PREDICTIVE DRUG PRESCRIPTION

By

Austin Gnanaraj, Yiyun Liang

December, 2014

Table of Contents

Abstract

Data mining has been used intensively and extensively by many organizations. In healthcare, Predictive Analytics is becoming increasingly popular, if not increasingly essential. Predicting the future can greatly benefit all parties involved in the healthcare industry. For example, healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services.

In this project, we have presented a visual model classifying patients belonging to their distinguishable features with demographics and have deduced the effective drug/ medication for curing Coronary Atherosclerosis. This will be aiding doctors to choose the medication according to the prediction based on trending treatment benefits from historical analysis of reliable data. Various assumptions and calculation criteria's have been set to arrive at the solution and they are elaborated. Limitations and hurdles faced in the process have also been documented. Map-reduce has been used in Data Transformation, data cleaning and Tableau was used in creating the reports.

Introduction

Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting patient encounters, diagnosis, lab results, courses of treatments with a history of drugs prescribed. We can data mine the extensive data to deliver a trend analysis of which course of action would prove effective. For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatment worked best and are most cost-effective.

Our project goal is to process the raw medical data, using data crunching techniques such as map-reduce and provide a visual representation that supports the doctor's treatment decisions for the Healthcare Industry. The output from the project will enable doctors to choose between treatment patterns based on the patient's disease, race, country and so on. Analysing the data, we will form treatment patterns that will contain the medicines prescribed, the lab results and the effectiveness of the treatment as its parameters.

The performance indicator (PI) is that whether the cure worked for a certain set of patients with similar features, and the key performance indicator (KPI) is the number of "cured" patients during the period of taking the medicine. Cured infers to patients who after taking some medicines came to a healthy state.

About Atherosclerosis

Atherosclerosis

Hardening of the arteries, also called atherosclerosis, is a common disorder. It occurs when fat, cholesterol, and other substances build up on the walls of arteries and form hard structures called plaques. Over time, these plaques can block the arteries and cause problems throughout the body.

Atherosclerosis Implications

Plaque may partially or totally block the blood flow through an artery in the heart, brain, pelvis, legs, arms or kidneys. Some of the diseases that may develop as a result of atherosclerosis include coronary heart disease, angina (chest pain), carotid artery disease, peripheral artery disease (PAD) and chronic kidney disease.

Two things that can happen where plaque occurs are:

(1) A piece of the plaque may break off.

(2) A blood clot (thrombus) may form on the plaque's surface.

If either of these occurs and blocks the artery, it may result in a heart attack or stroke. Atherosclerosis affects large and medium-sized arteries. The type of artery affected and where the plaque develops varies with each person.

Atherosclerosis is a slow, progressive disease that may start in childhood. In some people the disease progresses rapidly in their 30s. In others, it doesn't become dangerous until they reach their 50s or 60s. However, it is normal to have some hardening of the arteries as you get older.

Why Atherosclerosis

(1) Coronary artery disease is the No. 1 killer of Americans. Most of these deaths are from heart attacks caused by sudden blood clots in the heart's arteries.
(2) Good indicators (cholesterol, LDL, HDL, Triglycerides) to show the effectiveness of the medicines in patients.
(3) People between different age groups suffer from this disease.
(4) Coronary heart disease alone costs the United States $108.9 billion each year. 3 This total includes the cost of health care services, medications, and lost productivity.

Medicines in Use

Crestor

Crestor is in a group of drugs called HMG CoA reeducates inhibitors, or "statins." The study: "Rosuvastatin: a review of its effect on atherosclerosis" shows rosuvastatin(Crestor) delayed the progression of carotid atherosclerosis in patients with subclinical carotid atherosclerosis, moderately elevated cholesterol levels, and a low risk of cardiovascular disease in a primary prevention trial.

Crestor is used to lower cholesterol and triglycerides (types of fat) in the blood, and it is also used to lower the risk of stroke, heart attack, and other heart complications in people with diabetes, coronary heart disease, or other risk factors. In addition, Crestor is used in adults and children who are at least 10 years old and may also be used for other purposes not listed in this medication guide.

Lipitor

Lipitor (atorvastatin) belongs to a group of drugs called HMG CoA reductase inhibitors, or "statins." Lipitor reduces levels of "bad" cholesterol (low-density lipoprotein, or LDL) and triglycerides in the blood, while increasing levels of "good" cholesterol (high-density

lipoprotein, or HDL).

Lipitor is used to treat high cholesterol, and to lower the risk of stroke, heart attack, or other heart complications in people with type 2 diabetes, coronary heart disease, or other risk factors. Lipitor is used in adults and children who are at least 10 years old.

The study: *"Effect of intensive atorvastatin therapy on coronary atherosclerosis progression, composition, arterial remodelling, and microvascular function*" shows with moderate coronary artery disease, high-dose atorvastatin(Lipitor) resulted in alterations in coronary atheroma composition with corresponding changes in plaque phenotype and modest improvement in coronary microvascular function.

Zocor

Zocor is an HMG-CoA reductase inhibitor, also known as a "statin." Zocor (Zocar) is used to lower high cholesterol and triglycerides in certain patients. Lowering your cholesterol can help prevent heart disease and hardening of the arteries, conditions that can lead to heart attack, stroke, and vascular disease. Zocor (Zocar) also increases high-density lipoprotein (HDL, "good") cholesterol levels.

Zocor (Zocar) is used along with an appropriate diet. It is used in certain patients to reduce the risk of heart attack, stroke and death due to coronary heart disease. It also can reduce the risk of chest pain caused by angina. Zocor (Zocar) is also used to reduce the need for medical procedures to open blocked blood vessels. Zocor (Zocar) may also be used for other conditions as determined by your doctor. However, Zocor (Zocar) side effects include stomach upset. It may infrequently cause muscle problems (which can rarely lead to a very serious condition called rhabdomyolysis).

Health Indicators

All LDL cholesterol, HDL cholesterol and total cholesterol determine the cholesterol level. Blood cholesterol testing evaluates total cholesterol, with high total cholesterol and LDL cholesterol and HDL cholesterol readings suggesting coronary artery disease may be present.

LDL cholesterol

LDL is the bad cholesterol. LDL collects in the walls of blood vessels, causing the blockages of atherosclerosis. Higher LDL levels put you at greater risk for a heart attack from a sudden blood clot in an artery narrowed by atherosclerosis.

An LDL particle is a microscopic blob consisting of an outer rim of lipoprotein surrounding a cholesterol center. LDL is called low-density lipoprotein because LDL particles tend to be less dense than other kinds of cholesterol particles.

Different levels of the LDL cholesterol are shown in the table below.

LDL cholesterol (U.S. and some other countries)	LDL cholesterol (Canada and most of Europe)	
Below 70 mg/dL	Below 1.8 mmol/L	Ideal for people at very high risk of heart disease
Below 100 mg/dL	Below 2.6 mmol/L	Ideal for people at risk of heart disease
100-129 mg/dL	2.6-3.3 mmol/L	Near ideal
130-159 mg/dL	3.4-4.1 mmol/L	Borderline high
160-189 mg/dL	4.1-4.9 mmol/L	High
190 mg/dL and above	Above 4.9 mmol/L	Very high

HDL cholesterol

HDL cholesterol is the well-behaved "good cholesterol." This friendly scavenger cruises the bloodstream. As it does, it removes harmful bad cholesterol from where it doesn't belong. High HDL levels reduce the risk for heart disease -- but low levels increase the risk.

Different levels of the HDL cholesterol are shown in the table below.

HDL cholesterol (U.S. and some other countries)	HDL cholesterol (Canada and most of Europe)	
Below 40 mg/dL (men) Below 50 mg/dL (women)	Below 1 mmol/L (men) Below 1.3 mmol/L (women)	Poor
40-49 mg/dL (men) 50-59 mg/dL (women)	1-1.3 mmol/L (men) 1.3-1.5 mmol/L (women)	Better
60 mg/dL and above	1.6 mmol/L and above	Best

Total cholesterol

Different levels of the total cholesterol are shown in the table below.

Total cholesterol (U.S. and some other countries)	Total cholesterol* (Canada and most of Europe)	
Below 200 mg/dL	Below 5.2 mmol/L	Desirable
200-239 mg/dL	5.2-6.2 mmol/L	Borderline high
240 mg/dL and above	Above 6.2 mmol/L	High

Triglycerides

The role of triglyceride in the atherogenesis are largely unknown. Triglyceride-Rich Lipoprotein Remnant Particles and Risk of Atherosclerosis suggest that triglyceride-rich lipoprotein reduction by fabric acid derivatives, results in the reduction of atherosclerosis progression.

Different levels of the triglycerides are shown in the table below.

Triglycerides (U.S. and some other countries)	Triglycerides (Canada and most of Europe)	
Below 150 mg/dL	Below 1.7 mmol/L	Desirable
150-199 mg/dL	1.7-2.2 mmol/L	Borderline high
200-499 mg/dL	2.3-5.6 mmol/L	High
500 mg/dL and above	Above 5.6 mmol/L and above	Very high

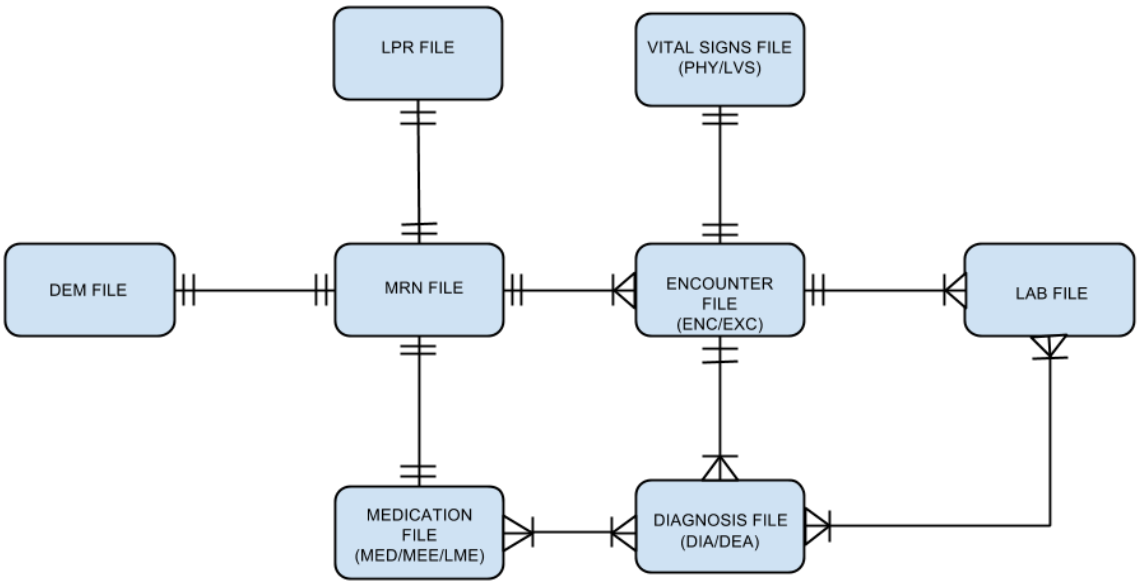
Data Description

Data File Description

File Name	**Description**
DEM File	Includes demographic info of the patients. ENC File: Encounters and diagnoses made for those encounters.
EXC File	Exclusive encounters and diagnoses made for those encounters.
MED File	Information of the medications given to the patients (including vaccinations)
MEE File	Information of the exclusive medications given to the patients (including vaccinations)
DEA File	Exclusive diagnosis with ICD9 codes
LAB File	Blood test (cholesterol, triglyceride, etc.) results of patients
DIA File	Diagnosis with ICD9 codes

ER Diagram

The entity relationship diagram (ERD) is shown in the picture below.



Data Integration

Extracted data from individual tables and loaded into tables. Below are the examples:

Table 1: DEA + Table 2: DIA = Table 3: DIAGNOSIS

Table 2: ENC + Table 2: EXC = Table 3: ENCOUNTERS

Data mining filters

(1) We filter data on disease by matching disease Name with ‘coronary atherosclerosis.

(2) We filter drugs by matching Medications Prescribed with any of ‘Lipitor’, ‘crestor’ or ‘zocor’.

(3) There are three criteria for filtering patients: a. Infected by the disease; b. Lab Results after diagnosis of the disease; c. Demographic details.

Data Limitations

(1) Gender: Data contains only Female patients

(2) Diversity of Race: more than 80% of the patients were white American

(3) The set of drugs that could be prescribed for coronary atherosclerosis to patients

(4) The list of diagnostic tests that may be prescribed to the patient for diagnosing coronary atherosclerosis

(5) The data doesn't specify the medication end date of the patients.

Project Assumptions

(1) A Patient can only be under one medicine at a time period, when she starts with another medicine, the start date of the later would mean the end date of the former intake. The last medication would have an end date till today.

(2) If the lab test results shows healthy levels of test parameter ranges we consider the patient to be healthy.

Performance Indicators

- Number of Patients getting cured based on suggestive charts

Key Performance Indicators

- Number of Patients getting cured in a week
- Number of Patients getting cured in a month
- Number of Patients getting cured in a quarter

Data Transformation Using Map Reduce

Medication End Date

Based on the assumption, a Mapreduce program was created with cloudera quickstart vm to create a new column in the medication table called “medication_end_date”.

We are conclusively deciding a medication end date based on,

1. Patient’s health indicators after tests fall within the healthy range.
2. Patient discontinues treatment abruptly.

Strategy:

- Read the data and recognise the medication prescribed for a particular patient with start and end dates
- Conclude a medication’s end date based on,
    - If the patient is prescribed a new medicine then the old medication’s end date is the start of the new medication’s start date
    - If the patient died during the medication the death date is the end date
    - If the medication date is greater than today or unspecified then today is the end date
- Comparison parameters
    - Patient ID
    - Medicine

Description:

At first, we got data come in line by line, for the first line, we saved the record and moved to the next line. Then, for the next line, we compare the patient ID and medicine with the

previous record. If it's the same patient ID and different medicine, then save the next medicine date as the previous end date. If patient ID is different, then find out if the precious record have a patient death date, if there is, then the medicine end date will be the death date, if there isn't, then today will be the end date of the medicine.

Finally, using this process will miss the last record, so we will add it by hand.

And there goes the medication_end_date column.

Calculate lab result Normality

Strategy:

- Read the data and recognise whether the patient's medication has worked based off the healthy levels
- Create a column that indicates this status with a Y or N
- Comparison parameters
    - TestID
    - Patient ID

Description:

Based on the assumption, a Mapreduce to calculate the health status of the patient for the lab result. At first, Map will read the date line by line. Then, get the name of the test by Test_ID. Finally, for the test, we compare the test result with the normal range, output "y" or "n" indicate the patient is normal or not.

Consolidation of Data

Strategy:

- Read the data and consolidate the last four test results
- Create a column that indicates this status with a Cured or Not Cured
    - If the last four tests results are Y then the patient is believed to be "Cured"
    - If the anyone of the last four results fail to make a Y then the patient is still in the recovery state and hence "Not Cured"
- Comparison parameters
    - Test ID
    - Patient ID
    - Test health status (Y or N)

Description:

Based on the assumption, Mapreduce logic calculates the status of the patient based on each lab test result of patient. First, gets the first record for each kind of test.

Then, saves the latest 10 records for each test. Finally, if the first result shows normal, we don't consider them. If the first result shows abnormal, then get the 10 latest record of the test. If the patient's test show she is getting better (last four records are Y), then it is a positive record, and we save it in a new column.

Data Reporting

The reports were developed in Tableau. The report's Data Definition is as follows

Dimensions:

Row 1: Age Group (in denomination of 10's from 10 to 100)

Row 2: Race

Row 3: Marital Status

Row 4: Language

Row 5: Is Veteran

Column1: Treatment Status (Cured / Not Cured)

Column2: Consumed Medication

Report Filter: Treatment Status = "Cured"

The report will provide the frame work required for the doctors to choose medications based on age group, race, marital Status, Language, Is a Veteran for Coronary atherosclerosis.

Conclusion

We have developed a predictive model with real medical information from a health organisation to predict the best drug to be prescribed to classified patients. The model has to be tested with real time data to measure its accuracy, which will be its future course.

References

1. http://www.webmd.com/heart-disease/atherosclerosis-and-coronary-artery-disease
2. http://www.himss.org/files/himssorg/content/files/jhim/19-2/datamining.pdf
3. http://www.heart.org/HEARTORG/Conditions/Cholesterol/WhyCholesterolMatters/Atherosclerosis_UCM_305564_Article.jsp
4. http://www.drugs.com/crestor.html
5. http://www.drugs.com/lipitor.html
6. http://www.drugs.com/misspellings/zocar.htm
7. https://en.wikipedia.org