Asian Journal of Applied Sciences (ISSN: 2321 – 0893)
Volume 03 – Issue 02, April 2015

Applying Data Mining Technology on Sepsis with National Health Insurance Research Database

Yi-Horng Lai

Department of Health Care Administration, Oriental Institute of Technology
New Taipei City, Taiwan
Email: FL006 [AT] mail.oit.edu.tw

ABSTRACT—*Sepsis was a whole-body inflammation caused by an infection. Common signs and symptoms include fever, increased heart rate, increased breathing rate, and confusion. There may also be symptoms related to a specific infection such as a cough with pneumonia or painful urination, with a kidney infection. Sepsis causes and pathogenic mechanism are still not fully grasped by the medical profession. Early symptoms of sepsis are very similar to common diseases. It lead miss the appropriate time of treatment because of ignorant or erroneous diagnosis easily, which lead to serious complications or even death, and also wastes a lot of medical resource. The purpose of this study was to identify characteristics of patients with sepsis and patient's medical information in the National Health Insurance Research database in Taiwan by using data mining technique in decision tree. The result can be used to assist health care workers to identify the patient groups which have high-risk to suffering from sepsis and progress the prevent strategies.*

Keywords—Sepsis, National Health Insurance Research Database (NHIRDB),Data Mining, C5.0 Decision Tree

1. INTRODUCTION

Sepsis was a serious illness. It happened when the body has an overwhelming immune response to a bacterial infection. The chemicals released into the blood to fight the infection trigger widespread inflammation. This leaded to blood clots and leaky blood vessels. Sepsis cause poor blood flow, which deprives your body's organs of nutrients and oxygen. In the worst cases, blood pressure drops and the heart weakens, leading to septic shock.Early symptomsof sepsis were similar to common diseases. It would lead miss the appropriatetime of treatment because of ignorant or erroneous diagnosis easily, whichlead to serious complications or even death, and also waste a lot of medicalresource. The purpose of this study was to identify characteristics of patients withsepsis, and explore whether there were certain rules associated with sufferingsepsis and patient's medical information in the National Health InsuranceResearch Database (NHIRDB) by using data mining technique [1]. The result can be used toassist health care workers to identify the patient groups which have high-risk tosuffering from sepsis and develop the prevent strategies.

Now, health care treatments were better than before, there were still alot of disease cannot be prevent and treat completely, such as Sepsis.According to the Taiwan Department of Health Statistics, Sepsis was one of the top 10causes of death in Taipei and a number of counties and cities after 2006. Mostof death growth rate of disease in top 10 causes of death was decreasing. Sepsisbecomes the highest growth rate disease of the causes of death in Taiwan, thereasons are worth exploring, especially the numbers of death not including thepatients in the last stage of disease such as diabetes, cancer, accident by but alsoinfect sepsis.

Sepsis was a blood infection disease, which wereeasier infected the peoplewho was weak or low leucocyte, such as infants , young children, chronic patient,and the people using steroid in treatment. The bacteria may be from outside orinside the body, but how to invasive was still not fully grasped. The commonreasons which lead to sepsis are pneumonia and urinary tractinfection; theothers are debridement surgery, tooth extraction, chronic diseases, burned, long-term inpatients caused by abdominaloperation or multiple intubations[2]. The nidusesof sepsis often occur in reproductive system, lung, soft tissue, hepatobiliarysystem, gastrointestinal tract.

If sepsis patients' pathogeny was common typical symptoms, such as fever,tachycardia, chills, mental status changes, shortness of breath and low bloodpressure, it's easier to distinguish in diagnosis. However, the early stages ofSepsis or under special conditions, the performance of symptoms may not showobviously or similar to deteriorates of patient's original disease. It's easily missthe appropriate time of treatment because of ignorant or erroneous diagnosis [6].

When a patient appears the symptoms of sepsis, it present the patient hadbeen infected, and it will deteriorate rapidly. The probability of death of thepatients was high in short time. This study result about the characteristics ofpatients and the rules associated with sepsis would help to increase awarenessand prevention for the health care workers and high-risk group of patients withsepsis.Sepsis was a blood infection disease which had highly mortality rate. It was aninfection which leukocytes in the body fight with bad germs, then it transportedto the body via the blood to other organs. It would lead to serious complications or even death ifwithout immediate treatment.Because of the symptoms of sepsis in early stage were not obviously, andhave higher progression risk, it lead to miss the appropriate time of treatmentbecause of ignorant or erroneous diagnosis easily [3]. The more organ failure willlead the higher mortality rate. Server sepsis mortality rate is around 30%, andthe septic shock mortality rate are up to 40% ~ 70%.When mentioned thetreatment of sepsis, Janes, Vangerow, Costigan, and Macias [4] thinkearly detection and early treatment can reduce mortality and better prognosis.

In 2008, Taiwan National Health Insurance (NHI) has been 99.19%Insurance rate, with 91.75% medical institutions join in. So that, The NationalHealth Insurance Research Database (MHIRDB) contains samples close to the overallpatients with long time and detail medical information. Many studies make useof National Health Insurance Research Database (NHIRDB) to explore the characteristicsof patients as disease prevention and health care resource allocation [5].

National Health Insurance (NHI) in Taiwanwas that the public paid premium and part of themedical fees, with the remaining part paid by the National Health Insurance (NHI) tomedical institutions. Medical institutions in order to reclaim the remainingamount; they must be declared medical details to the National Health InsuranceBureau. Therefore, Bureau of National Health Insurance Database (NHIRDB) contains anumber of valuable and detailed information, such as patient age, gender,medical treatment areas, medical records, medication records, etc., the amountof data up to millions and the time range is long. The use of data mining, it could be excavated the hidden and unknown knowledge to provide some help tomedical.This study used of National Health Insurance Research Database (NHIRDB) between1999 to 2005 providing by The National Health Research Institutes (NHRI) in Taiwan.The out-patient prescription and treatment data (CD), and insuranceidentity data (ID) were in thesample data system to analysis health care information and to analysis the statistics of patient medical information beforesuffering sepsis to explore the potential factors that leading tosuffering sepsis, and use decision tree model to create a model of huge patient datato find the important characteristics fields of sepsis patients.

2. METHODOLGY

Cross Industry Standard Process for Data Mining (CRISP-DM) wasproposed by DaimlerChrysler, SPSS, and NCR in 1996. This study adapted CRISP-DM tobe the process model to this study. The sequence of the CRISP-DM phases was not strict. It was always needmoving back and forth between different phases. It depend on the resultof each phase which phase, or which particular task of a phase, that has tobe performed next. The arrows indicate the most important and frequentdependencies between phases.According to CRISP-DM, the first step of data mining was businessunderstanding, which was the base of solving problems. Data mining wasbased on domain knowledge to find problems, and using computertechniques to explore the relationship between data to solving problemsand knowing the trends.So it need to understanding depth to theproblems that to continue next steps.After define the target, this study selectedrelated data based on the target. Through the selection of properinformation, the computer could build the correct data model.

This study use Bureau of National Health Insurance (NHI) reported databetween 1997~2010 to be the research data. The purpose of this study wasexploring the relationship between sepsis patients and their characteristic. The subjects were the patients suffering from sepsis (ICD-9-CM is 038) and using National Health Insurance Research Database (NHIRDB) to get medical treatment. The out-patient prescription and treatment data (CD), and insurance identity data(ID) were in the sample datasystem to analysis health care information.

2.1 Data Mining Tools

There are many data mining software such asIBM SPSS Modeler, SASEnterprise Miner, Microsoft SQLServer, and WEKA.This study adaptedPython 3.4 and Orange 2.7.8 to be the data mining tool in this study as Figure 1. Python 3 and Orange 2.7.8 canaccess, organize, and model all types of data from within a single intuitivevisual interface. Build reliable models and deploy results quickly to meetbusiness goals. Collaboration capabilities boost user productivity, andserver-based options dramatically increase scalability and performance.Orange 2.7.8 provides several models and can mix the models.

Orange 2.7.8was a component-based data mining and machine learning software suite, featuring a visual

programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is implemented in Python, andits graphical user interface builds upon the cross-platform Qt framework. Orange was distributed free under the GPL. It is maintained and developed at the Bioinformatics Laboratory of the Faculty of Computer and Information Science, University of Ljubljana, Slovenia.

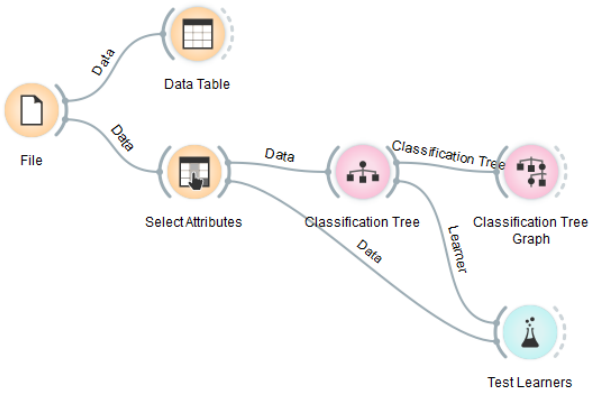

Figure 1: Orange 2.7.8 Data Mining Interface

2.2 Research Data

This study used Bureau of National Health Insurance reported data between1997~2010 to be the research data. The subjects were the patients suffering from sepsis (ICD-9-CM is 038) and using National Health Insurance Research Database (NHIRDB) to get medical treatment. This study selected the out-patient prescription and treatment data (CD), andinsurance identity data (ID)in the sample data system to analysis health care information.

There were 1001272 patients in this study. There were about 236730016 cases in CD and 25451 cases were sepsispatients. The ID file be jointedwith CD by subject's personal ID, and the resultbecomes the initial table for data mining.The data distribution of sepsis patients was as Table 1.

Table 1: The data distribution of sepsis patients

Variable		N	%
Sepsis	ICD9_0380	2178	8.56
	ICD9_0381	284	1.12
	ICD9_0382	63	.25
	ICD9_0383	26	.10
	ICD9_0384	4580	18.00
	ICD9_0386	23	.09
	ICD9_0388	624	2.45
	ICD9_0389	17673	69.44
AGEG	AGE_1	5189	20.39
	AGE_2	1673	6.57
	AGE_3	2207	8.67
	AGE_4	2473	9.72
	AGE_5	2846	11.18
	AGE_6	3550	13.95
	AGE_7	7513	29.52
PART_HOS	center	8629	33.90
	clinics	5283	20.76
	district	3068	12.05
	regional	8471	33.29
PART_TYP	EMR	12791	50.26
	OPD	12660	49.74
PART_HEA	barriers	1444	5.67
	normal	24007	94.33
Urban	N	25445	99.98
	Y	6	.02

Total		25451	100.00

3. RESAULT

Decision tree was a classification that can generalize rules from result.These rules were very important factor to affect data categorizing. Becausethe dataset have numerous field so that if input to clustering analysisimmediately may clustering analysis produce bad result that could notdetermine which fields are related to sepsis patients. Therefore, in thisstudy, it could be determined which fields were representative that aid with C5.0decision tree which it was good at handle set value.

C5.0 decision tree model provided simple modeand expert mode. Thestudy choose expert mode where satisfiedthe complex demand. Setting pruning severity in expert is to pruneunnecessary branch. With increasing value of pruning severity comessuccinct and low accuracy, otherwise comes complex and high accuracy.Settingoptions like Minimum records per child branch is to stop branching ifamount of records under a branch less than setting value. It would be able toavoid excessive training on noise or outliers. Because amount of recordsin the research was large and using decision tree in order to merelydetermine the importance of fields. It need to keep large numbers ofrecords containing the fields.

According to the result of data mining in this research using C5.0 decision tree in this study.This study eliminatedAGEG, PART_HOS, PART_TYP, PART_HEA, and Urban that are high discrimination will be the fields for clustering.

Reese, Betts, andGumustop[1] noted that the sepsis patients who were inpatient,drug resistance of the pathogeny is higher than common patients, doctorsshould consider the drug resistance when treat these patients. It was not easy to know all of these patients' past medical history, so thehospitals, especially medical centers, should consider the drug resistanceof medical areas when they making the infection control program.

The percentage of No co-payment and barrier patients of sepsis arehigher, it's likely the reason that the most of patients with No co-paymenthave long-term or serious disease, both of No co-payment and barrierpatients are low immunity patients, and have potential factors of sufferingfrom sepsis.

Decision tree was a classification that can generalize rules from result.These rules were very important factor to affect data categorizing. Becausethe dataset have numerous field so that if input to clustering analysisimmediately may clustering analysis produce bad result that could notdetermine which fields are related to sepsis patients. Therefore, in thisstudy, it could be determined which fields are representative that aid with C5.0decision tree which it was good at handle set value.

C5.0 decision tree model provided simple modeand expert mode. The study choose expert mode where satisfiedthe complex demand. Setting pruning severity in expert is to pruneunnecessary branch. With increasing value of pruning severity comessuccinct and low accuracy, otherwise comes complex and high accuracy.Settingoptions like Minimum records per child branch is to stop branching ifamount of records under a branch less than setting value. It is able toavoid excessive training on noise or outliers. Because amount of recordsin the research is large and using decision tree in order to merelydetermine the importance of fields. It should be need to keep large numbers ofrecords containing the fields. The result was as Figure 2.

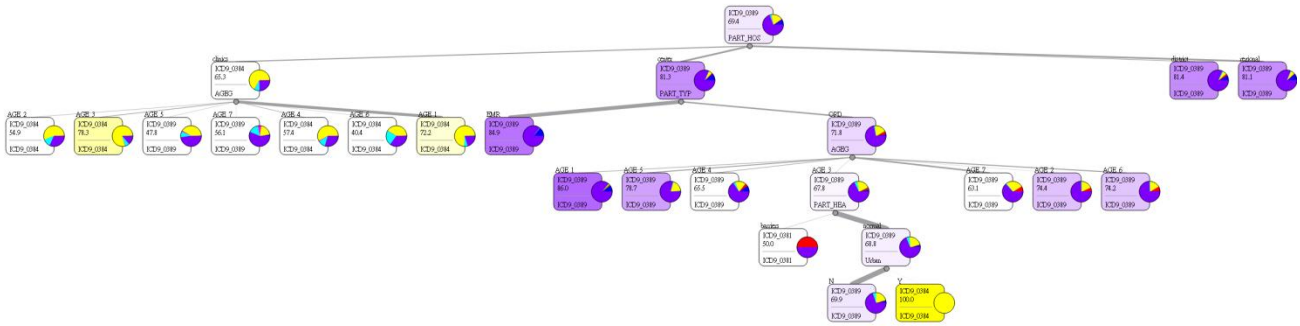

Figure 2: The Decision Tree of Sepsis

4. DISCUSSION AND RECOMMENDATION

According to the result of data mining, the drug resistance ofantibiotics, the percentage of infectious disease doctors

or specific doctors,the change of population structureand disease of patients may the factor that had influence to sufferingfrom sepsis.For hospitals managers and government, the study can improve the drug resistance ofantibiotics, the percentage of infectious disease doctors, infection control,and teach the high risk patients and their caregivers' health education ofsepsis.For the future research, the study provided an interpretation of the factor of sepsis;the future research could take in-depth study of the factors to help theprevention of sepsis.For the high risk patients and their caregivers, they could go to thehospitals that having higher rate of specific doctors or infectious disease,increasing their immunity, prepare sufficient medical history, and go tohospital when occurring the similar symptoms as soon as possible.

With the result, thedrug resistance of antibiotics has effect to infection disease, sepsis iscaused by infection disease, but there have not research to find therelation of sepsis and drug resistance of antibiotics directly. So this studyrecommendation future research to investigate the relationship between drugresistance of antibiotics and suffering from sepsis.

Besides, thelong-term disease and medical history are related to suffering fromsepsis. But National Health Insurance Research Database (NHIRDB) limits less than three diagnoses in one patient,if the follow-up researchers can get the detailed disease information onpatients, it be conducive to development the direction of preventionstrategies.

5. ACKNOWLEDGEMENT

This study is based in part on data from the National Health Insurance Research Database provided by the Bureau of National Health Insurance, Department of Health and managed by National Health Research Institutes. The interpretation and conclusions contained herein do not represent those of Bureau of National Health Insurance, Department of Health or National Health Research Institutes.

6. REFERENCES

[1]Reese, R.E., Betts, R.F.,&Gumustop, B.,Handbook of antibiotics (3rdEdition).Lippincott Williams & Wilkins, USA, 2000.

[2]Starr, M. E., Takahashi, H., Okamura, D., Zwischenberger, B.A.Mrazek, A.A., Ueda, J., Stromberg, A.J., Evers, B.M. , Esmon, C.T., &Saito, H., "Increased coagulation and suppressed generation of activated protein C in aged mice during intra-abdominal sepsis", American Journal of Physiology - Heart and Circulatory Physiology,vol. 308, no. 2, pp. 83-91, 2015

[3] Rivers, E., Nguyen, B., Havstad, S., Ressler, J., Muzzin, A., Knoblich, B., "Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock", New England Journal of Medicine, vol. 345, no. 19, pp. 1368-1377, 2001.

[4]Janes, J.M., Vangerow, B., Costigan, T.M. &Macias, W.L., "Drotrecoginalfa (activated) in severe sepsis", The Lancet Infectious Disease, vol. 13, no. 2, pp. 108-109, 2013.

[5] Lai, Y.H.,"Applying Data Mining Technology on National Health Insurance Research Database in Taiwan: HIV/AIDS as an Example", Asian Journal of Applied Sciences, vol. 2, no. 6, pp. 922-927, 2014.

[6] Monti, G., Landoni, G., Taddeo, D., Isella, F., &Zangrillo, A., "Clinical Aspects of Sepsis: An Overview", Sepsis, vol. 1237, pp. 17-33, 2015.