2010 IEEE International Conference on Data Mining Workshops

Towards Data-Oriented Hospital Services: Data Mining-based Hospital Management

Shusaku Tsumoto and Shoji Hirano and Yuko Tsumoto
Department of Medical Informatics,
Shimane University, School of Medicine
89-1 Enya-cho, Izumo 693-8501 Japan

***Abstract*—It has passed about twenty years since clinical information are stored electronically as a hospital information system since 1980's. Stored data include from accounting information to laboratory data and even patient records are now started 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. In this paper, we applied temporal data mining and exploratory data analysis techniques to hospital management data. The results show several interesting results, which suggests that the reuse of stored data will give a powerful tool for hospital management.**

I. Introduction

It has passed about twenty years since clinical information are stored electronically as a hospital information system since 1980's. Stored data include from accounting information to laboratory data and even patient records are now started 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. 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[1], [2].

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 univer-sity 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.

In this paper, we applied several exploratory data analysis techniques to data extracted from hospital information systems. 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.

II. Our Goal

Fig 1 shows our goal for hospital services, which consists of the following three layers of hospital management: services for hospital management, sevices for medical staff and services for patients. Data mining in hospital information system plays a central role in achieving these layers.

III. Background

A. Hospital Information System: Cyberspace in Hospital

On the other hand, clinical information have been stored electronically as a hospital information system (HIS). The database stores all the data related with medical actions, including accounting information, laboratory examination, treatement and patient records described by medical staffs. Incident or accident reports are not exception: they are also stored in HIS as clinical data. For example, Figure 2 shows the structure of the HIS in Shimane University Hospital. As shown in the figure, all the clinical inputs are shared through the network service, where medical staff can retrieve their information from their terminal [3], [4].

Thus, this system can be a cyberspace in a hospital where all the results of medical actions are stored with temporal information. This cyberspace can be viewed as distributed large-scale and multimodal spatiotemporal databases.

Thurs, dealing with cyberspace in a hospital will a new challenging problem in hospital administration, and of course spatiotemporal data mining play a central role in this challenge.

B. Basic Atom in HIS: Order

The basic atom in HIS is an "order", which is a kind of document which conveys an order from medical staff to medical staff. For exmaple, prescription can be viewed as order from a doctor to a pharmacist and an prescription order is executed as follows.

1) Outpatient Clinic
2) A prescription given from a doctor to a patient

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.178

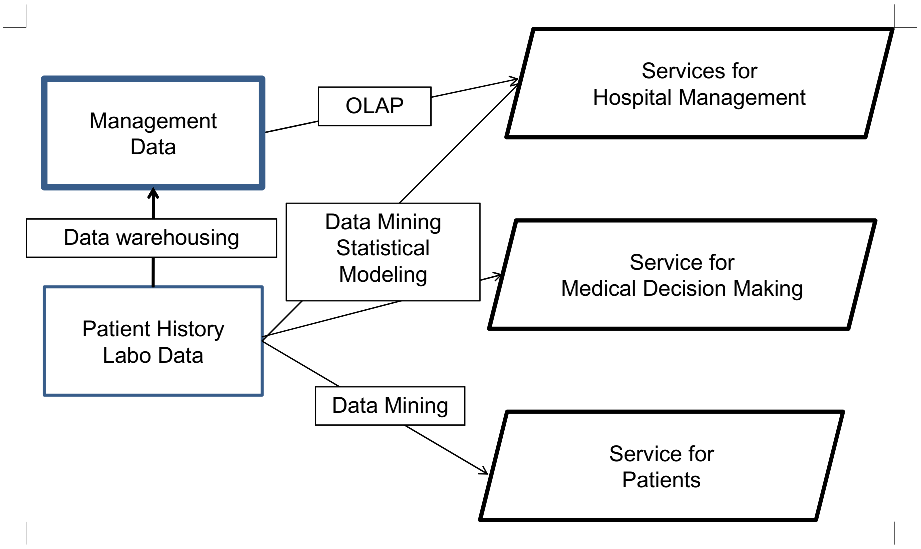

Figure 1. Service-Oriented Hospital Management

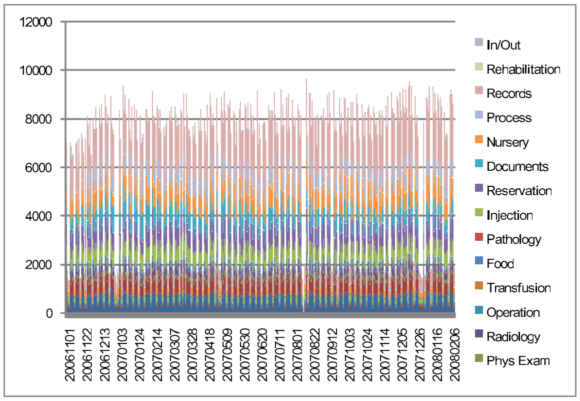

Figure 3. Total Number of Orders

3) The patient bring it to medical payement department
4) The patient bring it to pharmaceutical department
5) Exceution of order in pharmacist office
6) Delivery of prescribed medication
7) Payment

The second to fourth steps can be viewed as information propagation: thus, if we transmit the prescription through the network, all the departments involved in this order can easily share the information and execute the order immediately. This also means that all the results of the prescription process are stored in HIS.

These sharing and storing process can be easily stored as a database by using conventional IT and DB technologies: thus, HIS can also be viewed as cyberspace of medical orders.

C. Visualizing Hospital Actions from Data

Let us show the primitive mining results of HIS. Figure 3 depicts the chronological overview of the number of orders from November 1, 2006 to February 8, 2008. Vertical axis denotes the number of orders for each day, classfied by the type of orders. Horizonal axis gives each date. The plot shows that the temporal behavior of each order is periodical with respect to holidays and very stationary.

Let us show the primitive mining results of HIS. Table I shows the averaged values of the number of orders during the same period. Although these values do not remove the effects of holidays, all the characteristics reflects that shown in Figure 2. Records and Neursery accounts for 39% of orders and except for them, prescription, reservation of clinics, injection are top three orders in the hospital.

Although the above figure and table overview the total behavior of the hospital, we can also check the temporal trend of each order as shown in Figure 4 and 5. The former figure depicts the chronological overview of the number of each order from June 1 to 7, 2008, and the latter shows that of June 2, 2008. Vertical axes denote the aveeraged number of each order, classfied by the type of orders. Horizonal axie give each time zone. The plots show the characteristics of

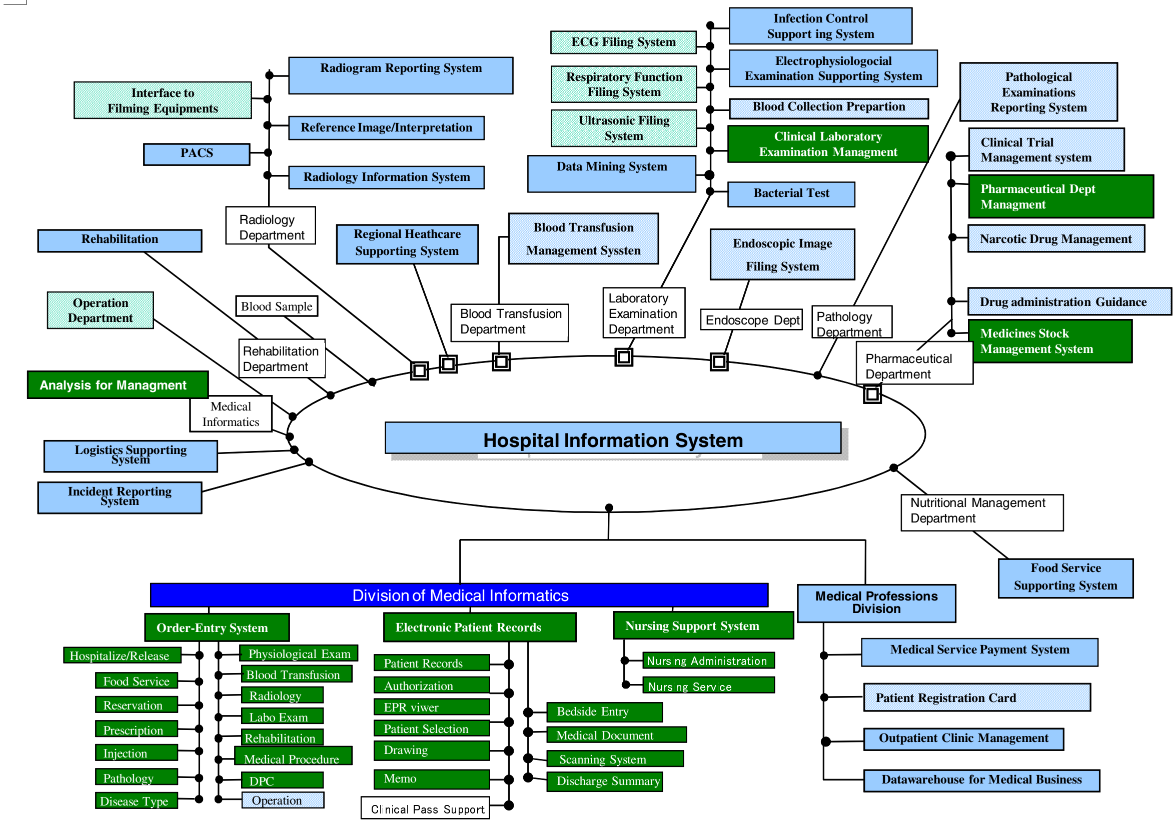
Infection Control Support ing System
ECG Filing System
Electrophysiologocial Examination Supporting System
Radiogram Reporting System
Respiratory Function Filing System
Blood Collection Prepartion
Pathological Examinations Reporting System
Interface to Filming Equipments
Reference Image/Interpretation
Ultrasonic Filing System
Clinical Laboratory Examination Managment
Clinical Trial Management system
PACS
Radiology Information System
Data Mining System
Bacterial Test
Pharmaceutical Dept Managment
Radiology Department
Regional Heathcare Supporting System
Blood Transfusion Management Sysstem
Endoscopic Image Filing System
Narcotic Drug Management
Rehabilitation
Operation Department
Blood Sample
Blood Transfusion Department
Laboratory Examination Department
Endoscope Dept
Pathology Department
Drug administration Guidance
Medicines Stock Management System
Rehabilitation Department
Pharmaceutical Department
Analysis for Managment
Medical Informatics
Hospital Information System
Logistics Supporting System
Incident Reporting System
Nutritional Management Department
Food Service Supporting System
Division of Medical Informatics
Medical Professions Division
Order-Entry System
Electronic Patient Records
Nursing Support System
Hospitalize/Release
Physiological Exam
Patient Records
Nursing Administration
Medical Service Payment System
Food Service
Blood Transfusion
Authorization
Nursing Service
Reservation
Radiology
EPR viwer
Patient Registration Card
Prescription
Labo Exam
Patient Selection
Bedside Entry
Injection
Rehabilitation
Drawing
Medical Document
Outpatient Clinic Management
Medical Procedure
Scanning System
Pathology
DPC
Memo
Discharge Summary
Datawarehouse for Medical Business
Disease Type
Operation
Clinical Pass Support

Figure 2. Hospital Information System in Shimane University

Table I
AVERAGED NUMBER OF ORDERS

Order	Average	Percentage
Prescription	580.6017131	9.91%
Labo Exam	427.4282655	7.30%
Phys Exam	94.90149893	1.62%
Radiology	227.5053533	3.88%
Operation	9.751605996	0.17%
Transfusion	10.74732334	0.18%
Food	157.2077088	2.68%
Pathology	29.37259101	0.50%
Injection	507.7002141	8.67%
Reservation	555.5010707	9.48%
Documents	481.4239829	8.22%
Nursery	677.0599572	11.56%
Process	432.1541756	7.38%
Records	1611.571734	27.51%
Rehabilitation	4.35117773	0.07%
In/Out	51.67237687	0.88%
Total	5858.950749	

each order. For example, the number records of doctors have its peak in 11am, which correponds to the peak of outpatient clinic, whose trend is very similar to reservation of outpatient clinic. The difference between these two orders is shown in 1pm to 5pm, which corresponds to the activities of wards.

Data in HIS can also capture the trend of the usage of terminals shown in Figure 6. Vertical axis denotes the averaged ratio of usage of terminals from February 3 to 8, 2008. Horizonal axis gives time zone. The plot gives the activity of outpatient clinic and ward. The usage of terminals of ward has two peaks before and after opening of outpatient clinic, which reflects our tuition on the activity in university hospital.

These results show that we can measure and visualize the dynamics of clinical activities in the university hospital by data analysis techniques. If we can detect some abnormalities different from usual behavior in these measurements, this may give some knowledge about risks in the clinical activities. Thus, it is highly expected that data mining methods, especially spatiotemporal data mining techiniques play crucial roles in analyzing data in hospital information system and understanding dynamics of hospital.

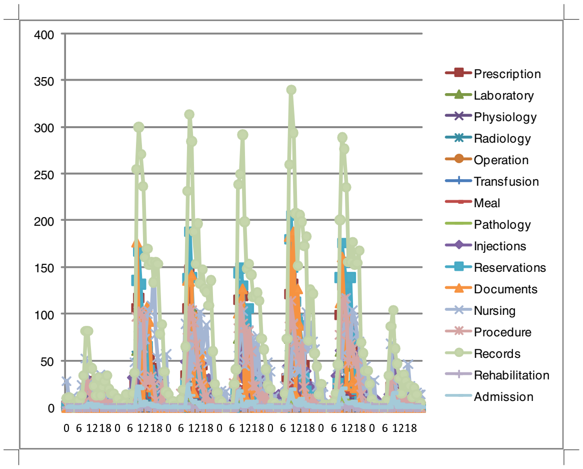

Figure 4. Trends of Number of Orders (June 1 to 6, 2008)

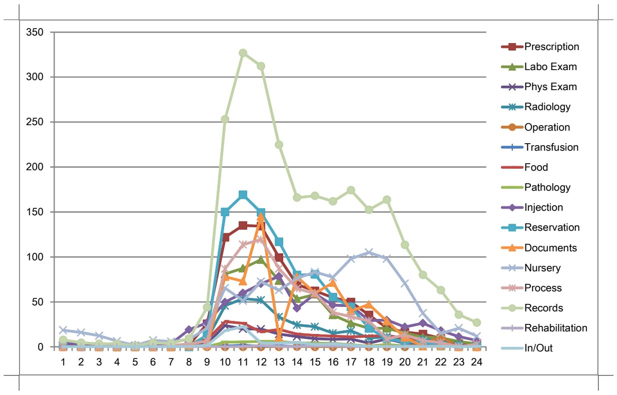

Figure 5. Trends of Number of Orders (June 2, 2008)

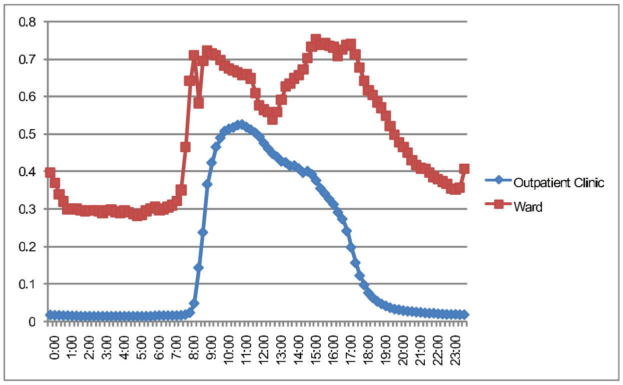

Figure 6. Averaged Usage of Terminals

IV. Analysis of Clinical Process for Long-term Follow Up Patients

One advantage of counting the total number of orders is that it enables us to capture the global nature of clinical process. For example, if a doctor has a patient with more than fiver years follow up, the total number of orders for each outpatient clinic gives an overview of its clinical process.

Table II shows the statistics of total number of orders for patients who were followed up for more than 5 years. The number of orders per visit is 2.042, and 1.700 of them consists of prescription, laboratory examinations and reservation, all of which are top 3 orders. This shows that $1.700/2.042 \sim 83\%$ of clinical orders can be explained by these three orders. Figure 7 shows the decision tree induced for the relations between long-term follow up patients and their visiting clinical departments. This tree also shows that these the number of three clinical orders characterizes those patients. For example, if the number of orders of laboratory examinations is more than 40, 76.2% of the patients are visiting the department of internal medicine.

V. Analysis of Trajectory of #Orders

If we take two variables of each orders shown in Figure 5, then we can depict the trend of two attributes as a trajectory, as shown in Figures 11 and 12.

Tsumoto and Hirano proposes a clustering method of trajectories, which calculates dissimarlity measures via multiscale matching and apply clustering methods to trajectories by using the dissimilarities between trajectories[5]. By applying this method to the data shown in Subsection III-C, the dendrogram shown in Figure 8 was obtained. Examples of clusters are shown in Figure 9 and Figure 10. The first one gives a pattern where orders are given both in wards

Table II
Averaged Number of Orders for Long-Term Follow-up Patients

Orders	Average	#Orders per visit
Total Orders	219.75	2.042
Prescription	90.80	0.845
Labo Exam	51.89	0.483
Reservation	39.81	0.371

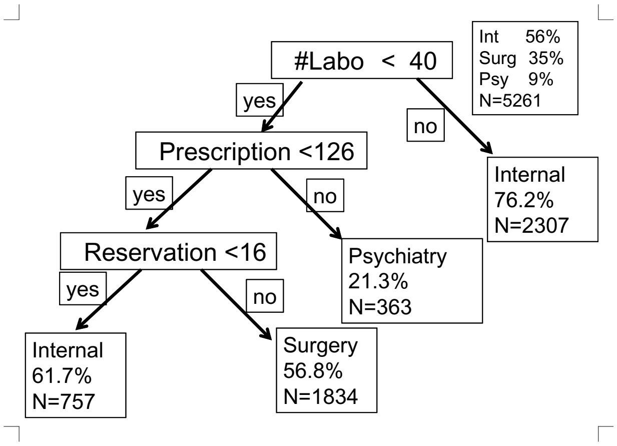

Figure 7. Decision Tree for Long-term Follow-up Patients

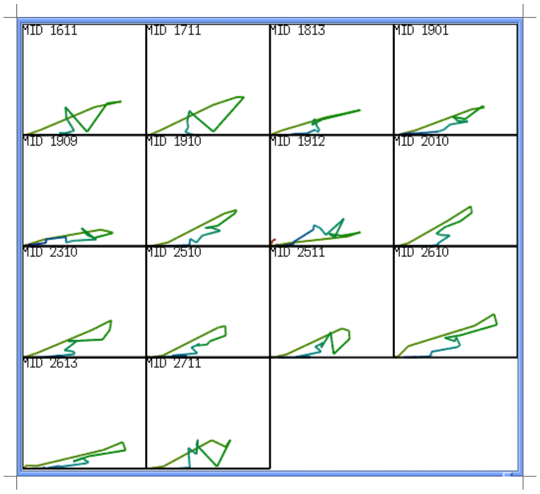

Figure 9. Cluster No.1 (June 1 to 6, 2008)

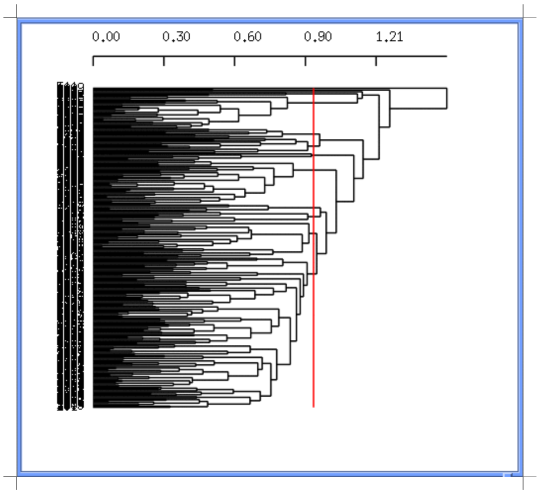

Figure 8. Dendrogram of Trajectories (June 1 to 6, 2008)

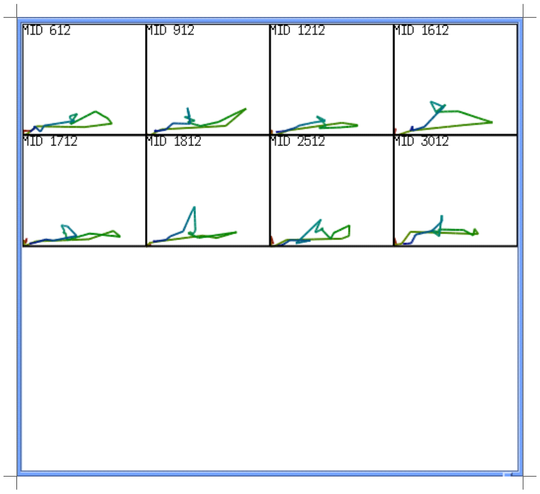

Figure 10. Cluster No.2 (June 1 to 6, 2008)

and outpatient clinics. The other one gives a pattern where orders are provided mainly in wards. A typical example in the first cluster is shown in Figure 11, while one in the second cluster is in Figure 12.

VI. Analysis of Hospital Management Data

A. Objective

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.

B. Method: Exploratory Data 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

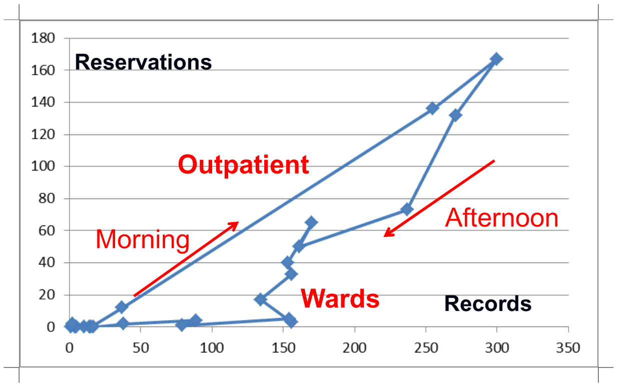

Figure 11. Trajectory between #Reservations and #Records (June 2, 2008)

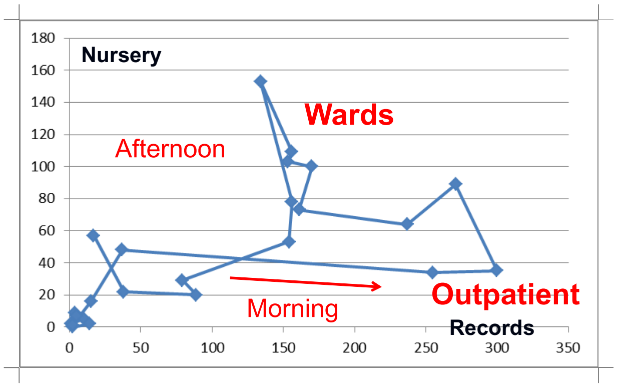

Figure 12. Trajectory between #Nursery Orders and #Records (June 2, 2008)

discharge summaries were analyzed by descriptive statistics, exploratory data analysis, statistical tests, regression analysis and generalized linear model. R was used for these analyses..

C. 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

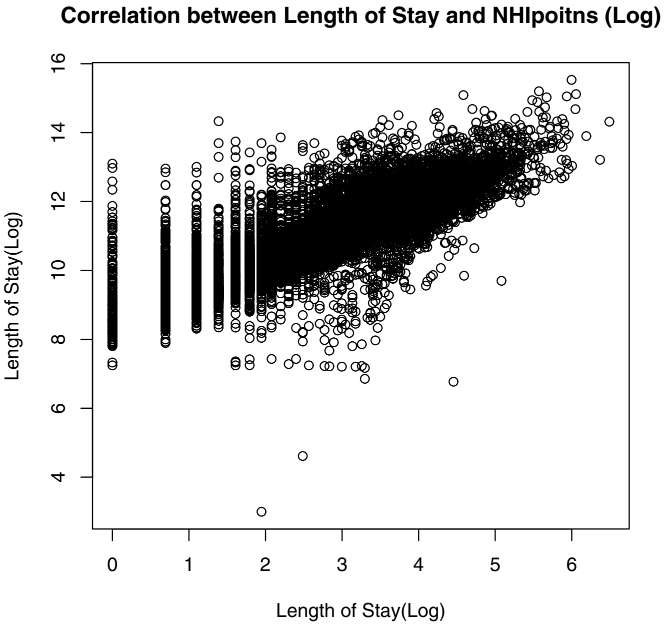

Figure 13. Scattergram of Length of Stay and NHI Points (Logarithm Transformation, Total Data)

denotes the above three divisions for short.

Distribution of Length of Stay.

Table III summarizes the descriptive statistics of length of stay with respect to the whole cases, major and minor divisions. The natures of these distributions are not significantly changed ifthese cases are stratified by the condition whether a surgical operation is applied to a case or not.

Correlation between Length of Stay and NHI Points

Figure 13 depicts the scattergram between the length of stay and NHI points of total cases, which suggest high correlation between two variables. For simplicity, the vertical and horizontal axes show the logarithm of raw values. Actually, The coefficient of correlation are calculated as 0.837 and 0.867, which mean that the correlation is very strong.

Table IV 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.

Table III
Descriptive Statistics of Length of Stay

	Average	Median	SD	Skewness	Kuritosis
Whole Cases					
Raw Data	26.46	16.00	33.67	4.34	34.15
Logarithmic Transformation	2.74	2.77	1.06	-0.06	-0.28
Neoplasm					
Raw Data	37.54	25.00	38.72	2.90	13.21
Logarithmic Transformation	3.19	3.22	0.98	-0.32	0.08
Malignant Neoplasm of Lung					
Raw Data	49.65	39.00	43.42	2.57	10.82
Logarithmic Transformation	3.57	3.66	0.88	-0.79	2.00
Malignant Neoplasm of Stomach					
Raw Data	36.44	36.00	19.18	0.46	0.37
Logarithmic Transformation	3.40	3.58	0.72	-1.42	2.55
Malignant Neoplasm of Liver					
Raw Data	35.93	33.00	21.40	1.19	2.70
Logarithmic Transformation	3.38	3.50	0.71	-1.18	3.03

Table IV
Correlation between Length of Stay and NHI Points (After Logarithm Transformation)

	Total	With Operation	Without Operation
Total Cases	0.837	0.829	0.779
Neoplasm	0.867	0.844	0.826
Lung Cancer	0.838	0.648	0.903
Stomach Cancer	0.827	0.738	0.801
Liver Cancer	0.711	0.577	0.755

1) 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 [6] 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. Figure 14 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.

VII. Conclusions

In this paper, we applied several exploratory data analysis techniques to data extracted from hospital information systems. 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. The results obtained from the dataset show that both the length of stay and NHI points follows log-normal distribution and these two variables are strongly correlated with each other. Thus, from the macro point of view, the characteristics of the length of stay decide those of NHI costs. Since these properties hold even in the major and minor divisions of malignant neoplasm, there exists a hidden factor which determines the features of the length of stay, i.e., the log-normal distribution of the length of stay. If this factor can be specified, it enables us to control the length of stay. Although the above results in generalized linear model do not give any significant models for the length of stay from the given database, the further analysis should be conducted by gathering more information into the hospital information system.

The results obtained above also show the possibility that combination of data from discharge summaries and data from accounting information system enables us to discuss

```
Call:
glm(formula
     = lognhi0 ~ (loglos0 + sex + age + outcome + adtime)^2,
    data = table)
% \end{vertabim}

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-7.17179  -0.34257  -0.05306   0.26980   4.37032

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.5134576  0.0517890 164.388  < 2e-16 ***
loglos0          0.8556933  0.0160016  53.476  < 2e-16 ***
sexM             0.1609546  0.0405073   3.973 7.12e-05 ***
age             -0.0038052  0.0008674  -4.387 1.16e-05 ***
outcome         -0.0083361  0.0181751  -0.459 0.646487
adtime          -0.0071641  0.0207570  -0.345 0.729992
loglos0:sexM    -0.0076588  0.0094779  -0.808 0.419061
loglos0:age      0.0006624  0.0001925   3.441 0.000581 ***
loglos0:outcome -0.0081192  0.0048621  -1.670 0.094960 .
loglos0:adtime  -0.0091114  0.0052452  -1.737 0.082392 .
sexM:age        -0.0003907  0.0004071  -0.960 0.337163
sexM:outcome    -0.0265756  0.0117403  -2.264 0.023611 *
sexM:adtime      0.0049953  0.0110712   0.451 0.651850
age:outcome      0.0011690  0.0002427   4.816 1.48e-06 ***
age:adtime       0.0011459  0.0002167   5.289 1.25e-07 ***
outcome:adtime   0.0136464  0.0056265   2.425 0.015304 *
---
Signif. codes:
        0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family
         taken to be 0.3421949)
Null deviance: 18087.9  on 15425  degrees of freedom
Residual deviance:  5273.2  on 15410  degrees of freedom
AIC: 27253
Number of Fisher Scoring iterations: 2
```

Figure 14. GLM Anaysis on NHI Points (Total Cases, Logarithmic Transformed Data)

the profitability of the university hospital in a quantitative way and provides us a basic tool for the management analysis of the university hospital, although more na?ve indices are used for the conventional management analysis [8,9].

References

[1] S. Tsumoto, “Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic,” *Information Sciences*, no. 124, pp. 125–137, 2000.

[2] ——, “G5: Data mining in medicine,” in *Handbook of Data Mining and Knowledge Discovery*, W. Kloesgen and J. Zytkow, Eds. Oxford: Oxford University Press, 2001, pp. 798–807.

[3] E. Hanada, S. Tsumoto, and S. Kobayashi, “A hubiquitous environmenth through wireless voice/data communication and a fully computerized hospital information system in a university hospital,” in *E-Health*, ser. IFIP Advances in Information and Communication Technology, H. Takeda, Ed. Springer Boston, 2010, vol. 335, pp. 160–168.

[4] S. Tsumoto and S. Hirano, “Risk mining in medicine: Application of data mining to medical risk management,” *Fundam. Inform.*, vol. 98, no. 1, pp. 107–121, 2010.

[5] ——, “Detection of risk factors using trajectory mining,” *Journal of Intelligent Information Systems*, pp. 1–23, 2009, 10.1007/s10844-009-0114-7. [Online]. Available: http://dx.doi.org/10.1007/s10844-009-0114-7

[6] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed. Boca Raton: CRC Press, 1990.