmacol* 71(6):751-758.
- Rosemann, T., Joos, S., Szecsenyi, J., Laux, G., Wensing, M. 2007. Health service utilization patterns of primary care patients with osteoarthritis. *Bmc Health Services Research* 23 (7):169.
- Rousseeuw, P. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53-65.
- Shi, T., Horvath, S. 2006. *Journal of Computational and Graphical Statistics* 15(1): 118-138.
- Smith, M., Saunders, R., Stuckhardt, L., McGinnis, M. 2012. Best care at lower cost: The path to continuously learning healthcare in America. Institute of Medicine (U.S.), Committee on the learning healthcare system in America, Wahington, D.C.