CASE 78

Forecasting Future Health from Existing Medical Examination Results Using the MTS

Abstract: For our research, we used medical checkup data stored for three years by a certain company. Among 5000 examinees, there were several hundred sets of data with no missing measurements. In this case, since the determination of a threshold between A and B was quite difficult, a different doctor sometimes classified a certain examinee into a different category, A or B. It was the aim of this study to improve the certainty of such judgment.

1. Introduction

The medical checkup items can be classified roughly into the following 19:

1. Diseases under medical treatment (maximum of three types for each patient)
2. Diseases experienced in the past (maximum of three types for each patient)
3. Diseases that family members had or have (categorized by grandparents, parents, and brothers)
4. Preferences (cigarette, alcohol)
5. Subjective symptoms
6. Time since meals
7. Physical measurements (height, weight, degree of obesity)
8. Eyesight
9. Blood pressure
10. Urinalysis
11. Symptoms of other senses
12. Hearing ability
13. Chest x-ray
14. Electrocardiogram
15. Blood (19 items)
16. Photo of eyeground
17. Additional blood test items
18. Gender
19. Age

A doctor's overall judgment should be added to a medical diagnosis, but there is difficulty when there are too many examinees. In this study, a doctor analyzes a medical checkup list (blood work, electrocardiogram, chest x-ray, eyesight, medical checkup interview, blood pressure, etc.), then categorizes an examinee in one of four categories:

A: normal

B: observation needed

C: treatment needed or under treatment

D: detailed examination needed

In this case, since the determination of a threshold between A and B is quite difficult, a different doctor sometimes classifies a certain examinee into a different category, A or B. It was our aim in this study to improve the certainty of such judgment.

Characteristics such as gender are quantified into numerals as a category data, such as "male = 1" or "female = 2." Age or biochemical data are used as they are. Items with the same value among all data are excluded (e.g., all of the examinees have "1" or "normal" for hearing ability because they are

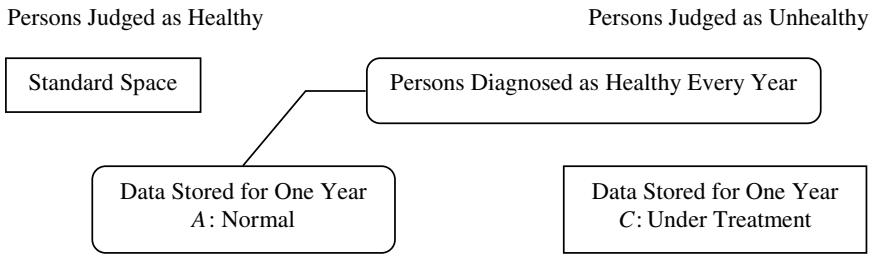

Figure 1
Base space I for one-year-healthy persons

healthy). In addition, items that lead to multicollinearity are also omitted. Eventually, the original nearly 100 items were reduced to 66.

2. Preparation of the Base Space

The following base spaces were prepared to observe the judgment being affected by different base space preparation.

Base Space I

Primarily, we investigated whether a medical checkup judgment could be made using only the one-year healthy persons' data (685 persons diagnosed as A and 735 as C). According to Figure 1, we formed base space I with the data for 685 persons judged as A and studied whether the data could be discriminated from those for 735 under-treatment patients identified as C. Furthermore, to improve discriminability, we selected examination items.

Base Space II

With the data for two-year-healthy persons, we looked for further improvement in accuracy of discriminability. According to Figure 2, we created base space II with the data for two-year-healthy 159 per-

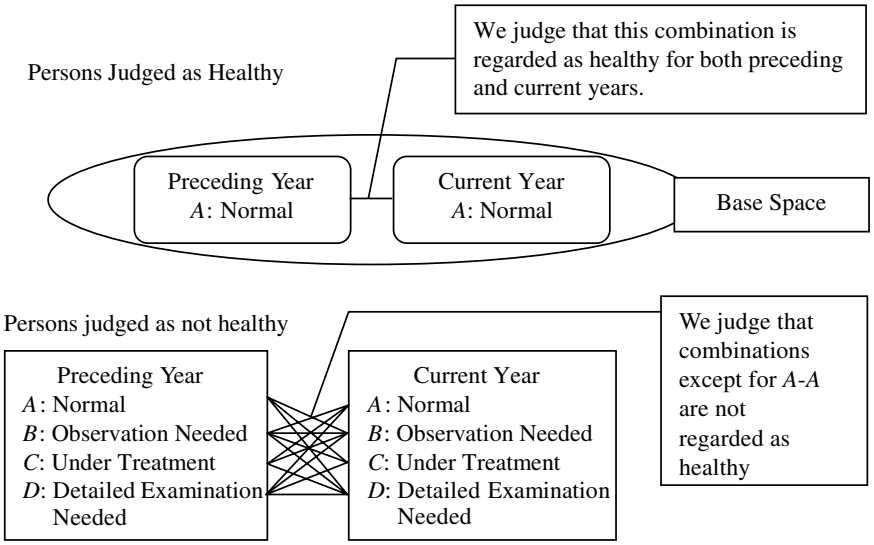

Figure 2
Base space II for two-year-healthy persons

sons diagnosed as *A* and calculated the distances for 159 persons diagnosed as *A* and the ones for 37 persons as *C*. Since the result within one year was treated as one data point, from a two-year-healthy person diagnosed as *A*, we obtained two data points. Since missing data leads to lower accuracy in judgment, we did not use situations with missing data.

Base Space III

After studying discriminability using the Mahalanobis–Taguchi system, we investigated predictability, which characterizes a Mahalanobis distance. That is, we studied the possibilities not only of judging whether a certain person is currently healthy but also foreseeing whether he or she will be healthy in the next year based on the current medical checkup data.

First, combining the two-year medical checkup data, we created base space III based on Figure 3. The persons who were diagnosed as category *A* this year were defined as healthy persons. Base space III was prepared from the preceding year's data for these healthy persons. In other words, a person categorized as healthy for the current year could be healthy or unhealthy the preceding year. Since the medical checkup data for the preceding year were used, only one-year data are available, although two-year data (for the preceding and current years) are available. No matter what category they belonged to in the preceding year, those who are judged as unhealthy are viewed as *C* in the current year. Through this analysis it is possible to predict what type of base space should be used to predict a healthy person in the following year.

Base Space IV

From the perspective of reliability in prediction, prediction with three-year-long data is more accurate than that with one- or two-year data. Whereas one- or two-year data do not reflect time-series elements, with three-year data, we studied the time-lapsed change for medical checkup data. As a matter of course, the number of data covering all measurements in three-consecutive-year medical checkups is rather limited. We formed base space IV with data only for the preceding two years from 120

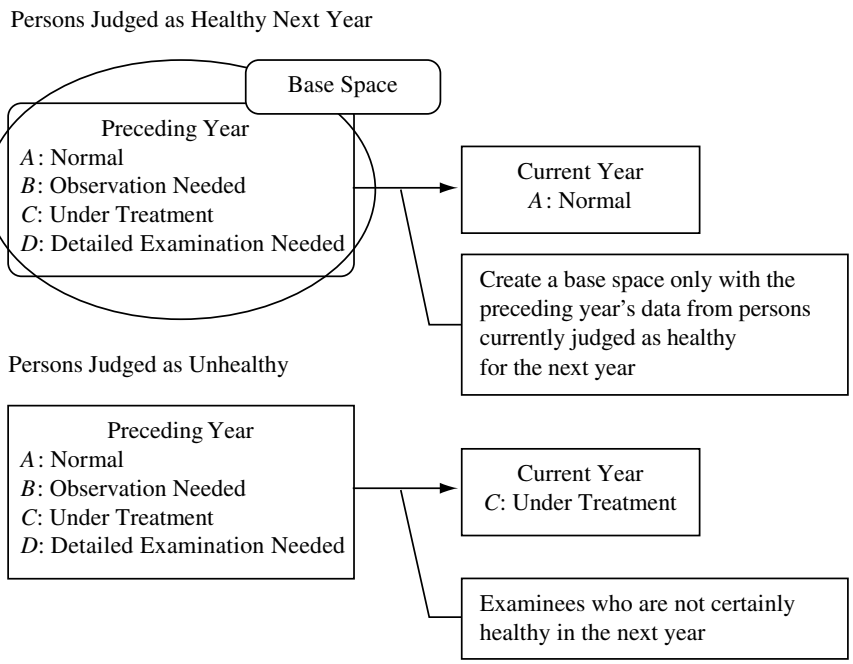

Figure 3
Base space III: predicting current year's health condition using two-year data

patients. The number of medical checkup items was approximately 200 as a two-year time-based item. However, because the number of datapoints was quite small, considering the stability of the base space, we reduced it to 101. Base space IV was prepared according to Figure 4. We also studied base space IV after item selection. If the data were judged as *A* for the current year, it does not matter what category they are diagnosed before. In addition, those who were diagnosed as unhealthy for the current year were regarded as *C*-diagnosed examinees, no matter what data they had for the preceding two years.

3. Selection of Examination Items

After having predicted and judging a health condition by using a Mahalanobis distance, we distinguished between necessary and unnecessary items for prediction and judgment based on the base space to reduce measurement cost. That is, without lowering predictability and discriminability grounded on a Mahalanobis distance, we selected essential examination items.

To design parameters of all items, we set "use of an item for creating a standard space" to level 1 and "no use of an item for creating a standard space" to level 2. Because 100 items were used for our medical checkup, an L_{128} orthogonal array was selected.

It is all right to leave some columns empty. Since the first row has only one item allocated, we created a base space using all items. For the second row, about half the row is assigned a 1, so we formed a base space using approximately half of all the items. Similarly for the remaining rows, we created base spaces, and finally, obtained 128 different spaces.

As a next step, for the 128 base spaces created thus far, we computed a distance for each of the

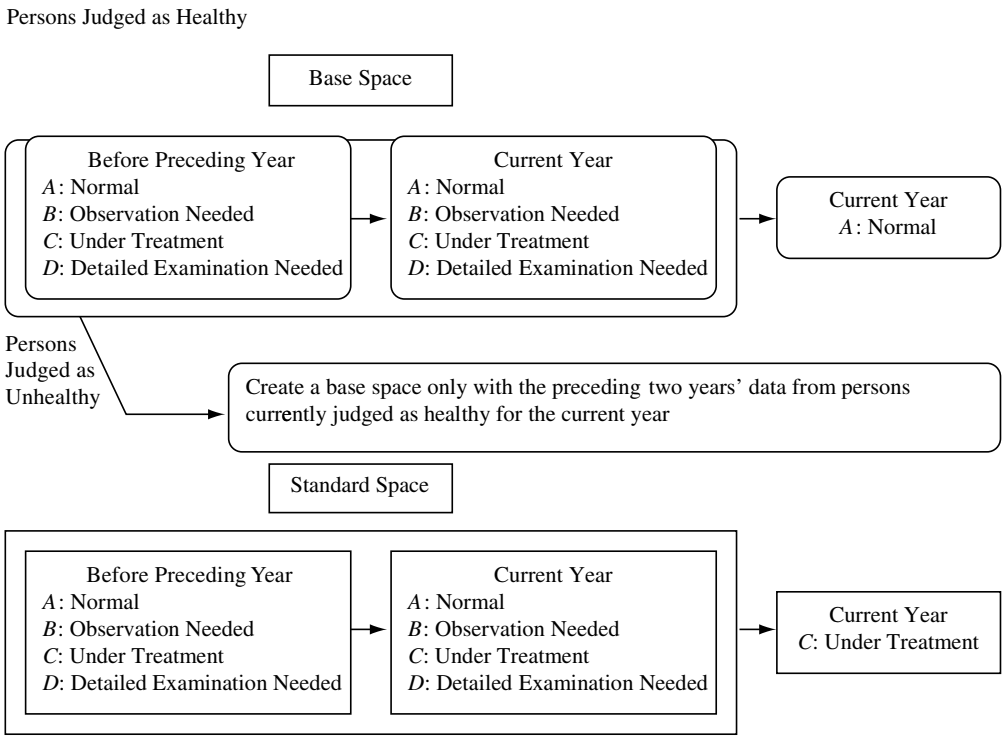

Figure 4
Base space IV: predicting current year's health condition with three-year-healthy persons' data

data that were expected to be distant (abnormal data). The greater the number of abnormal data, the more desirable. However, under some circumstances, we cannot collect a sufficient number of abnormal data. In this case, even a few data are regarded as applicable. We defined the distance for each abnormal data item, which was calculated from the base space for the first row, as D_1, D_2, ... , D_n. Now, since the abnormal data should be distant from the normal space, we used a larger-the-better SN ratio for evaluation:

$$\eta = -10 \log \frac{1}{n} \left(\sum_{i=1}^{n} \frac{1}{D_i^2} \right)$$

To select necessary items, we calculated the average for each level and created the response graphs where an item whose value decreases from left to right was regarded as essential, whereas one with a contrary trend was considered irrelevant to a judgment.

4. Results of Analysis

The possibility of judging health examinations using base spaces prepared in different ways was studied.

Prediction by Base Space I

Using only the one-year data, we investigated whether or not we could make a medical checkup judgment. The data have no missing data and consist of those for 685 persons judged as healthy (A) and 735 persons judged as unhealthy (C), with 109 items in total. In the hope of improving discriminability, we selected the following necessary items for a judgment by excluding unnecessary ones:

- *Blood test:* GOT, GPT, ALP, γ-GTP, T.Bil, ZTT, TG, T.Chol, HDL, UA, BS, HbA_1c, WBC, RBC, Hb, Ht, MCHC
- *Medical checkup interview:* gender, age, height, diseases experienced, diseases cured
- *Blood pressure:* maximum, minimum
- *Others:* Chest x-ray electrocardiogram

Since the discriminability deteriorates if we iterate item selection over four times, by creating base space I with items that came up with the highest discriminability without reaching a downward trend in discriminability, we computed the distance for the base space and summarized the results in Table 1. This table reveals that there is an overlap area between judgments A and C, which leads to inability of judgment, even though the number of data points is quite large.

Prediction by Base Space II

Second, we investigated whether we could make a judgment using the previous two-year data. According to Figure 2, we created base space II with the 67-item data for 159 persons judged as healthy. Figure 5 and Table 2 show the results of the Mahalanobis distances for 159 persons judged as healthy (A) and 37 persons under treatment (C). These results demonstrate that if a certain examinee has a Mahalanobis distance less than or equal to 2.0 from base space II, he or she is diagnosed as healthy. On the other hand, if the distance becomes greater than 2.0, he or she can be judged as unhealthy or possibly suffering some disease even without a doctor's diagnosis. However, for judgment C discussed here, the one-year data for those who were judged as C in two years were used. For people who were diagnosed as A in the preceding year but as C in the current year, it was difficult to make an accurate

Table 1
Distribution of one-year data's distances from base space I after item selection

Data Class	*A*	*C*	Data Class	*A*	*C*
0	0	0	6	0	11
0.5	17	0	6.5	0	7
1	356	42	7	0	9
1.5	261	121	7.5	0	4
2	49	141	8	0	7
3	1	71	8.5	0	6
3.5	0	44	9	0	4
4	0	21	9.5	0	5
4.5	0	21	10	0	3
5.5	0	17	Next class	0	77

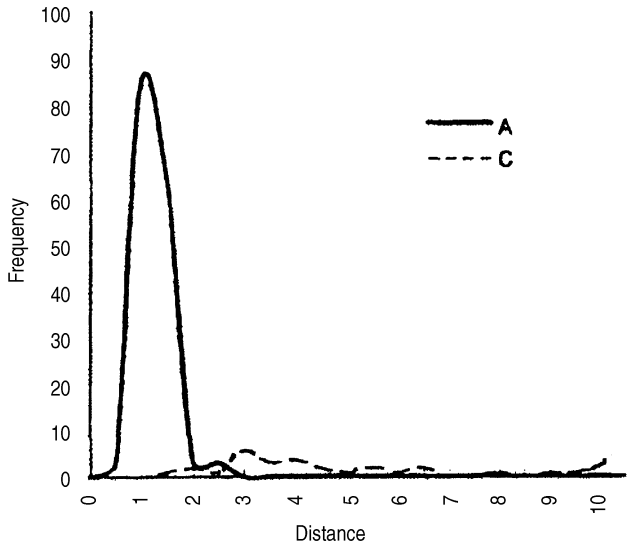

Figure 5
Distances from base space II using two-year data

Table 2
Distribution of distances from base space II for next year using two-year data

Data Class	*A*	*C*	Data Class	*A*	*C*
0	0	0	6	0	1
0.5	3	0	6.5	0	2
1	85	0	7	0	0
1.5	65	1	7.5	0	0
2	3	2	8	0	1
2.5	3	1	8.5	0	0
3	0	6	9	0	1
3.5	0	3	9.5	0	1
4	0	4	10	0	2
4.5	0	2	Next class	0	7
5	0	2			
5.5	0	2			
			Total	159	37

judgment on their health conditions because their data were distributed between judgments *A* and *C*, shown in Figure 2.

Prediction by Base Space III

Using the two-year data, we studied predictability. The data used for base space III were composed of the complete 100-item data in the preceding year for 336 persons judged as healthy (*A*) and in the previous year for 137 persons under treatment (*C*). Table 3 shows the distances from the base space for judgments *A* and *C*. Although we could not clearly separate *A* and *C*, despite being able to make a judgment on whether a certain person was currently in good health or not, when we selected 1.5 as the threshold, the accuracy in predictability was regarded as approximately 80%.

Prediction by Base Space IV

We investigated predictability based on the time-series data. Using the three-year data for 120 persons regarded as healthy, we formed base space IV. While the number of the two-year time-series items as medical checkup items was approximately 200, considering the stability of base space IV, we nar-

Table 3
Distribution of distances from base space III using two-year data for the prediction of next year

Data Class	*A*	*C*	*A*(%)	*C*(%)
0	0	0	100	0
0.5	40	0	100	0
1	140	7	88	5
1.5	112	14	46	15
2	39	24	13	33
2.5	4	12	1	42
3	0	18	0	55
3.5	1	18	0	68
4	0	8	0	74
4.5	0	12	0	82
5	0	7	0	88
5.5	0	6	0	92
6	0	1	0	93
6.5	0	2	0	94
7	0	0	0	94
7.5	0	0	0	94
8	0	0	0	94
8.5	0	1	0	95
9	0	0	0	95
9.5	0	0	0	95
10	0	1	0	96
Next class	0	6	0	100
Total	336	137		

Table 4
Distribution of distances from base space IV from three-year data before item selection

Data Class	*A*	*C*	*A′*	*A* (%)	*C* (%)	*A′* (%)
0	0	0	0	100.0	0.0	100.0
0.5	0	0	0	100.0	0.0	100.0
1	52	0	0	100.0	0.0	100.0
1.5	68	0	0	56.7	0.0	100.0
2	0	0	1	0.0	0.0	100.0
2.5	0	0	0	0.0	0.0	98.8
3	0	1	1	0.0	2.6	98.8
3.5	0	0	7	0.0	2.6	97.6
4	0	0	4	0.0	2.6	89.4
4.5	0	1	4	0.0	5.3	84.7
5	0	1	4	0.0	7.9	80.0
5.5	0	1	7	0.0	10.5	75.3
6	0	1	6	0.0	13.2	67.1
6.5	0	1	4	0.0	15.8	60.0
7	0	1	7	0.0	18.4	55.3
7.5	0	2	1	0.0	23.7	47.1
8	0	0	4	0.0	23.7	45.9
8.5	0	0	4	0.0	23.7	41.2
9	0	2	6	0.0	28.9	36.5
9.5	0	0	4	0.0	28.9	29.4
10	0	0	1	0.0	28.9	24.7
Next class	0	27	20	0.0	100.0	23.5
Total	120	38	85			

rowed them down to 101 because the total number of data points was small. Table 4 shows the results of judgments *A* and *C*. Those below 1.5 were judged as *A*, whereas the ones above 3.0 were seen as *C*. The two classes are separated perfectly. *A′* indicates the case where, after excluding one person's data from the 120 persons' judged as *A* and creating base space IV using only the remaining 119 persons' data, we calculated the distance using the data for this extended person, from base space IV generated from the 119 persons' data. Similarly, we calculated the distances for the 85 persons' data judged as *A* by using such a base space. We can see that some data judged as *A′* deviated from base space IV. The reason is that because of a small number of data in

base space IV, the space becomes so sensitive that a part of the data judged as A' are not separated from those in judgment C. Therefore, through item selection, by eliminating ineffective items for a judgment on abnormal data, we reduced the number of items to 60. Then, using 119 of the 120 normal data, we created base space IV′. The distance of the one data point excluded was computed using the base space formed with the 119 data points. Next, removing another data point from the 120 data, we created base space IV′ with the remaining 119 data once again. And then the distance for the one data item removed was calculated. Iterating this calculation process for all 120 data, we arrived at Figure 6 and Table 5, which reveal that judgment A' becomes closer to judgment A. Therefore, we came to the conclusion that considerable improvement in accuracy was obtained.

5. Selection of Examination Items by a Larger-the-Better SN Ratio

To improve prediction accuracy, through item selection with the two-year data, we excluded items that are not closely related to prediction and formed base space III′ once again. While the number of the data is still 336, the number of the items is cut down from 100 to 66. Based on the data, we created response graphs for item selection. The distances from the base space after item selection are shown in Table 6 and Figure 7.

That is, with the preceding year's data for those judged as A in the current year, we created a base space. Then we computed the distances using this base space for persons judged as A and C. Setting the threshold to 1.5 and classifying all the distances D's by this criterion, we summarize them in Table 7. According to this table, we can see that the proportion of accurate prediction of judgment A results in 90%, whereas the proportion of wrong prediction of judgment C, which are judged as A, is only 15%. As a result, these numbers prove that our predictability was fairly high.

6. Discussion

In our first analysis, we studied base space I generated from the one-year data for persons judged as healthy and created a base space with items that maximized discriminability after item selection. However, despite a large number of data, including those for 685 persons judged as A and 735 as C, an overlap area in the distribution has occurred and caused insufficient discriminability.

In the case where we took advantage of a Mahalanobis distance from base space II, which were formed by the two-year data for persons diagnosed as healthy, judgments A and C can be distinguished.

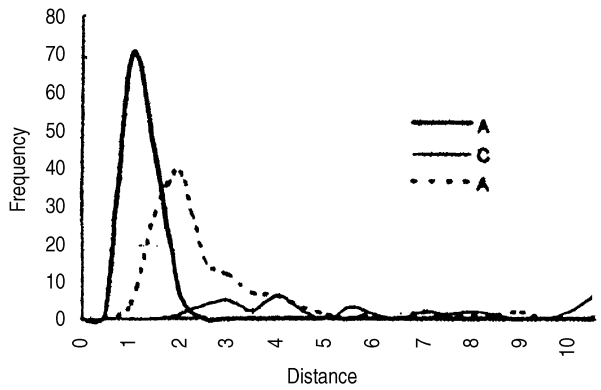

Figure 6
Distances from base space IV′ using three-year data after item selection

Table 5
Distribution of distances from base space IV′ using three-year data after item selection

Data Class	*A*	*C*	*A′*	*A* (%)	*C* (%)	*A′* (%)
0	0	0	0	100.0	0.0	100.0
0.5	1	0	0	100.0	0.0	100.0
1	69	0	4	99.2	0.0	100.0
1.5	44	0	27	41.7	0.0	96.7
2	6	1	38	5.0	2.6	74.2
2.5	0	4	15	0.0	13.2	42.5
3	0	5	12	0.0	26.3	30.0
3.5	0	2	7	0.0	31.6	20.0
4	0	6	6	0.0	47.4	14.2
4.5	0	2	3	0.0	52.6	9.2
5	0	0	1	0.0	52.6	6.7
5.5	0	3	0	0.0	60.5	5.8
6	0	1	1	0.0	63.2	5.8
6.5	0	0	0	0.0	63.2	5.0
7	0	2	2	0.0	68.4	5.0
7.5	0	1	0	0.0	71.1	3.3
8	0	2	1	0.0	76.3	3.3
8.5	0	1	1	0.0	78.9	2.5
9	0	0	2	0.0	78.9	1.7
9.5	0	0	0	0.0	78.9	0.0
10	0	2	0	0.0	84.2	0.0
Next class	0	6	0	0.0	100.0	0.0
Total	120	38	120			

Table 6
Distribution of distances from base space III′ using two-year data after item selection

Data Class	*A*	*C*	*A* (%)	*C* (%)
0	0	0	100	0
0.5	15	0	100	0
1	170	3	88	4
1.5	117	15	45	15
2	31	28	10	35
2.5	3	25	1	53
3	0	17	0	66
3.5	0	12	0	74
4	0	10	0	82
4.5	0	3	0	84
5	0	6	0	88
5.5	0	3	0	91
6	0	2	0	92
6.5	0	0	0	82
7	0	2	0	83
7.5	0	1	0	84
8	0	0	0	94
8.5	0	0	0	94
9	0	0	0	94
9.5	0	1	0	95
10	0	0	0	95
Next class	0	7	0	100
Total	338	137		

If the threshold for a Mahalanobis distance is set to 2.0, we can predict that the data below the threshold will be judged as healthy, whereas those above the threshold will be judged as unhealthy. Consequently, we may make a medical checkup judgment without a doctor's diagnosis. As a next step, using the preceding year's data of those judged as *A* in the current year, we calculated the distances from base space III to classify *A* and *C*. As a result, in the case of choosing 1.5 as the threshold, we have been able to obtain predictability of approximately 80%.

Since we have confirmed that judgment and prediction can be attained by using a Mahalanobis distance to reduce measurement cost without worsening predictability and discriminability, we selected necessary and unnecessary items for creating

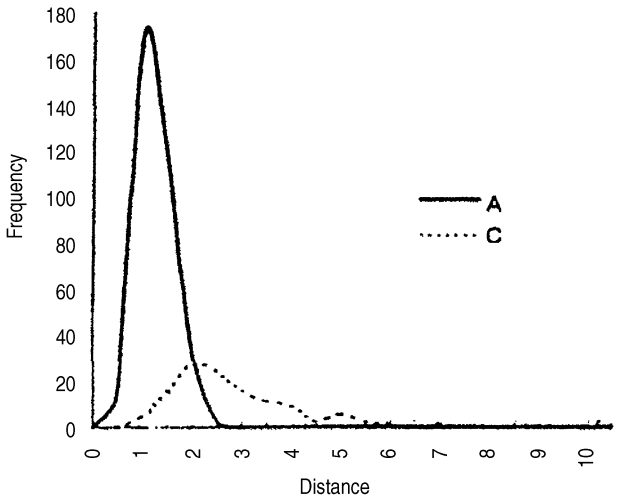

Figure 7
Distances from base space III′ using two-year data after item selection

a standard space. After item selection, utilizing the preceding year's data for persons diagnosed as *A* in the current year, we created base space III′ and calculated the distances for judgments *A* and *C*. In this case, if 1.5 is chosen as the threshold and all of the distances *D*'s are categorized, the proportion of the data judged accurately as *A* amounts to 90%, whereas the proportion of the data diagnosed erroneously as *A* even though they belong to *C* is only 15%. As a result, we have obtained fairly good predictability.

On the other hand, three-year data are obviously more effective than two-year data to improve the accuracy in prediction of health conditions. Thus, we have investigated whether we can raise data reliability by including time-series data in the calculation. However, since there are numerous changes in examination items for a medical checkup at a company because of budgetary issues, job rotations, or employees' ages, the number of data that can cover all medical checkups for three years in a row without missing data is very limited. Although we have created a base space by using the three-year data for 120 persons, eventually, the sum of the items has decreased because the number of data turned out to be small. Therefore, considering the stability of base space IV, we reduced the number of items to 101. As a result, we have been able to separate judgments *A* and *C* completely. But the situation occurred where the data that should be judged as *A*′ were distant from a base space. That is, a small number of data in the base space resulted in the space becoming sensitive, so judgments *A*′ and *C* overlapped.

In most cases, to confirm whether abnormal data's distances become distant after creating a base space, we make sure that normal data's distances that are not used to construct the base space lie around 1.0. In this case we faced two typical phenomena, an overlap between abnormal data and the base space and an increase in normal data's dis-

Table 7
Prediction of medical checkup for threshold = 1.5 (%)

	Threshold		
Expected Judgment for Next Year	<1.5	>1.5	Total
A	90	10	100
C	15	85	100

tances. In our research, since the distances of the abnormal data have also increased, distance calculated from the normal data that are not used to construct the base space are not close to 1.0. One of the possible reasons is that the excessive number of items in the base space has led to a difficult judgment. Therefore, it was considered that we should select only items effective to discriminate abnormal data. Yet the most important issue is how many items should be eliminated. Removal of too many items will blur the standard space. We left 60 items in our study. Through the item selection addressed above, we eventually obtained fairly good improvement in prediction accuracy, even though the normal data that are not used for the base space were somewhat distant from the base space. By increasing the number of data, we can expect to realize a highly accurate prediction in the future. In addition, all of the items sieved through the item selection are regarded as essential from the medical viewpoint. By improving the accuracy in item selection with an increased number of data points, we can obtain a proof of medically vital items, thereby narrowing down examination items and streamlining a medical checkup process.

7. Conclusions

According to the results discussed above, we confirmed primarily that a Mahalanobis distance is highly applicable to medical checkup data. For the prediction of health conditions for the next year, despite its incompleteness, it was concluded that we obtained relatively good results. Additionally, although we have attempted to perform an analysis of the three-year data, we considered that a larger number of data are needed to improve the prediction accuracy.

We summarize all of the above as follows:

1. Although we cannot fully separate healthy and unhealthy persons by using Mahalanobis distances based on the one-year medical checkup data despite a large amount of data used, if the two-year data are used, we can distinguish both groups accurately.
2. If the two-year data are used and necessary items are selected, it is possible to predict the current year's health conditions for most people on the basis of the preceding year's data. Furthermore, if the three-year data are used, we could improve the predictability drastically.
3. The effective examination items sieved through item selection cover all the items that are regarded as essential for a daily medical consultation.
4. The MTS method used in this research not only is effective for improvement of a medical checkup system but will also be applicable to the diagnosis and medical treatment of diseases if the methodology is advanced further.

References

Yoshiko Hasegawa, 1997. Mahalanobis distance application for health examination and treatment of missing data: III. The case with no missing data. *Quality Engineering*, Vol. 5, No. 5, pp. 46–54.

Yoshiko Hasegawa, 1997. Mahalanobis distance application for health examination and treatment of missing data: II. A study on the case for missing data, *Quality Engineering*, Vol. 5, No. 6, pp. 45–52.

Yoshiko Hasegawa and Michiyo Kojima, 1994. Prediction of urinary continence recovery among patients with brain disease using pattern recognition. *Standardization and Quality Control*, Vol. 47, No. 3, pp. 38–44.

Yoshiko Hasegawa and Michiyo Kojima. 1997. "Prediction of urinary continence recovery among patients with brain disease Using Mahalanobis Distance," *Quality Engineering*, Vol. 5, No. 5, p. 55–63.

Tatsuji Kanetaka, 1997. An application of Mahalanobis distance to diagnosis of medical examination. *Quality Engineering*, Vol. 5, No. 2, pp. 35–44.

Hisato Nakajima, Kei Takada, Hiroshi Yano, Yuka Shibamoto, Ichiro Takagi, Masayoshi Yamauchi, and Gotaro Toda, 1998. Forecasting of the future health from existing medical examination results using Mahalanobis–Taguchi system. *Quality Engineering*, Vol. 7, No. 4, pp. 49–57.

This case study is contributed by Hisato Nakajima, Kei Takada, Hiroshi Yano, Yuka Shibamoto, Ichiro Takagi, Masayoshi Yamauchi, and Gotaro Toda.