Applying Data Mining Technology on National Health Insurance Research Database in Taiwan: HIV/AIDS as an Example

Yi-Horng Lai

Department of Health Care Administration, Oriental Institute of Technology
New Taipei City, Taiwan
Email: FL006 {at} mail.oit.edu.tw

ABSTRACT— ***Human immunodeficiency virus (HIV) had great impact on global medicine and public Health since it was found in 1980. Studies related to utilization of medical resource on HIV/AIDS were popular in the other countries; however, such studies are rare in Taiwan. The purpose of this study is to identify characteristics of patients with HIV/AIDS and patient's medical information in the National Health Insurance Research database (NHIRD) by using data mining technique in CHAID decision tree and Two-stage clustering analysis. The result can be used to assist health care workers to identify the patient groups which have high-risk to suffering from HIV/AIDS and develop the prevent strategies. The result separated patients to three groups. With the result of this study, practitioners pay more attention to the patient's past history and long-term treatment of the disease, particularly in high-relevance department.***

Keywords—HIV/AIDS Patients, data mining, CHAID decision tree, Two-stage clustering analysis

1. INTRODUCTION

The HIV/AIDS epidemic was one of the most important and crucial public health risks facing governments and civil societies in the world. Adolescents were at the center of the pandemic in terms of transmission, impact, and potential for changing the attitudes and behaviors that underlie this disease. Therefore, HIV/AIDS prevention has become a priority all in the world.

The first HIV/AIDS case in Taiwan was reported in 1984. As of the end of 2013, the total number of HIV/AIDS cases had been accumulated to 26475. The number of HIV/AIDS infections began to outpouring since 2004 was due to a major increase of infection among injecting drug users. Faced with this serious situation, Taiwan Centers for Disease Control worked with other departments and dedicated a tremendous amount of effort and resources to introduce harm reduction programs. Total reported cases dropped in 2006, which was the first trend reversal since 1984. In 2008 and thereafter, the epidemic took a turn; infections mainly occurred through sexual encounter. In face of the rising HIV/AIDS epidemic, the most pressing course of action is the reinforcement of the health education campaign and intervention plans for the targeted population [1].

In terms of age, the largest number of infections in 2013 was in the 20-29 age group, accounting for 51.00% of all cases. The second largest group was the 30-39 age group, numbering 29.40% of all cases. Of Taiwanese nationals infected by HIV in 2013, 98% were males and 2% were females. The ratio of infected males to females was 42:1. An analysis of risk factors showed that in 2013, the highest proportion of HIV infections was a result of unsafe sexual transmission, with men who have sex with men accounting for 80% of all cases. The second largest proportion of infections was heterosexual contact, accounting for 12.00% [2].

Youth, who was in sexually active age, was under higher risk of contracting the HIV/AIDS. HIV-positive persons in the 15-24 age group account for 19.95% of the total number. Younger generations in Taiwan have become more and more open-minded about sex; most sexual intercourse between youngsters occurs without proper protection. According to data from the National Taiwan Normal University, only 30% of college students use condoms every time they have sex. The lack of awareness may trigger a disastrous outbreak among young adults [3].

In order to get data to guide future HIV/AIDS strategies, this study uses outpatients' data of National Health Insurance Research Database in 1997 to 2010 for data mining. The result about the characteristics of patients and the rules associated with HIV/AIDS would help to increase awareness and prevention for the health care workers and high-risk group of patients with HIV/AIDS.

2. METHODOLOGY

2.1. Research Framework

Cross Industry Standard Process for Data Mining (CRISP-DM) was proposed by DaimlerChrysler, SPSS, and NCR in 1996. It was an industry and tool-neutral data mining process model. So this study adapts CRISP-DM to be the process model to this study.

The sequence of the CRISP-DM phases was not strict. It was always need moving back and forth between different phases. It depend on the result of each phase which phase, or which particular task of a phase, that has to be performed next. The arrows indicate the most important and frequent dependencies between phases. The outer circle in the figure symbolizes the cyclic nature of data mining itself.

2.2. Research Tools

Recently, most of business information has been computerized, through the appropriate model and calculation, this valuable information can help companies to understand trends and improve decision-making quality. However, the ever-growing amount of data cause the difficulty in the used of artificial to analysis information, so the market of automatic analysis software that can automatically retrieve large amounts of data to useful knowledge (such as knowledge discovery and data mining) developed rapidly. In recent years, there are more used in the experiment and research [4].

Recently, there are many famous data mining software such as SAS Enterprise Miner, WEKA, and Microsoft SQL Server. This study adapted IBM SPSS Modeler 14.1 to be the data mining tool in this study. This computer software can access, organize, and model all types of data from within a single intuitive visual interface. Build reliable models and deploy results quickly to meet business goals. Collaboration capabilities boost user productivity, and server-based options dramatically increase scalability and performance. Clementine provides several models and can mix the models. Clementine also combine with CRISP-DM, so user can understand models and trends more easily, then become the leader of data mining field. The IBM SPSS Modeler data mining interface was as Figure 1.

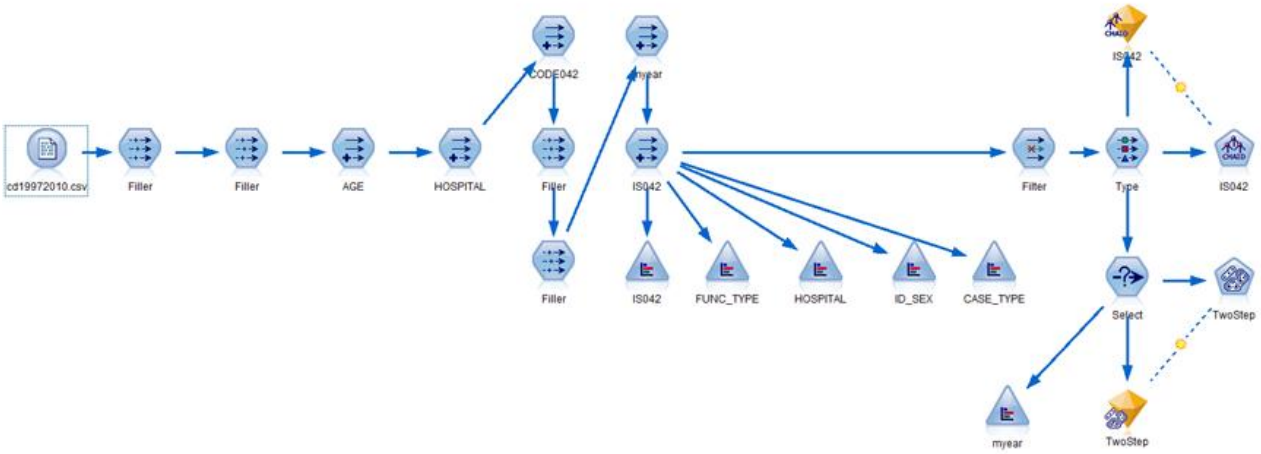

Figure 1: IBM SPSS Modeler Data Mining Interface

2.3. Research Data

According to CRISP-DM, the first step of data mining is business understanding, which is the base of solving problems. Data mining is based on domain knowledge to find problems, and using computer techniques to explore the relationship between data to solving problems and knowing the trends. So it needs to understanding depth to the problems that to continue next steps. After define the target, it would be selected related data based on the target. Through the selection of proper information, the computer could build the correct data model.

This study used Bureau of National Health Insurance reported data (NHIRD) between 1997 and 2010 to be the research data. The purpose of this study was exploring the relationship between HIV/AIDS patients and their characteristic. The subjects are the patients suffering from HIV/AIDS (first three ICD-9-CM is “042”) and using NHI to get medical treatment. The out-patient prescription and treatment data (CD), insurance identity data (ID), and the basic data of medical institutions (HOSB) in the sample data system to analysis health care information.

3. RESULTS

3.1. Data Understanding

This study use Bureau of National Health Insurance reported data (NHIRD) between 1997 and 2010 to be the research data. The subjects were the patients suffering from HIV/AIDS (first three ICD-9-CM is "042") and using NHI to get medical treatment. This study selected the out-patient prescription and treatment data (CD) and insurance 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 46200 cases are HIV/AIDS patients. It would be joined ID to CD by subject's personal ID, and the result becomes the initial table for data mining. The next step of the stage is removing the unrelated fields, such as drug code, card number, insurance date, and apply date.

3.2. Data Analysis

This study use STATA 13 and IBM SPSS Modeler 14.1 data audit to show the data distribution of patients with HIV/AIDS.

3.2.1. Data analysis of Patients with HIV/AIDS

The mean of patients' age was 40.92, and S.D. was 423.76 in 2010. This study used STATA 13 data audit to show the data distribution of HIV/AIDS patients as shown as Table 1 and Table 2.

Table 1: The data distribution of patients

Variable		N	%
Sex	Female	504961	50.43
	Male	496311	40.57
HIV/AIDS	Yes	866	.09
	No	1000406	99.91
Total		1001272	100.00

Table 2: The data distribution of patients in Outpatient department

Variable		N	%
HIV/AIDS	Yes	22839	.01
	No	191146128	99.99
Sex	Female	106754671	55.84
	Male	84265805	44.08
	Unknown	148491	.08
Case type	Common western medicine (01)	60854096	31.83
	Other western medicine (09)	54225850	28.37
	Chronic of western medicine (04)	27388425	14.33
	Other dentist medicine (19)	15076868	7.89
	Others	33623728	17.58
Hospital	Private clinics (35)	98478272	51.51
	Foundation hospital (11)	16941728	8.86
	Traditional Chinese medicine hospital (38)	16708827	8.74
	Private Hospitals (15)	16608830	8.69
	Private dental clinic (37)	14462022	7.57
	Others	27969288	14.63
Department	General Medicine (00)	23157546	12.11
	Family Medicine (01)	20888745	10.93
	Traditional Chinese medicine (60)	18999223	9.94
	Internal Medicine (02)	18946747	9.91
	Otolaryngology (09)	17392249	9.10
	Others	91784457	48.01
Total		191168967	100.00

In Table 3, Human immunodeficiency virus (HIV) disease that 85.45% of HIV/AIDS patients are unspecified

HIV/AIDS.

Table 3 HIV/AIDS type and percentage

Code	Code Type	N	%
042	HIV disease	740	85.45
042.0	HIV infection with specified infections	28	3.23
042.1	HIV infection with other specified infections	1	.12
042.2	HIV infection with specified malignant neoplasms	0	.00
042.9	Acquired immunodeficiency syndrome, unspecified	97	11.20
Total		866	100.00

In Table 4 and Figure 2, shows the cases were increasing in 1997 and 2010, and hit the peak in 2010.

Table 4: The trend of HIV/AIDS

Year	Cases	%
1997	60	.26
1998	136	.60
1999	390	1.71
2000	413	1.81
2001	589	2.58
2002	764	3.35
2003	1002	4.39
2004	1265	5.54
2005	1439	6.30
2006	2585	11.32
2007	2941	12.88
2008	3192	13.98
2009	3757	16.45
2010	4306	18.85
Total	22839	100.00

3.3. Model Introduction

After data pre-processing, data table has become appropriate data sets for this study. CHAID decision tree and two steps clustering analysis would be applicate as approach to perform this research.

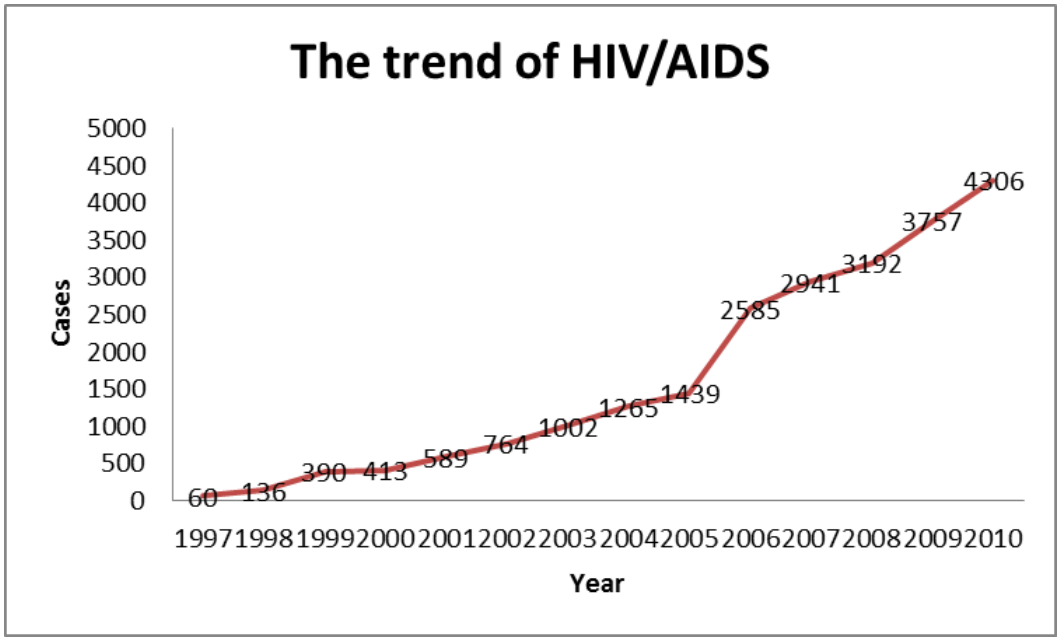

Figure 2: The Trend of the Type of Sepsis (unit: cases)

3.3.1. Decision Tree

Decision tree is a classification that can generalize rules from result. These rules are very important factor to affect data categorizing. Because our dataset have numerous field so that if input to clustering analysis immediately may clustering analysis produce bad result that could not determine which filed are related to HIV/AIDS patients. Therefore, in this research, it could be determined which fields are representative that aid with CHAID decision tree which it is good at handle set value.

According to the result of data mining in this research with application of IBM SPSS Modeler CHAID, the decision tree was with depth 4. Furthermore, the discrimination for each fields were as shown in Table 5.

However, it had set value of pruning severity and stop branching when amount of records within a branch is less than 100 so that may not find some fields in the tree. According the result, it could be eliminated AGE and ID_SEX.

Table 5: CHAID discrimination power

Field	Discrimination power
AGE	.52
ID_SEX	.48

3.3.2. Clustering Analysis

The study used two-stage clustering as our clustering analysis approach. Two-stage clustering was combined hierarchical clustering and non-hierarchical clustering. First, to use Ward's methods for determine number of clusters. The Wards's methods, also known as Minimum Variance Method, treated every data point as a group, then merged any two cluster or data point into one cluster by estimate change of variance and select the least one if merged closer cluster or data point This method can judge suitable number of clusters according to clustering coefficient. The second stage is further to cluster by K-means which are hierarchical clustering on the result of first stage.

The results were shown in Table 6. For these patients with HIV/AIDS, the discrimination power of CASE_TYPE, FUNC_TYPE, and HOSPITAL was 1, ID_SEX was .50, and AGE was .06.

There were 8696 cases in Cluter-1 group, accounted for 38.10% of all cases. The average of age was 38.36. 99.70% of them were Case of HIV. 97.90% of them were in the department of infectious diseases. 52.50% of them were in public hospital. 89.20% of them were male. There were 7608 cases in Cluter-2 group, accounted for 33.40% of all cases. The average of age was 39.82. 54.30% of them were case of HIV. 32.80 of them were in the medical department. 37.10% of them were in public medical school hospital. 91.70% of them were male. There were 6493 cases in Cluter-3 group, accounted for 28.50% of all cases. The average of age was 38.41. 100.00% of them were case of HIV. 100.00 of them were in the department of infectious diseases. 55.20% of them were in Private medical school hospital. 100.00% of them were male.

Table 6: Clustering analysis result

	Discrimination Power	Cluster-1	Cluster-2	Cluster-3
Size		8696 (38.10%)	7608 (33.40%)	6493 (28.50%)
CASE_TYPE	1.00	Case of HIV (99.70%)	Case of HIV (54.30%)	Case of HIV (100.00%)
FUNC_TYPE	1.00	Infectious Diseases (97.90%)	Medical Department (32.80)	Infectious Diseases (100.00%)
HOSPITAL	1.00	Public Hospital (52.50%)	Public medical school hospital (37.10%)	Private medical school hospital (55.20%)
ID_SEX	.50	Male (89.20%)	Male (91.70%)	Male (100.00%)
AGE	.06	38.36	39.82	38.41

3.4. Result of Data Mining

Based on the result of CHAID decision tree, it could find that age and sex play an important role in the classification

of patients with HIV/AIDS and patients without HIV/AIDS.

Based on the result of Two-stage clustering analysis, it could find that there were three groups (clusters) in these patients with HIV/AIDS. The most obviously characteristic of cluter-1 were HIV/AIDS patients that younger than other two groups. Most female HIV/AIDS patients were in this group. Most of them usually go to public hospital. This group was youth HIV/AIDS group. The most obviously characteristic of cluter-2 were HIV/AIDS patients that older than other two groups. Most of them were not Case of HIV, so they may go to hospital for other disease. Most of them usually go to Public medical school hospital. This group was complication group. The most obviously characteristic of cluter-3 were they all case of HIV. They all were male. This group was pure HIV/AIDS group.

4. CONCLUSION

The cases of HIV/AIDS in Taiwan were increasing in 1997 and 2010. Based on the result of this study, it could find that age and sex play an important role in the classification of patients with HIV/AIDS and patients without HIV/AIDS. Besides, it could find that there were three groups (clusters) in these patients with HIV/AIDS. They were youth HIV/AIDS group, complication group, and pure HIV/AIDS group. Youth HIV/AIDS group was HIV/AIDS patients that younger than other two groups. Most female HIV/AIDS patients were in this group. Most of them usually go to public hospital. This group was youth HIV/AIDS group. Complication group was HIV/AIDS patients that older than other two groups. Most of them were not Case of HIV, so they may go to hospital for other disease. Most of them usually go to Public medical school hospital. This group was complication group. Pure HIV/AIDS group was they all case of HIV. They all were male. This group was pure HIV/AIDS group.

According to literature review and the discussion of the result, the long-term disease and medical history are related to suffering from HIV/AIDS. But NHIR database limits less than three diagnoses in one patient. If the follow-up researchers can get the detailed disease information on patients, it is conducive to development the direction of prevention strategies.

5. ACKNOWLEDGMENT

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] Taiwan Centers for Disease Control, "Communicable Diseases & Prevention - HIV/AIDS HIV/AIDS," Health topics, Retrieved May 20, 2014 from http://www.cdc.gov.tw

[2] Li, N., Li, X,, Wang, X., Shao, J., & Dou, J., "A Cross-Site Intervention in Chinese Rural Migrants Enhances HIV/AIDS Knowledge, Attitude and Behavior," International Journal of Environmental Research and Public Health, 11(4), pp. 4528-4543, 2014.

[3] Ford, K., Chamratrithirong, A., Apipornchaisakul, K., Panichapak, P., & Pinyosinwat, T., "Social Integration, AIDS Knowledge and Factors Related to HIV Prevention among Migrant Workers in Thailand," AIDS and Behavior, 18(2), pp. 390-397, 2014.

[4] Grupe, F. H., & Owrang, M. M., "Data Base Mining Discovering New Knowledge and Competitive Advantage," Information Systems Management, 12(4), pp. 26-31, 1995.