uOttawa

L'Université canadienne
Canada's university

PREDICTING HIGH-COST PATIENTS IN GENERAL POPULATION USING DATA MINING TECHNIQUES

By

Seyed Abdolmotalleb Izad Shenas

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN HEALTH SYSTEMS

Thesis Supervisors

Bijan Raahemi, Ph.D.

Craig Kuziemsky, Ph.D.

October 2012

Abstract

In this research, we apply *data mining* techniques to a nationally-representative expenditure data from the US to predict *very high-cost* patients in the top 5 cost percentiles, among the general population. Samples are derived from the *Medical Expenditure Panel Survey*'s Household Component data for 2006-2008 including 98,175 records. After pre-processing, partitioning and balancing the data, the final MEPS dataset with 31,704 records is modeled by *Decision Trees* (including C5.0 and CHAID), *Neural Networks*. Multiple predictive models are built and their performances are analyzed using various measures including *correctness accuracy*, *G-mean*, and *Area under ROC Curve*. We conclude that the CHAID tree returns the best G-mean and AUC measures for top performing predictive models ranging from 76% to 85%, and 0.812 to 0.942 units, respectively. Among a primary set of 66 attributes, the best predictors to estimate the top 5% high-cost population include individual's overall health perception, history of blood cholesterol check, history of physical/sensory/mental limitations, age, and history of colonic prevention measures. It is worthy to note that we do not consider number of visits to care providers as a predictor since it has a high correlation with the expenditure, and does not offer a new insight to the data (i.e. it is a trivial predictor). We predict high-cost patients without knowing how many times the patient was visited by doctors or hospitalized. Consequently, the results from this study can be used by policy makers, health planners, and insurers to plan and improve delivery of health services.

In memory of my father

For my mother, who needs no reason to love me; and for my wife and our daughter who gave me two reasons to love.

Acknowledgements

I am sincerely and heartily grateful to my supervisors, Professor Bijan Raahemi and Professor Craig Kuziemsky of the University of Ottawa, for the guidance, supports, and direction they showed me throughout my thesis writing. This dissertation wouldn't have been completed successfully without their continuous and invaluable helps. I am also truly indebted and thankful to the University of Ottawa for the generous supports through granting full Admission and Excellence scholarships for my master's studies. I would like to show my earnest thankfulness to the Government of Ontario for granting me the prestigious Ontario Graduate Scholarship to support my master's researches.

I wish to thanks all my colleagues and friends in the University of Ottawa's Knowledge Discovery and Data mining Laboratory, who helped me in different steps of this research.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The continued growth of health care spending and the widespread implementation of quality performance initiatives have created a growing need for tools to identify high-cost populations. In the US, health spending reached 14- 17% of the nation's GDP in 2005-2009 period, equal to $2.5 Trillion in 2009 (CMS 2011; CDC 2009).Similar numbers and trends are visible in other developed countries which shows rising healthcare costs as an outstanding problem worldwide. During the years 2008-2010, healthcare spending in Canada reached to 11-12% of GDP. Health spending equaled $192 billion in 2010, growing an estimated $9.5 billion or 5.2% since 2009. This represents an increase of $216 per Canadian, bringing total health expenditure per capita to an estimated $5,614 (CIHI 2010).

This high-cost problem becomes more complicated when we explore the highly-skewed nature of the healthcare costs i.e. only a small portion of the population cause a major portion of the costs. In the US, major chronic diseases including Diabetes, heart attacks, cancer, and stroke cause around 70% of all deaths and over 75% of all healthcare costs (CDC 2009, DeVol & Bedroussian 2007). In Canada, over 40% of direct illness costs and over 50% of all health costs including indirect costs are incurred by persons diagnosed with seven major mortal diseases including cardiovascular, cancer, chronic obstructive lung disease (COLD), arthritis, and diabetes among others (Fauci 2008, Mirolla 2004; Patra et al. 2007).

1-1 Motivation

Researchers have examined cost issues in healthcare by focusing on high-cost profiles among patients diagnosed with specific medical conditions including cardiovascular diseases (Farley et

al. 2006), diabetes (Govan et al. 2011) and bronchial asthma (Khan et al. 2006). Based on our literature survey, only few research studies examined the healthcare costs, among the general population, irrespective of their health status and disease background, in order to identify the high-cost profiles. Some researcher used statistical modeling (Fleishman & Cohen 2010) and other relied on data mining techniques (Moturu 2007, 2010). All these studies introduce trivial measures including comprehensive disease categories (Fleishman & Cohen 2010) or visits counts (Moturu 2007, 2010; Farley et al. 2006) to predict high cost instances. In this study, we introduce non-trivial predictors and employ data mining predictive modeling to forecast high-cost individuals among general population. Data mining algorithms are superior to traditional statistics in predictive model building, because they focus on individual units of analysis and predicting its final assignment to a specific class; they follow a bottom-up approach which is not concerned with hypothesis formation and testing; they are not affected by multicollinearities among numerous predictors; and they strongly handle multiple independent variables in a large dataset with exhaustive details (Zhao & Luan 2006).

Understanding future health expenditures by predicting high-cost patients among different demographic populations, health planners can better allocate available resources to different initiatives. Predictive models can greatly enhance decision making related to resource allocation (Deslavo 2009) and helps researchers and policymakers to better evaluate efficacy of different clinical intervention programs, and to measure effects of policy changes. Predictive models also help to better customize different case-management or disease-management programs by enrolling potential high-cost patients earlier (Fleishman and Cohen 2010).

1-2 Research Objectives and Questions

The research objectives for the current research study are:

1- To build *predictive models* including Decision Tree (DT) and Artificial Neural Network (ANN) to classify patients to high-cost/low-cost groups.
2- Evaluating the accuracy of the resultant predictive data mining models.
3- Discovering the *minimal set of attributes* (MSA) which can reasonably predict the high-cost population.

Through these objectives, the current research will answer the following research questions:

1- Can we build predictive models to estimate high-cost patients in general population, and compare their performances?
2- What is *the minimal set of attributes* from the Medical Expenditure Panel Survey (MEPS) database to predict high-cost/low-cost patients? How are they ranked?

These objectives and questions are designed to respond to the needs of health planners, health policy makers, and insurance providers to better model cost issues. As such, we expect to come up with a set of recommendations specifically applicable to health systems.

1-3 Research Methodology

We studied a sample of data for adult-age non-institutionalized US population taken from the *American Agency for Healthcare Research and Quality* (AHRQ)'s *Medical Expenditure Panel Survey* (MEPS) for three years of survey results in 2006-2008. This is a valid database and is available to public through the MEPS' website at http://meps.ahrq.gov/mepsweb. We employ a quantitative approach in a systematic way for data preparation, preprocessing, model building, and performance evaluation of models which is based on data mining tools. We exploit DT and NN algorithms to build *supervised* predictive models, and K-Means algorithm to build

unsupervised models. The models performances are evaluated by various metrics and the research objectives and the research question are addressed based on the modeling results.

1-4 Structure of the Thesis

This study is organized into six chapters:

Chapter 1, Introduction, the motivation behind this study, research questions & objectives, and the research methodology,

Chapter 2, Literature Review, reviews the existing literature and presents the background of our study,

Chapter 3, Research Design and Methodology, discusses the study design and methodology in detail,

Chapter 4, Data Preparation, entails data pre-processing steps in depth.

Chapter 5, Modeling and Analysis of the Results, discusses research findings in details along with their implications.

Finally, in Chapter 6, The Conclusions, summarizes the research findings and answers the research questions along with the recommendations, and suggestions for future works.

Chapter 2 : Literature Review

In this chapter, we review the existing literature and discuss different aspects of data mining applications in medical datasets. First of all, we briefly discuss the general importance of data mining tools and their relevance to healthcare field. In the second part, reviewing major machine learning tools in medical field, we discuss four major areas of data mining techniques application in healthcare i.e. *Clinical Medicine*, *Public Health*, *Medical Text Mining*, and *Policy & Planning*. In the third part, highlighting present healthcare cost and expense issues in US and Canada, we discuss the role of data mining techniques in addressing these concerns.

2-1 Data mining and Biomedicine: Importance

Data mining is "the practice of searching through large amounts of computerized data to find useful patterns or trends" (Merriam & Webster 2011). It is a part of knowledge discovery and contributes to the whole process by analyzing the raw data and capturing novel and useful patterns in datasets (Chen 2005). During the process of knowledge discovery, an iterative sequence of cleaning, integration, selection, transformation, and mining of data is done which results in possible patterns and knowledge presentations, and the data mining is considered an essential step thereof (Han et al 2012).

In general, major areas of application of data mining techniques fall either under science or business fields. The scientific uses are extensive and may include diverse domains including astronomy or healthcare. On the other hand, in business field, data mining has fueled much more media interest. *Database Marketing* and *Market Basket Analysis* focus on forecasting customers' purchasing behavior by grouping them into multiple clusters (segments) and identifying items mostly purchased together in each segment. Modern techniques e.g. *Genetic*

Algorithms, *Neural Nets*, and different *Expert Systems* have been used by investment services firms for managing very large capital portfolios (Fayyad et al 1996).

Over the past few decades, healthcare sector experienced a rapid growth. New breakthroughs in gene sequencing, protein identification, imaging modalities, and digitized patient records all add to the amount of available data in biomedical field tremendously, and data mining techniques may be applied to these huge databases in order to discover new useful patterns out of them and to extract knowledge (Chen 2005). In addition to large volume, healthcare data have some other specifications which warrant the use of modern data mining techniques. Incompleteness and lack of some parameter values, high noise and incorrectness, sparseness of data for a given subject, inexactness of parameters selected for a given task, and working with time-series require us to apply knowledge discovery techniques in order to work with and derive new knowledge out of healthcare data (Lavrac 1999).

2-2 Major Applications of Data mining in Healthcare

Applications of machine learning systems in healthcare sector fall in four categories: Clinical medicine, Public health, Text mining, and Policy and planning.

2-2-1 Data mining Applications in Clinical Medicine

Modern hospitals and clinical centers surpassed their traditional role as a place for diseases' diagnosis and treatment and now are also acting as a mass database and a unique source of complex clinical, laboratory, equipment use, and drug management data. Due to the polymorphic, integrated, redundant, and sequential nature of these data and privacy issues accompanied with its use, researchers need modern mining and knowledge discovery tools to manipulate these data (Ang et al. 2010). Among different data mining tools applied on clinical

data, predictive models including Decision Tree (DT), Artificial Neural Networks (ANN), and Support Vector Machine (SVM) are more commonly used to build clinical decision support systems (CDSS) (Bellazzi & Zupan 2008).

Some researchers deployed data mining as a disease diagnostic tool. Electrocardiographs (ECG) data has been used in a Bayesian ANN (Haraldsson 2004) or a back-propagation neural network (BPNN) (Muhammad et al. 2010) model in order to automate the diagnosis of Myocardial infarction (MI). Sabouri et al (2010) ran an ANN algorithm on a body surface potential map (BSPM) graphs database to localize the MI lesions accurately.

To facilitate decision making in cumbersome clinical situations, complex data mining algorithms may be developed to learn from past experiences of expert clinicians and decisions, and then exactly mimic their inferences in order to be used wherever such expertise is needed. Kotel'nikova et al. (2004) used patient data including medical history and physical examination, cardiac risk factors, and instrumental results as input variables and predicted the likelihood of possible future patient outcomes by running an ANN model. Yang et al. (2010) showed that SVM is superior to regression when applied on input data from a small cohort of *membranoproliferative glomerulonephritis* (MPGN) type II patients, to determine likelihood of fast progression toward end-stage renal disease (ESRD).

Many researchers have deployed different mining tools to develop models that automate the diagnosis and control of select chronic diseases which are epidemiologically important including diabetes mellitus or asthma. ECG data (P.T. et al. 2011), data from skin conductance response (Rivera Farina et al. 2009), and electroencephalograph (EEG) waves data (Nguyen et al. 2010) has been used successfully to detect autonomic neuropathy, peripheral neuropathy, and

hypoglycemic episodes in diabetics patients. Some researchers modeled the diagnosis of bronchial asthma which is the most common cause of emergency room (ER) visits and hospitalizations during childhood in the US (Fauci 2008). Rietveld and colleagues (1999) trained an ANN to use spectrograms data, a complex visual graph of patients' breath sounds, and successfully screened asthmatics patient from normal cases. A Bayesian network that uses routine patient information including vital signs, triage data, and past medical and medication history of patients (Sanders & Aronsky 2006), or a DT that uses patient daily bronchial airflows, symptoms and allergic history (Lee et al. 2011) were employed in asthmatics, and both algorithms predict the possibility of future asthma attacks accurately.

2-2-2 Data mining Applications in Public Health

Biomedical surveillance is an important application area for data mining and related tools. Kwok-Leung et al (2008) recognizes three types of surveillances in this field. *Healthcare surveillance* deals with mining and analyzing clinical data as hospital databases in order to track and monitor incidents and performance measures including medical errors and disease frequency; *public surveillance* is concerned with monitoring existing trends for chronic and infectious diseases including disease incidence and death rates and to use this new information for public health planning, implementation and evaluation purposes; and *early outbreak detection* or *syndromic surveillance* detects large outbreaks in a community before the first case of that specific is diagnosed by a clinician (Kwok-Leung et al. 2008; Buckeridge et al. 2005). The latter type is more data mining intensive and uses two streams of data including observed data of daily counts of ER visits, and daily counts of pharmaceutical prescriptions from the same ER. Different algorithms are suitable to extract patterns of disease outbreaks from these data. If the database is single and no spatial information is included, as it is the case with most public

health data, statistical process control, with or without pre-processing stage (e.g. regression, Kalman Filter, or wavelet analysis) is a suitable algorithm. When covariates are added, use of a Bayesian Network becomes necessary (Buckeridge et al. 2005).

An important application of data mining in public health field is *population screening*. Grassi et al (2001) used demographic, clinical interview, respiratory function tests, and allergy profile as the input data to develop a predictive model for screening patients more probable of having asthma. Hirsch et al. (2001; 2004) gathered the input data by means of a patient-completed questionnaire and selected a threshold of positive answer counts to include a subject for more analysis. They labeled the selected pool by clinical experts, and trained an ANN classifier with this labeled data, and ran the classifier on the same pool. Their model ranks subjects according to their likelihood of having an asthma diagnosis.

2-2-3 Data mining Applications in Healthcare Text mining

Data mining tools have been applied extensively for text mining purposes in healthcare. Two major areas of text mining applications are either its use on medical and healthcare literature, or on clinical data e.g. patients' clinical records. When applied on research literature itself, text miner finds any hidden relationship between existing or emerging biomedical entities which are not visible to us otherwise; as each researcher is familiar with the entities belonging to a very small sub-domain. When applied on a large clinical data, text miner tracks the presence of a specific disease entity in a target population, or realizes the hidden patterns of a specific symptom e.g. fever, to detect potential outbreaks (Chen 2005).

Yildirim et al. (2011) targeted to find disease-drug relationship for osteoporosis, which is a costly disease among seniors in developed countries. They used a search engine tool leveraged

by text mining algorithms to find which drug has the highest number of articles published on MEDLINE and seems dominant in use by clinicians. According to their results, *Alendronate* and *Raloxifene* are the two top ranked drugs for the treatment of osteoporosis. Other researchers used shallow information extraction method to grab drug names, dose, mode and frequency of usage from free text for patient records in Bulgarian language (Boytcheva 2011). They used 1300 and 6200 patient records as training and test datasets, respectively, and showed a high accuracy for extracting drug name and dose.

2-2-4 Data mining Applications in Healthcare Policy and Planning

Data mining tools have been used to help policy makers and planners in decision making and program evaluation. They showed usefulness for underpinning the hidden and complex relationships between important disease categories in the outbreaks, and a set of complex risk factors. Ghosh and Guha (2011) employed a computational neural network (CNN) in order to analyze potential relationships between different risk factors (environmental, socioeconomic, built-environment, and existing mosquito abatement policies) and the incidence of West Nile Virus (WNV) infection in birds.

Data mining techniques are useful in improving disease management programs, helping the disease management organizations to better implement and evaluate disease-specific programs. Mougiakakou et al. (2010) introduced SMARTDIAB which uses different intelligent algorithms, information processing, and communicating tools on a web-based platform to provide clinicians with the best clinical choice for diabetic patients. Selecky (2008) used innovative IT tools and data mining techniques to provide best practices, guidelines and advices for patients and professionals dealing with chronic obstructive lung disease (COLD) which is the fourth leading cause of death in US (Selecky 2008). Bereznicki et al. (2008) analyzed a community-wide

pharmacy medication records with customized data miner software, to accurately find the asthmatic patients with suboptimal disease management.

Data mining has been used to detect expensive clinical profiles among patients diagnosed with a specific chronic illness which has a high disease's burden e.g. diabetes. Concaro et al. (2009) applied a temporal association rule (TAR) model on diabetic patients' data including body mass index, systolic blood pressure, hemoglobin A1C (HbA1C), and serum lipid profile of patients, and successfully profiled three high-cost DM patients groups. Weinstein et al. (2009) applied a Bayesian belief network on demographic, pharmacy usage, utilization history, co-morbidities, and chronic medical conditions data, from a pool of enrollees with substance use disorder (SUD). Their model successfully stratifies the SUD enrollees according to their next year's health cost using current year's expenditure data. This cost stratification helps health planners to allocate enrollees to the appropriate substance abuse management program available.

Data mining helps health planners to solve resource allocation problems and capacity issues. Teow et al. (2011) used logistic regression, DT, and ANN to identify patterns behind bed overflow in a hospital in Singapore. Isken & Rajagopalan (2002) used simulation to capture the flow of patients in a tertiary care hospital using a K-means clustering tool to create a logical set of patient types with similar profiles to circulate in the proposed simulation model.

2-3 Cost Issues in Healthcare

The *Total Health Expenditures (THE)* in the United States has experienced a huge growth in recent two decades. Total health spending was roughly over a trillion dollar in 1997; doubled by 2005; and then approached to 2500B in 2009. This figure is more than twice of the average for the remaining OECD countries' health spending. In 2009, health spending comprised more than

17 percent of the US' GDP, while its share in gross domestic product in 1997 and 2005 was 13.7% and 16%, respectively. The picture becomes clearer if we note that the nation's GDP increases from 8.3 trillion in 1997 to more than 14 trillion dollars in 2009; and the population experiences only 12 percent growth in the same period which is traceable by doubling in per capita health expenditure when we move from 4166 dollars in 1997 to 8086 dollars in 2009 (CMS 2011; CDC 2009).

Numbers are smaller when we review the health expenditures in Canada, but the trends are the same. Canada finances its healthcare through a mix of private and public sources. Public sector including federal, provincial/territorial, municipal governments and governmental agencies paid for 70-76 percent of total health expenditure in 1975-2010. The share of the private sector including out of pocket household payments, private insurance, and non-consumption categories, ranged between 24-30 percents for the same period. In this period, the share of public sector diminishes gradually, in favor of private sector payments, but the absolute numbers for each of public and private sectors, and therefore, *THE* experiences continuous growth. Total health expenditure was 40 billion dollars (40B) in 1975 and reached to 137B in 2010 (constant 1997 dollars). In current dollars, *THE* in 2008 through 2010 equals 172B, 182B, and 192B, respectively; which equals *THE* per capita of 5154, 5397, and 5613 Canadian dollars; and *THE* as a % of GDP of 10.7, 11.9, and 11.7, respectively (CIHI 2010).

Useful information could be derived by analyzing the places of use for health expenditures. In the US, the top user of health dollars is hospital sector with 31% share, but spending for physicians and clinics falls in second place with 20% share, and is followed by drug expenditure which represents 10% of costs. With respect to the sources, no nation-wide health coverage program exists. Private sector including private insurers, businesses, household payments, and

other private sponsors 56% of health dollars; and the government, including different public insurance programs including Medicare, Medicaid, CHIP (Children Health Insurance Program), DOD (Department of Defense), and VA (Veterans Affairs) contribute by 44%, collectively (CMS 2011).

Similar to the US, in Canada, three major destinations of health dollars are hospitals, drugs, and physicians. Canadians spent over 55B in 2010 for hospital expenses. The hospitals have traditionally been a major place for healthcare provisions and in mid-1970s they account for 45 percent of national health expenditures, but their share has fallen gradually in next 30 years until new millennium. Since 2001, hospitals' share remains constant around 29% of the Canadian THE. Drugs and physician services occupy 2nd and 3rd place in Canada's health spending with a 16% and 13% share in 2008, respectively. All other health professionals and health institutions collectively represent 22 % of *THE* in the same year (CIHI 2010).

2-3-1 Cost of Chronic Conditions

If we drill down into details of spending by disease categories, in Canada, seven major chronic illnesses including cardiovascular diseases (hypertension, stroke, and heart attacks), cancers, COLD, DM & other major endocrinopathies, musculoskeletal disorders (Osteoarthritis and osteoporosis), central nervous system, and mental illnesses account for 42% of all the direct spending on hospitals, drugs, and physicians services. Four of these conditions, i.e. cardiovascular diseases, cancers, COLD and DM cause 70 percent of all mortalities in Canada annually. If we add indirect costs (i.e. the decreased life quality of the affected persons, and economic consequences for their families, communities and societies in general) of these seven chronic diseases, their total burden will approach to 50% of all costs incurred by all illnesses in the country. In 2002, the absolute dollars amount spent for top seven chronic illnesses was 93B,

compared to the *THE* of the same year (174B) (Mirolla 2004; Patra et al. 2007). The same figure is evident for the burden of the chronic diseases worldwide. As in Canada, cardiovascular diseases are the leading cause of death around the world causing 30 percent of all global mortalities. The next prevalent mortal illnesses are cancers, COLD, and DM, causing 13%, 7%, and 2% of all deaths around the globe, respectively (Patra et al. 2007).

Chronic medical conditions have had considerable impacts on lives of the US citizens. Cardiovascular diseases (hypertension, heart attacks and stroke), cancers, diabetes, chronic mental disorders, chronic obstructive lung disease (COLD), arthritis, and obesity are the leading cause of mortality in this country (by causing 7 out of 10 deaths annually); and people with chronic conditions are a destination for 75% of all health dollars annually. In 2005, near half of the adult American population (more than 133 million persons) suffered from at least one chronic medical condition which are predisposed by lack of physical activity, poor nutrition, tobacco use, and excessive alcohol intake, as four major modifiable risk factors predisposing to chronic diseases (CDC 2009). In another conservative picture derived from the MEPS 2003 database, which only reports on the non-institutionalized US population's health expenditures, cost of therapy (not including other disease costs including follow-on costs for the control of a disease) for seven major illnesses (cancer, heart disease, hypertension, mental illnesses, DM, pulmonary conditions, and stroke) reached 227B in 2003; and the number of chronic conditions was over 162 millions nation-wide which equaled to 109 million afflicted Americans as some persons might have more than one condition at the same time (DeVol & Bedroussian 2007).

2-4 Medical Expenditure Panel Survey (MEPS)

American Agency for Healthcare Research and Quality (AHRQ) conducts the MEPS survey to respond to the increasing needs of policymakers and researchers for an accurate and nationally

representative expenditure data, which enable them to better explore healthcare cost issues (Cohen et al. 2009). AHRQ initiated MEPS in 1996 which collects data on the US civilian non-institutionalized (household) population's characteristics and their uses of health services including in terms of frequency, costs and expenditures, and method of payments detailed by type of encounters. It also provides detailed data about private health insurance held by household population in terms of cost, scope, and breadth of services. The MEPS data is unique with respect to the detailed information it provides, and its ability to link health services usage, health expenditures, and insurance data to all other demographic, employment, economic, health status, and other survey attributes (Cohen 2002).

The MEPS has three interrelated survey components: Household component (HC), Medical provider component (MPC), and Insurance component (IC). The HC surveys household member for demographics, health status & health conditions, all medical events by type (hospital including inpatient, outpatient, and ER; office-based; dental; home health; prescribed medications; and other including glasses, ambulance, equipment, etc.), date and other details of an event, charges & payments by source for each event, employment profile, health insurance profile, income, access to care, level of satisfaction, and opinions. The MPC reports on the activities of the providers of care to the household sampled persons including dates of visits/admissions, diagnoses based on the ICD-9 codes, services provided based on the CPT-4 codes, full charges, detailed payments by source, and reasons for any difference between charges and payments. The IC surveys employers either private or public, for their business establishment's full profile, offered plan and its profile, premiums, contributions of the employees, and resultant total costs incurred to the organization (Cohen et al. 2009).

The MEPS uses a subsample of the households participating in the National health interview survey (NHIS) each year, which has been conducted by CDC since 1957. The usual sample size for the MEPS is to 12-14 thousand families/30-32 thousand individual. Every year, one panel of the new MEPS sample is introduced into the survey and five rounds of computer-aided interviews are conducted for each panel during next 24-30 months in order to gather data for two calendar years for each panel; therefore, the HC part of the MEPS has an overlapping structure, in which two consecutive panels are participating each year (Cohen 2002; Cohen et al. 2009).

2-4-1 The MEPS' HC Component

The final HC data files are released annually; are free for the public use; and are accessible through the MEPS website at its internet address at http://www.meps.ahrq.gov. Each year different types of the HC data files are released in ASCII and SAS file formats. Main data files include the HC full year files (including Full year consolidated data, Full year population characteristics, Medical conditions, jobs, Person round plan, longitudinal weight, Supplemental variables, Health insurance plan abstraction, and long term care files), and the HC event files (Prescribed medicines, Dental visits, Hospital inpatient stays, ER visits, Outpatient visits, Office-based medical provider visits, Home health, and other medical expenses files) among others. Two data linkage files are published annually in order to make possible to link the HC files to either the NHIS survey data or the IC files (MEPS Website). By providing comprehensive data on how the population uses health services, how much does health care cost, and who pays for this spending, the HC is considered an unparalleled research data for exploring the macro issues including predicting health care costs, evaluating health reform policies, estimating the public health impact of nation-wide programs as Medicare and Medicaid, examining the effects of big

decisions as tax code changes on tax revenue and health expenditures, and developing the US economic indicators and projections (Cohen 2002; Cohen et al. 2009).

The MEPS HC data has been used in different ways to explore health cost issues and medical expenditure problems in the US. It was used to document either the highly concentrated nature of health spending, persistence of high costs over time, drill down into the individual characteristics, the behavioral attributes, the financial incentives and other factors or institutional arrangements, and to detect individuals with high-cost profiles (Cohen et al. 2009).

MEPS' HC component has some limitations including the absence of the expenditure data for institutionalized population, and its self-report nature. There also a risk of misreporting for medical conditions, because it uses a non-technical reporting by the family members and then converts their verbatim into the ICD-9 codes. It converts more than 17000 five-digit codes to a set of 285 three-digit codes to prevent privacy issues, and reduces the amount of facts transferred (Cohen et al. 2009).

2-4-2 The MEPS' MC Component

The MEPS data on Medical Conditions (MC) as a part of HC survey, has received special attention from different researchers according to its unique specifications. This survey is conducted by questioning a member of each family who responds to questionnaires and interviews on behalf of herself and all other household members. The questions are open-ended and designed to better find medical conditions, not disease symptoms, by recording and analyzing person's verbatim. An expert staff analyses the text results of interview information and assigns a suitable ICD-9 code to each identified condition (Fleishman & Cohen 2010; Machlin et al. 2009).

Based on prevalence, expense, and relevance to policy, MC part of HC reports some medical conditions as the Priority Conditions which includes hypertension, coronary heart disease, angina pectoris, heart attack & MI, other heart diseases, stroke & transient ischemic attack, emphysema, chronic bronchitis, high blood cholesterol, cancers & malignancies, DM, joint pains and arthritis, asthma, and, attention deficit hyperactivity disorder. For panels before panel 12, the Priority Conditions include HIV & AIDS, gallbladder diseases, stomach ulcers, back problems, and mental disorders including depression, anxiety, and Alzheimer's disease (The MEPS website).

2-5 Cost Prediction Models in Healthcare

Cost Prediction has been an interesting subject for different stakeholders in healthcare field in recent decade. Some researchers focus on the cost problems among patients with a specific disease including asthma or diabetes, while other explore health costs in general population. In both cases, researchers must deal with big challenges for cost modeling and prediction including proper selection of relevant input attributes/independent variables from a large database, creation of a balanced train/test dataset, proper model learning, and selection of an efficient performance evaluation method (Moturu et al. 2007).

Two big classes of attributes have been used by different researchers as predictors of high cost patients: Demographic and Clinical. Among demographics, age and sex are most commonly used as the basic predictors, along with other additional demographic attributes including race, region, marital status, etc. Clinical attributes have been used as the second-line predictors in order to improve models' predictive performance. Expert knowledge of a research team influences their ability to better select attributes from a bigger pool of available clinical information.

2-5-1 Cost Prediction: Disease-Specific

Most of existing literature on health cost prediction is devoted to profiling high-cost instances among individuals who have a specific disease diagnosis including diabetes, asthma, or cardiovascular diseases. Govan et al. (2011) use logistic regression and general linear models to predict the probability of hospital admission, and inpatient costs among a group of patients diagnosed with DM type 1 and 2 in Scotland in 2005-2007. Three different datasets in Scotland were used including Scottish Care Information- Diabetes Collaboration (SCI-DC), Scottish Morbidity Records (SMR), and General Register Office for Scotland (GROS) to build the final database. They use demographic predictors including age and sex; and clinical information predictors including BMI, serum creatinine, HBA1C, date and length of hospital admission, and date of discharge in order to build a model and predict the dependent variables, i.e. possibility of hospital admission for any specific year, and total cost of admission for that year. An increase in age, a higher serum creatinine, a female sex, a presence of previous history of vascular events, and a lower BMI is associated with a higher admission rate and higher costs for DM patients. A higher HBA1C concentration which is indicative of poorer glycemic control among diabetic patients (Fauci 2008) is associated with a lower admission rate and a lower cost among DM type 2 patients.

Logistic regression works well on large samples (Govan et al. 2011). Better correlative results and ranking of predictors is obtainable by using the data mining tools including DT or ANN on these huge dataset. These latter two tools better manipulate variables with more than two discrete values and also handle interactions between variables which are quiet common in clinical settings, and may lead to an un-interpretable result if the applied tool is not strong enough to deal

with complex inter-variables effects. They can handle more disease related variables and minimize effect of confounders.

Khan et al.'s work (2006) is a sample of data mining use in predicting cost issues for a specific disease. They used a pool of asthma patients, from the Louisiana Medicaid's retrospective claims dataset in order to predict high cost asthma recipients. In their model input variables include demographic predictors (age, sex, race, and location of residence); clinical predictors (including number of hospitalizations, emergency visits, and physician office visits; medication usage including quick relief, long-term, and total medication use; continuity of care; enrolment in community care; and physician specialty) and prior health costs in year 1. In a multiple logistic regression model, these variables predict high cost asthmatics in year 2. Threshold cost for labeling a patient as high cost in Khan's work is 95^{th} percentile of costs in one year which is equal or greater than $2,985.42 for total costs in year 2. Their model predicts 29.66% of the high cost asthmatics accurately. Khan et al (2006) attributes their model's overall low accuracy partly to the inherent weakness of claim databases especially when they are not maintained stable and suffered from some inconsistencies like policy changes over time. Another justification for their model's low accuracy is their failure to include a full range of co-morbid conditions which is vital when a researcher is working on cost issues of a specific disease. Although, they incorporate some important co-morbidity indicators including prior visits, hospitalization, and medication usage, they fail to include more important co-morbidities e.g. the presence and counts of number of chronic medical conditions in elderly patients, duration of hospital stays for any given patient, and sub-type of asthma (either extrinsic or intrinsic asthma).

Farley et al. (2006) test different co-morbidity indexes to predict future medical expenditures among 20378 hypertensive patients of a managed care organization in the US. They conclude

that a simple count of physician's visits is the best cost predictor especially when we are aiming to predict the highest cost percentiles (90^{th}), the next important predictors in their work are a count of diagnosis clusters, and a count of hospital claims. Prescription claims are not an accurate predictor of very high cost hypertensive.

2-5-2 Cost Prediction: General Population

In contrast to disease-specific cost modeling, few researches have been conducted to estimate high-cost individuals among general population. This scarcity is mostly attributable to the scarcity of a nation-wide or jurisdictional health survey itself, or lacking access to few available surveys by most researchers. The MEPS is an exception among others, and provides researchers with a free access to a nationally representative multi-dimensional survey data from the US.

Fleishman and Cohen (2010) ran a series of logistic regression models on the MEPS HC datasets in order to predict top 10 percentile of health costs in year 2, using predictors from year 1. They use simple demographics including age, gender, and type of insurance for any individual to build a *baseline* model, and evaluate the predictive performance of the model by several goodness-of-fit indices. This baseline model returns a correlation index equal to 0.310 to high cost label in year 2. They improve the baseline model's performance by incorporating some clinical predictors (including Key condition indicators, Perceived health and functioning, Count of chronic medical conditions, and Diagnostic cost group (DCG) category). By adding these clinical predictors (in above-mentioned order) to the baseline model, a better correlation index returns (0.380, 0.385, 0.414, and 0.419, respectively). The DCG categories yields the highest improvement (correlation index of 0.419), followed by a simple count of chronic medical conditions (correlation index of 0.414). The latter is very easier to calculate, and is directly

accessible through the MEPS MC file, compared to costly, time consuming, and complex nature of calculating DCG categories (Fleishman & Cohen 2010).

Moturu et al. (2007; 2010) uses the Arizona's Medicaid claim database in order to extract a broader range of demographic predictors (including age, gender, region, race, and marital status), combined with two clinical predictors: overall counts of medical visits for each person, and a binary attribute showing the presence or the absence of visits in different medical categories. They further detail these latter predictors by inpatient, emergency, outpatient (each summarized in 20 medical categories based on ICD classification), and pharmacy visits (summarized in 136 medical categories based on National Drug Code-NDC). They extract these predictors from year 1 data (2002, and 2003 for train and test sets, respectively), and the cost classes data from year 2 data (2003, and 2004 for train and test sets, respectively), in order to train and test five machine learning algorithms including LogitBoost, AdaBoost, logistic regression, logistic model trees, and SVM. They evaluate their models performance using G-Mean, along with models' sensitivity and specificity, and conclude that adding the pharmacy data and inpatient visits will add to overall models' predictive performance and all five data mining classifiers did well with marginally better results for the LogitBoost algorithm.

Chapter 3 : Research Design and Methodology

The current research follows a quantitative approach in its research design to build a set of predictive models using data mining tools by applying a set of classifiers and clustering algorithms on the MEPS database from the US.

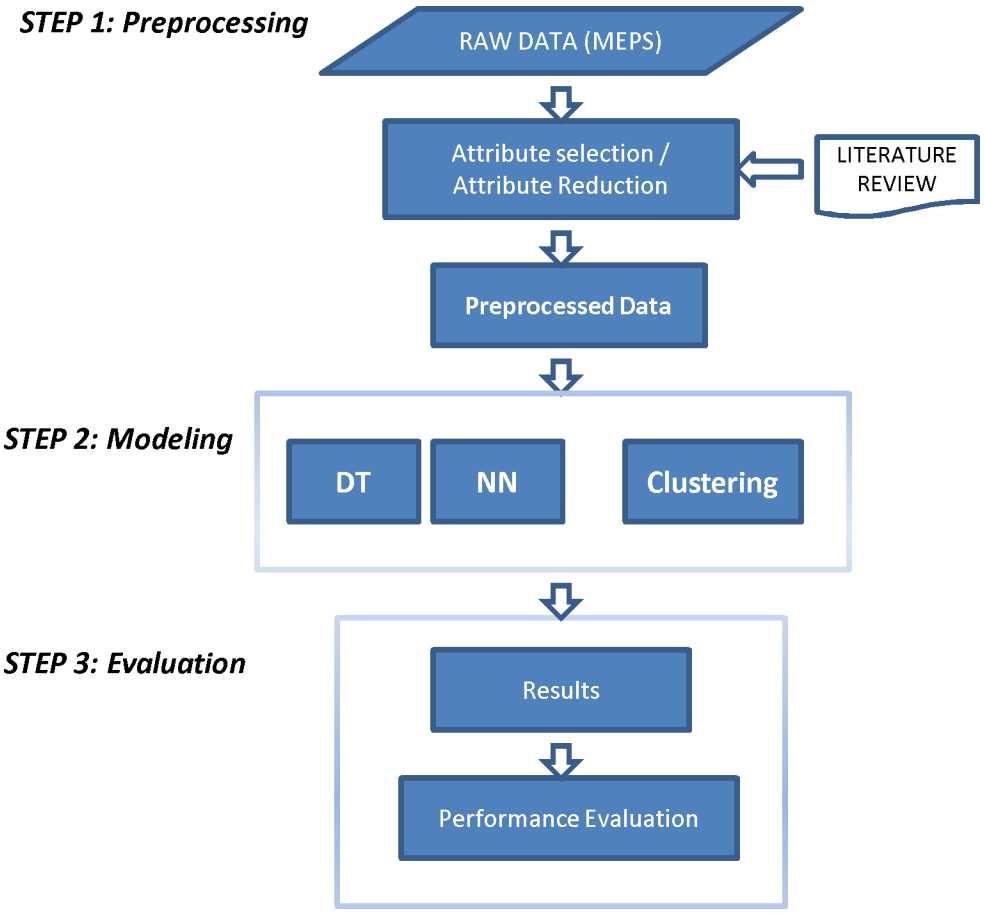

Figure 3-1: Research Design

A schematic representation of our research design is shown in Figure 3-1. The research design consists of 3 consecutive steps. The first step is *Preprocessing* that includes Raw Data extraction, Attribute Selection, and Preparation of different versions of final dataset with select set of pertinent attributes that are used in DT, and NN classifiers and K-Means clustering models. The

second step is *Modeling* in which we build, train, and run multiple models on test sets in order to get the results. The third step, *Evaluation*, deals with analyzing models' performance using relevant measures including Correctness Accuracy, Sensitivity, Specificity, G-Mean, Area under ROC Curve (AUC), and Silhouette measure of separation and cohesion. Using this step's results we may choose to return to step 2 to build more models in order to get the final results and answer our research questions and reach to the study's objectives.

3-1- Preprocessing

In this phase, we constructed the final dataset along with the different versions of it that are suitable for importing into the relevant models.

3-1-1 Literature-Driven Attribute Selection

We use online databases including MEDLINE, PubMed, Elsevier, SCOPUS, Web of Science, IEEE and Xplore through University of Ottawa's Library, and the MEPS website in order to find relevant research on health expenditure modeling. Any of or different combinations of the following search-words have been used in order to explore the existing literature:

a. Medical Cost*/Medical Expenditure*/Health* Cost*/Health* Expenditure*
b. Disease* Cost*/Disease* Expenditure*
c. Cost Prediction/ Cost Model*/Predictive/Cost Modeling/ Data Min*/Medical Data Min*/Health* Data Min*
d. Decision Tree*/Neural Net*/Neural Network*/ Cluster*/K-Means/ CHAID/CHAID Tree*
e. MEPS/Medical Expenditure Panel Survey, Expenditure Panel Survey

Based on the literature survey and medical expertise and insights taken from reviewing the MEPS documentation files for 2006-2008, we identify a primary set of input attributes, which are categorized under five modules:

a. Demographic attributes (or simply Demographics module)
b. Health Status Attributes (or simply Health module)
c. Preventive Care Attributes (or simply Preventive module)
d. Priority Conditions Attributes (or simply PrioC module)
e. Visits Counts Attributes (or simply Visits module)

We use ASCII format of both the MEPS FC (Full-Year Consolidated) data files for years 2006-2009 and the MEPS MC (Medical Conditions) data files for years 2006-2008 and employ a programming language to derive input/target attributes in a C*omma-separated* format and then transform it into the MS Excel file format for further processing. We use the MS Access and the MS Excel tools to further preprocess, and to integrate different data files from the MEPS survey.

3-1-2 Attribute Selection

One major objective of this study is to introduce the MSA that can be used effectively in order to predict high-cost cases with reasonable accuracy. We narrowed down the primary set of 66 attributes gradually in order to obtain the least possible counts of attributes while achieving a desirable accuracy. For this purpose, attribute reduction efforts start at the preprocessing stage and will continue throughout the modeling stage.

As part of preprocessing we removed attributes having more than 50% of missing values in an attribute's values. Inside each module of attributes, e.g. Demographics or Health, we use the results of *bivariate correlation* in order to reduce more attributes. Bivariate correlation is based

on different metrics including Pearson rho, Spearman's rho, or Kendall's T correlation coefficients and identifies possible association between two attributes, considering the effect of X attribute's variations (in terms of *values* or *rank of values*) compared to the Y attribute's changes. The most commonly used metric is Pearson rho (represented by r) which is the most suitable for normal distributions:

$$r = \frac{COV\ (X, Y)}{\sigma X \sigma Y}$$

Where, COV (X, Y) is the covariance between two variables X and Y; and σ stands for the standard deviation. We test the significance of these correlations in a either two-tailed or one-tailed direction as required.

As a part of our efforts to reduce attributes, we use the *Feature Selection* node (in the Modeling Palette of the IBM SPSS Modeler) in late modeling stage, in order to get to an overall understanding about the most relevant attributes to our targets.

3-1-3 Preprocessed Data

The final product of the *preprocessing* step is a dataset with select set of input attributes that are ready for model building. To prepare this dataset we need a few more operations to be executed on our raw data, including:

1- Removing Zero-weight records that are not representing any part of the US non-institutionalized population in MEPS survey

2- Defining all attributes' data properties, roles, and measurement levels as required by each model provisions and assumptions

3- Handling the *missing values* for all attributes, using the IBM SPSS Modeler itself and other software

4- Defining threshold for the high-cost cases, using insights from the literature survey, expert knowledge, and analysis of expenditures distribution pattern

5- Constructing the *target attribute* based on the agreed-upon high-cost threshold(s). As we predict both current year's and next year's expenditures, we construct as many target fields as needed to satisfy our results.

3-2 Modeling

We use the *IBM SPSS Modeler* software version 14.1 to build DT, NN, and Clustering models.

3-2-1 IBM SPSS Modeler

The IBM SPSS Modeler is a set of data mining tools which enable the users to build different predictive models taken from machine learning, artificial intelligence, and statistics. It has a user interface which makes it possible to quickly set up, customize, and run different models by building a stream of data. In technical terms, this tool provides a wide array of modern DM algorithms comparable to other software including WEKA and SAS Enterprise Miner.

A typical data mining task in the IBM SPSS Modeler starts with a *data stream* which is composed of a series of *nodes* (IBM SPSS Modeler's help documentation). In this study, we build multiple data streams by consecutive use of these palettes in the Modeler: *Source* palette, *Field operations* palette, *Record operations* palette, *Modeling palette*, and *Output* palette (Figure 3-2).

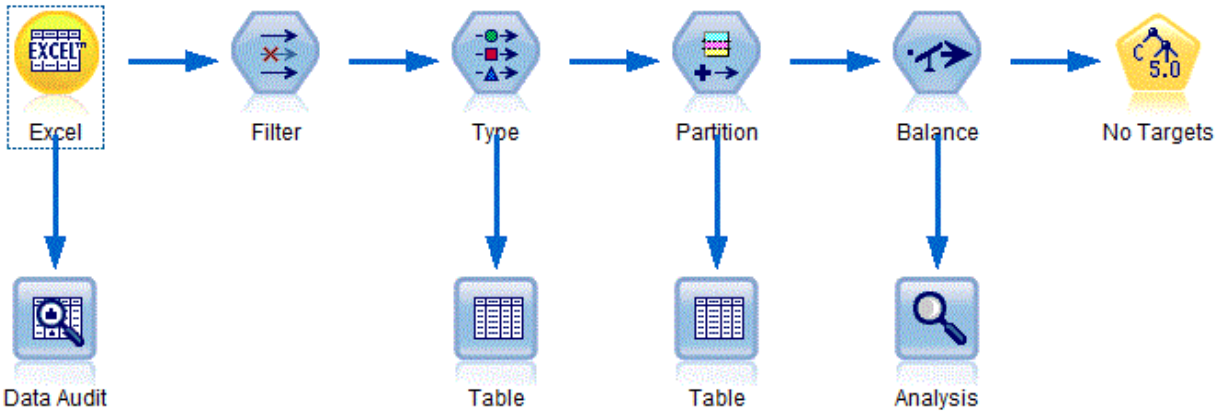

Figure 3-2: A typical data mining task in the IBM SPSS Modeler

After importing a data file in *source palette*, we use *field operations palette* to filter, clean, and transform the data fields for future analysis and model building. The Modeler rewords columns or variables as *fields*. We specially use the *partition node* to partition the final MEPS dataset into a *train* and a *test* set which comprises 70 and 30 percents of records, respectively (IBM SPSS Modeler's help documentation).

The record operations palette is used to transform data at the record level. The Modeler rewords instances or cases with records. Through relevant nodes, we select, sample, balance, sort, merge, aggregate, or append our data as needed. A *balance* node is specially used for oversampling in order to make a balanced dataset across all categories of outcomes (IBM SPSS Modeler's help documentation). We use it on the partitioned MEPS dataset data to balance the train subset in a way that the number of records in each categories of outcome i.e. high-cost or low-cost remained equal. A balanced train set helps a classifier to learn from equal population of outcomes.

The balanced partitioned datasets are imported to relevant classifier and clustering nodes through using the modeling palettes, which includes various DTs, ANNs, and K-means clustering in our study. In the final stage of any data mining task we extract the results the results in an output

node. For supervised models the results are shown by adding an *analysis* node to the data stream in order to get accuracy and other performance measures. For unsupervised models including K-Means, there is no actual result to be used as a basis for this comparison and other measures including the silhouette measure of clustering cohesion and separation are applied.

3-2-2 Decision Tree (DT)

In the very beginning of the modeling step, we run an *Auto-classifier* on the final dataset. The Auto Classifier node estimates and compares models for either nominal (set) or binary (yes/no) targets, using a number of different methods. Our target fields have also flag binary outcomes (high-cost or low-cost with a value label of 1 or 2, respectively). Based on the results of this classifier's modeling results, we build more DT models using two separate algorithms: C5.0, and CHAID.

A C5.0 model works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned. To train a C5.0 model, there must be one categorical (i.e., nominal or ordinal) target field, and one or more input fields of any type. The results would be shown as either a *tree* or a r*ule set* which are derived from the tree algorithm itself. We anticipate that the C5.0 will work on our data very well, because it handles the missing values very well, it works with large number of input fields, and finally, it yields best results when most input fields are categorical (IBM SPSS Modeler's help documentation).

We use another tree, the CHAID algorithm, in order to compare the performance of C5.0 with another strong tree builder. The CHAID (*Chi-squared Automatic Interaction Detection*) uses chi-square statistics to identify optimal splits. It can accept any type of input/target attributes, and can generate non-binary trees, meaning that some splits have more than two branches. It therefore tends to create a wider tree than the binary growing methods (IBM SPSS Modeler's help documentation).

We do not set misclassification costs in our modeling efforts as predicting a high-cost instance as a low-cost instance is not important to us at this stage. By setting a higher cost for any of misclassification, the correctness accuracy of a tree builder falls slightly.

In modeling step, we build trees in three stages. In the first stage, we build trees and NN by using all attributes in the final dataset to compare the performances of trees with those of NNs. In the second stage, we compare the performances of different trees. We build multiple CHAID and C5.0 models by using any of five modules of attributes (Demographics, Health, Preventive, PrioC, and Visits) separately. In the final stage of the modeling phase, we build more CHAID trees for all combination of the Demographics with any of other four modules / or other possible combinations, until we reach to the MSA.

3-2-3 Neural Network (NN)

In the first stage of the modeling phase, when we use all attributes in the final dataset, we compare the performance of tree builders with those of NNs by building multiple NN models. Neural Net's concept comes from the way that neurons work in human nervous system, i.e. a NN is a simplified model of the way the human brain processes information. It works by

simulating a large number of interconnected processing units that resemble abstract versions of neurons. The processing units in a NN model consist of input (representing the input fields), hidden, and output (a unit or units representing the target field) layers of neuronal units. The units are connected with varying connection strengths (or weights). Input data are presented to the first layer, and values are propagated from each neuron to every neuron in the next layer. Eventually, a result is delivered from the output layer (IBM SPSS Modeler's help documentation).

We manipulate the number of units in the hidden layer(s) based on trying different combinations, and using the IBM SPSS Modeler's automatic suggestion, in order to reach the best model's correctness accuracy for the test set data.

3-2-4 K-Means Clustering

Clustering models identify groups of similar records and label the records according to the group to which they belong. They do it without knowing about the groups and their characteristics from start. We use these models when we still have no idea about the best count of groups that could be found in our data. Therefore, they use an unsupervised learning because they use no target field at all; and their value is determined by their ability to capture interesting groups in the data and provide useful descriptions of those groups (IBM SPSS Modeler's help documentation).

When we reduce our attribute list to 10 (from starting pool of 66 attributes), we run a K-means clustering model in order to get a better understanding of hidden patterns in our data. We

compare these understandings with those of classifier models to see how an unsupervised learning algorithm as K-means differs from a supervised one as a DT.

3-3 Evaluation of Models' Performances

Evaluation is the most important step in any data mining task, because it gives us a clear view regarding the strength of each model individually, and enables us to compare different models based on common performance measures. We analyze the models' performance (either unsupervised clustering e.g. K-means or supervised classification e.g. DT or NN) based on a series of measures that are recommended most in the literature. Based on performance evaluation results, we build more models in order to reach to the MSA.

We use six performance measures in order to evaluate our models performance. Five performance measures are used for evaluation of classifiers (DT and NN models) and include *sensitivity*, *specificity*, *correctness accuracy*, *G-mean*, and *area under the ROC (Receiver Operations Characteristics) curve*. One performance measure is used for evaluation of the unsupervised learning models (K-means Clustering) and is the s*ilhouette measure of separation and cohesion*.

3-3-1 Sensitivity, Specificity, and Correctness Accuracy

In evaluating a classifier's sensitivity, specificity, and correctness accuracy we use the confusion matrix which plots predicted vs. actual classifications. (Figure3-3)

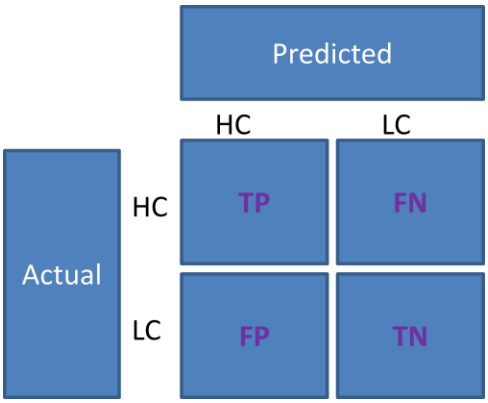

Figure 3-3: Confusion Matrix for classifying high-cost (HC)/low-cost (LC) outcomes. (TP: True Positive, FP: false Positive, TN: True Negative, FN: False Negative)

The sensitivity of a classifier is defined as its ability to correctly identify actual cases, i.e. for our study, it measures the proportion of high-cost instances which are correctly identified as such. The specificity of a classifier is defined as its ability to correctly identify negative cases, i.e. for our study, it measures the proportion of low-cost instances which are correctly identified as such. They are summarized in the following formulas:

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

Accordingly, the correctness accuracy for a data mining classifier is defined as the degree of closeness of its prediction to the actual values, either true or false, i.e. for our study, it measures the true results (both true positives or high-costs, and true negatives or low-costs) among all the test population:

$$Correctness\ accuracy = \frac{TP+TN}{TP+FN+TN+FP}$$

This measure is reported as an absolute value between 0-100 percent and is used to different models accuracies efficiently. Kubat et al. (1997) claimed that this measure is not adequate when the absolute count of actual negative cases is much larger than actual positives. This is the case with our study where for example, when we define the high-costs as top 5% of the test population, the proportion of high-costs vs. low-costs is 1:19. It will bias the correctness accuracy toward specificity, not sensitivity. To address this shortfall we use two other performance measures which give better trade-off between sensitivity and specificity. These two measures are G-mean and Area under the ROC curve.

3-3-2 Geometric Mean (G-mean)

First introduced by Kubat M., Holte R., and Matwin S. of University of Ottawa in 1997 (Kubat et al. 1997), is a geometric mean of sensitivity and specificity and is only the highest when both of these measures are high.

$$G - mean = \sqrt{(acc+) * (acc-)} = \sqrt{TPr * TNr} = \sqrt{Sensitivity * Specificity}$$

Where TPr equals to percentage of positive examples correctly recognized, and TNr equals percentage of negative examples correctly recognized. These measures are defined as acc+ and acc- in the original article (Kubat, Holte, and Matwin, 1997) and are synonym to sensitivity and specificity measures, respectively.

3-3-3 Area under ROC Curve (AUC)

The area under the ROC curve is a convenient way to compare different classification models which takes into account the trade-off between sensitivity and specificity of a model and gives a

clear picture of the model's accuracy. This measurement is equivalent to the *Gini index* and the *Mann-Whitney-Wilcoxon test* statistic for comparing two distributions and is referred in the literature in many ways, including c-statistic, "Θ", and "AUC" (Area under the ROC Curve) as shown in figure 3-4 (Hanley 1982).

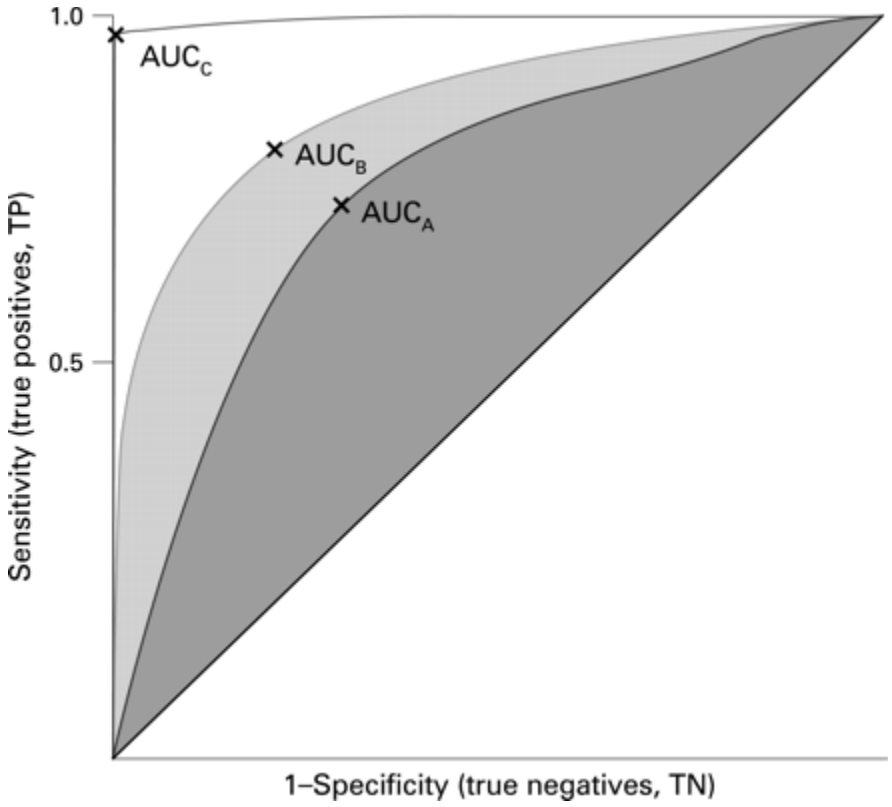

Figure 3-4: Area under the ROC Curve (Soreide 2009)

As we move from AUC_A to ward AUC_C the accuracy increases and the AUC measure can be considered as an averaging of the misclassification rates over all possible choices of the various classification thresholds. In other words, AUC measure is an average of the diagnostic performance of a particular model over all possible values for the relative misclassification severities (Hanley 1982). The interpretation of AUC measure, where a "high-cost" instance is

scored as a 1 and a "low-cost" patient is scored as a 2 is the answer to the question – "using this model, what is the probability that a true 1will be scored higher than the threshold expenditure for true 2?

In our study we have not weighed the misclassification costs very highly, but we still report on the AUC measure which is sensitive to this provision. In fact we need to have a good sense about models' performances from a perspective other than the correctness accuracy measure (which fails when our test set abounds in negative values).

We calculate this measure for all DT (C5.0 and CHAID) and NN models that are built in different stages of the modeling phase. We try to be consistent to use the adjusted propensity scores reported in order to draw ROC Curve and calculate AUC. The raw propensities are based purely on estimates given by the model, which may be over-fitted, leading to over-optimistic estimates of propensity. Adjusted propensities attempt to compensate by looking at how the model performs on the test or validation partitions and adjusting the propensities to give a better estimate accordingly. A propensity score is the likelihood of a specific *yes* (high-cost) or *no* (low-cost) outcome on a scale from 0.0 to 1.0. If the adjusted score is not available, we continue to calculate AUCs based on other scores available, but it may challenge the comparability of results. We discuss these issues if this happens anywhere.

3-3-4 Silhouette measure of cluster cohesion and separation

This measure is an overall indication of clustering performance which is illustrated in the model summary view as a shaded diagram to indicate poor, fair, or good results. It averages, over all records, (B−A)/max (A, B), where A is the record's distance to its cluster center and B is the record's distance to the nearest cluster center that it doesn't belong to. A silhouette coefficient of 1 would mean that all cases are located directly on their cluster centers. A value of −1 would

mean all cases are located on the cluster centers of some other cluster. A value of 0 means, on average, cases are equidistant between their own cluster center and the nearest other cluster (IBM SPSS Modeler's help documentation).

Chapter 4 : Data Preparation

In this chapter we review the data preparation or the preprocessing phase in detail. We start by introducing the raw data and its related processing, and then we explain attribute selection, attribute reduction, and other preprocessing steps taken to reach to the preprocessed data.

4-1 Preprocessing of the Raw Data

The most comprehensive data file which is released by the AHRQ's MEPS is the MEPS Full-year Consolidated file (FC, abbreviated). For each year, the pertinent file consists of the MEPS survey data obtained in rounds 1, 2, and 3 of the new panel and rounds 3, 4, and 5 of the previous Panel, and contains variables pertaining to survey administration, demographics, employment, health status, disability days, quality of care, patient satisfaction, health insurance, income, and person-level medical care uses and expenditures. Released as an ASCII file (with related SAS and SPSS programming statements and data user information) and a SAS transport dataset, this public use file provides information collected on a nationally representative sample of the civilian non-institutionalized population of the United States for the relevant calendar year.

Table 4-1 shows the summary of retrieved data before any preprocessing and shows the panels, rounds, count of attributes (Fields), logical length of records, and number of records. The total number of records is 98,175 records for 3 years of the MEPS survey. We used the MS Access and the MS Excel software tools for further data preparation.

Table 4-1: Details of records by year of survey (before preprocessing)

Year	Starting panels (Rounds)	Ending panels (Rounds)	Fields (Attributes)	Record Length	Records (All Weights)	Records (Positive Weights)
FC2008	13 (1,2,3)	12 (3,4,5)	1823	5256	33,066	31,262
FC2007	12 (1,2,3)	11 (3,4,5)	1787	5173	30,964	29,370
FC2006	11 (1,2,3)	10 (3,4,5)	1672	4612	34,145	32,577
				Total	98,175	93,209

All the records belonging to the previous panel from each year's data file have been removed to discard all the redundant cases; and all 3 years data were merged in a single file. This file include Panel 11 for year 2006, Panel 12 for year 2007, and Panel 13 for year 2008, which contains the total expenditure data for the same year (TOTEXP1) along with all input variables. The total expenditures data for the second-year for Panel 11-13 have been extracted from the FC files of their second year in 2007-2009, respectively.

According to the MEPS website, data for persons with a positive person-level weight (PERWT*F) are estimates for the civilian non-institutionalized U.S. population for each specific survey year; and records with a zero person level weight should not be considered in any study conducted on the US household population because they don't represent any part of the population. By including all records with a PERWT*F greater than zero, we have 17119, 13015, and 18948 records from panel 11-13, respectively (49082 records total). An additional 3099 records had no report for their TOTEXP2 and have to be removed because they became ineligible somewhere in the second year of their survey (45983 records remains). Finally, we discard all non-adult population from our study, because there are too many attributes from our pool of final attributes that are not applicable to persons aged 17 or less (14279 records discarded). The final count of population in our study is 31,704 individuals (Table 4-2).

Table 4-2: Details of records by year of survey (after preprocessing)- Adult population

Year	Panel	Records #	Records (%)
2008	13	12,171	38%
2007	12	8,441	27%
2006	11	11,092	35%
	Total	**31704**	**100%**

4-2 Attribute Selection

In first iteration, based on insights taken from literature review, expert knowledge from medical field, and an extensive review of all attributes detailed in MEPS documentation files for years 2006-2008, we select 66 attributes. We categorize those under five different modules: Demographics (9 attributes), Health Status (17 attributes), Preventive Care (16 attributes), Priority Conditions (14 attributes), and Visits Counts (10 attributes). We further analyze each module attributes and introduce the selected set of attributes in each category that are finally used in the modeling phase. We simplify each attribute's name by removing the round-specific or year-specific number suffixes from each attribute's name.

4-2-1 Demographic attributes

Table 4-3 summarizes the titles and descriptions of 9 demographics attributes as shown in the MEPS database documentations.

Table 4-3: Demographic attributes

#	Attribute	Description
1	*AGE31X*	AGE - R3/1 (EDITED/IMPUTED)
2	*SEX*	SEX
3	*RACEX*	RACE (EDITED/IMPUTED)
4	*EDUCYR*	YEARS OF EDUC WHEN FIRST ENTERED MEPS
5	*HIDEG*	HIGHEST DEGREE WHEN FIRST ENTERED MEPS
6	*REGION*	CENSUS REGION AS OF 12/31/0?
7	*MARRY*	MARITAL STATUS-12/31/0? (EDITED/IMPUTED)
8	*TTLP*	PERSON'S TOTAL INCOME
9	*POVCAT*	FAMILY INC AS % POVERTY LINE-CATEGORICAL

All these attributes are categorical (either flag, Ordinal, or Nominal), except AGE, EDUCYR and TTLP which are continuous (Scale) variables.

AGE represents the age data relevant to Round 1 of the relevant panel. In the MEPS, it has been top-coded to 85 in original data and ranges from 18 to 85. *SEX* indicates the gender of each person, either male or female, with a value equal to 1 or 2, respectively. *RACE* indicates the race of the person and includes whites, blacks, American Indians/Alaska natives, Asians, native Hawaiian/Pacific Islanders, or multiple race report by the values of 1 to 6, respectively. *EDUCYR* is the number of years of education completed, which is based on the first round in which the number of years of education is collected for a person. In the MEPS, it has been top-coded to 17, where 0 shows no school attendance; 1-8 shows elementary grades 1 to 8; 9-11 shows high-school grades 9 to 11; 12 shows grade 12; 13-16 shows one to four years college attendance; and 17 stands for 5+ years of college. *HIDEG* indicates the highest degree of education for every person, and is based on the first round in which the highest degree was collected for a person. It accepts values of 1 to 7 for no degree, GED (General Educational Development), high-school diploma, bachelor's, master's, doctorate, or other degrees, respectively. *REGION* indicates the census region of each living unit including the survey participant. It accepts values of 1 to 4 for Northeast (Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont) , Midwest (Indiana, Illinois, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin), South (Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia) , and West (Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah,

Washington, and Wyoming) regions, respectively. *MARRY* shows the marital status of the person in the year of survey, and accepts values of 1 to 5 for married, widowed, divorced, separated, and never married. *TTLP* represents the total person-level income is the sum of all income components and accepts scale values either negative or positive. *POVCAT* is the categorical variable for family income as a percentage of poverty line. It accepts values of 1 to 5 for poor/negative incomes, near poor, low income, middle income, and high income, respectively.

Because of high level of bivariate correlation between HIDEG-EDUCYR (Pearson r= 0.706, significant, p-value<0.01) and between POVCAT-TTLP (Pearson r=0.478, significant, p-value<0.01), we remove EDUCYR and TTLP attributes from demographics set. The final attributes in the Demographics are **AGE, SEX, RACE, HIDEG, REGION, MARRY** and **POVCAT.**

4-2-2 Health Status attributes

Table 4-4 summarizes the titles and descriptions of 17 Health Status attributes as shown in the MEPS database documentations.

Table 4-4: Health Status attributes

#	Attribute	Description
1	*PCS*	SAQ:PHY COMPONENT SUMMRY SF-12V2
2	*MCS*	SAQ:MNT COMPONENT SUMMRY SF-12V2
3	*RTHLTH*	PERCEIVED HEALTH STATUS
4	*MNHLTH*	PERCEIVED MENTAL HEALTH STATUS
5	*IADLHP*	IADL SCREENER
6	*ADLHLP*	ADL SCREENER
7	*WLKLIM*	LIMITATION IN PHYSICAL FUNCTIONING
8	*LFTDIF*	DIFFICULTY LIFTING 10 POUNDS
9	*STPDIF*	DIFFICULTY WALKING UP 10 STEPS
10	*WLKDIF*	DIFFICULTY WALKING 3 BLOCKS
11	*MILDIF*	DIFFICULTY WALKING A MILE
12	*STNDIF*	DIFFICULTY STANDING 20 MINUTES
13	*AIDHLP*	USED ASSISTIVE DEVICES
14	*ACTLIM*	ANY LIMITATION WORK/HOUSEWRK/SCHL
15	*COGLIM*	COGNITIVE LIMITATIONS
16	*ANYLIM*	ANY LIMITATION
17	*BMINDX*	ADULT BODY MASS INDEX

All these attributes are categorical (either flag, Ordinal, or Nominal), except PCS, MCS, and BMINDX which are continuous (Scale) variables.

PCS and *MCS* are measures of general physical and mental health status perception by a person, and are derived from the well-known *short-form 12 version 2* (SF-12v2) health survey questionnaires which is self-administered. In analyzing data from the SF-12v2, the standard approach is to form two summary scores based on responses to these questions. The scoring algorithms for both the physical component (PCS) and the mental component (MCS) incorporate information from all 12 questions, but gives different weights to the questions. Both PCS and MCS values are recorded in the MEPS as a scale measure that accepts a range of 5.85-73.09 and 1.04-78.08 numerical values, respectively.

The MEPS survey uses a single question to record all family members' perceived health with respect to their physical (*RTHLTH*) and mental (*MNHLTH*) health. The answers are recorded on an ordinal basis between 1 and 5, representing excellent, very good, good, fair, or poor health perceptions, respectively. Some researchers claim that this single question can better represent overall health status than complex measures as SF-12v2 tool; and is still simpler to be collected and interpreted (Fleishman & Cohen 2010).

The MEPS investigates the presence of different types of limitations extensively, which includes limitations in daily activities (represented by *IADLHP* and *ADLHLP*), functional limitations (*WLKLIM, LFTDIF , STPDIF, WLKDIF, MILDIF,* and *STNDIF*) , use of assistive technology (*AIDHLP*), work, housework and school limitations (*ACTLIM*), cognitive limitations (*COGLIM*), hearing and visual limitations (which we did not include in this study directly). Finally, the MEPS aggregate all these attributes and reports presence or absence of any of these limitations in a person (*ANYLIM*).

IADLHP is the indicator of *Instrumental Activities of Daily Living* (IADL) help or supervision; and is true if anyone received help or supervision with IADLs including using the telephone, paying bills, taking medications, preparing light meals, doing laundry, or going shopping; and indeed, this was the result of an impairment or physical or mental health problem. ADLHLP shows the *Activities of Daily Living* (ADL) help or supervision, and includes routine personal activities in everyday living that does not need any instrument use. Both IADLHP and ADLHLP are flag attributes.

The functional limitations are defined as any difficulty in performing certain specific physical activities and the MEPS filters it in the family through a single question: "Does anyone have difficulties walking, climbing stairs, grasping objects, reaching overhead, lifting, bending or

stooping, or standing for long periods of time?" This question identifies the category of possible related attributes. If the answer is positive, the following limitations are further investigated by using more questions: LFTDIF i.e. difficulty lifting 10 pounds, STPDIF i.e. difficulty walking up 10 steps, WLKDIF i.e. difficulty walking 3 blocks, MILDIF i.e. difficulty walking a mile, and STNDIF i.e. difficulty standing 20 minutes. The filtering question for the WLKLIM attribute has a flag response; and other five functional limitations attributes take ordinal values between 1-4, showing no difficulty, some difficulty, a lot of difficulty, and unable to do, respectively. Remaining attributes are more straightforward. AIDHLP indicates use of assistive technology; ACTLIM shows any limitation in work, housework, or school; and COGLIM shows if any adult family member is experiencing confusion or memory loss, having problems in making decisions, or needs supervision for her own safety. ANYLIM summarizes whether a person has any ADL, IADL, activity, functional, or sensory limitations in any of the pertinent rounds. *BMINDX* is a measure of body mass index and person's obesity and weight. BMI in adults ranges from underweight (<18.5), normal weight (18.5–24.9) inclusive, overweight (25.0–29.9), to obesity (>= 30.0). The MEPS reports it as a continuous measure and it accepts a value between 9.4 and 256.3 in the final dataset.

We analyze Health status attributes further to reduce their count as much as possible. RTHLTH and MNHLTH are easily gathered in any survey with asking only one questions from participants, compared to PCS and MCS which are more difficult to use in a survey and very difficult to interpret its variations among different respondents. Both PCS and MCS show a significant correlation with either RTHLTH (Pearson r is -0.378, and -0.220 respectively) or MNHLTH (Pearson r is -0.227, and -0.261 respectively) in a two-tailed test (p-values<0.01). The inverse correlation is due to inverse scoring of RTHLTH in the MEPS. The correlation with an

expenditure target (TOTEXP1 i.e. the total expenditures in the current year when setting the flag label at 95 percentile) is grater for RTHLTH (-0.225) compared to PCS (0.179), but both are significant (p-values<0.01) in a two-tailed test and we cannot conclude more. Considering these results and discussions for four attributes representing overall health perceptions, we pick RTHLTH and MNHLTH and discard PCS and MCS from final attributes list.

With respect to the functional limitations attributes, we follow a simple logic. ANYLIM is a true aggregate of all other attributes in this category and has been designed intentionally to do so in the MEPS survey. ANYLIM has a strong bivariate correlation (Pearson rho coefficient) to IADHLP (0.425), ADLHLP (0.296), WLKLIM (0.641), AIDHLP (0.361), ACTLIM (0.538), and COGLIM (0.321), which is always statistically significant in a two-tailed test (p-value<0.01). Some functional limitations attributes have too many missing values to be useful for model building (Table 4-5). We discard all sub-attributes and keep only ANYLIM which represent aggregate predictive strengths of all functional limitations attributes together.

Table 4-5: Health Status attributes: Counts and percentage of missing values

Attribute	Missing #	Missing (%)
IADLHP	12	<1
ADLHLP	8	<1
WLKLIM	83	<1
LFTDIF	27433	87
STPDIF	27433	87
WLKDIF	27444	87
MILDIF	27472	87
STNDIF	27442	87
AIDHLP	9	<1
ACTLIM	29	<1
COGLIM	61	<1
ANYLIM	374	1

The final attribute list for the Health Status attributes includes four attributes including **RTHLTH, MNHLTH, ANYLIM, and BMINDX.**

4-2-3 Preventive Care attributes

Table 4-6 summarizes the titles and descriptions of 16 Preventive Care attributes as shown in the MEPS database documentations.

Table 4-6: Preventive Care attributes

#	Attribute	Description	Missing (%)
1	*CHECK*	HOW LNG LST ROUTNE CHECKUP	4
2	*BPCHEK*	TIME SNCE LST BLOOD PRES CHK	3
3	*BPMONT*	# MOS SNCE LST BLOOD PRES CHK	20
4	*CHOLCK*	HOW LNG CHOLEST LST CHCK	7
5	*NOFAT*	RESTRICT HGH FAT/CHOLES FOOD	2
6	*EXRCIS*	ADVISED TO EXERCISE MORE	2
7	*ASPRIN*	TKE ASPIRN EVERY (OTHR) DAY	<1
8	*PSA*	HOW LONG SINCE LAST PSA	75
9	*BOWEL*	SIGMOIDOSCOPY/COLONOSCOPY	2
10	*STOOL*	BLD STOOL TST KIT/CRDS HOME	3
11	*MAMOGR*	HOW LNG SNCE LST MAMMOGRAM	58
12	*BRSTEX*	HOW LNG SNCE LST BREAST EXAM	48
13	*PAPSMR*	HOW LNG LST PAP SMEAR TST	49
14	*HYSTER*	HAD A HYSTERECTOMY	46
15	*DENTCK*	HOW OFTEN DENTAL CHECK-UP	1
16	*LSTETH*	LOST ALL UPPR AND LOWR TEETH	<1

All these attributes are gathered during round 3 of relevant panel's survey. The MEPS reports all these attributes under health status section, but we group them together in because all are indicators of routine clinical check-ups (CHECK, BPCHEK, BPMONT, CHOLCK, DENTCK), primary measures employed to prevent diseases initiation and/or early detection of diseases (PSA, BOWEL, STOOL, MAMOGR, BRSTEX, PAPSMR, HYSTER), or secondary measures for disease prevention/risk factor control/risk factor indicator (BPCHEK, BPMONT, CHOLCK, NOFAT, EXRCIS, ASPRIN, LSTETH). For the purpose of this study, we label them altogether as the Preventive Care attributes.

CHECK, BPCHEK, CHOLCK, PSA, MAMOGR, BRSTEX, and *PAPSMR* show frequency of a diagnostic intervention on ordinal basis, accepting values between 1-6, i.e. 1 through 5 shows the intervention has been done for the person within past year, past two years, past three years, past 5 years, and more than five years, respectively, and 6 stands for never having that intervention done in past. *DENTCK* is ordinal, and accepts values between 1-4 that shows the dental check-up has been done for a person twice a year or more, once a year, less than once a year, and never. *BPMONT* is a continuous version of BPCHEK which shows the number of months from last blood pressure check-up ranging between 0-24 months. All other attributes in this module are flag attributes accepting 1 (yes) or 2 (no) values.

We discard PSA, MAMOGR, BRSTEX, PAPSMR, HYSTER attributes from our study because they have a high percents of missing values that make them unfavorable to use in model building, and they are gathered over a specific age range and/or in a specific gender. BPCHEK and BPMONT are coding same data in different measures, and we remove BPMONT in favor of BPCHEK. We use DENTCK as a single dental care attribute, as it has more preventive care meaning than LSTETH.

The final set of selected attributes in this module includes **CHECK**, **BPCHEK**, **CHOLCK**, **NOFAT**, **EXRCIS**, **ASPRIN**, **BOWEL**, **STOOL,** and **DENTCK.**

4-2-4 Priority Conditions attributes

Table 4-7 summarizes the titles and descriptions of 14 Priority Conditions attributes as shown in the MEPS database documentations.

Table 4-7: Priority conditions attributes

#	Attribute	Description
1	*HIBPDX*	MULT DIAG HIGH BLOOD PRESS
2	*CHDDX*	CORONARY HRT DISEASE DIAG
3	*ANGIDX*	ANGINA DIAGNOSIS
4	*MIDX*	HEART ATTACK (MI) DIAG
5	*OHRTDX*	OTHER HEART DISEASE DIAG
6	*STRKDX*	STROKE DIAGNOSIS
7	*ASTHDX*	ASTHMA DIAGNOSIS
8	*EMPHDX*	EMPHYSEMA DIAGNOSIS
9	*CANCERDX*	CANCER DIAGNOSIS
10	*DIABDX*	DIABETES DIAGNOSIS
11	*ARTHDX*	ARTHRITIS DIAGNOSIS
12	*CHOLDX*	HIGH CHOLESTEROL DIAGNOSIS
13	*PC*	IS THERE ANY PC AT ALL?
14	*PCCOUNT*	COUNT OF PC IN AN INDIVIDUAL

All priority conditions attributes are of flag type, showing either having or lacking a specific medical condition's diagnosed for the person; besides *PCCOUNT* which is continuous and its value ranges from 0-12. This attribute and the PC attribute have been derived from other 12 disease conditions attributes. PC shows if a person has any priority conditions at all, and PCCOUNT represents count of priority conditions reported in each person. They help to build different verities of predictive models by using all 14 attributes including disease conditions attributes along with both PC and PCCOUNT, or using 13 attributes including disease conditions attributes along with PC (and discarding PCCOUNT), or using 12 attributes including only disease conditions alone (and discarding both PC and PCCOUNT).

As CANCERDX is not reported directly in the MEPS surveys before 2008, we use the same preprocessing steps taken for the FC files2006-2008 and apply them to the MEPS Medical Condition data in 2006-2007. We import two attributes *DUPERSID* and *CCCODEX* from these files to relational database software e.g. the MS Access. DUPERSID is used as the key identifier,

and CCCODEX lists all conditions reported for each person in multiple rows. We extract the CCCODEX for all types of cancers from relevant MEPS codebooks, and group them by using *IF*, *WHERE* and *OR* commands in a new single column. We replace CANCERDX field in year 2006-2007's dataset with this column. We did not delete any attribute from this module and will use them all in model building process.

4-2-5 Visits Counts attributes

Table 4-8 summarizes the titles and descriptions of 10 Visits Counts attributes as shown in the MEPS database documentations.

Table 4-8: Visits Counts attributes

#	Attribute	Description	Range
1	*OBTOTV*	# OFFICE-BASED PROVIDER VISITS	261
2	*OBDRV*	# OFFICE-BASED PHYSICIAN VISITS	163
3	*OBOTHV*	# OFFICE-BASED NON-PHYSICAN VISITS	261
4	*OPTOTV*	# OUTPATIENT DEPT PROVIDER VISITS	156
5	*OPDRV*	# OUTPATIENT DEPT PHYSICIAN VISITS	156
6	*OPOTHV*	# OUTPATIENT DEPT NON-DR VISITS	156
7	*ERTOT*	# EMERGENCY ROOM VISITS	22
8	*IPDIS*	# HOSPITAL DISCHARGES	10
9	*IPNGTD*	# NIGHTS IN HOSP FOR DISCHARGES	137
10	*RXTOT*	# PRESC MEDS INCL REFILLS	292

All Visits Counts variables are continuous data and show the counts of an individual's visits to the different level of health care. The names of these attributes are indicative of data they are representing, for example OBTOTV is the **o**ffice-**b**ased **tot**al **V**isits counts which is provided by a physician (OB**DR**V) or ano**the**r non-physician (OB**OTH**V). We include medicine usage in this module because it simply represents an individual's counts of visits to pharmacies (RXTOT). It is a utilization variable for medicine usage and is a count of all prescribed medications purchased during pertinent survey year, and includes initial purchases and refills.

We normalize all these variables in a new scale measure between 0-1, in order to prevent variables with a higher range of values (RXTOT and OBTOTV) from overshadowing variables that accept only a short range of values (IPDIS and ERTOT). The range of values for each variable is shown in relevant column in Table 4-8.

We discard OBDRV and OBOTHV in favor of OBTOTV, and similarly OPDRV and OPOTHV in favor of OPTOTV. We delete IPNGTD and retain IPDIS as they confer same information.

The final set of attributes for Visits Counts module is **OBTOTV**, **OPTOTV**, **ERTOT**, **IPDIS**, and **RXTOT**.

4-3 Selected attributes in the final MEPS dataset

After attribute reduction efforts on primary set of 66 attributes, the final dataset has 39 attributes which are categorized in 5 modules:

- Demographics (7 attributes): AGE, SEX, RACE, HIDEG, REGION, MARRY, and POVCAT.
- Health Status (4 attributes): RTHLTH, MNHLTH, ANYLIM, and BMINDX.
- Preventive Care (9 attributes): CHECK, BPCHEK, CHOLCK, NOFAT, EXRCIS, ASPRIN, BOWEL, STOOL, and DENTCK.
- Priority Conditions (14 attributes): HIBPDX, CHDDX, ANGIDX, MIDX, OHRTDX, STRKDX, ASTHDX, EMPHDX, CANCERDX, DIABDX, ARTHDX, CHOLDX, PC, and PCCOUNT.
- Visits Counts (5 attributes): OBTOT, OPTOT, ERTOT, IPDIS, and RXTOT (we removed letter “V” from end of these variables if existed, to make their use easier.)

From now on, we reword Health Status, Preventive Care, Priority Conditions, Visits Counts variables as Health, Preventive, PrioC, and Visits, respectively.

4-3-1Preprocessing of the selected attributes

By completing preliminary processing of attributes, the final dataset has 39 fields for 31704 records. Before feeding into models, we bin values in all fields to make models work better. With respect to the Demographics module, we change AGE's measurement type from continuous to nominal by binning it into three categorical values, i.e. 1 for 18-49 years, 2 for 50-65 years, and 3 for ages over 65. We re-bin the RACE into a new nominal attribute with values of 1-3 for whites, blacks, and others, respectively. It reduces the number of categories for RACE attribute from six to three and makes interpretations much easier. HIDEG is re-binned from its original 7-categories nominal type to a new flag (binary or dummy) attribute and accepts1 for *having a higher education*, and 2 for *lacking* it. We re-bin MARRY and POVCAT variables into a two new attributes which accept three nominal values. The new nominal attribute for MARRY, accepts values of 1, 2, and zero for married, widowed/separated/divorced, or never married states. In the new nominal POVCAT attribute, families are categorized to negative/poor/near poor, low/middle, or high incomers by accepting a value of 1 to 3, respectively.

With respect to the Health module, besides BMINDX, we leave all attributes unchanged. BMINDX attribute has been changed from continuous to ordinal by using the same definitions used in the MEPS documentations that are compatible with the US-CDC's definitions for body mass index in adults. New BMINDX attribute accepts values from 1 to 4 in an ordinal basis for underweight, normal weight, overweight, and obese persons, respectively.

With respect to the Preventive module, we leave all flag (binary, or dummy) attributes unchanged, and re-bin all nominal attributes (CHECK, BPCHEK, and CHOLCK) in three values i.e. 1 for *within past year*, 2 for *within 2 years or more*, and 0 for *never*. We re-bin DENTCK

values a little different, where 1, 2, and 0 values represent *twice a year* or more, *less than twice a year*, and *never*.

With respect to the PrioC module, we leave all attributes in their original flag state, besides binning PCCOUNT to a nominal measure with three values i.e. zero value for having no disease conditions, 1 to 3 values for having one to three disease condition, and 4 for having four diseases conditions or more.

With respect to the Visits module, we use the normalized 0-1 value state. We build a separate flag (binary or dummy) attribute for each of these attribute in order to compare its efficiency with that of continuous attributes. We use a threshold for a true (1) value for these new flag attributes which is an absolute count of 5 visits for OBTOT, 12 visits for RXTOT, and greater than zero visits/encounters for OPTOT, ERTOT, and IPDIS. These thresholds are cut-off points for top 30 percentiles.

Table 4-9 shows all 39 attributes from five modules along with each attribute's measurement type, range, and values they accept in the final dataset.

Table 4-9: The final dataset- attributes by modules

	Attribute	Data Type	Values	Missing #
Demographics	*AGE3*	Nominal	1,2,3	0
	SEX	Flag	1,2	0
	RACEX	Nominal	1,2,3	0
	HIDEG	Flag	1,2	199
	REGION	Nominal	1,2,3,4	10
	MARRY	Nominal	0,1,2	2
	POVCAT	Ordinal	1,2,3	0
Health	*RTHLTH*	Ordinal	1,2,3,4,5	0
	MNHLTH	Ordinal	1,2,3,4,5	0
	ANYLIM	Flag	1,2	374
	BMINDX	Ordinal	1,2,3,4	895
Preventive	*CHECK*	Nominal	0,1,2	1183
	BPCHEK	Nominal	0,1,2	1011
	CHOLCK	Nominal	0,1,2	2079
	NOFAT	Flag	1,2	658
	EXRCIS	Flag	1,2	571
	ASPRIN	Flag	1,2	282
	BOWEL	Flag	1,2	703
	STOOL	Flag	1,2	999
	DENTCK	Nominal	0,1,2	395
PrioC	*HIBPDX*	Flag	1,2	99
	CHDDX	Flag	1,2	91
	ANGIDX	Flag	1,2	90
	MIDX	Flag	1,2	78
	OHRTDX	Flag	1,2	86
	STRKDX	Flag	1,2	73
	ASTHDX	Flag	1,2	63
	EMPHDX	Flag	1,2	67
	CANCERDX	Flag	1,2	1
	DIABDX	Flag	1,2	69
	ARTHDX	Flag	1,2	119
	CHOLDX	Flag	1,2	189
	PC	Flag	1,2	0
	PCCOUNT	Nominal	0,1,2,3,4	0
Visits	*OBTOT*	Continuous	[0.0-1.0]	0
	OPTOT	Continuous	[0.0-1.0]	0
	ERTOT	Continuous	[0.0-1.0]	0
	IPDIS	Continuous	[0.0-1.0]	0
	RXTOT	Continuous	[0.0-1.0]	0

4-4 The missing values

In the final dataset, 26 out of 39 attributes have missing values (Table 4-10). The highest percentage of missing values in our data was 7 percent, while 12 fields had less than 1% missing and remaining 14 fields ranged between these two extremes. Table 4-10 shows the value labels for missing values in each pertinent field.

Table 4-10: Missing values in the final dataset

#	Attribute	Values	Missing Labels
1	HIDEG	1,2	8, -7, -8, -9
2	REGION	1,2,3,4	-1
3	MARRY	0,1,2	-9
4	ANYLIM	1,2	-9
5	BMINDX	1,2,3,4	-1, -9
6	CHECK	0,1,2	-1, -7, -8, -9
7	BPCHEK	0,1,2	-1, -7, -8, -9
8	CHOLCK	0,1,2	-1, -7, -8, -9
9	NOFAT	1,2	-1, -7, -8, -9
10	EXRCIS	1,2	-1, -7, -8, -9
11	ASPRIN	1,2	-1, -7, -8, -9
12	BOWEL	1,2	-1, -7, -8, -9
13	STOOL	1,2	-1, -7, -8, -9
14	DENTCK	0,1,2	-1, -7, -8, -9
15	HIBPDX	1,2	-1, -7, -8, -9
16	CHDDX	1,2	-1, -7, -8, -9
17	ANGIDX	1,2	-1, -7, -8, -9
18	MIDX	1,2	-1, -7, -8, -9
19	OHRTDX	1,2	-1, -7, -8, -9
20	STRKDX	1,2	-1, -7, -8, -9
21	ASTHDX	1,2	-1, -7, -8, -9
22	EMPHDX	1,2	-1, -7, -8, -9
23	CANCERDX	1,2	-9
24	DIABDX	1,2	-1, -7, -8, -9
25	ARTHDX	1,2	-1, -7, -8, -9
26	CHOLDX	1,2	-1, -7, -8, -9

In all attributes the missing values are labeled the same, and -1, -7, -8, and -9 shows *inapplicable*, *refused (to answer)*, *don't know*, and *not ascertained* values, respectively. Only in HIDEG attribute, an 8 value represents inapplicable which includes persons aged fewer than 16 that are excluded from our study pool from the beginning.

The IBM SPSS Modeler provides extensive options to manage missing values. Choice of handling approach depends on a number of factors including size of the dataset, number of fields containing blanks or missing values, and amount of missing information. Based on these criteria,

we can exclude fields or records with missing values, or impute, replace, or coerce missing values using a variety of methods. Some DM models handle the missing values efficiently, but it is better to don't include those attributes with too many missing values in all models. It means that when most missing values are concentrated in a small set of fields, it is better to handle them at the field level, instead of record level. In the primary MEPS dataset, in some attributes including LFTDIF, STPDIF, WLKDIF, MILDIF, STNDIF, PSA, MAMOGR, BRSTEX, PAPSMR, and HYSTER, the missing values abound and they have been discarded from the final dataset in this study.

Some DM models handle missing data better than others. For example, most DTs including C5.0 and Apriori cope well with values that are explicitly declared as missing. Some other DM models as NN have trouble dealing with missing values and experience longer training times, resulting in less-accurate models. For a better modeling performance, we should either discard all the records with the missing values, discard a complete field from analysis, or impute them in a suitable way.

We have very few missing values in the final dataset and most models can handle them efficiently. We keep all fields with the missing values and change all of their labels to -1. We flag them during model building to let the DM model identify them correctly. This change does not affect model functioning and accuracies. All DT models handle the missing values in their modeling process, but for NN and Clustering (K-Means), we delete all the records with missing label and we do not use them in training process.

4-5 Target attributes

With respect to the target attributes, as we aim to predict high-cost instances among general population in the US, the target attributes are ideally flag attributes identifying *high-cost* and

low-cost instances by *true* and *false* values, respectively. We define two sets of target attributes to build two sets of predictive models. In the first set, the models predict the *current year*'s total expenditure (we name it *TOTEXP1*) using the *current year*'s input variable. In the second set, the models predict the *next year*'s total expenditure (we name it *TOTEXP2*) using the *current year*'s input variable

For both TOTEXP1 and TOTEXP2 targets, we build a flag (binary or dummy) attribute, and we define a threshold for the high-cost individuals using both the insights from the literature and the nature of expenditures distributions among the MEPS survey. A rapid look into TOTXP1 data shows highly skewed values in this field as shown in its histogram (Figure 4-1).

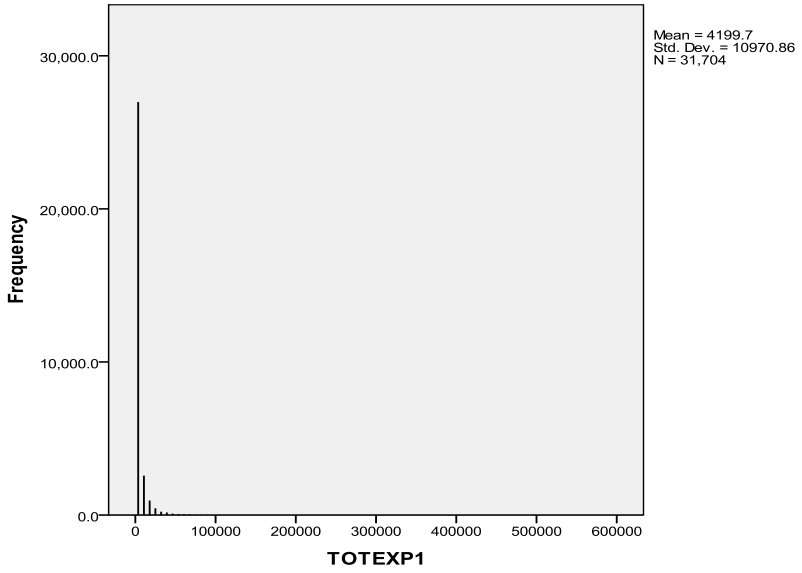

Figure 4-1: Histogram of TOTEXP1 data

A further descriptive analysis of central measures for TOTEXP1 confirms a highly-skewed data distribution (Table 4-11), which does not conform the normal distribution as shown in normal probability plot (P-Plot) of TOTEXP1 (Figure 4-2). Probability plots draw a variable's

cumulative proportions against the cumulative proportions of a test distribution and are generally used to determine whether the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points cluster around a straight line.

Table 4-11: Central measures for TOTEXP1 data

Descriptive Statistics			Percentiles	
N	Valid	31704	10	0
	Missing	0	20	60
Mean		4199.7	25	151
Median		1052.5	30	260
Mode		0	40	575
Std. Deviation		10970.86	50	1052.5
Skewness		12.98	60	1822
Std. Error of Skewness		0.014	70	3027.5
Range		521209	75	3920.75
Minimum		0	80	5263
Maximum		521209	90	10385.5
Sum		133147351	95	17923.25

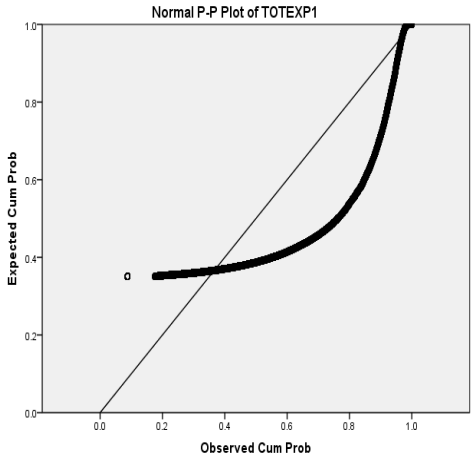

Figure 4-2: Normal distribution's P-Plot for TOTEXP1data

Further analysis for *goodness of fit* for TOTEXP1 data and major distribution patterns (Table 4-12) reveals the Weibull distribution is closer to TOTEXP1 data with the highest p-value (p=.010) which is still non-significant.

Table 4-12: Goodness of Fit statistics for TOTEXP1 data

Distribution	AD	p-value
Weibull	1850.515	<0.010
Gamma	933.376	<0.005
Logistic	3212.953	<0.005
Normal	5593.789	<0.005
Loglogistic	2733.354	<0.005
Exponential (2P)	14727.837	<0.003
Lognormal	7198.449	0.000

To identify individual distribution fit with Weibull, we replaced all non-positive values (lowest possible cost = 0) with 0.01. The fit is more evident when we ignore zero values (Figure 4-3).

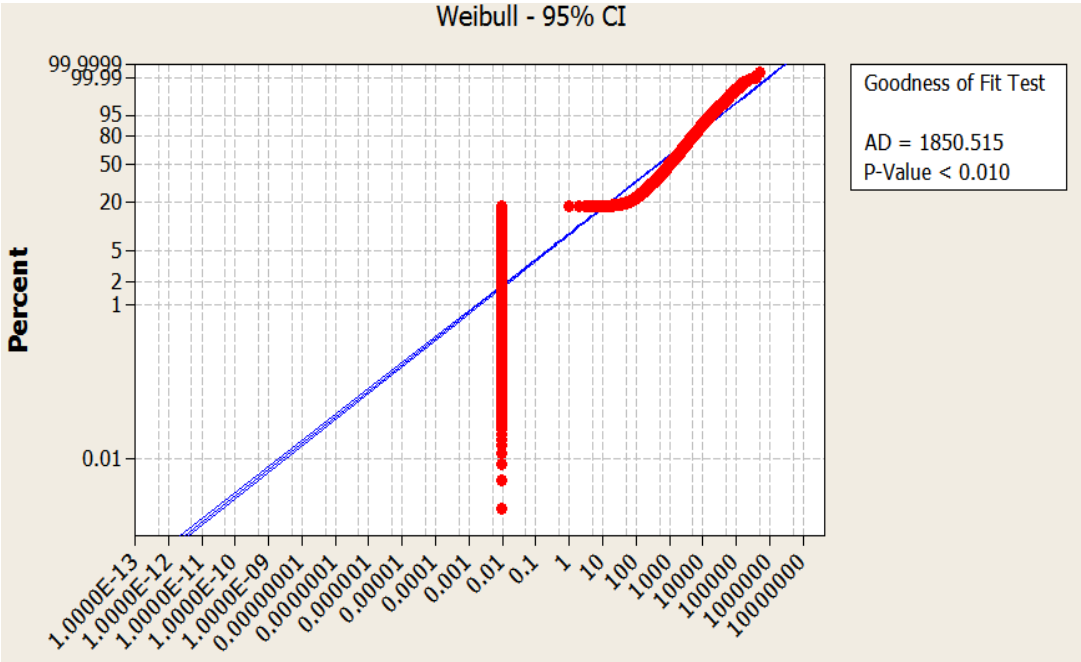

Figure 4-3: Weibull distribution fit to TOTEXP1 data

Simple testing of TOTEXP1 data against 20/80 rule is useful. In the final dataset, out of $133147351 cumulative cost of TOTEXP1 for three years, 45%, 61%, and 78% is incurred by top 5, 10, and 20 percentiles, respectively. Similar cost figures are evident for TOTEXP2 (Tables 4-13 and 14).

Table 4-13: Test of 20/80 rule for TOTEXP1 data

Percentile	Expenditures ($)	Expenditures (%)
Top 5%	59,575,116	45
Top 10%	80,914,106	61
Top 20%	104,156,078	78
All Records (100%)	133,147,351	100

Table 4-14: Test of 20/80 rule for TOTEXP2 data

Percentile	Expenditures ($)	Expenditures (%)
Top 5%	67,365,197	48
Top 10%	88,896,390	64
Top 20%	113,365,014	81
All Records (100%)	140,149,629	100

Based on results of TOTEXP data analysis, we select three targets for current year and three target for next year expenditure. In each year we set three separate thresholds for identifying high-cost cases at 95, 90 and 80 percentiles, and name them TOTEXP 95, 90, and 80, respectively (Table 4-15).

Table 4-15: Target fields for Current year/Next year expenditure data

Target Title	Target Year	High-cost Threshold
TOTEXP1-95	Current Year	> 95 Percentile
TOTEXP1-90	Current Year	> 90 Percentile
TOTEXP1-80	Current Year	> 80 Percentile
TOTEXP2-95	Next Year	> 95 Percentile
TOTEXP2-90	Next Year	> 90 Percentile
TOTEXP2-80	Next Year	> 80 Percentile

Chapter 5 : Modeling and Analysis of the Results

In this chapter, we present the results of the modeling phase in four parts. In the first three parts, we present the classification models including DTs and NNs; and in the fourth part we review the clustering results. First, we compare the performances of DT and NN classifiers modeled with 39 attributes using multiple evaluation metrics. In the second part, we compare two DTs (C5.0 and CHAID algorithms) by building models based on different sets of attributes (Demographics, Health, Preventive, PrioC, and Visits), then we choose the best tree algorithm for further model building according to their performances. In the third part, we use the best selected tree classifier, to build models on various combinations of different sets of attributes (i.e. Demographics / Health, or Demographics / Preventive, etc.) and based on their performance results, we introduce the minimal set of attributes for predicting high-cost instances. Finally, in the fourth part, we review the results of the clustering analysis (K-Means algorithm) in details.

5-1 Classification results: DT and NN models using 39 input attributes

The modeling efforts start with building classification models including DT and NN classifiers and uses all 39 input attributes, including 7, 4, 9, 14, and 5 attributes from Demographics, Health, Preventive, PrioC, and Visits modules, respectively. A list of these attributes is shown in Table 5-1.

Table 5-1: List of 39 input attributes in the final MEPS dataset

1- Demographics	2- Health	3- Preventive	4- PrioC		5- Visits
AGE3	RTHLTH	CHECK	HIBPDX	EMPHDX	OBTOT
SEX	MNHLTH	BPCHEK	CHDDX	CANCERDX	OPTOT
RACEX	ANYLIM	CHOLCK	ANGIDX	DIABDX	ERTOT
HIDEG	BMINDX	NOFAT	MIDX	ARTHDX	IPDIS
REGION		EXRCIS	OHRTDX	CHOLDX	RXTOT
MARRY		ASPRIN	STRKDX	PC	
POVCAT		BOWEL	ASTHDX	PCCOUNT	
		STOOL			
		DENTCK			

All 39 attributes show significant correlation to the expenditure targets i.e. TOTEXP1 and TOTEXP2. As an example, all Visits attributes show a high and significant correlation to current year's (TOTEXP1) or to next year's (TOTEXP2) expenditure. It is rational to have higher health costs when you visit health system more frequently and at different level of care (office-based visits, outpatient, emergency rooms, and inpatient stays). All other modules show significant correlations to expenditure targets (when α=0.05) which is shown in their Pearson rho coefficients (Table 5-2).

Table 5-2: Pearson rho for correlations to expenditure targets (significant correlations with α=0.05 are shown in bold.)

	TOTEXP1	TOTEXP2		TOTEXP1	TOTEXP2
AGE3	**.216**	**.228**	*HIBPDX*	**-.197**	**-.200**
SEX	**.052**	**.039**	*CHDDX*	**-.195**	**-.168**
RACE	**-.017**	**-.014**	*ANGIDX*	**-.130**	**-.119**
HIDEG	.009	.007	*MIDX*	**-.163**	**-.157**
REGION	**-.040**	**-.036**	*OHRTDX*	**-.171**	**-.153**
MARRY	**-.062**	**-.052**	*STRKDX*	**-.137**	**-.135**
POVCAT	.000	-.003	*ASTHDX*	**-.081**	**-.090**
RTHLTH	**.254**	**.240**	*EMPHDX*	**-.116**	**-.119**
MNHLTH	**.149**	**.144**	*CANCERDX*	**-.160**	**-.155**
ANYLIM	**-.259**	**-.250**	*DIABDX*	**-.174**	**-.186**
BMINDX	**.060**	**.072**	*ARTHDX*	**-.210**	**-.211**
CHECK	**-.138**	**-.134**	*CHOLDX*	**-.165**	**-.165**
BPCHEK	**-.145**	**-.131**	*PC*	**-.247**	**-.233**
CHOLCK	**-.174**	**-.172**	*PCCOUNT*	**.324**	**.320**
NOFAT	**-.132**	**-.120**	*OBTOT*	**.417**	**.310**
EXRCIS	**-.113**	**-.106**	*OPTOT*	**.234**	**.172**
ASPRIN	**-.158**	**-.163**	*ERTOT*	**.253**	**.142**
BOWEL	**-.202**	**-.197**	*IPDIS*	**.566**	**.221**
STOOL	**-.150**	**-.134**	*RXTOT*	**.416**	**.371**
DENTCK	**-.032**	**-.025**	*TOTEXP1*	**1**	**.347**

5-1-1 Performances of DT and NN models using 39 input attributes

The performance measures for all classifiers when using 39 input attributes are shown in Table 5-3 and Figure 5-1 detailed by six targets from current and next years. We use both *Accuracy* and *AUC* to evaluate the modeling results in the test partition.

Table 5-3: Performance evaluation of classifiers using all 39 attributes

	Accuracy			AUC		
Target	C5.0-Acc	CHAID-Acc	NN-Acc	C5.0-AUC	CHAID-AUC	NN -AUC
TOTEXP1-.95	**93.7**	**86.3**	**76.2**	**0.816**	**0.946**	**0.956**
TOTEXP1-.90	90.3	85.8	75.2	0.759	0.933	0.946
TOTEXP1-.80	86.7	84.0	74.6	0.729	0.913	0.940
TOTEXP2-.95	90.1	77.7	67.7	0.688	0.828	0.866
TOTEXP2-.90	83.9	80.1	66.1	0.659	0.827	0.842
TOTEXP2-.80	75.6	75.6	66.7	0.607	0.831	0.843

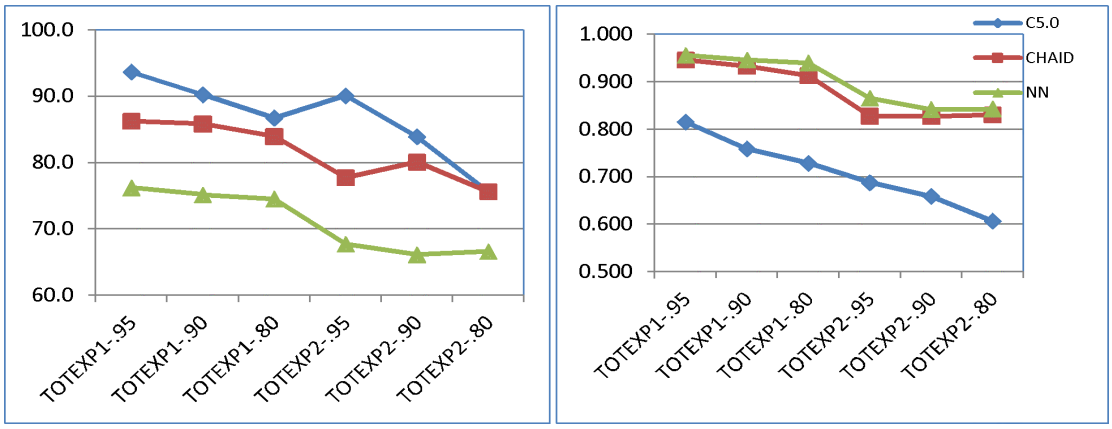

Figure 5-1: Correctness accuracy (Left) and AUC (Right) for classifiers using all 39 attributes

As all input attributes are significantly correlated to target attributes, it was expected that the classifiers should perform well when using all 39 input attributes. The resultant correctness accuracy and AUC of all classifiers are much higher than that of a random guess and it confirms proper model functioning. Good performances of these DT and NN classifiers show that the final MEPS dataset has been prepared properly and the dataset is valid.

The test accuracies of these models varies between 66.1-93.7 percent, and the average correctness accuracy for the 18 models built on 39 input attributes is 79.8%. It includes a lower score for NN and the CHAID and higher scores for C5.0 classifiers. The AUCs range between

0.607-0.956 and the average AUCs for the 18 models built on 39 input variables is 0.829. It includes a lowest score for C5.0 and higher scores for NN and the CHAID.

Whatever the performance measure is, all models predict better when the threshold for high-cost is set at the 95 percentile (top 5 percent or *very high-cost* instances) than *high-cost* instances (threshold at 90 or 80 percentiles).

All models predict the current year's cost better than next year's cost, which is considered a given fact i.e. all models built for the current year, show a higher correctness accuracy or AUC compared to next year's models. All TOTEXP1-95, TOTEXP1-90, and TOTEXP 1-80 models perform better than their counterparts in TOTEXP2-95, TOTEXP2-90, and TOTEXP 2-80 models, respectively. Based on these results, we use the TOTEXP1-95 as the single cost target throughout the remaining parts of our analysis.

Regarding the performance measures, although C5.0 yields the highest correctness accuracy rates (93.7%) comparing to CHAID (86.3%) and NN (76.2%), the AUC measure shows the opposite, with the lowest areas for C5.0 (0.816), and much higher areas for CHAID (0.946), and NN (0.956). A quick look into the confusion matrices for these three models explains that how the AUC measure works better to rank these classifiers (Figure 5-2). (Note: we discarded null values from NN confusion matrix for ease of comparison.)

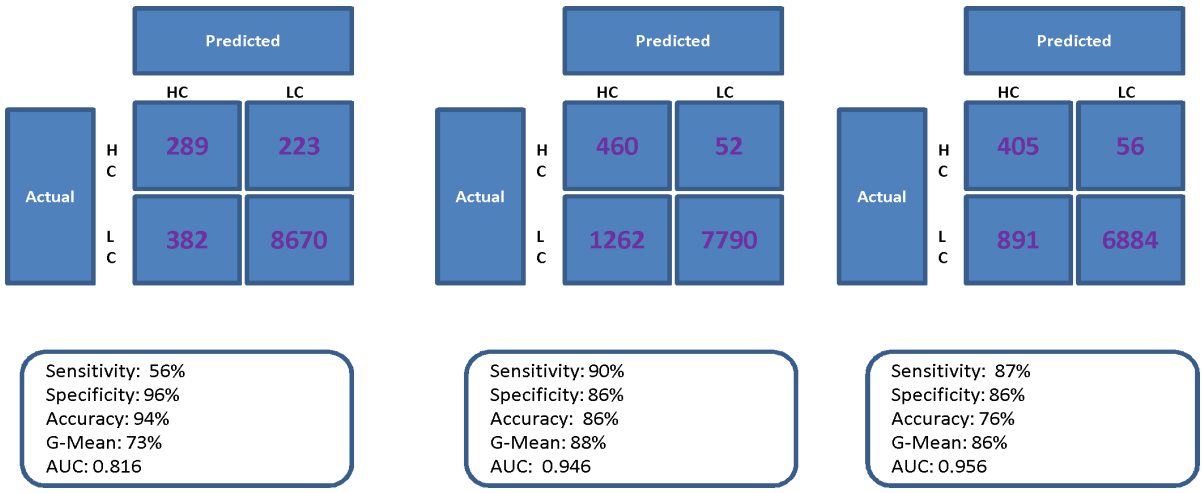

Figure 5-2: Confusion Matrices for C5.0 (Left), CHAID (Middle), and NN (Right) models (Target=TOTEXP1-95)

In the final MEPS dataset, we set the high-cost at 95 percentile; therefore, the target's outcomes abound in false values. The low-costs instances comprise 95% of all cases, and the high-cost instances comprise only 5% of cases. In this situation, the model's accuracy is a poor measure of its performance. For example, while the C5.0 has a very high correctness accuracy at 94%, most of this high accurate prediction comes from better prediction of false instances (specificity= 96%), while the model nearly performs like a random guess when it comes to predicting true cases (sensitivity= 56%). Both the CHAID and NN models yield lower accuracies, but as evident by their high AUC and G-Mean, they predict high-cost instances as good as the low-cost ones, and they exhibit a better trade-off between sensitivity and specificity measures.

5-1-2 Predictor Importance of DT and NN models using 39 input attributes

Classifiers also generate the *predictors importance* calculated from the test population. They basically indicate the relative importance of each input attributes (the predictor) in estimating the model. Since the values are relative, the sum of the values for all attributes is 100 percent. This measure does not relate to model accuracy and it just relates to the importance of each input

attribute in making a prediction. Table 5-4 summarizes the predictor importance for DT and NN models when the target is TOTEXP1-95 with all 39 input attributes.

Table 5-4: Predictor importance for classifiers using all 39 attributes

Rank	C5	C5%	CHAID	CHAID%	NN	NN%
1	IPDIS	38	IPDIS	48	IPDIS	16
2	RXTOT	17	OBTOT	31	OBTOT	15
3	OBTOT	9	RXTOT	12	OPTOT	15
4	OPTOT	7	OPTOT	3	ERTOT	15
5	AGE	3	RTHLTH	1	RXTOT	12
6	POVCAT	3	MNHLTH	1	RTHLTH	2
7	PC	3	POVCAT	1	CHOLCK	1
8	RTHLTH	3	PCCOUNT	1	CHECK	1
9	REGION	2	AGE	<1	PCCOUNT	1
10	ARTHDX	2	BOWEL	<1	MNHLTH	1

All models rank 4 to 5 attributes from the Visits module as their top 5 predictors. This is reasonably expected. Persons with higher number of visits to different levels of health system cause higher expenditure. Confirmation of this fact in our modeling efforts with 39 input attributes verifies the proper functioning of our classifiers and the validity of the final MEPS dataset which represents the US non-institutionalized population correctly.

NN models yield very good performance level for predicting TOTEXP1-95 outcomes, but it only works with continuous attributes. NNs return better AUC and G-Mean results when we test the classifiers on the Visits module only. This classifier doesn't return any results when we run it on other four modules, including Demographics, Health, Preventive, and PrioC, because most of these attributes are categorical and NNs do not properly work with categorical data. Based on these findings, we will not use NN for further classification in this study.

5-1-3 Best classifier among DT models using 39 input attributes

Based on classification results, DTs perform better than NNs when we consider categorical inputs, and the CHAID tree, when compared to C5.0, gives better trade-off for misclassification costs by returning higher AUC and G-Mean results. To compare the CHAID with other important tree classifiers including C&R tree and QUEST algorithm and for a better selection among tree classifiers, we ran an *Auto Classifier* algorithm using 39 input attributes. The Auto Classifier node in IBM Modeler creates and compares a number of different models for binary outcomes, allowing us to choose the best approach for a given analysis. A number of modeling algorithms are supported, making it possible to select the methods we want to use, the specific options for each, and the criteria for comparing the results. The node generates a set of models based on the specified options, and ranks the best candidates according to the criteria. As shown in figure 5-3, the modeling results with Modeler's Auto Classifier node shows that among different trees, the CHAID algorithm returns better trade-off between sensitivity and specificity by returning better G-Mean measure.

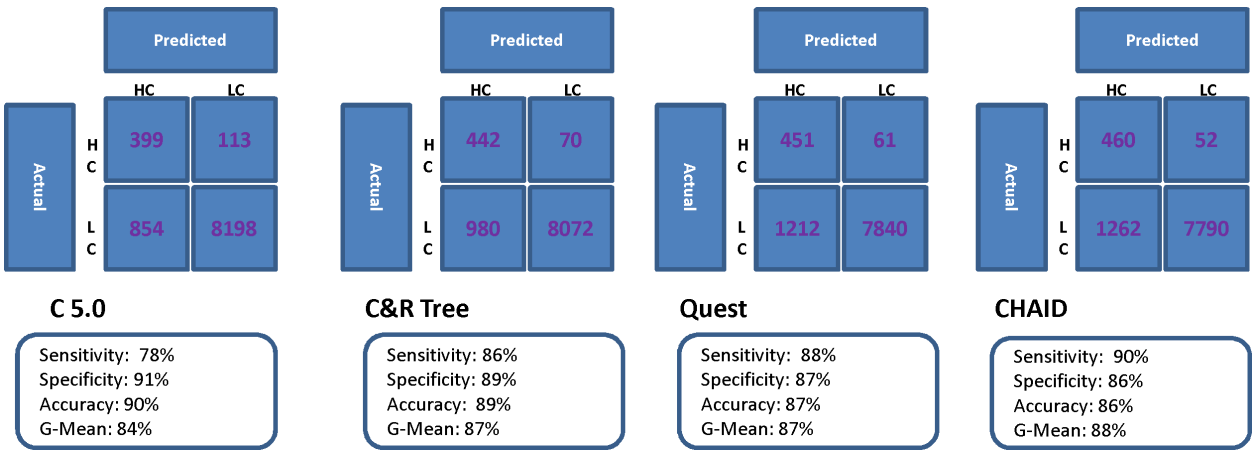

Figure 5-3: Auto Classifier results: Comparison of Tree classifiers (Target: TOTEXP1-95)

5-2 Classification results: DT models using separate modules

We ran the CHAID, a non-binary tree and C5.0, a binary tree, algorithms on each module of attributes in the final MEPS dataset in order to compare their performances. We use the predictor importance results of the better performing model (whether the CHAID or C5.0) in order to find best predictors in each module.

The list of attributes in each module is the same as in Table 5-1, but for PrioC module, PC and PCCOUNT attributes are excluded from this modeling run. This allows us to understand the relative rank of each priority condition in estimating the model. When we remove both PC and PCCOUNT attributes, a C5.0 model returns slightly better accuracy (77.2) compared to including either PC (73.8) or both PC and PCCOUNT (73.3) (Target is TOTEXP1-95).

5-2-1 Performances of DT models using separate modules

A summary of performance measures for the CHAID and C5.0 trees for different modules of attribute is shown in Table 5-5.

Table 5-5: Performance measures for CHAID and C5.0 models on separate modules

Demographics

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	67.0	0.784	56	68	62
Chaid	69.3	0.717	62	70	66

Health

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	75.8	0.772	69	76	72
Chaid	73.8	0.789	72	74	73

Preventive

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	70.6	0.787	65	71	68
Chaid	70.6	0.743	68	71	70

PrioC

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	77.3	0.799	61	78	69
Chaid	75.2	0.772	67	76	71

Visits

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	91.9	0.765	67	93	79
Chaid	88.3	0.944	87	88	87
NN	88.5	0.954	90	88	89

C5.0 has higher correctness accuracy than the CHAID tree for all modules. Major portion of higher accuracies comes from C5.0 higher specificities, while its sensitivities are always below the CHAID rates. The CHAID tree works better in Health and Visits modules compared to C5.0 when we compare them based on their AUC. We don't use AUC measures in other modules (Demographics, Preventive, and PrioC), because for C5.0 models, they are based on raw propensities and are more optimistic as they are calculating both test and training partitions' rates. For all modules, the CHAID tree gives a better G-Mean which is an indication of better trade-off between true positives and true negatives estimations. Collectively, the CHAID algorithm performs better with separate modules compared to tree maker, the C.50. Figure 5-4 summarizes the performance results of the CHAID models on different modules separately, and compares it to its performance on 39 input attributes set (reworded as ALL).

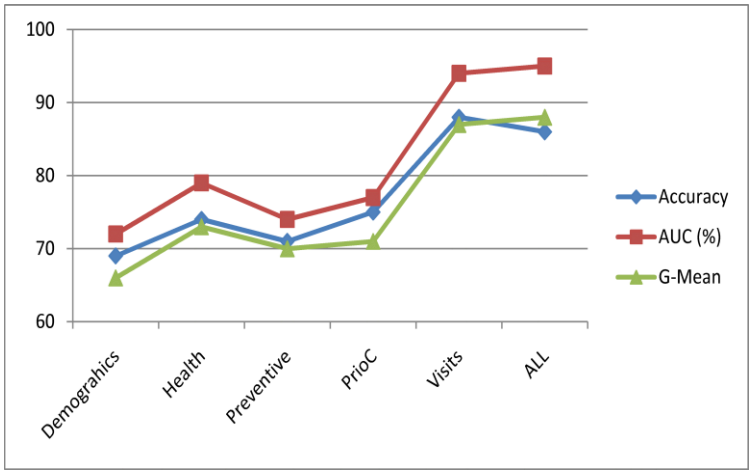

Figure 5-4: Performance results of CHAID on separate modules compared to ALL attributes

5-2-2 Predictor Importance of DT models using separate modules

The CHAID algorithms returns best predictive results among all trees when we build models on separate modules. Table 5-6 shows its predictor importance results on each of the Demographics, Health, Preventive, PrioC and Visits modules when we predict very high-cost patients in the current year i.e. TOTEXP1-95 (Numbers in parenthesis show relative importance in each module). We include predictor importance results for the CHAID algorithm on all 39 attributes for better comparison.

Table 5-6: Predictor importance for CHAID models on separate modules

Rank	Demographics	Health	Preventive	PrioC	Visits	All
1	AGE (69)	ANYLIM (51)	CHOLCK (44)	ARTHDX (27)	IPDIS (49)	IPDIS (48)
2	SEX (13)	RTHLTH (49)	BOWEL (24)	HIBPDX (27)	OBTOT (31)	OBTOT (31)
3	REGION (7)	BMINDX (<1)	BPCHEK (12)	OHRTDX (14)	RXTOT (17)	RXTOT (12)
4	MARRY (6)	MNHLTH (<1)	ASPRIN (8)	DIABDX (13)	OPTOT (3)	OPTOT (3)
5	POVCAT (4)		NOFAT (5)	CANCERDX (12)	ERTOT (1)	RTHLTH (1)
6	RACE (1)		CHECK (5)	CHDDX (3)		MNHLTH (1)
7	HIDEG (<1)		STOOL (2)	CHOLDX (1)		POVCAT (1)
8			EXRCIS (<1)	ASTHDX (1)		PCCOUNT (1)
9			DENTCK (<1)	STRKDX (1)		AGE (<1)
10						BOWEL (<1)

5-3 Classification results: CHAID models using combinations of modules

We ran the CHAID classifier on the combinations of different modules in order to find the minimum set of attributes. We set the Demographics as the base module, and add other modules (i.e. Health, Preventive, PrioC and Visits modules) to it consecutively, and name the resultant combination CHAID models as DHealth, DPreventive, DPrioC, and DVisits, respectively. List of attributes used in each of new models are shown in Figure 5-5.

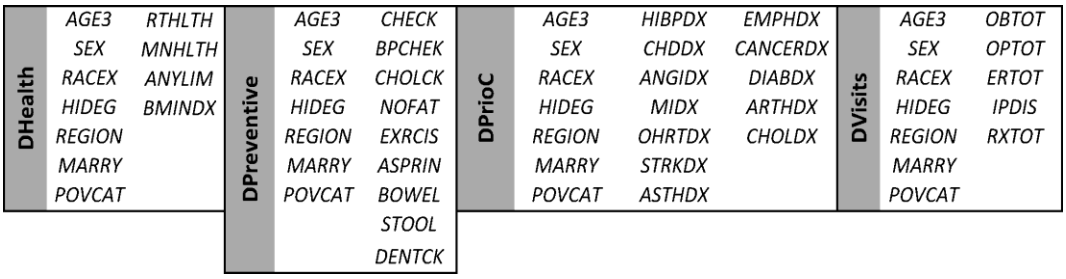

Figure 5-5: List of attributes used in combination models

5-3-1 Performances of CHAID models using combinations of modules

We use the G-Mean and AUC measures, along with correctness accuracy, to compare new combination models' performances with models using separate modules (Table 5-7). As all compared models use the same classifier (i.e. the CHAID algorithm), correctness accuracy can provide good insight into their relative performances. Since we include all attributes from each module in a 2-module combination model, we call them large *set*.

Table 5-7: Performance of the CHAID combination models (large set)

	Accuracy	G-Mean	AUC
DHealth	74	74	0.801
Health	74	73	0.789
DPreventive	69	70	0.763
Preventive	70	70	0.743
DPrioC	71	70	0.768
PrioC	75	71	0.772
Dvisits	88	87	0.944
Visits	88	87	0.944

The base module alone (Demographics) yields G-Mean and AUC equal to 66% and 0.717, respectively. By adding the base module to each of Health, Preventive, PrioC, and Visits modules, their performance remains either approximately the same. The DHealth combination model performs slightly better when compared to the Health module itself. All other possible combinations between different modules were tried, too. For example, we used Health module as the base module and ran the CHAID tree on its combination with other modules i.e. HealthPreventive, HealthPrioC, and so forth. All resultant models perform below the DHealth combination.

Using the predictor importance results in each combination model; we ran the CHAID tree on different sets of input attributes and used the G-mean and correctness accuracy to compare the resultant models. The attributes list in each combination model were reduced gradually in order to reach to the smallest set of attributes that returns better accuracies, G-means, and AUCs compared to original large sets. Figure 5-6 shows the reduced list of attributes in each combination models that give best performances.

DHealth	DPreventive	DPrioC	DVisits
AGE3	AGE3	AGE3	AGE3
SEX	SEX	SEX	SEX
RTHLTH	CHOLCK	REGION	IPDIS
ANYLIM	BOWEL	PCCOUNT	OBTOT

Figure 5-6: Reduced attribute lists in combination models

Each of the new combination models uses only four attributes and we call them as *small set*. AGE and SEX are from the Demographics module, and the two other attributes are from the relevant module. We use the G-Mean and AUC measures, along with correctness accuracy to compare performances of the new combination models with models using large sets (Table 5-8). As all compared models use the same classifier i.e. the CHAID algorithm, correctness accuracy can provide good insight into their relative performances.

Table 5-8: Performance of the CHAID combination models (small set)

	Accuracy	G-Mean	AUC
DHealth	78 (74)*	74 (74)	0.804 (0.801)
DPreventive	69 (69)	71 (70)	0.752 (0.763)
DPrioC	75 (71)	71 (70)	0.784 (0.768)
Dvisits	87 (88)	86 (87)	0.937 (0.944)

*Numbers in parentheses show the performance Measures for the combination models when they use all attributes (large set).

By retaining only four input attributes in any of the combination models (small set), the performance measure is still similar or slightly better than using all attributes (large set). Instead of using 11, 16, 19, and 12 attributes in DHealth, DPreventive, DPrioC, and DVisits models, respectively, we use four attributes in any of these combination models and they still return similar performances. AGE and SEX is shared between all new combination modules. Figure 5-7 shows the ROC curve of the combination models using the small set of attributes.

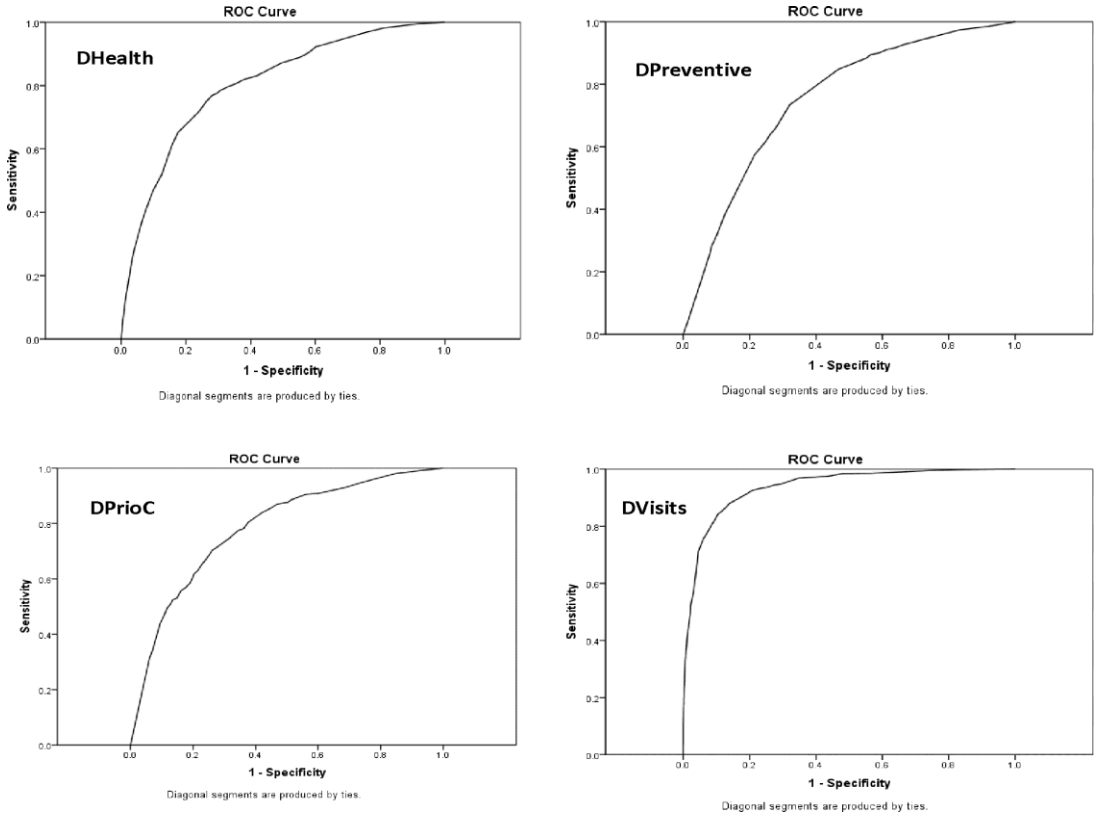

Figure 5-7: ROC Graphs for the combination models (Small set)

In DVisits model, we use IPDIS and OBTOT attributes from the Visits module, along with AGE and SEX from Demographics module. The resultant combination model performs nearly as good as the Visits module itself. It is helpful because we drop three complex input attributes i.e. OPTOT, ERTOT, and RXTOT from the Visits module; replace them with two simple demographics i.e. AGE and SEX; and still achieve similar performances. In DPrioC model, we use PCCOUNT from PrioC module, along with AGE, SEX, and REGION from Demographics modules, and the resultant combination performs slightly better than the PrioC module itself. It is helpful because we drop eleven attributes from the PrioC module; replace them with three simple demographics i.e. AGE and SEX, and REGION; and still achieve comparable performances.

Figure 5-8 summarizes the predictor importance results for the combination models (small set), along with an overall ranking for top 10 attributes according to their relevance to the TOTEXP1-95 target (ALL).

DHealth	
RTHLTH	45
AGE	28
ANYLIM	22
SEX	5

DPrioC	
PCCOUNT	81
AGE	12
SEX	4
REGION	3

DPreventive	
CHOLCK	47
AGE	32
BOWEL	15
SEX	5

DVisits	
IPDIS	47
OBTOT	45
AGE	8
SEX	1

ALL
PCCOUNT
CHOLCK
IPDIS
RTHLTH
OBTOT
AGE
ANYLIM
BOWEL
SEX
REGION

Figure 5-8: Predictor importance results for the combination models (small set)

5-4 Classification results: Best set of predictors for TOTEXP1-95

The CHAID tree models for combination models rank *PCCOUNT, CHOLCK, IPDIS, RTHLTH, OBTOT. AGE, ANYLIM, BOWEL, SEX,* and *REGION* as the top 10 predictors estimating the TOTEXP1-95. We remove two Visits attributes from this set (IPDIS and OBTOT) and then we test different combinations of the remaining 8 attributes until we reach to the best and smallest set of attributes. Table 5-9 shows the performance results for these new models along with the input attributes in each model. We use the G-Mean and AUC measures, along with correctness accuracy to compare their performances. As all compared models use the same classifier i.e. the CHAID algorithm, correctness accuracy can provide good insight into their relative performances.

Table 5-9: The Best 5-10 attributes (TOTEXP1-95)

MODEL (INPUTS)	Accuracy	G-Mean	AUC (%)
AGE, RTHLTH, ANYLIM, CHOLCK, BOWEL, SEX, PCCOUNT, REGION, IPDIS, OBTOT	89	85	0.942
AGE, RTHLTH, ANYLIM, CHOLCK, BOWEL, SEX, PCCOUNT, REGION	78	75	0.816
AGE, RTHLTH, ANYLIM, CHOLCK, BOWEL, SEX	75	76	0.813
AGE, RTHLTH, ANYLIM, CHOLCK, BOWEL	***75***	***76***	***0.812***

The last CHAID classifier that uses the smallest set of attributes including AGE, RTHLTH, ANYLIM, CHOLCK, and BOWEL performs acceptably well, with correctness accuracy, G-mean and AUC equal to75%, 76% and 0.812, respectively. This model performs superior to all single or combination models that use any of Demographics, Health, Preventive, and PrioC attribute sets (Figure 5-9). This set does not use any Visits attributes and is not based on a count of priority conditions as it excludes PCCOUNT. It is an advantage because a count of important medical conditions (PCCOUNT) and an individual's number of visits to different health providers (Visits attributes) are rationally related to the overall health expenditures. We don't need to build a sophisticated data mining model to re-confirm this common belief.

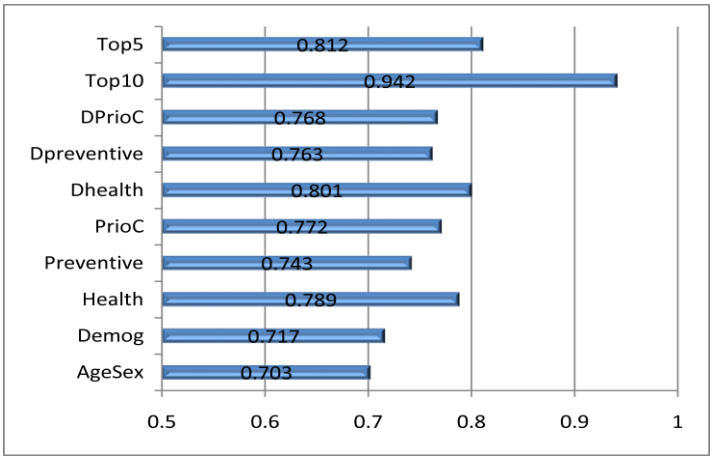

Figure 5-9: AUC measure for CHAID models using top 5-10 attributes, compared to other models

We use Feature Selection node in the modeler to compare its results with the best set of attributes selected in this study. We ran this node twice, once on the primary dataset with 66 attributes and once on the smallest set with 39 attributes. Table 5-10 lists top 10 predictors of feature selection nodes (called FS70 and FS39, respectively), and compare their performance with that of selected top 10 and top 5 sets in this study.

Table 5-10: Comparison of top CHAID models with Feature Selection node results

	FS70	FS39	Top 10	Top 5
Attributes (Ranked)	**IPDIS** **IPNGTD** **RXTOT** **OBTOTV** PCS42 PCCOUNT RTHLTH WLKLIM31 **OPTOTV** **ERTOT**	**IPDIS** **RXTOT** **OBTOTV** PCCOUNT RTHLTH **OPTOTV** **ERTOT** ANYLIM PC ARTHDX	**OBTOT** **IPDIS** PCCOUNT RTHLTH ANYLIM AGE REGION SEX CHOLCK BOWEL	RTHLTH CHOLCK ANYLIM AGE BOWEL
Accuracy	87	87	89	75
G-Mean	87	87	85	76
AUC	0.946	0.945	0.942	0.812

Feature selection lacks expert knowledge and insights from literature, and always returns Visits attributes in its top attributes list. The CHAID models using FS70 and FS39 attribute sets return higher G-mean and AUC compared to the selected Top10 and Top 5 models introduced in this study, but these better results are based on the strength of the Visits attributes. The *Top 10* and *Top 5* models use either two of or none of the Visits attributes and return comparable results.

5-5 Clustering results: K-Means algorithm

We ran K-means algorithm on the final MEPS dataset with 11 attributes including top 10 attributes and TOTEXP1-95. This model doesn't use a target field and is considered an unsupervised learning. All fields are labeled as input for the clustering model. K-Means works by defining a set of starting cluster centers derived from data. It then assigns each record to the cluster to which it is most similar, based on the record's input field values. After all cases have been assigned, the cluster centers are updated to reflect the new set of records assigned to each cluster. The records are then checked again to see whether they should be reassigned to a different cluster, and the record assignment/cluster iteration process continues until either the maximum number of iterations is reached, or the change between one iteration and the next fails to exceed a specified threshold.

When we set cluster numbers to 2 or 3 i.e. 2-cluster and 3-cluster models, the silhouette measure is 0.3 or 0.2, respectively. The resultant models' cluster sizes are shown in Figure 5-10.

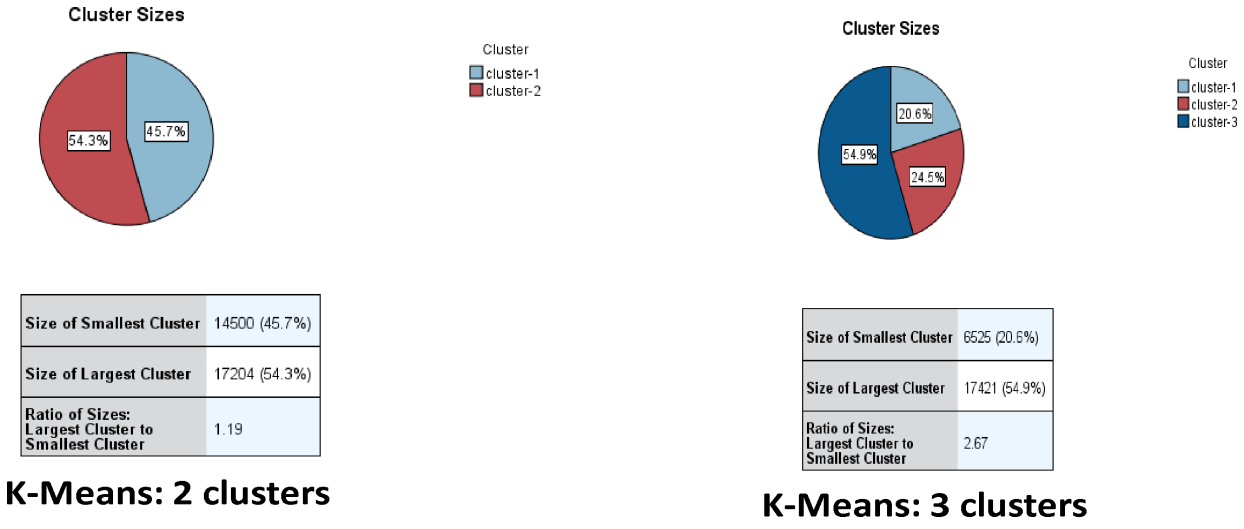

Figure 5-10: K-Means Cluster sizes: Top 10 attributes featuring TOTEXP1-95

A summary of cluster details for K-means models is shown in Figure 5-11. In 2-cluster model, the SEX attribute is the best clustering attribute; all males go to cluster 1 and all females go to cluster 2. Females are more aged, check their blood cholesterol more frequently, live slightly more with some type of limitations, and are labeled more as a high-cost individual. Remaining attributes don't confer a visible difference between females and males clusters. Same is evident from a frequency analysis of major attributes in two sexes, as shown in SAS output in table 5-11.

Table 5-11: Frequency of top input attributes by SEX

SEX	AGE				TOTEXP1-95			RTHLTH					
	1	2	3	Total	1	2	Total	1	2	3	4	5	Total
1	28	12	6	46	1.95	43.79	45.74	9	18	13	5	1	46
2	33	13	8	54	3.06	51.21	54.26	9	20	17	7	2	54
Total	61	25	14	100	5.00	95.00	100	18	38	29	11	3	100
SEX	CHOLCK				ANYLIM			PCCOUNT					
	0	1	2	Total	1	2	Total	0	1	2	3	4	Total
1	9	22	11	46	12	34	46	21	10	6	4	5	46
2	9	31	12	54	16	38	54	24	12	8	5	6	54
Total	18	53	23	100	28	71	100	45	21	14	9	11	100

K-means: 2 Clusters

C- 2	C- 1
SEX 2 (100.0%)	SEX 1 (100.0%)
OBTOTV 0.02	OBTOTV 0.02
CHOLCK53 1 (60.1%)	CHOLCK53 1 (52.4%)
RTHLTH 2 (37.3%)	RTHLTH 2 (39.1%)
IPDIS 0.01	IPDIS 0.01
ANYLIM 2 (70.4%)	ANYLIM 2 (74.3%)
PCCOUNT 0 (43.7%)	PCCOUNT 0 (46.9%)
TOTEXP1-.95 2 (94.4%)	TOTEXP1-.95 2 (95.7%)
AGE31X 1 (60.8%)	AGE31X 1 (61.9%)
REGION 3 (38.4%)	REGION 3 (37.4%)
BOWEL53 2 (75.1%)	BOWEL53 2 (76.0%)

K-means: 3 Clusters

C- 3	C- 2	C- 1
AGE31X 1 (95.4%)	AGE31X 2 (37.6%)	AGE31X 2 (66.0%)
ANYLIM 2 (94.1%)	ANYLIM 1 (100.0%)	ANYLIM 2 (100.0%)
BOWEL53 2 (98.4%)	BOWEL53 2 (54.4%)	BOWEL53 1 (60.8%)
CHOLCK53 1 (35.1%)	CHOLCK53 1 (81.1%)	CHOLCK53 1 (82.7%)
OBTOTV 0.01	OBTOTV 0.04	OBTOTV 0.02
PCCOUNT 0 (71.8%)	PCCOUNT 4 (31.6%)	PCCOUNT 1 (30.2%)
RTHLTH 2 (45.6%)	RTHLTH 3 (35.9%)	RTHLTH 2 (41.6%)
TOTEXP1-.95 2 (98.9%)	TOTEXP1-.95 2 (85.9%)	TOTEXP1-.95 2 (95.4%)
IPDIS 0.01	IPDIS 0.03	IPDIS 0.01
REGION 3 (36.8%)	REGION 3 (40.6%)	REGION 3 (37.8%)
SEX 2 (52.3%)	SEX 2 (58.8%)	SEX 2 (54.2%)

Figure 5-11: Predictor ranking and cluster details for 2- and 3-cluster K-means model

In 3-cluster model, the AGE attribute is a major predictor. The 3rd cluster is mostly from younger adults. Over 95% of all members in this cluster age 18 to 49. A dominant portion of individuals who have no limitations, have had no history of priority conditions, limitations, bowel instrumentation, and blood cholesterol check are segmented in this cluster. Their overall health perception is high and very few members belong to high-cost individuals.

5-6 Recommendations Derived from this Study

This study introduces the MSA for predicting top 5% high-cost group among general population who may or may not have any current medical conditions. Without relying on well- known and

trivial measures of health costs, including counts of visits to different care providers, the MSA includes non-trivial and easy-to-survey measures for self-perception of health, age, along with two preventive health indicators (history of blood cholesterol checks, and colonic preventive interventions) and presence or absence of any physical, sensory, or cognitive limitations.

Health planners can use our proposed MSA to design local, regional, or national surveys for tracking high-cost groups in general population, far before they seek medical helps. It will give the policy makers enough time to better plan various disease-management programs for high-risk groups, and will give them enough resources to tailor those programs to more localized demographic targets. The information from a representative sample of a high-risk population in a specific geographic area can be feed into a trained CHAID tree classifier in order to identify potential high-cost group among them, and to enroll the tracked group in suitable preventive or disease-control programs to closely monitor them for complications, recurrences, or exacerbations. Then the planners can better allocate available resources to potentially high-risk groups and may spend the healthcare budget in a more targeted manner.

At the national level, the results can help federal and provincial health authorities to devise similar but smaller panel surveys in order to track health expenditures at the macro level. When combined with data on the sources of healthcare budget and data on the spenders of the healthcare budget, a well-conducted survey can portray existing variations in geographical health spending by providing extensive information regarding demographical differences of potential high-cost groups compared to other parts of the population. Multiple years survey results will enable governmental bodies to better track health expenditure variations across a jurisdiction and to better evaluate the impacts of policy changes in a period of time.

Insurance companies always seek more efficient ways to spend their limited earnings from health premiums, and try to enroll their clients into the most suitable health protection plans. For this purpose, they use a few predictive models to estimate future health costs for each potential enrollee, mostly based on the current clinical status of the individual. Having accurate clinical information, insurance providers identify high-risk groups and predict high-cost cases. Based on a preliminary clinical interview, they screen if an individual have had a diagnosis for a major cost-bearing disease including diabetes or heart disease, and then they compare patient's clinical status with typical high-cost profiles in that diagnostic category, in order to estimate the risk of an individual to be in high-cost group. Having access to the accurate medical profile of a person is the mainstay for this cost-estimation process, but this is not always feasible. As such, use of predictive models relying on non-clinical data seems inevitable.

The MSA from this research can help insurers to identify potential high-cost or very high-cost individuals among a pool of potential clients, based on non-trivial and non-clinical data. The proposed attributes are easy to collect, and they can be used in a short screening questionnaire which may be employed as a preliminary screening tool to find the potential high-cost cases among the general population. To get more efficient results, the insurer may primarily decrease the confidence threshold for interpreting the screening survey results to identify a very high-cost person; i.e. they may extract 10% of a specific population as a potential pool for the top 5% costs, or they may set a lower threshold for G-mean or AUC of the predictive model in order to extend the boundaries of high-cost group. After narrowing the potential high-costs to 10% population size, extra runs of predictive modeling on the remaining pool, or further clinical interview can be conducted to better identify those with the highest possibility of becoming high-cost in the future.

Through implementing the predictive models from this study along with the MSA, policy makers and service providers can examine the relevance of the findings to real field practice conditions and provisions. Further customization in modeling provisions and alterations in the MSA itself may become necessary when researchers try to apply cost-modeling tools to various health services delivery settings.

Chapter 6 : Conclusions

In this chapter, we conclude the study in three parts: summary of the thesis; answers to the research questions; and research limitations and suggestions for future works.

6-1 Summary of the Thesis

The current thesis is a quantitative research study designed to apply different supervised and unsupervised machine learning algorithms on a valid dataset, the MEPS, which is a nationally representative medical expenditures survey for the non-institutionalized US population. The target attribute is either current year's or next year's total health expenditure, which are set at three different thresholds for identifying high-cost individuals i.e. 95, 90, and 80 percentiles. We test six targets, three for current year and three for next year. Two components of the MEPS database i.e. the HC (household component) and the MC (medical conditions component) for a three-year period 2006-2008 have been used to build the final MEPS dataset with 66 attributes and 31704 records. By finishing all preparation efforts, the attribute counts has been reduced to 39 from its original 66, by using results of correlation studies, medical expertise, and relevance to the targets. The final MEPS dataset now includes 7, 4, 9, 14, and 5 input attributes from Demographics, Health status, Preventive care, Priority conditions, and Visits counts modules, respectively.

Supervised modeling (classification) of the final MEPS dataset, containing all 39 attributes from five modules shows that the classifiers (NNs or DTs) predict *current year* costs more accurate compared to next year costs; and predict better when the high-cost threshold is set at its highest threshold (95 percentile, to predict top 5 percentile of costs) when compared to other thresholds for high-cost (90 and 80 percentiles, to predict top 10 and top 20 percentile of costs). NNs return

better AUCs, followed by the CHAID tree, and C5.0 algorithm. Higher AUCs for NNs using all 39 attributes verifies proper functioning of DM models and validity of the final MEPS dataset. When we remove Visits attributes from the final MEPS dataset, NNs fail to return any results because all other attributes are categorical. Both NNs and DTs rank 4 or 5 attributes from Visits module at the top of their predictor importance results which shows the overshadowing of other 34 attributes by this module's attributes. G-mean and AUC metrics give better trade-off between sensitivity and specificity of a classifier, compared to correctness accuracy.

Supervised modeling (classification) of the separate modules (Demographics, Health, Preventive, PrioC, and Visits) by DTs shows that the CHAID tree returns better G-mean compared to the C5.0 algorithms. A run of Auto-Classifier algorithm of the modeler shows that for the final MEPS dataset, the CHAID tree is the best classifier, returning higher AUCs. The CHAID tree better discriminates the true positives from the true negatives. It, therefore, reduces the misclassification rates through the trade-off between models' sensitivity and specificity. In the final MEPS dataset, only 5% of cases belong to true positive group. As such, the non-binary CHAID tree model that reasonably deals with outliers and reduces misclassification rate, performs better than C5.0 decision tree.

Supervised modeling (classification) of different combinations of modules by the CHAID tree is used to reach to the minimal set of attributes (MSA) (containing 5 to 8 attributes from any of Demographics, Health, Preventive, and PrioC modules). We intentionally omit any combination with the Visits module. The Visits module has a strong correlation with the target attribute, and this relevance is a given fact. When we run the Feature Selection algorithm, it confirms this relevance and ranks the Visits attributes at the top 3 or 4 attributes. Our directed approach helps to omit these attributes and introduce the MSA that contains no Visits predictor.

The results from this study can help health planners to design efficient surveys at local, regional, or national levels, and track potential high-cost groups among general population. They can use their customized survey results to better allocate scarce health resources to various disease-management programs and other health initiatives. At a macro level, provincial and federal health authorities can use similar insights to portray important demographical variations in health expenditures over time, and over different populations. Finally, insurance providers can use the MSA from this study along with its validated models, to screen and profile high-risk groups according to their future health costs, and to tailor their insurance plans according to varying needs of different clients.

6-2 Research Objectives/ Research Questions

This study successfully achieves its all three objectives, which includes building different predictive models, including DT and ANN to classify top 5% high-cost population; comparing the resultant models based on various performance measures; and discovering the minimal set of attributes (MSA) that can predict the high-cost population.

Successful predictive models' building and evaluation: CHAID Tree as the best classifier.

With respect to the first research question, we conclude that we can build various predictive models to estimate high-cost patients, and compare their performances by a combination of Correctness Accuracy, G-mean, and Area under the ROC Curve metrics. We conclude that the CHAID algorithm works more accurate with the MEPS dataset. It returns higher G-mean and AUC values compared to other classifiers including DTs and NNs. In the final MEPS dataset, when the high-cost threshold is set at 95 percentile (to predict top 5 percentile of costs) in current year, the true cases comprise 5% of the population. A G-mean and AUC measure provide a

better trade-off between models' sensitivity and specificity and is superior to correctness accuracy.

Specifying the minimal set of attributes excluding "visit counts" as the trivial attribute

With respect to the second research question, we recommend the following 8 attributes as the MSA in a CHAID tree to predict *very high-cost* percentile (top 5 percentile of costs) instances among the general population:

1- PCCOUNT: Count of priority conditions diagnosed in an individual

2- ***RTHLTH*****:** Perceived health status

3- ***ANYLIM*****:** Presence of any limitation (physical, sensory, or cognitive) in individual

4- ***CHOLCK*****:** Time elapsed since last blood cholesterol check

5- ***BOWEL*****:** Time elapsed since last sigmoidoscopy/colonoscopy

6- ***AGE:*** Age of the individual in years

7- Region: The US Census region of the individual

8- Sex: Sex of the individual

If we remove three of these predictors and keep five remaining attributes, the AUC drops only by 0.04 degree. Further removal of any of five remaining attributes reduces the performances remarkably. The minimal set of attributes is, therefore, ranked as follows:

1- ***RTHLTH*****:** Perceived health status

2- ***CHOLCK*****:** Time elapsed since last blood cholesterol check

3- ***ANYLIM*****:** Presence of any limitation (physical, sensory, or cognitive) in individual

4- ***AGE:*** Age of the individual in years

5- ***BOWEL*****:** Time elapsed since last sigmoidoscopy/colonoscopy

It is worthy to note that this minimum set of attributes does not include counts of visits to care providers. This is a well-known attribute that has high correlation with the expenditure, and does not offer a new insight to the data (i.e. it is a trivial predictor). We would like to predict high-cost patients before they are hospitalized or without knowing how many times the patient was visited by care providers. Consequently, the results from this study are useful for policy makers, health planners, and insurers to plan and improve delivery of health services.

The minimal set of attributes includes non-trivial and easy-to-survey measures for self-perception of health, age, along with two preventive health indicators (history of blood cholesterol checks, and colonic preventive interventions) and presence or absence of any physical, sensory, or cognitive limitations

6-3 Study Limitations

The MEPS database is a unique health expenditure data that stores medical spending data for the non-institutionalized US population detailed by the sources of payments and destinations of health dollars. Its extensiveness is limited by the categorical nature of data it stores, and limits application of strong DM algorithms which better work with continuous data. Normalization techniques can reshape the categorical data to scale type, but may leave the results less interpretable.

6-4 Future works

The current research introduces the top predictors estimating an individual's current year's total health expenditures by classifying her in high- or low-cost population by applying the CHAID algorithms on the MEPS dataset. Future research may estimate the likelihood of hospital admission for an individual by using the same dataset, or by using new input attributes based on

the model's requirement and literature insights. To get more insight into the MEPS data and its potential predictors, researchers may run multiple clustering algorithms on entire MEPS dataset for a single panel. Based on revealed patterns, a researcher may test different set of attributes and test their strength in estimating cost classes including cost-related targets as hospital admissions.

References

Ang Q.; Wang W.; Zhao B.; Li J.; Li K.; (2010), "Application of data mining based on clinical medicine database," *Signal Processing Systems (ICSPS), 2nd International Conference on* , vol.3, no., pp.V3-719-V3-723.

Bellazzi R., Zupan B.,(2008).Predictive data mining in clinical medicine: Current issues and guidelines, International Journal of Medical Informatics, V. 77-2, PP. 81-97.

Bereznicki, B. J., Peterson, G. M., Jackson, S. L., Walters, E. H., Fitzmaurice, K. D., & Gee, P. R. (2008). Data-mining of medication records to improve asthma management. *Medical Journal of Australia, 189*(1), 21-25.

Boytcheva, S. (2011). Shallow medication extraction from hospital patient records. *Studies in Health Technology and Informatics, 166*, 119-128.

Buckeridge, D. L., Burkom, H., Campbell, M., Hogan, W. R., & Moore, A. W. (2005). Algorithms for rapid outbreak detection: A research synthesis. *Journal of Biomedical Informatics, 38*(2), 99-113.

Canadian Institute for Health Information. National Health Expenditure Trends, 1975 to 2010. Ottawa, ON, CAN: Canadian Institute for Health Information, 2010.

Centers for Disease Control and Prevention (CDC). (2009). The power of prevention: Chronic disease…the public health challenge of the 21st century. Available from: http://www.cdc.gov/chronicdisease/pdf/2009-Power-of-Prevention.pdf

Centers for Medicare and Medicaid Services (CMS) (2011). National Health Expenditure Data. Accessed in 25 September 2011 at: https://www.cms.gov/NationalHealthExpendData/02_NationalHealthAccountsHistorical.asp#TopOfPage

Chen, Hsinchun ; Medical Informatics : Knowledge Management and Data Mining in Biomedicine. Boston, MA, USA: Kluwer Academic Publishers, 2005.

Cohen, J. W., Cohen, S. B., & Banthin, J. S. (2009). The medical expenditure panel survey: A national information resource to support healthcare cost research and inform policy and practice. *Medical Care, 47*(7 Suppl 1), S44-50.

Cohen, S. B. (2002). The medical expenditure panel survey: An overview. *Effective Clinical Practice : ECP, 5*(3 Suppl.

Concaro, S., Sacchi, L., Cerra, C., Stefanelli, M., Fratino, P., & Bellazzi, R. (2009). Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment. *AMIA ...Annual Symposium Proceedings / AMIA Symposium.AMIA Symposium, 2009*, 119-123.

DeVol R, Bedroussian A, (2007). An unhealthy America: The economic burden of chronic disease. Prepared for the Milken Institute. Available from: http://www.milkeninstitute.org/publications/publications.taf?function=detail&ID=38801018&cat=ResRep

Eggers, K. M., Ellenius, J., Dellborg, M., Groth, T., Oldgren, J., Swahn, E., et al. (2007). Artificial neural network algorithms for early diagnosis of acute myocardial infarction and prediction of infarct size in chest pain patients. *International Journal of Cardiology, 114*(3), 366-374.

Farley, J. F., Harley, C. R., & Devine, J. W. (2006). A comparison of comorbidity measurements to predict healthcare expenditures. *American Journal of Managed Care, 12*(2), 110-117.

Fauci AS, Braunwald E, Kasper DL, Hauser SL, Longo DL, JL Jameson, Loscaizo JL (eds), (2008). *Harrison's Principles of Internal Medicine, 17e*. New York, McGraw-Hill, 2008. URL address of the online site (http://www.accessmedicine.com).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-53.

Fleishman, J. A., & Cohen, J. W. (2010). Using information on clinical conditions to Gregori, D., Petrinco, M., Barbati, G., Bo, S., Desideri, A., Zanetti, R., et al. (2009). Extreme regression models for characterizing high-cost patients. *Journal of Evaluation in Clinical Practice, 15*(1), 164-171.

Ghosh, D., & Guha, R. (2011). Using a neural network for mining interpretable relationships of west nile risk factors. *Social Science & Medicine (1982), 72*(3), 418-429.

Govan, L., Wu, O., Briggs, A., Colhoun, H. M., McKnight, J. A., Morris, A. D., et al. (2011). Inpatient costs for people with type 1 and type 2 diabetes in scotland: A study from the scottish diabetes research network epidemiology group. *Diabetologia, 54*(8), 2000-2008.

Grassi, M., Villani, S., & Marinoni, A. (2001). Classification methods for the identification of 'case' in epidemiological diagnosis of asthma. *European Journal of Epidemiology, 17*(1), 19-29.

Han, J.; Kamber, M.; and Pei J.; (2012) 1 - Introduction, Data Mining (Third Edition), Morgan Kaufmann, Boston, 2012, Pages 1-38.

Haraldsson, H., Edenbrandt, L., & Ohlsson, M. (2004). Detecting acute myocardial infarction in the 12-lead ECG using hermite expansions and neural networks. *Artificial Intelligence in Medicine, 32*(2), 127-136.

Hirsch S., Shapiro JL, Turega MA, Frank TL, Mcl Niven R., Frank PI; (2001). Using a Neural Network to Screen a Population for Asthma. Annals of epidemiology. V.11- 6 PP. 369-376.

Hirsch, S., Frank, T. L., Hazell, M., & Frank, P. I. (2004). Screening for asthma by population ranking: A validation study. *Annals of Epidemiology, 15*(1), 64-70.

Isken, M. W., & Rajagopalan, B. (2002). Data mining to support simulation modeling of patient flow in hospitals. *Journal of Medical Systems, 26*(2), 179-197.

Khan, S., Rappaport, H. M., & Magoun, A. D. (2006). Predicting future high-cost asthma patients (recipients). *Journal of Pharmaceutical Finance, Economics and Policy, 14*(4), 43-52.

Kotel'nikova, E. V., Gridnev, V. I., Dobgalevskiĭ, P. I., & Bespiatov, A. B. (2004). Prognostication of coronary atherosclerosis for selection of tactics of management of patients with ischemic heart disease. [Prognozirovanie koronarnogo ateroskleroza dlia vybora taktiki vedeniia bol'nykh ishemicheskoi bolezn'iu serdtsa.v ambulatornoi praktike.] *Kardiologiia, 44*(3), 15-19.

Kubat, M, Holte R.C., Matwin S,. (1997). Learing when Negative Examples Abound: One-sided selection. Proceedings of the Ninth European Conference on Machine Learning.

Kwok-Leung T., Wenchi C., Gierlich P., Goldsman D., Xuyuan L., and Maschek T.; (2008). A Review of Healthcare, Public health, and Syndromic Surveillance. *Quality Engineering*, 20 (4), 435-450.

Lavrač N., (1999). Selected techniques for data mining in medicine, Artificial Intelligence in Medicine, Volume 16, Issue 1, May 1999, Pages 3-23.

Lee, C. -., Chen, J. C. -., & Tseng, V. S. (2011). A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring. *Computer Methods and Programs in Biomedicine,101*(1), 44-61.

Merriam and Webster Dictionary, Retrieved from http://www.merriam-webster.com/dictionary/data%20mining in September 2011.

Mirolla, Michael. The Cost of Chronic Disease in Canada. Glen Haven, NS, Canada: GPI Atlantic, 2004.

Mougiakakou, S. G., Bartsocas, C. S., Bozas, E., Chaniotakis, N., Iliopoulou, D., Kouris, I., et al. (2010). SMARTDIAB: A communication and information technology approach for the intelligent monitoring, management and follow-up of type 1 diabetes patients. *IEEE Transactions on Information Technology in Biomedicine, 14*(3), 622-633.

Moturu, S. T., Johnson, W. G., & Liu, H. (2007). Predicting future high-cost patients: A real-world risk modeling application. Paper presented at the *Proceedings - 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007,* 202-208.

Moturu, S. T., Johnson, W. G., & Liu, H. (2010). Predictive risk modeling for forecasting high-cost patients: A real-world application using Medicaid data. *International Journal of Biomedical Engineering and Technology,3*(1-2), 114-132.

Muhammad, A., Malagore, I. A., & Afsar, F. A. (2010). Automatic detection and localization of myocardial infarction using back propagation neural networks. Paper presented at the *2010 4th International Conference on Bioinformatics and Biomedical Engineering, iCBBE 2010.*

Nguyen, H. T., & Jones, T. W. (2010). Detection of nocturnal hypoglycemic episodes using EEG signals. *Conference Proceedings : ...Annual International Conference of the IEEE Engineering in Medicine and Biology Society.IEEE Engineering in Medicine and Biology Society.Conference, 2010*, 4930-4933.

P.T., A. S., Joseph, P. K., & Jacob, J. (2011). Automated diagnosis of diabetes using heart rate variability signals. *Journal of Medical Systems,* , 1-7.

Patra J, Popova S, Rehm J, Bondy S, Flint R, & Giesbrecht N; (2007). Economic cost of chronic disease in Canada 1995-2003. Prepared for the Ontario Chronic Disease Prevention Alliance and the Ontario Public Health Association.

Rietveld, S., Oud, M., & Dooijes, E. H. (1999). Classification of asthmatic breath sounds: Preliminary results of the classifying capacity of human examiners versus artificial neural networks. *Computers and Biomedical Research, 32*(5), 440-448.

Rivera Farina, P., Pérez Turiel, J., González, L., González Sarmiento, E., Herreros, A., & Higuero, S. (2009). Neural network application to the development of a novel diabetic neuropathy diagnosis tool using the valsalva index and the SCR. Paper presented at the *Final Program and Abstract Book - 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009.*

Sabouri, S., SadAbadi, H., & Dabanloo, N. J. (2010). Neural network classification of body surface potential contour map to detect myocardial infarction location. Paper presented at the *Computers in Cardiology, , 37*301-304.

Sanders, D. L., & Aronsky, D. (2006). Detecting asthma exacerbations in a pediatric emergency department using a bayesian network. *AMIA ...Annual Symposium Proceedings / AMIA Symposium.AMIA Symposium,* , 684-688.

Selecky, C. E. (2008). Disease management of chronic obstructive pulmonary disease from a disease management organization perspective: Providing technology and time to address gaps in care. *Disease Management and Health Outcomes, 16*(5), 319-325.

Soreide K. (2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. J. Clin Pathol, 2009; 62:1 1-5.

Teow, K. L., El-Darzi, E., Foo, C., Jin, X., & Sim, J. (2011). Intelligent analysis of acute bed overflow in a tertiary hospital in singapore. *Journal of Medical Systems,* , 1-10.

Weinstein, L., Radano, T. A., Jack, T., Kalina, P., & Eberhardt 3rd., J. S. (2009). Application of multivariate probabilistic (bayesian) networks to substance use disorder risk stratification and cost estimation. *Perspectives in Health Information Management / AHIMA, American Health Information Management Association, 6* Retrieved from www.scopus.com

Yang, C., Street, W. N., Lu, D. -., & Lanning, L. (2010). A data mining approach to MPGN type II renal survival analysis. Paper presented at the *IHI'10 - Proceedings of the 1st ACM International Health Informatics Symposium,* 454-458.

Yildirim, P., Çeken, C., Hassanpour, R., Esmelioglu, S., & Tolun, M. R. (2011). Mining MEDLINE for the treatment of osteoporosis. *Journal of Medical Systems,* , 1-9.

Zhao CM, Luan J.; (2006). Data Mining: Going Beyound Traditional Statistics. New Directions for Institutional Research 131. Wiley Online Library; Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/ir.184/pdf

Appendices

Appendix I: Sample Decision Tree diagram for a CHAID tree made on combination of demographics and health status attributes

Appendix II : Partial Rule sets for a CHAID tree made on combination of demographics and health status attributes

```
RTHLTH <= 1 [ Mode: 2 ]
        AGE31X = 1 [ Mode: 2 ]
                SEX = 2 [ Mode: 2 ] => 2.0
                SEX = 1 [ Mode: 2 ] => 2.0
        AGE31X = 2 [ Mode: 2 ]
                SEX = 2 [ Mode: 2 ] => 2.0
                SEX = 1 [ Mode: 2 ] => 2.0
        AGE31X = 3 [ Mode: 2 ] => 2.0
RTHLTH > 1 and RTHLTH <= 2 [ Mode: 2 ]
        AGE31X = 1 [ Mode: 2 ]
                ANYLIM = 2 or ANYLIM IS MISSING [ Mode: 2 ]
                        SEX = 2 [ Mode: 2 ] => 2.0
                        SEX = 1 [ Mode: 2 ] => 2.0
                ANYLIM = 1 [ Mode: 2 ] => 2.0
        AGE31X = 2 [ Mode: 2 ]
                ANYLIM = 2 or ANYLIM IS MISSING [ Mode: 2 ]
                        SEX = 2 [ Mode: 2 ] => 2.0
                        SEX = 1 [ Mode: 2 ] => 2.0
                ANYLIM = 1 [ Mode: 1 ] => 1.0
        AGE31X = 3 [ Mode: 1 ]
                ANYLIM = 2 [ Mode: 2 ] => 2.0
                ANYLIM = 1 or ANYLIM IS MISSING [ Mode: 1 ] => 1.0
RTHLTH > 2 and RTHLTH <= 3 [ Mode: 1 ]
        ANYLIM = 2 or ANYLIM IS MISSING [ Mode: 2 ]
                AGE31X = 1 [ Mode: 2 ]
                        SEX = 2 [ Mode: 2 ] => 2.0
                        SEX = 1 [ Mode: 2 ] => 2.0
                AGE31X = 2 or AGE31X = 3 [ Mode: 2 ] => 2.0
        ANYLIM = 1 [ Mode: 1 ]
                AGE31X = 1 [ Mode: 1 ]
                        SEX = 2 [ Mode: 1 ] => 1.0
                        SEX = 1 [ Mode: 2 ] => 2.0
                AGE31X = 2 or AGE31X = 3 [ Mode: 1 ]
                        SEX = 2 [ Mode: 1 ] => 1.0
                        SEX = 1 [ Mode: 1 ]
                                AGE31X = 2 [ Mode: 1 ] => 1.0
                                AGE31X = 3 [ Mode: 1 ] => 1.0
RTHLTH > 3 and RTHLTH <= 4 [ Mode: 1 ]
        ANYLIM = 2 [ Mode: 1 ]
                AGE31X = 1 [ Mode: 2 ] => 2.0
                AGE31X = 2 or AGE31X = 3 [ Mode: 1 ] => 1.0
```

ANYLIM = 1 or ANYLIM IS MISSING [Mode: 1]
AGE31X = 1 [Mode: 1] => 1.0
AGE31X = 2 [Mode: 1] => 1.0
AGE31X = 3 [Mode: 1] => 1.0
RTHLTH > 4 [Mode: 1]
AGE31X = 1 [Mode: 1] => 1.0
AGE31X = 2 [Mode: 1]
SEX = 2 [Mode: 1] => 1.0
SEX = 1 [Mode: 1] => 1.0
AGE31X = 3 [Mode: 1] => 1.0