Expert Systems With Applications 60 (2016) 118–129

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

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, 362 Leonhard Building, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 22 December 2015
Revised 2 May 2016
Accepted 2 May 2016
Available online 6 May 2016

Keywords:
Health market segmentation
Health promotion
Data analytics
Machine learning
Total joint arthroplasty
Value-based healthcare

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.

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." Centers for Disease Control and Prevention (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, Nielson, 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

* Corresponding author.
E-mail addresses: ers187@psu.edu (E.R. Swenson), ndbastian@psu.edu (N.D. Bastian), hbn2@engr.psu.edu (H.B. Nembhard).

http://dx.doi.org/10.1016/j.eswa.2016.05.006
0957-4174/Published by Elsevier Ltd.

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, 2000). 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 (Axén et al., 2011; Kim, Oh, Park, Cho, & Park, 2013; Newcomer, Steiner, & Bayliss, 2011) to segmentation studies on survey data (Berg et al., 2010; Kolodinski & Reynolds, 2009; Liu & Chen, 2009; Moss, Kirby, & Donodeo, 2009; Suragh, Berg, & Nehl, 2013). 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, 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 (Kolodinski & Reynolds, 2009; Moss et al., 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, Lin, and Liu (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, Chang, and Liu (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. (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.

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