Paper 095-2010

Gastrointestinal Diseases: Diagnoses, Misdiagnoses, and Comorbidities

Pedro G. Ramos, University of Louisville, Louisville KY

Abstract

The gastrointestinal track is a complex system of organs that are exposed to external elements on a daily basis. Some of the diseases that affect the digestive system are caused by external agents; yet some of these diseases have been theoretically explained as emotional responses due to the intricate nervous network embedded in the digestive system. This paper shows how various data mining techniques and statistical methods can be employed studying this subject matter. *SAS Enterprise Guide 9.2®* was used to explore the MarketScan data set for the years 2000-2001, containing information on inpatient and outpatient doctor visits as well as the consumption of prescription drugs used by millions of Americans with health care coverage. The MarketScan database records patient specific health care utilization and expenditure for inpatient and outpatient settings as well as prescription drugs. It includes about 100 private sector payers and over 500 million claim records. Specifically, this dataset represents the medical experience of privately insured employees and their dependents for the year 2000 and 2001. *SAS Enterprise Miner 6.1®* was used to process the data and conduct the data mining and statistical processes. Cluster analysis was used to define subpopulations while link analysis and association analysis were employed to define relationships between diseases and medications. Similarly, predictive models were designed to classify patients and predict how gastrointestinal diseases may affect them. In addition, by implementing categorical data analysis, risk factors were determined for several diseases such as gastric and colon cancer.

Introduction

The purpose of this data-mining project is to describe how data mining methods and statistical techniques can be used in the epidemiological study of gastrointestinal cancer, specifically colorectal cancer. Various, both prospective and retrospective, observational studies on this subject indicate the existence of an association between some gene variation and these types of cancer; however, such information is almost nonexistent in insurance claim databases. Other studies have provided strong support for an association between various diseases, such as diabetes and inflammatory bowel disease, as precursors of colorectal cancer. Furthermore, other studies have linked emotional stress to various types of cancer; however, researchers have speculated that the drugs used to treat stress symptoms may be the *true* risk factors associated with colorectal cancer.

Method

The MarketScan database, for years 2000-2001, containing information on outpatient doctor visits and the consumption of prescription drugs on millions of Americans with health care coverage was used. *SAS Enterprise Miner®* was used to process the data and conduct the data mining and statistical process. Patients with a diagnosis of rectal or colon cancer in 2001 were identified by ICD9 codes: 153.X and 154.X. All diagnosis from both inpatient and outpatient visits from the previous year were identified and compressed to a text string variable for cluster and link analysis. This was achieved using the *SAS data step* to filter the enrollee tables as well as to composed as string containing all the diagnosis codes from up to fifteen DX, diagnosis, variables per visit.

```
data cancerpatients;
set patients_2001;
array dx[*] dx_;
cancer = 0;
do i=1 to 15;
  if index(dx[i], '153') > 0 or index(dx[i], '154') > 0 then cancer = 1;
end;
if cancer = 1;

proc SQL;
create table allpatients as
select t1.id, t1.dx1, t1.dx2, t1.dx3, t1.dx4, t1.dx5, t1.dx6, t1.dx6, t1.dx7,
t1.dx8, t1.dx9, t1.dx10, t1.dx11, t1.dx12, t1.dx13, t1.dx14, t1.dx15 from
patients_2000 as t1, cancerpatients as t2
where t1.id = t2.id;

proc transpose data=allpatients out=allpatientsdiagnosis prefix=dx_;
by id;
run;
```

```
data diagnosisstring;
set allpatientsdiagnosis;
length dx $ 350;
array dxx[*] dx_: ;
dx = dxx[1];
do i=2 to dim(dxx);
  if index(dx, dxx[i]) = 0 then dx = catx(' ', dx, dxx[i]);
end;
keep patid dx;
run;
```

The prescription drug record for the year 2000 was also identified using the pharmaceutical claims table to match NDC records by patient ID's. This table was populated with drug names by NDC code tables from the FDA website, freely available for download as text files. This was achieved in *SAS Enterprise Guide* by using a macro written to access the drugs' name by NDC number from the tables provided by the FDA. Similarly using *proc transpose* and *proc sort* these records were compressed by patient ID into a string variable for cluster processing and link analysis.

```
data allprescriptions;
set allprescriptions;
length ndc $ 9;
ndc = substr(ndcnum,1,9);
drop ndcnum;
run;

data prescriptions_for_patients_ndc;
merge cancerpatients (IN = diagnosis) allprescriptions (IN = prescriptions);
by patid;
if diagnosis AND prescriptions;
keep patid ndc;
run;

%obtaindrugname(prescriptions_for_patients_ndc, fda.drugnames, ndc,
prescriptionsforcancerpatients);

%MACRO ObtainDrugName(intable, nametable, byvariable, outtable);
proc sort data=&intable;
by &byvariable;
run;

proc sort data=&nametable;
by &byvariable;
run;

data &outtable;
merge &intable (IN=presc) &nametable (keep = ndc name);
by &byvariable;
if presc=1 and length(name) > 2;
run;
%MEND;

proc sort data=prescriptionsforcancerpatients;
by patid;
run;

proc transpose data=prescriptionsforcancerpatients out=t_prescriptions prefix=name;
by patid;
var name;
run;

data prescriptionsperpatient;
set t_prescriptions;
length newname $ 50;
length px $ 4000;
array p_x[*] name:;
px = translate(left(trimn(p_x[1])),'_', ' ');;
do i=2 to dim(p_x);
```

```
newname = translate(left(trimn(p_x[i])),'_', ' ');
if index(px, newname) = 0 then px = catx(' ', px, newname);
end;
keep patid px;
run;
```

Because establishing associations between thousands of drug names and drug doses is not feasible, the prescription strings were compressed into clusters with a text-miner node. The resulting prescription clusters were examined with an association node to determine related prescription drug patterns. Likewise, the diagnoses strings were compressed into clusters, which were then analyzed with an association node to establish possible epidemiological patterns of disease. In order to examine the meaning of the resulting association rules, link analysis was utilized. Subsequently, the *SAS merge* procedure was used to merge these subsets to its *parent* sets respectively. The *SAS Kernel Density* function was used to compute the distribution of *Total Charges* for the year 2000 conditioned on a colorectal cancer diagnosis in 2001 as well as by cluster classification.

```
proc sort data=patients_2000;
by cancer_status;
run;

proc kde data=patients_2000;
univar totchgs / bwm=3 method=os out=kde_totchgs ;
by cancer_status;
run;
```

For classification purposes, predictive modeling was implemented within Enterprise Miner. These techniques help in the identifying of possible health insurance users who may develop colorectal cancer. First, a data partition node was used to partition the data into training, testing and validation sets so that the various predictive modeling methods can be compared and their results validated. For each type of cancer, colon or rectal, the regression node, dmine node, neural network node, decision tree node, and memory-based reasoning node were implemented (Diagram 1). To make a decision about the best model, a model comparison node was used to compare the model more directly.

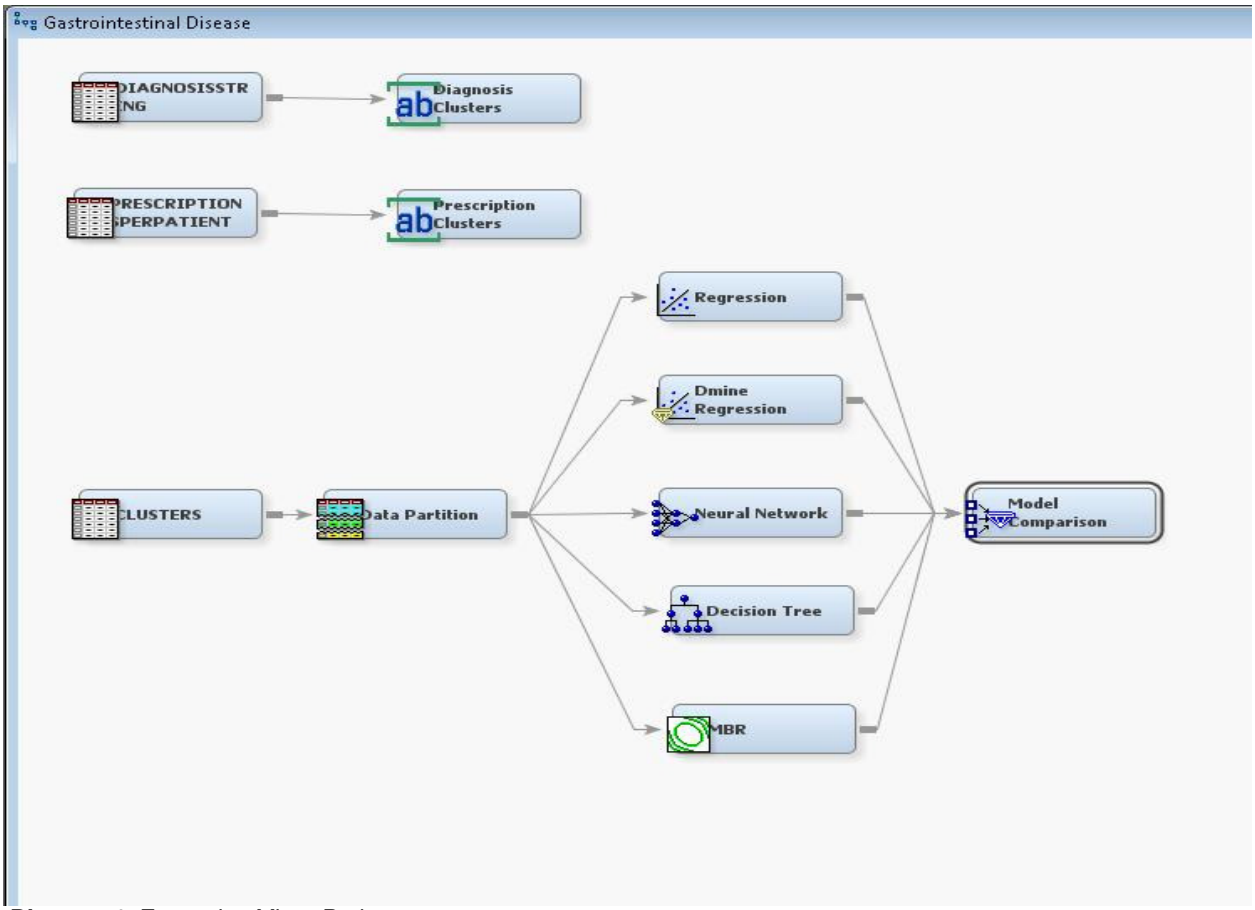

Diagram 1. Enterprise Miner Project

Results

The incidence of colorectal cancer was 49 per 100,000 for the year 2001. When adjusted for age, those who were 44 years old or younger had a rate of 18 per 100,000; while those who were over 44 years of age had a rate of 305 per 1000,000. The number of diagnosis per cluster was 4 distinct ICD9 codes. This resulted in 10 clusters. The top three clusters by percentage of observations were irritable bowel disease, diabetes, and gastric ulcer/dyspepsia. Because of the vast number of medications to treat a single condition, the cluster size for medication allowed up to twenty terms for a cluster definition. Under this condition, the prescription strings were compressed into eighteen categories. The top three clusters by percentage of observations were congestive heart failure, antibiotics, and diabetes.

Using link analysis of the prescription clusters, the most prominent node was *antibiotics*; it connected to other clusters with varying degrees of strength. The reason for this may be that gastrointestinal ulcers are being treated with antibiotic therapy for the presence of helicobacter pylori, various diarrhea types, and gastrointestinal parasites, which are associated with gastritis and dyspepsia. Similarly, the link analysis of the diagnosis clusters showed the irritable bowel disease and gastric ulcer/dyspepsia in the middle of the graph. They connected strongly with the nodes for diabetes and with the node of congestive heart disease with a lesser degree.

Kernel density estimated the probability density distribution for total charges for the year 2000 between those who had a colon cancer or rectal cancer diagnosis in the year 2001. Clearly, those who had a cancer diagnosis in the year 2001 also had a peak in their total charges for the year 2000 at much larger amount than those who did not have a colorectal cancer diagnosis in the year 2001. Likewise, those who had a colorectal cancer diagnosis in the year 2001 had a prescription utilization for the year 2000 centered to the right of the distribution of prescription utilization of those who did not have a colorectal cancer diagnosis in the year 2001.
Prescription clusters, diagnosis clusters, and age were used in the implementation of the various predictive model nodes. A model comparison node within Enterprise Miner was connected to the regression node, dmine node, neural network node, decision tree node, and memory-based reasoning node in order to compare the various models directly. This node uses the receiver operating curve (ROC) to visually compare the models in question; it also allows for comparison of fit statistics. The receiver operating curve uses the sensitivity on the y-axis and 1-specificity of the testing set on the x-axis; a curve with corresponding area under it close to 1 represents a better predictive model. Both the ROC and fit statistics indicated that the neural network node generated a better model that all the other predictive model nodes.

Conclusion

The incidence of colorectal cancer is very high in the United States. This makes this disease a top priority in epidemiology and health outcome analytics. Text mining and clustering proved to be excellent tools in the classification of patients and risk prediction of cancer and insurance utilization. Enterprise Miner predictive modeling nodes were great at easily and quickly implementing classification models for colorectal cancer based on previous diagnosis clusters, previous prescription clusters, and age of insurance users.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Pedro G. Ramos
University of Louisville
328 Natural Science Building
University of Louisville
Louisville, KY 40292
Phone: 502-852-6240
Fax: 502-852-7132
E-mail:pgramo01@louisville.edu