

Healthcare market segmentation and data mining: A systematic review

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

To cite this article: Eric R. Swenson, Nathaniel D. Bastian & Harriet B. Nembhard (2018): Healthcare market segmentation and data mining: A systematic review, Health Marketing Quarterly, DOI: [10.1080/07359683.2018.1514734](https://doi.org/10.1080/07359683.2018.1514734)

To link to this article: <https://doi.org/10.1080/07359683.2018.1514734>



Published online: 23 Nov 2018.



Submit your article to this journal [↗](#)





Article views: 29



View Crossmark data [↗](#)



Healthcare market segmentation and data mining: A systematic review

Eric R. Swenson^a , Nathaniel D. Bastian^b , and Harriet B. Nembhard^c 

^aPennsylvania State University, University Park, Pennsylvania, USA; ^bUnited States Military Academy, West Point, New York, USA; ^cOregon State University, Corvallis, Oregon, USA

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 that is premised upon individuals becoming more active participants in managing their health. A systematic review of the literature concerning healthcare market segmentation and data mining identified several areas for future health marketing research. Common themes included: (a) reliance on survey data, (b) clustering methods, (c) limited classification modeling after clustering, and (d) detailed analysis of clusters by demographic data. Opportunities exist to expand health-marketing research to leverage patient level data with advanced data mining methods.

KEYWORDS

Healthcare market segmentation; data mining; systematic review

Introduction

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.” Note that health marketing draws from traditional marketing theories and principles and adds science-based strategies to prevention, health promotion and health protection (CDC, 2011). 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).

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. Awareness of patients' preferences and styles needs to be taken into consideration. Strategies to encourage and support consumer engagement in healthcare are important for health care organizations (e.g., providers, health plans, pharmaceutical companies, etc.). 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 health costs. Healthcare 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).

Many patients are motivated to increase control over and improve their health based on individual circumstances, to include experience with a new medical problem, loss of employer-sponsored coverage, or their inability to obtain effective medical treatment due to cost or denial of coverage. As these circumstances increase across the patient population and as healthcare costs force many to go without insurance, it is anticipated that consumer activist segments will increase (Greenspun & Coughlin, 2012). Individuals' self-care is positively correlated to education and cultural perspectives about what constitutes health and healthcare. Further, with the onset of the Affordable Care Act and changes to employer-sponsored insurance coverage, individuals may experience higher levels of price sensitivity, forcing them to become more actively involved in their medical treatment decisions (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., antismoking, antiobesity, etc.). The authors proposed the importance of market segmentation in the healthcare services for better strategizing as per specific needs. As a result of improved information and communication technologies as well as health information technology (HIT), patients are now better empowered to improve their health.

Market segmentation is a critical step in health marketing which the CDC defines as a blending of social networking and health communication (CDC, 2015). Customer-based market segmentation provides the focus and precision required to enhance personalized healthcare by identifying the latent relationships between attributes found in individual health records, customer surveys, and or demographic data. These relationships help define patient

clusters or segments which hospitals, health systems, insurers, and affiliated health agencies can use to refine health marketing efforts. Understanding market segments can focus health communications, which are strategies to inform and influence health-based decision making (CDC, 2015). 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 health consumer participation in managing their own health. Additionally, understanding the uniqueness of market clusters can identify underserved segments and may help link existing health promotions to yet unexplored segments.

Market segmentation studies hold the potential to be a critical component of the National Institutes of Health translational research initiatives. Although the definition of translational research is not fully developed or defined and means different things to different people (Rubio et al., 2010), translational research is in essence the transfer of laboratory or bench-top research to larger and larger audiences. Ideally, research investments at a local level spawn best practices that ultimately become standard operating procedures that are widely adopted across the healthcare industry. Market segmentation allows translational researchers to efficiently locate desirable health market segments to target with new laboratory research; this will allow new clinical research to proliferate more rapidly to patient segments most in need.

Tynan and Drayton (1987) discussed the importance of market segmentation techniques in overall marketing strategy. They emphasized that segmentation helps marketers improve precision of the prediction of consumer responses to a marketing stimuli. They suggested that the main market segmentation bases could be geographic, demographic, psychological, psychographic, or behavioral. They argued 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, 2000). The authors argued that there are both driving and constraining forces acting for and against strategic market segmentation in any organization. These forces are mostly associated with limited resource availability and their optimum allocation along with the organizational culture.

There have been numerous health marketing research studies done over the past few decades. Common clustering methods include hierarchical and nonhierarchical 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. In several papers, the concept of market segmentation is discussed without a formal model or the application of data analytics. In this paper, we survey the data mining approaches to healthcare market segmentation. In addition to discussing the results and limitations, we provide recommendations for future opportunities in health marketing research.

Methods

Systematic search and article selection

In order to build the initial list of journal articles concerning healthcare market segmentation, we performed a systematic literature search using PubMed and PMC online database searches. Clustering and market segmentation are well-established and published methods; therefore, containing the search to medical related journals helped filter results. The search terms included clustering and market segmentation, health market segmentation, and healthcare market segmentation. After filtering queries initially by publishing date and key word search, further filtering via abstracts and ultimately full-text reviews reduced the number of articles to 12. [Figure 1](#) illustrates the article selection diagram.

Here are some descriptive statistics of 12 selected studies. Country breakdown: United States (6), Sweden (1), Korea (2), Denmark (1), Taiwan (2). Primary data mining method: Latent cluster analysis (1), hierarchical clustering (6), *k*-means (4), other (1). Type of data: survey (5), patient data/secondary use data/combo (7). Type of study: prospective (4), retrospective (8).

Description of data mining methods

Hierarchical clustering

A priori clustering. In a priori clustering, specific variables such as demographic, state of being, and geographic, are predetermined as the basis for clustering decisions. After all data is collected, clusters are formed around these specific predetermined variables. As compared with other clustering techniques, a priori clusters are easier to interpret, measure, and act upon given the observations fit the cluster. When the segmentation variables are not predetermined, resulting clusters must be interpreted to understand why they formed and what types of observations fit the cluster.

K-means clustering. *K*-means clustering is an unsupervised statistical learning technique that separates n multidimensional observations into k clusters based on the similarity between the observation and the centroid of the

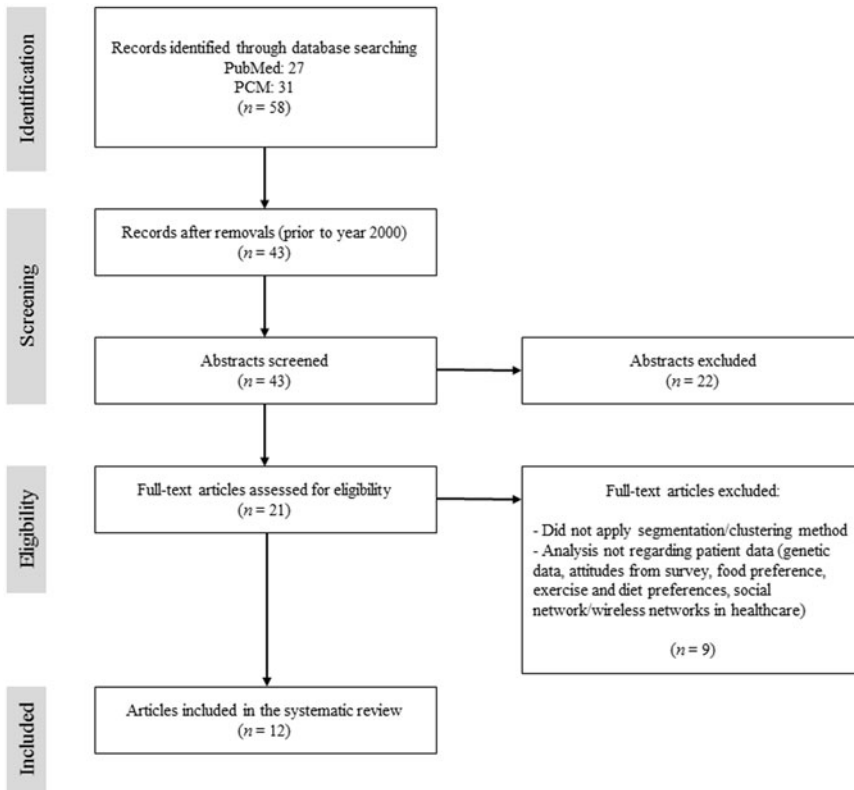


Figure 1. Article selection diagram.

cluster. The 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 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 most common for continuous attributes is the squared Euclidean distance (Jain, Murty, & Flynn, 1999).

The k -means clustering algorithm is iterative and at each step calculates the centroid of each cluster, then compares each observation to the centroid based on a similarity measure. Observations are reassigned to clusters based on maximizing similarity between the observation and clusters centroid. The process repeats until a predetermined 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 ease of use and ability to handle large data sets. The k -means clustering algorithm is susceptible to initial starting conditions, which can prevent it from reaching a global minimum. It works best when multiple starting points are used.

Hierarchical clustering

Hierarchical clustering covers both agglomerative and divisive clustering. In each case, the method starts with a set of n -multidimensional observations. The difference being that agglomerative hierarchical clustering starts with n clusters and terminates with one cluster and divisive clustering starts with one cluster and subdivides into n clusters. The methods are similar but approach the clustering from different sides, one being construction and one be division. 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 agglomerative clustering, clusters are traditionally joined based on a minimum distance measure or maximum similarity measure. 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 link which combines clusters that have the minimum of the maximum pairwise distance between any two points (from different clusters) and single link which combines two clusters if the distance between them is the minimum of the pairwise distances (Jain et al., 1999).

SPSS TwoStep cluster analysis

This approach is used in the SPSS software package. The clustering algorithm is a combination of several techniques. In the first step or precluster phase, sequential clustering is applied to each observation (SPSS, 2001). Observations are passed down a decision tree and are either assigned to a cluster of similar observations or the observation forms a new cluster. This output of step one is a set of subclusters, p , where p is less than or equal to n , the number of observations. In step 2, agglomerative hierarchical clustering is applied to the p subclusters to form the desired number of k clusters. By design, the subclustering step places observations into at most 512 subclusters. This reduction in size make subsequent hierarchical clustering feasible. This technique can be applied to large data sets (SPSS, 2001).

Latent class analysis

Latent class analysis (LCA) is a probability-based clustering technique that seeks to cluster observations based on unobserved variables. LCA uses a stochastic approach to find likely distributions with the data and the placement of observations within the distributions such that two or more observed

variables are conditionally independent of each other based on the condition that they are in the same latent class (Kent, Jensen, and Kongsted, 2014). The cluster model is

$$P(y_n|\theta) = \sum_{j=1}^S \pi_j P_j(y_n|\theta_j) \quad (1)$$

where S is the number of clusters, y_n is the n th observation of the observable (not latent) variable, and π_j is the prior probability of membership in cluster j . P_j is the probability of y_n given θ_j (cluster specific parameters) (Haughton et al., 2009). LCA takes a model based approach to clustering and has been used in market segmentation studies. It is fairly common in marketing, economics, and the social sciences and used as an alternative to the common distance based methods (hierarchical, k -means).

Description of distance/similarity measures

Ward's method: Ward's method is also known as 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).

Gower's dissimilarity coefficient: 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.

$$s_{ijk} = 1 - \frac{|x_{jk} - x_{ik}|}{R_k}$$

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

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

Results

A total of 12 studies were examined in significant detail based on the article selection diagram depicted in Figure 1. Table 1 shows the summary of the 12 articles.

There have been numerous papers written on healthcare market segmentation over the past 40 years. The advent of powerful computers and statistical learning software have expanded opportunities for exploring market segments through the use of big data sets. The 12 papers reviewed include many of the market segmentation and cluster techniques that are used in the broader literature regarding marketing studies. *K*-means clustering and hierarchical clustering are the predominate methods in these studies. Other methods such as a priori clustering and CHAID (or chi-squared automatic interaction detection) were cited in several of the articles published prior to 2000 (Carroll & Gagnon, 1983; Malhotra, 1989). The 12 papers included in this review started with a set a data and applied unsupervised learning techniques to find homogenous clusters or segments within the population.

Diversity of studies: All 12 studies are published in peer-reviewed journals and include a mix of professionals to include medical doctors and PhD researchers from economics, healthcare science, industrial engineering, economics, and marketing. The studies range from analysis of clinical populations (Axén et al., 2011; Kim et al., 2013; Newcomer, Steiner, & Bayliss, 2011) to segmentation studies on survey data (Kolodinsky & Reynolds, 2009; Liu & Chen, 2009; Moss, Kirby, & Donodeo, 2009; Suragh, Berg, & Nehl, 2013; Berg et al., 2010). Of the papers that used survey data, two looked at college student substance abuse behaviors (Berg et al., 2010; Suragh et al., 2013), one looked at customer preference for healthcare service and clustered patients based on their preference and demographic attributes (Liu and Chen, 2009), and the last two used large survey data from a combination of the Behavioral Risk Factor Surveillance System (BRFSS), U.S. Department of Agriculture funded nationwide polls, and a mix of public and U.S. census data (Kolodinsky & Reynolds, 2009; Moss et al., 2009).

Two of the studies based on patient data investigated RFM (or 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

Table 1. Summary of 12 articles.

Study	Retrospective/ prospective	Setting	Sample size	Country	Outcome(s) measured	Factors used	Results	Method
Newcomer et al. (2011): Identifying subgroups of complex patients with cluster analysis.	Retrospective	HMO population; 20% of care expenditures and with two or more chronic medical conditions; data from CY2006/2007	15,480	USA	How patient's cluster around coexisting conditions demographics	Obesity, mental health conditions, diabetes, cardiac disease, COPD, kidney disease, cancer, gastrointestinal bleeding, chronic pain stroke, skin ulcer, dementia, fall, abdominal surgery, orthopedic surgery, back surgery, hip fracture	10 clinically relevant clusters grouped around single or multiple anchoring conditions. Mental health and obesity prevalent in all clusters.	Agglomerative hierarchical clustering; Ward's Algorithm; SAS V9.2 software
Kolodinsky and Reynolds (2009): Segmentation of overweight Americans and opportunities for social marketing.	Prospective	National level polling data; patient survey conducted by authors regarding food and lifestyle behaviors	581	USA	How patients clustered around food and lifestyle behaviors	Behavioral variables; personal and environmental factors. Ht, wt, computer use, smoker, gender, education, income, children, age, geographic region, residence location (urban/rural); knowledge of food pyramid; exercise	Five clusters (highest risk, at risk, right behavior/wrong results, getting best results, doing OK). 99% in highest risk were overweight	Two step cluster analysis with Schwartz's Bayesian Criteria. ANOVA and Chi squared used to determine whether cluster membership related to demographics. Used SPSS software.
Berg et al. (2010): Using market research to characterize college students and identify targets for influencing health behaviors.	Prospective	Survey of college aged students from Minnesota; diverse sample	2,700	USA	Health related measures, confidence and motivation, market research, what are the influencers of behavior	Demographic, psychographic (attitudes and interests), health-related variables	Three clusters: stoic individuals, thrill seeking socialists, and responsible traditionalists.	Hierarchical cluster analysis using Ward's Method. Used Gower's general dissimilarity coefficient then clustered on distance matrix products. Used ANOVA and chi-squared tests to compare variables across segments. Used SAS and SPSS software.
Kent et al. (2014): A comparison of three clustering methods for finding subgroups in MRI, SMS or clinical data.	Retrospective	Secondary use data; longitudinal studies; multiple data sets to include real data and randomly generated test data.	3x MRI data sets (412, 631, and 4,162 patients); 1x self reported lower back pain intensity data set (n = 1121), clinical data set (n = 543) based on	Denmark	Consistency across methods	Number of subgroups detected, classification probability of individuals in a subgroup, reproducibility of results, ease of use of software	Number of subgroups detected varied by method; certainty of classifying individuals into subgroups varied; finding were reproducible; ease of use and	Comparison of three common clustering methods using 9 data sets (five actual data sets and four artificial). Methods used are SPSS's TwoStep CA, Latent

(continued)



Table 1. Continued.

Study	Retrospective/ prospective	Setting	Sample size	Country	Outcome(s) measured	Factors used	Results	Method
Avén et al. (2011): Clustering patients on the basis of their individual course of low back pain over a six-month period.	Prospective (observational)	Outpatient chiropractic care; based on survey and clinical data.	176 patients with low back pain	Sweden	Change in pain intensity over time	26 parameters reduced to four via spline (non linear) regression; slope and intercept of regression line in early course; difference in slope between two regression lines; intersection estimate	Four clusters with distinct clinical courses as best overall	Ward's method and hierarchical clustering; SPSS, STATA, and Sleipner software used.
Liu and Chen (2009): Using data mining to segment healthcare markets from patients' preference perspectives.	Retrospective	U.S. not-for-profit health-care group; inpatients; telephone interviews/surveys	1,561	USA	How patients clustered; most important to least important attribute by cluster. Survey questions were demographic and statements that measured healthcare service preference.	24 Attributes reduced to five factors through cluster analysis; communication and empowerment, compassionate and respectful care, clinical reputation, care responsiveness, efficiency	Three clusters emerge: reputation driven, performance driven, and empowerment driven.	Hierarchical cluster analysis, Pearson correlation, and average linkage to measure similarity. Used R software and a mix of hierarchical and nonhierarchical methods plus Enterprise Miner.
Suragh et al. (2013): Psychographic segments of college females and males in relation to substance use behaviors.	Prospective online survey	Six college campuses in Southeast, USA; diverse male and female student population; 230 question survey conducted in 2010.	3,469	USA	How students clustered according to psychographic characteristics and substance use behavior	15 psychographic measures (sensation seeking, personality traits (5), 9 measures adapted from tobacco industry)	Three psychographic distinct clusters: safe responsible, stoic individuals, thrill-seeking socializers.	Hierarchical clustering; Ward's method, Gower's dissimilarity coefficient (due to nominal and ordinal values). Used t-statistic to determine optimal number of clusters. Used SPSS software.
Moss et al. (2009): Characterizing and reaching high-risk drinkers using audience segmentation.	Retrospective survey data analysis	Combination of 2004 U.S. survey data (BRFSS data plus Simmons Market Research Bureau data that consists of public records data, U.S. Census data, etc.).	>30,000 people	USA	Clustering of population based on high risk drinking behaviors/attitudes	Self-reported drinking episodes, frequency, and demographics.	66 audience segments with top ten analyzed in depth. Cyber-millennial cluster has highest concentration of binge drinkers. Laid back towners, city producers, metro newbies are in descending order the clusters of highest risk	Proprietary PRIZM software. Audience segmentation that creates 66 clusters from nationwide database.
	Retrospective			Korea				

Kim, Oh, Cho, and Park (2013): Stratified sampling design based on data mining.	Single specialty clinics and hospitals that conduct either general surgery or ophthalmology; 2011 data; combination of hospital and insurance data merged into a single database for analysis	442 clinics/hospitals that did general surgery; 715 facilities with specialty of ophthalmology	Classification of health-care providers	Type of location, population density, number of specialists, number of beds, number of inpatients per specialist, length of stay, costliness index, case-mix index, rate of annual change in number of inpatients per specialist	Four clusters of ophthalmology facilities, three clusters of general surgery facilities.	K-means clustering and decision tree induction to segment and classify health-care providers then to stratify them into five stratum. Used MATLAB software.
Lee (2012): Data mining application in customer relationship management for hospital inpatients.	University hospital; data from Jan to Dec 2009.	14,072 discharge records	Customer loyalty, Customer relationship model	Recency, frequency, monetary, LOS, certainty of selectable treatment, surgery, number of accompanying treatments, kind of patient room, department from which discharged.	Customers were classified as either loyal or ordinary. Demographic characteristics were overlaid on two clusters. Decision tree showed most important factor is LOS. 12 clusters based on LRFM.	k-means clustering, group comparison via t-test. Decision tree and logistic regression used to predict patients who were clustered as "loyal customers."
Wu, Lin, and Liu (2014): Analyzing patients' values by applying cluster analysis and LRFM model in a pediatric dental clinic in Taiwan.	Pediatric dental clinic; data from July 2009 to June 2011	1,462 patients (under 18 years old)	How pediatric dental patients cluster	LRFM (length, recency, frequency, monetary model), gender, age	12 clusters based on LRFM.	k-means and self-organizing maps; used SPSS Modeler 14.2 software.
Cheng, Chang, and Liu (2005): Enhancing care services quality of nursing homes using data mining.	Nursing home; study period April to March 2003.	407 nursing home residents	Patient clustering and CRM	K scale, LOS, Times of stay, discharge reason, no. of diseases, special passageways brought, no. of rehab outpatient visits, age	Four clusters of patients/residents. Used clusters to determine best care strategies based on expert opinion.	Demographic clustering using K-means; cluster size determined using MANOVA test of the discriminant analysis. Used SPSS and Intelligent Miner V6.1.

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 postcluster 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. Berg et al. (2010), Kolodinsky et al. (2009), Moss et al. (2009), and Suragh et al. (2013) and 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 United States and arrived at the same number of clusters with strikingly similar names and cluster demographics. The prior study was in Minnesota and the latter was a larger study conducted at six universities in the Southeast. 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. (2010) 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. This study used a proprietary software called PRIZM that clusters large public data sets into 66 segments. The goal of this study is similar to the college surveys in that it tried to form homogenous subgroups, decompose each by the strength of their attributes, and then use that information to target at-risk segments with marketing strategies aimed at behavior modification.

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 Berg et al. (2010), Kolodinsky et al. (2009), 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. (2010) borrow from the tobacco industry and Kolodinsky et al. (2010) borrow survey methods from the food industry.

Liu and Chen (2009) and Kent et al. (2014) conduct market segmentation using different approaches but each applies multiple clustering techniques to verify the results. The prior uses survey data while the latter is based on secondary use data from a variety of studies. Liu and Chen (2009) use a mix of hierarchical and nonhierarchical methods and ultimately settle on a hierarchical clustering method that reduces the attributes from 24 to 5 yielding 3 distinct clusters. Kent et al. (2014) apply and compare three different methods to five real data sets and four randomly generated data sets to test reproducibility, likeness of outputs, and ease of use.

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

Country of origin, time frame and statistical software used in studies: Half of the studies were conducted in the United States, four in Southeast Asia and two Scandinavian countries. The majority of the 12 papers were published after 2009 and apply current data analytic software including SAS, SPSS, STATA, and R. All studies use data collected after year 2000. Three studies use SAS, six studies use SPSS, and R, MATLAB, PRIZM, STATA, SNOB LCA, and Latent Gold LCA are used less frequently. See Table 1 for specifics.

Methods used: The 12 papers in this review cover a breadth of subjects, methods, and outcomes. The common themes are market segmentation

and understanding how patients, clinics, students, or adults align with others of like attributes. The goal of these studies is to provide insight and an angle to better understand a population. The number of clusters or segments varies across studies which is consistent with cluster analysis in general. In most cases, the user must define the number of clusters ahead of time or must identify a condition upon which the algorithm stops. Hierarchical clustering is used in five studies (Axén et al., 2011; Berg et al., 2010; Liu and Chen, 2009; Newcomer et al., 2011; Suragh et al., 2013). In all but one of them, Liu and Chen (2009), Ward's method is used as the distance/similarity measure. In Liu and Chen (2009) Pearson's correlation is the similarity measure. Four of the studies use *k*-means clustering (Cheng et al., 2005; Lee, 2012; Kim et al., 2013; Wu et al., 2014).

Discussion

From the 12 articles investigated, we sought to learn how data mining techniques can be leveraged for conducting market segmentation with respect to patient preferences for healthcare attributes and exploring the patient segment demographic characteristics. The identification of gaps and opportunities provides the necessary direction for future health marketing research. A detailed discussion of the surveyed articles follows.

Liu and Chen (2009) employed cluster analysis techniques to conduct healthcare market segmentation using complicated psychographic variables and to reveal the benefits of data mining to understand consumers' psychological needs for improving healthcare services. The authors used survey data for patients who received care from a nonprofit healthcare group in 2006. Respondents were surveyed on 24 healthcare services attributes covering physiological care, psychological care, physical environment, and spiritual care. Factor reduction techniques reduced the number of factors to five and cluster analysis identified three segments. Factor reduction helped make the results more interpretable. Liu and Chen (2009) identified three healthcare market segments: reputation-driven, performance-driven, and empowerment-driven. Segments are subgroups with similar patient preferences in the whole healthcare market. Successfully identifying demographically well-defined consumer segments can assist hospital managers develop long-term business strategies and offer an optimal mix of products and services that meet customer needs and preferences (Ross et al., 1993; Woodside et al., 1998).

Kim et al. (2013) conducted a retrospective study using stratified sampling design based on *k*-means clustering and decision tree induction. Although their approach applied data mining techniques, they were focused on healthcare providers and not consumers. Their research was specific to

general surgery and ophthalmology into which they identified three clusters of general surgery clinics and hospitals and four clusters of ophthalmology clinics and hospitals. The three general surgery clusters were divided based on whether they were private or public and the number of inpatients. The ophthalmology hospitals clustered similarly with the additional factor of whether there were multiple specialists in the hospital. The authors' motivation was to improve sampling efficiency by creating homogenous strata of clinics and providers based on several factors including size and ratio of patient to specialist. After clustering, decision trees were applied to the two sets of data to further stratify hospital and clinics. For each type of hospital/clinic, the decision trees resulted in five strata based on three variables: number of inpatients per specialist, population density, and lengthiness index. The result of this study are intended to help with future healthcare policy decision making. The author's did not compare their method against other well-known classification methods nor did they discuss the robustness of their method nor stability of the clusters.

Lee (2012) applied data mining in a retrospective study to discover patient loyalty to a hospital and to model patient medical service usage. He studied customer relationship management marketing which is a process that segments customers to understand their behaviors with the goal of strengthening relationships with valuable customers. Patients were first classified into two groups: loyal and ordinary, based on recency, frequency, and monetary measures. Decision trees were then applied to each group (segment) to determine which factors/characteristics were most important in each segment. Logistic regression output was compared to the decision tree analysis and results were displayed on an ROC curve. This study is narrow on its approach to segmenting the market. It focuses on patient loyalty and uses frequency and monetary factors to determine segments. The author does not address why patients may use the same hospital frequently such as proximity to the next closest hospital, insurance considerations, or ability of patients to get to other facilities. Length of stay (LOS) is the leading factor in determining a patient's loyalty but LOS may be an unintended consequence of an unplanned hospitalization or a procedure gone wrong.

Chang et al. (2005) applied market segmentation, in particular *k*-means clustering, to a nursing home population in Taiwan to assist with customer relationship management. The goal of the study was to understand the characteristics of patient subgroups in a nursing home environment so that the staff can provide better, more customized, care to each patient. The authors use *k*-means clustering in combination with discriminant analysis to determine the appropriate number of clusters. Clustering was done with SPSS and Intelligent Miner V6.1. They showed that the population could

be clustered into four unique subgroups. Each subgroup was then analyzed by a team of professionals to determine the best care service strategy. Given the wide range of patient care needs in a nursing care setting, understanding how patients segment according to their conditions and needs can help management tailor care to existing and future residents.

Newcomer et al. (2011) applied hierarchical clustering, namely Ward's algorithm, to a large HMO patient data base to identify clinically similar subgroups. The patient population included over 15,000 adult patients who had at least two comorbidities and ranked in the top 20% for cost expenditure per year. Using agglomerative hierarchical clustering, Newcomer et al. (2011) merged clusters based on Ward's distance. To assess the stability of their algorithm, they divided the data set in half, create a dissimilarity matrix for each set using Jaccard's coefficient, then applied Ward's algorithm. Since the two data sets had similar cluster membership, the algorithm was applied on the entire data set. In 8 of the 10 resulting clusters with $k=10$ subjectively chosen, there was a clear dominate chronic condition that defined the segment. Newcomer et al. (2011) then analyzed each cluster by predominance of attributes and other comorbidities. The shortcomings in this study include the narrow focus on a single two-year data set and a lack of generalizability to other patient populations outside this HMO. Newcomer et al. (2011) did experiment with different clustering techniques but they do not show the results of the other methods nor how the outputs varied. The authors also do not discuss the relevance of their finding in mitigating chronic conditions or targeting at risk populations.

Kolodinsky et al. (2009) applied a social market segmentation approach in a behavioral study regarding peoples eating habits and the effect on body weight. The goal of this prospective study was to apply similar market segmentation techniques that the food industry uses to market products to understand people's behaviors and attitudes towards foods. Their survey questions were rooted in social learning theory and health belief model and interspersed with questions to understand socio-demographic attributes of the survey population. Kolodinsky et al. (2009) applied SPSS's TwoStep Cluster Analysis to the survey data initially excluding the demographic data. The 581 respondents clustered into five distinct segments primarily separated due to overweight risk. Segments were then analyzed using demographic data to better understand their composition. As in many of the health market segmentation studies, the study ends with a list of clusters distinguished based on a factor or series of factors directed related to the goal of the study. What is missing is the discussion on the relevance of the clusters and how machine learning can further help to classify new patients and match interventions to help with improved health outcomes.

Berg et al. (2010) and Suragh et al. (2013) each reported on the same topics with near identical results. Both considered college aged students and segmented them based on survey questions specifically designed to assess health behaviors and substance abuse. They both used hierarchical clustering albeit from different software packages (SAS and SPSS respectively) and they used Gower's general dissimilarity coefficient and Ward's method. Gower's coefficient was applied to handle both nominal and ordinal values in the survey results. Each research team concluded that their respective student population, which were drawn from different regions within the United States, segmented into the same three clusters: safe and responsible, stoics, and thrill seekers. Unfortunately, both studies conclude with three distinct segments. There is no discussion about the utility of each segment, what interventions could be used or have been used, and how statistical learning can further help classify new patients. Also, although Suragh et al. (2013) referenced the Berg et al. (2010) study, there were no parallels drawn or suggested.

Kent et al. (2014) is a comparative study of three different clustering methods on healthcare related data. In the study, the authors compare the clustering results of five real data sets and three artificial data sets across several criteria to include the number of segments or subgroups formed, the classification probability of observations into specific clusters, and the reproducibility of the clusters over 10 replications of each method on each data set. Kent et al. (2014) also compared methods for ease of use and interpretability of output. The methods tested in this paper included SPSS Two Step Cluster Analysis, Latent Class Gold, and SNOB latent class analysis. Although the results varied by methods and data set, the author's chose Latent Gold as the best method based on overall performance, sensitivity to determining the right amount of clusters, ease of use, and interpretability. All the methods provided highly reproducible results, but this could also be a function of starting seeds. The authors acknowledged that repeating test with different starting seeds could negatively impact reproducibility.

Axén et al. (2011) provides another example of a prospective market segmentation study using a hybrid mix of survey and clinical data. This study is based in Sweden and focused on 176 patients with low back pain. The authors used a SMS messaging service to track pain scores of patients over 26 weeks. This time series data was reduced using nonlinear spline regression to four measures that included the slope and intercept of the nonlinear regression line during the early part of the treatment course, the difference in slope between the early and late courses, and the intersection estimate. From this data, Axén et al. (2011) was able to cluster patients into four distinct segments. They used Ward's method, which is an agglomerative

hierarchical clustering method. Given the small size of the data set, this technique is computationally efficient. Given the nebulous nature of non-specific lower back pain, providing a clustering tool to categorize and segment the treatment population based on the change of pain related factors over time is a unique approach and application of data mining. As in many of the healthcare-related segmentation studies, the details of how data analytics can be used in the treatment or monitoring of treatment and intervention planning is missing.

Similar to the Berg and Suragh papers, Moss et al. (2009) apply market segmentation in a study of high-risk drinking behaviors. They use a combination of data from the BRFSS and other private and publically available survey data. The authors use a proprietary software called PRIZM that segments the data into 66 subgroups. The article analyzes the top 10 segments that are most likely to display highest risk behaviors. Each cluster is then dissected based on alcohol and tobacco use, digital communication use, sports and leisure activities, and media use to provide insight into how marketing strategies could be tailored to influence change in a subgroups behavior. Much of the details of the clustering technique are excluded from the paper.

Wu et al. (2014) conducted a market segmentation study of pediatric dental patients using SPSS's Modeler 14.2. The retrospective study applied *k*-means clustering and organizational maps to a sample of over 1,400 patients. The goal of the segmentation study was to understand how the patients clustered using attributes such as length of stay, recency of visits, frequency of visits, and monetary costs of visits. Demographic data such as age and gender were also included. The authors found 12 distinct clusters. The paper does not offer insight into how the clusters can or will be used to assist in better service or care delivery based on cluster assignment.

Gaps and opportunities in healthcare market segmentation

The predominance of healthcare market segmentation research over the past 26 years has focused on segmenting a healthcare population to identify segments for the purpose of behavior modification marketing and identifying subgroups within a larger but still specific group. There is a lack of studies based on patient-level electronic health record (EHR) data. In the 12 papers that met the inclusion criteria for this review, 5 were based on survey data and a sixth used a combination of survey data and clinical data. Three papers used RFM data in conjunction with customer responsiveness models, one used specific hospital/clinic data on facility usage, one used service specific data from both chiropractic care and imaging services, and the final paper used patient level data. Although EHRs have been in

existence for over a decade, only one study (Newcomer et al., 2011) took a large hospital data set and applied data mining techniques to cluster patients into meaningful segments. Understanding these segments will help health service providers, healthcare providers, and insurers target the right intervention and health services to “at risk” at “at benefit” subgroups.

Another gap in the healthcare market segmentation research is the lack of differentiation between market or audience segmentation and clustering. Many of the articles use clustering and segmentation interchangeably, whereas Liu et al. (2012) cite a few differences, namely, that clustering is a subset of segmentation that groups people or patients based on similarity (distance, likeness of needs, preferences, etc.). The clustering of people is a fundamental task of market segmentation and at one point in the late 1970s was synonymous with segmentation (Wind, 1978); however, market segmentation has evolved to include more than clustering or descriptive segmentation, and now includes predictive market segmentation (Liu et al., 2012). Furthermore, market segmentation research often involves multicriteria optimization because the goal often includes the application of the descriptive clusters into economic criteria related to responsiveness, identifiability, profitability, and accessibility (Liu et al., 2012). With multiple objectives, there may be no single optimal solution.

In the majority of the 12 papers reviewed, the authors stopped at the clustering solution. They applied some form of cluster analysis to define homogeneous or near homogeneous subgroups, but they did not use those clusters to aid in predictive market segmentation. The gap in methods is the absence of supervised statistical learning applied after the unsupervised methods assigned a cluster to each patient or observation.

Conclusion

The importance of market segmentation studies applied to healthcare cannot be understated. In fact, Kennett et al. (2005) discuss the importance of healthcare market segmentation and assess how well hospital executives understand and use various marketing tools to include market segmentation. They conducted a survey of healthcare executives and mid to upper level healthcare managers to assess how hospital leaders rate the importance of and their current level of knowledge of marketing. They found that although market segmentation was considered to be very important for hospitals it ranked in the top three tasks that that hospitals were least knowledgeable about (Kennett et al., 2005).

The majority of healthcare market segmentation studies over the past twenty years focus on either survey data or specific data sets with the purpose of segmenting a specific population. Although these studies help

define near homogenous clusters of patients, providers, or observations within the study, the studies end with defining the clusters. Market segmentation is more than just a study in defining a segment, it also includes predictive market segmentation in which the “decision maker seeks to optimize both within-segment homogeneity and segment level predictability” (Liu et al., 2012). Predictive segmentation is a key gap missing in most healthcare market segmentation papers.

Market segmentation is a well-known approach in marketing research and when applied to healthcare presents a great opportunity to identify subgroups of patients that share commonalities. 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 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 widespread use of integrated EHR databases across the United States presents an opportunity for healthcare providers to apply data mining methods to large healthcare data sets to enhance precision medicine. Hospitals, health systems and insurers already collect an enormous amount of patient data to include physical characteristics (age, weight, height), as well as past medical conditions, lab results, radiology reports and images, and a host of time-series data pertaining to each visit to a networked provider (those with access to the patient’s EHR). Modern EHRs store all patient data in a centralized and searchable database. The EHR 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. In an era of unprecedented demand for hospital services and rising health care costs, the old adage that an “ounce of prevention is worth a pound of cure” is more relevant than ever. Healthcare market segmentation holds the potential to enhance personalized and precision medicine by allowing health providers to efficiently find and target at-risk or at-benefit market segments. At-benefit is defined as a segment of the population that can greatly benefit from preventative care or interventions to help sustain or strengthen current health.

As an extension to this systematic review of healthcare market segmentation and data mining, future research will develop a two-phase healthcare market segmentation framework that uses EHR data to cluster

a hospital's patient population, then run a series of classification models to predict patient outcomes using their assigned cluster. This approach will combine both unsupervised and supervised statistical learning methods to big hospital data sets with the goal of increasing health promotion. The results of this analysis could benefit insurers, health systems, clinicians, and patients themselves as they seek better personalized healthcare solutions.

ORCID

Eric R. Swenson  <http://orcid.org/0000-0001-9044-0189>

Nathaniel D. Bastian  <http://orcid.org/0000-0001-9957-2778>

Harriet B. Nembhard  <http://orcid.org/0000-0001-6803-7641>

References

- Axén, I., Bodin, L., Bergström, G., Halasz, L., Lange, F., Lövgren, P. W., ... 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*(1), 99. doi:10.1186/1471-2474-12-99
- Berg, C. J., Ling, P. M., Guo, H., Windle, M., Thomas, J. L., Ahluwalia, J. S., & An, L. C. (2010). Using market research to characterize college students and identify targets for influencing health behaviors. *Social Marketing Quarterly*, *16*(4), 41–69. doi:10.1080/15245004.2010.522768
- 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 November 13, 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 September 19, 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. doi:10.1080/14783360500077476
- Greenspun, H., & Coughlin, S. (2012). The U.S. health care market: a strategic view on consumer segmentation. Deloitte Center for Health Solutions. Accessed November 20, 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., & Woolford, S. (2009). Review of three latent class cluster analysis packages: Latent GOLD, poLCA, and MCLUST. *The American Statistician*, *63*(1), 81–91. doi:10.1198/tast.2009.0016
- Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323. doi:10.1145/331499.331504

- 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), 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*(1), 13. 13.
- 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.
- Liu, Y., Kiang, M., & Brusco, M. (2012). A unified framework for market segmentation and its applications. *Expert Systems with Applications, 39*(11), 10292–10302.
- MacLennan, J., & Mackenzie, D. (2000). Strategic market segmentation: An opportunity to integrate medical and marketing activities. *Journal of 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–e332.
- Pires, G., & Stanton, J. (2008). Marketing issues in healthcare research. *International Journal of Behavioural and Healthcare Research, 1*(1), 38–60.
- 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., ... Baez, A. E. 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 November 26, 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*(301), 236–244.
- Wind, Y. (1978). Issues and advances in segmentation research. *Journal of Marketing Research, 15*(3), 317–338.
- Woodside, A., Nielson, R., Walters, R., & Muller, G. (1998). Preference segmentation of health care services: the old-fashioned, value conscious, affluent, and professional want-it-all. *Journal of Health Care Marketing, 8*(2), 14–24.

- World Health Organization. (2014). Health promotion. Accessed November 13, 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.