

Mining Time Dependency Patterns in Clinical Pathways

Fu-ren Lin, Shien-chao Chou
 Department of Information Management
 National Sun Yat-sen University
 Kaohsiung, Taiwan 804, R.O.C.
 frlin@cc.nsysu.edu.tw

Shung-mei Pan, Yao-mei Chen
 Department of Nursing
 Chung-Ho Memorial Hospital of
 Kaohsiung Medical University
 Kaohsiung, Taiwan, R.O.C.

Abstract

Clinical pathways are widely adopted by many large hospitals around the world in order to provide high-quality patient treatment and reduce the length of hospital stay of each patient. The development of clinical pathways is a lengthy process, and may require the collaboration among physicians, nurses, and staffs in a hospital. However, the individual differences cause great variances in the execution of clinical pathways. It calls for a more dynamic and adaptive process to improve the performance of clinical pathways. This paper proposes a data mining technique to discover the time dependency pattern of clinical pathways for curing 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, we can predict the paths for new patients when he/she is admitted into a hospital, and, in turn, the health care procedure will be more effective and efficient.

1. Introduction

Clinical pathways are structured, multidisciplinary care plans, in which diagnostic and therapeutic interventions performed by physicians, nurses and other staffs for a particular diagnosis or procedure, are sequenced on a timeline [7][9]. As the competition among healthcare institutes is getting strong, the competition advantage is not only on outstanding medical professional quality, but also on the agile clinical care processes. In order to provide high-quality patient treatment and reduce the length of hospital stay of each patient, clinical pathways are widely adopted by many large hospitals around the world.

The purposes of clinical pathways are to organize client care activities and interventions, reduce practice variations, minimize delays in treatments, and decrease resource use with a

goal of decreasing costs while maintaining or improving quality. The tasks of developing clinical pathways include seeking input from the client interdisciplinary team members, defining particular diagnoses or procedures, designing clinical management tools, and fostering flexibility dictated by client condition and medical provider preferences [10]. The application of clinical pathways becomes an efficient approach to analyze and control clinical care processes.

In this paper, we develop a data mining technique, *mining time dependency patterns*, to discover the patterns of clinical pathways. We focus on mining the medical treatment process from the patient records and clinical log data. We recognized that time is an important factor affecting the structure of clinical paths and the sequence among activities.

Agrawal and Srikant propose a data mining technique to discover sequential patterns of customer buying behavior in retailer stores [3]. Their study was mainly derived from their previous research in mining association rules [2].

The problem of mining sequential patterns is to find the maximal sequences among all sequences that have a certain user-specified minimum support. Each such maximal sequence represents a *sequential pattern* [4]. However, their algorithm cannot be used for mining clinical pathways because (1) the duration of each activity is not included in the dependency graph, (2) high frequent activities, which are lack of dependent relationship with others, should be kept in clinical pathways, and (3) activities may overlap in their time durations.

The application of data mining techniques to designing clinical pathways is based on the following rationales: the design of clinical pathways is knowledge intensive, and the knowledge can be learned from the data collected from the clinical processes. Moreover, clinical pathways may be embedded with unknown workflow patterns and exception handling. Therefore, it is suitable for investigating the use of data mining techniques

for improving the performance of clinical pathways.

The next section proposes the algorithm for mining time dependency patterns of workflows. Section 3 illustrates the proposed method for discovering the clinical pathways for brain stroke treatment. We conclude the paper and lay out future research in Section 4.

2. Mining time dependency patterns of process execution paths

In this section, we propose the algorithm for mining time dependency patterns of process execution paths, such as clinical pathways. *The mining of time dependency pattern is to discover patterns of activity execution sequences and identify the dependent relation between activities from process log data.* The sample of dependency graph is shown in Figure 1.

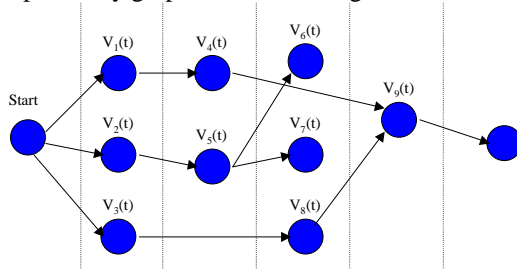


Figure 1. Sample of dependency graph

2.2. Definitions

Definition 1. Activity set (Denoted as \mathcal{A})

A process consists of a set of activities. All possible activities is denoted as $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$. An activity is an execution element that leads the transition of state in a business process. In this research, we only focus on the time sequence in process rather than other complex factors that trigger events or the state transition. We define two time parameters of an activity, one is the starting time and the other is the ending time. For example, an activity $A_i(2,8)$ represents activity A_i starting at time unit 2 and terminating at time unit 8.

Definition 2. Process log data (Denoted as L)

In a process log data L , an activity is a 4-tuple, $\langle P_i, A_j, T_{s,ij}, T_{e,ij} \rangle$, where P_i is the process i , A_j is an activities j in the process i , $T_{s,ij}$ is the starting time of activity A_j , and $T_{e,ij}$ is the ending time of activity A_j .

Definition 3. Time dependency

In a process, if activity A_j starts instantaneously after A_i terminates, we assert that activity A_j depends on activity A_i . “Start instantaneously” means that after A_i terminates, there is no other activities starting before A_j .

Definition 4. Process Graph (Denoted as G_p)

A process consists of a set of activities. In practice, a process can be a patient treatment flow or a customer buying flow. We consider a process as a graph with the following features:

- (1) A process can be modeled by a directed acyclic graph.
- (2) A vertex represents the activities performed by the process.
- (3) An edge represents the time dependency between two activities. There is no weight information in the edge.
- (4) Each process graph includes two pseudo vertices, starting and terminating vertices. The starting and ending times of a start vertex is set to 0, and the start vertex links to all of the first activities of the process. The activities that are not dependent on any other activities link to the terminate vertex.

We denote a process graph as $G_p(V_p, E_p)$, where V_p is a vertex set, and E_p is a Boolean edge set, which is represented as a Boolean value to indicate whether the time dependency exists among the two activities. Note that, each activity performs only once in a process by default.

Definition 5. Large graph (Denoted as LG)

Large graph is a time dependency graph that represents a pattern of execution sequence in a majority of cases, which is derived from the previous mining iteration. We name the large graph by its number of activities. For example, large *onegraph* is a graph with one vertex without edge and large *twograph* is a graph with two vertices. We denote the notation LG to indicate the set of large graphs with the same number of vertices. For example, the LG_3 is a set of large *threegraphs*, and the each graph consists of three activities. Any potential large graph generated by LG_2 would be denoted as a candidate graph CG_3 . While mining the large graph, the user gives a support threshold, which is defined as the fraction of total processes that contain this graph. If the support count of a graph equals to or greater than the minimum support, we call it a large graph. A large graph is composed of a list of activities represented by its subgraphs. For example, large graph LG is induced by three subgraphs G_{Pa}, G_{Pb}, G_{Pc} , so that the $LG.G_p_List$ is $\{G_{Pa}, G_{Pb}, G_{Pc}\}$.

Definition 6. Maximal graph

A maximal graph is a large graph that does not contain any other large graphs.

Definition 7. Complete graph

For a maximal graph, if both starting and terminating nodes are included in the graph, we call it a complete graph.

Definition 8. Window number (Denoted as W)

For a process, user can specify window number W to view the process in several periods. For an activity across different windows, we can partition the time duration of activities into different interval according to the window number. The interval is calculated as: *duration of process / window number + 1*. We trim the decimal parts of the interval.

2.2. Mining time dependency graph

The mining of time dependency graph consists of the following three phases: (1) process graph transformation phase, (2) large graph mining phase, and (3) graph maximization phase.

2.2.1 Process graph transformation. This phase transforms the original process log data to process graphs. The transformation follows the following steps:

- (1) Sort process log data L by process identification and starting time.
- (2) Identify activities in L . Each activity is denoted as $\langle P_i, A_j, T_{s,ij}, T_{e,ij} \rangle$.
- (3) Scan each process in L and establish graph Gp_i for each process P_i . For each activity A_j in the process P_i , find the time dependency relationship between A_j and other activities and add the process identification to the activity A_j . The edge set Ep_i of Gp_i will be assigned according to the dependency relationship.
- (4) Activities in A belong to LG_l if the count of process is equal to or greater than the support threshold. After the transformation phases, we will obtain a set of process graph Gp_i and LG_l , which will input to the next phase.

2.2.2. Large graph mining phase. This phase is the core of the mining process. Multiple passes are executed to generate large graphs. In each pass, the previous large graphs are used for generating new potentially large graphs, called *candidate graphs* (CGs). Each graph maintains a process list to link the process graph it belongs to. Large graphs can be identified by scanning their process list and counting their intersections. It saves much time because the number of process is much smaller than whole log data. For example, the number of large fourgraphs in Figure 2(a) (i.e., $LG_4(a).Gp_List$) is $\{P1, P2, P3\}$, and that in Figure 2(b) (i.e., $G_4(b).Gp_List$) is $\{P2, P3, P4\}$. The count of LG_5 (shown in Figure 3) is 2 because the intersection of processes ($LG_4(a).Gp_List \cap LG_4(b).Gp_List$) is $\{P2, P3\}$. However, while finding LG_2 , we still need to scan the G_p . Note that, CG_2 is composed of two

large *onegraphs*. We cannot get the count of LG_2 by simply finding the number of process intersection.

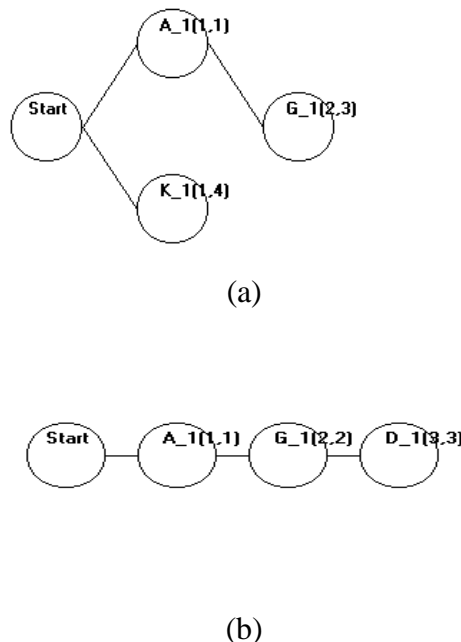


Figure 2. An Example of LG_4

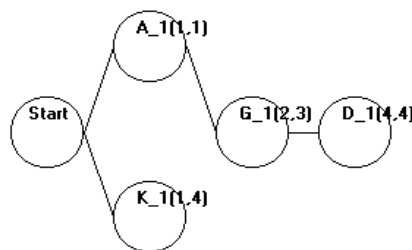


Figure 3. An Example of LG_5

We summarize the mining process of time dependency graph as the following steps:

- Step 1. For generating large *twograph*, we check any two activities A_i and A_j in LG_l and scan the process G_p that has both A_i and A_j . If the number of the dependency of $A_i \rightarrow A_j$ or $A_j \rightarrow A_i$ is larger than the support threshold, we add the activities A_i and A_j into LG_2 .
- Step 2. For generating large graphs with the number of vertices larger than two (e.g., the number of vertices is n , and n initially sets to 3), we check any two large graphs generated in the previous pass. If they have the same $n-2$ subgraph and the number of common process list is beyond the support threshold, the new large n -graph is generated.
- Step 3. Repeat Step 2 and increment n by 1 until no new large graphs are found. When

all large graphs are obtained, the maximization phase is aimed at deleting the graphs, which are subgraphs of another larger graphs.

3. Mining Time Dependency in Clinical Pathways for Curing Brain Stroke

Stroke is a major health problem in most industrialized countries. In addition to causing a huge number of deaths, stroke is the most important cause of physical disability in people over 60 years old. According to the World Health Organization, the stroke is defined as *rapidly developing clinical signs of focal disturbance of cerebral function, lasting more than 24 hours or leading to death, with no apparent cause other than vascular origin* [1].

The cost of brain stroke is a considerable burden on healthcare and social services. In the industrialized countries, it is the major killer and about one-half of survivors are left with a permanent handicap. In several studies in the USA, the annual costs of stroke have been estimated as varying between US \$6.5 and 11.2 billions [5][12]. An absence of effective therapies may be one of the reasons for this situation, and the clinical pathway is a new solution for this cost issue.

In the study of [11], the length of stay decreased from 7.52 to 6.33 days after the initiation of a critical pathway that begins its interventions in the Emergency Department. Quality of care was also improved in delivery time of carotid artery ultrasound examinations, as well as in timeliness of obtaining head computed tomography scans and reports.

Figure 4 depicts the steps we adopted in mining time dependency patterns in clinical pathways for curing brain stroke. It includes the following steps:

- (1) Data collection. We collect two types of data from patients' records. One is the clinical pathways log data and the other is patients' data.
- (2) Pre-processing. We select relevant data according to the domain experiences and knowledge. For example, some nursing care activities are not meaningful because they are usually executed everyday.
- (3) Mining time dependency patterns. The algorithm mentioned in Section 2 is used for mining time dependency patterns.
- (4) Prediction. An association technique is used to evaluate the relationship between the diagnosis data and treatment paths in order to predict the paths for new patients.

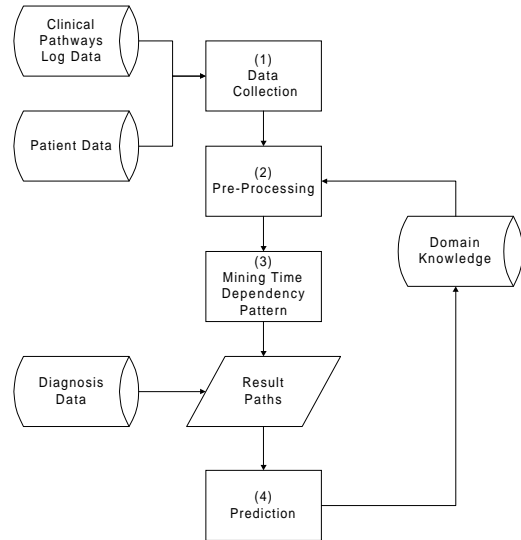


Figure 4. Steps of mining time dependency patterns for clinical pathways

3.1. Data Collection and Preprocessing

Due to the potential deviation among the different type of disease, we choose the brain stroke as our example data source. We gathered medical records from 113 patients during 1997 and 1998. We then induced around 400 activities from these medical records. The activities mainly include examination, prescription, treatment and nursing cares.

After collecting data, we filtered noisy data according to the domain knowledge. For example, the item Aq-dest was found in medical records most of the time. However, it is not meaningful to medical staffs. We also identified several routine activities occurring more than 50%, such as blood pressure checking. We view these as the routine activities, which are independent from the occurrence of other activities, but they may affect the mining performance extremely.

3.2. Mining Time Dependency

We implemented the algorithm of mining time dependency in clinical pathways. We also built a visualization interface to represent the mining results. An example clinical pathway is shown in Figure 5, and it is generated under the setting of 3% support threshold and 2 windows. The information in the inner brackets is the standard deviation. For example, the DM Diet_1(1(0.5), 6(0.87)) is the activity that starts at time unit 1 and terminates at time unit 6. The standard deviation of starting and ending times are 0.50 and 0.87 respectively.

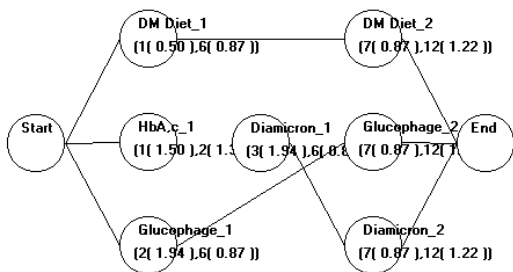


Figure 5. An example of clinical pathways

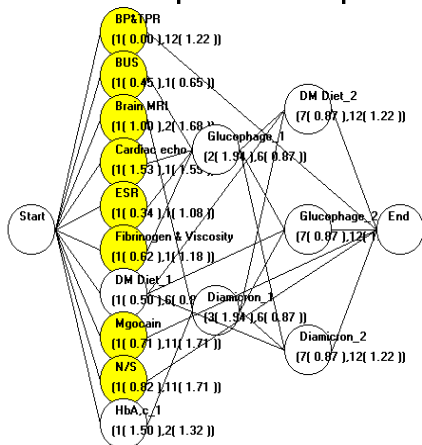


Figure 6. The clinical pathways with routine activities

The routine activities of the this example pattern of clinical pathways are: BP&TPR(1(0.0), 1(1.2)), Brain MRI(1(1.0), 1(1.7)), Mgcocain(1(0.7), 1(1.7)), N/S(1(0.8), 1(1.7)), BUS(1(0.4), 1(0.6)), ESR(1(0.3), 1(1.0)), Cardiac echo(1(1.5), 1(1.5)), and Fibrinogen & Viscosity(1(0.6), 1(1.1)). In order to understand this pattern easily, we can add these routine activities to Figure 5, and obtain the new graph as shown in Figure 6. Many of routine examinations are performed at the

admission day. For example, the item “BP&TPR” refers to the blood pressure, temperature, pulse, and respiration exam. The execution time of “BP&TPR” starts from day 1 to day 12 and its occurrence does not depend on any other activities. The execution of these activities may repeat until discharge. The activities “Glucoophage” and “Diamicron” are the medicine for DM patients and we can also find that these medicines are usually used within the second window of the admission period.

We summarize the size in each iteration and the related average support in mining clinical pathways as shown in Table 1. We can see that, after maximization, large amount of large graphs is removed, so that we can focus on more meaningful paths.

The path shown in Figure 6 is a typical brain stroke patient with diabetes. With this common treatment flow, a new medical staff could easily realize the potential activities and the length of stay for a diabetic patient. However, different complication will result in different paths. For example, a more complex treatment flow is shown in Figure 7. Such path may be assigned to the patients with disease codes 332, 434, 437, 486 and 582. (Note that the diagnosis data for a patient is usually represented as a set of standard international disease codes.) There are about 100 disease codes among 113 patients. In such a complex domain, it raises an issue that how to find the potential relationship between the disease code and the mined clinical pathways. We will use the other data mining technique called the association analysis to discover such kind of knowledge, and describe it in the next subsection.

Table 1. Results of mining time dependency graphs

Number of Vertex	Number of Paths (Before Maximal)	Number of Paths (After Maximal)	Average Support of Large Graph
2	278	29	7%
3	712	94	5%
4	848	115	4%
5	615	107	4%
6	273	92	3%
7	72	29	3%
8	14	5	3%
9	2	2	3%

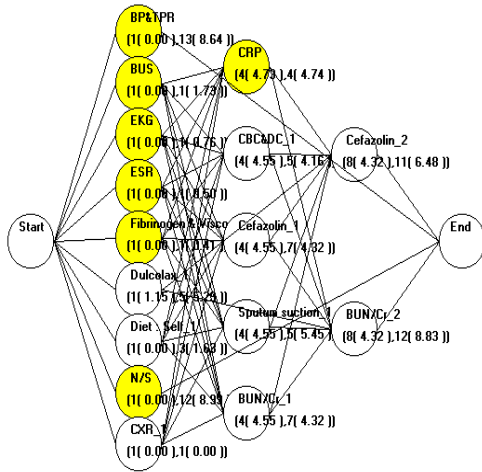


Figure 7. A more complex case

3.3. Prediction

The patterns of clinical pathways release insights and implications to the medical staffs. However, one more important mechanism for physicians and nurses is the ability to predict which paths should be assigned to a new patient when he/she is admitted to the hospital.

The prediction problem is formulated as the follows: *given a new patient's diagnosis data, we should be able to assign a specific clinical pathway to the patient.* However, each patient could be with multiple diagnosis and paths. We anticipate that the association analysis is a proper technique to conquer the prediction problem. Agrawal, Imielinske, and Swami first introduce the method of mining association rules [2]. The main purpose of mining association rules is to discover “what items are bought together in a transaction” over the market basket data. For example, cola and potato chips are purchased together frequently. We adopted the same concept to the tasks of predicting clinical pathways. We first transform all the diagnosis data and paths as items in a transaction. For example, patient “P001” has three disease codes 401, 434 and 250, and two paths P0 and P1. All of these items can be assigned to a single transaction, e.g., P001={401, 434, 250, P0, P1}.

After transformation, we can discover the linkage between the diagnosis and clinical paths.

$$Sim(P_i, P_j) = 2 * |E_i \cap E_j| / (|E_i| + |E_j|) \quad (1)$$

$$PredictionAccuracy = \frac{\sum_{i=1}^n \left[\max_{j=1}^{|Ptrn|} \left(Sim(P_i, Ptrn_j) \right) \right]}{n} \quad (2)$$

For example, under the setting of $GS = 3$, $W = 2$ and $PS = 3$, we then get the result of accuracy by selecting 30 patients randomly as

Similar to our algorithm, users should define a support threshold to indicate the user confidence level. We divide the 113 sample patients into two sets. One is called the training set and the other is the testing set. The large itemsets can be generated by the training set, and then we use the testing set to evaluate the prediction accuracy. The following notations are used in the prediction process.

- (1) Trn : Training set.
- (2) Tst : Testing set.
- (3) L : the large itemsets generated by Trn .
- (4) D : the disease code of patients.
- (5) P : the common paths of patients (large graph).
- (6) GS : the support threshold of mining time dependency graph. For example, when the value of GS is 3%, the path must be applied to at least three percent of patients.
- (7) PS : the support threshold of prediction. For example, when the value of PS is 2%, the association relationship exists for disease code 001 and path $P1$ if there are two percent of patients who are diagnosed to have disease 001 and treated by clinical path $P1$.

For each example of testing patient Tst_i , we then scanned L to calculate the prediction accuracy by the following steps:

- (1) We first identify the frequent clinical paths discovered from the training set. We scan L and get the set of paths $Ptrn$ and $Ptrn \in Ltrn$, where $Ltrn$ is the large itemsets that contains D_i .
- (2) If the Tst_i contains n paths, we then calculate the prediction accuracy by comparing the real paths P_i and the predicted paths $Ptrn_j$. We used the concept of similarity to indicate how precise the testing paths are. The similarity function is defined by formula (1). E_i and E_j in formula (1) are two sets of edges for path P_i and P_j . $|E_i|$, $|E_j|$, and $|E_i \cap E_j|$ denote the number of edges in path P_i , P_j , and $P_i \cap P_j$ respectively.
- (3) The prediction accuracy is defined as shown in formula (2). Note that, after comparing with each predicted path $Ptrn_j$, we pick the most similar paths as predicted paths.

testing set. We calculate accuracy in different threshold level and the result is shown in Figure 8.

We find that the accuracy will decrease when the *PS* increases and the *GS* decrease. This interesting trend may result from the following reasons:

- (1) The lower *GS* is, the more paths are generated. For the patients in the testing set, more paths need to be verified and the accuracy will be lower.
- (2) The lower *PS* is, the more association between the disease code and paths. The accuracy may increase when the number of predicted paths increases.
- (3) The higher *PS* is, the more common predicted paths are generated. However, the pathways for brain stroke is such a complex domain that the ideal accuracy cannot be held if the predicted paths are too general.

We can see an example with such kind of situation: When a patient with disease code 250, 272 and 434 admitted to hospital, we can find the following large itemsets from previous training: 250, 272, 434, P76 and P214. Paths P76 and P214 are the suggested paths that are predicted according to association analysis. Figure 9 and 10 show the resulting paths from this prediction.

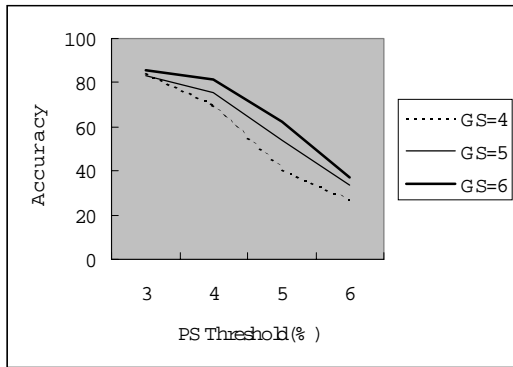


Figure 8. The prediction accuracy under different *GS* and *PS*

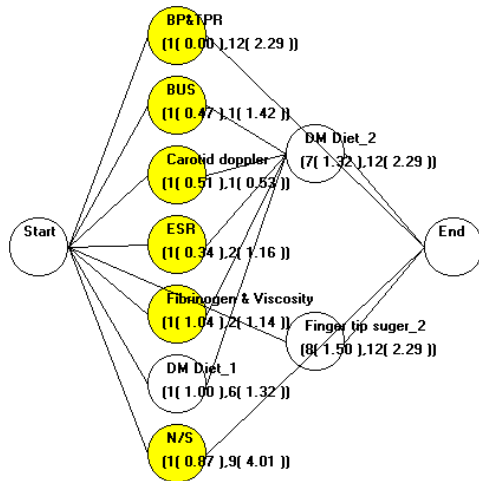


Figure 9. The Predicted Path "P76"

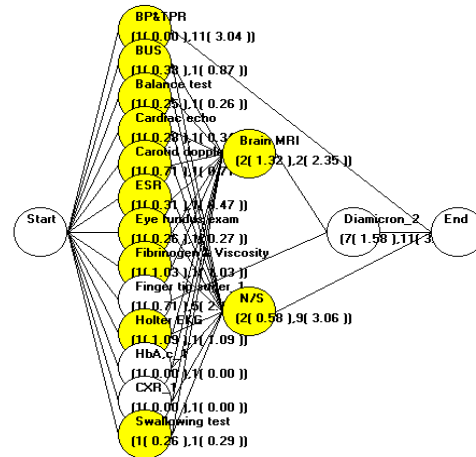


Figure 10. The Predicted Path "P214"

4. Conclusions and future research

Data mining is an application dependent technique. In this research, we focus on investigating the data mining techniques for discovering knowledge from process patterns, such as clinical pathways. We designed a new algorithm to discover the time dependency patterns based on the sequential patterns, and applied it to the development of clinical pathways in curing brain stroke. The resulting clinical pathways can facilitate the continuous improvement of assigning more suitable paths to patients.

However, when we tried to implement our tasks on the application in clinical pathways, we found that it is such a complex problem that time dependency may be one of major factors. In the future work of mining clinical pathways, the following directions may be good for further research.

- (1) More data should be collected, such as the daily states of patients, in order to identify the thinking model of doctors.
- (2) The definition of time dependency constraint could be relaxed because there are some minor activities, which decrease the support of candidate graphs. We can classify activities into different types according to the occurrence frequency and domain knowledge. Some minor activities can also be combined as a pseudo activity to increase the connectivity of different path segments.
- (3) Besides mining the common workflow patterns, the variance analysis is an important issue for clinical pathways. The advanced research can pay more attention on discovering the deviation patterns and providing a warning mechanism to help a manager to predict the potential variance.
- (4) Bayesian networks may be used as a

representation model. With Bayesian networks, we can present the probability between activities to predict the possibility of potential variance. It is also beneficial in analyzing and inferring further implication of mining results. Bayesian networks can be defined as a graphical model for probabilistic relationships among a set of variables [8]. The application of Bayesian networks has become a popular representation for expert knowledge over the last decade. More recently, researchers have developed methods for learning Bayesian networks to extract knowledge from data. For example, Ezawa and Norton developed the Advanced Pattern Recognition and Identification system (APRI) which automatically construct Bayesian network model to classify the uncollectible telecommunications accounts [6]. We believe that it will help us to perform a more complex task in clinical pathways.

- (5) The cost variable can be considered. After all, the cost is the most important reason why many hospitals implement clinical pathways. Cost also is a good evaluation criterion to know the effect of clinical pathways.
- (6) Resource allocation is an important factor affecting the outcome the care plan. In a well-designed Intranet environment, such data can be available, and can be used to improve the clinical pathways.

5. Acknowledgement

We would like to thank Dr. Shunsheng Chen at the Department of Neurology of Chung-Ho Memorial Hospital of Kaohsiung Medical University, Kaohsiung, Taiwan, R.O.C. With his support, we can obtain the medical records of patients with brain stroke treatment history.

6. References

- [1] Aho, K. et al. Cerebrovascular disease in the community: result of a WHO Collaborative Study. *Bull World Health Organ*, 58: 113-130, 1980.
- [2] Agrawal, R., and T. Imielinske, and A. Swami. Mining Association Rules between Sets of Items in Large Databases,

- Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, DC, USA, pp.207-216, May1993.
- [3] Agrawal, R., and R. Srikant, Mining Sequential Patterns, *Proceedings of the 11th International Conference on Data Engineering (ICDE)*, Taipei,Taiwan, May 1995.
 - [4] Agrawal, R., and D. Gunopulos and F. Leymann, Mining Process Models from Workflow Logs, *Research Report RJ 10100(91916)*, IBM Almaden Research Center.
 - [5] Dombovy, B. Sandok, and J. Basford. Rehab for stroke: a review. *Stroke* 17:363-365, 1986.
 - [6] Ezawa, Kazuo J. and Steven W. Norton, Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts, *IEEE Expert*, October 1996, pp.45-51.
 - [7] Healy, L.W., M.E. Ayers, R. Iorio, D.A. Patch, D. Appleby and B.A. Pfeifer, Impact of a Clinical Pathways and Implant Standardization on Total Hip Arthroplasty, *The Journal of Arthroplasty*, 13(3), 1998.
 - [8] Heckerman, David, Bayesian Networks for Data Mining, *Microsoft research*, 9S, Redmond, WA 98052-6399, 1996.
 - [9] Ireson C.L., Critical Pathways: Effectiveness in Achieving Patient Outcomes, *The Journal of Nursing Administration*, 27(6): 16-23, 1997.
 - [10] Quigley, P.A., S.W. Smith, and John Strugar, Successful Experiences with Clinical Pathways in Rehabilitation, *Journal of Rehabilitation*, April/May/June 1998, pp. 29-32.
 - [11] Ross, G., D. Johnson and M. Kobernick, Evaluation of a critical pathway for stroke. *J Am Osteopath Assoc*, , 97:5, 269-272, 1997 May.
 - [12] Vanclay F. Stroke rehabilitation. *Journal of Clinical Epidemiol* 44:22-28, 1991.