Applying Data Mining Technology on Survival Analysis for Arteriovenous Fistula in Taiwan

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

Abstract—**This study described the effect and change history after arteriovenous fistula operations in Taiwan by case linkage in the National Health Insurance Research Database (NHIRD) from 1997 to 2010. 3374 patients and 3866 cases with confirmed clinical details of arteriovenous fistulae were queried from the 1/20-sampled hospitalization dataset. The overall patency rate is above 80.00% in 600 weeks. The survival curves were different in fresh and non-fresh cases. The survival probability of non-fresh patient was 1.64 times on fresh patients. Survival of different age groups was shown in Figure 3 and Table 4. Age group 50-59, age group 60-69, and age group 70- were significantly different from age group -19. The survival probability of age 60-69 patient was .23 times on age group -19 patients. The survival curves were different in male and female cases. The survival probability of male patient was 1.39 times on female patients. With competing risks analysis, fresh and non-fresh was different in repeat even, different age (group) was different in these two even, and male and female were different in these two even.**

Keywords—Arteriovenous Fistula; Data Mining; Recurrent Event; Competing Risks; Cox Regression

I. INTRODUCTION

Patients' dependent on hemodialysis has steadily increased, with an estimated 31000 persons in Taiwan currently requiring long-term hemodialysis and more than two thousands of cases increased annually. Arteriovenous fistula (AV fistula) operation becomes the most performed operation of the reconstructive microsurgery team in Cathay General Hospital; about 150 to 200 fistulas are carried out annually.

Care of patients with end-stage renal disease has considerable economic burden to health insurance systems both in Taiwan and USA. The cost exceeds 1 billion US dollars and access related hospitalization accounts for 25% of all hospital admissions in the USA. Up to 17% total spending for hemodialysis per patient per year was associated with hemodialysis access- related morbidity. [1]

Continuous care and repeated interventions are necessary to maintain AV fistula patency. Quality assurance was especially important in the surgical intervention of this chronic disease. However, there are no population-based data available in Taiwan as the reference standards for quality management and improvement.

To understand the course of fistula patency and subsequent revisions, it could be drilled into datasets from the National Health Insurance Research Database (NHIRD) in Taiwan. The methodology of case follow-up and differentiation between fresh and revision, different age, and different gender cases would also contributed to other clinical studies of surgical procedures.

II. MATERIALS AND METHODS

National Health Insurance Research Database (NHIRD) covers 96% of people in Taiwan. Patient with major diseases were nearly completely included in NHI, and follow-up study of these subject would be more convincing.

A. Patient selection and the primary fistula operation

Based on the confirmed patients selected in the 1/20 sampled hospitalization database (SDD dataset) with more clinical information (DO dataset) including the detailed medical orders, it could further explore the non-sampled complete major disease dataset for the outcome after the confirmed vascular access operation, or explore the past history to differentiate between fresh and redo cases. The non-sampled, complete "major diseases" hospitalization (DD) and outpatient (CD) datasets were used for tracing the clinical course of these selected patients. The clinical details of non-sampled datasets are mainly derived from the five ICD diagnostic fields and the five ICD procedural fields, and could only be partly confirmed by the associated fee claims in different administrative categories. The clinical information in the latter (DD, CD) is relatively ambiguous, and should be used with reservation [4].

NHI procedure code "69032B" and "69032C" was the main criteria of inclusion of patients with AV fistula operation. 3866 cases were found in the 1/20-sampled hospitalization datasets from 1997 to 2010 with detailed ICD diagnostic and operation codes.

Considering the sampling rate of 1/20 for inpatient and outpatient datasets, the true case numbers performed in the 4 years were 77320. The latter should be used with conservation due to large variation.

It would be followed-the clinical course of these 3866 inpatient cases in complete major disease datasets 1997-2010. It would be use ICD-OP-Code to identified revision vascular operations according to different search strategy. It could be identified that 387 redo or revision operations had been performed.

B. Data Analysis

Selection used the 392.7, 394, 395, 393, 393.0 393.1, and 393.2 ICD-op-code for the query. The non-freshness of cases was determined by the presence of related ICD-OP-Code in the past history before the SDO confirmed events. The NHIRD data was kept in a MySQL data warehouse. The data mining were performed in IBM SPSS Modeler 14.1 package and R 3.1. 2.

C. Recurrent Event

Within the framework of the multiplicative or additive hazards regression models, a variety of models have been proposed and utilized in real applications. Among the rich selection of different models, the gap time model as an extension of the multiplicative Cox proportional hazards model [5] received the greatest attention due to easy interpretation of the covariate effects. This model assumed unspecified baseline hazards and constant covariate effects. In the models, we will assume that all censoring is non-informative and independent.

Suppose that there are n subjects and that each subject can experience K failures or recurrent events. Suppose that censoring is non-informative, which means that knowledge of a censoring time for a subject provides no further information about the subject's likelihood of survival at a future time. Let T_{ik} be the time when the kth failure occurs for the ith subject and C_{ik} be the corresponding censoring time. T_{ik} is measured from the subject's study enrollment and the censoring C_{ik} occurs after the subject has been entered into a study to the right of the last known failure time; thus, it is right censoring. When T_{ik} is subject to right censoring, the kth failure time X_{ik} is a minimum of (T_{ik}, C_{ik}), i.e., X_{ik} is equal to T_{ik} if the event was observed and is equal to C_{ik} if it is censored. Let δ_{ik} = $I(T_{ik} \le C_{ik})$, where I(.) is an indicator function and takes the value 1 when $T_{ik} \le C_{ik}$ and is 0 otherwise. Let Z_{ik} be a covariate vector of p-dimensions for the ith subject at the kth failure. For each of the K failures, the hazard function for the ith subject with respect to the kth failure $\lambda_{ik}(t)$, is assumed to take additive or multiplicative forms.

The gap time model requires the same assumptions as the Cox proportional hazards model, but they allow the baseline hazard to vary from recurrence to recurrence. Gap time is defined the time between two successive failures experienced by the same subject [5]. For the gap time model, the hazard function is

$$\lambda_{ik}(t) = \lambda_{ok}(t - t_{k-1})e^{\vec{\beta}'\vec{Z}_{ik}(t)}$$

where t is the time since a patient's study enrollment and t_{k-1} is the time of the (k-1)th failure. Note that $\lambda_{0k}(t)$ are unspecified baseline hazard functions varying with k = 1, .., K. The corresponding partial likelihood function [16,17] is

$$L(\vec{\beta}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left\{\frac{e^{\beta'\vec{Z}_{ik}(X_{j,k-1}+G_{ik})}}{\sum_{j=1}^{n}Y_{jk}(G_{ik})e^{\beta'\vec{Z}_{jk}(X_{j,k-1}+G_{ik})}}\right\}^{\delta_{ik}}$$

where $G_{i,k} = X_{ik} - X_{i,k-1}$ is the inter-event or gap time interval and $Y_{jk}(t) = I(G_{ik} \geqslant t)$ is a risk set indicator. $\vec{\beta}$ is a p-vector of regression coefficients of Z_i.

In order to draw a semi-parametric inference on $\vec{\beta}$ for the model (1), the score functions U($\vec{\beta}$) are obtained by differentiating the logarithm of L($\vec{\beta}$) with respect to $\vec{\beta}$.

The maximum partial likelihood estimator $\hat{\vec{\beta}}$ is obtained by solving the corresponding score equation, $\frac{\partial \ln L(\vec{\beta})}{\partial \vec{\beta}} = 0$. When failure times are independent, the variance-covariance matrix is estimated from the inverse of the information matrix, $I^{-1}(\hat{\vec{\beta}})$, called the naïve variance-covariance matrix; however, when failure times are dependent, $I^{-1}(\hat{\vec{\beta}})$ is not a good estimator of the variance-covariance matrix. When there are dependencies, the variance-covariance matrix $\hat{Q}(\hat{\vec{\beta}})$, the so-called sandwich or robust variance-covariance estimator, is obtained from $\hat{Q}(\hat{\vec{\beta}}) = I^{-1}(\hat{\vec{\beta}})V(\hat{\vec{\beta}})I^{-1}(\hat{\vec{\beta}})$, where $V(\hat{\vec{\beta}})$ is a data-based estimator, i.e., the cross-product of the empirical score residual matrix $W(\hat{\vec{\beta}})$. Here, $V(\hat{\vec{\beta}}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}W_{ik}(\hat{\vec{\beta}})W_{il}(\hat{\vec{\beta}})'$, and

$$W_{ik} = \delta_{ik}\left\{Z_{ik}(X_{i,k-1}+G_{ik}) - \frac{S_i^{(l)}(X_{i,k-1}+G_{ik})}{S_i^{(O)}(X_{i,k-1}+G_{ik})}\right\} - \sum_{j=1}^{n}\frac{\delta_{ij}Y_{ik}(G_{ik})^{\hat{\beta}'Z_{ik}(X_{k,m-1}+G_{ik})}}{S_i^{(O)}(X_{i,k-1}+G_{ik})} * \left\{Z_{ik}(X_{i,k-1}+G_{ik}) - \frac{S_i^{(l)}(X_{i,k-1}+G_{ik})}{S_i^{(O)}(X_{i,k-1}+G_{ik})}\right\}$$

where

$$S_i^{(l)}(X_{i,k-1}+G_{ik}) = \sum_{j=1}^{n}Y_{jk}(G_{ik})Z_{jk}(X_{i,k-1}+G_{ik})e^{\vec{\beta}'Z_{jk}(X_{jk}+G_{ik})}$$

and

$$S_i^{(l)}(X_{i,k-1}+G_{ik}) = \sum_{j=1}^{n}Y_{jk}(G_{ik})Z_{jk}e^{\vec{\beta}'Z_{jk}(X_{jk}+G_{ik})}$$

D. Competing Risks

Consider the competing risks setting where the data consist of failure times for different subjects and where failure is categorized into several distinct and exclusive types. In this paper a method is given for comparing over time the probability of failures of a certain type being observed among different groups. To be precise, suppose there are K independent groups of subjects, and let T^0_{ik} be the failure time of the ith subject in group k, i =1,.. .,n_k, and δ^0_{ik} be the type of failure, δ^0_{ik} = 1,.. . , J. The pairs (T^0_{ik}, δ^0_{ik}) from different subjects in a group are assumed to be independent and identically distributed. However, it is not assumed that the underlying processes leading to failures of different types are acting independently for a given subject. Rather, only quantities which can be identified from the observed data, regardless of whether or not the risks are independent, will be used.

Denote the sub-distribution function for failures of type j in group k by

$$F_{jk}(t) = P(T^0_{ik} \le t, \delta^0_{ik} - j)$$

This will be called the cumulative incidence function for failures of type j here. The main subject of this paper is to develop tests for the hypothesis

$$H_0 : F_{1k} = F^0_1, \text{ k=1,…,K,}$$

where F^0_1 is an unspecified sub-distribution function and where the failure type of special interest is taken to be type 1.To simplify the presentation, the $F_{jk}(t)$ are assumed to be continuous with sub-densities $f_{jk}(t)$ with respect to Lebesgue measure [6].

TABLE 1: AV fistula patient database analysis

Year	Patient Age	Surgery costs	Hospital costs	Hospitalization days
1997	45.86(10.09)	17305.67(16420.97)	80750.63(86086.64)	11.89(10.09)
1998	48.10(13.25)	17867.12(16740.25)	79295.73(95411.58)	12.10(13.25)
1999	48.55(10.73)	15133.71(12102.65)	62973.01(54867.29)	10.47(10.73)
2000	51.33(11.12)	15303.49(12308.77)	70738.03(69118.67)	11.61(11.12)
2001	51.39(9.99)	15365.91(11280.06)	77356.42(82875.72)	10.91(9.88)
2002	53.99(11.06)	16625.89(14536.57)	81600.51(99584.51)	11.88(11.06)
2003	55.71(13.63)	16865.58(16622.57)	92368.77(115123.36)	13.11(11.06)
2004	57.12(20.48)	18917.37(20339.20)	109251.93(129024.97)	13.76(13.63)
2005	57.99(16.80)	20016.72(20415.86)	130260.65(180002.25)	17.33(20.48)
2006	58.00(15.28)	20027.02(30763.78)	132855.80(234276.77)	15.46(16.80)
2007	59.75(13.93)	22209.53(28884.64)	142106.23(229555.37)	14.80(15.28)
2008	61.82(18.01)	23095.28(41850.53)	154911.72(224103.85)	16.91(13.93)
2009	60.20(18.01)	21242.00(23605.67)	147951.86(204165.62)	15.87(18.01)
2010	60.18(14.61)	24747.17(36005.08)	140162.71(190037.23)	15.04(14.61)

III. RESULTS

A. Characteristics of Patients and the Primary Fistula Operation

Average age of the patients was around 45-60 years old (as Table 1). Male and female were equally 2139 (55.33%) and 1724 (44.59%). The sex was unknown in 3 patients (.08%). Average length of stay, the total operation fees, and the ward expenses in the whole hospitalization periods were listed in Table 1.

The patients were mostly admitted to Nephrology departments (1817 cases, 47.00%). 404 cases were Cardiovascular Surgery department (10.45%). 381 cases were Surgery department (9.86%). 357 cases were Plastic Surgery department (9.23%) (Table 2).

Since the cases were derived from inpatient dataset, the numbers managed in public hospitals and private hospitals were 1152 (29.80%) and 2714 (70.20%), respectively.

TABLE 2: Admitted department of AV fistula patient

Department	Cases	%
Nephrology	1817	47.00
Cardiovascular Surgery	404	10.45
Surgery	381	9.86
Plastic Surgery	357	9.23
Medicine	342	8.85
Orthopedics	194	5.02
Cardiovascular Medicine	100	2.59

Others	271	7.00
Total	3866	100.00

TABLE 3: Age distribution of primary fistula operation

GROUP	1	2	3	4	5	6	7	
AGE	0-19	20-29	30-39	40-49	50-59	60-69	70-79	Total
1997	9	15	39	45	27	37	6	178
1998	11	16	41	51	28	35	27	209
1999	13	21	36	38	27	45	26	206
2000	14	16	25	43	44	41	41	224
2001	6	22	40	38	41	49	42	238
2002	4	19	27	49	50	67	47	263
2003	8	22	28	55	67	83	77	340
2004	8	14	25	44	70	77	77	315
2005	5	19	31	44	79	74	98	350
2006	15	18	27	29	69	69	110	337
2007	6	8	15	32	67	74	79	281
2008	7	7	13	34	64	68	107	300
2009	10	8	17	37	50	79	98	299
2010	7	9	22	41	67	76	104	326
Total	123	214	386	580	750	874	939	3866

B. The Vascular Revision Even

The vascular revision events after the primary fistula operation were illustrated by the number of events versus the interval to the primary operation (as Figure 1). The intervals between primary and revision fistula operation were analyzed in Table 4. The 387 events occurred in 3374 cases, and 387 cases received multiple revisions in the follow-up period. The total number of revision operations for each patient was listed in Table 5. 309 patients received two revisions, and 56 patients, three revisions, and 22 patients, more than four revisions

C. Recurrent Event

For calculation of the primary patency rate, it would be count the first revision for each patient in the subsequent presentations. The survival curves of cumulative proportion methods were drawn for all cases, for the fresh cases versus non-fresh cases, for different age groups, and for different gender groups.

It could be analyzed primary patency of AV fistula operation as revision-free survival time in Figure 1. The overall patency rate is above 80.00% in 600 weeks. The survival curves were different in fresh and non-fresh cases (as Figure 2 and Table 4). The survival probability of non-fresh patient was
1.64(1/.61) times on fresh patients. Survival of different age groups was shown in Figure 3 and Table 4. Age group 50-59, age group 60-69, and age group 70- were significantly different from age group -19. The survival probability of age 60-69 patient was .23(1/4.38) times on age group -19 patients. The survival curves were different in male and female cases (As Figure 4 and Table 4). The survival probability of male patient was 1.39(1/.72) times on female patients.

D. Competing Risks

NHI procedure code “69032B” and “69032C” was the main criteria of inclusion of patients with AV fistula operation. 68032B was vascular repair, and 68032C was repair and anastomosis of peripheral vascular. The 69032B-69032B and 69032C-69032C be coded with 1 (repeat). The 69032B-69032C and 69032C-69032B be coded with 2 (change). The 69032B-censor and 69032C-censor be coded with 0 (censor).

With Table 5 and Figure 4, fresh and non-fresh was different in repeat even. With Table 5 and Figure 5, different age (group) was different in these two even. With Table 5 and Figure 5, male and female were different in these two even.

TABLE 4: The result of Recurrent Event Survival Analysis

								95% CI for Exp(B)	
		B	SE	Wald	df	Sig	Exp(B)	Lower	Upper
Fresh				6.64					

	Non-Fresh	-.50	.20		1	.01	.61	.42	.89
Age				40.89					
	20-29	.11	0.508		6	.86	1.11	.35	3.58
	30-39	.78	0.438		6	.14	2.18	.78	6.12
	40-49	.94	0.426		6	.07	2.56	.93	7.01
	50-59	1.04	0.422		6	.04	2.84	1.04	7.72
	60-69	1.48	0.417		6	<.01	4.38	1.62	11.80
	70-	1.28	0.419		6	.01	3.59	1.32	9.71
Gender				12.3					
	Male	-.324	0.09		1	<.01	.72	.60	.87

TABLE 5: The Result of Competing Risks Analysis

	Event	stat	df	p-value
Fresh	1. Repeat	7.89	1	<.01
	2. Change	.40	1	.53
Age	1. Repeat	46.27	6	<.01
	2. Change	16.75	6	<.01
Gender	1. Repeat	5.83	1	.02
	2. Change	12.65	1	<.01

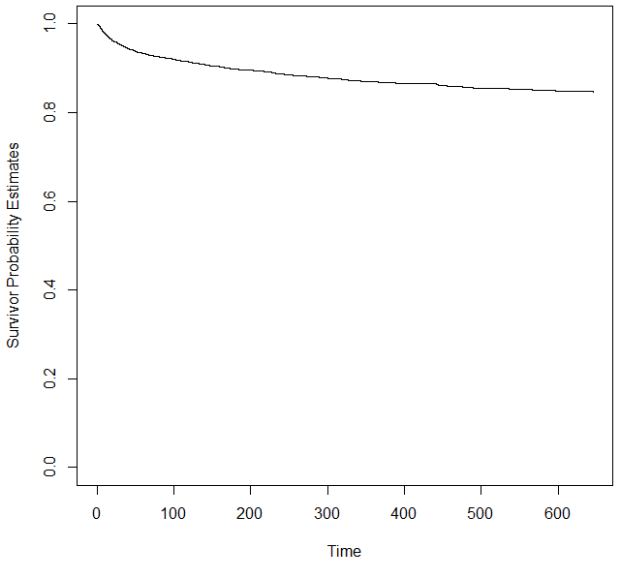

Figure 1: Survival curve of primary fistula operation

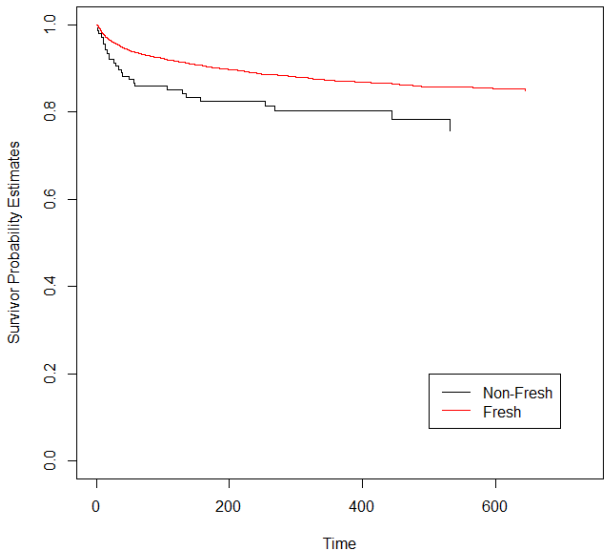

Figure 2: Survival curve of fresh/non-fresh fistula operation

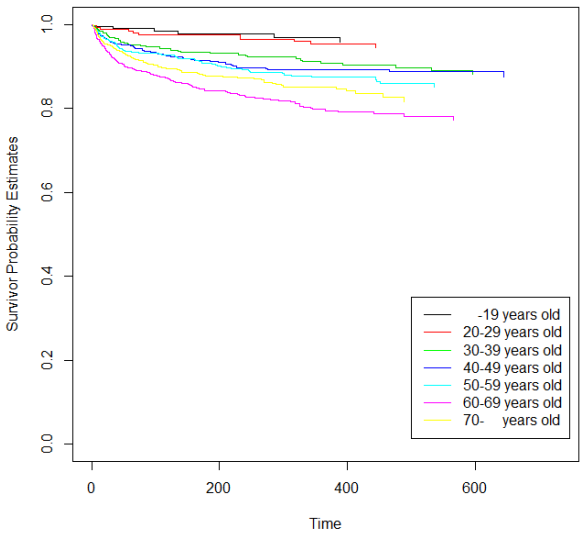

Figure 3: Survival curve of primary fistula operation in different age groups

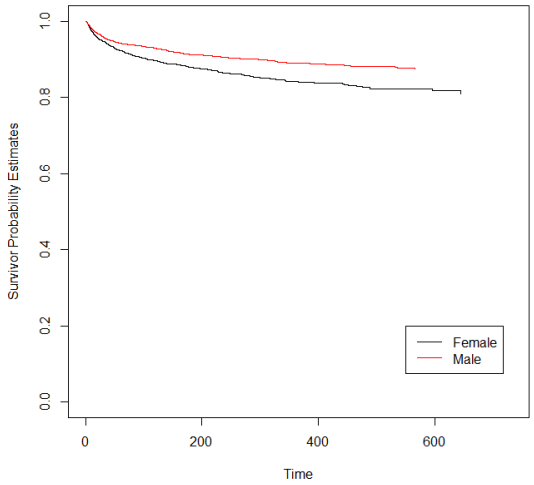

Figure 4: Survival curve of primary fistula operation in different sex groups

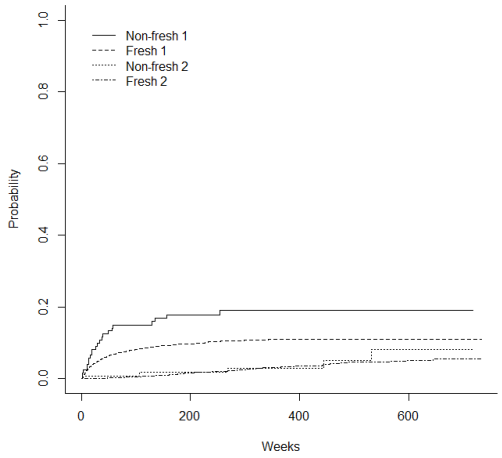

Figure 5: Cumulative incidence curves in fresh/non-fresh fistula operation

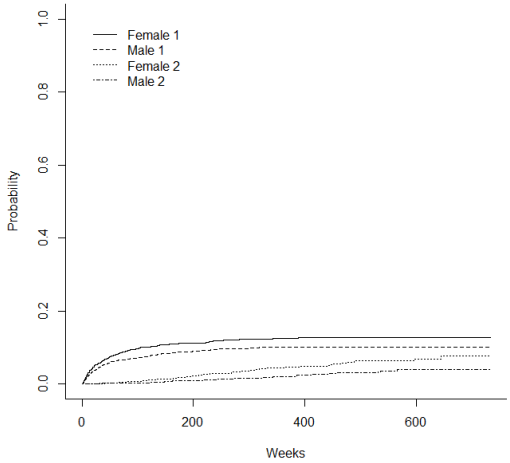

Figure 6: Cumulative incidence curves in female/male

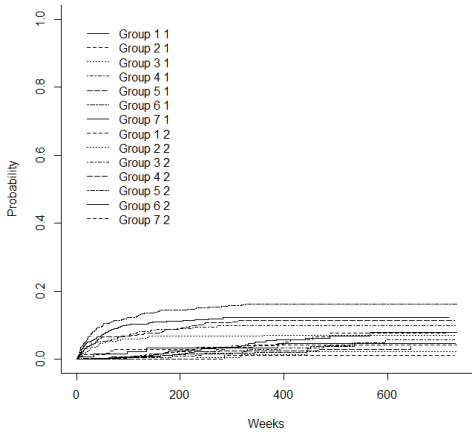

Figure 6: Cumulative incidence curves in age

IV. DISCUSSIONS AND CONCLUSION

This study focus on the outcome analysis of AV fistula operation in Taiwan based on insurance claim database. Since NHI covers 96% of population, and end-stage renal disease was one of the major diseases included in the initial setup of NHI, the high recruitment rate and coverage of various practice patterns of different hospitals were the most valuable part of this study.

The other contribution of this study was validation of the methodology of case follow-up in insurance claim database by the relatively homogeneous problem of the subsequent vascular revisions. Although the study of NHRID was limited by the more detailed clinical procedures in inpatient datasets than in outpatient datasets, the large case (big data) base and the inclusion of various practice patterns of all hospital types still makes this methodology important for follow-up study of surgical disease or procedures.

This data was limited mainly to the inpatients. The proportion performed in the medical center may be exaggerated. Accordingly, the severity of disease and the associated disease may be aggravated. Both factors may affect the patency rate and increase the complications, medical expenses, and length of hospital stay.

In addition to the inpatient-only approach, there are other limitations inherent in the secondary use of insurance claim database for clinical study such as ICD-only information and no control group.

The overall patency rate is above 80.00% in 600 weeks. The survival curves were different in fresh and non-fresh cases. The survival probability of non-fresh patient was 1.64 times on fresh patients. Survival of different age groups was shown in Figure 3 and Table 4. Age group 50-59, age group 60-69, and age group 70- were significantly different from age group -19. The survival probability of age 60-69 patient was .23 times on age group -19 patients. The survival curves were different in male and female cases. The survival probability of male patient was 1.39 times on female patients.

With competing risks analysis, fresh and non-fresh was different in repeat even, different age (group) was different in these two even, and male and female were different in these two even.

It could be prove the validity of follow-up methodology in the claim database by ICD-code selections. The events before the primary surgery could help the differentiation between freshness of cases.

V. 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.

REFERENCES

[1] Hakim, R., & Himmelfarb, J. (1998). Hemodialysis access failure: A call to action. Kidney International, 54(4), 1029-1040.

[2] Woods, J. D., & Turenne, M. N., Strawderman, R. L., Young, E. W., Hirth, R. A., Port, F. K., & Held, P. J. (1997). Vascular access survival among incident hemodialysis patients in the United States. American Journal of Kidney Diseases, 30(1), 50-57.

[3] Feldman, H. I., Held, P. J., Hutchinson, J. T., Stoiber, E., Hartigan, M. F., & Berlin, J. A. (1993) Hemodialysis vascular access morbidity in the United States. Kidney International, 43(5), 1091-1096.

[4] National Health Research Institutes (2014). National Health Insurance Research Database (NHIRD), Retrieved December 1, 2014 from http://w3.nhri.org.tw/nhird/

[5] Prentice, R.L., Williams, B.J., & Peterson, A.V. (1981). On the regression analysis of multivariate failure time data. Biometrika, (68), 373-379.

[6] Gray, R.J. (1988). A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. The Annals of Statistics, 16(3), 1141-1154.