DESIGN OF A REFERRAL AND MANAGEMENT TOOL FOR PATIENTS WITH COMORBIDITIES

DESIGNING A MANAGEMENT AND REFERRAL TOOL FOR PATIENTS WITH MULTIPLE CHRONIC ILLNESSES IN PRIMARY CARE SETTINGS

By FLAVIEN OWOLABI, B.A., MSC.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2016) Hamilton, Ontario (Health Informatics)

TITLE: Designing a Management and Referral Tool for Patients with Multiple Chronic Illnesses in Primary Care Settings

AUTHOR: Flavien Owolabi, B.A., MSC. (SUNY-ESF/Syracuse University)

SUPERVISOR: Dr. Ann McKibbon

COMMITTEE MEMBERS: Dr. Norm P. Archer, Dr. Cynthia Lokker

NUMBER OF PAGES: xii, 63

Abstract

Some local health organizations in Ontario (e.g., Local Health Integration Network or LHINs) have put forward a strategic objective to identify patients with preventable high cost healthcare service usage (e.g., hospitalizations, emergency department [ED] visits). To attain this goal, primary care service providers, who are considered the entry point to the health system, need tools to help diagnose, treat and refer those patients identified as being potential high users of the health care system.

The goal of this study was to develop a management and referral tool to identify, manage and refer patients living with multiple comorbidities to specialized care teams such as Health Links.

Data used in this analysis were obtained from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) primary care data holdings. The dataset created for this study contained 14,004 patient records.

Data analysis techniques included use of both statistical and predictive analytic tools. The base models included four data mining classification algorithms: Decision Tree, Naïve Bayes, Neural Network and Clustering. The predictive modeling approach was complemented by an association analysis.

The one-way ANOVA analysis indicated that age and health status (number of conditions, and individual medical conditions) identified statistically significant differences in patient utilization of health services.

Results from the predictive analytics showed that patient age and patient medical conditions, as well as number of medical conditions for each patient (5 or more) could be used as criteria to develop tools (e.g. searches, reminders). Specifically, Parkinson disease, dementia and epilepsy were found to be important predictors (i.e. most frequently associated with) the top 4 most prevalent conditions (hypertension, osteoarthritis, depression and diabetes) within the population of the study. The association analysis also revealed that chronic obstructive pulmonary disease (COPD) was closely associated with the top 4 most prevalent conditions. Based on the findings of this study, Parkinson Disease, dementia, epilepsy and COPD can be used to identify patients with complex medical needs who are likely to be high users of the healthcare system and to be considered for early, personalized intervention.

Acknowledgments

This research work would not have been possible without the invaluable assistance and support of many individuals who directly or indirectly contributed to its successful completion.

I would like to express my profound gratitude to Dr. Karim Keshavjee whose assistance helped me navigate the various cycles required to obtain access to CPCPSSN data holdings key ingredients for this research. In the same vein, I would to extend my gratitude to all the authorities at CPCSSN who approved my access to the CPCSSN data and to Ken Martin who helped create the research database.

I am also very grateful to my Supervisor Dr. Ann McKibbon, to Dr. Norm Archer and Dr. Cynthia Lokker who provided me with insightful guidance and assistance in very difficult moments where the thesis project was in jeopardy. Your adherence to academic high standards and your critical thinking had helped improve the readability of the text as well as other materials presented (e.g. tables).

I would like to extend my gratitude to Iris Kehler whose assistance and advice helped me chart various waves of troubles by providing me with practical solutions to issues at hand.

Dedication

To God,
Who has always been my help,
Who gave me knowledge and understanding,
Who filled my mind with unmatched inspirations to face all kind of adversities,
Who trained my hands for battle.
And who brought me victory,
Glory and honor be given for all ages

Contents

List of Tables and Figures

List of Abbreviations

ARRA: American Recovery and Reinvestment Act

CIHI: Canadian Institute for Health Information

CMG+: Case Mix Group Plus

COPD: Chronic obstructive pulmonary disease

CPCSSN: Canadian Primary Care Sentinel Surveillance Network

eCE: eHealth Center of Excellence

ED: Emergency Department

EMR: Electronic Medical Record Systems

HIT: Health Information Technology

HITECH: Health Information Technology for Economic and Clinical Health Act

HSD: Honest Significant Differences

NCD: Number of Chronic Diseases

NOE: Number of Encounters

NOEx: Number of Exams

NOL: Number of Labs

LHIN: Local Health Integration Network

MOHLTC: Ministry of Health and Long-Term care

PCP: Primary Care Provider

PHAC: Public Health Agency of Canada

PSS EMR: Practice Solution Suite EMR (by Telus Health)

1. Introduction

In recent years, some legislative initiatives such as the U.S. American Recovery and Reinvestment Health Information Technology for Economic and Clinical Health (HITECH) Act or ARRA-HITECH[1] have brought significant changes in the healthcare industry in the United States. This legislation made massive funding available to healthcare practitioners to adopt health information technology (HIT) (e.g., electronic medical records systems [EMRs]) to streamline diagnosis and treatment plans at the point of care(1, 2). One of the many goals of recent legislative actions in Canada (e.g., the promotion of EMRs by federal and provincial governments) was to contain the increased spending in the healthcare sector while maintaining good care. While the adoption of HIT would certainly have some impacts on healthcare spending, it has become evident that the adoption of efficient care management strategies (delivery of diagnosis and treatment plans) could also play a big role in containing spiraling healthcare costs (3, 4).

Both the executive and legislative branches of the US government show growing concerns regarding the sustainability of MEDICARE spending. From 1987-2002, MEDICARE expenditures for patients having 5 or more chronic conditions jumped from 52% to 76% of the total MEDICARE spending (3). It was also reported that the treatment cost for people with 5 or more chronic conditions is 17 times higher than for people with no chronic conditions (3).

In Canada, the rate of chronic medical conditions are rising at an alarming 14% every year as reported by the Public Health Agency of Canada in (PHAC) 2010 (4). This high prevalence of medical conditions is translated into escalating costs, which will likely not be sustainable for health system spending. Aging has been found to be a major contributor to the rapid increase of chronic conditions. The increase in the prevalence of disease is associated with the increase in the use of health services (3-6). In 2009, 14% of the Canadian population was over 65 years and this proportion will grow to 25% by 2030 (4). Because age is associated with multiple comorbidities, the current aging trend ultimately will result in higher health costs as demand for care increases (6-8). In Canada, treatment of chronic conditions accounted for 67% of direct heath care costs in 2009 (e.g., hospitals, physicians) (4).

---

[1] In aftermath of the 2008 economic recession, President Obama signed into law a $789 billion dollar economic stimulus package, known as the American Recovery and Reinvestment Act, or ARRA on February 17, 2009. Included in ARRA legislation was the Health Information Technology for Economic and Clinical Health (HITECH) Act, also known as HITECH.

A close look at the health spending in Canada revealed that health care costs are highly concentrated within certain strata of the population. In 2013, the top 1% of health system users in Ontario accounted for 49% of the total healthcare costs (combined hospital and home care costs) (7, 9). The top 5% of the health system users were also high cost users: they consumed 84% of total healthcare costs (combined hospital and home care costs)(7, 9).

It is projected that healthcare spending will increase for older people who are likely to be people with complex medical needs because of their multiple chronic conditions(9). The continuation of this trend is unsustainable for healthcare system spending. As a result, many initiatives have sprung up to help understand and contain the spiraling healthcare costs across Canada as well as in Ontario (10).

The Ontario Ministry of Health and Long-Term Care (MOHLTC) launched Health Links as a model to assist health system managers with different tools to create a coordinated care plan for patients' complex medical needs. With involvement of primary care providers (PCPs), the emphasis has been put on the successful identification of health systems users with complex medical needs (10). Health Link is a team of providers in a geographic area whose role is to create a care plans for patients with complex medical needs by insuring coordination and transition between services or providers for these patients.

The MOHLTC, along with other stakeholders, agreed on the identification of patients with 4 or more comorbidities (or 4+ patients) as one of the critical element to define the target population for Health Links (10).

Ontario is divided into 14 geographic areas called LHINs. These organizations have the mandate to "*plan, integrate and fund local health care, improving access and patient experience*". Some LHINs in Ontario aim to contain health resource utilization by identifying patients with preventable high cost healthcare service usage (e.g., hospitalizations, ED visits) (11). To achieve this strategic objective, health managers believe that a meaningful partnership with PCPs who are the entry point to the health system will be pivotal.

PCPs face many challenges identifying high users of healthcare services. Their ability to do so could help health system managers in the implementation of care management strategies that will hopefully cut health care expenditures while optimizing care quality. Among challenges regarding intelligent cost cutting are data availability, tools to help predict potential high cost users of health services, and defining appropriate health care management plans for them. While the data availability challenge has been somewhat addressed with several initiatives to standardize health data in primary care setting, the necessary tools to manage and refer patients with complex medical needs in primary care setting are still lacking. This research effort hopes to start filling that gap by using coded primary care health data and applying predictive analytics tools to possibly provide PCPs with a tool to identify patients who are high users of the system and to potentially reduce health services uses/costs through better care management.

2. Research Question and Objectives

To arrive at the research question and the objectives of this study, which focuses on the early identification of patients with complex medical needs (i.e. patients with multiple comorbidities), we explored previous work.

2.1 Literature Review

Health services utilization

Even though increasing healthcare costs drew, and continue to draw, the attention of policy-makers and decision-makers of the healthcare system, the issue of health care resource utilization has not garnered major attention in the research world (e.g., academia and research institutions). There has been more published scientific literature on health care services utilization before 2000s.

Socio-demographic factors of health services utilization

An analysis of health services utilization data can help answer important questions about the health care system: health care access and disparity, efficiency and quality of care delivery of different components of the healthcare system (e.g. hospitals, ED visits, general practices) and predictions about the propensity of the use of health services and the total costs of care for various groups (e.g. patients with coronary artery disease or congestive heart failure) for a specific time (12).

Health services utilization can be measured either by the number of uses or the cost associated with a specific health service (12). Very often, analyses presented in the literature have focused on assessing the relationships between the cost of health services and certain patient characteristics such as age and sex (5, 12-14).

While most of the analyses on healthcare services tend to indicate a linear relationship between cost, age and sex, Diehr et al. (11) posited rather that the relationship is non-linear and thus recommended analytic approaches to account for the complexity of relationships among those determinants of health services utilization. In reviewing different approaches to assess health services utilization behavior, Anderson(13) suggested that health status is another key variable of interest.

Health status and medical conditions

Health status, as defined in the literature, can emphasize various dimensions of health. According to Statistics Canada, health status is a broader term that cannot be captured by a single measure (15). There have been different attempts to capture health status by using indicators. Those indicators point out both the subjective and objective nature of the assessment of health. Thus, health status can be defined as "*The level of health of the individual, group, or population as subjectively assessed by the individual or by more objective measures*."(16) Among the various attempts at measuring health status, researchers have developed a variety of health status surveys (16, 17). Those surveys and their measurements of health tend to capture both the objective and the subjective nature of health. Examples of health status indicators are perceived health status, role limitations, functional limitations and restricted-activity days (17). This latter indicator is directly linked to medical conditions. In the literature, some researchers confined health status to an individual's medical condition. Hence, heath status is defined by Rumsfeld(16) as "*a range of manifestation of disease in a given patient including symptoms, functional limitation, and quality of life, in which quality of life is the discrepancy between actual and desired function*". However limiting that definition may seem to restrict health status to the overall impact of medical conditions of an individual's health, the definition does capture most of the states which health status indicators attempt to measure. The working definition of health status used in this study is the same proposed by Rumsfeld, with an emphasis on medical conditions as determined by a clinician's diagnosis of a patient's conditions.

The universe of health status in this study is limited to eight medical conditions diagnosed by clinicians and captured (coded) in the research database obtained for this research. These medical conditions are hypertension, osteoarthritis, depression, diabetes, chronic obstructive pulmonary disease (COPD), dementia, Parkinson's disease and epilepsy[2].

[2] CPCSSN chose to code the eight listed medical conditions from all the medical conditions obtained from the EMRs of participating physicians in the project.

Medical conditions and health services utilization

Pope (17) examined the use of ambulatory services in the USA and found a strong association between health status and medical conditions, and also between ambulatory service utilization and medical conditions. Ischemic heart disease, conditions of pulmonary circulation and other heart and mental disorders are among the most frequent medical conditions associated with extensive utilization of ambulatory care services (17).

In general, a person's medical condition was found to drive his or her utilization of health services (5, 17). A more specific characterization of this fact is that certain specific medical conditions (e.g., chronic conditions) and age (aging population) were found to increase the use of health services and consequently drive up costs (5).

Age, chronic conditions and utilization of health services

Age and chronic conditions

The literature shows a strong association between age and medical conditions. People who are young (12-39 years) or adults (40-59 years) tend to have fewer medical conditions than those who are older (60+) who tend have an exceptionally high number of medical conditions (2 or more)(5). Age and number of chronic conditions are strongly related to the utilization of health services (5, 6, 14).

It is expected that the growing trend of Canada's aging population will be accompanied by a steady increase in people living with chronic conditions, leading to more pressure on health services(18). Variations in the use of health services have been seen across chronic conditions. A high proportions of people living with high blood pressure, diabetes or heart disease tend to have two or more other chronic conditions(5, 18). Sex differences also have been observed. Women are more likely than men to have two or more chronic conditions(5).

The analysis of the rates of medical conditions and the use of health services indicate that people living with one or more medical conditions make many visits to a family doctor, nurse and specialist (5, 18). This trend toward increased utilization is also seen with the use of expensive health services such as hospitalizations, ED visits and re-admissions(3).

Medical conditions and costs of health services utilization

The literature indicates that "High Cost Users" of healthcare services are highly concentrated among patients with multiple chronic illnesses (3, 5, 7). In Ontario, in 2007-2008, the top 1% of healthcare users accounted for one third of healthcare spending while the lower 50% users consumed 1% of healthcare expenditures(7). Most of the healthcare services utilized by patients with multiple medical conditions include in-patient hospital stays, ED visits and 30-day re-admissions among other uses (5, 7, 19).

Prevention and management of medical conditions

Attention to high cost users of healthcare resources– patients living with multiple chronic illnesses – has grown as they are more likely to use healthcare resources heavily (through hospitalizations, ED visits and re-admissions) (7, 9, 19). It has become evident that some of the heavy uses of healthcare resources could be prevented if those patients could be identified and referred to the care of specialized teams (e.g. Health Links[3])(5, 8).

The need for better management and appropriate prevention of high uses of high cost health services provides justification for this study.

2.2 Research Question

In line with several themes explored in the literature, which have a clear bearing on the topic central to this study – the management of patients with complex needs in primary care settings – we formulated our research question as follows:

In adults and seniors living with multiple medical conditions, which characteristics influence their use of healthcare services in primary care settings in Canada?

In other words, can we use the number of medical conditions, individual medical conditions, age, sex and other risk factors (e.g. smoking, obesity) to predict primary care health resource utilization of adults and older patients living with multiple medical conditions (but who are not yet critically ill and hospitalized)?

[3] Health Links are initiatives from the MOHLTC to create specialized teams of health care providers whose role is to bring together various health care providers (hospital, family doctor, long-term care home, community organization etc.) to better and quickly coordinate care for high-needs patients.

2.3 Study Objectives

The goal of this research is two-fold:

The first goal was to test the use of predictive analytics tools to predict which patients with multiple chronic illnesses were likely to be high cost users of healthcare resources (measured by multiple hospitalizations, multiple ED visits and 30-day re-admissions). In the context of this study, encounters or visits, laboratory tests (labs) and examinations (exams) were used as proxies for acute care health services utilization data (e.g. hospitalizations), which are not available in the primary care database used in this research.

Based on the characteristics of patients identified by predictive analytics, the second goal was to inform the design of a management and referral tool. Ideally such a tool would be incorporated into an EMR through searches linked to reminders and custom forms. Once designed and tested, this tool could help Canadian PCPs to manage and refer patients with complex medical needs to specialized care teams (e.g. Health Links).

3. Data Analysis and Methodology

3.1 Data Preparation

3.1.1 CPCSSN

The research database used for this research was obtained from CPCSSN. The Canadian Primary Care Sentinel Surveillance Network *"is a primary care research initiative—it is the first pan-Canadian multi-disease electronic medical record surveillance system. CPCSSN collects health data from electronic medical records in the offices of participating primary care providers (i.e.) family physicians). CPCSSN's aim is to improve the quality of care for Canadians suffering from five chronic and mental health conditions (hypertension, Osteoarthritis, Diabetes, chronic obstructive pulmonary disease—COPD—and depression) and three neurologic conditions (Alzheimer's and related dementias, epilepsy and Parkinson's disease*"(20).

CPCSSN collects health data from EMRs in the offices of participating PCPs –called sentinels. As of July 2015, CPCSSN included 752 sentinels with a total of 985,176 patients. Currently health data are submitted by sentinels from 8 provinces (British Columbia, Alberta, Manitoba, Ontario, Quebec, New Brunswick, Nova Scotia and Newfoundland) and the Northwest Territories(20).

3.1.2 Research database

For this study, a database was obtained from the data holdings of CPCSSN, which contained data extracted from PCPs EMRs across Canada. The original data extracted from PCP EMRs were subjected to various data transformations and organized into different tables.

The research database that was created for this study contained 12 tables. Table 3-1 below shows the list of 7 tables used for this study.

Table 3-1: Tables in the research database used for this study

Table name	Table content or definition
Patient	List of EMR patients whose PCP is a consenting physician in the CPCSSN project (columns: *Patient_ID, Sex, BrithYear* etc.)
Lab	Results of lab tests relevant to Index Diseases or targeted medical conditions (columns: *Patient_ID, PerformedDate, Name_Orig* etc.)
Exam	Results of physical exams performed on the patient (columns: *Patient_ID, DateCreated, Exam1, Result1, Exam2, Result2* etc.)
Encounter	All encounters of the patient (e.g. visits, phone calls), (columns: *Patient_ID, EncounterDate, EncounterType, Reason_Orig* etc.)
DiseaseCase	Patients in the Patient table who have one or more of the Index Diseases or targeted medical conditions were chronic obstructive pulmonary disease (COPD), depression, dementia, diabetes mellitus, epilepsy, hypertension, osteoarthritis, and Parkinson's Disease
RiskFactor	Risk factors recorded for the patient (smoking, drinking, obesity, diet, exercise and stress)
PatientDemographic	Demographics characteristics of the patients (age and sex)

EMR Electronic medical records system
PCP Primary care provider

3.1.3 Tables and variables creation and variables recoding

To create the appropriate dataset to perform the analysis for this study, the 7 tables of interest were transformed to create new tables, new columns (variables)[4] and recode some of the existing columns. The process of tables and columns or variables transformations was completed using Transact-SQL[5] in SQL Server Management Studio of Microsoft SQL Server 2012. This process of data manipulation and transformation addressed some data quality issues, checking for "*Null*" values and excluding them before joining tables.

The column (or variable) Number of Chronic Condition (or *NCD)* is a dichotomic column taking "*Basic*" and "*Intermediate*" as values. This column was created using the MN Tiering disease scoring scale. The Minnesota (MN) Tiering Disease Categorization system categorizes a patient's complexity level by their number of conditions. It has 5 levels(19).

Table 3-2: Minnesota Tiering disease categorization

Tier	Level of complexity	Number of conditions
Tier 0:	Low	0 Condition
Tier 1:	Basic	1-3 Conditions
Tier 2:	Intermediate	4-6 Conditions
Tier 3:	Extended	7-9 Conditions
Tier 4:	Complex	10+ Conditions

The number of medical conditions for patients in the Research database ranged from 1-7. Selection criteria for the study were such that each patient had 1 or more of the 8 conditions studied. Applying the MN Tiering to the NumberOfConditions column yielded only 3 categories: "*Basic*", "*Intermediate*" and "*Extended*"*(19)*. There are only 5 patients in the "*Extended*" category. These 5 patients were categorized in the "*Intermediate*" category for ease of analysis. Binomial logistic regression takes only 2 values. In addition, putting those 5 patients in the "Extended" category barely meets standards for total numbers for a meaningful statistical analysis, especially performing regression analysis.

[4] In database language, a table is made of rows and columns. Columns from a scientific or research perspectives, columns are considered variables.

[5] Transact-SQL or T-SQL is Microsoft's and Sybase's proprietary extension to SQL. SQL, the acronym for Structured Query Language, is a standardized computer language that was originally developed by IBM for querying, altering and defining relational databases, using declarative statements. (Wikipedia)

a. Creating health services utilization tables

Three tables in the Research database were identified as health services utilization tables: *Encounter*, *Exam* and Lab tables. The analysis performed using these three health services categories is based on one year of data (January to December 2014).

An encounter (as defined by CPCSSN in their data dictionary) is an "*interaction of the patient with a provider in some fashion*". A close look at some of key data elements captured under this term indicated an encounter can be grouped in two categories: contacts (face-to-face or not) and the venue of the contact if it is not face-to-face. There were nine unique types of encounters recorded for the one-year time period of this study (Appendix Table 0-6). During encounters, patients also discussed the reason for their encounters or contacts, which was also included in the data. Over 59,000 of unique reasons were recorded for patients included in the database for the specified 1-year time-period.

Examinations are defined (according to the CPCSSN data dictionary) as "*Results of physical exams performed on the patient*." There were 23 unique types of exams performed for patients included in this research (Appendix Table 0-7).

Laboratory tests are defined (according to the CPCSSN data dictionary) as "*Results of lab tests relevant to Index Conditions*." The data dictionary indicated that only laboratory tests applicable to diabetes mellitus were collected in the CPCSSN data holdings. There were 133 unique types of lab tests collected for patients under consideration (Appendix Table 0-6).

Three new health services utilization tables (***Encounters***, ***Exams*** and ***Labs***) were created by counting the number of times the health services targeted were accessed by patients for the year 2014 (January to December). The 3 targeted health services were Laboratory Tests, Visits or Encounters and Examinations. The newly created health services utilization tables were joined with the *DiseaseDemo* table to create the ***ResUtilization*** table (this was the final dataset used for analysis).

3.2 Analytic Approaches

Analyses of health services utilization data had mostly used a "one-part" model adjusting for common covariates such as age and sex and including some interactions terms. These "one-part" models do not necessarily account for the complexity of relationships that may exist within the data between the outcome (cost) and predictors (e.g. sex, age, number of conditions). Diehr et al(12) recommended the use of "two-part" models for a better fit between models and model assumptions. The idea in this method is to allow a better understanding of health services utilization by first checking the propensity of the use of health services using a logistic regression or probit model and then regress covariates on utilization cost using only those patients who actually accessed the health services. We therefore applied Diehr's methods for two-part models.

The main goal of this study is to make predictions about the use of a health service (whether a patient would have any use of a health service). We especially wanted to predict which patients were likely to be high cost users of health services such as hospitalizations or ED visits.

The Diehr two-part model was partially used for this study. To adhere to the goal of the study – which was to predict the propensity of use of a health service – only the first part of the two-part model was used. A binomial logistical regression was performed using the *Resource Utilization* dataset. The dataset was split randomly into training and validation datasets. Three data mining algorithms (Naïve Bayes, Clustering, and Decision Tree) were applied to the original dataset. We compared each algorithm's level of accuracy and selected the most appropriate model to be used for predictions (based on model performance criteria). The four outcome variables (Number of labs, exams, encounters and chronic diseases) in the *ResUtilization* dataset were used to run each of these models. In total, 12 models (3 models for each of the 4 outcomes variables) were run on both the training and validation datasets.

The predictive modeling analysis was complemented by an association analysis using the Microsoft Association Rules algorithm(21). The association analysis finds interactions between conditions and helps validate the predictive power of the most significant predictors identified in the predictive modeling analysis.

The medical conditions used in this study are limited to 8 medical conditions which were the 8 coded by the CPCSSN project after obtaining data from the EMRs of physicians who participated in the project. These eight medical conditions are hypertension, osteoarthritis, depression, diabetes, COPD, dementia, Parkinson's disease and epilepsy.

4. Defining the population

4.1 Demographics

The final dataset used for this study contains 14,004 patients' records. Women represent more than half (54%) of the study population (see Table 4.1). The age variable was recoded in four age groups: the adult age group represents two fifths (41%) of the patient population while the other age groups (all senior age groups) were more or less evenly represented. The average age for the population was approximately 68 years. From table 4-1 we found that one quarter of the patient population was 59 years old or younger and 75% of the population was 75 years old or younger.

Table 4-1: Distribution of population study by sex and age

Demographics	# of Cases	%	Cumulative %
Sex			
Women	7718	53.56	53.56
Men	6691	46.44	100
Age Category			
Adult (49 – 64)	5950	41.29	41.29
Senior1 (65 – 69)	2461	17.08	58.37
Senior2 (70 – 74)	2077	14.41	72.79
Senior3 (75 – 79)	1649	11.44	84.23
Senior4 (>= 80)	2272	15.77	100

Demographics	**Mean**	**25th Percentile**	**75th Percentile**
Age	67.73	59	75

4.2 Health status

In line with the discussion on health status, this study is concerned with measures related to medical conditions and the relationships between those measures and health services utilization. Table 4.2 gives a summary of the patient population under consideration with regards to their specific medical conditions and the number of conditions they have.

In the research database 8 medical conditions were captured: hypertension, osteoarthritis, depression, diabetes, COPD, dementia, Parkinson's disease and epilepsy. These medical conditions were retained for analysis. Each patient studied had at least 1 of these 8 conditions. The medical condition with the highest prevalence rate were: hypertension (69%), diabetes (50%), osteoarthritis (33%) and depression (24%). Parkinson's disease, epilepsy and dementia are the medical conditions with the lowest prevalence rate respectively 1%, 2% and 4% (Table 4-2).

Table 4-2: Health status of the study population

Conditions	# of cases	%	Ranking
Hypertension	9890	69%	1
Diabetes	7198	50%	2
Osteoarthritis	4808	33%	3
Depression	3497	24%	4
COPD	1596	11%	5
Dementia	599	4%	6
Epilepsy	249	2%	7
Parkinson	111	1%	8

Number of Medical conditions	# of Cases	%	Cumulative %
Basic (1-3 medical conditions)	13487	93.6	93.6
Intermediate (4 or more medical conditions)	922	6.4	100

Medical conditions	Mean	25th Percentile	75th Percentile
NumberOfConditions	1.94	1	2

Chronic obstructive pulmonary disease (COPD)

With regards to the number of medical conditions and based on the MN Tiering comorbidity complexity scale, 9 in 10 patients are categorized as having a basic level of comorbidity complexity (1-3 conditions) (19). Six percent are categorized as having intermediate to extended level of comorbidity complexity (four or more conditions)[6]. The average number of medical conditions within the population is 2 conditions. Twenty five percent of the patient population have only 1 medical condition and 75% of the population have only 2 medical conditions.

[6] Based on MN Tiering comorbidity complexity scale, only one patient was categorized in the extended level (7-9 diseases). This lone patient was included in the intermediate level for smoother analysis.

4.2 Health Services' use

Table 4-3: Health Services' Utilization

Health Services' Use	# of Cases	%	Cumulative %
NOL (Number of lab use)			
BasicUse (1-5 labs)	8026	55.7	55.7
HighUse (6 or more labs)	6383	44.3	100
NOEx (Number of exams use)			
BasicUse (1-4 exams)	9157	63.55	63.55
HighUse (5 or more exams)	5252	36.45	100
NOE (Number of visits to physician)			
BasicUse (1-2 visits)	8794	61.03	61.03
HighUse (3 or more visits)	5615	38.97	100

Health Services' Use	**Median**	**Mean**	**25th Percentile**	**75th Percentile**
NumberOfLabs	5	5.18	3	7
NumberOfExams	4	4.04	1	5
NumberOfEncounters	2	2.93	1	3

Three types of health services were captured in the research database: labs, exams and encounters. This study is more concerned about the propensity (or the intensity) of health services use. As a result, the three types of health services retained for analysis were recoded based on their intensity of use. The median for each of these three health services was calculated to determine their level of use.

Encounters have the lowest average number (3 visits) of health service's use and labs have the highest average (5 labs) of health service's use. Twenty five percent of the patients used 3 labs, 1 exam and 1 visit in 2014. Seventy five percent of the patient population used 7 labs, 5 exams and 3 visits for the same time period. This made lab use the most intense health service used. This is reflected in the general patient population in this study. Forty four percent of patients had had an intense use (5 or more labs services used) of lab services while 36% and 39% of the patients had used respectively exams and visits at the same level (intense use: 5 or more exams and 3 or more visits).

Note that only laboratory tests that were related to diabetes were included in the study data set.

5. Assessing differences in health services utilization

5.1 Hypothesis testing criteria

The main goal of this study is to characterize the health service utilization in primary care settings by identifying key patient characteristics that influence their use of health services. Before arriving at the model building stage that will help achieve the study's goal, it is important to assess whether differences exist in the ways patients identified for this study use health services.

The following sections examine differences in the use of the three health services identified (labs, exams and encounters or physician's visits) and patients' demographic characteristics (sex and age) as well as their medical conditions.

Hypothesis testing helped assess differences in health services utilization and certain patients' characteristics by stating a null and an alternative hypotheses and selecting a risk level (or α) of 0.05 (which represents a 5% chance of making error). The hypothesis testing used the analysis of variance method.

The graphical visualization along with Shapiro test provided evidence that the data used for the three dependent variables of interest (*NumberOfLabs*, *NumberOfExams* and *NumberOfEncounters*) are not normally distributed.

To meet the requirements of performing the Shapiro test in R, 20% of the 14,404 patients in the population were randomly sampled giving a sample size of less than 3,000. The Shapiro test in R requires a sample size ranging between 3,000 and 5,000. The graphical normality test was performed using the original sample size of the dataset.

The p-value for all the three variables being less than 0.05 (see Table 0-1 in Appendix), the null hypothesis that the sample data for the three variables are normally distributed was rejected and the alternative hypothesis that the sample data for all three variables were not normally distributed was retained. However, having a sample size big enough as stipulated by Central Limit Theorem[7], the analysis of variance was used to test differences in mean use of health services among the patients' population(22).

5.2 Differences in the use of health services and demographic characteristics

5.2.1 Health services' utilization by sex

This section examined differences in the use of the three health services identified (labs, exams and encounters or physician's visits) and patients' sex. A null and an alternative hypotheses are stated below.

H_0: $\mu_{Men} = \mu_{Women}$ (means of health services' use were equal for Men and Women)
H_1: means were not equal
$\alpha=0.05$

The comparison of the mean use of health services by sex showed that the mean use of each type of health service is similar to the mean use of the overall population without accounting for sex differences.

Table 5-1: Assessing differences in mean use of health services and patients' sex for 2014

Health Services	Women	Men	Df	F value	Pr(>F)
Labs	5.17	5.19	1	0.131	0.717
Exams	3.99	4.1	1	0.131	0.717
Encounters	2.95	2.92	1	0.252	0.616

Df: Degree of freedom
Pr: Probability

The analysis of variance reinforced the fact that sex introduces no difference in terms of the use of health services considered in this analysis (Table 5-1). The F static is very small indicating no difference between men and women with regards to the means of lab use, exam use and visits for one year. The p-values for lab use, exam use and visit are greater than the chosen level of (p=0.05).

[7] Central Limit Theorem states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution (Wikipedia).

As a result of a small F value and P> α, the null hypothesis of equality in the means of the use of the labs, exams and visits was retained.

5.2.2 Health services utilization by age category

This section examined differences in the use of the 3 health services identified (labs related to diabetes, exams and encounters or physician's visits for one year) and patient age (hypotheses were tested for each of the 5 categories of age). A null and alternative hypotheses are stated below.

H_0: $\mu_{adult} = \mu_{senior1} = \mu_{senior2} = \mu_{senior3} = \mu_{senior4}$ (means of health services' use were equal for all age categories)
H_1: means were not equal for at least one age category
$\alpha=0.05$

Table 5-2 shows differences in terms of mean use of labs, exams and visits among the 5 age categories considered.

Table 5-2: Assessing differences in means use of health services and patients' age categories for 2014

Health Services	Adult (49-64)	Senior1 (65-69)	Senior2 (70-74)	Senior3 (75-79)	Senior4 (80+)	Df	F value	Pr(>F)
Labs	5.27	5.37	5.21	5.13	4.76	4	12.45	4.17e-10 ***
Exams	4.03	3.99	4.02	4.14	4.05	4	0.574	0.682
Encounters	2.84	2.82	2.88	3.09	3.23	4	8.817	4.17e-07 ***

Significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Df: Degree of freedom
Pr: Probability

Table 5-2 supported the existence of variations in mean use of health services under analysis among age categories. There were statistically significant differences in the mean use of labs and visits by age category ($p < 0.000$). The F statistic is bigger for these two health services as compared to the F statistic for mean use of exams for which no statistical difference was found among various age categories $p > 0.05$).

Mean use of labs was lower for Senior4 (80+) than the other four age categories. Senior3 (75-79) and Senior4 (80+) mean use of labs was lower than the mean use of labs for the overall patient population (Table 5-2). In contrast, the other 3 age categories had a slightly higher mean use of labs as compared to the patient population sample mean (5.18).

Senior3 (75-79) had the highest mean use of exams while Senior1 (65-69) had the lowest mean use of exams for all the five age categories. These two means were respectively higher and lower compared to overall patient population mean use of exams (4.04).

Senior3 (75-79) and Senior4 (80+) had a slightly higher mean use of visits as compared to the population sample mean (2.93). These two means were also higher compared to the other age categories whose means were lower than the patients' population sample mean (Table 5-2).

With a smaller F statistic for exam use indicating no statistically significant difference in the use of this health service among the various age categories and a p-value > α (0.05), the null hypothesis of equal mean of exam use for all age categories was retained.

However, with a bigger F statistic for both diabetes related lab use and visit use indicating a variation in the use of those services among age categories and the p-values < α (0.05), the null hypothesis of equal means of lab use and visit use among age categories was rejected.

These findings were further investigated by performing the *Tukey HSD (Honest Significant Differences) test* to determine for which age categories differences in the means of labs' and visit use exist.

Statistically significant age groups for lab use

Table 5-3: Tukey Honest Significant Differences for diabetes related Lab use in 2014

Age Category	Difference	Adjusted p values
Senior4-Senior2 (80+) (70-74)	-0.451239	0.0000689
Senior4-Senior3 (80+) - (75-79)	-0.371472	0.0047107
Senior4-Adult (80+) - (49-64)	-0.510801	0.0000000
Senior4-Senior1 (80+) - (65-69)	-0.610051	0.0000000

Differences in mean use of labs occurred between Senior4 (80+) and all the other 4 age categories (Adult (49-64), Senior1 (65-69), Senior2 (70-74) and Senior3 (75-79)) (table 5-3). This is confirmed by another HSD test which shows two major groupings in terms of labs use as a result of age: patients in Senior4 (80+) age category used fewer labs (table 5-4). These patients were statistically different from patients from the other four age categories who used more labs services in one year (Table 5-4). Patients in these 4 age groups were similar in their use of lab service.

Statistically significant age groups for visit use

Table 5-4: Tukey Honest Significant Differences for Visit use in 2014

Age Category	Difference	Adjusted p values
Senior3-Adult (75-79) - (49-64)	0.2529221	0.0231034
Senior4-Adult (80+) - (49-64)	0.3858721	0.0000026
Senior3-Senior1 (80+) - (70-74)	0.2717665	0.0393902
Senior4-Senior1 (80+) - (65-69)	0.4047166	0.0000454
Senior4-Senior2 (80+) - (70-74)	0.34503	0.0017011

As for visits, significant differences in the use of this health service were observed between Senior3 (75-79) and Senior4 (80+) and the younger senior age categories

Table 5-5: Statistically significant age groups for visit use in 2014

Groups	Treatments	Means of number of visits
a	Senior4 (80+)	3.228
ab	Senior3 (75-79)	3.095
bc	Senior2 (70-74)	2.883
c	Adult (49-64)	2.842
c	Senior1 (65-69)	2.823

Note: Means with the same letter are not significantly different

(Senior1 (65-69) and Senior2 (70-74)) and adult (49-64) age category (Table 5-5). The HSD Test (for identifying statistically significant groupings) shows 3 groupings (Table 5-6). The first group is made of Senior4 (80+) and Senior3 (75-79) who had more visits than the other age groups. The second group made of Senior3 (75-79) and Senior2 (70-74) had fewer visits than the first group. The latter group is made of Senior1 (65-69) and Adult (49-64) who had fewer visits than the other two groups. Patients in each of these three groupings had similar patterns of visit use, but clearly the visit use for Senior3 (75-79) and Senior4 (80+) were statistically different for Adult (49-64), Senior1 (65-69) and Senior2 (70-74) (Table 5-6).

5.3 Differences in the use of health services and health status

5.3.1 Health service utilization and patients' number of medical conditions

This section examined differences in the use of the three health services identified (labs, exams and encounters or physician's visits) and patients' medical conditions (measured by number of medical conditions). A null and alternative hypotheses are stated below.

H_0: $\mu_{basic} = \mu_{intermediate}$ (means of health service use were equal regardless of the number of medical conditions)
H_1: means were not equal
$\alpha=0.05$

The number of medical conditions (or NCD) in this study was measured using the MN Tiering comorbidity complexity level(19). The patient population for this study was categorized in Basic (1-3 conditions) and Intermediate (4-7 conditions) levels.

Table 5-6: Assessing differences in means use of health services and patients' Number Medical Conditions

Health Services	Basic (1-3 Conditions)	Intermediate (4+ conditions)	Df	F value	Pr(>F)
Labs	5.22	4.67	1	23.82	1.07e-06 ***
Exams	4.01	4.45	1	15.59	7.92e-05 ***
Encounters	2.86	4.02	1	126	<2e-16 ***

Significance: 0 '***' 0.001

Df: Degree of freedom

Pr: Probability

Table 5-7 shows statistically significant differences in the mean use of the 3 health services depending on patient's number of medical conditions (p< .000). The mean use of lab service for one year was higher for patients with less than 4 conditions as compared to those having 4-7 conditions. Lab mean use for this latter group was lower than the population sample mean (5.18). The reverse trend was observed for exams and visit use. Mean use for these two services was higher for patients with 4-7 conditions than for those with less than 4 conditions. Patients living with 1-3 conditions had fewer exams and visits. Their mean use for exams and visits was lower than the population mean (respectively 4.04 and 2.98) for one year.

These observed differences in means of health services utilization are an indication that the number of conditions of a patient had influenced his or her utilization of health services. Patients with 1-3 conditions had fewer visits and exams in a year while patients with 4 or more conditions tend to have more visits and exams.

The F statistic for each of these services is bigger and their respective p-values are less than the chosen α (0.05) level. Thus the null hypothesis of equal mean use for the 3 health services was rejected. The number of medical conditions introduced statistically significant patterns of use of the three health services. The F statistic for visit use is the biggest. This indicates that patients living with multiple conditions had more visits.

5.3.2 Health services utilization by patients' medical conditions

This section examined differences in the use of the three health services identified (labs, exams and encounters or physician's visits) and patients' medical conditions (measured by the type of medical conditions). A null and alternative hypotheses are stated below.

H_0: $\mu MC_{no} = \mu MC_{yes}$ (means of health services use were equal regardless of the type of medical conditions)
H_1: means are were equal
α=0.05

Note:
MC or medical conditions
Four medical conditions are evaluated: hypertension, diabetes, Osteoarthritis and Depression.

The analysis of means of health services and medical conditions was performed for the four most prevalent chronic conditions: hypertension, diabetes, osteoarthritis and depression.

Table 5-7: Assessing differences in means use of health services for (patients with or without) hypertension

Health Services	Hypertension_No	Hypertension _Yes	Df	F value	Pr(>F)
Labs	5.18	5.18	1	0.007	0.935
Exams	4.06	4.03	1	0.258	0.611
Encounters	2.84	2.98	1	5.711	0.0169 *

Significance: '*' 0.05 '
Df: Degree of freedom
Pr: Probability

The results of ANOVA test (Table 5-8) showed that there was not a statistically significant difference in the means of lab use and exam use for patients with or without hypertension. However, the results indicated a statistically significant difference in the means of number of visits. Patients with hypertension had more visits as compared with those without hypertension. The null hypotheses of equal mean for lab use and exam were retained while the null hypothesis of equal mean for visit use was rejected.

Table 5-8: Assessing differences in means use of health services for (patients with or without) diabetes

Health Services	Diabetes_NO	Diabetes _Yes	Df	F value	Pr(>F)
Labs	5.23	5.14	1	2.444	0.118
Exams	3.69	4.39	1	165.6	<2e-16 ***
Encounters	3.69	4.39	1	65.8	5.39e-16 ***

Significance: '***' 0.001

Df: Degree of freedom

Pr: Probability

The means of exam use and visit use for patients with diabetes were statistically different for those who did not have diabetes for 1 year (Table 5-9). Patients with diabetes used more exams and made more visits than patients who did not have diabetes. There was no statistical difference in the use of labs for these two groups. The null hypotheses of equal exam use and visit use were rejected and the null hypothesis of equal mean for lab use was retained when analyzing the presence of diabetes.

Table 5-9: Assessing differences in means use of health services for (patients with or without) osteoarthritis

Health Services	Osteoarthritis_NO	Osteoarthritis _Yes	Df	F value	Pr(>F)
Labs	5.23	5.08	1	7.113	0.00766 **
Exams	4.02	4.07	1	0.62	0.431
Encounters	2.81	3.17	1	44.98	2.07e-11 ***

Significance: 0 '***' 0.001 '**' 0.01

Df: Degree of freedom

Pr: Probability

There was a statistically significant difference in the means of lab use and visits in 1 year for patients with osteoarthritis as compared with patients who did not have the condition (Table 5-10). On one hand, patients without osteoarthritis used more labs than those with osteoarthritis. Patients with osteoarthritis made more visits than those without the disease. There was no statistically significant difference between the two groups for exam use. The null hypotheses of equal mean for lab use and visit use were rejected and null hypothesis of equal mean for exam use was retained.

Table 5-10: Assessing differences in means use of health services for (patients with or without) depression

Health Services	Depression_No	Depression _Yes	Df	F value	Pr(>F)
Labs	5.26	4.94	1	24.11	9.22e-07 ***
Exams	4.03	4.08	1	0.786	0.375
Encounters	2.83	3.26	1	53.37	2.91e-13 ***

Significance: '***' 0.001
Df: Degree of freedom
Pr: Probability

The pattern of use of exams was similar for patients with depression and those without depression (Table 5-11). However, the 2 groups exhibited a statistically different pattern for labs use and visits ($p< .000$). Patients without depression used more labs than patients with depression. As for the mean number of visits, patients with depression made more visits as compared with those who did not have depression. The null hypothesis of equal mean for exam use was retained and the null hypotheses of equal mean for lab use and visit use were rejected.

The one-way ANOVA was performed to look at differences between patients based on their demographic characteristics (sex and age) and health status (number of conditions and patients' medical conditions) and their health services' utilization patterns. The test found that except for sex, age and health status introduced significant differences in patients' utilization of health services.

The multivariate analysis through model building helped untangle the relationships between the outcome variables (health services of lab use, exam use and visits) and the independent variables (sex, age and health status).

6. Characterization of patients' health services utilization in primary care settings

6.1 Model building and model selection process

6.1.1 Problem definition

The main goal of this study was to use predictive analytics tools to help identify individuals who might be prospective high users of health services (i.e. hospitalizations and ED visits) health services use and patients characteristics in primary care. The bivariate analysis indicated statistically significant associations between some patient characteristics and health services use in primary care. This section presents the modeling approaches and tools used to link health services utilization and the most prominent patient characteristics that could potentially predict who are the high users of health services—in this case labs, exams and visits rather than hospitalizations or ED visits.

6.1.2 Model selection process

The health resource utilization dataset with 14,004 patient records was used for model development. The dataset was randomly split into training and validation or test datasets with 70% and 30% respectively of the original dataset.

Previous research, by the author, had shown weak association between the selected predictors for this study and the selected health services utilization outcome variables Number of Labs (**NOL**), Number of Exams (**NOEx**) and Number of encounters (**NOE**)). For that reason the Number Chronic Conditions (**NCD**), which showed strong association with the selected predictors for this study, was added to the model building process.

Two groups of models were run and compared to select the best performing model for making predictions.

The first group of models contained the three original outcome variables of health utilization services: Number of diabetes related Labs (**NOL**), Number of Exams (**NOEx**) and Number of encounters (**NOE**).

Base models using classification algorithms (Decision Tree, Naïve Bayes and Neural Network)
NOL, **NOEx** and **NOE**= **NCD**, medical conditions (8 conditions), Sex, Age, risk factors

The second group of models actually used the Number of Chronic Condition (**NCD**) as outcome variable and included the other three health utilization variables (NOL, NOEx and NOE) as predictors along with the other predictors.

NCD-Based model using classification algorithms (Decision Tree, Naïve Bayes and Neural Network)
NCD= medical conditions (8 conditions), Sex, Age, risk factors, **NOL**, **NOEx** and **NOE**

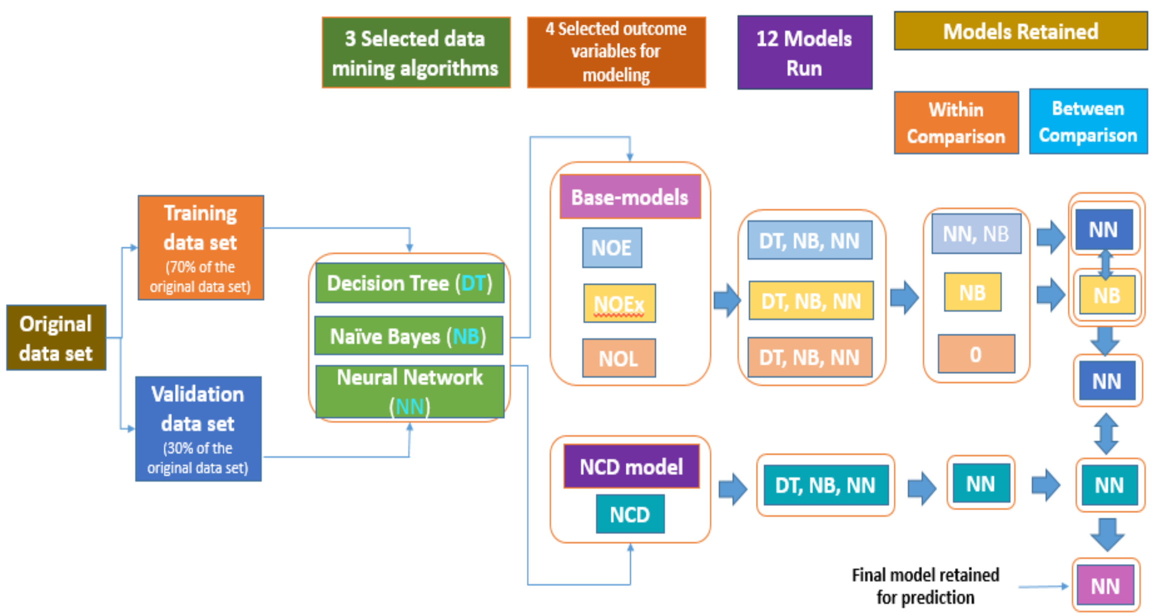


Figure 6-1: Model development and model selection diagram flow

Number of Encounters (NOE)
Number of Exams (NOEx)
Number of Labs (NOL)

These two groups of models used the same 3 classification-mining algorithms (Decision Tree, Naïve Bayes and Neural Network). The based models contained 3 groups of models (3 models for each of the 3 health services outcome or predictable variables). The NCD-based models contained 3 models. For the Base models, within groups and between groups comparison were made (see Figure 6-1). First, within groups comparisons were made for the 3 groups of the Base models to identify the best performing group of models. That process yielded the 2 best performing groups of models. Then a between group comparisons helped identified the best performing model in each group. Finally, the best model from Base models – after the within and between groups comparisons – was compared to the best model of the NCD-based model for another between group comparison to select the best model of the 2 best models from each group (see Figure 6-1). Within and between groups' comparisons were made based on model performance criteria such as lift chart/score and classification matrix (gauging the sensitivity, specificity, and accuracy of each model). The model performance criteria analysis is presented in the subsequent section.

6.2 Model performance criteria and analysis

6.2.1 Model's performance criteria

Two categories of criteria generated from data mining model output were used to assess model performance or accuracy: lift score/chart and classification matrix.

Lift score and lift chart

Lift chart/score

Lift chart and lift score are closely related terms and used to analyze model accuracy in MS SSAS[8].

The *lift chart* is a graphical representation of model performance as compared to a random guess and measured by the *lift score* showing model improvement on a scale of 0-1(23). The lift chart used in MS SSAS model building is an alternative to the Receiver Operating Characteristic (ROC) for a logistic regression. The lift chart along with lift score are used to assess the quality of a classifier (23).

Figure 6-2 shows both the lift chart and lift score for 5 different models. Every time a model is created, there is always a baseline model (random guess) and an ideal model

[8] Microsoft SQL Server Analysis Services (also referred to as Analysis Services)

(the perfect model used as benchmark)(23, 24). When models built are closely aligned or perfectly aligned with the ideal model, the lift score will be close or equal to 1(23, 24). The lift score is also used to quickly assess the performance of models and select the best one. A higher lift score indicates a better model for making predictions(23).

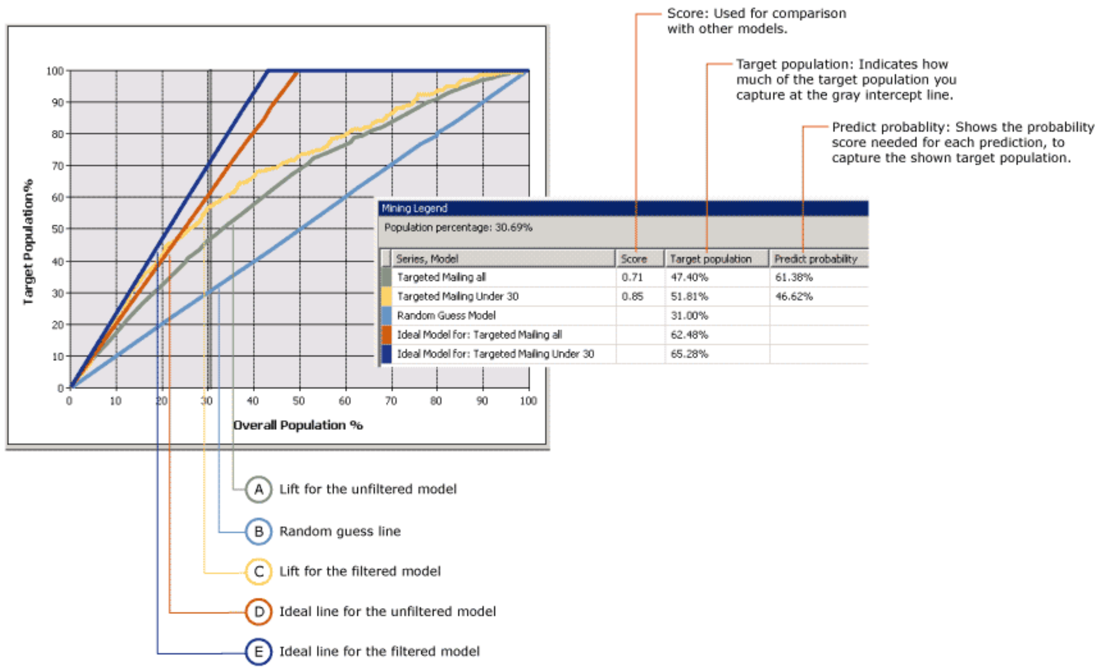


Figure 6-2: Lift chart and lift score illustration as used in Analysis Services

Source: Microsoft Corporation –MSDN-Library/ Lift Chart (Analysis Services – Data Mining)

Model's accuracy measures

In addition to lift chart/score, a classification matrix (confusion matrix) based on the validation dataset was also used to assess model performance or accuracy.

In Analysis Services (MS SSAS), a classification matrix is generated using the validation or test dataset which sorts all cases from a model built into categories (predicted and actual) by determining whether the predicted value matched the actual value. For a binary classifier or model, the cross tabulation of the predicted and actual values are labelled as: *true positive, false positive, false negative* and *true negative (21)*.

These values were used to compute a model's performance measures in terms of sensitivity, specificity and accuracy.

Sensitivity

The model's sensitivity was measured by the proportion of patients correctly identified as having the specified condition (True Positive) by the total number of patients with the specified condition (True Positive and False Negative). False Negative represents a patient incorrectly identified by the model as not having the specified condition when they do(25).

Sensitivity = True Positive / (True Positive + False Negative)

Specificity

The model's specificity was measured by the proportion of patients correctly identified by the model as not having the specified condition (True Negative) by the total number of patients who did not have the specified condition but are noted as having it (True Negative and False Positive). False Positive represents a patient incorrectly identified by the model as having the condition while in reality the patient did not have the specified condition(25).

Specificity = True Negative / (True Negative + False Positive)

Accuracy

Model accuracy was measured by the proportion of all correctly identified cases or patients (TP+TN) by the total number of cases or patients (25).

Accuracy = (True Positive) + (True Negative) / (True Positive) + (True Negative) + (False Positive) + (False Negative)

6.2.2 Lift chart analysis for the base-models

In this section the lift score and other criteria to assess models' performance are shown in graphs. However, only the lift score is the sole criterion of relevance for this analysis.

Comparing the 3 groups of models within the base-models based on the lift score, the NOL-based models were the least performing models with a lift score < .65 (Figure 6.4). The NOE-based models (Figure 6.2) and the NOEx-based models (Figure 6.3) had higher lift score (> .65) and therefore were retained for the next stage of comparison which involved a within comparison for each group (of 3 models).

- NOE-based models

Table 6-1: Lift score of selected models for number of encounters (NOE)

Selected Models	Lift Score
Decision Tree	0.68
Naïve Bayes	0.69
Neural Network	0.69
Ideal Model	1

Of all 3 three models within the NOE-based models, Neural Network model (or NN-model) and Naïve Bayes model (or NB-model) had the same lift score and were retained as best model. The best model for this group was select using the results of the classification matrix performance measures.

- NOEx-based models

Table 6-2: Lift score of selected models for number of exams (NOEx)

Selected Models	Lift Score
Decision Tree	0.7
Naïve Bayes	0.71
Neural Network	0.7
Ideal Model	1

As for the NOEx-based models, the NB-model had the highest lift score (.71) and was retained for the next phase of comparison.

- NOL-based models

Table 6-3: Lift score of selected models for number of labs (NOL)

Selected Models	Lift Score
Decision Tree	0.63
Naïve Bayes	0.64
Neural Network	0.63
Ideal Model	1

The NOL-based models had the lowest lift scores for all the 3 groups of models within the base models and was not retained for next phase of comparison.

6.2.3 Classification matrix analysis for the base-models

The next step of model comparison and selection involved the use of the classification matrix data which were used to compute model performance measures such as sensitivity, specificity and accuracy.

Table 6-4: Accuracy of selected models for predicting number of encounters (NOL)

Counts for Naïve Bayes model for encounter

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	**611**	463
BasicUse	1090	2158

Model's Performance Measures

Sensitivity	Specificity	Accuracy
0.36	0.82	0.64

Counts for Neural Network model for encounter

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	**690**	543
BasicUse	1011	2078

Model's Performance Measures

Sensitivity	Specificity	Accuracy
0.41	0.79	0.64

Comparing the two models retained for the NOE-based models (NN-model and NB-model), the NN-model had the higher sensitivity score and thus was retained as the best model for the NOE-based models (table 6.1). The NN-model (of the NOE-based models) compared to the best performing model of NOEx-based model (NB-model) (Table 6.2), appeared to be the best performing models of the 2 models based on their respective performance measures. The NN-NOE-model had the higher sensitivity score.

Table 6-5: Accuracy of selected models for predicting number of exams (NOEx)

Counts for Naïve Bayes model for exams

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	**468**	398
BasicUse	1071	2385

Models' Performance Measures

Sensitivity	Specificity	Accuracy
0.30	0.86	0.66

In conclusion, using model lift scores and performances measures, the within model comparison yielded the NN-NOE-based model as the best performing model for the base-model series. This retained model was compared to the NCD-based best performing model using the same criteria.

Lift chart and classification matrices analysis for the NCD-based-models

Of the 3 models, the NN_NCD model yielded the highest lift score. However, the lift scores being almost the same for all the 3 models, the best model selection relied on the comparison of the performance measures based on classification matrices.

Table 6-6: Lift score of selected models for number of chronic diseases (NCD)

Selected Models	Lift Score
Decision Tree	0.99
Naïve Bayes	0.99
Neural Network	1
Ideal Model	1

Table 6-3 shows that the NN_NCD model yielded the highest model accuracy scores (on sensitivity, specificity and accuracy). Thus the model was retained as the best performing model for the NCD-based models.

Table 6-7: Accuracy of selected models for predicting number of chronic diseases (NCD)

Counts for Decision Tree model for Number of Chronic Diseases

Predicted	Intermediate (Actual)	Basic (Actual)
Intermediate	**238**	5
Basic	35	4044

Models' Performance Measures

Sensitivity	Specificity	Accuracy
0.87	1.00	0.99

Counts for Naïve Bayes model for Number of Chronic Diseases

Predicted	Intermediate (Actual)	Basic (Actual)
Intermediate	**123**	0
Basic	150	4049

Models' Performance Measures

Sensitivity	Specificity	Accuracy
0.45	1.00	0.97

Counts for Neural Network model for Number of Chronic Diseases

Predicted	Intermediate (Actual)	Basic (Actual)
Intermediate	**273**	1
Basic	0	4048

Models' Performance Measures

Sensitivity	Specificity	Accuracy
1.00	1.00	1.00

6.2.4 Selection of prediction model

The final model selection involved the comparison of the best performing model for each of the 2 groups of models generated (the Base models and NCD-based models).

The lift chart score and classification matrix analysis helped determine the best model to use for prediction of the 2 best performing models selected: NN-NOE-based model and NN-NCD-based model.

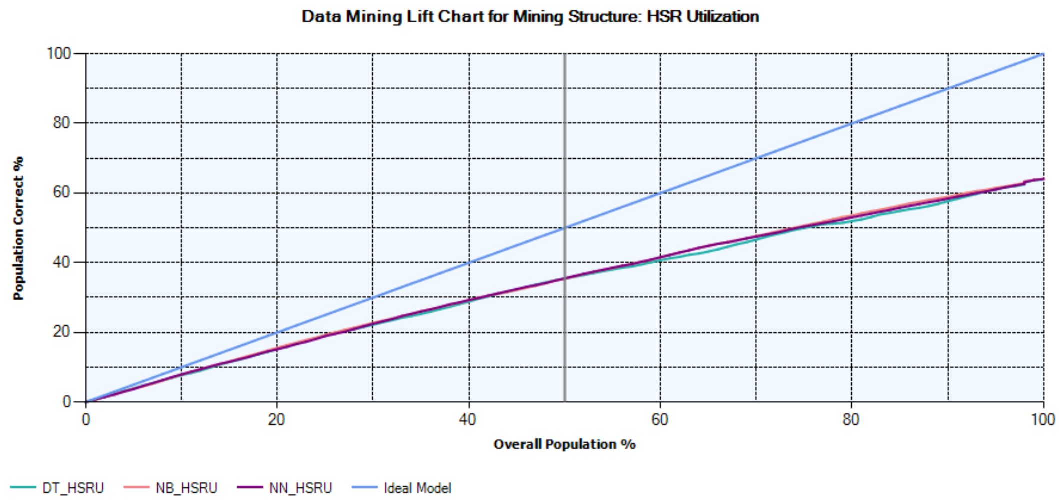


Figure 6-3: Lift Chart for NOE-based models

HSR Health services utilization

The lift score for the NN-NOE-based model and the NN-NCD-based model was respectively 0.69 and 1.00. In addition, the NN-NCD-based model also had higher scores on all the 3 performance measures based on their respective classification matrix. Both the lift scores and models' accuracy measures, without doubt, indicate that the NN-NCD-based model is the best model of the 2 groups of models created (base models and NCD-based models).

These conclusions were further confirmed by the lift chart graphs of the 2 groups of models.

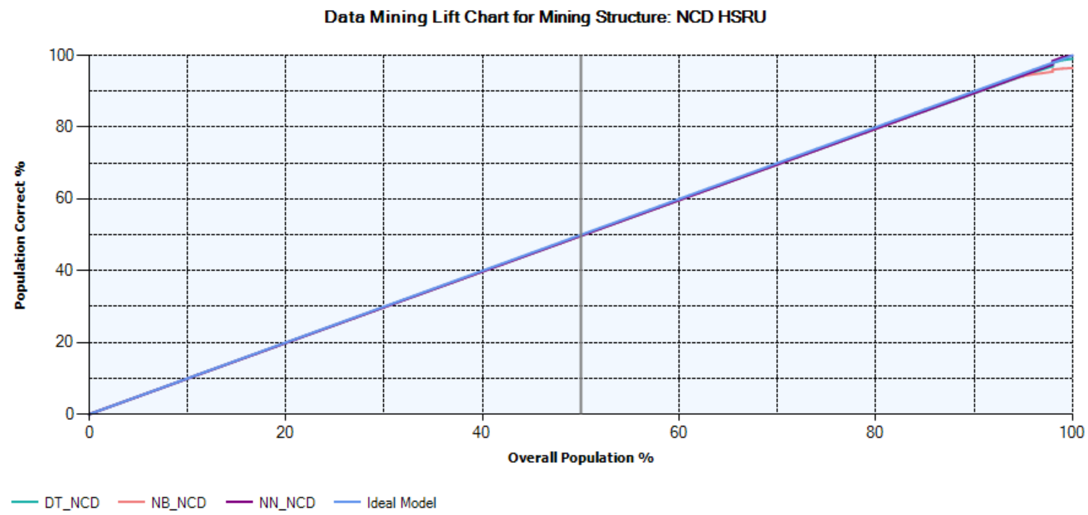


Figure 6-4: Lift Chart for NCD-based models

The NOE-based models' lift scores being lower than 1.00 appeared to be below the ideal model (the diagonal line across the chart). In contrast, the NCD-based models lift scores being close to 1.00 or 1.00 appeared to be aligned with the ideal model.

The within and between models' comparison based on the lift scores and accuracy measures of classification matrices yielded the NN-NCD-based model as the best model for the 2 series of models created. The NN-NCD-based model was therefore used to make predictions about potential high cost users of health services.

6.3 Making predictions

To make predictions, a predictive modeling approach was used by randomly splitting the dataset into training and validation datasets. Three data mining algorithms were selected using 4 outcome or predictable variables: *NOE*, *NOEx*, *NOL* and *NCD*. The first 3 outcome or predictable variables were variables of health services utilization and the last outcome or predictable variable is a health status variable (representing the number of medical conditions of a patient). The 3 selected data mining algorithms (*Decision Tree*, *Naïve Bayes* and *Neural Network*) were applied to the training dataset to identify patterns within the data and make predictions using the validation dataset based on the patterns identified. Predictions were made using the best performing data mining algorithm by comparing the accuracy or performance of the 3 data mining algorithms for each outcome or predictable variable.

The following sections present the descriptive analysis of the best performing model from the ***base models*** which used health services variables as outcome or predictable variables and the ***NCD- based models*** using health status variable number of chronic diseases (*NCD*). The descriptive analysis used the training dataset. The predictive modeling results analysis was based on the original dataset and used the NN-NCD-based model retained as the best model for making prediction.

6.3.1 Descriptive model analysis

Best performing model from the base models

Based on the results of lift score/chart and classification matrix comparing the 9 base models (3 models run for each of the 3 outcome or predictable variables), *Neural Network* algorithm using number of examinations (*NOE)* as outcome or predictable variable is the best model (see section 6.2 for discussion on model performance).

Table 6.8 has 4 columns:

The first 2 columns represent the independent variables or predictors.

Attribute: are the predictors included in the predictive model
Value: the value of the different predictors included in the model

The last 2 columns represent the dependent or outcome variable (Number of Encounters). It has 2 values (Basic Use and High Use). The value associated with each of them is a ranking score to determine the importance of a predictor variable associated with each of the 2 values.

Favors Basic Use: indicates attributes or predictor variables associated with Basic Use
Favors High Use: indicates attributes or predictor variables associated with High Use

Highlighted in red are the predictors associated with the outcome variable being predicted (in this case “***High Use***”).

Table 6-8: Neural Network output for number of encounters (NOE)

Attribute	Value	Favors BasicUse	Favors HighUse
Stress	YES		100
Parkinson	YES		68.7
NOEx	HighUse		64.31
COPD	YES		58.52
Depression	YES		52.22
Obesity	YES	52.02	
NOEx	BasicUse	36.41	
Dementia	YES		35.22
Smoking	NO	35.17	
Hypertension	NO	31.87	
Obesity	NO		30.2
Age Category	Senior4 (80+)		26.3
Epilepsy	YES		21.55
NCD	Intermediate		21.2
Osteoarthritis	YES		20.09
Smoking	YES		19.59
Diabetes	YES		16.84
Diet	YES	15.7	
Depression	NO	15.27	
diabetes	NO	14.49	
Exercise	YES		14.36
Hypertension	YES		13.94
Osteoarthritis	NO	10.77	
Age Category	Senior2 (70-74)		10.43
COPD	NO	9.9	
NOL	HighUse		8.98
Exercise	NO	8.52	
Alcohol	NO		8.42
Alcohol	YES	7.47	
Age Category	Senior1 (65-69)	6.99	
Diet	NO		6.51
NOL	BasicUse	6.2	
Sex	Women		5.92
Sex	Men	1.73	
Age Category	Adult (49-64)	1.51	

Epilepsy	NO	0.77	
NCD	Basic	0.61	
Parkinson	NO	0.48	
Stress	NO	0.45	
Age Category	Senior3 (75-79)	0.44	
Dementia	NO	0.42	

Number of Exams (NOEx)
Chronic obstructive pulmonary disease (COPD)
Number of Labs (NOL)
Number of Chronic diseases (NCD)

Table 6-4 compares the basic users and high users of physician visits (*NOE*) to identify what characteristics differentiate the two groups of patients. The table shows that more than 50 % of physician visits (*NOE*) were high users with Parkinson, COPD and depression; they were high users of exams and they were stressed.

The results also show that, between 15-35% of high users of physician visits (*NOE*) had dementia, epilepsy, osteoarthritis and diabetes; they were Senior4 (>= 80 years), they had 4-7 medical conditions, they were smokers and they were not obese.

Lastly, table 6-4 also shows that 5-15% of high users of physician visits (*NOE*) compared to basic users had hypertension, they were Senior2 (70-74 years), they exercised, they were high users of lab services, they were not drinkers, they did not diet and they were Women.

These findings were supported by a logistic regression analysis using *NOE* as dependent variable (see table 0-2 in Appendices). The logistic regression found that high users of physician visits were more likely to have COPD, diabetes, osteoarthritis, dementia, and depression; they were likely to be high users of exams, they were classified as Senior4 (75-79 years), smokers, drinkers and stressed. These predictors were found to be statistically significant with physician visits (*NOE*).

Best performing model from the NCD-Based models

As the case for the base models, Neural Network was the best performing model for the *NCD-Based models* after applying the 3 selected data mining algorithms to the training dataset. Though all of the three models had almost the same score based on the lift score/chart results (0.99-1.00), Neural Network had higher sensitivity, specificity and accuracy compared to the other two models (Decision Tree and Naïve Bayes) (see section 6.2.2).

Results from table 6-5 – comparing the characteristics of patients with 1-3 conditions (or Basic) and those with 4-7 conditions (Intermediate) – indicate that more than 50% of patients with 4-7 conditions (NCD-intermediate) had Parkinson, epilepsy, dementia, COPD, depression and osteoarthritis. Between 4-35% of NCD-intermediate patients had diabetes and hypertension; they were Senior2 (70-74 years), Senior3 (75-79 years) and Senior1 (65-69 years); they were obese. These characteristics are the best predictors of patients with 4-7 conditions (NCD-intermediate) in contrast to those patients with 1-3 conditions (NCD-Basic) who tended to have the opposite and other characteristics different from patients with more than 3 conditions.

Of the two best performing models using *NOE* and *NCD*, the *NCD Neural Network* model is the best and thus was used to make predictions.

Table 6.9 has 4 columns:

The first 2 columns represent the independent variables or predictors.

Attribute: are the predictors included in predictive model
Value: the value of the different predictors included in the model

The last 2 columns represent the dependent or outcome variable (Number of Chronic Conditions). It has 2 values (Intermediate and Basic). The value associated with each of them is a ranking score to determine the importance of a predictor variable associated with each of the 2 values.

Favors Intermediate: indicates attributes or predictor variables associated with ***Intermediate (4 or more conditions)***
Favors Basic: indicates attributes or predictor variables associated with ***Basic (1-3 conditions)***

Highlighted in red are the predictors associated with the outcome variable being predicted (in this case “***Intermediate***”).

Table 6-9: Neural Network output for number of chronic diseases (NCD)

Attribute	Value	Favors Intermediate	Favors Basic
Parkinson	YES	100	
Epilepsy	YES	83.06	
Dementia	YES	80	
COPD	YES	71.26	
Depression	YES	62.61	
Osteoarthritis	YES	51.31	
Diabetes	YES	33.04	
Stress	YES	27.67	
Diabetes	NO		24.47
Osteoarthritis	NO		20.99
Hypertension	NO		20.92
Hypertension	YES	17.98	
Depression	NO		17.28
Age Category	Senior2 (69-74)	10.23	
COPD	NO		8.12
Obesity	YES	5.54	
Age Category	Senior3 (75-79)	4.82	
Age Category	Senior1 (65-69)	4.2	
Dementia	NO		2.77
Epilepsy	NO		2.01
Parkinson	NO		1.5
Alcohol	YES		0.72
Obesity	NO		0.66
Smoking	NO		0.32
Stress	NO		0.23

Chronic obstructive pulmonary disease (COPD)

6.3.2 Predictive modeling results

The *NCD-Neural Network* model, the best model of the 2 best performing models from the 12 models run for the four outcome or predictable variables (*NOE*, *NOEx*, *NOL* and *NCD*), was used to make predictions.

The *NCD-Neural Network* run on the original dataset (14,004 patients' records) generated a list of 927 patients with probability scores (ranging for 50% - 100%) and other selected characteristics. The predictive model probability score was an indicator measuring the likelihood of every selected patient to be a high user of health services (physician visits, labs and exams or hospitalizations and ED visits).

6.4 Tool development criteria

6.4.1 Top conditions (most prevalent conditions)

A key goal for this study was to design a management (or assessment) and referral process, which could help clinicians in primary care settings by providing them with an appropriate course of action to handle patients with multiple comorbidities.

The predictive modeling approach used to identify patients with complex medical needs who could benefit from a referral to a specialized care team such as Health Links is the required first step in dealing with issue. To help clinicians identify, on an ongoing basis, patients with complex medical needs, it is necessary to design a process that aligns with the capabilities of the EMR and their workflow.

Current limitations with EMRs could not allow for a deployment of a predictive model to identify patients with complex medical needs. However, the predictive model provides some criteria that could be used to create tools that are easily implementable in the EMR and align well with clinicians' workflow.

The NCD-based model – using neural network – indicated that 7 out of the 8 medical conditions of this study has the predictive power to identify patients with complex medical needs. The medical conditions with a predictive score greater than 70 are Parkinson's disease, epilepsy, dementia and COPD. These four conditions are not the most common 4 conditions in both the study population as well as in the 4+ patients' population as shown in the Table 6-6.

To define and help identify a potential "Health Link" patient, the MOHLTC proposed the criterion of "4+ patient" which means a patient who has four or more chronic conditions(10).

In the author's previous work, the analysis of data on "4+ patient" population showed that certain conditions could be good predictors of a "4+ patient". We therefore define a "4+ condition" as a condition that is associated with or is likely to be a good predictor of 3 or more conditions.

Table 6-10: Comparison of prevalence of medical conditions between the study population and the 4+ patients' population

Study population

Medical Condition	Number of patients	Ranking	Proportion (%) N=108054
Hypertension	102123	1	95%
Osteoarthritis	55096	2	51%
Depression	45040	3	42%
Diabetes Mellitus	43963	4	41%
COPD	21155	5	20%
Dementia	15370	6	14%
Epilepsy	3849	7	4%
Parkinson	2439	8	2%

4+ patients' population

Medical Condition	Number of patients	Ranking	Proportion (%) N=4927
Hypertension	4558	1	93%
Osteoarthritis	3937	2	80%
Depression	3599	3	73%
Diabetes Mellitus	3486	4	71%
COPD	2479	5	50%
Dementia	1814	6	37%
Epilepsy	424	7	9%
Parkinson	369	8	7%

Chronic obstructive pulmonary disease (COPD)

Diseases listed in green are the less prevalent present in 50% or less of the patients' population as opposed to the most prevalent diseases present in more than 50% of the patients' population

More than 40% of the study population and more than 70% of the 4+ patient population have the top four medical conditions. These medical conditions are expected to be predicted by the bottom three conditions Parkinson, Epilepsy and Dementia. Patients with COPD account for 50% of the 4+ patients' population. Nevertheless, COPD could still predict the top 4 medical conditions, the most common of the 8 conditions studied.

The following section presents the results of the association analysis mapping the predictive power of the bottom 4 medical conditions and their interactions with the top four medical conditions.

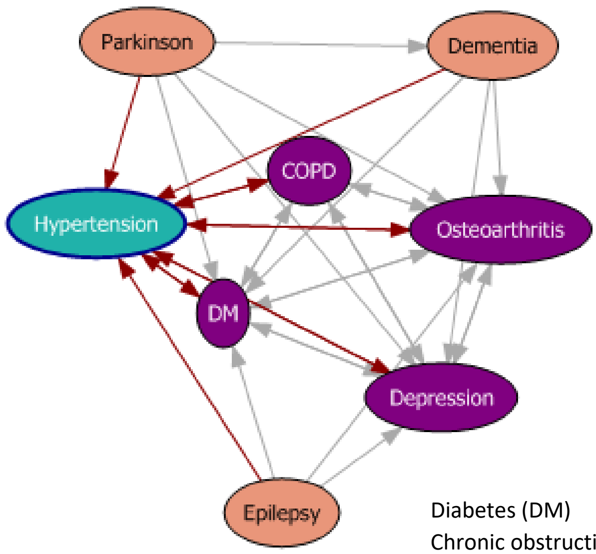
6.4.2 Association analysis: mapping the predictive power of "4+ conditions"

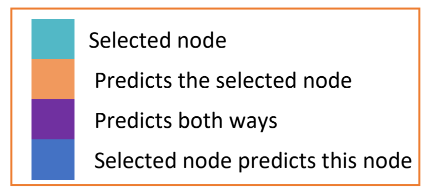
To generate the set of criteria to be used for tool development, we used association analysis which helps understand the relationships between the medical conditions under consideration and determine those medical conditions with greater predictive power.

Mapping the interactions of the top four medical conditions with other medical conditions

Figures 6-9 to 6-12 depict the relationships between the 8 medical conditions included in the association analysis model. The first 4 profiles looking at the interactions of the top 4 medical conditions with the bottom 4 conditions showed that each of these 4 medical conditions (hypertension, osteoarthritis, depression and diabetes) and COPD are closely related as they can predict the presence of each other. The four maps (Figure 6-9 to 6-12) reveal that the bottom three conditions (Parkinson's disease, epilepsy and dementia) are very good predictors for the top 4 medical conditions (hypertension, osteoarthritis, depression and diabetes). This is an indication that these predicted medical conditions are found in high to very high proportions in patients with Parkinson's disease, epilepsy and dementia.

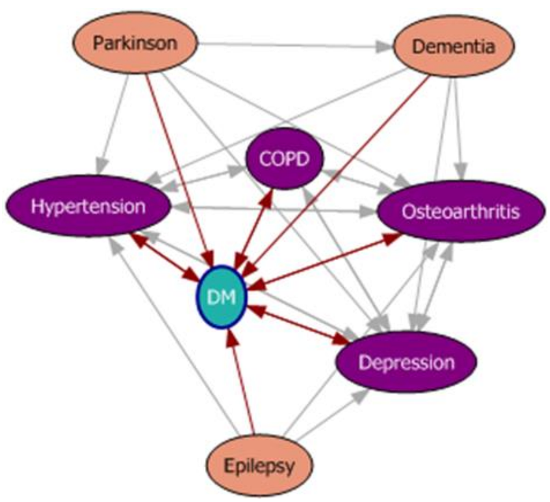
Figure 6-5: Hypertension's interactions with the other 7 conditions

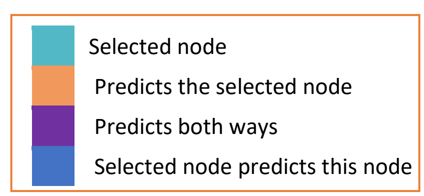




Diabetes (DM)
Chronic obstructive pulmonary disease (COPD)

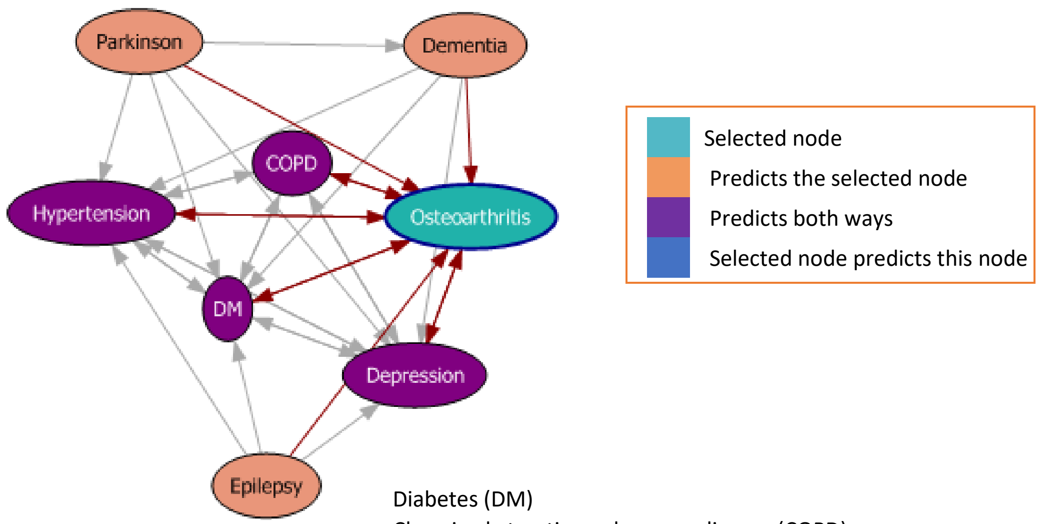
Figure 6-6: Diabetes' interactions with the other 7 conditions





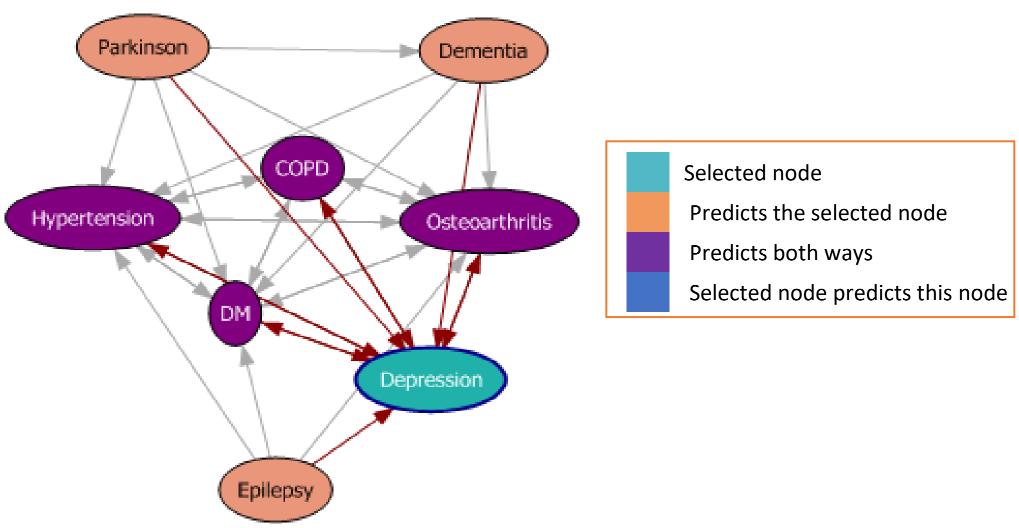
Diabetes (DM)
Chronic obstructive pulmonary disease (COPD)

Figure 6-7: Osteoarthritis' interactions with the other 7 conditions



Diabetes (DM)
Chronic obstructive pulmonary disease (COPD)

Figure 6-8: Depression's interactions with the other 7 conditions



Diabetes (DM)
Chronic obstructive pulmonary disease (COPD)

Next we looked at the bottom 4 medical conditions and their interactions with the top 4 medical conditions.

COPD predictive profile

Figure 6-9: COPD's interactions with the other 7 conditions

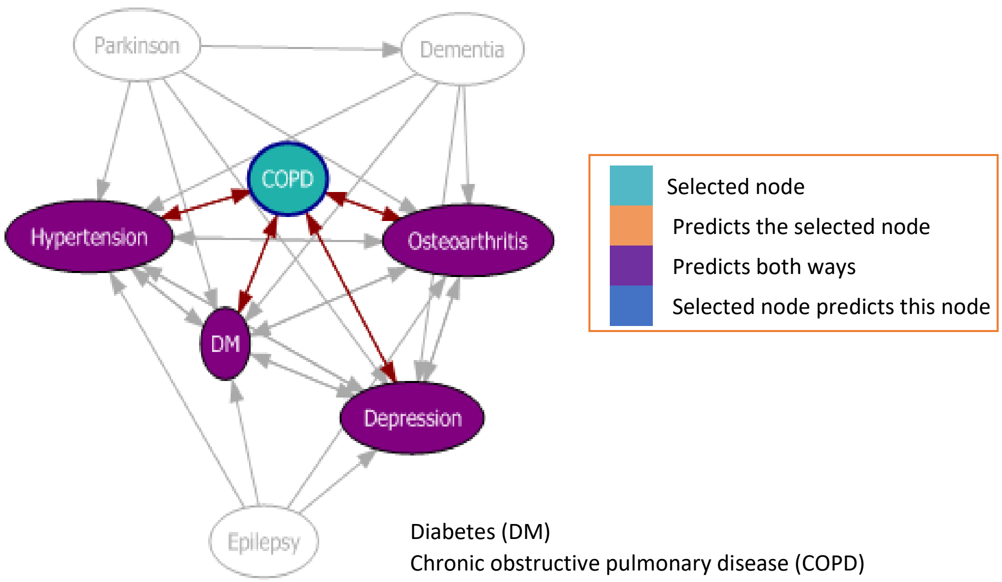
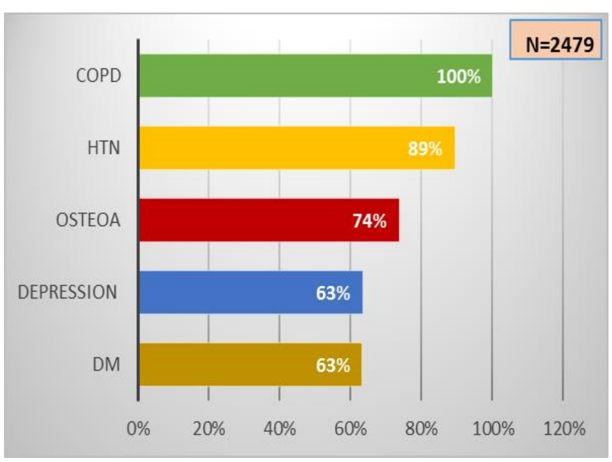


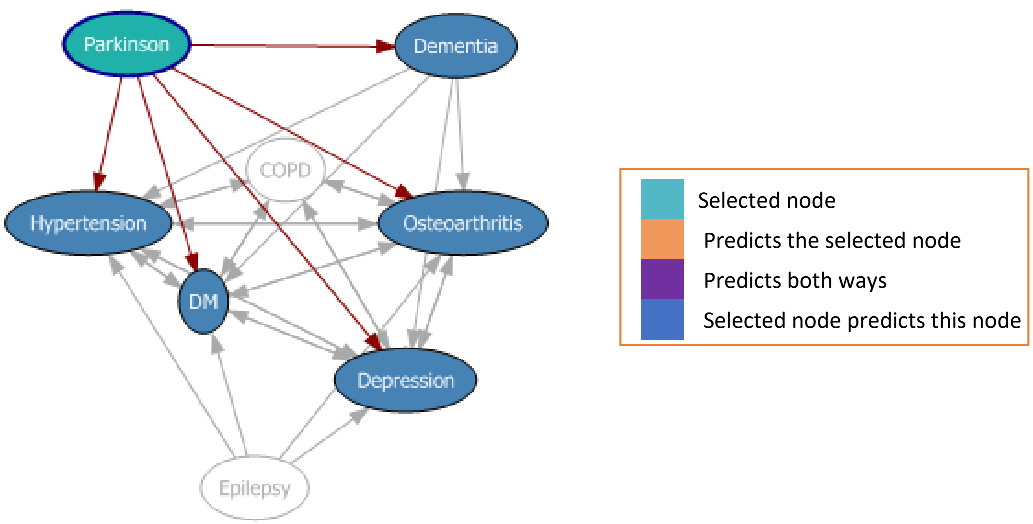
Figure 6-10: Co-occurrence of conditions among 4+ patients with COPD



Contrary to the top 4 medical conditions, the bottom 3 medical conditions do not predict COPD. This is an indication of a difference between COPD and the top 4 medical conditions. However, the COPD map (Figure 6-12) shows that COPD tends be closely related to the top 4 medical conditions. COPD predicts each of these conditions and they in turn predict COPD. COPD predicting the top 4 medical conditions is an indication that, each of these 4 medical conditions are found in high to very high proportions in patients with COPD. More than 60% of patients with COPD also have all the 4 medical conditions within the 4+ patient population. More than 70% of patients also have 2 of the top 4 medical conditions (hypertension and osteoarthritis) (Figure 6-13).

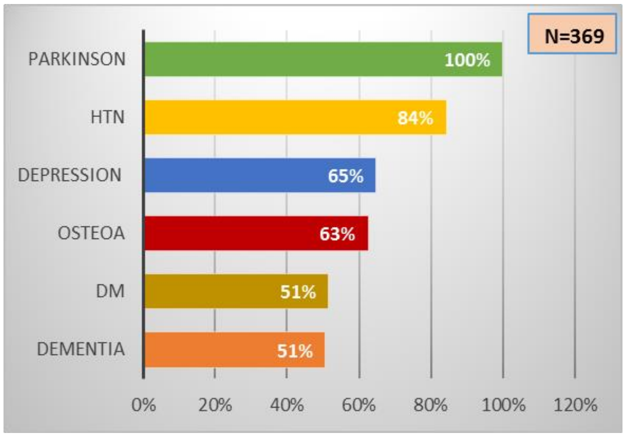
Parkinson's disease predictive profile

Figure 6-11: Parkinson's disease interactions with the other 7 conditions



Diabetes (DM)
Chronic obstructive pulmonary disease (COPD)

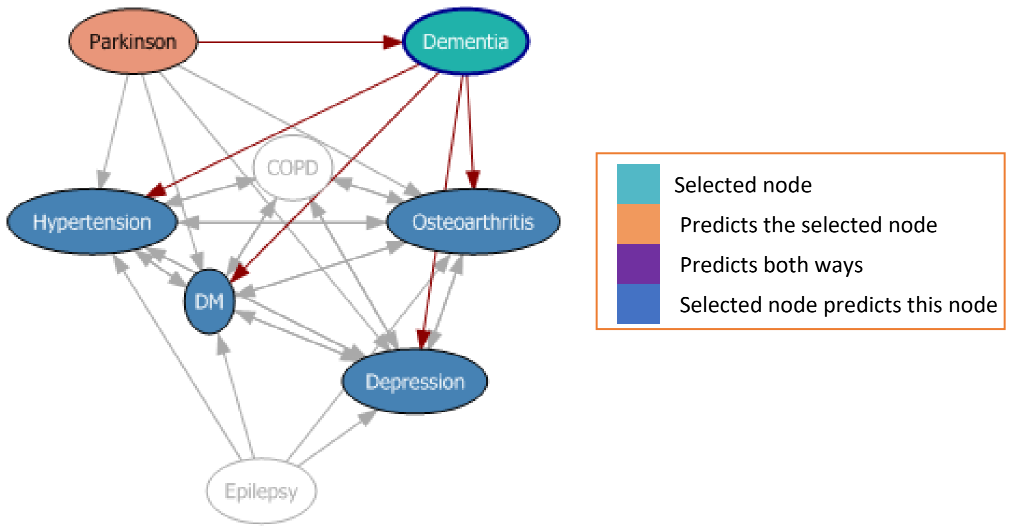
Figure 6-12: Co-occurrence of conditions among 4+ patients with Parkinson



Parkinson's disease Predictive Profile (Figure 6-14) shows that this medical condition predicts all the top 4 medical conditions and one of the bottom 4 medical conditions (dementia). All the 5 medical conditions predicted by Parkinson were found in high to very high proportions among patients with Parkinson's disease. More than 50% of patients with Parkinson's disease also have the 5 medical conditions shown in Figure 6-15. More than 60% of patients with Parkinson's disease also had 3 conditions (osteoarthritis, depression and hypertension). More than 80% of patients with Parkinson's disease within the 4+ patients' population also have hypertension.

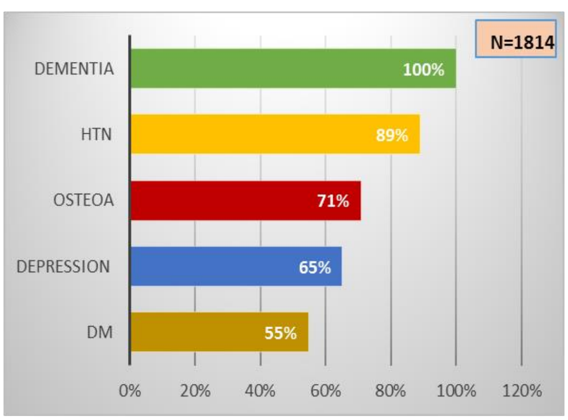
Dementia's predictive profile

Figure 6-13: Dementia's interactions with the other 7 conditions



Diabetes (DM)
Chronic obstructive pulmonary disease (COPD)

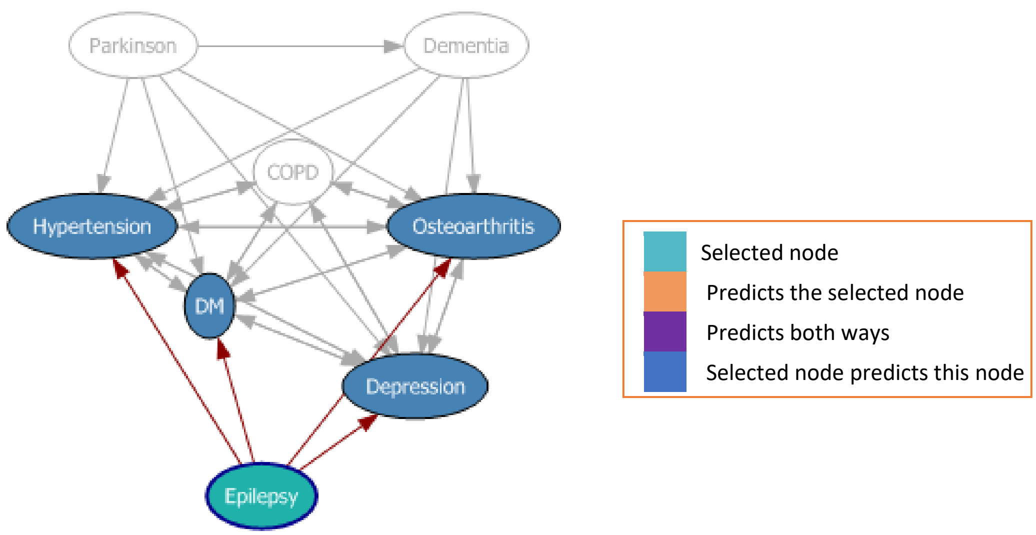
Figure 6-14: Co-occurrence of conditions among 4+ patients with Dementia



Dementia predicts the top four medical conditions (Figure 6-17), which as shown in Figure 6-18, and are found in high to very high proportion in patients with dementia. More than 50% of patients with dementia within the 4+ patients' population also have the top four medical conditions. More than 70% of patients with dementia also have hypertension and osteoarthritis.

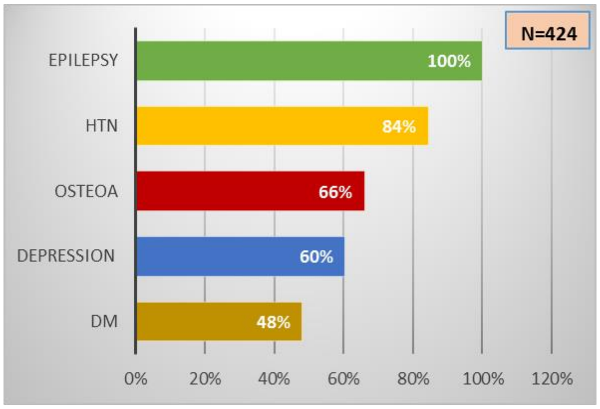
Epilepsy's predictive profile

Figure 6-15: Epilepsy's interactions with the other 7 conditions



Diabetes (DM)
Chronic obstructive pulmonary disease (COPD)

Figure 6-16: Co-occurrence of conditions among 4+ patients with Epilepsy



As seen for Parkinson's disease and dementia predictive profile, epilepsy predicts the top 4 medical conditions (Figure 6-19). These 4 medical conditions are therefore found in high to very high proportions among patients with epilepsy within the 4+ patients' population. Almost 50% of patients with epilepsy also have all the top 4 medical conditions (Figure 6-20). More than 80% of patients with epilepsy also have hypertension (Figure 6-20).

Parkinson's disease, dementia and epilepsy stood out as very good predictors of the top four medical conditions as shown in the analysis of their respective predictive profile. COPD, could be considered as a good predictor of the top 4 medical conditions even though each these 4 conditions can also predict COPD. Each of the bottom four conditions can then be considered as a "4+ condition" which means a medical condition that can predict a 4+ patient. These 4 conditions could make the set of criteria that could be used for building the different tools required for managing and referring potential 4+ patients.

6.4.3 Generating search criteria

Based on the findings of the predictive modeling refined by the association analysis, the generic structure of the search to be built and implemented in the EMR is depicted in ?

Figure 6-17: Generic structure of search criteria

Without the underlying conditions	With the underlying conditions
Parkinson	**Parkinson**
OR	OR
Dementia	**Dementia**
OR	OR
Epilepsy	**Epilepsy**
AND	AND
COPD	**COPD**
	AND
	HTN
	OR
	Osteoarthritis
	OR
	Depression
	OR
	Diabetes

The top 4 conditions (hypertension, osteoarthritis, depression and diabetes) being the most prevalent conditions among the 4+ patients' population which are also predicted by the bottom 4 conditions are considered underlying conditions. Thus the proposed search criteria are presented in 2 versions: with and without the underlying conditions.

6.4.4 Proposed tools for implementation

There are several approaches to turn the insights gained from the predictive modeling and the association analysis into actionable tools for patients' management and referral. In Practice Solution EMRs, searches, reminders and custom forms can be used as tools to create a management and referral process for identifying and referring patients with complex medical needs.

Searches

Findings from the predictive modeling and the association analysis' results could be used as criteria to build a "*search*" tool to run on EMRs and thus help clinicians to find patients with special or complex medical needs in their roster of patients. If the search generates a list, the clinician could review those patients' records and determine whether further actions need to be taken (e.g. customized treatments, referral). In PSS EMRs, a search can be turned into a reminder and linked to referral and assessment forms to streamline the referral process for patients with complex medical needs.

Reminders and Custom forms

Reminder: A reminder is a search that triggers or prompts a clinician to act or perform an intervention. In the EMR world, a search -- like common search engines – searches and finds a patient or group of patients based on defined criteria (or search criteria). When a search turned into a reminder (to display or prompt the clinician for a specific action) is enabled in the EMR, it will insert in every patient's chart that meets the search criteria – running in the background – the specific intervention which is in this case will be linked to the complex patient's management and referral form.

Custom form: a custom form is an electronic version of any from designed to collect information on individuals (26). A complex patient management and referral form can be built to help the clinician assess the recommended patients for referral and determine whether or not the recommended patient meets the complex patient's referral criteria.

A clear referral process of patients with complex medical needs should be laid out and the different tools needed e.g. (reminders linked to different patients' management, assessment and referral forms) should be aligned with referral process all the stakeholders agree to do.

7. Discussion

The healthcare sector in north America has witnessed tremendous acceleration of health data supply with both legislative and financial levers either mandating or incentivizing the use of HIT especially EMRs(1, 2). Intense efforts of health record digitization starting in the early 2000s are materializing with the generation of enormous amounts of health data(27). This trend marks the opening of a new era of health information as powerful data mining tools are made available to turn these enormous amounts of health data into a new wealth of knowledge and insights for better decision-making (2).

To accelerate the value from using the immense amount of health data being generated, Groves et al proposed a holistic patient-centered framework that defines "*new values pathways*" that is aimed to strike the right balance between healthcare spending or cost and maximizing patient outcomes(2).

The current study proposes a tool to identify patients with 4 or more medical conditions who are likely to be high users of healthcare resources. The study objectives correspond with 2 out 5 concepts outlined in the "*new value pathways*" proposed by Groves et al: *Right Value* (balancing cost reduction and quality of care) and *Right Innovation* (using knowledge and insights derived from health data to develop new approaches to treatment or new tools or applications for better clinical decisions(2). Once implemented, the results of the tool could also address *Right Living* (build value by including patients in their own treatment plan in order to foster disease's prevention) and *Right Care* (patients get timely and adequate treatment). The key findings of this research aimed to build on the principles of the "*Right Innovation*" which benefits trickle down to the other three "*new value pathways*".

The characterization of potential high cost users of health services based on predictive analytics revealed that demographic factors such as age and sex (to some extent), medical conditions, number of chronic conditions (four or more medical conditions) and certain risk factors (e.g. smoking, obesity) are predictors of high use of health services.

Drawing from the "*Right Innovation*" concept of the "*new value pathways*", findings of this research can be turned into actionable tools that can help primary clinicians in managing and referring certain patients with special or complex medical needs to specialized teams such as Health Links(2).

The ultimate motivation of this study was to develop a management and referral tool to identify, manage and refer patients living with multiple comorbidities to specialized care teams such as Health Links. Ideally the tool proposed would be tested with both CPCSSN dataset as well as other datasets to determine its performance. However time and resources constraints and other factors (e.g. test site) have not permitted me to take this work into the design and test stages.

There are also limitations which taken into consideration could improve the outcomes of this analysis.

There are some data quality issues around some of tables used for this analysis which limited the precision of the analysis performed. For example, the encounter table did not differentiate between type of encounters putting “face-to-face visits” and “phone calls” or other types of encounters on the same footing. The same limitation is found with the risk factors table which lacks key information on factors’ definition: none of the 6 risk factors – smoking, drinking, obesity, diet, exercise and stress – used in the analysis was defined. This lack of definition is a key issue as their proper definition could have helped establish some categorizations and ultimately could have helped refine the analysis. Risk factors such as “exercise”, “drinking” or “smoking” – for instance – could have been defined with some metrics to differentiate their level of severity. This lack of definition could have diminished the predictive power of the risk factors. As a result the risk factors included in the predictive modelling did not show a major influence on health resources utilization as expected. Thus in the second round of predictive modelling using the association analysis, the risk factors were dropped for showing a weaker predictive power. The other issue around the risk factors table is the lack of standardization: more than one term were used to capture risks such as “smoking”, “drinking” and “stress”. We were able to address this lack of standardization issue by recoding these items as captured in the risk factors table.

Another major limitation is that the laboratory data used only related to tests associated with diabetes. Many more tests are used in primary care. A fuller evaluation needs to be done using multiple laboratory tests to inform tool development….?

This research ideally was aimed at using data analytics to predict the use of health services (i.e. hospitalizations and ED visits). These acute care data were not available in the research database obtained. As a result, the proxy variables of labs, exams and encounters were used instead. Note also that the laboratory data is also limited.

The small number of diseases available in the research database was also an impediment which certainly limited the accuracy of the proposed search algorithm for tool development. Previous work, by the author showed that by adding congestive Heart failure – which is missing from the list of 8 conditions used in this analysis – to the diseases under consideration has the potential to improve the proposed search algorithm and at same time reveals congestive heart failure as a “4+ condition”.

Future research steps could include the development and testing of EMR search algorithms based on criteria identified by this study in partnership with PCPs, once these algorithms are refined with better quality data. Testing the search algorithm will enable a “complex patient” reminder as an essential stage to build the proposed referral and management tool for identifying Health Links patients. There were some initiatives to build a management and referral tool for Health Links patients using hospital data (i.e. hospitalization data and ED visits data). It would be beneficial to compare the results of search algorithms based on hospital data and search algorithm based on the criteria identified by this research and other confirmatory studies. Results from a comparison study could help refine and create a more robust search algorithm for Health Links patient management and referral.

8. Conclusion

Health record digitization through the use of EMRs and the availability of powerful data mining tools to explore massive health data, discover knowledge and generate insights has brought forth a promising era of health information that could benefit both patients and health system managers.

This study has shown that insights from data mining and predictive analytics tools can start to be used to develop actionable tools to help PCPs in the management and the referral of patients with complex medical needs.

Results from the both the predictive modeling and association analysis showed that factors such as age, specific medical conditions, as well as the number of medical conditions (4 or more) could be considered to be used as criteria to develop predictive tools to identify high healthcare resource users. Such tools can be integrated into searches and reminders within EMRs, and linked to custom forms, to assist clinicians in having better knowledge of their patients with special needs and recommend some course of actions to better manage those patients.

To help PCPs prevent unnecessarily high uses of health resources (hospitalizations, ED visits etc.) by certain categories of the patient population, it would be necessary to assist them in acquiring the "*Right Innovative*" tools needed for that task.

Future research should confirm the findings of this study and expand on these innovation ideas (design and development of clinical decision tools based on insights from data mining and predictive analytics) suggested in this research and bring them to fruition through practical implementation in primary care settings. That will require a collaborative partnership between clinicians, data scientists and IT professionals as well as high quality health data. Such collaborative teams working on innovative health information projects will generate greater value creation that will benefit all actors in the healthcare system.

References

1. Hoyt RE, Yoshihashi A, Bailey NJ. Health informatics: practical guide for healthcare and information technology professionals. 5th ed. Raleigh, N.C.: Lulu.com; 2012. 534 p.
2. Groves P, Kayyali B, Knott D, Van Kuiken S. The big data revolution in healthcare: Accelerating the value of innovation.: McKinsey&Company; 2013.
3. Bodenheimer T, Berry-Millett R. Care management of patients with complex health care needs. Unites States of America: Robert Wood Johnson Foundation; 2009.
4. Kimberly E. Against the Growing Burden of Disease. Ottawa, Canada: Public Health Agency of Canada; 2010.
5. Broemeling AM, Watson DE, Prebtani F. Population Patterns of Chronic Health Conditions, Co-morbidity and Healthcare Use in canada: Implications for Policy and Practice. Healthcare Quaterly [Internet]. 2008; 11(3):[70-6 pp.].
6. Redondo-Sendino Á, Guallar-Castillón P, Banegas JR, Rodríguez-Artalejo F. Gender differences in the utilization of health-care services among the older adult population of Spain. Bio-Med Central. 2006;6(155):9.
7. Rosella LC, Fitzpatrick T, Wodchis WP, Calzavara A, Manson H, Goel VH-chcuiO, Canada:. High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. BMC Health Services Research [Internet]. 2014; 14(532):[1-13 pp.].
8. Mirolla M. The Cost of Chronic Disease in Canada. Canada: The Chronic Disease Prevention Alliance of Canada; 2004.
9. Wodchis WP, Austin P, Newman A, Corallo A, Henry D. The Concentration of Health Care Spending: Little Ado (yet) About Much (money) Institute for Clinical . Canada: Evaluative Sciences (ICES); 2012.
10. Robertson M, Bains N. Health Links Target Population. Toronot, Canada: Ministry of Health and Long-Term Care; 2015.
11. LHIN T. A Plan to Transform Local Health Care for All. Toronto: Toronto Central LIHN; 2013.
12. Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. Methods for Analyzing Health Care utilization Costs. Annual Review of Public Health. 1999;20(1):125-44.
13. Anderson JG. Health services utilization: Framework and review. Health Services Research [Internet]. 1973; 8(3):[184-99. pp.].
14. Laporte A, Nauenberg E, Shen L. Aging, social capital, and health care utilization in Canada. Health Economics, Policy and Law. 2008;3(4):393-411.
15. Canada S. Health Status 2015 [Available from: http://www.statcan.gc.ca/pub/82-229-x/2009001/status/int4-eng.htm.
16. Rumsfeld JS. Health Status and Clinical Practice. When Will They Meet? Circulation 106: 5-7 doi: [Internet]. 2002; (106):[5-7 pp.].

17. Pope GC. MEDICAL CONDITIONS, HEALTH-STATUS, AND HEALTH-SERVICES UTILIZATION. Health Services Research. 1988;22(6):857-77.
18. Information CIfH. Seniors and the Health Care System: What Is the Impact of Multiple Chronic Conditions? Ottawa; 2011.
19. Haas LRea. Assessment of Risk Stratification Methods: Identifying Patients for Care Coordination within a Medical Home. Unites States of America: Mayo Clinic; 2012. p. 1-27.
20. CPCSSN. Potential Sentinels 2015 [
21. Microsoft. Classification Matrix (Analysis Services - Data Mining): Microsoft Developer Network (MSDN); 2015 [Web article]. Available from: https://msdn.microsoft.com/en-us/library/ms174811.aspx.
22. Chao-Ying JP. Data Analysis Using SAS. California, USA: SAGE Publications, Inc; 2009. 640 p.
23. Microsoft. Lift Chart (Analysis Services - Data Mining): Microsoft Developer Network (MSDN); 2015 [Web artcile]. Available from: https://msdn.microsoft.com/en-us/library/ms175428.aspx.
24. Snider D. Explaining The SQL Server 2014 Analysis Services Data Mining Model Lift Chart: Dallas Snider; 2015 [Web article]. Available from: https://www.mssqltips.com/sqlservertip/3401/explaining-the-sql-server-2014-analysis-services-data-mining-model-lift-chart/.
25. Zhu W, Zeng N, Wang N. Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations. North East SAS Users Group [Internet]. 2010:[9 p.].
26. MD Practice Software LP. PS Suite Software v5.2**:** Encounter Assistants Training. MD Practice Software LP; 2012.
27. Collier R. National Physician Survey: EMR use at 75%. CMAJ. 2015;187(1):2.

Appendices

Table 0-1: Shapiro-Wilk normality test

Dependent Variables	Test's results
NumberOfLabs	W = 0.873, p-value < 2.2e-16
NumberOfExams	W = 0.7589, p-value < 2.2e-16
NumberOfEncounters	W = 0.5669, p-value < 2.2e-16

Table 0-2: Logistic regression output using NOE as outcome variable

```
> summary(fit2)

Call:
glm(formula = NOE ~ COPD + Parkinson + DIABETES +
Osteoarthritis +
    Dementia + Hypertension + Depression + NOEx + Sex +
AgeCategory +
    Smoking + Alcohol + Exercise + Obesity + Stress, family
= binomial(logit),
    data = trainSet)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8789  -0.9608  -0.7030   1.1782   2.0860

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.07095    0.07618 -14.058  < 2e-16 ***
COPDYES             0.50579    0.06696   7.553 4.25e-14 ***
ParkinsonYES        0.33720    0.23189   1.454 0.145914
DIABETESYES               0.15356    0.04343    3.536
0.000407 ***
OsteoarthritisYES   0.25251    0.04622   5.463 4.68e-08 ***
DementiaYES         0.32932    0.10903   3.020 0.002524 **
HypertensionYES     0.21011    0.04764   4.410 1.03e-05 ***
DepressionYES       0.41114    0.05112   8.042 8.83e-16 ***
NOExHighUse         0.74062    0.04419  16.759  < 2e-16 ***
SexWomen           -0.10197    0.51994  -0.196 0.844514
SexMen             -0.10665    0.04394  -2.427 0.015221 *
AgeCategorySenior1 -0.04730    0.06195  -0.764 0.445162
AgeCategorySenior2  0.04278    0.06721   0.637 0.524448
AgeCategorySenior3  0.10898    0.07244   1.504 0.132465
AgeCategorySenior4  0.17491    0.06636   2.636 0.008393 **
SmokingYES          0.20657    0.05406   3.821 0.000133 ***
AlcoholYES         -0.17495    0.05081  -3.443 0.000575 ***
```

```
ExerciseYES          0.07493    0.04786   1.566 0.117433    
ObesityYES          -0.70259    0.05109 -13.753  < 2e-16 ***
StressYES            0.66006    0.18322   3.603 0.000315 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13501  on 10085  degrees of freedom
Residual deviance: 12642  on 10066  degrees of freedom
AIC: 12682

Number of Fisher Scoring iterations: 4
```

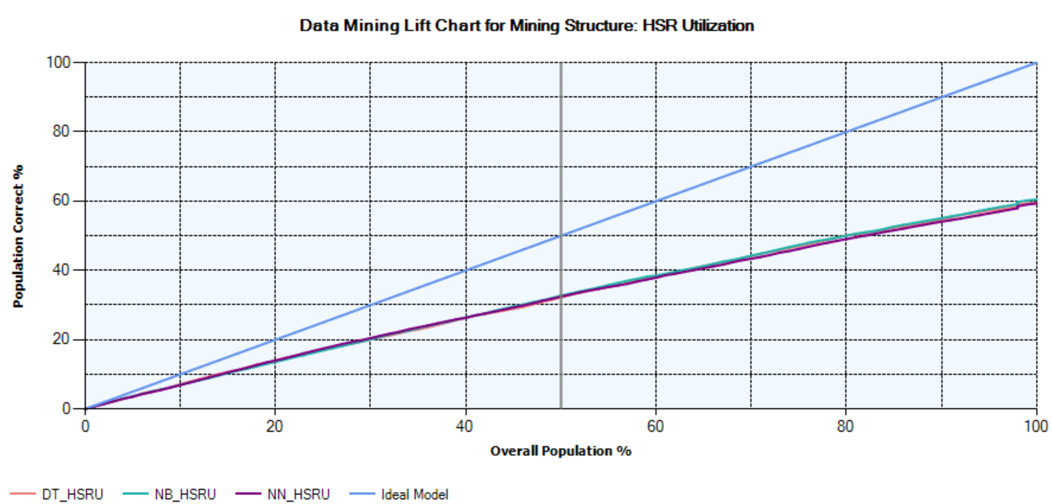


Figure 0-1: Lift Chart for NOL-based models

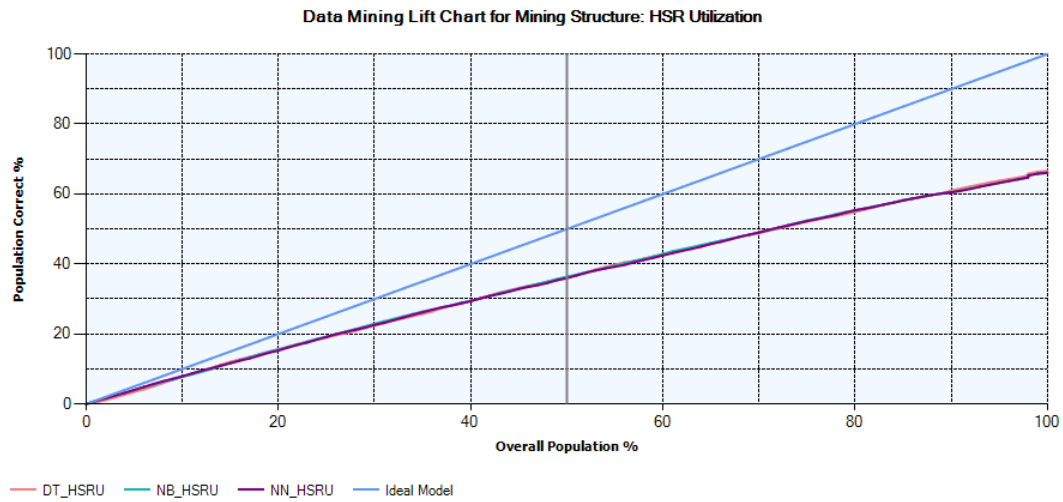


Figure 0-2: Lift Chart for NOEx-based models

Table 0-3: Accuracy of selected models for predicting encounter

Counts for Decision Tree model for encounter

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	661	519
BasicUse	1040	2102

Model's Performance Measures

Sensitivity	Specificity	Accuracy
0.39	0.80	0.64

Counts for Naïve Bayes model for encounter

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	611	463
BasicUse	1090	2158

Model's Performance Measures

Sensitivity	Specificity	Accuracy
0.36	0.82	0.64

Counts for Neural Network model for encounter

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	690	543
BasicUse	1011	2078

Model's Performance Measures

Sensitivity	Specificity	Accuracy
0.41	0.79	0.64

Table 0-4: Accuracy of models selected for predicting exams

Counts for Decision Tree model for exams

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	**216**	119
BasicUse	1323	2664

Models' Performance Measures

Sensitivity	Specificity	Accuracy
0.14	0.96	0.67

Counts for Naïve Bayes model for exams

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	**468**	398
BasicUse	1071	2385

Models' Performance Measures

Sensitivity	Specificity	Accuracy
0.30	0.86	0.66

Counts for Neural Network model for dxams

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	**341**	268
BasicUse	1198	2515

Models' Performance Measures

Sensitivity	Specificity	Accuracy
0.22	0.90	0.66

Table 0-5: Accuracy of models selected for predicting labs

Counts for Decision Tree model for labs

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	**432**	291
BasicUse	1457	2142

Models' Performance Measures

Sensitivity	Specificity	Accuracy
0.23	0.88	0.60

Counts for Naïve Bayes model for labs

Predicted	HighUse (Actual)	BasicUse (Actual)
HighUse	**813**	633
BasicUse	1076	1800

Sensitivity	Specificity	Accuracy
0.43	0.74	0.60

Counts for Neural Network model for labs

Predicted	HighUse (Actual)	BasicUse (Actual)

Sensitivity	Specificity	Accuracy

HighUse	596	464				
BasicUse	1293	1969		0.32	0.81	0.59

Table 0-6: List of the type of encounters captured in the CPCSSN database

Type of Encounter
E-mail
ER Visit
Home visit
Hospitalization
Nursing Home Visit
Phone
Primary Care - Academic
Primary Care - Community
Primary Care Clinic

Table 0-7: List of exams captures in the CPCSSN database

List of exams
BMI
BMI (kg/m^2)
Diastolic BP
Foot Exam
GLUCOSE
Glucose meter
Head Circ.
Hearing
Height (cm)
Joint pain
O2Sat
PEFR (L/min)
Pulse
resp
sBP (mmHg)
Temperature
URINALYSIS
Vision

Waist
Waist Circumference
Waist Circumference (cm)
Waist Hip ratio
Weight (kg)

Table 0-8: List of labs captures in the CPCSSN database

List of Labs
GLUCOSE-FASTING
Glucose FAST 'g Pl (PST)
A1C
Albumin
ALBUMIN (R U)
ALBUMIN CREATININE RATIO U
Albumin/Creatinine
ALBUMIN/CREATININE RATIO
Albumin/Creatinine Ratio; Urine
ALBUMIN/CREATININE:RATIO:PT:URINE:QN
ALBUMIN/CREATININE:RATIO:PT:URINE:QN:DETECTION LIMIT = 20 MG/L
Calculated GFR
CHOLESTEROL
Cholesterol Fasting
Cholesterol HDL
Cholesterol In HDL
Cholesterol In LDL; Calculated
Cholesterol LDL Calc
Cholesterol Plasma
Cholesterol Random
Cholesterol Reference Range Note
Cholesterol Serum
CHOLESTEROL.IN HDL:SCNC:PT:SER/PLAS:QN
CHOLESTEROL.IN LDL:SCNC:PT:SER/PLAS:QN:CALCULATED
CHOLESTEROL:SCNC:PT:SER/PLAS:QN
Creatinine
FAST.BL.GLUCOSE
FASTING BLOOD GLUCOSE
FASTING CHOLESTEROL

Fasting Glucose
FASTING SERUM GLUCOSE
FBS
GFR, Calculated
GFR, Estimated
GLU.TOL.TEST
GLUCOSE
GLUCOSE - FASTING
glucose (fasting)
Glucose AC
Glucose Fasting
GLUCOSE FASTING-SER
GLUCOSE PLASMA FASTING
GLUCOSE PRE CHALLENGE
GLUCOSE SERUM FASTING
GLUCOSE TOLERANCE - 2.0h
GLUCOSE TOLERANCE - FAST
GLUCOSE TOLERANCE 2.0H.
Glucose Tolerance 75g
GLUCOSE TOLERANCE FASTING
GLUCOSE TOLERANCE TEST
GLUCOSE TOLERANCE:DOSAGE
GLUCOSE, FASTING
GLUCOSE-FASTING
Glycohemoglobin(HbA1c)
Haemoglobin - glycosylated: A1C
Hb A1C
HbA1C
HBA1C YCAHB
HDL
HDL CHOL
HDL cholesterol
HDL Cholesterol Pl
HDL Cholesterol Ser
HDL Reference Range Note
HDLC
HDLC_Cholestérol-HDL
HDL-Cholesterol

Hemoglobin A1c
HG A1C
HgA1C
Hgb A1c
High Density Lipoprotein Cholesterol
LDL
LDL (Sans résultat) 1238Q
LDL CALCULATED
LDL CHOL
LDL Chol Pl Calc
LDL Chol Ser Calc
LDL Cholesterol
LDL CHOLESTEROL CALC.
LDL CHOLESTEROL DIRECT
LDL CHOLESTEROL(CALCULATED)
LDL Direct
LDL Reference Range Note
LDL, CALCULATED
LDLC_Cholestérol-LDL
Low Density Lipoprotein Cholesterol
Microalb Daily Exc
Microalb Excr Rate
Microalb/creat
MICROALB/CREAT RATIO
Microalb/Creat Ratio RDiabetes
MICROALB/CREAT.
Microalbumin
MICROALBUMIN (24H U)
MICROALBUMIN (RDIABETES U)
MICROALBUMIN (RUR)
MICROALBUMIN R U
Microalbumin Random
Microalbumin to Creatinine, Urine
MICROALBUMIN, 24 HOUR URINE
Microalbumin, conc.
MICROALBUMIN, URINE
Microalbumin,ur
MICROALBUMIN,URINE

MICROALBUMIN,URINE (RANDOM)
MICROALBUMIN,URINE RANDOM
MICROALBUMIN/CREATININE RATIO
MICROALBUMIN/CREATININE RATIO, URINE
RANDOM CHOLESTEROL
TG
Thyroglobulin (TG) Ser
Total Cholesterol
Triglyceride
Triglyceride Reference Range Note
TRIGLYCERIDE:SCNC:PT:SER/PLAS:QN
Triglycerides
TRIGLYCERIDES LEVEL
Triglycerides Plasma
Triglycerides Serum
uALB/CR/RATIO
UR 24 MICROALBUMIN
UR.MICROALB./L
UR.MICROALB/d
UR.MICROALB/L
Urine ACR (Albumin/Creatinine Ratio)
Urine Albumin Creatinine ratio
URINE ALBUMIN/CREATININE RATIO
URINE CHEMISTRY:MICROALB/CREAT.
URINE CHEMISTRY:MICROALBUMIN (RUR)
URINE CHEMISTRY:MICROALBUMIN (UR)
Urine Microalbumin
Urine Microalbumin/Creatinine Ratio