Case Grading Method for Getting Maximum Diagnostic Accuracy by Multiple Tests

Sadayasu Shibata, M. Sci.
Kitasato University School of Allied Health Sciences
16-26 Sumiyoshidai Aobaku Yokohama, Japan 227-0035
email: sshibata@excite.co.jp

***Abstract*----** **Since my presentation in Vancouver, the study of finding the optimum method for getting the maximum diagnostic accuracy by the multiple tests has been continued. This paper focuses on how to judge the remaining subjects (so called grey-zone subjects) who are between all positive and all negative in multiple tests. In order to solve this problem I proposed "the case grading method". This method is to define the weight on the each test, to calculate the sum of the weights on the positive tests for all possible cases and to define the grade of each case. The point of this study is how to define the weight, the grade and the thresholds of the final diagnosis.**

I. INTRODUCTION

The final aim of this study is to earn "maximum outcome with minimum cost" in the field of medical diagnosis. Figure 1 shows situation of this study in which multiple tests are used for diagnosis of one disease.

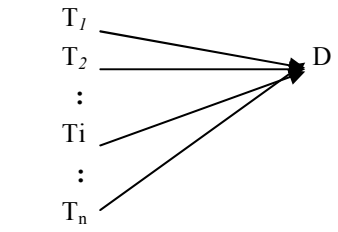

Fig. 1. Situation of this study.

The problem how to classify the overlapped cases in the multi-dimensional space is an old but new problem in the medical diagnosis. J. Swets and R. Pickett founded the basis of evaluation of diagnostic Systems in 1982. [1] P. Polistar made the definition of the decision rules and the value of repeated tests in 1982. [2] R. Miller reviewed medical diagnostic decision support systems in 1994. [3] S. Walter, S. Les Irving and P. Glaziou proposed the use of meta-analysis of diagnostic tests with imperfect reference standards in 1999. [4]

II. KINDS OF METHODS OF MULTIPLE TESTS

The following five kinds of combination methods were defined in my previous paper.

(a). Sequential tests (repeating tests in series).

(b). Sequential tests after changing the order in (a).

(c). (new A) Simultaneous tests using the Believe the Negative Rule which evaluates those subjects who obtained "negative" results even in one of the tests as "normal" and as "abnormal" the rest of the subjects. [2]

(d). (new B) Simultaneous tests using the Believe the Positive Rule which evaluates those subjects who obtained "positive" results even in one of the tests as "abnormal", and as "normal" the rest of the subjects. [2]

(e). (new C) Simultaneous tests using the Believe the all Positive and the all Negative e" in all the tests as "abnormal", and as "normal" those who obtained "negative" in all the tests, and as "borderline or not judged" for the rest of the subjects. [5]

As it was verified in my previous study [3] that the methods of the multiple tests (a), (b) and (c) are equivalent concerning the over-all diagnostic accuracy, five kinds of the multiple tests are attempted to summarize into three kinds in this paper. That is the former (a), (b) and (c) methods become present A method, and the former (d) and (e) method becomes B and C respectively as shown in Table 1.

The hatched part of "Not Judged" in this table is the concerning matter of this paper, which is the target cases to be processed by the case grading method.

Table 1. Three evaluation rules for multiple tests.

Result of Tests		Evaluation Rules		
Test 1	Test 2	A Believe Negative Rule	B Believe Positive Rule	C Believe Positive Negative Rule
+	+	Abnormal	Abnormal	Abnormal
+	-	Normal	Abnormal	**Not Judged**
-	+	Normal	Abnormal	**Not Judged**
-	-	Normal	Normal	Normal

III. DEFINITION OF DIAGNOSTIC ACCURACY

The diagnostic accuracies of the test i are defined by sensitivity (α_i) and specificity (β_i) which are calculated from the 2×2 table as shown in Table 2.

$$\text{Sensitivity}(\alpha_i) = \frac{a_i}{a_i + c_i} \quad (1)$$

$$\text{Specificity}(\beta_i) = \frac{d_i}{b_i + d_i} \quad (2)$$

Table 2. A 2×2 table of test i to define its sensitivity and specificity.

Result of Test i \ True Diagnosis	Abnormal	Normal
+	a_i	b_i
-	c_i	d_i

IV. METHODS

The method to calculate overall diagnostic accuracy of multiple tests including correlation coefficients between the tests, is based on making an overall 2×2 table for multiple tests, which is inducted from 2×2 tables for each test. Table 3 shows the process of getting an overall 2×2 table in case of two tests, including three kinds of evaluation rules. At first, this overall table indicates not a 2×2 but a 2×4 style. After adopting the evaluation rules, the table becomes a 2×2 table. The single-dotted line in this overall table corresponds to the method A, that is the believe-the-negative rule. The double-dotted line corresponds to the method B, that is the believe-the-positive- rule. In the method C, the subjects between the single-dotted line and the double-dotted line are not judged, and should go to be processed by the Case Grading Method.

Table 3. A 2×4 table between true diagnosis and two test results including correlation between test 1 and test 2.

Test Results		True Diagnosis	
Test 1	Test 2	Abnormal	Normal
+	+	$\alpha_1\alpha_2+\Phi_\alpha$	$(1-\beta_1)(1-\beta_2)+\Phi_\beta$
	-	$\alpha_1(1-\alpha_2)-\Phi_\alpha$	$(1-\beta_1)\beta_2-\Phi_\beta$
-	+	$(1-\alpha_1)\alpha_2-\Phi_\alpha$	$\beta_1(1-\beta_2)-\Phi_\beta$
	-	$(1-\alpha_1)(1-\alpha_2)+\Phi_\alpha$	$\beta_1\beta_2+\Phi_\beta$

Table 4 shows a 2×2 table to define correlation coefficient between test 1 and test 2 concerning abnormal subjects.

Table 4. A 2×2 table between test 1 and test 2 concerning abnormal subjects in order to define correlation coefficient concerning abnormal subjects. In this table, "m" is number of abnormal subjects, and "p" is occurrence probability.

Test 2 \ Test 1	+(a_2)	−(c_2)
+(a_1)	mp_{11}	mp_{10}
−(c_1)	mp_{01}	mp_{00}

Correlation coefficient φ_α (4 point correlation coefficient) between test 1 and test 2 concerning abnormal subjects is expressed by formula (3).

$$\varphi_\alpha = \frac{m^2 p_{11} p_{00} - m^2 p_{10} p_{01}}{\sqrt{m^4 (p_{11} + p_{10})(p_{01} + p_{00})(p_{11} + p_{10})(p_{10} + p_{00})}} = \frac{p_{11} p_{00} - p_{10} p_{01}}{\sqrt{\alpha_1 \alpha_2 (1-\alpha_1)(1-\alpha_2)}} \quad (3)$$

And the compensation term Φ_α and Φ_β with the correlation coefficients between test 1 and test 2 are expressed by formula (4) and (5) respectively.

$$\Phi_\alpha = \varphi_\alpha \sqrt{\alpha_1 \alpha_2 (1-\alpha_1)(1-\alpha_2)} \quad (4)$$

$$\Phi_\beta = \varphi_\beta \sqrt{\beta_1 \beta_2 (1-\beta_1)(1-\beta_2)} \quad (5)$$

Where φ_β is correlation coefficient between test 1 and test 2 concerning normal subjects.

V. RESULTS

The generalized formulas for the overall Sensitivity (α) and Specificity (β) in the case of A, B and C as a function of number of the tests (n) are calculated from the overall 2×2 table of each case shown as formula (6) through (8)'. The compensation term Φ_α and Φ_β in the formula (6), (7), (7)', (6)', (9), (10), (10)' and (9)' are restricted only for the case of n=2.

In the method C, which corresponds to "the Believe the all Positive and the all Negative Rule", the sensitivity and specificity are obtained with the same type of formula as (8) and (8)'.

$$\alpha_A = \prod_{i=1}^{n} \alpha_i + \Phi_\alpha \qquad (6)$$

$$\beta_A = 1 - \prod_{i=1}^{n} (1 - \beta_i) - \Phi_\beta \qquad (7)$$

$$\alpha_B = 1 - \prod_{i=1}^{n} (1 - \alpha_i) - \Phi_\alpha \qquad (7)'$$

$$\beta_B = \prod_{i=1}^{n} \beta_i + \Phi_\beta \qquad (6)'$$

$$\alpha_C = \frac{\prod_{i=1}^{n} \alpha_i}{\prod_{i=1}^{n} \alpha_i + \prod_{i=1}^{n} (1 - \alpha_i)} \qquad (8)$$

$$\beta_C = \frac{\prod_{i=1}^{n} \beta_i}{\prod_{i=1}^{n} \beta_i + \prod_{i=1}^{n} (1 - \beta_i)} \qquad (8)'$$

The boldly assumption $\alpha_n = \beta_n$ and $\varphi_\alpha = \varphi_\beta$ was adopted in order to understand the tendency of the change of overall sensitivity and specificity when the number of tests is changed. Then, three kind of formulas (6) through (8)' became the simplified formulas (9) through (11)', respectively.

$$\alpha_A = \alpha^n + \Phi_\alpha \qquad \textbf{(9)}$$

$$\beta_A = \Sigma\,(-1)^{n-1}\, {}_nC_i\, \beta i - \Phi_\beta \qquad \textbf{(10)}$$

where $n\,C\,r = n!\,/\,(n-r)!\,r!$

$$\alpha_B = \Sigma\,(-1)^{n-1}\, {}_nC_i\, \alpha i - \Phi_\alpha \qquad \textbf{(10)'}$$

$$\beta_B = \beta^n + \Phi_\beta \qquad \textbf{(9)'}$$

$$\alpha_C = \alpha^n \,/\, \{\alpha^n + (1-\alpha)^n\} \qquad \textbf{(11)}$$

$$\beta_C = \beta^n \,/\, \{\beta^n + (1-\beta)^n\} \qquad \textbf{(11)'}$$

These simplified generalized formulas for the overall sensitivity and specificity were expressed by the graphical expressions shown in Figure 2. The upper two graphs show in case of the method A. Where is two kinds of curves, in case of $\alpha_n = \beta_n = \mathbf{0.8}$ (solid line) and **0.6** (dotted line). The "curves" show in case of correlation coefficient of tests is "0", and "horizontal lines" shows in case of correlation coefficient is "1". The axis of the ordinate of the lowest graph in this Figure shows both sensitivity and specificity in the method C, the Believe-the-all-Positive-and-the-all-Negative Rule. It can be seen in these three graphs that for method C both sensitivity and specificity are excellent, although the problem of how to judge the not-judged subjects remains.

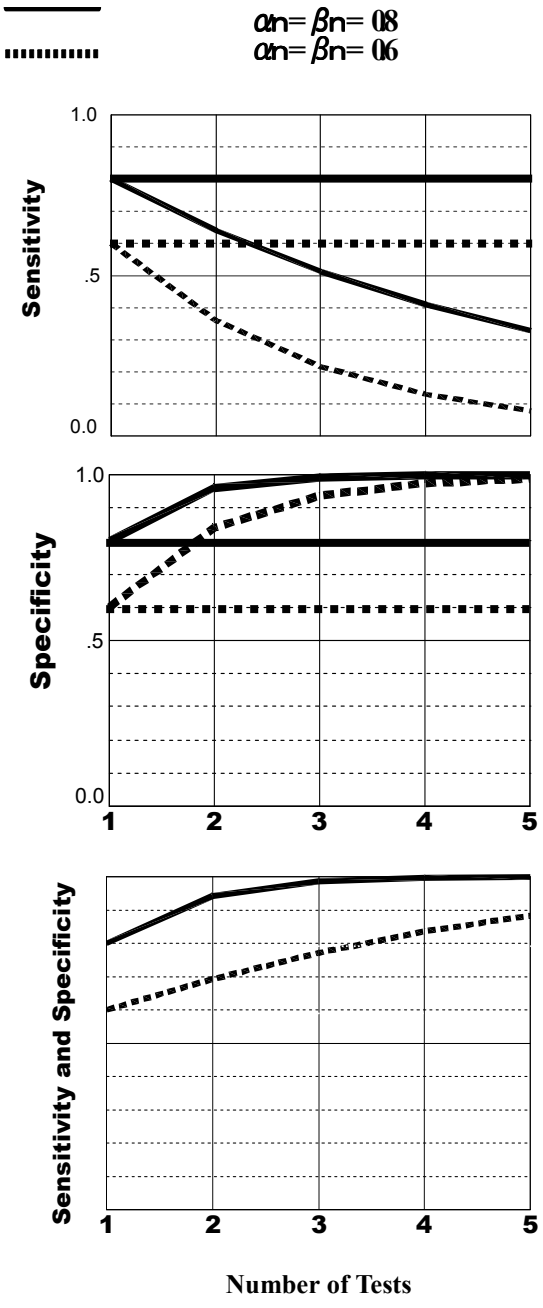

Fig. 2. Three graphs of sensitivity and specificity vs. number of tests, corresponding to the formula (9), (10) and (11) = (11)' from the top to the bottom respectively. The ordinate axis of the top graph shows sensitivity in the method A and also specificity in the method B which is the formula (9)'. The ordinate axis of the middle graph shows specificity in the method A and also sensitivity in the method B which is the formula (10)'. The ordinate axis of the bottom graph shows both sensitivity and specificity in the method C.

VI. CASE GRADING METHOD

I propose the Case Grading Method to solve the problem how to judge the borderline subjects. This method is to grade all the cases occurring between abnormal and normal. The all-positive case is defined as "the most abnormal" and its **NDP** (Normalized Degree of Positiveness) is "1", and the all-negative case is defined as "the most normal" and its NDP is "0". To simplify the explanation, the number of tests will be limited to three as shown in Table 5. This shows all the possible combinations of the test results. Test 1 is positive (indicated by a plus sign) in cases 1, 2, 3 and 5, and negative (indicated by space) in cases 4, 6, 7 and 8. Test 2 is positive in cases 1, 2, 4 and 6; and Test 3 is positive in cases 1, 3, 4 and 7. The right column shows the sum of the number of positive tests for each row.

Table 5. Definition of the case.

Cases	Test 1	Test 2	Test 3	Number of Positive Test
Case 1	+	+	+	3
Case 2	+	+		2
Case 3	+		+	2
Case 4		+	+	2
Case 5	+			1
Case 6		+		1
Case 7			+	1
Case 8				0

However, the problem how to grade the cases having the same number of positive tests remains. The number of positive tests in the previous table was weighted by the "Unified Diagnostic Accuracy Index", "Contribution Rate Index for Diagnosis" in another word.

Table 6 shows the process of case grading and normalization for degree of positivenessness. The degree of positivenessness is defined as the sum of the weights of each positive tests. The normalized degree of positiveness are calculated such as the maximum degree of positiveness ($w_1+w_2+w_3$) has been normalized to **1**.

Table 6. Case grading by Degree of Positiveness and its normalization.

Cases	Weights			Degree of Positiveness	Normalized Degree of Positiveness
	w_1	w_2	w_3		
Case 1	+	+	+	$w_1+w_2+w_3$	1
Case 2	+	+		w_1+w_2	$\frac{w_1+w_2}{w_1+w_2+w_3}$
Case 3	+		+	w_1+w_3	$\frac{w_1+w_3}{w_1+w_2+w_3}$
Case 4		+	+	w_2+w_3	$\frac{w_2+w_3}{w_1+w_2+w_3}$
Case 5	+			w_1	$\frac{w_1}{w_1+w_2+w_3}$
Case 6		+		w_2	$\frac{w_2}{w_1+w_2+w_3}$
Case 7			+	w_3	$\frac{w_3}{w_1+w_2+w_3}$
Case 8				0	0

Finally, the normalized degree of positiveness are categorized into three categories. For example, when the normalized degree of positiveness is greater than the upper threshold (**0.7** for instance), those cases are categorized as abnormal. When less than the lower threshold (**0.3** for instance), they are categorized as normal, and the rest are categorized as borderline or not judged.

VII. WEIGHTS FOR CASE GRADING METHOD

The six kinds of "Unified Diagnostic Accuracy Index" as the weight for the case grading method are comparatively examined. "Unified" means to seek the one index of how much contribute to the diagnosis for the aiming disease instead of two indexes of sensitivity (**α**) and specificity (**β**) as diagnostic accuracy. Six indexes include odds-ratio, positive likelihood ratio, negative likelihood ratio, average, root product and root mean square, which are defined as the following formulas.

$$\text{Odds-ratio} = \frac{a_i d_i}{b_i c_i} = \frac{\alpha_i \beta_i}{(1-\alpha_i)(1-\beta_i)}$$

$$\text{Positive-likelihood-ratio} = \frac{a_i/(a_i+c_i)}{b_i/(b_i+d_i)} = \frac{\alpha_i}{1-\beta_i}$$

$$\text{Negative-likelihood-ratio} = \frac{c_i/(a_i+c_i)}{d_i/(b_i+d_i)} = \frac{1-\alpha_i}{\beta_i}$$

$$\text{Average} = \frac{\alpha_i + \beta_i}{2}$$

$$\text{Root product} = \sqrt{\alpha_i \beta_i}$$

$$\text{Root mean square} = \sqrt{\alpha_i^2 + \beta_i^2}$$

Figure 3 shows simplified feature of the six kinds of indexes. This graph was made by the assumption $\boldsymbol{\alpha_i} = \boldsymbol{\beta_i}$ changing from **0.6, 0.7** to **0.8** as variables in order to see the tendency of the change of the unified indexes. In these assumptions, the curve of average is behind of the curve of RMS, because these have the same values. So, the number of graphs is not six but five in this Figure.

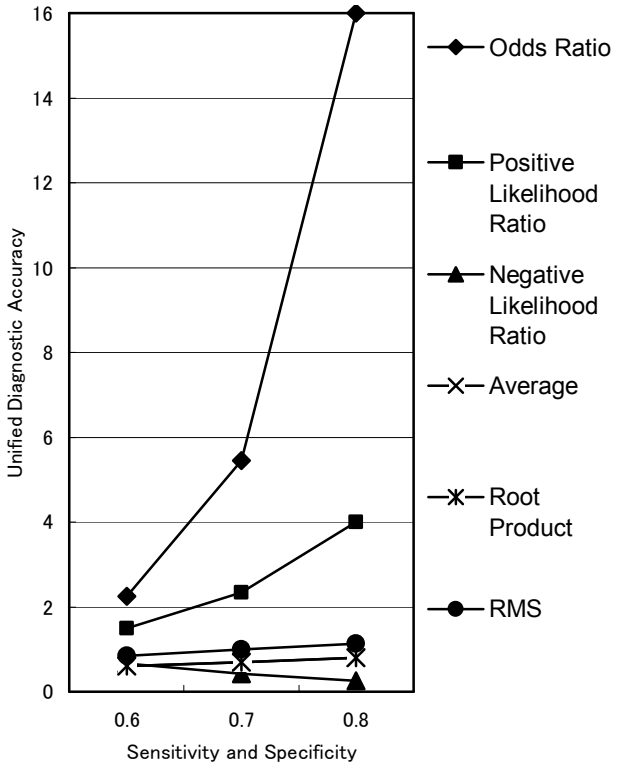

Fig. 3. Unified indexes of diagnostic accuracy.

The difference of **NDP** in five kinds of indexes which removed negative-likelihood-ratio from the six kinds is compared as shown in Figure 4 and 5. The boldly assumption of $\boldsymbol{\alpha_1} = \boldsymbol{\beta_1} = \mathbf{0.8}$, $\boldsymbol{\alpha_2} = \boldsymbol{\beta_2} = \mathbf{0.7}$ and $\boldsymbol{\alpha_3} = \boldsymbol{\beta_3} = \mathbf{0.6}$ was made to calculate the weights w1, w2 and w3, respectively.

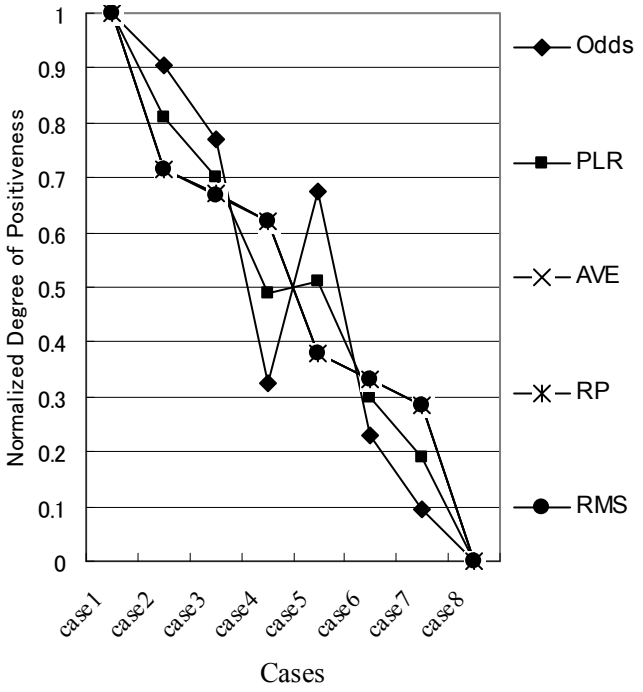

Fig. 4. The difference of NDP in five kinds of indexes along the ordinary case order. (In case of n=3)

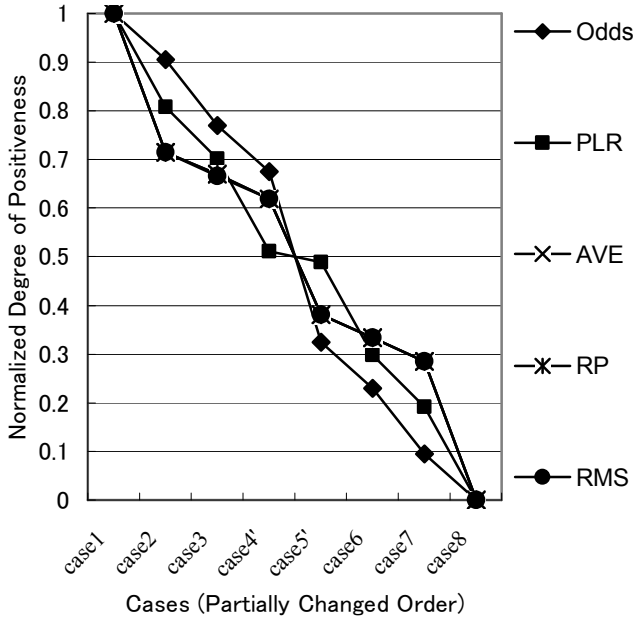

Fig. 5. The difference of NDP in five kinds of indexes along the changed case order. (In case of n=3)

The abscissas axis of Figure 4 shows eight cases in ordinary order which is corresponding to the Table 5 and 6. The curves of average and root product are behind of the curve of RMS (almost central position in this graph), because these three values are the same in these assumption. So, the number of graphs is not five but three in this Figure.

The abscissas axis of Figure 5 was changed from Figure 4 along the magnitude of Odds and PLR.

VIII. DISCUSSION

In order to make clear the difference of the weighting indexes, the case which the number of tests is four is tried to be calculated by the same way and be shown in Figure 6. The assumption of $\boldsymbol{\alpha_1 = \beta_1 = 0.9}$, $\boldsymbol{\alpha_2 = \beta_2 = 0.8}$, $\boldsymbol{\alpha_3 = \beta_3 = 0.7}$ and $\boldsymbol{\alpha_4 = \beta_4 = 0.6}$ was made to calculate the weights w_1, w_2, w_3 and w_4, respectively. These graphs in Figure 5 and 6 look similar curves to the membership function in the fuzzy method. Odds ratio seems the most emphasizing index to separate normal and abnormal subjects.

On the other hand, "average" is equal to "true rate" and "efficiency" when "prevalence" is **0.5**.

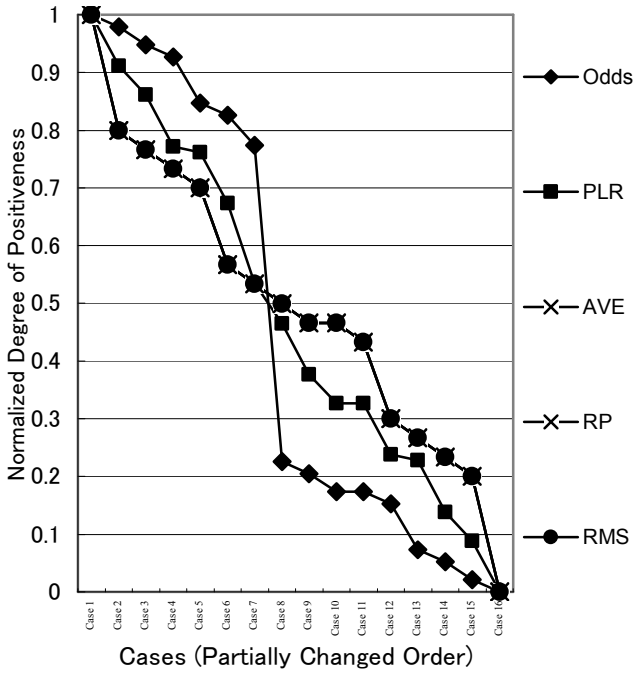

Fig. 6. The difference of NDP in five kinds of indexes along the changed case order. (In case of n=4)

IX. CONCLUSION AND FURTHER POSSIBILITY

Adopting the Case Grading Method in order to judge the "borderline" subjects in the multiple test method C (the Believe-the-all-Positive-and-the-all-Negative Rule), it is also proved by the actual data that the maximum over-all diagnostic accuracy can be realized.

In this paper, the "test" means not only the laboratory examination but also the questionnaire and the physical findings, etc.

This method seems to be the more useful, the more plenty numbers of cases. This also can easily expand to such case as having multiple quantized values for one test result. Table 7 shows the number of cases (m^n) corresponding to the number of tests (n) and the number of quantized values (m) of each test .

Table 7. The number of cases corresponding to the number of tests and quantized values of each test.

No. of Values / No. of Tests	2	3	4	5	…	m
2	4	9	16	25	…	m^2
3	8	27	64	125	…	m^3
4	16	81	256	625	…	m^4
5	32	243	1024	3125	…	m^5
⋮	⋮	⋮	⋮	⋮		⋮
n	2^n	3^n	4^n	5^n	…	m^n

ACKNOWLEDGEMENT

The author would like to appreciate Professor Noriaki Ikeda, Department of Medical Informatics, Kitasato University School of Allied Health Sciences for his supervision of this study.

REFERENCES

[1] J. Swets and R. Pickett. *Evaluation of Diagnostic Systems*: Academic Press, New York, 1982.

[2] P. Polistar. Reliability, Decision Rules, and the Value of Repeated Tests, *Med Decision Making,* Vol.2 No.1, pp47-69, 1982.

[3] R. Miller. Medical Diagnostic Decision Support Systems -Past, Present and Future: A Threaded Bibliography and Brief Commentary, *Journal of the American Medical Informatics Association*, Vol.1 No.1, pp8-27, 1994.

[4] S. Walter, S. Les Irving and P. Glaziou, Meta-Analysis of Diagnostic Tests with Imperfect Reference Standards, *J Clinical Epidemiology,* Vol.52 No.10, pp943-951, 1999.

[5] S. Shibata. Soft-computing-Considerations on Diagnostic Accuracy of Multiple Tests, *Abstracts of Joint 9th IFSA (International Fuzzy System Association) World Congress and NAFIPS (North American Fuzzy Information Processing Society) International Conference*, p109, Vancouver, 2001.

[6] S. Shibata. Fundamental Considerations on the on the Health Checkup Accuracy of Multiple Testing, *Abstract of Biennial Conference of the IHEA (International Health Evaluation Association)*, Taipei, pp47-48, 2000.

[7] S. Shibata. Theoretical Considerations on the Health Checkup Accuracy of Combination Testing, *Methods of Information in Medicine*, Vol.41 No.3, pp216-219, 2002.

[8] S. Shibata. Fundamental Considerations on the on the Health Checkup Accuracy of Multiple Tests, *Health Evaluation and Promotion*, Vol.29 No.4, pp814-818, 2002.