J. Soc. Korea Ind. Syst. Eng Vol. 42, No. 4 : 126-134, December 2019
https://doi.org/10.11627/jkise.2019.42.4.126
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Diagnosis Analysis of Patient Process Log Data

Joonsoo Bae†

Department of Industrial and Information Systems Eng. Jeonbuk National University

환자의 프로세스 로그 정보를 이용한 진단 분석

배 준 수†

전북대학교 산업정보시스템공학과

Nowadays, since there are so many big data available everywhere, those big data can be used to find useful information to improve design and operation by using various analysis methods such as data mining. Especially if we have event log data that has execution history data of an organization such as case_id, event_time, event (activity), performer, etc., then we can apply process mining to discover the main process model in the organization. Once we can find the main process from process mining, we can utilize it to improve current working environment. In this paper we developed a new method to find a final diagnosis of a patient, who needs several procedures (medical test and examination) to diagnose disease of the patient by using process mining approach. Some patients can be diagnosed by only one procedure, but there are certainly some patients who are very difficult to diagnose and need to take several procedures to find exact disease name. We used 2 million procedure log data and there are 397 thousands patients who took 2 and more procedures to find a final disease. These multi-procedure patients are not frequent case, but it is very critical to prevent wrong diagnosis. From those multi-procedure taken patients, 4 procedures were discovered to be a main process model in the hospital. Using this main process model, we can understand the sequence of procedures in the hospital and furthermore the relationship between diagnosis and corresponding procedures.

Keywords : Process Mining, Patient, Diagnosis Analysis, Healthcare

1. Introduction

With the growth in electronic health records (EHRs), more and more facilities are gathering large amounts of digitized patient data. Much of the responsibility for patient data input has been taken on by nurses who previously recorded patient information in paper-based formats. Although accurate documentation is essential for patient care, computerized patient data also enhances quality for the entire healthcare system. Healthcare providers can use data mining to uncover previously unknown patterns from vast data stores and then use this information to build predictive models. Since the 1990s, businesses have been using data mining for activities such as credit scoring, fraud detection, and maintenance scheduling. Now, healthcare organizations are also seeing value in data mining. Healthcare organizations can use data mining to make better patient-related decisions. For instance, they provides information to guide patient interactions by determining patient preferences, usage patterns, and current and future needs–all of which help to improve patient satisfaction.

All clinicians have shouldered a large portion of the responsibility for recording patient data in the EHR, but their efforts will contribute to a significant potential benefit to patients and to the health delivery system. As more data becomes

Received 24 October 2019; Finally Revised 7 December 2019; Accepted 13 December 2019
† Corresponding Author : jsbae@jbnu.ac.kr

available to data miners, clinicians will also benefit by having more opportunities to provide appropriate, well planned, and cost-effective patient care. But there is a more suitable mining method discovered. Process mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets.

Process mining is a process management technique that allows for the analysis of business processes based on event logs. The basic idea is to extract knowledge from event logs recorded by an information system. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs [1]. Process mining techniques are often used when no formal description of the process can be obtained by other approaches, or when the quality of an existing documentation is questionable. For example, the audit trails of a workflow management system, the transaction logs of an enterprise resource planning system, and the electronic patient records in a hospital can be used to discover models describing processes, organizations, and products. Moreover, such event logs can also be used to compare event logs with some prior model to see whether the observed reality conforms to some prescriptive or descriptive model. In present era, process mining is becoming popular in health care field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data.

2. Related Research

As electronic health records are accumulated and available to enhance healthcare service, data mining techniques are applied for decision making in hospital with the two objectives. One is for doctors and the other one is for hospital management. For doctors, data mining can help to determine the diagnosis of patients based on available big datasets. Aljumah et al. [2] concentrated upon predictive analysis of diabetic treatment using a regression-based data mining technique. Lee et al. [6] adapt and extend association rule mining and clustering algorithms to extract useful knowledge regarding diabetes and high blood pressure from the 1999–2008 survey results. Santos et al. [15] presented an automated data mining system that allows public health decision makers to access analytical information regarding brain tumors. The emphasis in this study is the use of ontology in an automated data mining process. The non-experts who tried the system obtained useful information about the treatment of brain tumors. Bilge et al. [3] aimed to explore rules and relationships that might be used to detect possible asymptomatic carotid stenosis by using data mining techniques. For this purpose, Genetic Algorithms (GAs), Logistic Regression (LR), and Chi-square tests have been applied to the patient dataset. Mookiah et al. [10] discussed the data mining system for the automated identification of normal and glaucoma classes using Higher Order Spectra (HOS) and Discrete Wavelet Transform (DWT) features. The extracted features are fed to the Support Vector Machine (SVM) classifier with linear, polynomial order 1, 2, 3 and Radial Basis Function (RBF) to select the best kernel function for automated decision making.

Data mining can be used for decision making of hospital management. Yang et al. [17] proposed a data-mining framework that utilizes the concept of clinical pathways to facilitate automatic and systematic construction of an adaptable and extensible detection model of fraudulent and abusive cases. The proposed approaches have been evaluated objectively by a real-world data set gathered from the National Health Insurance (NHI) program in Taiwan. The empirical experiments showed that our detection model is efficient and capable of identifying some fraudulent and abusive cases that are not detected by a manually constructed detection model. Lavraca et al. [5] proposed an innovative use of data mining and visualization techniques for decision support in planning and regional-level management of Slovenian public healthcare. The results are applicable to healthcare planning and support in decision making by local and regional healthcare authorities. In addition to the practical results, which are directly useful for decision making in planning of the regional healthcare system, the main methodological contribution of the paper is the developed visualization methods that can be used to facilitate knowledge management and decision making processes.

There are some research about time data in healthcare environment in connection with data mining. Yeh et al. [18] combined temporal abstraction with data mining techniques for analyzing dialysis patients' biochemical data to develop a decision support system. The mined temporal patterns are helpful for clinicians to predict hospitalization of hemodialysis patients and to suggest immediate treatments to avoid hospitalization. Lin et al. [7] reported a data mining technique have developed to discover the time dependency pattern

of clinical pathways for managing brain stroke. The mining of time dependency pattern is to discover patterns of process execution sequences and to identify the dependent relation between activities in a majority of cases. By obtaining the time dependency patterns, doctors can predicted the paths for new patients when he/she is admitted into a hospital; in turn, the health care procedure will be more efficient.

Process Mining can be also applied to healthcare data to support decision making of doctors and hospital management [9, 13]. Homayounfar [4] describes process mining in relation to hospital information systems and shows which direction the challenges can take if the two areas are combined by applying process mining in hospital information systems. Combining the field of process mining with the hospital information systems is the modern and recommendable approach, and provides a lot of meaningful results if the event logs are appropriately maintained and the structure of the data is a-priori known. Mans et al. [8] demonstrated the applicability of process mining using a real case of a gynecological oncology process in a Dutch hospital. They applied process mining techniques to obtain meaningful knowledge about these flows, e.g., to discover typical paths followed by particular groups of patients. This is a non-trivial task given the dynamic nature of healthcare processes. Rebuge et al. [12] introduced a methodology for the application of process mining techniques that leads to the identification of regular behavior, process variants, and exceptional medical cases. The approach is demonstrated in a case study conducted at a hospital emergency service. Tsumoto et al. [16] presented process mining results in which the temporal behavior of global hospital activities is visualized. The results show that the reuse of stored data will provide a powerful tool for hospital management and lead to improvement of hospital services. All of the above existing research are for hospital management using process mining. In this paper, process mining can be applied to help doctors by finding the principal and most complex medical procedures, then we will propose an algorithm how to determine the diagnosis from the discovered process.

3. Methodology

3.1 Problem Description

Most patients need to take medical procedures (medical test and examination) in hospital to diagnose disease. Furthermore some patients can be diagnosed by only one procedure, but there are many patients who are not classified into the general diagnosis procedures and they need to take several extra steps until the name of disease is identified. Although this case is not majority, but it is very important to discover such a patient who has very difficult disease to diagnose and recognize what is the main process being followed in the hospital. Concerning the processes, there are important questions that require an answer;

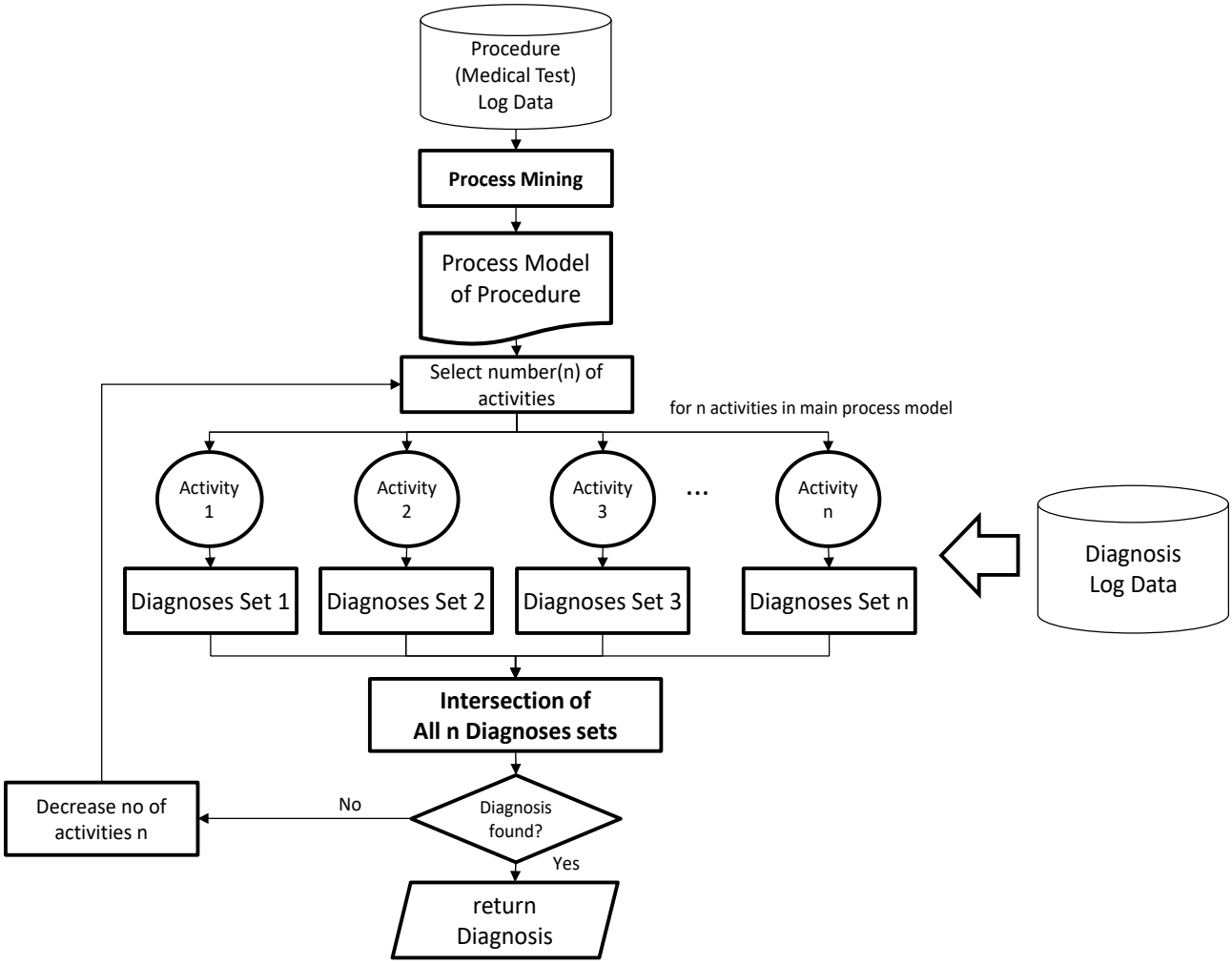
- What is main procedure process that multi-procedure patients take in the hospital?
- If the procedure process is discovered, what kind of diseases is diagnosed?

In this research we used a data set of US National trauma data bank [11] published in 2010. The data set includes 6,359 kinds of diagnosis, 3,021 kinds of procedures. Especially there are 55% patients took multiple (2 and more) procedures to find a final disease.

<Table 1> shows the number of patients who took 2 and more procedures. The total number of patients in this dataset is 716,460, and they took 2,645,132 procedures to diagnose, which proves that more than half of total patients took 2 and more procedures. The number of 91,280 patients took 2 kinds of procedures and the number of procedure code becomes 182,560. The same way, 3 kinds of procedures, 4 kinds of procedures, 5 kinds of procedures, and 6 and above are in the table. We will utilize the patient's data which have 2 and more procedures, because we try to discover process model of procedures in the hospital. In other words, 2,325,716 (88%) procedures of 397,044 (55%) patients are used for process mining.

<Table 1> Number of Patients who Took 2 and More Procedures

	No. of dataset	No. of Pcode (Procedure code)						Ratio of multiple Pcode
		1	2	3	4	5	≥ 6	
Procedure(Medical Test)	2,645,132	319,416	182,560	196,446	213,364	214,390	1,518,956	2,325,716 (88%)
Patient (Incident)	716,460	319,416	91,280	65,482	53,341	42,878	144,063	397,044 (55%)



<Figure 1> Algorithm Flowchart

<Figure 1> shows that flowchart of the algorithm. The first step is process mining of procedure (medical test and examination) log data using a commercial software package, Disco [14] for process mining, which is one of the powerful tools in process mining. The result is shown in the process diagram map, statistic and cases. The second step is selecting a number of activities for diagnosis determination. It is a very important step for diagnostic analysis, because the final result will depend on the number of activities. If we select more activities, the final diagnosis set becomes smaller. On the contrary if we select less activities, the final diagnosis set becomes larger. The activities can be selected based on the max frequency. The third step is building diagnosis set. For each selected activity, there are related diagnosis set which can be extracted from the diagnosis log data. That means that if one patient take a procedure, he/she may have certain set of diagnosis in the log data. We can understand easily that what kind of diseases diagnosed by one activity (procedure) in the discovered main process model. The fourth step is an intersection of diagnosis sets. If the intersections are found satisfactorily, we can conclude our algorithm to find diagnosis in the main process is success. If not, we have to decrease number of activities and reselect the number of activities. And doing the step 3 and 4 repeatedly until the diagnosis are found successfully.

Using this algorithm we can answer the above 2 questions. The main process of procedures in trauma center can be discovered using process mining. This new knowledge can be utilized not only for management and operation improvement of facility, but also useful for doctors to decide diagnosis of patients who has very complex procedures and difficult disease. This research will prevent wrong diagnosis of difficult patients by using process mining technique

3.2 US NTDB Data

US National Trauma Data Bank (NTDB) is trauma related data voluntarily reported by participating trauma centers. The Research Dataset (RDS) [11] is a set of relational tables and consists of 18 data files as in <Table 2>. Among them, only two files RDS_DCODE and RDS_PCODE are used in this paper. RDS_DCODE has diagnosis information for each patient using ICD-9-CM code. ICD-9-CM means to be ninth revision of the International Classification of Diseases, Clinical Modification. ICD-9-CM is the official system used in the United States to classify and assign codes to health conditions and related information. The International Classification of Diseases (ICD) is the standard diagnostic tool for epidemiology, health management and clinical purposes. This includes the analysis of the general health situation of population groups. It is used to monitor the incidence and prevalence of diseases and other health problems, proving a picture of the general health situation of countries and populations. ICD is used by physicians, nurses, other providers, researchers, health information managers and coders, health information technology workers, policy-makers, insurers and patient organizations to classify diseases and other health problems recorded on many types of health and vital records, including death certificates and health records. In addition to enabling the storage and retrieval of diagnostic information for clinical, epidemiological and quality purposes, these records also provide the basis for the compilation of national mortality and morbidity statistics by WHO Member States.

The example of RDS_DCODE are as follows; 001~139 (Infectious And Parasitic Diseases), 280~289 (Diseases Of The Blood And Blood-Forming Organs), 320~389 (Diseases Of The Nervous System And Sense Organs), 390~459 (Diseases Of The Circulatory System), 520~579 (Diseases Of The Digestive System), 710~739 (Diseases Of The Musculoskeletal System And Connective Tissue), 800~999 (Injury And Poisoning). Since our data set is about trauma, all of the diagnosis codes used in this paper are in 800~999.

RDS_PCODE file includes procedure codes using ICD-9-CM code. A medical procedure is a course of action intended to achieve a result in the care of persons with health problems. A medical procedure with the intention of determining, measuring or diagnosing a patient condition or parameter is also called a medical test. For trauma patient, the example procedure codes are 76~84 (Operations On The Musculoskeletal System), 85~86 (Operations On The Integumentary System), 87~99 (Miscellaneous Diagnostic And Therapeutic Procedures), etc.

<Table 2> NTDB RDS 18 Data Files and Descriptions

File Name	Description
RDS_AISPCODE	The AIS (Abbreviated Injury Scale) code submitted by the hospital
…	
RDS_COMPLIC	Any NTDS complications
RDS_DEMO	Demographic information
RDS_DCODE	**ICD-9-CM Code of Diagnosis Information**
RDS_DISCHARGE	Includes discharge and outcome information
RDS_ECODE	Includes the ICD-9 external cause of injury code.
RDS_ED	Emergency Department information
RDS_FACILITY	Facility Information
RDS_PCODE	**Procedure codes**
RDS_PROTDEV	Protective devices
RDS_VITALS	Vital signs from EMS and ED

<Table 3> shows the RDS_PCODE file descriptions of field. The file included 5 fields with incident key, procedure code, year of procedure, days to procedure and hours to procedure with their definition, data type, length and valid values. From this file, two fields are used in process mining; incident key for id case, procedure code for an event. If there are multiple records with the same incident key, it means that one patient took multiple procedures.

<Table 3> RDS_PCODE File Description

Field name	Definition	Data type	Length
Incident Key (INC_KEY)	Unique identifier for each record	Numeric	10
ICD-9-CM Procedure code (PCODE)	ICD-9-CM Procedure code	String	5
Year of Procedure (YOPROC)	Year in which the procedure occurred	String	100
Days to Procedure (DAYTOPROC)	No. of days until the beginning of procedure	String	10
Hours to Procedure (HOURTOPRO)	No. of hours until the beginning of procedure	String	10

<Table 4> shows the RDS_DCODE file that includes 2 different fields, incident key and diagnosis code with definition, data type, length and valid values. In this paper, incident key is used for each patient, and this can be multiple in this file, which means this patient has two or more diagnoses. This incident key can link RDS_PCODE and

RDS_DCODE files. One patient with the same incident key can take procedures in RDS_PCODE and can have diagnoses in RDS_DCODE. The cardinality between RDS_PCODE and RDS_DCODE relationship is m:n. We are going to find most relevant diagnosis in RDS_DCODE of the main process model discovered from RDS_PCODE in following sections.

<Table 4> RDS_DCODE File Description

Field name	Definition	Data type	Length
Incident Key (INC_KEY)	Unique identifier for each record	Numeric	10
ICD-9-CM Diagnosis (DCODE)	ICD-9-CM Diagnosis Code	String	6

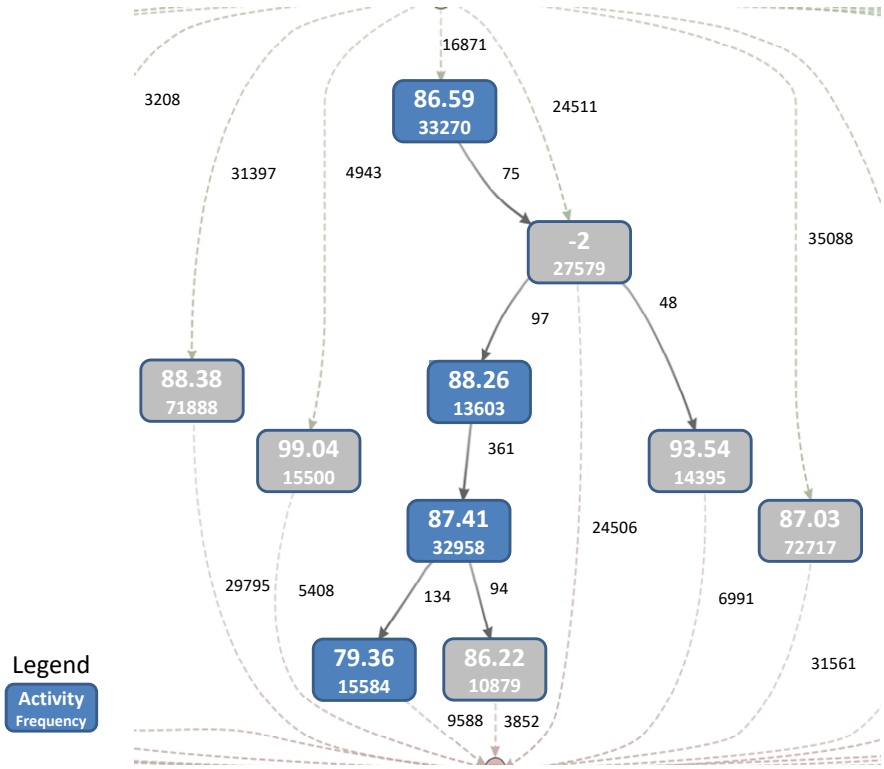
4. Discovering Process Models

<Figure 2> shows discovered process model of procedures in RDS_PCODE with the same incident key. That was result of commercial software package, Disco [14]. There are many different flows from start to end. But we can find that there is a main process that has several activities in the flow. It is called main process, and there are 5 activities (procedures) on the main process. That is, 86.59 → -2 → 88.26 → 87.41 → 79.36. This means 5 activities are most related pattern based on the precedence rule determined by process mining algorithm. If we consider only procedure data, these 5 activities may happen together with this sequence. If one patient arrives at hospital, he/she may take these 5 procedures. But second procedure code, -2 (unknown) is meaningless, so only 4 procedures except -2 are considered from now on.

<Table 5> Procedures in the Main Process of Discovered Process Model

Procedure Code	Description
86.59	Closure of skin and subcutaneous tissue of other sites
88.26	Other skeletal x-ray of pelvis and hip
87.41	Computerized axial tomography(CAT) of thorax
79.36	Open reduction of fracture with internal fixation, tibia and fibula

<Table 5> shows the description of procedures in the main process model. They are closure of skin, skeletal x-ray of pelvis and hip, CAT, reduction of fracture, etc. Even though we cannot understand medical terminology completely, we can see that 4 procedures are related and becomes main process in the hospital.



<Figure 2> Discovered Process Model of Procedures

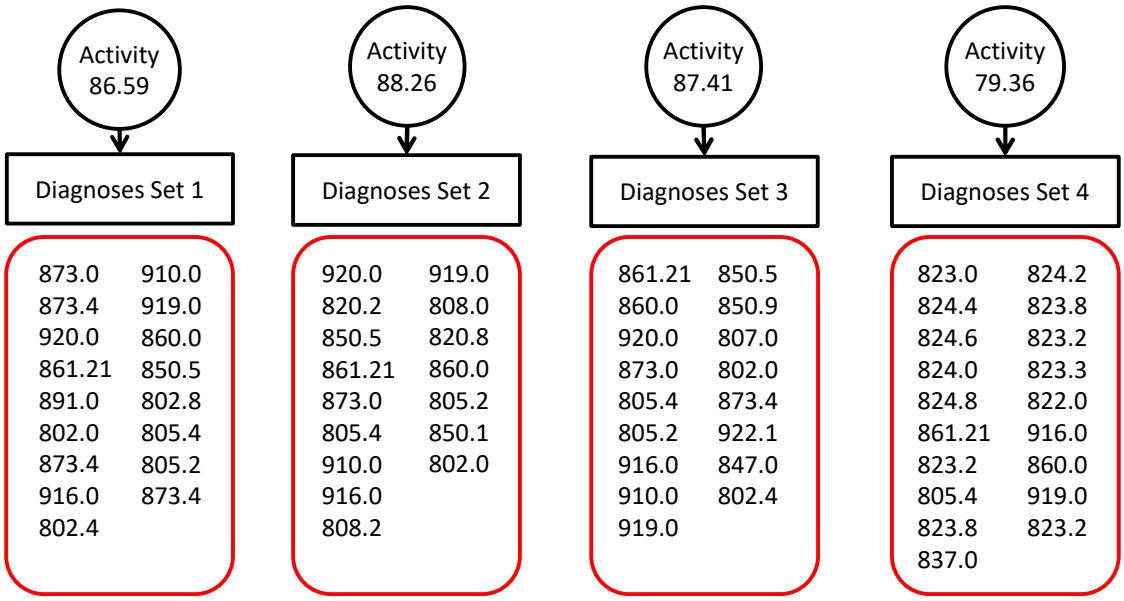
5. Diagnosis Analysis from Process Models

Now we have discovered main process which has 4 activities (procedures), 86.59, 88.26, 87.41, and 79.36. Then next step is finding diagnosis from the discovered main process. Using the frequency criteria, 5.71% of all patients who took the procedure code 86.59 is diagnosed with diagnosis code 873.0. Likewise, 17 kinds of diagnosis codes (873.0, 873.42, 920, ⋯, 873.43) are selected according to relative frequency that is one or more percent. These 17 diagnosis codes becomes diagnosis set 1 of procedure 86.59 in <Figure 3>. In <Figure 3>, the other 3 diagnosis sets are generated using the same method. Procedure 88.26 generates diagnosis set 2 having 16 diagnoses (920.0, 820.2, ⋯, 802.0). Procedure 87.41 generates diagnosis set 3 having 17 diagnoses (861.21, 860.0, ⋯, 802.4). Procedure 79.36 generates diagnosis set 4 having 19 diagnoses (823.0, 824.4, ⋯, 823.2).

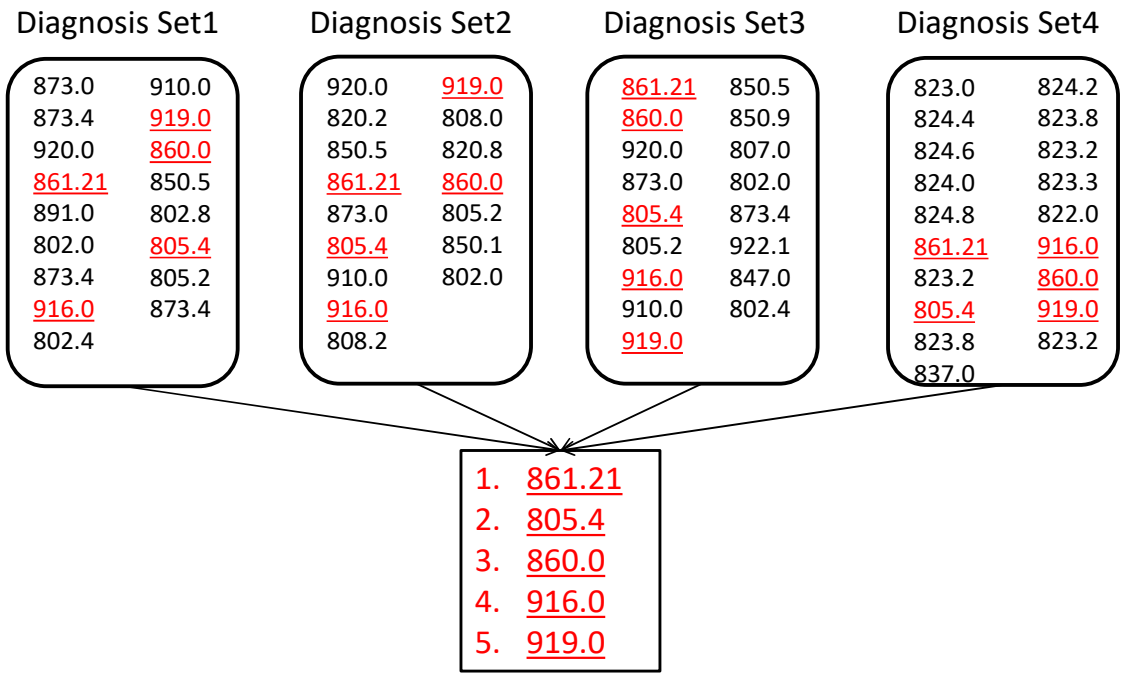
Once diagnosis sets are generated for each activities in the main process, we need to extract common diagnosis set of whole main process. We have to determine final diagnosis of main process. There are so many methods to integrate several diagnosis sets, but the simplest method is intersection. If the intersection set of common diagnosis is satisfactory, we can return the final diagnosis of main process. But otherwise, we must go back to the previous step then reselect the number of activities. Most possible method for successful intersection is to decrease the number of activities. If the number of activities is decreased from 4 to 3, common diagnosis set will be bigger and it becomes easier to find intersection set.

<Figure 4> shows the intersection of diagnosis sets in <Figure 3>. We found 5 different diagnosis codes (861.21 Lung contusion closed; 805.4 Fracture lumbar vertebra, closed; 860.0 Traumatic pneumothorax and hem thorax; 916.0 Abrasion in hip, thigh, leg or ankle; 919.0 Superficial injury of other multiple and unspecified sites) in the order of frequency that are included in all of 4 diagnosis sets. From this result, we can conclude that 861.21 is the most possible diagnosis from this analysis. This means that patients diagnosed by those 5 diagnosis codes took the same procedures from our main process model. If there is new patient who took the same procedures in the main process model, we can conclude this patient has high possibility of the same diagnosis code. This case may be not so common. But it is very important to give a suggestion to doctors in case the patient took complex multiple procedures, which will prevent wrong diagnosis and reduce wrong treatment.

In order to validate the contribution of our approach, process mining result is compared with data non-procedural data analysis in <Table 6>. When we do not consider process mining, the most possible diagnosis is 920 (Contusion of face scalp and neck except eye(s)). But when we apply more than 4 activity patient, the most possible diagnosis becomes 861.21, which are the same result from process mining.



<Figure 3> Diagnosis Set of Each Procedure in Main Process



<Figure 4> Intersection of Diagnosis Sets

<Table 6> Diagnosis Analysis with Process Mining

All Patient		More than 4 Activity Patient	
DCODE	Count	DCODE	COUNT
920	57,755	**861.21**	23,255
873.0	51,019	873.0	22,148
861.21	41,136	920	20,974
910	38,102	805.4	18,727
805.4	37,109	860.0	17,952

6. Conclusions

This paper proposed a new method to find a final diagnosis of a patient, who needs several procedures (medical test and examination) to diagnose disease of a patient by using process mining approach. We used 2 million procedures log data and there are 397 thousands patients who took 2 and more procedures to find a final disease. From those multi-procedure taken patient data, 4 procedures were discovered to be a main process model in the hospital by using process mining method. Using this main process model, we can understand the sequence of procedures in the hospital and furthermore we can find the most probable diagnosis of main process. This will support doctors to decide diagnosis of patients who took very complex multiple procedures. For a future work, the practical validation of result will be possible by collaboration with medical experts and furthermore patient behavior analysis and hospital structure design using this kind of dataset are possible.

Acknowledgement

This paper was supported by Industry Cooperation R&D project 2019 of Spatial Information Research Institute, LX Corporation.

References

[1] Aalst, W. Van Der, Process Mining : Discovery, Conformance and Enhancement of Business Processes, Springer Verlag, Berlin, 2011.

[2] Aljumah, A.A., Ahamad, M.G., and Siddiqui, M.K., Application of data mining : Diabetes health care in young and old patients, *Journal of King Saud University -Computer and Information Sciences*, 2013, Vol. 25, No. 2, pp. 127-136.

[3] Bilge, U., Bozkurt, S., and Durmaz, S., Application of data mining techniques for detecting asymptomatic carotid artery stenosis, *Computers and Electrical Engineering*, 2013, Vol. 39, No. 5, pp. 1499-1505.

[4] Homayounfar, P., Process mining challenges in hospital information systems, *in Proceedings of the Federated*

Conference on Computer Science and Information Systems (FedCSIS), 2012, pp. 1135-1140.

[5] Lavraca, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M., and Kobler, A., Data mining and visualization for decision support and modeling of public health-care resources, *Journal of Biomedical Informatics*, 2007, Vol. 40, No. 4, pp. 438-447.

[6] Lee, J.W. and Giraud-Carrier, C., Results on mining NHANES data : A case study in evidence-based medicine, *Computers in Biology and Medicine*, 2013, Vol. 43, No. 5, pp. 493-503.

[7] Lin, F.-R., Chou, S.-C., Pan, S.-M., and Chen, Y.-M., Mining time dependency patterns in clinical pathways, *International Journal of Medical Informatics*, 2001, Vol. 62, No. 1, pp. 11-25.

[8] Mans, R.S., Schonenberg, M.H., Song, M., and Aalst, W.M.P. van der, Process mining in healthcare: a case study, *in Proceedings of the First International Conference on Health Informatics*, 2008, Madeira.

[9] Mans, R.S., Schonenberg, M.H., Song, M., Aalst, W.M. P. van der, and Bakker, P.J.M., Application of Process Mining in Healthcare-A Case Study in a Dutch Hospital, *International Joint Conference on Biomedical Engineering Systems and Technologies*, 2008, pp 425-438.

[10] Mookiah, M.R.K., Acharya, U.R., Lim, C.M., Petznick, A., and Suri, J.S., Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features, *Knowledge-Based Systems*, 2012, Vol. 33, pp. 73-82.

[11] National Trauma Data Bank, NTDB Research Data Set Admission Year 2010 User Manual, October 2011.

[12] Rebuge, Á. and Ferreira, D.R., Business process analysis in healthcare environments : A methodology based on process mining, *Information Systems*, 2012, Vol. 37, No. 2, pp. 99-116.

[13] Rojas, E., Munoz-Gama, J., Sepúlveda, M., and Capurro, D., Process mining in healthcare : A literature review, *Journal of Biomedical Informatics*, 2016, Vol. 61, pp. 224-236.

[14] Rozinat, A., Disco User's Guide, fluxicon, 2019.

[15] Santos, R.S., Malheiros, S.M.F., Cavalheiro, S., and Oliveira, J.M.P. de, A data mining system for providing analytical information on brain tumors to public health decision makers, *Computer Methods and Programs in Biomedicine*, 2013, Vol. 109, No. 3, pp. 269-282.

[16] Tsumoto, S., Iwata, H., Hirano, S., and Tsumoto, Y., Similarity-based behavior and process mining of medical practices, *Future Generation Computer Systems*, 2014, Vol. 33, pp. 21-31.

[17] Yang, W.-S. and Hwang, S.-Y., A process-mining framework for the detection of healthcare fraud and abuse, *Expert Systems with Applications*, 2006, Vol. 31, No. 1, pp. 56-68.

[18] Yeh, J.-Y., Wu, T.-H., and Tsao, C.-W., Using data mining techniques to predict hospitalization of hemodialysis patients, *Decision Support Systems*, 2011, Vol. 50, No. 2, pp. 439-448.

ORCID

Joonsoo Bae | http://orcid.org/0000-0001-8872-5169