POSSIBILITY OF INTEGRATED DATA MINING OF CLINICAL DATA

Akinori Abe[1*, 2*], Norihiro Hagita[1, 3], Michiko Furutani[1], Yoshiyuki Furutani[1], and Rumiko Matsuoka[1]

*[*1] International Research and Educational Institute for Integrated Medical Science (IREIIMS), Tokyo Women's Medical University, 8-1 Kawada-cho, Shinjuku-ku, Tokyo 162-8666 JAPAN*
E-mail: {michi,yoshi,rumiko}@imcir.twmu.co.jp
*[*2] ATR Knowledge Science Laboratories, 2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 JAPAN*
E-mail: ave@ultimaVI.arc.net.my
[3] ATR Intelligent Robotics and Communication Laboratories, 2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 JAPAN
E-mail: hagita@atr.jp

ABSTRACT

In this paper, we introduce integrated data mining. Because of recent rapid progress in medical science as well as clinical diagnosis and treatment, integrated and cooperative research among medical researchers, biology, engineering, cultural science, and sociology is required. Therefore, we propose a framework called Cyber Integrated Medical Infrastructure (CIMI). Within this framework, we can deal with various types of data and consequently need to integrate those data prior to analysis. In this study, for medical science, we analyze the features and relationships among various types of data and show the possibility of integrated data mining.

KEYWORDS: Integrated data mining, Clinical data, Influential data, Decision tree, Cyber Integrated Medical Infrastructure (CIMI)

1 INTRODUCTION

Medical science as well as clinical diagnosis and treatment has progressed rapidly in recent years with each field becoming more specialized and independent. As a result, cooperation and communication among researchers in the two fields has decreased, which has led to problems between both communities, not only in terms of medical research but also of clinical treatment. An integrated and cooperative approach to research between medical researchers and biologists is needed. Furthermore, we are living in a changing and quite complex society, so important knowledge is always being updated and becoming more complex. Therefore, integrated and cooperative research needs to be extended to include engineering, cultural science, and sociology. As for medical research, the integration of conventional (Western) and unconventional (Eastern) medical research, which should be fundamentally the same but in fact are quite different, has been suggested.

With this situation in mind, we propose a framework called Cyber Integrated Medical Infrastructure (CIMI). This framework of integrated management of clinical data on computer networks consists of a database, a knowledge base, and an inference and learning component, which are connected to each other over the network. In this framework, medical information (e.g. clinical data) is analyzed or data mined to build a knowledge base for predicting all possible diseases and to support medical diagnosis.

For medical data mining, several techniques such as Inductive Logic Programming (ILP), statistical methods, decision tree learning, Rough Sets, and KeyGraph have been applied (e.g. (Ichise & Numao, 2005), (Tsumoto, 2004) and (Ohsawa, 2003)), and acceptable results have been obtained. For data mining, generating plausible or correct results is of course important, but if the results are trivial or well known, they are not so important for physicians. Thus, the research focus has recently shifted from how to obtain proper and plausible results to how to obtain interesting results. In this context, ``interesting'' means ``interesting for physicians,'' that is, to discover knowledge which doctors were unaware of or had previously ignored and which represents the mechanisms which lead to serious disease. Of course, the generated knowledge should be correct. In fact, in the knowledge discovery field, we focus on discovering not only frequently occurring trends but also rare or novel events. The aims of research in ``active mining'' (Tsumoto et al., 2005) and ``chance discovery'' (Ohsawa & McBurney, 2003) are to establish techniques or procedures for discovering interesting, rare, or novel knowledge or events. Previously, we applied C4.5 (Quinlan, 1993) to medical data to discover hidden relationships (Abe, Kogure, & Hagita, 2003) and pointed out the importance of discovering knowledge covering gray areas. In fact, general induction or statistical analysis such as C4.5 might discover such knowledge, but such rare or novel knowledge tends to be hidden among general knowledge when we conduct general data mining.

In the above analyses, we dealt with clinical data that consists of mostly the same data types, so we could analyze the data regardless of the relationships between data from different categories. However, if we deal with multiple categorized data, it is rather difficult to discover such hidden or potential knowledge. Furthermore, the discovery of normal knowledge can be disturbed by influential data in other categories. It might be necessary to analyze the features of data sets to determine their relationships and influence patterns. In this study, we analyze actual clinical data, consisting of multiple categorized items collected by the International Research and Educational Institute for Integrated Medical Sciences (IREIIMS) project, to show features of the data and suggest methods to discover their hidden or potential features. In section 2, we introduce the Cyber Integrated Medical Infrastructure (CIMI). In section 3, we describe the features of clinical data. In section 4, we present initial results obtained by analyzing the clinical data to find relationships among the differently categorized data. In section 5, we review the possibility of integrated data mining.

2 CYBER INTEGRATED MEDICAL INFRASTRUCTURE

Recently, for medical research, integration and cooperation among the various fields has been advocated. Therefore, we propose a framework called Cyber Integrated Medical Infrastructure (CIMI), which is a framework of integrated management of clinical data on computer networks. Figure 1 is an image of CIMI. As shown, CIMI consists of a database, a knowledge base, and an inference and learning part, which are connected to each other in the network. Various types of data are stored in the database to be analyzed by machine learning techniques. The analyzed results are regarded as medical knowledge and stored in the knowledge base to be used in computational inferences to support medical diagnosis. In addition, the knowledge base is used with the training of medical students and doctors, which will be described elsewhere.

Thus, in the CIMI framework, medical, clinical, and other information (e.g. personal information, interview results) are analyzed or data mined to discover relationships among the medical, clinical, and other data and all possible diseases.

Although, CIMI includes various types of data, as a first step we mainly deal with clinical data. The clinical data includes liver, pancreas, and kidney test results, tumor markers, and blood test results. Clinical interviews,

Ryodouraku results, and plethysnographic analysis (Eastern medicine) data will be added. Thus, the database will contain data from both Western and Eastern medicine. As shown in the previous section, Western and Eastern medical research are conducted separately. One of our aims is to discover relationships between Western and Eastern medical treatment.

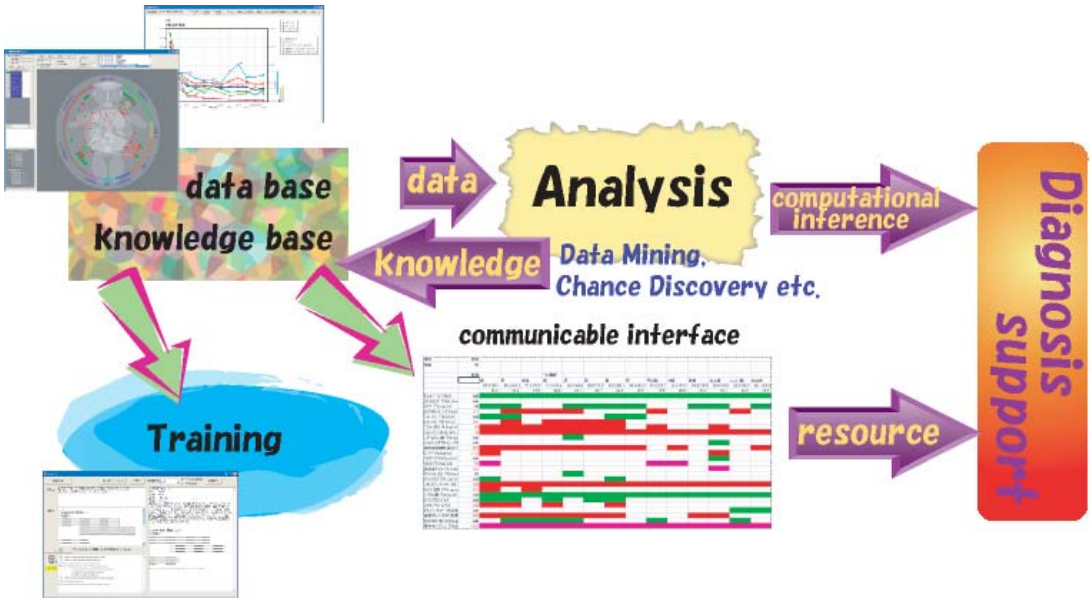

Figure 1. Cyber Integrated Medical Infrastructure}

3 FEATURES OF THE CLINICAL DATA

In this section, we describe the features of the clinical data collected for CIMI. Although various types of data will be stored in the database, we are mainly dealing with clinical data here.

3.1 Clinical data

To construct the database in CIMI, we are now collecting various types of clinical data, such as those obtained in blood and urine tests. Currently, more than 100 items are included in the clinical data. In fact, they are clinical data, but they can be categorized more precisely as follows (for the data described in this paper):

1. liver, pancreas, and kidney test data: 24 items
2. metabolic function test data: 29 items
3. general urine test data: 11 items
4. blood and immunity test data: 31 items
5. tumor markers: 36 items

These categories include the following items.

1. Total protein, albumin, serum protein fraction, alpha-globulin

2. Na, K, Ferritin, total acid phosphatase
3. Urobilinogen, urine acetone
4. Mycoplasma pneumoniae antibody, cellular immunity
5. Immunosuppressive acidic protein, Sialyl Le X-i antigen, urine β2-microglobulin

In addition, data from clinical interviews, family trees, and lifestyles are collected. Although these data are relevant to the health status, many factors, which were neither formalized nor coherent, were included in the interview data, so we did not fully analyze this data but will do so in a future study.

Currently, we have collected data from about 1000 persons (for some, the data were collected more than once.). It is quite hazardous to directly analyze such a large amount of data. Therefore, we analyzed data from only 77 persons. In addition, health levels are assigned by doctors from the clinical data and by an interview. Health levels that express the health status of patients are defined based on *Tumor stage* (Kobayashi & Kawakubo, 1994) and modified by Matsuoka. Categorization of the health levels is shown in Fig. 2 (``%" represents the percentage of persons in the level.). Persons at level I and II can be regarded as being healthy, but those at levels III, IV, and V can possibly develop cancer. In (Kobayashi & Kawakubo, 1994), level III is defined as the stage before the shift to precliminal cancer; level IV is defined as conventional stage 0 cancer (G0), and level V is defined as conventional stages 1-4 cancer (G1-G4).

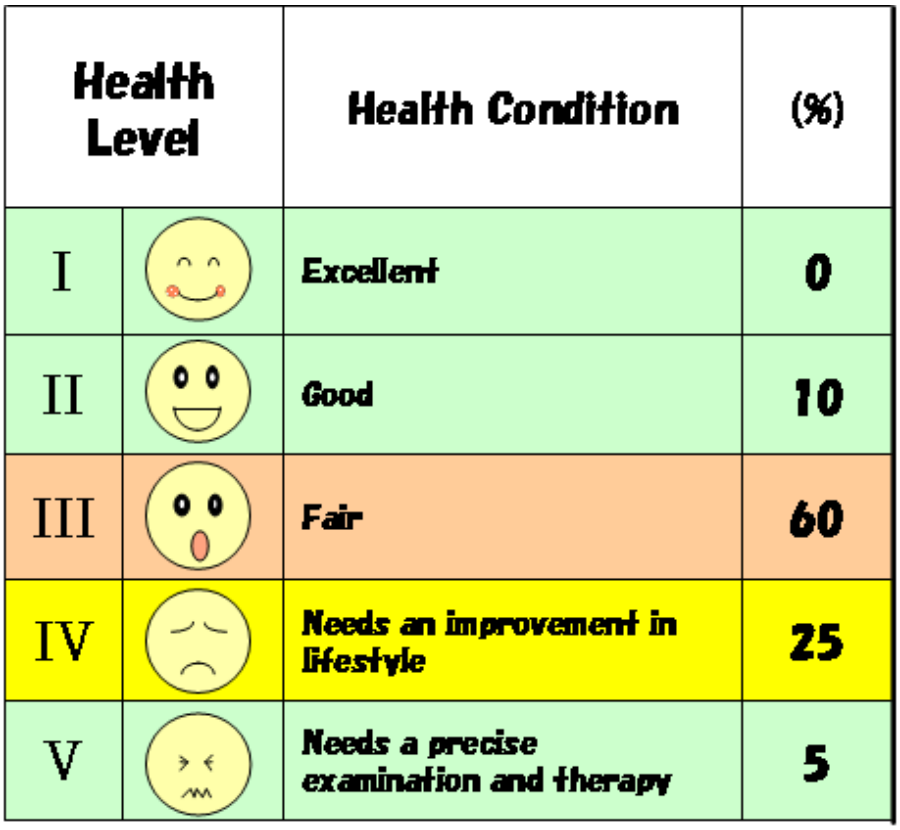

Figure 2. Health levels

Table 1. Health levels.

health level	1	2	3	4	5
ratio (%)	0.0	0.0	12.0	58.7	29.3

3.2 Features of the clinical data

The percentage of persons in each health level is shown in Table 1. The pattern of Table 1 is quite different from that of Fig. 2 and the data shown by (Kobayashi & Kawakubo, 1994). In Table 1, the percent of persons in level II is quite low and that of persons in level IV is quite high. Thus in Table 1, the health level distribution pattern seems to shift to higher levels in parallel. This tendency is similar to that seen in the data of 1000 persons. The distribution of age is shown in Fig. 3. The percentage of persons in their 50's is very high which might be the reason for the different distribution of health levels. Currently, we collect data from office workers (aged 40 to 50 years old) but not from students. Therefore, our results might not reflect social tendencies across all age groups.

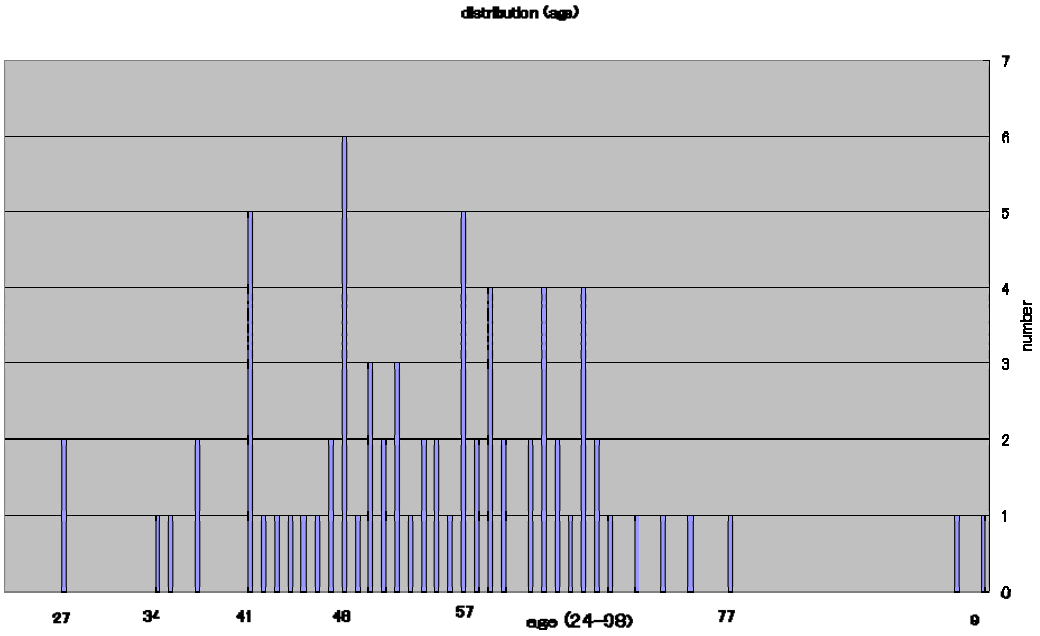

Figure 3. Distribution of age (Clinical data)

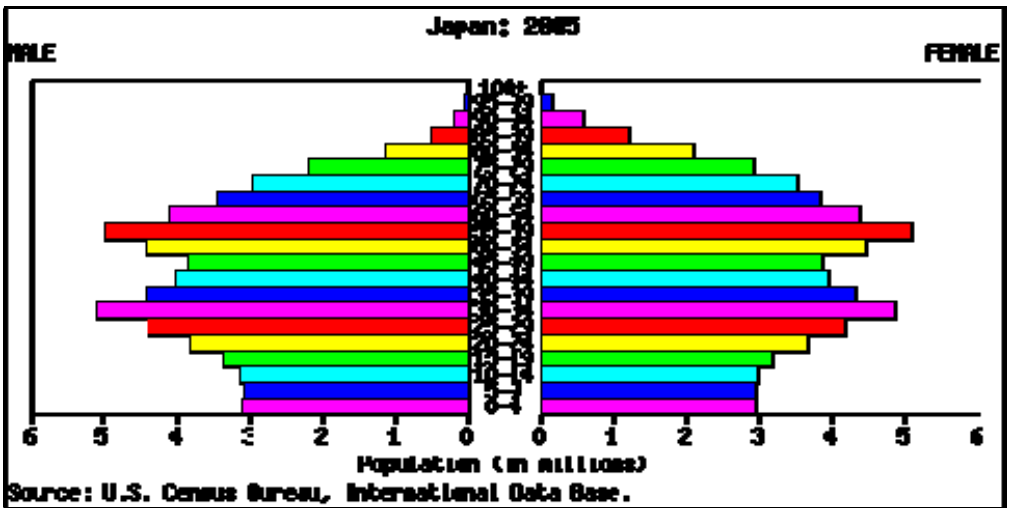

Figure 4. Population Pyramids for Japan (2005)

The age distribution pattern is not the same as the Japanese standard age distribution pattern (Fig. 4). This is because we did not collect samples from persons from all over Japan but only from those who applied to take part in our study. Most of our subjects are office workers and live in or close to Tokyo. Figure 5 shows the distribution pattern of age according to health. For health levels IV and V, the peak is in the 50's. Although the data are not sufficiently large, the distribution pattern seems to be a normal distribution. Therefore, taking account of the pattern shown in Fig. 3, if we were to collect more clinical data from persons in their teen or twenties, the distribution pattern might be different.

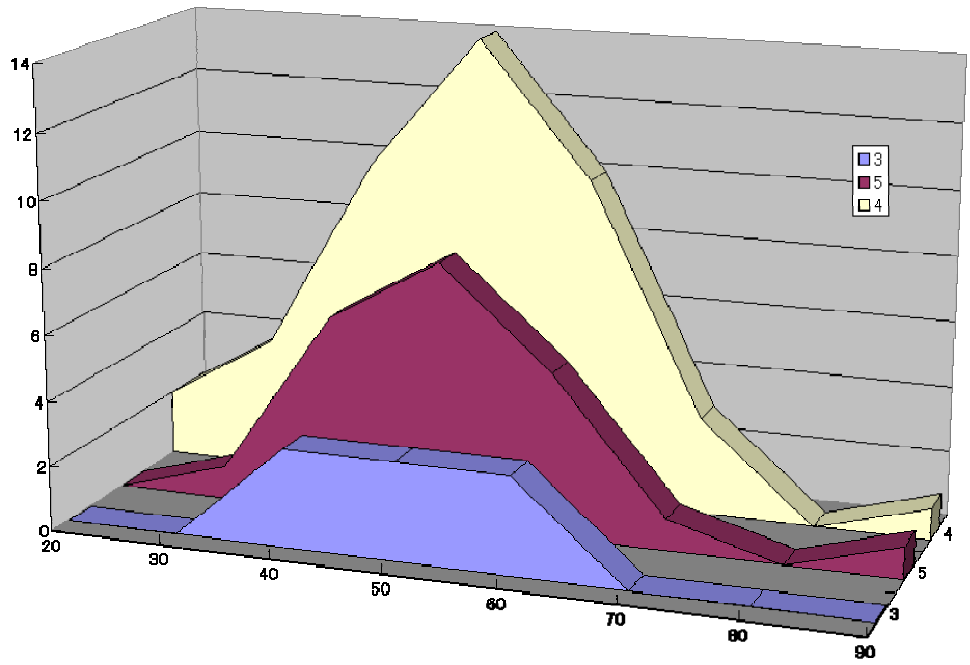

Figure 5. Distribution of age according to health levels.

3.3 Other data

Although not fully dealt with in this study, the data set also includes information from a clinical interview. For instance, worrying symptoms, nonessential foods consumed (coffee, alcohol etc.), medicine, length of exercise, meal style, and family history were included in the interview. Of course, analysis of the clinical data shown above, together with these interview data, is very important, as the clinical data includes information about a patient's physical condition which results from his or her lifestyle and family history. The interview data set is a record of a patient's daily life and family history. Thus the clinical and interview data sets help explain each other. It is quite significant to analyze both data sets as results and reasons, but in this study, we focused on the relationships between a person's clinical data and health status.

4 ANALYSIS OF CLINICAL DATA

In this section, we analyzed results of the relationships between clinical data and health levels. For the analysis, we applied C4.5 (Quinlan, 1993), which is a decision tree learner. The generated decision tree can be regarded as logical formulae representing the relationships between clinical data and health levels. We apply C4.5 because it generates logical formulae and does not require any background knowledge. Because we try to discover new relationships, it is

rather difficult to prepare background knowledge, and sometimes-improper background knowledge leads to incorrect results.

4.1 Relationship between health levels and clinical data

First, we analyzed all the data described in 3.1 without any modification. That is, all data are equally treated in the analysis. Parts of the results are shown below.

```
β2-microglobulin (mg/l) > 1.8 : 5
β2-microglobulin (mg/l) <= 1.8 :
|   γ-GTP (U/l) > 119 : 5
|   γ-GTP (U/l) <= 119 :
|   |   Creatinine (mg/dl) > 1 : 3
|   |   Creatinine (mg/dl) <= 1 :
|   |   |   γ-seminoprotein (ng/ml) <= 0.8 :...
```

The results are almost acceptable but have been generated by analyzing large numbers of items from various categories. Therefore, certain items might be too influential and hide the effects of less influential ones. As shown below, we then analyzed relationships between health levels and data in each category.

```
1) liver, pancreas, and kidney test data
Cholinesterase (U/l) <= 4811 :
|   Creatinine (mg/dl) > 0.9 : 3
|   Creatinine (mg/dl) <= 0.9 :
|   |   TP (g/dl) > 6.9 : 4
|   |   TP (g/dl) <= 6.9 :...

2) metabolic function test data
Total acid phosphatase <= 9.5 :
|   Non-esterified fatty acid (mEq/l) <= 0.3 :
|   |   Fe (μg/dl) <= 69 : 4
|   |   Fe (μg/dl) > 69 : 3 ...

3) general urine test data
Urine acetone > 0 : 4
Urine acetone <= 0 :
|   Urine sedimentary test, squamous epithelium > 1 : 4
|   Urine sedimentary test, squamous epithelium <= 1 :
|   |   Urine sediment-bacteria <= 2 :
|   |   |   Urine sedimentary test, squamous epithelium <= 0 :...

4) blood and immunity test data
C3 (mg/dl) <= 105 :
|   Cellular immunity (T CELL CD2) (%) <= 84 :
```

```
|   |  Cellular immunity (T CELL CD2) (%) > 76 : 4
|   |  Cellular immunity (T CELL CD2) (%) <= 76 :
|   |  |  Leukocyte classification Mono (%) <= 5.8 : 4
|   |  |  Leukocyte classification Mono (%) > 5.8 : 3...
```

5) *tumor markers*

```
β2microglobulin (mg/l) > 1.8 : 5
β2microglobulin (mg/l) <= 1.8 :
|   Carcinoembryonic antigen (ng/ml) <= 4.1 :
|   |  CA72-4 (U/ml) > 3 : 5
|   |  CA72-4 (U/ml) <= 3 :...
```

The above results represent relationships between the data in each category and health levels. The first classification of the analysis of whole data and that of tumor markers is the same. This means that tumor markers obviously influence the classification results. To obtain influential power relationships of each category, we analyzed relationships between health levels and mixed category data. Although we analyzed all the possible combinations, only typical relationships are shown below.

✧ *liver, pancreas, and kidney test data+metabolic function test data*

```
Cholinesterase (U/l) <= 4811 :
|   Creatinine (mg/dl) > 0.9 : 3
|   Creatinine (mg/dl) <= 0.9 :
|   |  TP (g/dl) > 6.9 : 4
|   |  TP (g/dl) <= 6.9 :...
```

✧ *liver, pancreas, and kidney test data+general urine test data*

```
Urine acetone > 0 : 4
Urine acetone <= 0 :
|   Urine sedimentary test, squamous epithelium > 1 : 4
|   Urine sedimentary test, squamous epithelium <= 1 :
|   |  Urine sediment-bacteria <= 2 :
|   |  |  Urine sedimentary test, squamous epithelium <= 0 :....
```

✧ *liver, pancreas, and kidney test data+blood and immunity test data*

```
C3 (mg/dl) <= 105 :
|   Cellular immunity(T CELL CD2) (%) <= 84 :
|   |  Cellular immunity(T CELL CD2) (%) > 76 : 4
|   |  Cellular immunity(T CELL CD2) (%) <= 76 :
|   |  |  Leukocyte classification Mono (%) <= 5.8 : 4
|   |  |  Leukocyte classification Mono (%) > 5.8 : 3 ...
```

✧ *liver, pancreas, and kidney test data+tumor markers*

```
β2microglobulin (mg/l) > 1.8 : 5
β2microglobulin (mg/l) <= 1.8 :
```

```
|   γ-GTP (U/l) > 119 : 5
|   γ-GTP (U/l) <= 119 :
|   |   Creatinine (mg/dl) > 1 : 3
|   |   Creatinine (mg/dl) <= 1 :....
```

✧ *metabolic function test data+general urine test data*

```
Total acid phosphatase <= 9.5 :
|   Urine sediment-bacteria <= 2 :
|   |   Urine sediment-protein quantitative > 0 : 4
|   |   Urine sediment-protein quantitative <= 0 :
|   |   |   Non-esterified fatty acid (mEq/l) > 0.5 : 4
|   |   |   Non-esterified fatty acid (mEq/l) <= 0.5 :....
```

✧ *blood and immunity test data+tumor markers*

```
β2microglobulin (mg/l) > 1.8 : 5
β2microglobulin (mg/l) <= 1.8 :
|   Carcinoembryonic antigen (ng/ml) <= 4.1 :
|   |   CA72-4 (U/ml) > 3 : 5
|   |   CA72-4 (U/ml) <= 3 :
|   |   |   Erythrocyte counts (×10^6/μl) > 526 : 5
|   |   |   Erythrocyte counts (×10^6/μl) <= 526 :....
```

The number of items in each category is quite different (from 11 to 36). This imbalance might influence the results. In addition, we only dealt with a small number of data as a pre-examination. It would be hazardous to determine relationships among the categories by analyzing the results of pre-examinations. However, we can determine a simple relationship, such as the influence of the category on health levels, by comparing the root of decision trees, as results from the most influential factors usually come at or near the root of decision trees. Thus we found the following influential order of the health levels:

$$\textit{metabolic function test data} \prec \textit{liver, pancreas, and kidney test data} \prec \textit{general urine test data} < \textit{blood and immunity test data} \prec \textit{tumor markers}$$

Health levels are assigned according to the possibility of the presence of disease, for instance, cancer. Therefore, it would be reasonable that a tumor marker is the most influential factor. In addition, the diagnosis (assignment of health level) is performed for those who are not believed to be suffering from cancer. As a result, factors such as internal organs play a less influential role in health levels. That is, the internal organs would not be badly damaged.

As for the decision trees of factors other than tumor markers, health level 5 cannot be observed, and the classification points occur within normal values. In addition, the concept of health levels has been introduced to pinpoint the period before a patient's condition becomes a disease during which the patient is moving toward disease (presymptomatic stage).

4.2 Relationship between health levels and interview data

For the preview experiment, we added the interview data to the clinical data to analyze relationships among health levels and clinical data and interview data. The interview data contains various types of information such as family history, diet, and lifestyle data. Currently, it is neither well formalized nor coherent, so it is difficult to analyze without any modification. In fact, if we use the data without proper modification, we only obtain meaningless results. Some of the results obtained after applying data cleaning technique (removing data that causes meaningless results) are shown below:

```
type(alcohol) = sour: 5
type(alcohol) = beer: 4
type(alcohol) = wine: 4
type(alcohol) = sake: 4
type(alcohol) = sake, shochu: 4
type(alcohol) = 0:
|   Diabetes(mother) = 1: 5
|   Diabetes(mother) = 0:
|   |   dinner = 0: 4
|   |   dinner = 1:
|   |   |   start_age(health food(else(1))) <= 30 : 4
|   |   |   start_age(health food(else(1))) > 30 : 5 ....
```

A mother's diabetes might affect her children's health, and alcohol intake might affect health. However, it is difficult to evaluate the results. Also the values in interview data are mostly discrete. In contrast, values from the clinical data are mostly continuous. When analyzing by C4.5, discrete values can easily be classified. Thus, most of the results from clinical data disappear when we add the interview data. It is rather difficult, therefore, to determine whether the clinical data has less influence than interview data on health levels, and further studies are needed.

5 TOWARD INTEGRATED DATA MINING

In the previous section, we presented initial results from analyzing clinical data according to categories such as tumor markers. In addition, we analyzed the influence of these categories on health levels. The results suggest that it would be better to analyze the clinical data according to categories or by considering the influence of the powerful categories. To classify the clinical data, we applied Principal Component Analysis, but we could not find any significant classifications. Thus, it is rather difficult to determine automatically meaningful classification of the clinical data. However, as mentioned before, the influence on the health levels differs according to the categories. In addition, we discussed an influence power order in the previous section. Agrawal proposed an association rule that represents relationships between items in databases (Agrawal et al., 1993). The association rule is frequently used when analyzing Point of Service (POS) data to discover tendencies of users' shopping patterns (basket analysis). However, from the analysis, we can only discover frequently co-occurring patterns. Also, relational data mining has recently been proposed (Džroski & Lavrač, 2001). This determine an effective paradigm also discovers relationships between items in a (relational) database by using ILP techniques. Their approaches are important for complex data mining. However our major aim is not to discover relationships between each category but to classification for data mining while considering influential power. By conducting pre-experiments, we found the following needs for (integrated) data mining of clinical data sets.

- ✓ We need to classify the data not according to statistical or associated patterns but rather according to their influence on health levels. Currently we do not have a clear answer to the problem, but by applying partial data mining more than once and comparing the results, we might be able to solve the problem.
- ✓ After classification, we could determine the influence of category on health levels. By removing more powerful influential categories, we could then find hidden, potential, rare, or novel relationships between the clinical data and health levels. Therefore, we need to conduct multiple partial data mining.
- ✓ In addition, we need to integrate the results from partial data mining. This integration will enable us to discover complex relationships between the clinical data and health levels. Similarly, we can combine the analysis of other types of data, such as interview data and that from Eastern medicine (e.g. Ryodouraku). Then relationships between Western and Eastern medicine can also be discovered by observing the results obtained by persons from various fields.

6 CONCLUSIONS

In this paper, we present initial results from the analysis of relationships between clinical data and health levels. In fact, we used given (authorized) categorizations such as blood test data and tumor markers. We found that the most influential category for the health level is tumor markers. As a result, when the other categories are mixed with tumor markers, the tumor markers might hide relationships among items in the other categories. We, therefore, suggest integrated data mining that categorizes the clinical data into multiple categories to discover relationships among the items in each category and the health levels and integrates the results. We can, of course, add data from Eastern medicine. In addition, integrated data mining can be applied to chance discovery in discovering rare or novel events that might cause serious disease, as powerfully influential factors that might disturb the influence of powerless factors can be removed by the categorization. We currently have data sets from more than 1000 persons, and additional data are still being collected from some of these persons and from new subjects. Therefore, we will be able in the future to conduct temporal data mining. We think the discovery of temporal patterns in the health levels is very important in protecting the subjects from developing diseases. In addition to integrated data mining, we should take account of temporal data mining to discover a temporal pattern to disease.

7 ACKNOWLEDGEMENTS

This research was supported in part by the Program for Promoting the Establishment of Strategic Research Centers, Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sports, Science and Technology (Japan).

8 REFERENCES

Abe, A., Kogure, K. & Hagita, N. (2003) Discovery of Hidden Relations from Medical Data. *Proc. of HCI2003 3rd. Int'l Workshop on Chance Discovery,* pp. 37-43.

Agrawal, R., Imielinski, T., & Swami, A. (1993) Mining association rules between sets of items in large databases. *Proc. of ACM SIGMOD Int'l Conf. on Management of Data*, pp. 207-216.

Džroski, S. & Lavrač, N. eds. (2001) *Relational Data Mining*, Springer Verlag

Ichise, R. & Numao, M. (2005) First-Order Rule Mining by Using Graphs Created from *Temporal Medical Data. LNAI, Vol. 3430*, pp. 112-25.

Kobayashi, T. & Kawakubo, T. (1994) Prospective Investigation of Tumor Markers and Risk Assessment in Early Cancer Screening. *Cancer, Vol. 73, No. 7*, pp. 1946-1953.

Ohsawa, Y., Okazaki, N., & Matsumura, N. (2003) A Scenario Development on Hepatics B and C. *Technical Report of JSAI, SIG-KBS-A301*, pp. 177-182.

Osawa, Y. & McBurney, P. eds. (2003) *Chance Discovery*, Springer Verlag

Tsumoto, S. (2004) Mining Diagnostic Rules from Clinical Databases Using Rough Sets and Medical Diagnostic Model. *Information Sciences, Vol. 162*, No. 2, pp. 65-80.

Tsumoto S., Yamaguchi T., Numao M., & Motoda H. (2005) Active Mining Project: Overview. *LNAI, Vol. 3430*, pp. 1-10.

Quinlan, J. R. (1993) *C4.5 Programs for Machine Learning*, Morgan Kaufman