# Temporal Abstraction and Data Mining with Visualization of Laboratory Data

**Katsuhiko Takabayashi [a], Tu Bao Ho [b], Hideto Yokoi [c], Trong Dung Nguyen [b], Saori Kawasaki [b], Si Quang Le [b], Takahiro Suzuki [a], Osamu Yokosuka [d],**

[a] *Division for Medical Informatics and Management, Chiba University Hospital, Inohana, Chuou-ku, Chiba, 260-8677 Japan*
[b] *Japan Advanced Institute of Science and Technology,Tatsunokuchi, Ishikawa, 923-1292 Japan*
[c] *Division for Medical Informatics, Kagawa University Hospital, Ikedo, Miki, Kida-gun, Kagawa, 761-0793 Japan*
[d] *Department of Gastroenteology, Chiba University Hospital, Inohana, Chuo-ku, Chiba, 260-8677 Japan*

### Abstract

*To analyze the laboratory data by data mining, user-centered universal tools have not been available in medicine. We analyzed 1,565,877 laboratory data of 771 patients with viral hepatitis in order to find the difference of the temporal changes in laboratory test data between Hepatitis B and Hepatitis C by the combination of temporal abstraction and data mining. The data for one patient is temporal for more than 5 years. After pretreatment the data was converted to abstract patterns and then selected into sets of data combination and rules to identify Hepatitis B or C by D2MS and LUPC which were originally produced by ourselves. Not only data pattern, but also temporal relations were considered as a part of the rules. In the course of evaluating the results by domain experts, even though there were not so remarkable hypotheses, visualization tools made it easier for them to understand the relations of the complicated rules.*

***Keywords:***

databases, liver function test, hepatitis.

## Introduction

Data mining is beginning in medical application. Even though there are many data mining techniques to analyze the database, most of them are still experimental and in the hands of computer scientists. Medical doctors cannot use the tools for their own data analysis individually. User-centered universal tools should be applied for medical researchers to analyze their own data. In Japan we started the national project of active mining to analyze the same database of viral hepatitis and evaluate the availability and validity to implement it in medicine [1]. In this approach from the medical point of view, not only providing data, but we performed pretreatment of medical data and attended evaluation with mathematical researchers as active mining. One of the keys of laboratory data analysis in medicine is how to treat temporal data. By using temporal abstraction, we aimed to solve this problem.

### The Goals

One of the concrete goals of this project was to discover the differences in the temporal patterns of hepatitis B (HBV)and C (HCV) which has not been clearly defined, and, more importantly, to examine whether the methods we applied here can work well and be applied to other fields.

## Materials and Methods

### Database

The hepatitis data are located in two databases, one is the laboratory database of the hospital, and another is the biopsy and clinical data in the department of Medical School, Chiba University. The contents are as follows:

- Basic information of patients (total 771 records)
- Results of biopsy (total 960 records)
- Information about measurements in in-hospital tests (total 459 records)
- Results of out-hospital tests (total 30,243 records) and those of in-hospital tests (total 1,565,877 records)

The data in the hospital are more than 20 years. Those patients were performed liver biopsy once at least.

### Preprocessing

Our preprocessing of hepatitis data includes data cleaning, data integration, data reduction, deidentification and data transformation. We removed only redundant and not unsuitable suffix data for further processing and we eliminate noisy data in the next temporal abstraction step. For the purpose of temporal abstraction, we have to integrate original relational data tables into one data table, where each column represents laboratory examination. By combining the expert guidance and the frequencies of attributes presented from 983 examinations, we selected 15 most significant examinations. The numbers of examinations for each patient are different, and the examination periods are irregular.

### Temporal Abstraction

Temporal abstraction (TA) is one approach to deal with time-related data in medical research. TA methods are those able to derive an abstract description of temporal data by extracting their most relevant features [2]. The key idea is to transform time-stamp points by abstraction into an interval-based representation of data by extracting their most relevant features [3]. TA task can be defined as follows.

The input includes a set of time-stamped data points (events). The output includes a set of interval-based, context-specific unified values or patterns (usually qualitative) at a higher level of abstraction. TA can be generally considered in two phases: *basic* TA for abstracting time-stamped data from given episodes (which are significant intervals for the investigation purpose) and *complex* TA for investigating specific temporal relationships between episodes that can be generated from a basic TA or from other complex TAs.

### Basic Temporal Relations

We started by a separation of two groups of tests, one with values that change rapidly in the short term such as GOT, GPT, TTT and ZTT and the other with values that change slowly in the long term such as T-CHO, CHE, ALB.

Basic temporal abstractions typically extract *states* (e.g., low, normal, high), and/or *trends* (e.g., increase, stable, decrease) from a uni-dimensional temporal sequence.

The essential ideas of our temporal abstraction methods here is to deal with long and irregular time-stamp sequences, and doing abstraction in efficient. We introduce the notion of “changes of state” to characterize the slowly changing tests, and the notions of “base state” and “peaks” to characterize the rapidly changing tests.

#### *Temporal abstraction primitives*

From observation and analysis, we defined the following temporal abstraction primitives:

1. State primitives: N (normal), L (low), VL (very low), XL (extreme low), H (high), VH (very high), XH (extreme high).
2. Trend primitives: S (stable), I (increasing), FI (fast increasing), D (decreasing), and FD (fast decreasing).
3. Peak primitives: P (peaks occurred).
4. Relations: “>” (change state to), “&” (and), “–“ (and then), “/” (“X/Y” means majority of points are in state X and minority of points are in state Y).

The thresholds to distinguish the state primitives of tests are given by medical doctors, for example, those of the test GOT are 40, 100, 200, respectively. We define structures of abstraction patterns as follows:

<pattern> ::= <state primitive>
<pattern> ::= <state primitive> <relation>
<pattern> ::= <state primitive> <relation> <peak>
<pattern> ::= <state primitive> <relation> <state primitive>

Examples of abstracted patterns in a given episode are like follows:

- “GOT = H” (all the values of GOT in one case are above the normal region as shown in the upper left in Figure 1),
- “GPT = H&P” ( The values of GPT in one case are always high and with peaks like lower left of Fig 1),
- “I-BIL = N>L>N” (I-BIL first is normal, then changed to the low region, and finally changed to the normal region in one case like the right bottom in Fig 2), etc.

Figure 1 shows typical possible patterns (8 and undetermined) for rapidly changing tests, and Figure 2 shows typical possible patterns (21 and undetermined) for slowly changing tests [4].

#### *Abstraction of rapidly changing test results*

From our observation and analysis, especially GPT and GOT were defined as rapidly changing attributes, which can go up in a very short period and go back to a “stable” state. Thus two most representative characteristics of these tests are a “stable” base state (BS), and the position and value of peaks, where the attributes suddenly go up. Based on this, we formulated the following algorithm to find the base state and peaks of a test. Rapidly changing tests applied also to TTT and ZTT and they showed 9 patterns.

**Algorithm 1** (for rapidly changing tests)
**Input:** A sequence of patient’s values of a test with length N denoted as $S_{00} = \{s1, s2, \ldots, sN\}$ in a given episode.
**Output**: Base state and peaks, and an abstraction of the sequence derived from them.
Parameters: NU, HU, VHU, XHU: upper thresholds of normal, high, very high, extreme high regions of a test, a (real).
Notation:

- Mi: Set of local maximum points of S
- BS: base state of S
- PEi: set of peaks of S

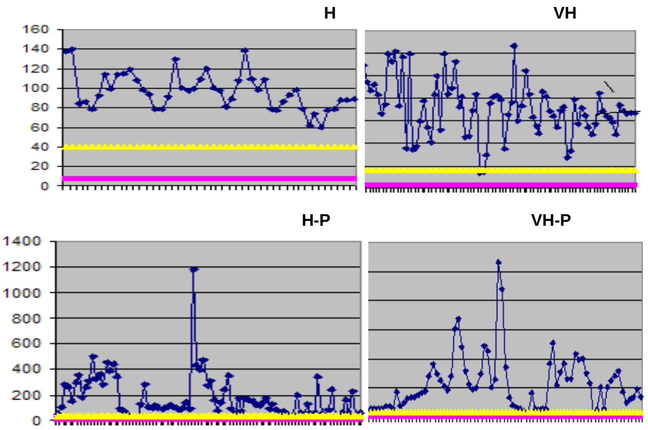


*Figure 1 rapidly changing test patterns*

### *Abstraction of slowly changing test results*

The key idea is to use the "change of state" as the main feature to characterize sequences of the tests. It can be seen that the "change of state" characterize information of both state and trend of the sequences.

From the beginning of a sequence, the first data points can be at one of the three states "N", "H", or "L". It will happen that:

Either the sequence changes from one state to another state, smoothly or variably (at boundaries), or the sequence remains in its state without changing.

We provided 22patterns for slowly changing data.

N/H normal & high
H-I high with increase
L>N low to normal etc

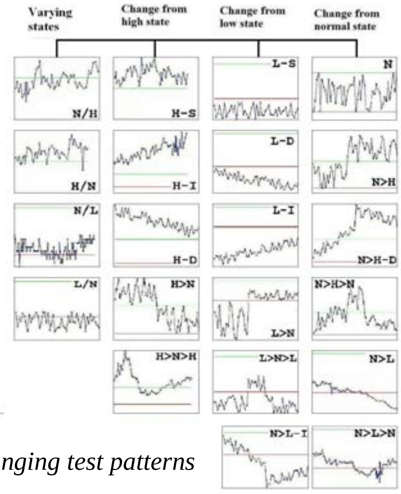
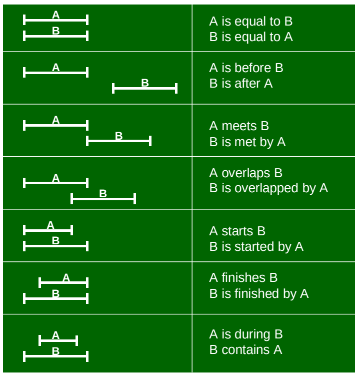


*Figure 2 slowly changing test patterns*

## Temporal relationships

The temporal relations between the abstracted events of laboratory data were also treated here as phenomena and a part of the rules by comparing the period of the state. They are classified into seven relations by Allen [8].



- Relations between two basic patterns each happens in a period of time,

for example

ALB is Normal to Low

after

GOT is Normal to high

**(Allen's Temporal Logic, 1984)**

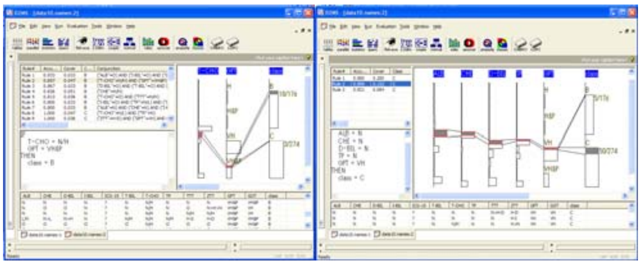
*Figure 3 temporal relations*

## Complex Temporal Abstraction

### *Mining abstracted hepatitis data with system D2MS*

The authors developed an interactive visualization tool in decision tree construction called D2MS (data mining for Model Selection) for supporting an effective cooperation of the user and the computer in classification. D2MS shares many features with WinViz [5] and Cviz that both use parallel coordinates. WinViz allows the user to visually examine a tabular database and to formulate query interactively and visually. Cviz is an attempt to integrate visualization into the knowledge discovery process.

D2MS facilitates the trials of various alternatives of algorithm combinations and their settings. The data mining methods in D2MS consists of programs CABRO for tree learning and LUPC for rule learning [6]. CABRO produces decision trees using R-measure and graphically represents them in particular with T2.5D tool (trees 2.5 dimension). As shown in *Figure 4*, visualization made us easily recognized the different pattern of HBV and HCV.



T-cholesterol is mostly normal partly high
And GPT is extremely high with peaks
is HBV hepatitis (85.7% confidence)

Albumin normal, Ch-E normal, bilirubin normal, TP normal and GPT is extremely high without peak
is HCV (90.7% confidence)

*Figure 4 Rules for HBV and HCV by D2MS*

$$Z = \frac{conf\,(R) - p(C)}{\sqrt{p(C)(1 - p(C)/n_A}}$$

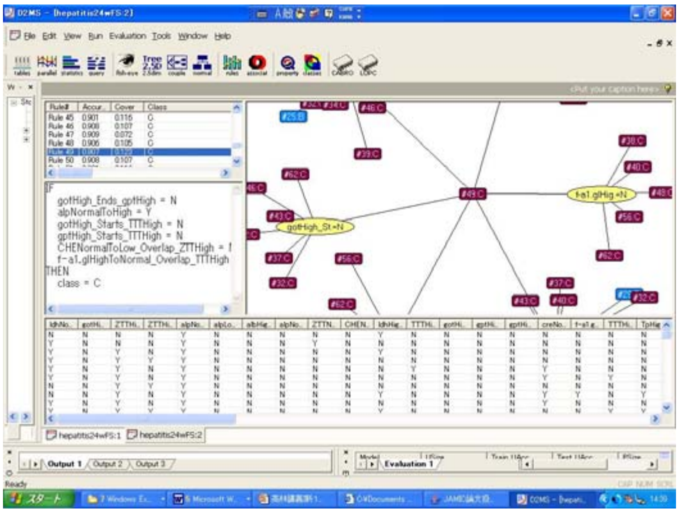
We examined statistical significance of the consequence according to the method of [7], which prunes discovered rules statistically as follows.

Assume a rule R: A → C (or R: A → ¬C) with confidence conf(R). If conf (R) = p(C) then R is eliminated. To test whether conf (R) = p(C), we use the following test statistic where $n_A$ is the number of cases satisfying C.

### *Mining abstracted hepatitis data with LUPC*

LUPC is a separate-and-conquer algorithm that controls the induction process by several parameters. The parameters allow obtaining different results and this ability allows the user to play a central role in the mining process [6].

LUPC is developed to learn prediction rules from supervised data. Each rule found by LUPC is a conjunction of attribute-value pairs that may present an interesting pattern. The main features of LUPC are (1) its ability of finding rules with associate domain knowledge (such as finding rules containing or not containing specified attribute-value pairs), as well as finding rules for minority classes; (2) it is integrated with D2MS's rule visualizer and thus supports the user in selecting the appropriate rules which result from different possible settings of parameters. The performance of LUPC depends on several parameter specified by the user: α for min accuracy of rules, β for min coverage of rules, γ for maximal number of candidate rules in the beam search, and η for maximal number of attribute-value pairs to be consider. By varying these parameters we can find different sets of rules [6]. When using the setting with default parameters of α = 80%, β = 3, γ = 200, and η = 100, we found 119 rules characterizing the hepatitis B and 152 rules characterizing hepatitis C.


D2MS – [hepatitis24wFS:2]
File Edit View Run Evaluation Tools Window Help
tables parallel statistics query fish-eye 2.5dim couple normal rules associat
gotHigh_St=N
f-a1.glHig=N
hepatitis24wFS:1
hepatitis24wFS:2
Output 1
Output 2
Output 3
Evaluation 1
Ready

| Rule# | Accur.. | Cover | Class |
|---|---|---|---|
| Rule 45 | 0.901 | 0.116 | C |
| Rule 46 | 0.908 | 0.107 | C |
| Rule 47 | 0.909 | 0.072 | C |
| Rule 48 | 0.906 | 0.105 | C |
| Rule 49 | 0.907 | 0.129 | C |
| Rule 50 | 0.908 | 0.107 | C |

```
IF
   gotHigh_Ends_gptHigh = N
   alpNormalToHigh = Y
   gotHigh_Starts_TTTHigh = N
   gptHigh_Starts_TTTHigh = N
   CHENormalToLow_Overlap_ZTTHigh = N
   f-a1.glHighToNormal_Overlap_TTTHigh
THEN
   class = C
```

*Fig 5    LUPC    Rules can be illustrated in a left figure*

**Evaluation**

The produced rules were evaluated by three experts.

## Results

By using D2MS, we discovered 33,477 rules for type B and C difference. These rules are complicated and sometimes contradictory to each other. For example, there is a rule "if T-Bilirubin(Bil) is N(normal) , ZTT is N, and GOT is N with P(peak), then HBV", while there is another rule "if D-Bil is N, TTT is N, GOT is N with P, then HCV", which are almost the same but the results are completely different.

After pruning by statistical aspects between HBV and HCV, there are only 27 rules (0.08%) left (*Table1*). However, these rules seemed not attractive for medical doctors even if they are statistically significant. For example, "T-cholesterol is normal is HCV" in 171/260 cases (66%), or "GPT is high with a peak and ZTT is mostly high partly normal is HBV". They are too simple or vague and must be carefully assessed.

Different datasets were found by using LUPC with various parameters including temporal relations between laboratory tests. *Table2* presents the top five rules by LUPC from the point of coverage and confidence. From this table, especially rule3 and 5 are similar and could be merged as a rule that "if TP decreasing high to normal and both ZTT and TTT is high, then it is HCV". In the evaluation of medical doctors, though most of them seemed not crucial or not useful in clinical medicine even the discovered rules covering many cases with high accuracy. However, some of the rules could be reasonable from the different clinical course of two types of hepatitis, especially when the experts checked and integrated the rules in the illustration.

| No. | class | T-CHO | CHE | GOT | GPT | TTT | ZTT | D-BIL | T-BIL | I-BIL | TP | acc | ratio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | C (95%) | N | | | | | | | | | | 0.66 | 171/260 |
| 2 | C (95%) | | N | | | | | | | | | 0.72 | 183/256 |
| 3 | C (95%) | | | | | | | N | | | | 0.73 | 180/248 |
| 4 | C (95%) | | | H | | | | | | | | 0.76 | 89/117 |
| 5 | C (95%) | | | | H | | | | | | | 0.76 | 78/103 |
| 6 | C (95%) | | N | | | | | N | | | | 0.82 | 142/173 |
| 7 | C (95%) | | | | | N | H/N | | | | | 0.92 | 11/12 |
| 8 | C (95%) | | | | H | | | | N/H | | | 0.93 | 14/15 |
| 9 | B (95%) | | | | H&P | | H/N | | | | | 0.92 | 11/12 |
| 10 | B (90%) | | | | | | N | | | | | 0.68 | 63/93 |
| 11 | B (90%) | | H/N | | | | | | | | | 0.7 | 14/20 |
| 12 | B (90%) | | N/H | | | | | | | | | 0.74 | 23/31 |
| 13 | B (90%) | | | N&P | H&P | | | | | | | 0.7 | 16/23 |
| 14 | B (90%) | | N | | | | N/H | N | | | | 0.88 | 7/8 |
| 15 | C (90%) | | N | | | N | | | | | | 0.8 | 67/84 |
| 16 | C (90%) | | | | | | H-I | N | | | | 0.95 | 63/66 |
| 17 | C (90%) | | | H | XH | | | | | | | 0.92 | 11/12 |
| 18 | C (90%) | | N | N | | | | | N | | | 0.93 | 26/28 |
| 19 | C (90%) | | | | | | | | | | N/H | 0.81 | 35/43 |
| 20 | C (90%) | | | | | H-I | | | | N | | 0.93 | 25/27 |
| 21 | C (85%) | | | | N&P | N | | | | | | 0.69 | 33/40 |
| 22 | C (85%) | | N | N | | | | | | | | 0.84 | 41/49 |
| 23 | C (85%) | | N | H | | | | | | | | 0.87 | 58/67 |
| 24 | C (85%) | | | | | H/N | | | | N | | 0.79 | 23/29 |
| 25 | C (85%) | | | | XH | | | | | | | 0.72 | 18/25 |
| 26 | C (85%) | N/L | | | | | | | | | | 0.8 | 28/35 |
| 27 | C (85%) | | | | | | H-D | | | | | 0.83 | 33/40 |

*Table 1 27 pruned rules produced by D2MS and satisfying chi-square test*

Rule 1 (Coverage ; 4.098%  confidence ; 100%  coverage ; 25 cases)

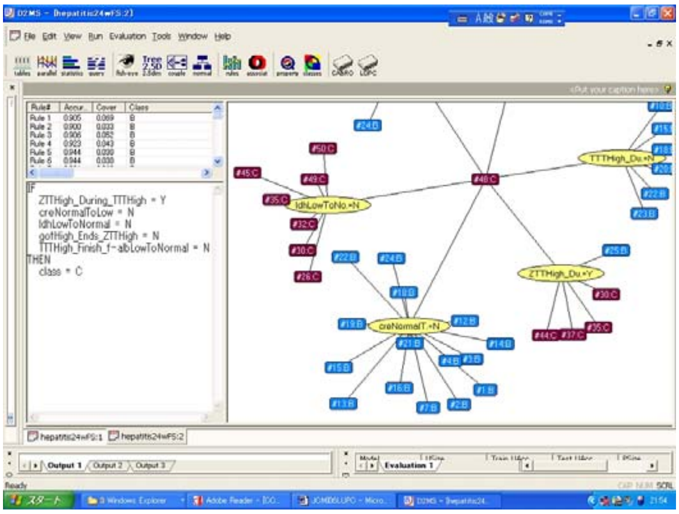
creNormalToLow = Y  
gptHigh_Start_gotHigh = Y  
-> Class = C

Creatinine decreasing from normal to low  
and GPThigh start with GOT high  
is HCV

Rule 2 (Coverage 3.443%  confidence ;100%  coverage;21 cases)  
Bilirubin decreasing from normal to low before TTT elevates  
is HCV  
Rule 3 (Coverage 3.443% confidence ;100% coverage;21 cases)  
TP decreasing high to normal and ZTT goes up to high after TTT up to high  
is HCV  
Rule 4 (coverage 3.279% confidence 100% coverage; 20 cases)  
Creatinine decresing normal to low and bilirubin increasing normal to high  
is HCV  
Rule 5 (coverage 2.951% confidence 100 % coverage; 18 cases)  
TP decreasing high to normal and ZTT goes up to high before TTT up to high  
is HCV

*Table 2 Top 5 rules selected by LUPC*

In the viewer of LUPC we can see the accuracy and coverage on the upper left, rule itself in the middle and the relation of the rules and attribute value pairs in the figure to the right which can be manipulated by users (*Figure 6*). By handling it users can see the relations of each item. In this figure doctors could easily recognize "if LDH is low to normal is false", then all the rules are related to HCV, while "if creatinine is normal to low is false", then it is related to the rules of HBV except rule #48(center), so that the doctors could understand from a more comprehensive point. This technique was highly evaluated by medical doctors and some rules such as the top 5 were considered as meaningful.



| Rule# | Accur... | Cover | Class |
|---|---|---|---|
| Rule 1 | 0.905 | 0.069 | B |
| Rule 2 | 0.900 | 0.033 | B |
| Rule 3 | 0.906 | 0.052 | B |
| Rule 4 | 0.923 | 0.043 | B |
| Rule 5 | 0.944 | 0.030 | B |
| Rule 6 | 0.944 | 0.030 | B |

```
IF
  ZTTHigh_During_TTTHigh = Y
  creNormalToLow = N
  ldhLowToNormal = N
  gotHigh_Ends_ZTTHigh = N
  TTTHigh_Finish_f-albLowToNormal = N
THEN
  class = C
```

*Figure 6 LUPC makes it easier for the users to understand the relations comprehensively*

## Discussion

We have presented a temporal abstraction approach to data mining the temporal hepatitis data. Most doctors do not believe that they can distinguish HBV and HCV from the laboratory data change, so it might be true when we cannot obtain any new findings from this data mining. However, we in fact obtained many rules to identify HBV or HCV with statistical significance. Some of them look very simple. To confirm, we need to analyze them by stratified analysis, removing the modification of other factors such as treatment and then compare again. By LUPC, we can estimate the patient data comprehensively and expect new findings in many diseases because it is difficult for human beings to find data changes over a long period of time. some experts of liver diseases mentioned that the cases with HBV and HCV were apt to show different clinical changes, and our results would reflect these changes.

One of the major problems in rule based data mining is that there are too many rules deduced for us to evaluate. To select most important rules we introduced the chi-square test which was effective to decrease the number of rules as well as statistical reasoning. Another is by LUPC, not only selecting minority classes from large unbalanced datasets but visualization, it is not difficult to separate the important ones from many rules for medical doctors.

Other studies of data mining in medicine are mostly in the field of genomics and epidemiology and the analysis of laboratory data is quite limited. We provided the data of anti-phospholipid antibody syndrome to PKDD 1999 as model data in order to establish a new technique as well as hepatitis data.

Current medical study is deeply inclined to use a prospective way or Cohort as a scientific study design, which implies a carefully planed experiment. However, when we think of a long-term experiment lasting over 10 years, it is not realistic to study prospectively. There is a great possibility of new paradigms appearing before the study is completed. Retrospective studies are expected for these long term studies and data mining techniques will play a major role in this filed by creating high potential hypotheses.

Even though we did not discover crucial rules to show the difference of laboratory data change between HBV and HCV, we proved to show that this combination of TA and data mining with visualization is useful and effective. Furthermore, it could be applied to other fields of medicine and would be a basic model for the universal analysis of data mining for temporal data analysis in medicine.

## Conclusion

The rules that show the difference of the laboratory changes in the long clinical course between the HBV and HCV could be deduced by D2MS. Pruning by statistical significance could decrease the number of rules but obtained rules were not interesting in individuals. Visualization made it easier for doctors to find the relations and led them to find reasonable results. This combination technique of temporal abstraction and data mining with visualization could be applied universally.

## References

[1] Motoda, H., Active Mining: New Directions of Data Mining, IOS Press, 2002.

[2] Horn, W., Miksch, S., Egghart, G., Popow, C., Paky, F., "Effective Data Validation of High-Frequency Data: Time-Point-, Time-Interval-, and Trend-Based Methods", *Computer in Biology and Medicine, Special Issue:Time-Oriented Systems in Medicine*, 27(5), 389-409, 1997.

[3] Bellazzi, R., Larizza, C., Magni, P., Monntani, S., and Stefanelli,M. Intelligent Analysis of Clinic Time Series: An Application in the Diabetes Mellitus Domain *Artificial Intelligence in Medicine*, 20, 37-57, 2000.

[4] Ho, T.B., Nguyen, T.D., Kawasaki, S., Le, S.Q., Nguyen, D.D., Yokoi, H., Takabayashi, K. Mining Hepatitis Data with Temporal Abstraction. *ACM Inter. Conf. on Knowledge* Discovery and Data Mining KDD-03, 369-377, 2003.

[5] Lee, H.Y., Ong, H.L., and Quek, L.H. Exploiting Visualization in Knowledge Discovery. *First Inter. Conf. on Knowledge Discovery and Data Mining,* 1995, pp. 198–203.

[6] Ho T.B., Nguyen T.D., Nguyen D.D., Kawasaki S., Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining. *Int. Journal Artificial Intelligence Tools,* 2001; 10(4): 691-713.

[7] Bruzzese, D. and Davino, C. Statistical Pruning of Discovered Association Rules. *Computational Statistics,* 16 (3), pp. 387 – 398, 2001.

[8] Allen E and Jai S. A decidable temporal logic to reason about many processes. Proceedings of the ninth annual ACM symposium on Principles of distributed computing table of contents. Quebec City, Canada 233-246,1990.