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

2) ATR Knowledge Science Laboratories

3) ATR Intelligent Robotics and Communication Laboratories

Medical science as well as clinical diagnosis and treatment has *progressed rapidly* in recent years with each field becoming more specialized and independent.

⇓

An integrated and cooperative approach to research between medical researchers and biologists is needed.

⇓

Cyber Integrated Medical Infrastructure (CIMI)

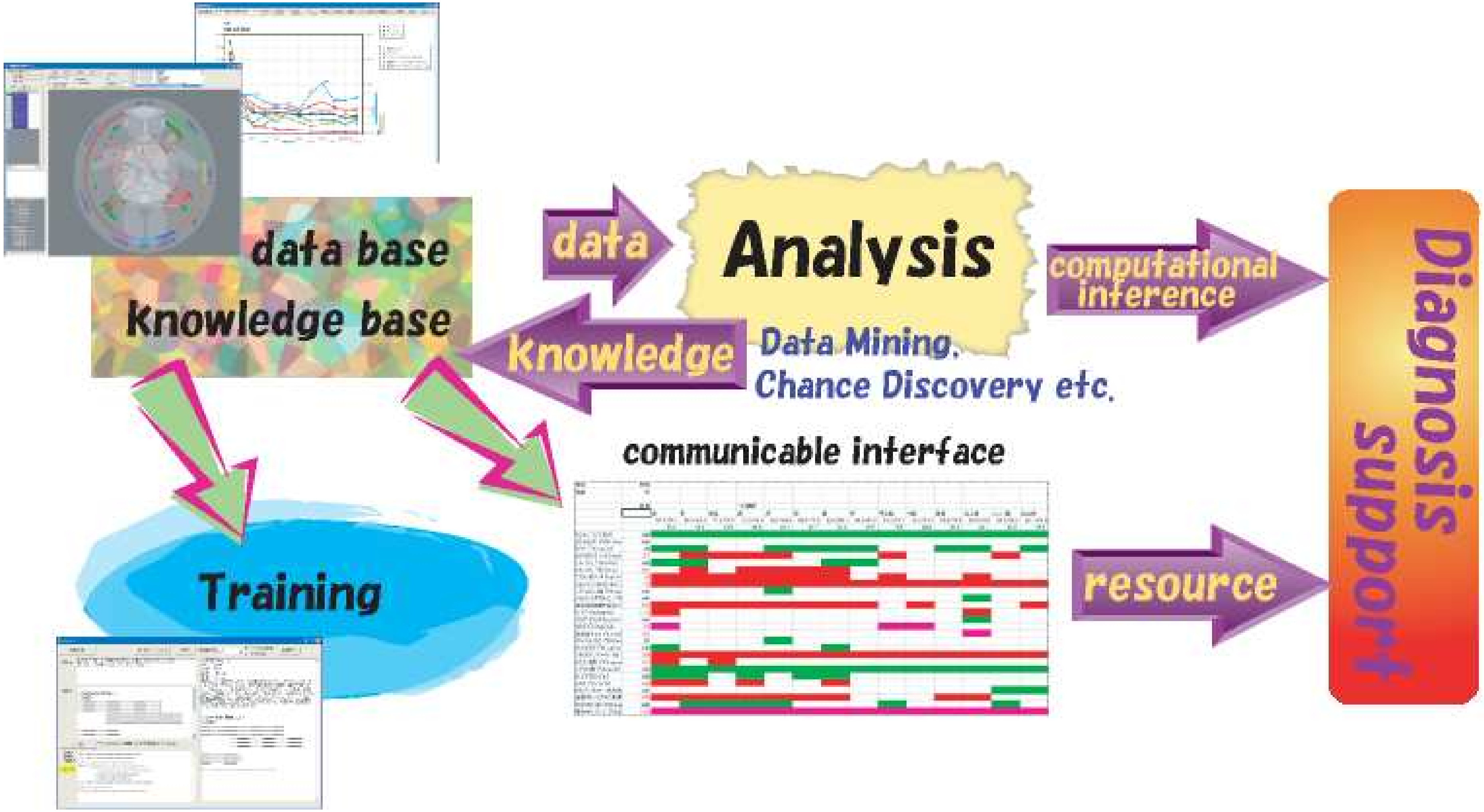
A framework of integrated management of clinical data on computer networks consisting of a database, a knowledge base, and an inference and learning component, which are connected to each other in the network.

Medical information (e.g. clinical data) ⇒

A knowledge base ⇒ *final goal*

Predicting all possible diseases and to support medical diagnosis

Cyber Integrated Medical Infrastructure (CIMI)



CIMI framework: Medical, clinical and other information (e.g. personal information, interview) is analyzed or data mined to discover relationships between the medical, clinical and other data and all possible diseases.

Feature of clinical data

1) **liver, pancreas, and kidney test data:** 24 items

   Total protein, albumin, serum protein fraction-α1-globulin

2) **metabolic function test data:** 29 items

   Na, K, Ferritin, total acid phosphatase

3) **general urine test data:** 11 items

   Urobilinogen, urine acetone

4) **blood and immunity test data:** 31 items

   Mycoplasma pneumoniae antibody, cellular immunity

5) **tumor markers:** 36 items

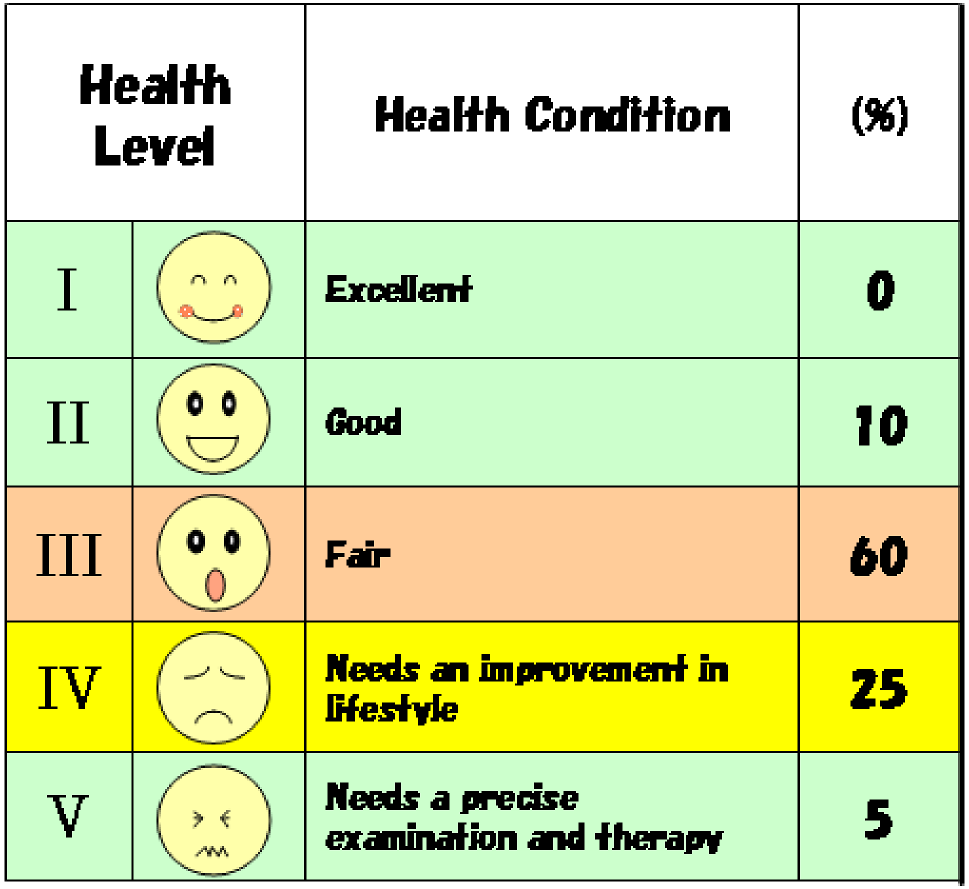
Immunosuppressive acidic protein, Sialyl Le X-i antigen, urine β2-microglobulin

else) **clinical interviews, family tree, and lifestyle...**

worrying symptoms, nonessential foods (coffee, alcohol etc.), medicine, length of exercise, meal style, and family history etc.

Health levels (manually assigned)

express the health status of patients are defined based on *Tumor stage* [Kobayashi 1994] and modified by Matsuoka.



Health Level		Health Condition	(%)
I		Excellent	0
II		Good	10
III		Fair	60
IV		Needs an improvement in lifestyle	25
V		Needs a precise examination and therapy	5

- level I & II: can be regarded as being healthy
- level III: defined as the stage before the shift to precliminal cancer
- level IV: defined as conventional stage 0 cancer (G0)
- level V: defined as conventional stages 1–4 cancer (G1–G4)

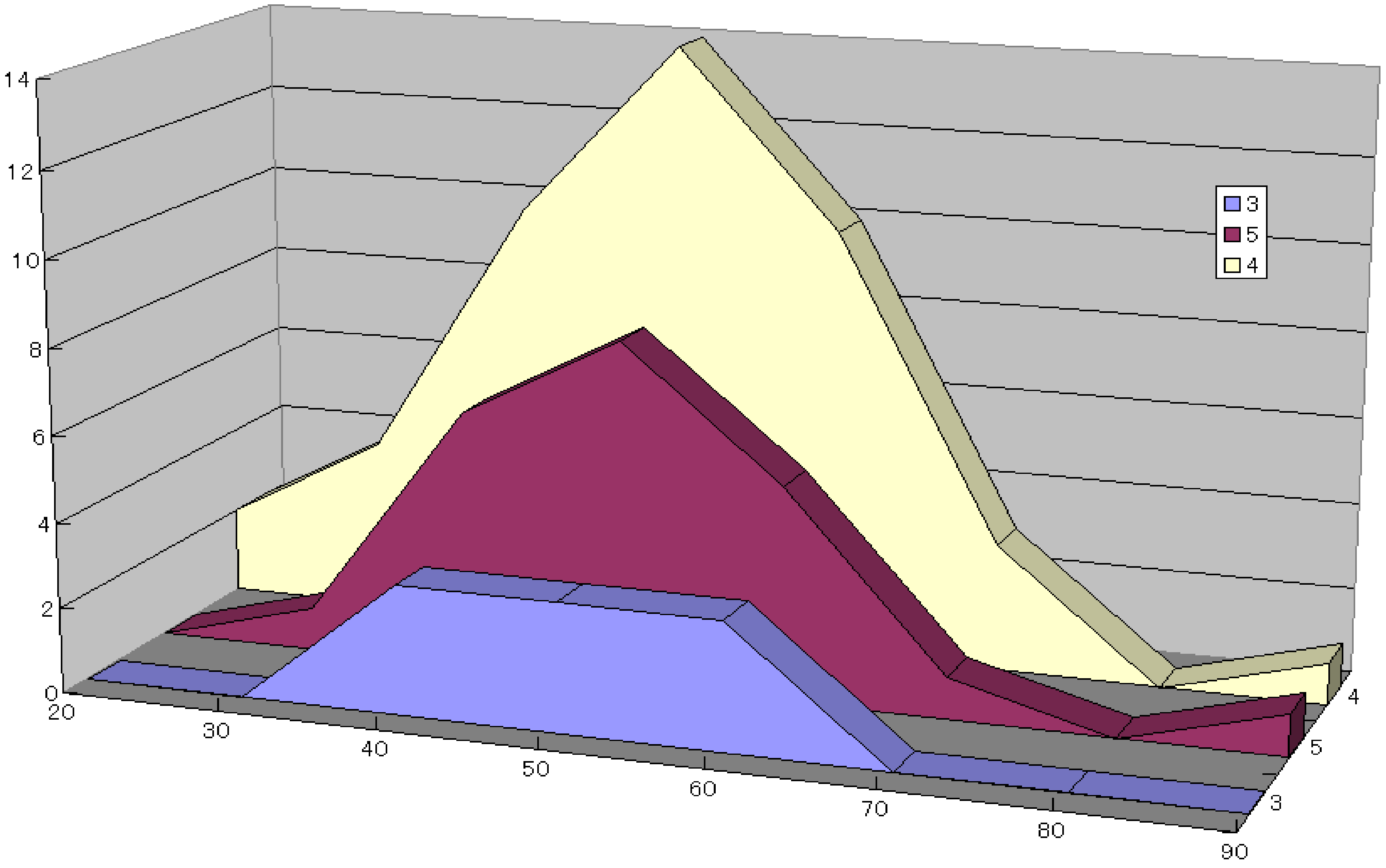
Data: 77 persons

Ratio of persons in each health level

health level	1	2	3	4	5
ratio (%)	0.0	0.0	12.0	58.7	29.3
standard distribution					
ratio (%)	0	10	60	25	5

shift to level IV

Distribution of age according to health levels



Most of our subjects are office workers and live in or close to Tokyo.

For health levels IV and V, the peak is in the 50's.

Persons in the health level III are not so many.

If we would collect more clinical data from persons in their teen or twenties, the distribution pattern might be different (close to standard distribution).

data analyses (C4.5)

whole data

```
β 2-microglobulin (mg/l) > 1.8 : 5
β 2-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....

$\Downarrow$

analysis of 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 (μ g/dl) <= 69 : 4
|   |   Fe (μ g/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

```
β 2microglobulin (mg/l) > 1.8 : 5
β 2microglobulin (mg/l) <= 1.8 :
|   Carcinoembryonic antigen (ng/ml) <= 4.1 :
|   |   CA72-4 (U/ml) > 3 : 5
|   |   CA72-4 (U/ml) <= 3 :...
```

The first classification of the analysis of whole data and that of tumor markers is the same.

⇒ tumor markers obviously influence the classification results.

analysis of relationships between health levels and mixed category data

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 :....
```

Whole

```
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 :...
```

liver, pancreas, and Kidney test

```
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 :...
```

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 :...
```

liver, pancreas, and Kidney test+general urine test

```
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 :....
```

influential order of the health levels

metabolic function test data $\prec$

liver, pancreas, and kidney test data $\prec$

general urine test data $\prec$

blood and immunity test data $\prec$ *tumor markers*

Health levels are assigned according to the possibility of the presence of disease, for instance, cancer.

$\Rightarrow$ "*tumor markers* is the most influential factor" is reasonable

The diagnosis (assignment of health level) is performed for those who are not believed to be suffering from cancer.

$\Rightarrow$ Factors such as internal organs play a less influential role in health levels.

Toward integrated data mining

- Need to classify the data not according to statistical or associated patterns but rather according to their influence on health levels.

  By applying partial data mining more than once and comparing the results?

- After classification, the influence of category on health levels can be determined.

  By removing more powerful influential categories, we could then find hidden, potential, rare or novel relationships between the clinical data and health levels.

- Need to integrate the results from partial data mining ⇒ complex relationships between the clinical data and health levels.

  Analysis of other types data such as interview data and that from Eastern medicine (e.g. Ryodouraku) can be combined