TITLE PAGE

**Data Analytics in Health Promotion: Health Market Segmentation and
Classification of Total Joint Replacement Surgery Patients**

Eric R. Swenson, Nathaniel D. Bastian, Harriet B. Nembhard

Center for Integrated Healthcare Delivery Systems
Department of Industrial and Manufacturing Engineering
The Pennsylvania State University
University Park, PA 16802, USA

ers187@psu.edu; ndbastian@psu.edu; hbn2@engr.psu.edu

Corresponding author: Nathaniel D. Bastian, 362 Leonhard Building, University Park, PA 16802, USA, (570) 809-3619, ndbastian@psu.edu

Data Analytics in Health Promotion: Health Market Segmentation and Classification of Total Joint Replacement Surgery Patients

Abstract

Providing insight into healthcare consumers' behaviors and attitudes is critical information in an environment where healthcare delivery is moving rapidly towards patient-centered care. We apply a two-stage methodology using both supervised and unsupervised machine learning methods to a patient data set from the electronic medical records of an academic medical center located in central Pennsylvania. The data are from patients who had total joint replacement surgery between December 2013 and September 2015. Two clustering methods and four classification algorithms were applied to the data set. Patients cluster into six distinct health market segments from which the cluster assignment is used as the response variable in supervised learning to classify patients. The classification model accurately predicts the cluster assignment for out-of-sample patients, while offering insight into patient behaviors and attributes to help clinicians, health marketers, and healthcare consumers move toward the goal of patient-centered and value-based healthcare.

Keywords: health market segmentation, health promotion, data analytics, machine learning, total joint arthroplasty, value-based healthcare

1. Introduction

1.1 Background

According to the World Health Organization (WHO), “health promotion is the process of enabling people to increase control over, and to improve, their health. It moves beyond a focus on individual behavior towards a wide range of social and environmental intervention” (WHO, 2014). Further, the Centers for Disease Control and Prevention (CDC) state that health marketing involves “creating, communicating and delivering health information and interventions using customer-centered and science-based strategies to protect and promote the health of diverse populations.” The purpose of market segmentation is to find specific well-defined, homogenous customer groups in a larger population, some of which are likely to respond positively to promotions or service offers (Woodside, Nielsen, Walters, & Muller, 1998).

Health market segmentation offers insights into healthcare consumers' behaviors and attitudes, which is critical information in an environment where healthcare delivery is moving rapidly towards patient-centered care that is premised upon individuals becoming more active participants in managing their health. Strategies to encourage and support consumer engagement in healthcare are important for healthcare organizations (e.g., providers, health plans, pharmaceutical companies). Increased access to health information can help patients make better and more informed decisions leading to better quality of care, health outcomes, and satisfaction with care. Providing individuals in a community with more useful information may change their behavior in a way that reduces healthcare costs. Health market segments may provide valuable clues as to how healthcare organizations may more specifically target and personalize products and services for healthcare consumers (Greenspun & Coughlin, 2012).

As a means to improve health promotion for patients in a given community, effective health marketing strategies should be developed and employed. Pires and Stanton (2008) discuss the

application of marketing knowledge to healthcare services, arguing that social marketing has played a crucial role in acceptability and awareness regarding key health issues by campaigns (e.g., anti-smoking, anti-obesity). Customer-based market segmentation provides the focus and precision required to enhance personalized healthcare by identifying the latent relationships between attributes found in individual electronic medical records, customer surveys, and/or demographic data. These relationships help define patient clusters or segments which hospitals, health systems, insurers, and affiliated healthcare agencies can use to refine health marketing efforts. Targeting health promotions to specific market segments increases efficiency, decreases health promotion costs, enhances patient-centered care and personalized healthcare goals, and is more likely to increase healthcare consumer participation in managing their own health. Market segmentation studies also hold the potential to be critical components of the National Institutes of Health translational research initiatives by allowing researchers to efficiently locate desirable health market segments to target with new laboratory research.

Tynan and Drayton (1987) discuss the importance of market segmentation techniques in overall marketing strategy. They suggested that the main market segmentation bases can be: geographic, demographic, psychological, psychographic or behavioral and that market segmentation leads to closer association with the targeted set of consumers. In addition, strategic market segmentation plays a key role in discovery, innovation and development of medical products and services (MacLennan & Mackenzie, 2010). There have been numerous health marketing research studies done over the past few decades. Common clustering methods include hierarchical and non-hierarchical clustering, chi-squared automatic interaction detection, and CART (or classification and regression trees). Additionally, market segmentation studies normally fall in to one two categories: a priori or data-driven (Wind, 1978). In healthcare, a majority of the papers also use either surveys or interviews to gather the data, as opposed to actual patient data.

Advances in the electronic medical record and in analytic capabilities in health systems has led to the implementation of machine learning algorithms and statistical models to identify underlying patterns and garner actionable insight from readily available patient data. This enormous amount of patient data include patients' physical characteristics (age, weight, height, etc.), as well as past medical conditions, lab results, radiology reports and images, and a plethora of time-series data pertaining to each visit to a network provider. Further, these patient health records provide real-time access to providers in the clinical setting, and they hold the potential to tell a much bigger story about a patient's past, current, and future health, such as what types of treatments or health promotions they may respond to, whether they value customer service, prefer messages via an interactive personal health record, or value routine care.

This paper demonstrates the value of using data analytics in a healthcare setting, while serving as a tutorial to guide researchers in the area of health analytics on how to integrate both unsupervised and supervised machine learning methods in a robust framework. In addition to helping improve health promotion by leveraging patient data and machine learning for health market segmentation, the results provide an invitation for subsequent expert systems applications in health care. In an era of rising healthcare costs and demand for services, the use of expert systems holds the potential to enhance personalized health care by allowing healthcare providers to efficiently find and target at-risk or at-benefit health market segments so that health promotion strategies can be tailored effectively to positively impact patient health outcomes.

1.2 Literature Review

Diversity: Health market segmentation studies range from analysis of clinical populations (Newcomer et al., 2012; Axen et al., 2011; Kim et al., 2013) to segmentation studies on survey data (Liu and Chen, 2009; Suragh et al., 2013; Moss et al., 2009; Kolodinski and Reynolds, 2009; & Berg et al., 2010). Of the papers that used survey data, two looked at college student substance abuse behaviors (Berg, 2010; Suragh, 2013), one looked at customer preference for healthcare service and clustered patients based on their preference and demographic attributes (Liu, 2009), and the last two used large survey data from a combination of the Behavioral Risk Factor Surveillance System (BRFSS), US Department of Agriculture funded nationwide polls, and a mix of public and US Census data (Moss, 2009; Kolodinski and Reynolds, 2009).

Two of the studies based on patient data investigated RFM (recency, frequency, and monetary) models. Lee (2012) studied customer loyalty in a university hospital setting in Korea. He analyzed patient demographics and hospital visit data to understand which patient types were loyal or ordinary users. Wu et al. (2014) conducted a similar study in Korea where they looked at a tenth of the sample size as Lee (1462 vs 14,072), but studied LRFM which is RFM plus length. The goal of Wu et al. (2014) was to cluster the under-18 year old patient population in a dental clinic based on demographics, length of stay, frequency of visits, and proximity of recent visits.

Outcomes measured: Two of the retrospective studies from Taiwan and Korea focused on customer loyalty and Customer Relations Management (CRM). Cheng et al. (2005) applied k -means clustering to demographic data regarding nursing homes. The goal was to cluster patients based on demographics, specialty care required, rehabilitation services, etc. and then develop care service strategies based on provider feedback. Lee (2012) conducted a similar study in Korea using a CRM. Lee (2012) also applied k -means clustering with k equal to two. The two clusters divided the population into loyal and ordinary patients. After clustering, Lee (2012) applied decision trees to stratify the loyal patients to determine which factors were most important in determining how a patient is classified.

Lee (2012) was not alone in his post-cluster stratification approach. Kim et al. (2013) used k -means clustering and decision tree induction to segment and classify healthcare providers. In this study of hospital providers, Kim et al. (2013) looked at location, population density, beds, patient to provider ratio and other costing data to segment both single specialty and hospitals that conduct either general surgery or ophthalmology services. After clustering both types of hospital services, they applied a stratification approach using decision trees to develop homogenous strata. Determining homogenous strata allows for better sample approaches that aid in future policy studies (Kim et al., 2013).

Four of the papers that applied market segmentation to survey data measured health and behavior outcomes. Kolodinsky et al. (2009), Berg et al. (2010), Suragh et al. (2013), and Moss et al. (2009) all looked for influential behaviors with the end state of being able to identify distinct segments and then use specific techniques to target those segments in order to modify behaviors. Berg et al. (2010) and Suragh et al. (2013) conducted almost the same study in different regions in the US and arrived at the same number of clusters with strikingly similar names and cluster demographics. The congruency of results despite different time frames, locations, statistical software programs, and sample sizes indicates the strength of cluster analysis to deliver repeatable findings given similar data sets. Although not specifically addressing college students, Moss et al. (2009) conducted a larger version of Suragh et al. (2013) and Berg et al. (2009) studies. Moss et al. (2009) used various large data sets from the CDC,

publicly available data, and BRFSS to look at the attitudes and behaviors regarding high risk drinking.

Similarly, but on a much smaller scale, Kolodinsky et al. (2009) used national poll survey data to cluster based on behavioral, environmental, geographic, food knowledge, and education factors. Kolodinsky et al. (2009) was interested in obesity and the role of food and lifestyle behaviors on population health. A striking similarity in Kolodinsky et al. (2009), Berg et al. (2010), and Suragh et al. (2013) is how they use the same industry practices that created the problems they are studying to counter the problems. Both Suragh et al. (2013) and Berg et al. (2009) borrow from the tobacco industry and Kolodinsky et al. (2009) borrow survey methods from the food industry.

Liu and Chen (2009) and Kent et al. (2014) conducted market segmentation using different approaches but each applied multiple clustering techniques to verify the results. The prior used survey data while the latter was based on secondary use data from a variety of studies. Liu and Chen (2009) used a mix of hierarchical and non-hierarchical methods and ultimately settled on a hierarchical clustering method that reduced the attributes from 24 to 5 yielding 3 distinct clusters. Kent et al. (2014) applied and compared three different methods to five real data sets and 4 randomly generated data sets to test reproducibility, likeness of outputs, and ease of use.

The final two papers used patient data sets to cluster patient populations based on a specific condition or set of conditions. In Axen et al. (2011), a composite data set based on questionnaires and self-reported pain score data were analyzed. The self-reported data were via time-series SMS text messages over a 26-week period. These patient pain progress scores were cleverly reduced to four parameters through the use of nonlinear spline regression. These four parameters (developed for all 176 patients) were segmented using hierarchical clustering. In Newcomer et al. (2011), hierarchical clustering was also used, however, in this study, the sample size was relatively large (15,480 patients) and pulled from a health maintenance organization (HMO) database of patients with at least two chronic medical conditions that fell into the top 20% of healthcare expenditures. The goal of the study was to further segment high-risk and high-cost patients to enable clinicians to target specific at-risk populations with appropriate health interventions and healthcare management plans.

Methods used: Hierarchical clustering was used in five studies (Newcomer et al., 2011; Berg et al., 2010, Axen et al., 2011; Liu and Chen, 2009; & Suragh et al., 2009). In all but one of them, Liu and Chen (2009), Ward's method was used as the distance/similarity measure. In Liu and Chen (2009), Pearson's correlation was used as the similarity measure. Four of the studies used *k*-means clustering (Kim et al., 2013; Lee, 2012, Wu et al., 2014, & Cheng et al., 2005).

1.3 Research Objectives

Numerous health market segmentation studies were conducted over the past few decades, but few used electronic medical record (EMR) patient data and none applied supervised learning to help predict patients' health market segment (i.e., cluster). In this research, we conduct health market segmentation using EMR data from over 700 total joint replacement surgery patients from an academic medical center located in central Pennsylvania. The research objectives are: 1) to identify distinguishable patient clusters based on age, BMI, pain scores, and other descriptive variables; 2) to use those clusters to build predictive models that can accurately classify future patients into their health market segment; and 3) to discuss how patient clusters

can provide useful information for helping project future patient outcomes. At a minimum, clustering patients into distinguishable segments will help clinicians, health marketers, hospitals, and health insurers identify specific and targetable health market segments. Efficiently targeting at-risk or at-benefit health market segments will increase patient-centered care, help translational researchers expand their work from benchtop to bedside, and improve value-based care initiatives.

A goal of this research is to demonstrate the value and application of expert and intelligent systems in healthcare. The explosion of big data in healthcare coupled with the need for real-time or near real-time decision support necessitates the use of advanced data analytics. This has set the conditions for new expert and intelligent systems healthcare-related applications. Although this is a small data set, it represents a large revenue generating portion of a hospital's orthopedic service line. The demonstration of expert system tools applied to total joint arthroplasty could be adapted to other service lines such as oncology and cardiology. Healthcare data is vast, but with the proper tools from expert and intelligent systems it can provide great value to the healthcare industry.

2. Materials and Methods

2.1 Study Setting

The Penn State Hershey Medical Center (PSHMC) is an academic medical center located in Harrisburg, PA. It is part of the Pennsylvania State University and the Penn State Health Care Partners clinically integrated network of regional care providers. In 2012, PSHMC admitted over 27,000 in-patients and treated over 890,000 outpatients (PSHMC, 2015). The data used in this study were approved under PSHMC institutional review board *STUDY00002554: Total joint replacement process flow*.

2.2 Study Population

Unlike many of the previous health-related market segmentation studies, this study uses actual patient data from EMRs. The patients were treated at the PSHMC between 30 December 2013 and 30 September 2015, during which time they had total joint arthroplasty (joint replacement surgery) of the lower extremity (hip or knee). The de-identified clinical data are divided into two main categories: training and testing. The training data include 814 patients and over 75 attributes ranging from age, BMI, procedure code, length of stay, pain scores, ICD-9 codes (International Classification of Diseases and Related Health Problems, Version #9) and numerous other fields related to surgical side, approach, and times that various steps in the surgical process started and ended. There are 14 ICD-9 variables representing documented disease or health problems. In total, over 650 unique ICD-9 codes are referenced.

To account for common disease codes across the patient sample, each unique ICD-9 code is re-coded as a binary variable. Patients with that ICD-9 code listed are coded as a 1, or 0 otherwise. The data include patients from diagnosis readiness groups (DRG) 461, 462, 469, and 470 (bilateral joint replacement or reattachment of the lower extremity without and with major complications and comorbidities (MCC) and major joint replacement or reattachment of the lower extremity without and with MCC) and both procedure codes 81.51 (hip) and 81.54 (knee). For this study, we excluded DRG 461/2 patients to focus on single surgical site patients. Removing DRG 461/2 reduced the patient count to 767 (loss of 47 records). Patient records that

were missing key data elements for this analysis were also excluded, bringing the sample size down to 741 (26 more deletions). The final training data set includes 741 patients spanning 719 variables (columns). The testing data set includes 63 patients from the month of September 2015. After removing 3 DRG codes 461/2 and 9 entries with missing data, the final testing data set contains 49 complete patient records. Table 1 below summarizes the data.

Table 1. Summary of Total Joint Replacement Surgery Patient Data

Original Data	Training Data	Testing Data
Dates	30/12/2013-30/08/2015	1/09/2015-30/09/2015
Sample Size	814	61
DRG 461	1	0
DRG462	46	3
DRG469	17	1
DRG470	750	57
Hip (81.51)	286	17
Knee (81.54)	528	44
Male	286	24
Female	528	37
Average BMI	33.21	35.17
Average LOS	65.21	56.12
Average Age	63.12	61.66
Cleaned Data	Training Data	Testing Data
Sample Size	741	49
Number removed due to DRG	47	3
Number removed due to incomplete data	26	9

2.3 Methodological Approach

This study investigates the use of a two-stage methodology that employs both unsupervised and supervised machine learning techniques applied to clinical patient data from the EMRs of total joint replacement surgery patients. The goal is to apply two different unsupervised learning (clustering) methods to a large subset of the patient data to identify unique clusters of patients. Once the clusters are identified, each patient's cluster label is appended to his/her record as a new variable. Supervised learning is then applied to the same training data set. The training data are further broken down into two subsets: a training set and a validation set. Four supervised learning (classification) methods are applied to the new training data (80% of data). The four methods include decision trees, random forests, bootstrap aggregation (bagging), and non-linear support vector machines. Each classification model is evaluated for accuracy in its ability to

predict a patient cluster. The four models are compared based on their classification accuracy using the validation data.

3. Machine Learning Methods

3.1 Unsupervised Machine Learning Methods

We applied both k -means clustering and hierarchical clustering to the patient data. K -means clustering can handle high-dimensional data and is generally computationally faster than hierarchical clustering with large data sets, but the user must specify a value of k up front. To determine the best value of k , we initially selected k such that the marginal change in the within-cluster sum of squares, as k increases, diminished. In hierarchical clustering, the data are clustered completely and the user must determine the best place to cut the resulting dendrogram to distinguish the number of clusters. The following subsections provide the details and theory on k -means and hierarchical clustering models.

3.1.1 K -means Clustering

K -means clustering is an unsupervised machine learning technique that separates n multidimensional observations into k distinct, non-overlapping clusters based on the similarity between the observation and the centroid of the cluster. The main idea behind k -means clustering is to find a good clustering where the within-cluster variation (i.e., dissimilarity) is as small as possible. The most commonly used similarity measure for continuous attributes is the squared Euclidean distance (Jain, Murty, & Flynn, 1999). This technique requires an initial value of k from which k initial clusters are formed. Depending on the variant of the algorithm, each observation is either randomly assigned a cluster number first or k observations are randomly selected as initial centroids. In either case, every observation is assigned to a cluster based on a similarity measure.

The algorithm starts by placing each of the n observations into exactly one of k pre-determined clusters. Let $C_1, C_2, \dots, C_k$ denote sets containing the indices of the observations in each cluster. The following two conditions must hold at every iteration of the algorithm: $C_1 \cup C_2 \cup \dots \cup C_k = 1, 2, \dots, n$ and $C_k \cap C_{k'} = \emptyset \quad \forall k \neq k'$ (James, Witten, Hastie, & Tibshirani, 2013). That is, each observation i belongs to at least one of the k clusters, and the clusters are non-overlapping (i.e., no observations belongs to more than one cluster). After the initial placement of observations into one of the k clusters, a cluster centroid (i.e., the vector of the p feature means for the observations in the k th cluster) for each of the k clusters is computed, where the mean for feature j in cluster C_k is calculated in Equation (1).

$$x_{kj} = 1/|C_k| \sum_{i \in C_k} x_{ij} \quad \forall k, j \quad (1)$$

where $|C_k|$ represents the number of observations in the k th cluster. In the next step of the algorithm each observation is assigned to the cluster whose centroid is the closest, which is equivalent to minimizing the total within-cluster variation; this objective function Z is denoted in Equation (2). The value of Z is guaranteed to decrease at each step.

$$\min_{C_1, C_2, \dots, C_k} \sum_{k=1}^K \sum_{i \in C_k} \sum_{j=1}^p (x_{ij} - x_{kj})^2 \quad (2)$$

At each iteration of k -means clustering, the algorithm first calculates the centroid of each cluster and then compares each observation to the centroid of each cluster as to minimize the total within-cluster variation. Note that observations are reassigned to clusters based on maximizing similarity between the observation and clusters centroid. The k -means clustering algorithm continues until some pre-determined convergence criterion is achieved. Convergence criteria could be based on iterations, when no more reassignments occur, or when no significant change in squared error from one iteration to the next (Jain et al., 1999).

K -means clustering is widely used due to its ease of use and ability to handle large data sets. The k -means clustering algorithm converges to a local optimum and is susceptible to initial starting conditions, which can prevent it from reaching a global minimum. It works best when multiple starting points are used.

3.1.2 Agglomerative (Hierarchical) Clustering

Agglomerative (hierarchical) clustering is another common approach to unsupervised machine learning. The method starts with a set of n -multidimensional observations which each comprise a cluster (of size one). Clusters are merged based on a similarity measure until there is one cluster. Unlike k -means clustering, there is no predetermined value of k . The user must determine an appropriate value of k . The output from hierarchical clustering is displayed in a dendrogram which “represents the nested grouping of patterns and similarity levels at which groupings change” (Jain et al., 1999).

In hierarchical clustering, clusters are traditionally joined based on a minimum distance measure or maximum similarity measure (similar to k -means). The similarity between pairs of observations, one from each cluster, are compared and clusters are merged based on a maximum similarity criteria (normally minimum distance). Different algorithms use different methods to determine minimum distance; two common techniques are complete linkage which combines clusters that have the minimum of the maximum pairwise distance between any two points (from different clusters) and single linkage which combines two clusters if the distance between them is the minimum of the pairwise distances (Jain et al., 1999).

3.1.3 Algorithm Input Matrix

As a prerequisite for clustering, a distance or similarity matrix is required. In the case of continuous variables, the predictor variables are standardized then clustered based on Euclidean distance in the case of the k -means algorithm or Ward’s method in the case of hierarchical clustering. Ward’s method is also known as the minimum variance criterion. This method is applied in hierarchical clustering algorithms where the objective is to minimize the total within-cluster variance. The algorithm starts with n clusters representing the n observations. Then, $n - 1$ clusters are formed out of n clusters by combining the pair of observations that results in the smallest increase in within cluster variance. Ward’s method uses a squared Euclidean distance measure to determine minimum variance (Ward, 1963).

When the variables are categorical or logical (binary) or a mix of types, using a standard distance measure such as Euclidean or Manhattan distance is not appropriate. As is the case in this patient data set, there is a mix of binary, categorical, and continuous variables which requires

a method that first defines the similarity between patients. We apply Gower's coefficient to create a dissimilarity matrix which is then used in both k -means and hierarchical clustering. Gower's dissimilarity coefficient is a general similarity measure, S_{ij} , that Gower (1971) developed to determine similarity between two observations, i and j . This coefficient can be applied to ordinal, continuous, and dichotomous data. In determining Gower's coefficient, the similarity between two observations on the k th dimension are calculated for all k dimensions: $s_{ijk} = 1 - x_{jk} - x_{ik} R_k$,

where R_k is the range of k . The overall similarity coefficient is:

$$S_{ij} = \frac{1}{k} \sum_{k=1}^k s_{ijk} = \frac{1}{k} \sum_{k=1}^k \delta_{ijk} \quad \forall i, j \text{ where } \delta_{ijk} = \begin{cases} 0 & \text{if there is a missing value in } i \text{ or } j \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

3.2 Supervised Machine Learning Methods

Classification modeling is a component of supervised machine learning that seeks to predict a response for an observation where the response falls into a particular class. At this point in the study, the first phase of the two-stage methodology is complete, where each patient in the training data set is now assigned to the cluster such that the total within-cluster variation is minimized. In order to assign out-of-sample patients to a particular segment, one potential approach could be to assign them to the closest cluster center. While this procedure would be relatively simple to implement, it lacks inherent flexibility and interpretability compared to traditional supervised machine learning models. Therefore, to predict patient segments in the second phase of the methodology, four classification models are tested, including decision trees, random forests, bagging, and non-linear support vector machines. The goal of the classification modeling is to identify the best method given the data that will accurately predict a patient's cluster using the validation data. Recall that the original training data are randomly split into two subsets: training (80%) and validation (20%). The training data are used to train each classification model and the validation data are used to evaluate each model against the others. The area under the (ROC) curve (AUC) is used to evaluate the four models. Note that an AUC of 0.5 indicates that a classifier performs no better than chance, where as an AUC of 1.0 perfectly classifies each observation.

3.2.1 Decision (classification) Trees

Decision trees, random forests, and bagging are all tree-based classification models, whereas support vector machines (SVM) seek to find a non-linear boundary between classes. In the tree-based methods, the predictor space is segmented into smaller regions through a series of rules that define a split at each level of the tree. Recursive binary splitting is used to grow the tree with one of three common criteria: classification error rate, Gini index, or cross-entropy, used for binary splits (James, et al., 2013). The following three equations show each criterion:

$$\text{Classification error rate:} \quad E = 1 - \max_k p_{mk} \quad (4)$$

where p_{mk} is the proportion of observations from the k th class in the m th group (James et al., 2013).

$$\text{Gini index:} \quad G = -\sum_{k=1}^K p_{mk}(1-p_{mk}) \quad (5)$$

$$\text{Cross-entropy:} \quad D = -\sum_{k=1}^K p_{mk} \log(p_{mk}) \quad (6)$$

Cross-entropy and Gini index are often referred to as node purity measures since values of p_{mk} near one or zero (observations predominately from a single class or pure) result in an index value near zero. This approach creates splits based on the one that maximizes the reduction in impurity (Cran, 2105) or classification error rate.

3.2.2 Bootstrap Aggregation (Bagging) and Random Forests

Although decision trees are easy to interpret, they commonly suffer from high variance. This can be offset using bootstrap aggregation and random forests. Both bagging and random forests are ensemble classifiers that bootstrap the training data set to create Z multiple training data sets. The average of the Z decision trees produce a classification model with lower variance than a single tree. Random forests proceed one step beyond bagging by regulating the number of predictors available for consideration at each split. Random forests overcomes the high correlation that can occur in bagging when a dominate predictor causes all subsequent bootstrapped trees to look similar (James et al., 2013).

In bagging (classification trees), M different training sets are generated by bootstrapping the original training data. Bootstrapping is random sampling with replacement. M decision tree prediction models, $f_i(x)$, are developed, one for each bootstrapped data set i ($i \in 1, 2, \dots, M$). For a given test set observation, the predicted class is determined based off the class most commonly returned by the M prediction models. In essence, bagging is an ensemble classifier that determines a classification based majority rules.

Random forests are a modification to the bagging method described above that reduces the correlation between bagged trees. A drawback of bagging occurs when one or two predictors dominate the rest resulting in very similar looking trees. Random forests deviate from bagging in that it restricts the number of available predictors that a given tree can use. For instance, if there are p dependent variables, bagging will build a tree considering p variables. In random forests, the user selects the number of variables (less than p) that are randomly selected for each bagged tree. Dominate variables will not be able to influence every bagged tree.

3.2.3 Support Vector Machines

Although tree-based methods can provide excellent classification results, they are not the only classification models. SVMs are a good classification approach especially when observations can be divided by a separating hyperplane. When this condition does not hold and the boundaries are non-linear, non-linear SVMs using kernels can accurately classify data.

SVM classification for multi-class problems starts with the understanding of how it applies to a two-class problem. In linear form, SVM seeks to find a maximum marginal classifier with

decision rule, Dx such that if $Dx > 0$, x is in class 1, and if $Dx < 0$, x is in class 0. For an unknown sample, u , the maximal marginal classifier can be written in parametric form using a slope and intercept, as in Equation (7). This is performed in such a way that includes every point in the sample (Kuhn & Johnson, 2013).

$$Du = \beta_0 + \sum_{j=1}^n \alpha_j x_{ju} = \beta_0 + \sum_{i=1}^n \alpha_i x_{iu}; \quad \alpha_i \geq 0 \quad (7)$$

In the case where the classes are not separable, the set of positive α 's define points on the boundary of the margin. These points define the support vectors and the prediction equation (Kuhn & Johnson, 2013). If the boundary between classes is not linear, then a kernel function, K , is substituted as shown in Equation (8).

$$Du = \beta_0 + \sum_{i=1}^n \alpha_i K(x_{iu}); \quad \alpha_i \geq 0 \quad (8)$$

A kernel is a function that describes the similarity between observations (James et al., 2013). In non-linear SVMs, there are several kernel functions to include polynomial, radial, hyperbolic tangent, and others (Kuhn & Johnson, 2013). By adjusting parameters within each function, the SVMs can be tuned to prevent over-fitting. In multiclass problems where C , the number of classes, is greater than two, there are several methods to extend the binary classifier such as one-verse-all and one-verses-one. In the prior, C SVMs are fit with each one considering two classes: a class c and a class consisting of all other classes. Test observations are assigned a classification based on the largest fitted value from the C SVMs evaluated at each observation. In one-verse-one, K^2 SVMs are created with each one comparing a pair of classes. Each test observation is evaluated by each SVM and the test observation is assigned a classification based on the most frequently occurring class (James et al., 2013).

4. Results and Discussion

4.1 Descriptive Statistics of the Patient Population

Before discussing the results obtained from applying unsupervised and supervised machine learning methods to the patient data, we first describe the EMR data elements to understand better the study population. Table 2 provides descriptive statistics of several patient characteristics, Hip/Knee Disability and Osteoarthritis Outcome Scores (HOOS/KOOS), and the most prevalent medical and disease diagnoses found in the patient data.

Table 2. Description of the Study Population

Clinical Data Overview			
Dates: 30/12/2013-30/08/2015			
Prevalence (or absence) of clinical conditions	#	%	
Obesity (BMI>30 by ICD9 code)	544	66.8%	
Hypertension	482	59.2%	
Never smoked	457	56.1%	
Osteoarthritis lower leg	450	55.3%	
Esophageal Reflux	369	45.3%	
Anemia acute (blood loss)	222	27.3%	
Osteoarthritis pelvis	214	26.3%	
High Cholesterol	148	18.2%	
Obstructive Sleep Apnea	144	17.7%	
Hypothyroid	123	15.1%	
Depression disorder	118	14.5%	
Diabetes	117	14.4%	
Prior TJA	105	12.9%	
Long term aspirin use	105	12.9%	
Anemia	93	11.4%	
Current Smoker	92	11.3%	
Asthma	89	10.9%	
Anxiety State	80	9.8%	
Coronary Atherosclerosis of native coronary artery	59	7.2%	
Retention of Urine	58	7.1%	
Hip and Knee Scores (HOOS/KOOS) from patient survey	Median	Mean	StdDev
Pain	32.50	29.99	24.46
Symptom	35.71	32.86	26.00
Function Daily Living	32.35	30.67	27.09
Function Sports, Recreational Activities	0.00	15.47	22.20
Quality of Life	12.5	16.17	18.03
Patient Characteristics	Median	Mean	StdDev
Age	65	63.122	11.40
BMI	32.45	33.21	7.06

4.2 Results of Unsupervised Machine Learning

Given the number of available predictors, we first experimented with dimension reduction techniques in order to find the right combination of predictors that could accurately represent the key relationships between patients while still being interpretable. Of the 724 available variables, the majority of them are binary (over 650 are binary) and, of the ICD-9 disease codes, only 20 of them occurred with frequency greater than 55 (6%). To reduce the size of the initial patient data set, we combined several of the weight-related variables into an obesity variable. A patient is obese if either his/her BMI is greater than 30 or he/she is diagnosed as obese or morbidly obese (ICD-9 codes 278.0 and 278.01). Since there were five patients without a BMI score, creating a new obesity variable allowed us to keep those patients in the data set. We also combined four of the asthma related ICD-9 codes into one binary variable for asthma.

The EMR data had redundant fields for smoking and tobacco use, so we combined three into two: patient never smoked and patient is a current smoker. The two separate ICD-9 codes for prior knee and hip replacement were combined into one binary variable: prior joint replacement. Finally, we excluded all remaining ICD-9 codes that occurred with a frequency less than 55 times and removed the patients with missing HOOS/KOOS scores. The final data set included 29 variables and 746 observations (741 if you exclude patients without a BMI score).

We ran both k -means and hierarchical clustering using Ward's method on two data sets. All methods were performed in R 3.2.0 using the package "stats" version 3.3.0 (R-Core Team, 2016). Both data sets contained 746 patients. The first had all 29 variables discussed above and the second was a subset that only contained the continuous variables (age, BMI, LOS, and five HOOS/KOOS scores). Since the larger data set contained mixed variables, we converted the matrix of 29 variables into a dissimilarity matrix using Gower's coefficient. The best value of k that maximized average silhouette distance was $k = 2$. The resulting two clusters were different but not significantly different in factors such as age, BMI, or HOOS/KOOS scores. We then applied the hierarchical clustering algorithm using Ward's distance. We cut the dendrogram at a height that gave six clusters. The clusters are fairly balanced and have some separation between key demographic and HOOS/KOOS scores. Figure 1 shows the dendrogram with $k = 6$.

Figure 1. Dendrogram of Patient Data using Ward's Method with $k = 6$

With the exception of cluster 2 (green), the clusters are evenly balanced. Table 3 shows the summary of each cluster. The mean age per cluster ranges between 61 and 65. Every cluster has a high percentage of obese patients and clusters 1 and 2 are likely total knee patients given their high percentage with osteoarthritis of the lower leg, whereas clusters 3, 4 and 5 are likely comprised of hip patients as osteoarthritis of the pelvis is a prevalent medical condition. Also cluster 2 and 5 contain the patients with the most painful conditions just prior to surgery as their

HOOS/KOOS scores are very low (low score indicates significant deviation from normal lifestyle).

Table 3. Summary Data from Hierarchical Clustering using Ward's Method

Cluster	Freq.	Age	BMI	LOS	Pain	Symptoms	Function Daily Living	Function Sports Rec.	Quality of Life	Most Prevalent Medical Condition
1	109	65.62	32.61	63.36	46.03	48.98	47.14	27.28	28.23	Hypertension (50%) Reflux (51%) Osteoarthritis: lower leg (86%) Obese (57%)
2	254	63.54	34.64	63.59	36.45	39.75	37.48	19.15	19.65	Never smoked (72%) Not smoked in >1 year (98%) Obese (79%) Osteoarthritis: lower leg (88%) Reflux (44%) Hypertension (57%)
3	135	61.79	30.90	69.73	8.84	12.01	7.85	2.99	2.87	Obese (53%) Osteoarthritis: pelvis (63%) Hypertension (46%)
4	87	63.32	33.87	62.46	38.67	41.71	40.30	17.18	19.33	Not smoked in >1 year (98%) Never smoked (68%) Obese (76%) Osteoarthritis: pelvis (57%) Hypertension (66%)
5	79	63.19	30.92	64.24	35.18	38.61	36.30	19.78	21.52	Obese (48%) Reflux (47%) Osteoarthritis: pelvis (89%) Hypertension (66%)
6	82	61.51	34.69	64.83	8.44	8.72	8.91	2.36	2.52	Obese (76%) Reflux (62%) Hypertension (67%) Sleep Apnea (41%)

Although the results from the hierarchical clustering were distinguishable, there was not significant distance between age, BMI, and pain scores. As an alternative method, we also conducted clustering on a reduced variable matrix where we clustered on just 8 variables (age, BMI, length of stay, and the five HOOS/KOOS measures). We applied k -means clustering to the reduced data set and evaluated the within-group sum of squares over a range of k values. Figure 2 shows the results.

Figure 2. Within-Group Sum of Square Errors for k -means Clustering using Reduced Data Set

Based on Figure 2, we selected $k = 6$ despite the significant decrease in the marginal reduction in within-group sum of squares seen at $k = 2$. Again, we are looking to pull out some of the nuisance differences in the clusters. The results of the clustering are shown in Table 4. Again, we present six clusters. This time, we did not use any binary variables and were, therefore, able to use Euclidean distance versus Gower's dissimilarity measure. By using eight scaled continuous variables, six distinct clusters are presented below.

Table 4. Results of K -means Clustering on Reduced Data Set

Cluster	Freq	Age	BMI	LOS	Pain	Symptoms	Function Daily Life	Function Sports & Rec	Quality of Life
	.								

1	84	60.14	32.58	58.10	59.08	56.45	63.83	63.68	40.25
2	124	73.06	29.25	65.35	6.66	9.75	5.43	1.61	1.92
3	130	53.75	33.78	55.61	3.50	5.29	4.31	0.98	1.11
4	43	67.67	33.74	136.86	14.11	18.79	11.32	6.57	8.58
5	200	68.96	31.12	61.14	48.38	54.07	48.22	18.83	29.94
6	160	56.63	38.52	59.82	35.57	38.10	37.51	10.87	11.60

Cluster 4 has the least number of patients at 43, but it also has the longest average length of stay (LOS). This is a significant cluster for clinicians to understand as these patients' average age is within the coverage of Medicare, and they require the most resources of all the patients who have a joint replacement. Under value-based care policies such as the new comprehensive joint replacement program (CJR) by the Center for Medicare and Medicaid Services, hospitals will receive a fixed reimbursement for each patient regardless of the number of days they spend in the hospital. Cluster 4 patients also have relatively low HOOS/KOOS scores, indicating that their quality of life is severely impacted by their joint pain. Cluster 3 and cluster 6 patients are the youngest of the patients but differ in their BMI. Cluster 6 patients have an average BMI nearing morbid obesity. Cluster 3 patients are the youngest, obese on average, and have the lowest HOOS/KOOS scores. These are likely middle-aged men and women who suffer from joint pain but are also still relatively young and value their mobility. Their low length of stay indicates that they are motivated to return to work and their families. With the average age below 65, cluster 3 patients may also be conscience of their stay in the hospital, as they likely have private insurance and high copays. Cluster 2 and cluster 5 patients contain the oldest patients with average ages of 73 and 69 years old, respectively. These two clusters differ in that the older patients (cluster 2) are in more pain and discomfort.

Table 5 shows the results of hierarchical clustering using Ward's method on the reduced data set. The accompanying dendrogram is in Figure 3. The results are similar to the output from the *k*-means clustering output in Table 4. Hierarchical clustering also found a small size cluster (6) with relatively low HOOS/KOOS scores and a high length of stay. Cluster 6 in Table 5 differs from cluster 4 in Table 4 by age, whereas the average age in the latter table is slightly lower. Cluster 1 and cluster 2 in Table 5 correspond to cluster 6 and cluster 2 in Table 4, respectively. With a small data set and few predictor variables, the results from *k*-means and hierarchical clustering methods are similar; however, the *k*-means results have slightly better distribution amongst the clusters. Additionally, the computer processing time for both hierarchical clustering and *k*-means was less than 1.2 seconds when run on a Hewlett-Packard laptop with an Intel Core i7 processor. With a small patient data set, computational time is not a significant factor.

Table 5. Results of Hierarchical Clustering using Ward's method

Cluster	Freq.	Age	BMI	LOS	Pain	Symptoms	Function Daily Life	Function Sports &	Quality of Life

Rec									
1	185	59.48	36.75	61.25	33.83	35.33	34.85	8.89	10.38
2	76	77.03	27.57	68.29	9.21	17.86	8.26	1.68	3.62
3	231	64.37	31.86	57.79	49.48	53.35	50.38	26.06	31.48
4	163	58.63	33.22	58.44	1.86	3.36	2.08	0.41	1.11
5	41	65.34	33.53	57.34	64.38	63.40	69.80	75.46	46.36
6	45	64.38	34.48	136.63	19.73	22.00	18.07	11.39	9.59

Figure 3. Dendrogram of Hierarchical Clustering on Small Data Set

4.3 Results of Supervised Machine Learning

There are many supervised machine learning models readily available and each has its own strengths. We chose four, three of which are tree-based classification models and common in the literature. Since there is no one best model for all data sets, we tested several and selected the one with the best AUC. As mentioned previously, AUC is the area under the receiver operator curve which is a plot that traces the performance of a classifier as the threshold value is changed. In other words, ROC is a plot of the true positive rate (y-axis) versus the false positive rate (x-axis) at different thresholds. An ideal predictor, or one that can perfectly separate the classes of a response variable, has an AUC of one, whereas a classifier that is no better than random guessing would have a AUC around 0.5. The ideal classifier hugs the upper left corner of the unit square. AUCs greater than 0.8 are generally considered good classifiers. We also divided our data into a training set and testing set. As in regression modeling, creating training, validation, and test data is an important step to prevent the problem over-fitting.

After performing unsupervised machine learning to determine homogeneous subgroups of patients within the TJA patient population, we proceeded with the two-stage methodology by appending the generated cluster label to each patient observation to then apply supervised machine learning. As mentioned previously, we built four classification models namely decision

trees, random forests, bagging, and support vector machines for predicting which segment a new patient observation would fall into. Figure 4 shows a plot of the one-versus-all ROC curves derived from the four classification models that were trained on 80% (randomly assigned) of the original training data, then validated on the remaining 20% of the training data. All models performed very well, with the decision trees having the lowest multi-class AUC at 83.2% and random forests having the highest at 92.8%. Note that we improved the decision tree fit with a pruned version via cross-validation using a miss-classification scoring method. Table 6 provides the multi-class AUCs derived from the four classification models. Note that these same type of classification models can then be modified to predict other patient health outcomes, such as patient readmission risk, length of stay, and medical supply costs. Here, the patients' health market segment identifier becomes a new predictor in the prediction model.

Figure 4. One-versus-all ROC Curves for Four Classification Models

Table 6. Multi-class AUC by Classification Model

Model	Decision Tree	Random Forest	Bagging	Support Vector Machine
AUC	0.8322	0.9278	0.9006	0.8673

Although applying supervised learning to predict a patient's health market segment after performing *k*-means clustering may seem redundant, supervised learning methods, namely decision trees, provide a useful advantage in healthcare. Decision trees, such as the one shown in Figure 5, are user-friendly and help improve the interpretability and flexibility of the predictive model. Since decision trees are based on the actual predictor variables, clinicians can easily determine a patient's cluster by simply following the tree branches until the classification is evident. This does not require calculating distances between cluster centers and new patients. Additionally, since some patients have complex medical conditions and require care from multiple clinicians in a single visit, decision support tools must be easy-to-use and universally understood.

Figure 5: Decision Tree for Total Joint Replacement Surgery Patient

The decision tree in Figure 5 above provides a graphical representation of the classification model. For example, if a new patient is recovering from a total knee replacement surgery and his/her pre-surgery KOOS pain score is greater than 19.44, HOOS sports and recreation activity score is less than 41.88, HOOS quality of life score is less than 15.63, and BMI is greater than 33.4, then the patient is assigned to cluster 6.

During the clustering phase, all observations are assigned to the nearest cluster center. Since we are using the reduced data set, the closeness measure is based on Euclidean distance. In the classification phase, every patient's cluster assignment number is appended to their record and used as the response variable. Up to this point, the same data is used for both classification modeling and clustering.

To test the effectiveness of this two-stage methodology, we acquired a new data set (from September 2015), assigned every patient to the closest cluster center, and then predicted the cluster using the random forests model (since it had the largest AUC value compared to the other classification models). Table 7 below shows the results. The predicted cluster from the random forests model are on the left side, and the assigned clusters based on distance to nearest cluster center are across the top. We see that nine patients were misclassified out of 49, resulting in an accuracy rate of 81.6%.

Table 7. Predicted versus Assigned Segment for Testing Data

Predicted\Assigned	1	2	3	4	5	6
1	11	2	2	0	0	0

2	1	14	1	0	1	0
3	0	0	3	0	0	0
4	0	1	0	1	0	1
5	0	0	0	0	2	0
6	0	0	0	0	0	9

In this research, we demonstrated that EMR data is suitable for clustering total joint replacement patients into distinct clusters that describe the likeness of their attributes. This information alone is applicable to the healthcare industry, as it identifies patients' segments that may require post-operative interventions to improve discharge times. In Table 4, we showed that patients in cluster 4 have an average BMI around 33, are on average Medicare patients (over 65), and have low HOOS/KOOS scores. Although this cluster represents only 6% of the total population, they spend over twice as long in the hospital. In an era of hospital overcrowding and ever-increasing healthcare costs, identifying cluster 4 patients during pre-op enables the hospital to use lower-cost interventions to speed discharge.

Our study also determined that there are two large segments of the population (cluster 3 and 6) that contain younger patients who are likely still working full-time, have private insurance, and want to get back to normal activities. The clusters differ in the degree of obesity and pain/discomfort levels, but they present an opportunity for healthcare organizations and health insurers as they look to increase preventative care and focus on wellness and value-based initiatives. The next step in analyzing these clusters is to research patient insurance types, frequency of routine care appointments, and other touchpoints that healthcare providers have with these patients.

After clustering each patient, we showed that classification models can accurately predict a patient's segment with the original data and with new data. Applying the best classification model (random forests) to a new 30-day sample of patient data accurately predicted a patient's cluster better than 80% of the time. This implies that in our small test, a new patient's health market segment can be accurately predicted without performing clustering.

Future health market segmentation research should include larger datasets that span multiple service lines within the hospital. Additionally, merging cluster information with patient attributes holds great potential for predicting various outcomes such as patient length of stay, discharge location, and expected costs of treatment. In an era of reduced reimbursements and increased demand for services, predicting patient outcomes offers insights to clinical decision makers that can inform pre- and perioperative healthcare decisions. With larger data sets comes the potential for longer computational time. As shown in Table 8, the methods used in this study were easily managed on a standard laptop computer. With big patient data, however, processing times will increase and could quickly become a discriminating factor in determining which methods to use.

Table 8. Processing Time for all Machine Learning Methods Applied to Reduced Data Set

Method	Computational Time (sec)
K-means Clustering	1.09
Hierarchical Clustering	0.03
Decision Trees	0.01

Bagging	0.44
Random Forests	0.42
Support Vector Machines	0.11

5. Conclusions

The advent and widespread use of the EMR not only enables clinicians to access patient data in real-time, but it also provides the data warehouse to support health market segmentation efforts. Through secondary use of EMR data, healthcare organizations can efficiently target specific segments of their populations without requiring additional patient information. This study is a small step toward empowering patients, providers, and health marketers with valuable health market segment information with which to target select groups of patients. Clinicians can use this information to adopt specific protocols during the pre-, peri-, and post-operative phase of TJA. Health insurers can use this information to isolate high costs and high-risk patients and then incentivize them and their primary care providers to take appropriate, patient-customized interventions to reduce health risks. Patients who are interested in their health outcomes can learn from their segment what actions to take now to delay or prevent common ailments.

This study is small in scope and focuses on one specific surgical procedure. Health market segmentation holds great potential if it can be applied to large patient populations without the need to conduct time-consuming surveys, and if it can provide detailed demographic, geographic, and clinical information regarding each segment. Additionally, exploiting the classification models for new patients and to help predict health outcomes for existing patients is a powerful tool for clinical decision support. EMR technology is quickly becoming the norm in US healthcare delivery, and, despite the size of patient databases, many of these clustering and classification algorithms remain computationally inexpensive.

Hospitals, health systems and health insurers already collect an enormous amount of patient data to include physical characteristics (e.g., age, weight, height), as well as past medical conditions, lab results, radiology reports and images, and a plethora of time-series data pertaining to each visit to a network provider (those with access to the patient's EMR). Modern EMRs store all patient data in a centralized and searchable database. The EMR provides real-time access to providers in the clinical setting, but it also holds the potential to tell a much bigger story about a patient's past, current, and future health, such as what types of treatments or health promotions they may respond to, whether they value customer service, prefer messages via an interactive personal health record, or value routine care.

The two-stage methodology presented in this study demonstrates that the integration of both unsupervised and supervised machine learning methods show congruence in their ability to predict a patient's segment. Applying the predictive model is computationally efficient and, when using a decision tree, provides an intuitive visual aid for clinicians. The future of this research entails incorporating the predictive modeling directly into the EMR to provide clinicians with real-time access to a patient's cluster and real-time decision support as to what interventions would best suit the patient. We believe that computationally efficient predictive models could easily run in the background of the EMR and update when new information or a new patient is added. Although larger patient databases would increase computation time to generate the initial clusters and predictive models, the application time of the predictive models to each patient should remain constant.

In an era of skyrocketing healthcare costs and demand for services, understanding how patients cluster and respond to health promotions presents an opportunity to efficiently and effectively target segments of the market with health promotions tailored specifically to positively impact health outcomes. As healthcare costs increase, the trend for employers to shift more of the financial burden to individuals will continue and, as a result, will cause some consumers to seek personalized healthcare solutions to minimize their risks. The old adage that an “ounce of prevention is worth a pound of cure” is more relevant than ever. Health market segmentation holds the potential to enhance personalized and precision medicine by allowing healthcare providers to efficiently find and target at-risk or at-benefit health market segments.

Acknowledgments

This study is based upon work supported by the National Science Foundation under Grant No. IIP-1361509. In addition, the second author was supported by the National Science Foundation under Grant No. DGE1255832. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Pennsylvania State University. The authors would like to thank Dr. Charles M. Davis, III, M.D., Ph.D, and his team at the Penn State Hershey Medical Center Hip and Knee Joint Arthroplasty Division.

References

- Axén, I., Bodin, L., Bergström, G., Halasz, L., Lange, F., Lövgren, P., & Jensen, I. (2011). Clustering patients on the basis of their individual course of low back pain over a six month period. *BMC Musculoskeletal Disorders*, 12:99.
- Berg, C., Ling, P., Guo, H., Windle, M., Thomas, J., Ahluwaia, J., & An, L. (2010). Using market research to characterize college students and identify targets for influencing health behaviors. *Social Marketing Quarterly*, 16(4), 41-69.
- Carroll N., & Gagnon J. (1983). Identifying consumer segments in health services markets. An application of conjoint and cluster analysis to the ambulatory care pharmacy market. *Journal of Health Care Marketing*, 3(3), 22-34.
- Centers for Disease Control and Prevention (2011). What is health marketing? Accessed 13 November 2014, available from <http://www.cdc.gov/healthcommunication/toolstemplates/whatishm.html>.
- Centers for Disease Control and Prevention (2015): Gateway to health communication & social marketing practice 2015. Accessed 19 September 2015, available from <http://www.cdc.gov/healthcommunication/healthbasics/whatishc.html>.
- Cheng, B., Chang, C., & Liu, I. (2005). Enhancing care services quality of nursing homes using data mining. *Total Quality Management & Business Excellence*, 16(5), 575-596.

- Cran (2015). R website. Accessed 5 December 2015 at <https://cran.r-project.org/web/packages/tree/tree.pdf>
- Greenspun, H. & Coughlin, S. (2012). The U.S. health care market: a strategic view on consumer segmentation. Deloitte Center for Health Solutions. Accessed 20 November 2015, available from <http://www2.deloitte.com/content/dam/Deloitte/us/Documents/life-sciences-health-care/us-lhsc-mhealth-in-an-mworld-103014.pdf>.
- Haughton, D., Legrand, P., & Wooford, S. (2009). Review of three latent class cluster analysis packages: Latent GOLD, poLCA, and MCLUST. *The American Statistician*, 63(1), 81-91.
- Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3), 264-323.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.
- Kennett P., Henson, S., Crow S., & Hartman, S. (2005). Key tasks in healthcare marketing: assessing importance and current level of knowledge. *Journal of Health and Human Services Administration*, 24(4), 414-427.
- Kent, P., Jensen, R., & Kongsted, A. (2014). A comparison of three clustering methods for finding subgroups in MRI, SMS or clinical data: SPSS TwoStep Cluster analysis, Latent Gold and SNOB. *BMC Medical Research Methodology*, 14(113).
- Kim, Y., Oh, Y., Park, S., Cho, S., & Park, H. (2013). Stratified sampling design based on data mining. *Healthcare Informatics Research*, 19(3), 186-195.
- Kolodinsky, J. & Reynolds, T. (2009). Segmentation of overweight Americans and opportunities for social marketing. *International Journal of Behavioral Nutrition and Physical Activity*, 6:13, 1-13.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
- Lee, E. (2012). Data mining application in customer relationship management for hospital inpatients. *Healthcare Informatics Research*, 18(3), 178-185.
- Liu S. & Chen, J. (2009). Using data mining to segment healthcare markets from patients' preference perspectives. *International Journal of Health Care Quality Assurance*, 22(2), 117-134.
- Lie, S., Kiang, M., & Brusco, M. (2012). A unified framework for market segmentation and its applications. *Expert Systems with Applications*, 39, 10292-10302.

- Lynn, J., Straube, B., Bell, K., Jencks, S. & Kambic R. (2007). Using population segmentation to provide better health care for all: The Bridges to Health Model. *The Milbank Quarterly*, 85(2), 185 – 208.
- MacLennan, J. & Mackenzie, D. (2000). Strategic market segmentation: An opportunity to integrate medical and marketing activities. *International Journal Medical Marketing*, 1(1), 40 – 52.
- Malhotra, N. (1989). Segmenting hospitals for improved management strategy. *Journal of Health Care Marketing*, 9(3), 45-52.
- Moss, H., Kirby, S., & Donodeo, F. (2009). Characterizing and reaching high-risk drinkers using audience segmentation. *Alcoholism Clinical and Experimental Research*, 33(8), 1336-1345.
- Newcomer, S., Steiner, J., & Bayliss, E. (2011). Identifying subgroups of complex patients with cluster analysis. *The American Journal of Managed Care*, 17(8), e324-332.
- Penn State Hershey Medical Center (PSHMC, 2015) website, accessed on 10 December 2015 at <http://www.pennstatehershey.org/web/guest/home/aboutus/history>
- Pires, G. & Stanton, J. (2008). Marketing issues in healthcare research. *International Journal on Behavioral and Healthcare Research*, 1(1), 38 – 60.
- R-Core Team and contributors worldwide. (2016). The R Stats package. R package version 3.3.0. Retrieved January 28, 2016.
- Ross, C., Steward, C, & Sinacore, J. (1993). The importance of patient preferences in the measurement of health care satisfaction. *Medical Care*, 31(12), 1138 – 1149.
- Rubio, D., Schoenbaum, E., Lee, L., Schteingart, D., Marantz, P., Anderson, K., Platt, L., Baez, A., & Esposito K. (2010). Defining translational research: implications for training. *Academic Medicine: Journal of the Association of American Medical Colleges*, 85(3), 470-475.
- SPSS. (2001). The SPSS TwoStep Cluster Component: A scalable component enabling more efficient customer segmentation. Technical Report. Accessed on 26 November 2015 from http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf.
- Suragh, T., Berg, C., & Nehl, E. (2013). Psychographic segments of college females and males in relation to substance use behaviors. *Social Marketing Quarterly*, 19(3), 172-187.
- Tynan, A. & Drayton, J. (1987). Market segmentation. *Journal of Marketing Management*, 2(3), 301-335.

- Ward, J. (1963), "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association*, 58, 236–244.
- Wind, Y. (1978). Issues and advances in segmentation research, *Journal of Marketing Research*, 15, 317-338.
- Woodside A., Nielson R., Walters R., & Muller G. (1998). Preference segmentation of health care services: the old-fashioned, value conscious, affluents, and professional want-it-alls. *Journal of Health Care Marketing*, 8(2), 14-24.
- World Health Organization. (2014). Health promotion. Accessed 13 November 2014, available from http://www.who.int/topics/health_promotion/en/.
- Wu, H., Lin, S., & Liu, C. (2014). Analyzing patients' values by applying cluster analysis and LRFM model in a pediatric dental clinic in Taiwan. *The Scientific World Journal*, 2014, 1-7.