Hospital Management Data Mining towards Chance Discovery

Yuko Tsumoto[1] and Shusaku Tsumoto[2]

[1]Department of Fundamental Nursing, Shimane University, School of Nursing , Izumo, Japan
[2]Department of Medical Informatics, Shimane University, School of Medicine, Izumo, Japan

Abstract

Rapid progress in information technology has come to enable us to store all the information in a hospital information system, including management data, patient records, discharge summary and laboratory data. Although the reuse of those data has not started, it has been expected that the stored data will contribute to analysis of hospital management. In this paper, the discharge summary of Chiba University Hospital, which has been stored since 1980's were analyzed to characterize the university hospital. The results show several interesting results, which suggests that the reuse of stored data will give a powerful tool to support a long-period management of a university hospital.

Keywords: Medical Data Mining, Chance Discovery

1. Introduction

It has passed about twenty years since clinical information are stored electronically as a hospital information system since 1980's. Stored data includes from accounting information to laboratory data and even patient records are now stared to be accumulated: in other words, a hospital cannot function without the information system, where almost all the pieces of medical information are stored as multimedia databases [1]. Especially, if the implementation of electronic patient records is progressed into the improvement on the efficiency of information retrieval, it may not be a dream for each patient to benefit from the personal database with all the healthcare information, "from cradle to tomb". However, although the studies on electronic patient record has been progressed rapidly, reuse of the stored data has not yet been discussed in details, except for laboratory data and accounting information to which OLAP methodologies are applied. Even in these databases, more intelligent techniques for reuse of the data, such as data mining and classical statistical methods has just started to be applied from 1990's [2,3].

Human data analysis is characterized by a deep and short-range investigation based on their experienced "cases", whereas one of the most distinguished features of computer-based data analysis is to enable us to understand from the different viewpoints by using "cross-sectional" search. It is expected that the intelligent reuse of data in the hospital information system provides us to grasp the all the characteristics of university hospital and to acquire objective knowledge about how the hospital management should be and what kind of medical care should be served in the university hospital.

This paper focuses on the following two points for analysis. One is what kind of knowledge can be extracted by statistical methods from the datasets stored for about twenty years in Chiba University Hospital. The other is how these pieces of knowledge are useful for the future hospital management and decision support. The analysis gives interesting and results. For example, the discharge summaries show that most of the patients are admitted to the university hospital because of the diseases which requires the advanced treatment, such as malignant neoplasm. Combination of the discharge summaries and data in the hospital accounting systems shows that the profitability significantly differs in each disease, but that within each disease, the number of days in the hospital is a principal factor for the profitability. Also, most of the distributions of the number of days in the hospital for the diseases do not follow the normal distributions, but log-normal distributions, which influence the profitability of the university hospital. The reason why the distributions follow the log-normal distribution should be investigated in the near future.

The objectives of this research is to investigate what kind of knowledge can be extracted by statistical methods from the datasets stored in the hospital information system of Chiba University Hospital, especially useful for future hospital management and decision support. Especially, since the revenue of Japanese hospital is based on NHI points of Japanese medical care, it is important to investigate the factor which determines the amount of NHI points.

2. Methods

2.1. Data

When the hospital information system for discharge summaries is introduced in Chiba University Hospital in 1978, a discharge summary is distributed to doctors as a paper sheet for each patient admitted to the hospital. Doctors fill in each sheet just after the patient leaves the hospital, the parts of this sheet which can be coded are stored electronically. A sheet for discharge summary is composed of the items common to all the departments and the items specific to each department. For example, the items specific to neurology consists of the results of neurological examinations and the severity of diseases. The common items consist of those in which codes or numerical values should be filled in and those in which texts should be input. After the doctor in charge fill in those items and submit to the division of medical records, the staff input codes and numerical values into a database system. These processes are continued until a new hospital information system was introduced in 2000, which means that the non-text items common to all the departments has been stored for about 20 years.

There are 16 items for codes or numerical values: patient ID, the department in charge, occupation, height and weight on admission, height and weight just before hospital discharge, a motivation for visit, outcome, autopsy or not, cause of death, the date of first visit, the date of admission, the date of discharge, the name of disease (ICD-9 code [4]), treatment method. However, height and weight just before hospital discharge are not input in the database.

Concerning the items specific to each department, only those of surgery and ophthalmology are stored electronically.

2.2. Extraction of Datasets from Hospital Information System

Datasets for analysis are extracted from the database on discharge summaries and the database on patient basic information by using patient ID and the date on admission as keys. The program for extraction is developed by the first author due to the following reasons. Since NHI points, which stands for National Healthcare Insurance points, are stored for each patient ID and each month, the total points for each admission for each patient are calculated from NHI points for each month. The total points are combined with the dataset extracted from the discharge summaries by using patient ID and the date on admission as keys.

The number of the records of the dataset extracted from the global: MRMG, which is a database on discharge summaries, is 157,636 for 21 years from 1978.4 to 2000.3. The time needed for computation is about one hour by SUN Workstation (Enterprise 450). Concerning the dataset combined with NHI points, the number of the records is 20,146 for three years from 1997.4 to 2000.3.

2.3. Statistical Analysis

Descriptive statistics, exploratory data analysis and statistical tests were applied to the dataset extracted only from the discharge summaries for the analysis of patient basic information (gender, age and occupation), outcome, the number of the days in hospitals and diseases, including their chronological trends. Concerning the datasets combined with accounting information for three years (1997.4 to 2000.3), the relations among NHI points and items in the discharge summaries were analyzed by descriptive statistics, exploratory data analysis, statistical tests, regression analysis and generalized linear model. SPSS 11.0J for windows was used for these analyses..

3. Results

Due to the limitation of the spaces, the most interesting results are shown in this section. In the subsequent subsections, the results of the whole cases, and two levels of ICD-9 code, called major and minor divisions, are compared. Especially, concerning the results for the major and minor divisions, malignant neoplasm and the following three largest minor divisions of the malignant neoplasm are focused on: neoplasm of trachea, bronchus, and lung, neoplasm of stomach, and neoplasm of liver and intrahepatic bile ducts. In the subsequent sessions, neoplasm of lung, stomach and liver denotes the above three divisions for short.

3.1. Correlation between Length of Stay and NHI Points

Figure 1 depicts the scattergram between the length of stay and NHI points of the whole cases, which suggests a high correlation between two variables. For simplicity, the vertical and horizontal axes show the logarithm of raw values. Actually, The coefficient of correlation is calculated as 0.837, which means that the correlation is very strong. It is notable

that the coefficient of malignant neoplasm, whose coefficient is 0.867.

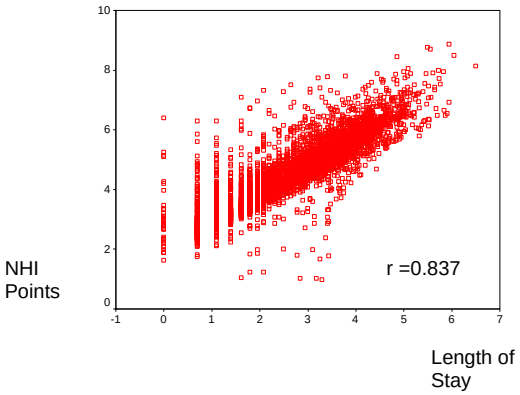

Fig. 1. Correlation between Length of Stay and NHI Points (Whole Cases)

Concerning the three largest minor divisions of neoplasm, neoplasm of lung and stomach have strong positive correlations, whose coefficients are larger than 0.8. On the other hand, the correlation of neoplasm of liver is weaker than them, whose coefficient is 0.71. These results are shown in Table III with those of major and minor revisions.

TABLE III
CORRELATION BETWEEN NHI POINTS AND LENGTH OF STAY

	Total Cases	With Operation	Without Operation
Total Cases	0.837	0.829	0.779
Neoplasm	0.867	0.844	0.826
Malignant Neoplasm of Trachea, Bronchus, and Lung.	0.838	0.648	0.903
Malignant Neoplasm of Stomach.	0.827	0.738	0.801
Malignant Neoplasm of Liver and Intrahepatic Bile Ducts.	0.711	0.577	0.755

Table III summarized the correlation coffecients between NHI points and Length of Stay with respect to the whole cases, neoplasm and three major types of malignant neoplasm: lung, stomach and liver. Comparison of the coefficient of correlation between the group with and without a surgical operation shows that the group without an operation has higher correlations than that with an operation, which suggests that NHI points of the treatment methods other than surgical operations should be strongly dependent on the lengths of stay.

Surprisingly, such strong correlations not only in general category, but also minor divisions are not expected in hospital management field.

3.2. Generalized Linear Model

Since all the items except for the length of stay are categorical variables, conventional regression models cannot be applied to the study on relations between NHI points and other items. For this purpose, generalized linear model [7] was applied to the dataset on combination of accounting data and discharge summaries. NHI point was selected as a target variable and the following four variables were selected as explanatory variables: outcome, treatment method, major division of ICD-9 codes and the categorized length of stay. The length of stay is categorized so that the distribution of the transformed variable is close to normal distribution, where the width of windows is set to 0.5 for the logarithmic value of the length of stay. Treatment, outcome and major divisions of ICD codes are transformed into dummy variables to clarify the contributions of these values to a target variable. For example, the outcomes of discharge are split into the following six dummy variable: D1: recovered, D2: improved, D3: unchanged, D4: worsened, D5: dead and D6: others.

Table VI shows the results of GLM on the total cases, whose target variable is NHI points. All the variables are sorted by the F value. The most contributing factor is the length of stay, whereas the contributions of the other factors are small. The adjusted R-square value is 0.704.

Table VII gives the results of GLM on major division (malignant neoplasm), and three minor divisions, whose target variable is NHI points. Compared with Table V, the number of the factors which gives the significant contributions to NHI points are very small, which suggest that the variabilities of NHI points in major and minor divisions are very low, compared with that of total cases.

4. Conclusion

This paper analyzes the following two datasets extracted from the hospital information system in Chiba University Hospital. One is the dataset extracted from the database on discharge summaries stored for about twenty years from 1978 to 2000. The other is combination of data from discharge summaries and accounting information system. The analysis gave the following results: (1) malignant neoplasm is the first major category which determines

the profitability of Chiba University Hospital, which is stable for twenty years. (2) In a global view, the length of stay is the principle factor for the revenue of the hospital, whose distribution follows the log-normal distribution. (3) Treatment method may be a secondary factor to determine the distribution of the length of stay for each disease, which may be correlated with the property that the length of stay follows log-normal distribution for each minor division in total. (4) Treatment without a surgical operation should be more examined by additional information, which is also important to evaluate the profitability of the university hospital.

Acknowledgements

The authors would like to thank Prof. Okazaki, Prof. Mitsuoka, Prof. Takabayashi and Prof. Satomura for insightful discussions and preparation for the manuscript. They also thank Ms. Koya and Ogawa for support on this research.

TABLE VII
RESULTS OF GLM ANAYSIS (MAJOR AND MINOR DIVISIONS)

Neoplasm		
	F value	P
Length of Stay	625.8	<0.001
Death	36.0	<0.001
Operation	23.2	<0.001
Changeless	16.4	<0.001
Other Therapies	9.8	0.002
Radiation	8.5	0.003
Exacerbation	8.2	0.004
	R square =	0.753

Malignant Neoplasm of Stomach.		
	F value	P
Length of Stay	45.5	<0.001
Other Consequences	4.7	0.031
	R square =	0.719

Malignant Neoplasm of Trachea, Bronchus, and Lung.		
	F value	P
Length of Stay	58.8	<0.001
Operation	16.2	<0.001
Changeless	8.6	0.004
Exacerbation	5.9	0.016
	R square =	0.751

Malignant Neoplasm of Liver and Intrahepatic Bile Ducts.		
	F value	P
Operation	27.4	<0.001
Length of Stay	20.5	<0.001
Death	5.0	0.026
	R square =	0.582

References

[1] *[1] Institute of Medicine Committee on Improving the Patient Record. The Computer-based Patient record: An Essential Technology for Health Care. Washington DC: National Academy Press; 1997.*

[2] *Tsumoto S. Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. Inf. Sci 2000; 124(1-4): 125-137.*

[3] *Tsumoto S. Chapter G5: Data mining in medicine, In: Kloesgen W, Zytkow J, editors. Handbook of Data Mining and Knowledge Discovery. Oxford: Oxford University Press; 2001. p.798-807.*

[4] *Online ICD9/ICD9CM codes and Medical Dictionary. URL: http://icd9cm.chrisendres.com/*

[5] *Richard Walters. M Programming: A Comprehensive Guide. Woburn: Butterworth-Heinemann; 1997*

TABLE VI
RESULTS OF GLM ANAYSIS (WHOLE DATA)

Whole Data		
	F value	P
Length of Stay	1590.3	<0.001
Death	347.7	<0.001
Mental Disorders	264.4	<0.001
Complications of Pregnancy, Childbirth, and the Puerperium	241.7	<0.001
Diseases of the Circulatory System	228.2	<0.001
Operation	119.4	<0.001
Certain Conditions Originating in the Perinatal Period	98.2	<0.001
Exacerbation	56.2	<0.001
Other Therapies	53.9	<0.001
Diseases of the Nervous System and Sense Organs	52.1	<0.001
Medication	42.9	<0.001
Artificial Organ	42.2	<0.001
Diseases of the Skin and Subcutaneous Tissue	42.0	<0.001
Changeless	38.6	<0.001
Diseases of the Genitourinary System	30.7	<0.001
Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders	29.0	<0.001
Symptoms, Signs, and Ill-Defined Conditions	28.5	<0.001
Radiation	19.3	<0.001
Dietary Care	17.3	<0.001
Rehabilitation	14.8	<0.001
Diseases of the Respiratory System	11.8	0.001
Lighthearted	11.6	0.001
Transfusion/Infusion	6.6	0.010
Diseases of the Blood and Blood-Forming Organs	6.3	0.012
Infectious and Parasitic Diseases	5.6	0.018
Diseases of the ear and mastoid process	3.9	0.048
Diseases of the Musculoskeletal System and Connective Tissue	3.9	0.049