Applying Data Mining Technology on Risk Analysis for Transitional Cell Carcinoma on End-Stage Renal Disease Patients with Maintenance Hemodialysis

Yi-Horng Lai[1],

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

ABSTRACT— ***The purpose of this study was application the general insurance claims data to investigate the risk of transitional cell carcinoma (TCC) for patients with hemodialysis (HD). This study surveyed 4577 patients with end-stage renal disease undergoing maintenance HD and 970339 patients without HD identified from claims data of National Health Insurance from 1997 to 2010. Incidence densities of TCC in upper urinary tract (UUT) and bladder only were estimated for these two cohorts. Hazard ratios (HRs) of TCC were measured in association with HD, covariates and comorbidities. Based on the result, the risk rate of TCC and UUT of patients with HD was higher than patients without HD. The risk of female was lower than male. The risk in TCC of female was .63 times of male. The age was positive with the risk of TCC an UUT. One more unit of age of patients, and the risk of TCC increase 1.04 times. One more unit of age of patients, and the risk of UUT increase 1.04 times.***

Keywords— Hemodialysis, transitional cell carcinoma, insurance data, urinary tract disease, data mining, Cox regression

1. INTRODUCTION

The incidence of transitional cell carcinoma (TCC) in Taiwan was greater than other countries. TCC was most prevalent in the bladder followed by the upper urinary tract (UUT), including the ureter and renal pelvis. According to the latest cancer registry annual report from the Bureau of Health Promotion Department in Taiwan, the incidence of TCC of UB and UUT in 2007 was 2050 versus 1174 among the 23-millioned population [1]. The UUT-TCC ratio here was higher than other countries' investigations that they account for about 5% of all urothelial tumors [2].

More and more study indicated that the correlation between over intake of Aristolochic acid and nephropathy as well as urothelial cancer among renal insufficiency or kidney transplantation patients in the recent decades. Except for the Asia area, numerous researches in western countries have previously disclosed the same phenomenon that Aristolochic acid was a risk factor for both the urothelial cancinoma and progressive renal interstitial fibrosis disease [3, 4, 5, 6, 7].

No matter it was the Aristolochic acid or other compounds that make renal damage with potential risk of TCC development, numerous clinical observation studies have disclosed the correlation between ESRD and TCC formation. The hazard ratio (HR) of UUT and bladder TCC between the uremic status undergoing hemodialysis (HD) patients and non-HD population has not been studied. The aim of this study was to applicate the claims data from the Bureau of National Health Insurance (NHI) to analyze the incidence, comorbidities and characteristics of TCC patients including UUT and urinary bladder TCC among these two cohorts.

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. Data source

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 TCC, UTT, and HD.

The subjects are the patients suffering from TCC (first three ICD-9-CM is "188"), UUT (ICD-9 is 189.1 or 189.2) and using NHI to get medical treatment. The out-patient prescription and treatment data (CD), insurance identity data (ID) in the sample data system to analysis health care information.

2.3. 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 developed rapidly. In recent years, there are more used in the experiment and research [9].

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 14.1 data mining interface was as Figure 1.

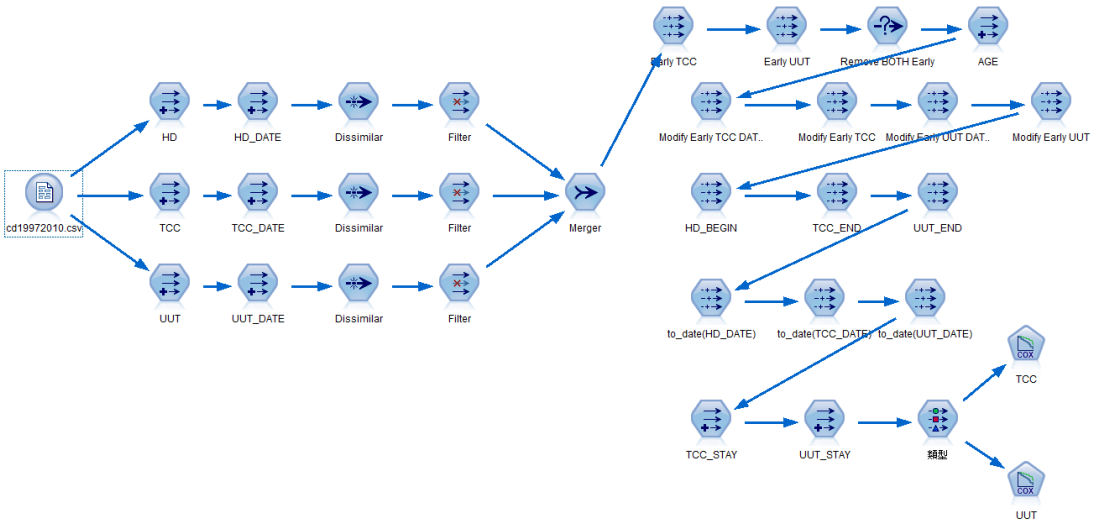

Figure 1: IBM SPSS Modeler Data Mining Interface

2.4. Statistical analysis

The methodology of data analysis in this study was survival analysis. Survival analysis is a branch of statistics which deals with death in biological organisms and failure in mechanical systems. The data analysis in the study was done with IBM SPSS Modeler 14.1.

3. RESULTS

3.1. Descriptive Statistics

There were 1001272 patients in this study. There were about 236730016 cases in CD and 974916 cases are HD patients. Comparisons in sociodemographic factors between cohorts with and without hemodialysis (HD as Table 1. Comparisons in sociodemographic factors between cohorts with and without transitional cell carcinoma (TCC) as Table 2. Comparisons in sociodemographic factors between cohorts with and without upper urinary tract (UUT) as Table 3.

Table 1: Comparisons in sociodemographic factors between cohorts with and without HD

Hemodialysis	No	Yes	Total	%
Gender				
Female	487729	2297	490026	50.26
Male	482610	2280	484890	49.74
Total	970339	4577	974916	100.00

Table 2: Comparisons in sociodemographic factors between cohorts with and without TCC

TCC	No	Yes	Total	%
Gender				
Female	489201	825	490026	50.26
Male	483572	1318	484890	49.74
Total	972773	2143	974916	100.00

Table 3: Comparisons in sociodemographic factors between cohorts with and without UUT

UUT	No	Yes	Total	%
Gender				
Female	489952	74	490026	50.26
Male	484736	154	484890	49.74
Total	974688	228	974916	100.00

3.2. Data Analysis-TCC

Based on overall modeling test, Chi-square was3922.95 (P-value<0.05) and Cox Regression was significant.

Table 4 Summary of Cox Regression Model

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Sex*	-.47	.04	111.54	1	<.01	.63	.57	.68
Age	.04	<.01	6100.26	1	<.01	1.04	1.04	1.04

*Reference: male

Based on the summary of the regression model, it could be found that sex and age were significant. The regression model of this study could be shown as follow:

$h(t)=h0(t)exp(-.47\times Sex+.04\times Age)$

Base on the regression model, the risk of female was .63 times (Exp(-.47)) of male. One more unit of age of patients, and the risk of TCC increase 1.04 times (Exp(.04)).

The Chi-square of the change between each blocks was 3016.77 (p<.05)). The survival function line of patients with HD and The survival function line of patients without HD was different. The risk of patients with HD (HD=1) was higher that the risk of patients without HD (HD=0) in TCC as Figure 2.

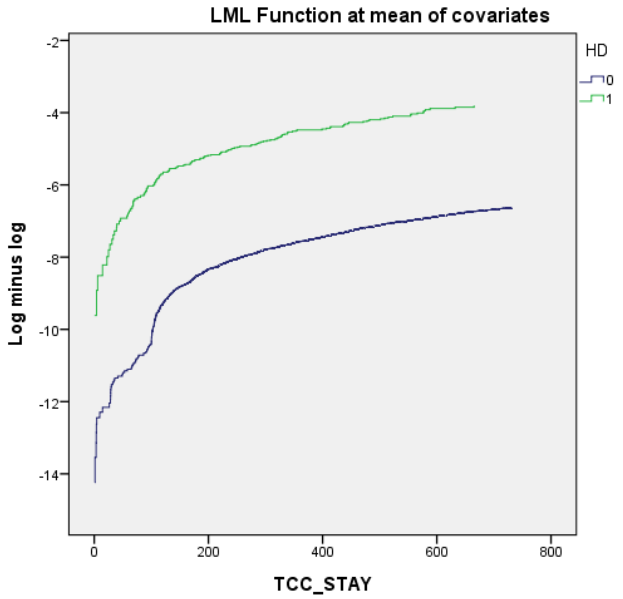

Figure 2: Plot of LML Function in TCC

The curve of Log(-log(t)) of Hazard function of patients with HD and patients without HD were as Figure 2, and both curve of two groups were two parallel lines. The ratio of the two function line was a constant, and it means the model of this study was followed the assumption of Cox regression model rule. According to Figure 3, the up one was the Hazard function of the hospice patients and the down one was the Hazard function of the non-hospice patients. It could be found that the mortality rate of patients with HD was higher than the mortality rate of patients without HD.

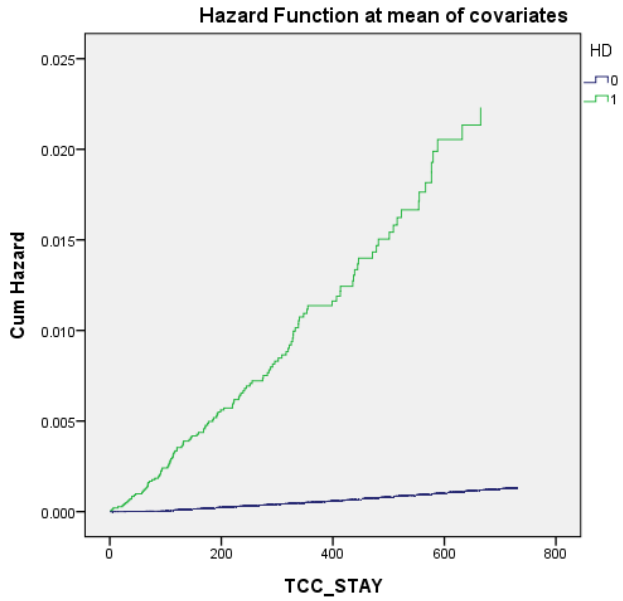

Figure 3: Plot of Hazard Function in TCC

3.3. Data Analysis-UUT

Based on overall modeling test, Chi-square was 539.803 (P-value<0.05) and Cox Regression was significant.

Table 4 Summary of Cox Regression Model

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Sex*	-.73	.14	26.63	1	<.01	.48	.37	.64
Age	.04	<.01	831.02	1	<.01	1.04	1.04	1.05

*Reference: male

Based on the summary of the regression model, it could be found that sex and age were significant. The regression model of this study could be shown as follow:

$h(t)=h_0(t)\exp(-.73\times Sex+.04\times Age)$

Base on the regression model, the risk of female was .48 times (Exp(-.73)) of male. One more unit of age of patients, and the risk of UUT increase 1.04 times (Exp(.04)).

The Chi-square of the change between each blocks was 3016.77 ($p<.05$)). The survival function line of patients with HD and The survival function line of patients without HD were different. The risk of patients with HD (HD=1) was higher that The risk of patients without HD (HD=0) in UUT as Figure 4.

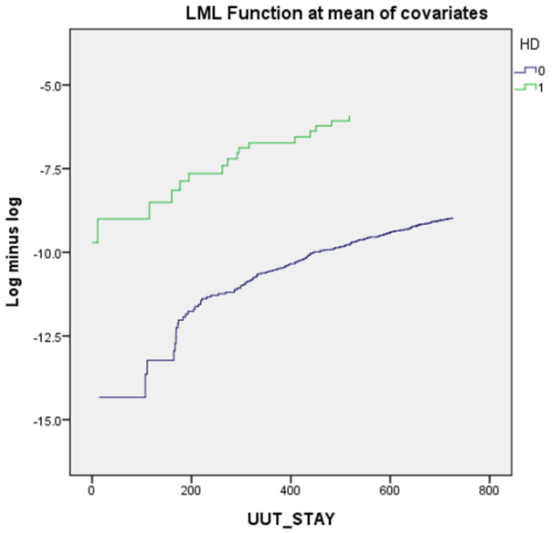

Figure 4: Plot of LML Function in UUT

The curve of Log(-log(t)) of Hazard function of non-hospice patients and hospice patients were as Figure 4, and both curve of two groups were two parallel lines. The ratio of the two function line was a constant, and it means the model of this study was followed the assumption of Cox regression model rule. According to Figure 5, the up one was the Hazard function of the hospice patients and the down one was the Hazard function of the non-hospice patients. It could be found that the mortality rate of patients with HD was higher than the mortality rate of patients without HD.

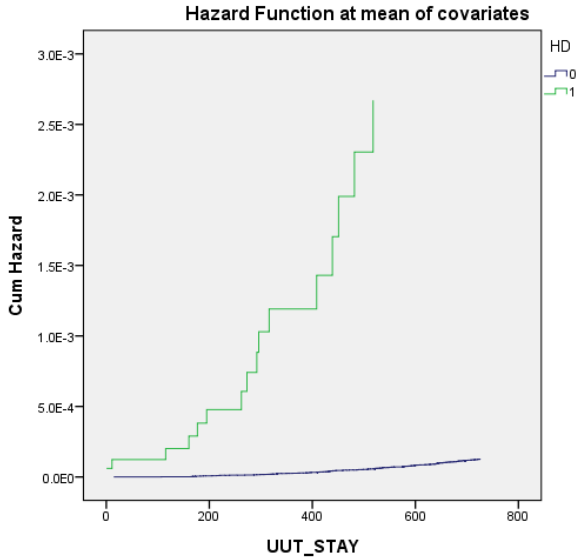

Figure 5: Plot of Hazard Function in UUT

4. CONCLUSION

Based on the result, the risk rate of TCC and UUT of patients with HD was higher than patients without HD. The same as Huang's study [2]. The patient with HD should pay attention to the risk of TCC and UUT. The risk of female was lower than male. The risk in TCC of female was .63 times of male. The risk in UUT of female was .48 times of male. The same as Huang's study [2]. The male patient with HD should pay attention to the risk of TCC and UUT. The age was positive with the risk of TCC an UUT. One more unit of age of patients, and the risk of TCC increase 1.04 times. One more unit of age of patients, and the risk of UUT increase 1.04 times. The same as Huang's study [2]. The elderly patient with HD should pay attention to the risk of TCC and UUT.

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

List and number all bibliographical references in 10-point Times New Roman, single-spaced, at the end of your paper. For example, [1] is for a journal paper, [2] is for a book and [3] is for a conference (symposium) paper.

[1] Hall, M.C., Womack, S., Sagalowsky ,A.I., Carmody, T., Erickstad, M.D., & Roehrborn, C.G. "Prognostic factors, recurrence, and survival in transitional cell carcinoma of the upper urinary tract: a 30-year experience in 252 patients. Urology", vol. 52, pp. 594-601, 1998.
[2] Huang, C. P. (2010). "Risk Analysis for Urothelial Transitional Cell Carcinoma on End-Stage Renal Disease Patients with Maintenance Hemodialysis: A Population-Based Study", Master Thesis of Executive MBA Program at College of Management, National Chiayi University, 2010.
[3] Babaian, R. J., Johnson, D.E. "Primary carcinoma of the ureter", The Journal of Urology, vol. 123, pp. 357-359, 1980.
[4] Yuan, M., Shi, Y.B., Li, Z.H., Xia, M., Ji, G.Z., Xu, G.X., Han, Y. "De novo urothelial carcinoma in kidney transplant patients with end-stage aristolochic acid nephropathy in China", Transplant Proc. vol. 41, No. 5, pp. 1619-1623, 2009.
[5] Vanherweghem, J. L., Tielemans, C., Abramowicz, D., Depierreux, M., Vanhaelen-Fastre, R., Vanhaelen, M., Dratwa, M., Richard, C., Vandervelde, D., Verbeelen, D., & Jadoul, M. "Rapidly progressive interstitial renal fibrosis in young women: association with slimming regimen including Chinese herbs", The Lancet, vol. 341, no. 8842, pp. 387-391, 1993.
[6] Lord, G.M., Tagore, R., Cook, T., Gower, P., Pusey, C.D. "Nephropathy caused by Chinese herbs in the UK", Lancet, vol. 354, pp. 481-482, 1999.
[7] Cosyns, J.P., Jadoul, M., Squifflet, J.P., Wese, F.X., van Ypersele, de Strihou C. "Urothelial lesions in Chinese-herb nephropathy", American Journal of Kidney Diseases, vol. 33, no. 6, pp. 1011-1017, 1999.
[8] Debelle, F.D., Vanherweghem, J.L., Nortier, J.L. "Aristolochic acid nephropathy: a worldwide problem", Kidney International, vol. 74, no. 2, pp. 158-169, 2008.
[9] Grupe, F. H., & Owrang, M. M., "Data Base Mining Discovering New Knowledge and Competitive Advantage", Information Systems Management, vol. 12, no. 4, pp. 26-31, 1995.