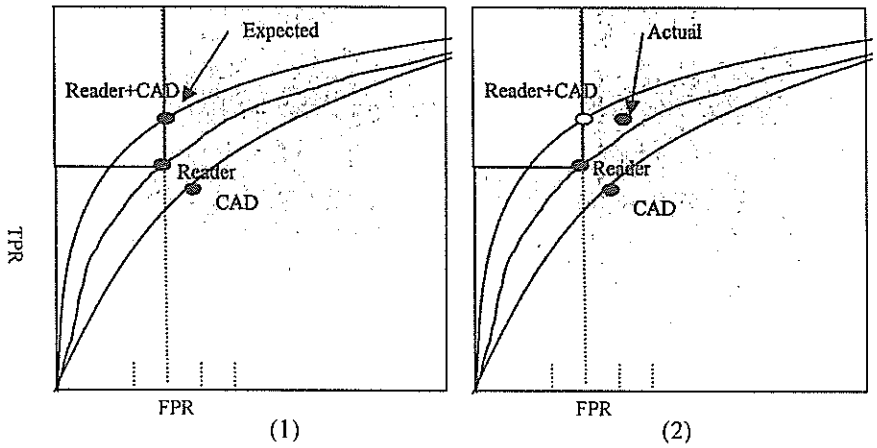


Figure 11 Expected FPR and TPR(1) and actual results (2) in the CAD use of Fig.10(2)

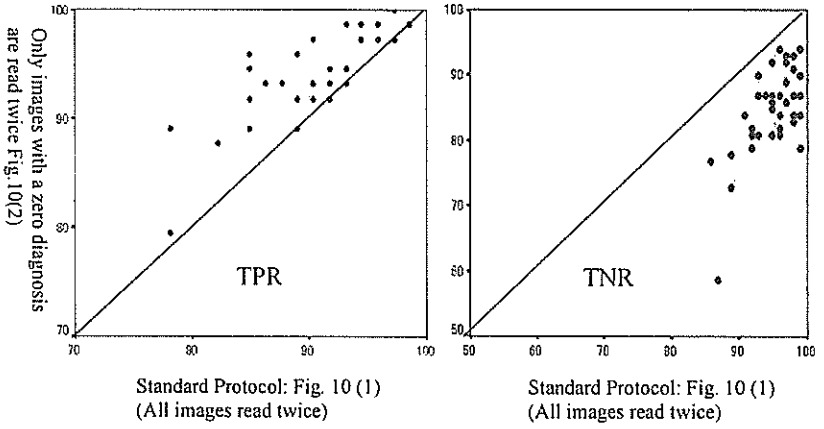


Figure 12 Comparison of TPR and TNR of the CAD use of Fig.10 (1) with that of Fig.10 (2)

However, after the opportunity of a second reading, as shown in Figure 11 (2), in which a CAD-noncompliant response is also possible, the reader +CAD TNR3 (FPR3) decreases (increases) relative to the TNR1 (FPR1) for the reader alone. Based on the risk of a TNR decrease (FPR increase), it is a protocol for CAD use that aims to increase the TPR (FNR: decrease in oversight). We compared in Figure 12 the diagnostic accuracy (TNR3, TPR3) for the standard protocol and that of Figure 10 (2). While the TPR for Figure 10 (2) may be the same or greater than that of the standard protocol for many readers, TNR3 (FPR3) was smaller (larger) than FNR1 (FPR1) for all readers.

However, cases in which the reader makes a diagnosis of 1 will not require CAD consultation. It reduces the inconvenience of making a diagnosis by comparison of the CAD output with one's own diagnosis. As well, the potential for CAD-noncompliant responses is reduced. If the quality control when a person reads images using CAD is refined the near future, and we consider that the reader attempts to perform the reading process with a constant FPR1 (=1–TNR1), the protocol for CAD use in Figure 10 (2) may be more practical than the standard protocol of Figure 10 (1).

Figures 10 (3) and (4) represent further refinements of (1) and (2). The diagnostic accuracy of these methods may be in accordance with Figures 11 (1) and (2) respectively, but have the advantage in that the inconvenience of CAD usage is reduced and CAD-noncompliant responses are further decreased.

Figure 10 (5) is an attractive protocol for CAD use that should be the goal for CAD, but it will not be possible unless the CAD system has exceptionally high performance.

The prime advantage of the protocol for CAD use in Figure 10 (6) is that all images only need to be read once. One issue that warrants investigation is whether the results obtained by this protocol will be the same or differ in some way from the results for reader +CAD (TNR3, TPR3) obtained for the protocol in Figure 10 (1).

5.4 Application to re-evaluation of the CAD assistance effect reported in previous papers

The protocol described in this report can be applied to re-evaluate the assistance provided by CAD systems published in previous papers under actual clinical conditions. In brief, in the paper, if the three ROC curves are obtained for reader, CAD, and reader +CAD, or if three pairs of TNR and TPR data are demonstrated, the difference between reader and reader +CAD can be evaluated by the method described in this report. For example, in the case of Figure 11 (1), it is possible to evaluate the difference in TPR (exact p-value) when the horizontal axis FPR for reader and reader +CAD is constant. However, by varying the FPR (or TNR), a difference in TPR may or may not be detected. It is possible to evaluate its practicality by ascertaining where the horizontal axis FPR produces the largest difference between reader and reader +CAD.

6. CONCLUSION

With three pairs of TPR values and three pairs of TNR value for the reader, CAD, and reader +CAD as known quantities, we proposed a method for estimating contingency tables for reader/CAD, and reader/reader +CAD to achieve the actual reader +CAD TPR or TNR. The estimated values showed a state in which the reader has used CAD according to rational criteria, and that CAD is used in actual practice with deviations from that state. The degree of difference between the two is an individual difference, and readers with smaller differences have high complementarity with CAD, leading to superior diagnostic performance. Via these investigations, we have been able to reveal a part of the mechanism by which the use of CAD can improve diagnostic accuracy.

ACKNOWLEDGMENTS

This study was conducted as part of the joint study of Niigata University and Fujitsu Limited. This work was supported in part by a Grant-in-Aid for Cancer Research (15-25) from the Ministry of Health, Labor and Welfare and in part by a Grant-in-Aid for Scientific Research in Priority Areas (15070205) from the Ministry of Education, Science, Sports and Culture in Japan.

REFERENCES

1. http://www.r2tech.com/
2. TW. Freer et al. Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. *Radiology,* 220(3):2001; 781-786
3. N. Karssemeijer et al. Computer-aided detection versus independent double reading of masses on mammograms. *Radiology,* 227(1):2003; 192-200
4. D. Gur et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J. of National Cancer Institute,* 96(3):2004;185-190
5. T. Hara et al. Current status of mammography CAD, *Proceedings of Japanese Society of CADM '05*:2005;263-264
6. T. Matsumoto et al. An assessment of the potential for interpretation of CT images by radiological technologists, *Proceedings of SPIE 2005*, 5749:2005; 590-600.
7. T. Matsumoto et al. Methods of evaluating the effectiveness of double-checking in interpreting mass screening images. *Proceedings of SPIE 2004*, 5372:2004;496-508.

A Study on the Performance Evaluation of Computer-aided Diagnosis for Detecting Pulmonary Nodules for the Various CT Reconstruction

Shinichi Wada,[a] Toru Matsumoto,[b] Kohei Murao,[c] Shusuke Sone,[d]
[a] Niigata University; 746 2-bancho, Asahimachi-dori, Niigata-shi, Niigata, JAPAN 951-8518
[b] National Institute of Radiological Sciences; 9-1, Anagawa-4, Inage-ku, Chiba-shi, Chiba, JAPAN 263-8555, [c] Bio IT Business Development Group, Fujitsu Ltd; 17-25, Shinkamata 1-chome,Ota-ku, Tokyo-to, JAPAN 144-8558, [d] JA Azumi General Hospital; 3207-1 Ikeda-ooaza, Ikeda-cho, Kitaazumi-gun, Nagano-ken, JAPAN 399-8695

ABSTRACT

The purpose of this study was to evaluate the performance of computer-aided diagnosis (CAD) system detecting pulmonary nodules for the various CT image qualities of the low dose CT cancer screening. Sixty three chest examinations cases with sixty-four pulmonary nodules consisting mainly ground-glass opacity (GGO) were used. All the CT images were acquired by using a multi-slice CT scanner Asteion with 4 detector rows system (Toshiba Medical Systems Co. Ltd, Japan) with 0.75-second rotating time and 30mA. After the examination, CT image reconstructions were performed for every CT data set using seven reconstruction kernels and three sorts of slice thickness, i.e. 5mm, 8mm and 10mm thickness. Totally twenty-one data sets for a patient, namely 1323 data sets with about 60 thousands CT images which is 30.1GB data sets were investigated. The seven reconstruction kernels consist of three types of standard or smoothed kernels, and four types of high resolution kernels. Nodule detections were carried out using a computer-aided diagnosis system for automatic detection of pulmonary nodule developed by Fujitsu Ltd, Tokyo, Japan. The mean nodule size was 0.69±0.28 (SD)[cm](range,0.3-1.7cm). The CAD system identified 42 to 48 nodules out of the 64 nodules, in the slice thickness of 8mm for the seven reconstruction kernels, yielding a true-positive rate (TPR) of 65% to 75%. In the slice thickness of 5mm our CAD system indicates a TPR from 70% to 80%. In the slice thickness 10mm, TPR were resulted from 50% to 64%. Some kernel indicated relatively high TPR with high false positive fractions (FPF), other kernel showed high sensitivity with relatively low FPF. CT image data sets with multi-reconstruction conditions are useful in assessing the robust characteristics of a CAD system detecting pulmonary nodule by multi-slice low dose CT screening.

Keywords: CAD performance, low dose CT screening, reconstruction condition, lung cancer screening, CT reconstruction kernel

1. INTRODUCTION

Among all deaths from malignant neoplasm, lung cancer is the most common causes of death in Japan. Chest computed tomography (CT) is the most sensitive diagnostic imaging modality for the detection of lung cancer. With helical CT and, more recently, multi-slice CT further sensitive detection of pulmonary nodules has been accomplished.

Recently helical or multi-slice CT techniques have been applied to screening detecting early lung cancer[1-2]. Consequently the burst numbers of CT images would be appeared for the diagnosis. In this circumstance the computer-Aided Diagnosis (CAD) system have been developed to detect the pulmonary nodules.[3-8]

Although there are many reports for CAD system for CT-lung cancer screening, very few study are observed about the relationship between CT image qualities or CT reconstruction factors and CAD performance. In the actual use of the CT-CAD system, the CT image quality will have much variety on the clinical use in the imaging or image reconstruction process. Consequently the evaluation method should be investigated about the CAD performance and CT image quality. The robustness of CAD performance should be mentioned.

This study proposes a scheme of the method for the performance evaluation of CAD system for detecting pulmonary nodules for various CT image qualities of reconstructed images using MDCT raw data from low-dose CT cancer screening study.

*swada@clg.niigata-u.ac.jp, Phone 81 25 227 2398; fax 81 25 227 2398

Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment, edited by Yulei Jiang, Miguel P. Eckstein, Proc. of SPIE Vol. 6146, 61461C, (2006) · 1605-7422/06/$15 · doi: 10.1117/12.654418

2. MATERIALS AND METHODS

In order to evaluate the performance of the CAD in detecting pulmonary nodules for various CT image qualities, we reconstructed the low-dose CT cancer screening data sets with various reconstruction conditions. The reconstructions were performed with three slice thickness, and seven reconstruction kernels for 63 cases with 64 nodules. The CAD detections for the nodules were performed. The true positive ratio and false positive fractions were discussed.

2.1. Low-dose CT screening and reconstructions

CT image were acquired using a multi detector row CT scanner Asteion-4 (Toshiba Medical, Japan). Tube voltage was 120 kV, tube current was 30 mA with 0.75sec per rotation. Table feed distance was 27.5mm per rotation, and detector configuration was 5mm. That is the detector pitch is 5.5 and the collimation pitch is 1.375. Image reconstructions were underwent with three conditions of slice thickness and seven conditions of reconstruction kernels. These conditions are 5, 8 and 10mm in slice thickness and FC01, FC10, FC20, FC30, FC50, FC51 and FC52 on the kernels which correspond to standard kernel or smoothed kernel (FC01,FC10,FC20), moderate high resolution kernels (FC30,FC50,FC51) and sever high resolution kernel respectively(FC52). Because the scanning ranges of the whole lung–field was about 30 cm, the resulting reconstructed CT images for a data set was 60, 40 and 30 slices in respective slice thickness, and reconstructed images were about 900 CT images for an examination In order to determine the characteristics of the kernels, the line spread function of every reconstruction kernel was measured using a simple phantom of thin aluminum foil (50μ m) in the tissue equivalent material slab phantom proposed by JM.Boone[9-10].

2.2. Patients and nodules

Sixty three chest examination cases with sixty four pulmonary nodules were used in this study, which were diagnosed as lung cancer suspicious or, as case excluding as non-cancer by a board certificated radiologist in the period from September 2004 to September 2005 in the low dose cancer CT screening in Azumi-Genearl Hospital, Nagano Japan. The institutional review board approval was obtained.

2.3. CAD Scheme

The CAD scheme investigated in this study was a computerized scheme for automated detection of pulmonary nodules developed by Fujitsu Ltd (tentatively call f-lung CAD: TOKYO, JAPAN). The another data sets of 8mm slice thickness and FC50 kernel of a low-dose CT cancer screening were used for the tuning of the f-Lung CAD system.

3. RESULTS

3.1 Characteristics of the reconstruction kernel

Figure1. show the measured results of LSFs corresponding to the kernels of the CT-scanner

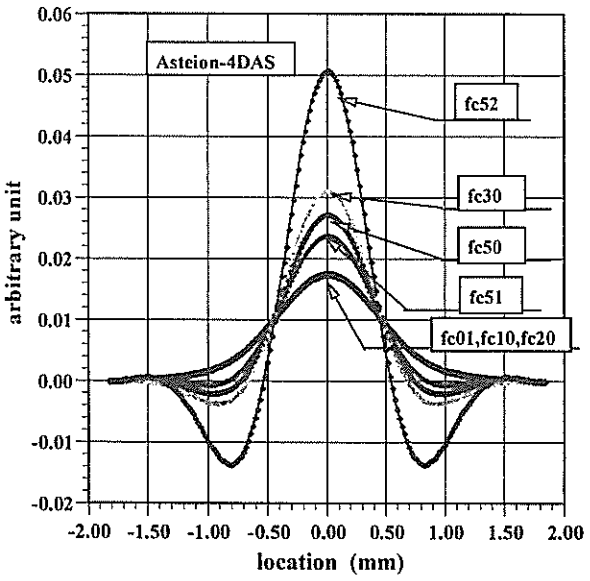


Figure1. The LFSs measured using Boone's phantom. FC50-to 52 is the kernel for lung imaging.

3.2. Case and nodules.

The size distribution of 58 nodules was shown in the figue.2 with minimum diameters and large diameters. The sizes were determined in clinical observation of CT image by a radiologist. The mean of minimum diameters was 6.9mm (2.8mm SD) and large diameter was 7.3mm (3.0mm SD). The others out of the 64 nodules were 5 non-globular with

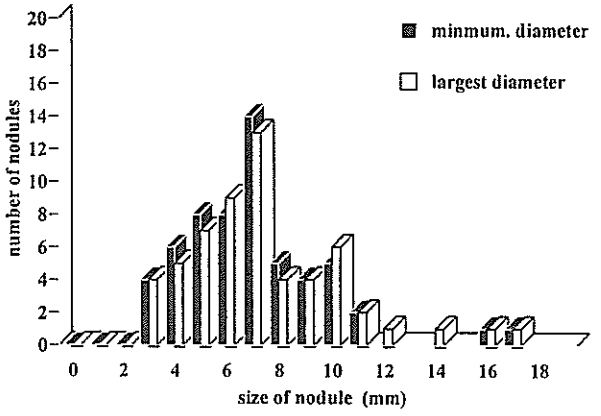
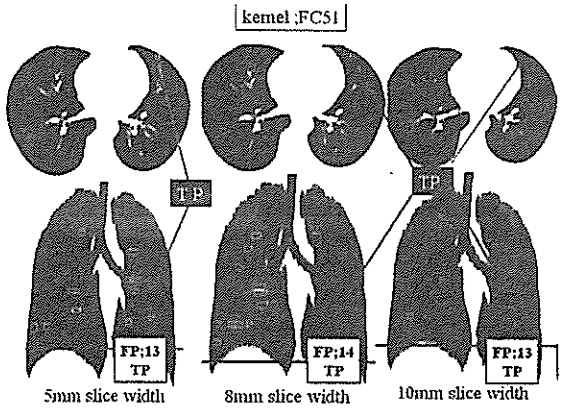
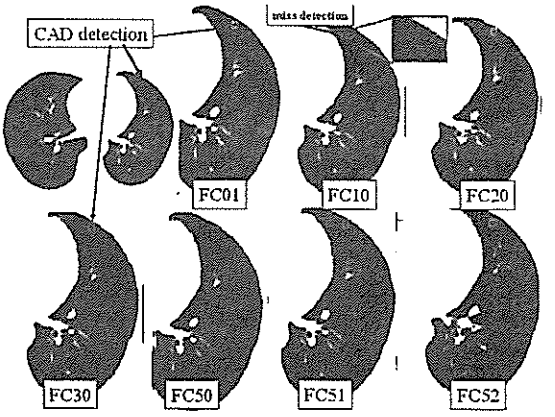
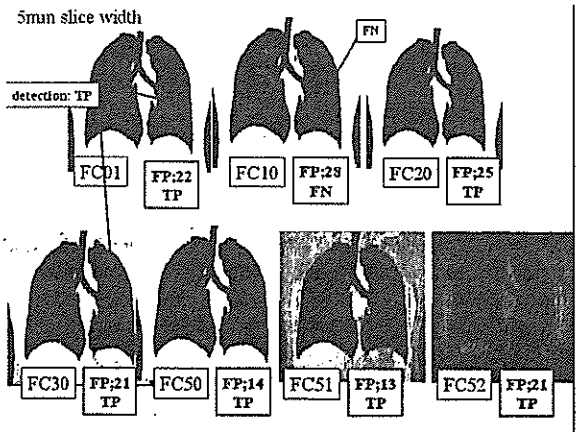


Figure2. The size distribution of the 58 nodules







(a) | (b)
(c)

Figure.3-1. Case result of the nodule detection by f-Lung CAD. Determined nodule size by a radiologist was 5 mm big. In the all reconstruction data sets in 5mm slice thickness and FC51 kernels are detected nodule in true-positive successfully except for the condition of 5mm slice thickness with FC10.

(a) shows the result for the kernel FC51 with 5, 8,10mm slice thickness. In every slice thickness CAD detected the nodule successfully. The falls positive numbers in this case shows nearly equal for these slice thickness. (b) shows axial image of 5mm slice thickness with 7 kernels. The data set FC10 missed the detection. (c) shows the coronal image with CAD detection. The numbers of falls positive decreased in FC50 and FC51.

irregular shaped abnormality with the size of 11 to 25mm in largest diameter. The remainder was a case of small scattered nodules.

3.3. Results of the f-Lung CAD detection and the characteristics.

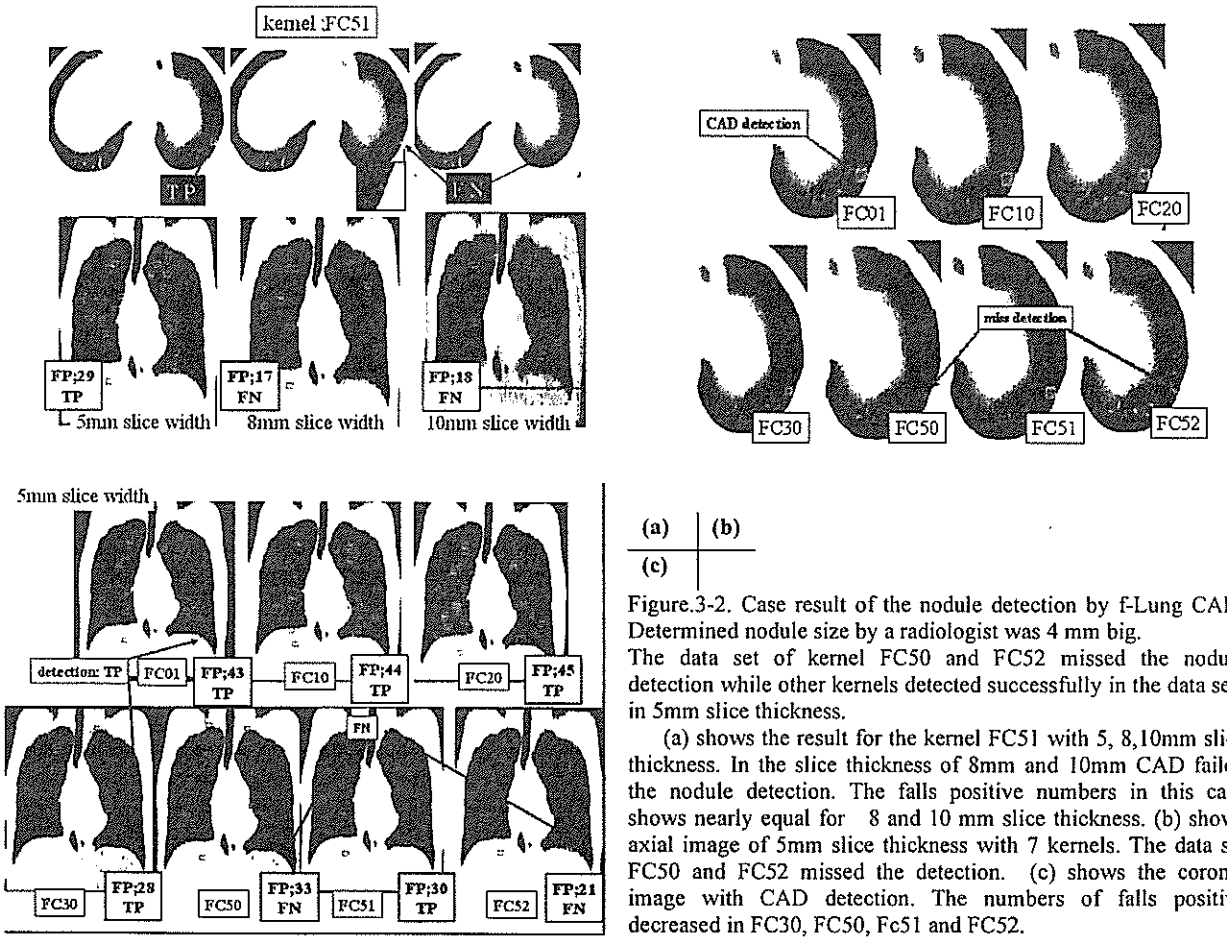
The nodule detection with f-lung CAD was performed for 63 cases of low-dose cancer screening. The totally about 60,000 numbers of reconstructed CT slice images were analyzed. The time spending for nodule detection of an 8mm slice sickness data set, that is about 40 slices, was about 70-80 second with use of the 2.4GHz-CPU Pentium(R) 4 Windows-XP. The total machine time for CAD detection of 60,000 slices of CT image was about 26hrs.

Figure 3 shows the results of two cases CAD detection. The nodules were appeared with the finding of grand-glass opacity of relatively small size on each case. The detection characteristic shows a sort of difference.

True positive ratio (TPR) for the 64 nodules was investigated (Figure 4) for the corresponding slice thickness and kernel. The TPR of the slice thickness 5, 8 and 10 mm were distributed from 70 to 80%, from 66 to 75% and form 50 to 64% respectively. In every slice thickness the FC52 which is recommended as a suitable for lung imaging by CT manufacturer was pointed out the lowest TPR. TPR showed the tendency to go up as the slice thickness decrease

The statistical significant differences were proved among some conditions (figure. 4).

The false positive fractions (FPF) of the CAD system in detecting pulmonary nodules were investigated (Figure 5). The FPF demonstrated a tendency to increase in the standard kernels, while it decreased in the high resolution kernels.



(a) | (b)
(c)

Figure.3-2. Case result of the nodule detection by f-Lung CAD. Determined nodule size by a radiologist was 4 mm big.
The data set of kernel FC50 and FC52 missed the nodule detection while other kernels detected successfully in the data sets in 5mm slice thickness.

(a) shows the result for the kernel FC51 with 5, 8,10mm slice thickness. In the slice thickness of 8mm and 10mm CAD failed the nodule detection. The falls positive numbers in this case shows nearly equal for 8 and 10 mm slice thickness. (b) shows axial image of 5mm slice thickness with 7 kernels. The data set FC50 and FC52 missed the detection. (c) shows the coronal image with CAD detection. The numbers of falls positive decreased in FC30, FC50, Fc51 and FC52.

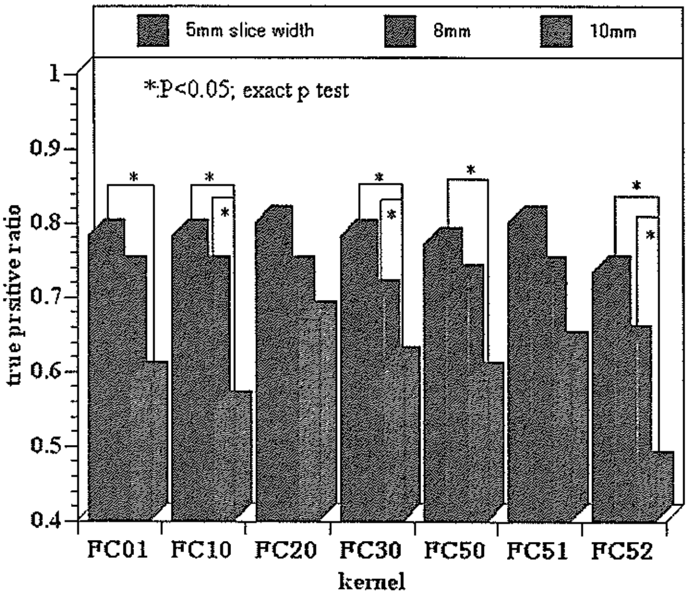


Figure 4.
True positive ratio (TPR) for the 64 nodules The significant differences were proved among some slice thickness.
The kernel recommended as a suitable for lung imaging by CT manufacturer was pointed out the lowest TPR in the respective slice thickness.

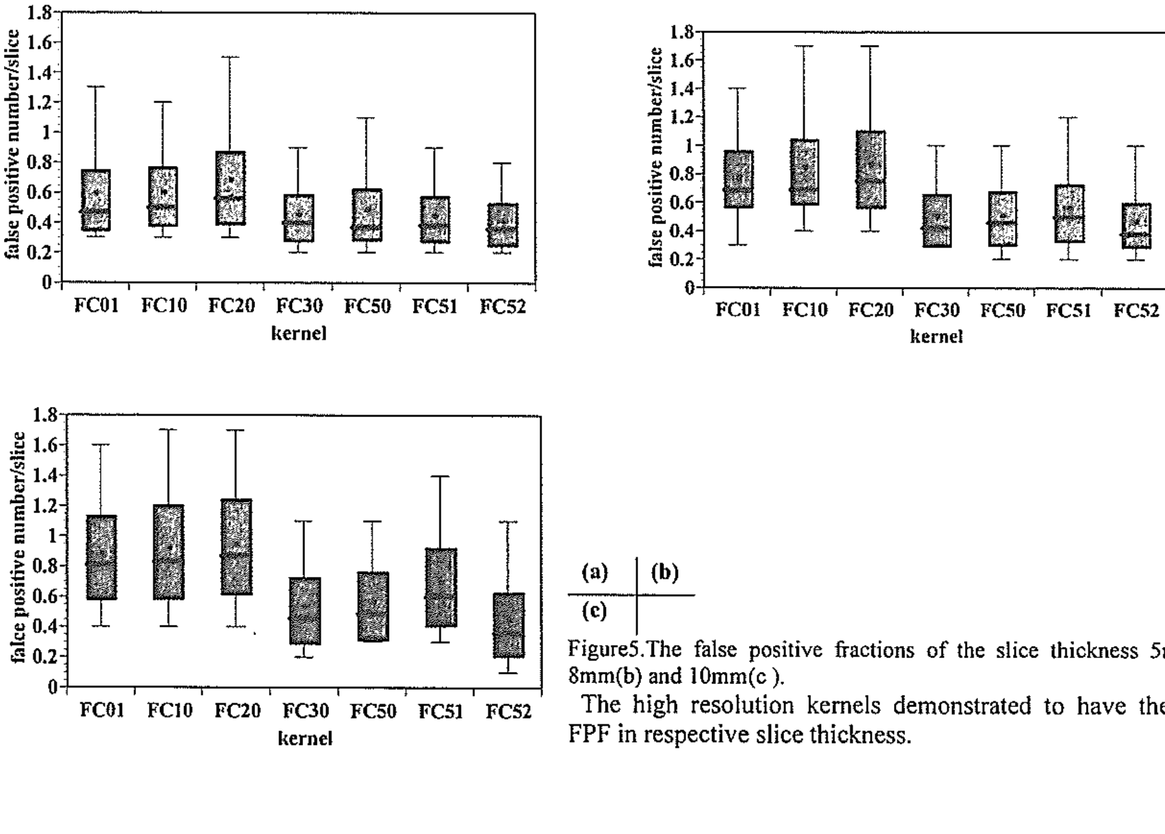


Figure5.The false positive fractions of the slice thickness 5mm(a), 8mm(b) and 10mm(c).
The high resolution kernels demonstrated to have the fewer FPF in respective slice thickness.

4. DISCUSSIONS

The FPF per slice and TPR distribution for the respective slice thickness or kernel was demonstrated in the figure 6. We call this picture a FP-TPR diagram. In the FP-TPR diagram, x-axis indicates the median of the FPF per slice (Figure 6) or the FPF per study (Figure 7), y axis indicates the TPR.

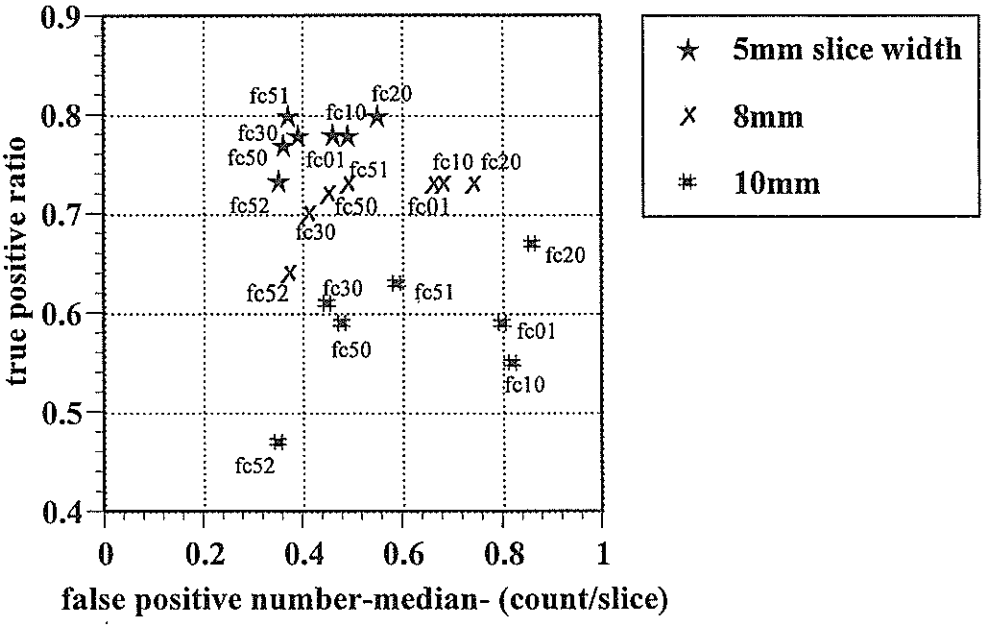


Figure6. FP-TPR diagram per slice. The higher performance in 5mm slice thickness with the reconstraction kernel FC51, FC30 or FC50 was demonstrated in this diagram.

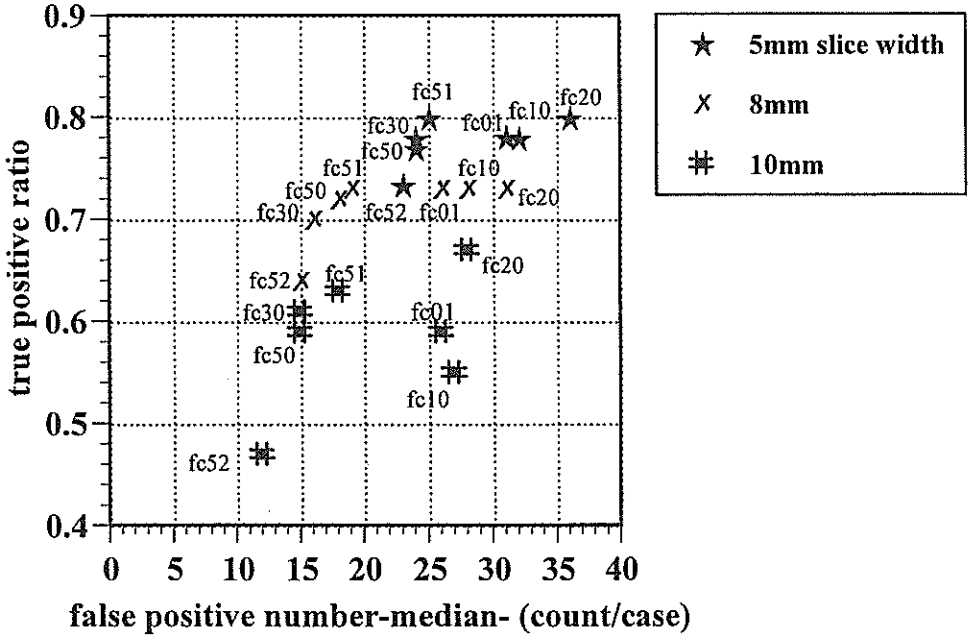


Figure7. FP-TPR diagram per study.

The difference of the slice thickness didn't show so much difference. The FPR in a study shows slight increase in thinner slice thickness.

In the Figure 6 the data 10mm slice thickness distributed in right-lower while that of 5mm distributed left-high position in the diagram. This picture will be convenient to show the relation between CAD and CT image quality characteristics at a glance.

FP-TPR diagram per slice show that the kernel FC52 demonstrated the low in TPR and FPF. It is also demonstrated that the kernel FC20, FC10 or FC01 was relatively high in TPR and FPF. FC51, FC50 or FC30 was distributed in the low FPF and high TPR. On the analogy of ROC curve the low FPF and the high TPR image will be desirable, so in this CAD the kernel FC51 or FC30 will be show the good performance. FC50 also shows a good. As the slice thickness become smaller the FPF-TPR characteristics showed better performance. And these data also shows the robust characteristics for the CT image quality for the CAD system.

The measured LSF of the kernels were correlated to the FPF, that is the higher the spatial resolution resulted to the lower FPF. It was resulted that the TPR was high in the moderate high resolution kernels.

Thinking about the actual use of CAD system for low dose CT lung cancer screening, there are many CT manufacturer, and father more the same manufacturer have very many kind of reconstruction kernels. As we shown in this paper 7 kinds of kernels are evaluated, but there are more many kinds in the kernels. So in using CAD system we should evaluate of the FT-TPR CAD performance between CAD system and CT image quality.

We also think FT-TPR diagram is useful for investigating the CAD development in its performance. The figure 7 shows that there are relatively small valiance for CT slice thickness. That is the numbers of false positive would not largely change for the difference of slice thickness or CT image numbers. This fact also shows some particular characteristics of the CAD system. There are many kinds of CAD algorisms for low dose CT lung cancer screening. So in the future walk we would evaluate the other CAD system by FPF-TPR diagram using the multi-reconstruction low-dose CT screening data sets. We are expecting that the FPF-TPR diagram for the

4. CONCLUSION

We described a method of the performance evaluation of computer-aided diagnosis system for detecting pulmonary nodules for various CT image qualities form reconstructions. The low-dose lung cancer multi-slice CT data sets with many reconstructing conditions suggested to useful on this evaluation. The FPF-TPR diagram will be convenient to describing the relationship between CAD and CT image reconstructing conditions.

ACKNOWLEDGMENTS

The authors are grateful to Masahiro Koyama and respective radiological technologists in the Azumi General Hospital and Y. Kose, a student of Niigata-University in the processing data.
This study was supported in part by a Grant-in-Aid of Cancer Research (15-25) from the Ministry of Health, Labor and Welfare, and in part by a Grant-in-Aid for Scientific Research on Priority Areas (15070205) from the Ministry of Education, Science, Sports and Culture in Japan. This study was also supported by joint studies by the Niigata University and Fujitsu Limited.

REFERENCES

1.Sone S, Takashima S, Li F, et al. Mass screening for lung cancer with mobile spiral computed tomography scanner, Lancet, 351:1242-1245, 1998
2.Henschke CI, MacCauley DI, Yankelevitz DF,et al. Early Lung Cancer Action Project :overall disign and findings from baseline screening, Lancet, 354:99-105,1999
3. Toshioka S, Kanazawa K, Niki N et al. Computer aided diagnosis system for lung cancer based on helical Ct images. Proc SPIE,3034:975-984,1997
4. Reeves AP, Kostis WJ. Computed-aided diagnosis for lung cancer. Radiol Clin North Am,38:497-509,2000
5.Lee Y. Hara T. Fujita H, Itoh S, Ishigaki T. Automated detection of pulmonary nodules in helical CT image based on an improved template-matching technique. IEEE Trans Med Imaging,20:595-604,2001
6.Armato SG III, Li F, Giger ML, MacMahon SG, Sone S, Doi K, Performance of automated CT nodule detection on missed cancer from a lung cancer screening program, Radiology, 225,685-692,2002
7.Kazuo Awai, Kohei Murao, Akio Ozawa, Masanori Komi, Haruo Hayakawa, Shinchi Hori, Yasumasa Nishimura, Pulmonary Nodules at Chest CT: Effect of Computer-aided Diagnosis on Radiologists' Detection Performance, Radiology,230,347-352,2004

8. Arimura H, Katsuragawa S, Suzuki K,et al. Computerized scheme for automated detection of lung nodules in low-dose CT images for lung cancer screening. Acta Radiol, 11:617-619, 2004
9. J. M. Boone, Determination of the presampled MTF in computed tomography, Med. Phys. 28: 356-360 ,2001.
10. Shinichi Wada, Masaki Ohkubo, T.Matsumoto, Measurements and precisions of point spread function of multislice CT, Physics of Medical Imaging ,Proc. SPIE , 5745: 1209-1216, 2005,

A proposal for a diagnosis-dynamic characteristic (DDC) model describing the relation between search time and confidence levels for a dichotomous judgment, and its application to ROC curve generation

Toru Matsumoto[*a], Nobuo Fukuda[a], Akira Furukawa[a], Koji Suwa[b], Shinichi Wada[c], Mitsuomi Matsumoto[d], Shusuke Sone[e]
[a]National Institute of Radiological Sciences; 9-1,Anagawa-4-chome, Inage-ku, Chiba-shi, 263-8555, JAPAN
[b]Koji Suwa (School of Density at Tokyo, Nippon Dental University, 1-9-20, Fujimi, Chiyoda-ku, Tokyo-to, 102-8159, JAPAN
[c]Niigata University;746 2-bancho, Asahimachi-dori, Niigata-shi, 951-8518, JAPAN
[d]Diichi Hospital; 1227, Shimokotori-cho, Takasaki-shi, Gunma-ken, 370-0074, JAPAN
[e]JA Azumi General Hospital; 3207-1 Ikeda-ooaza, Ikeda-cho, Kitaazumi-gun, Nagano-ken, 951-8518,JAPAN

ABSTRACT

When physicians inspect an image, they make up a certain degree of confidence that the image are abnormal; p(t), or normal; n(t)[n(t)=1-p(t)]. After infinite time of the inspection, they reach the equilibrium levels of the confidence of $p^*=p(\infty)$ and $n^*=n(\infty)$. There are psychological conflicts between the decisions of normal and abnormal. We assume that the decision of "normal" is distracted by the decision of "abnormal" by a factor of $k(1 + ap)$, and in an inverse direction by a factor of $k(1 + bn)$, where $k (> 0)$ is a parameter that relates with image quality and skill of the physicians, and a and b are unknown constants. After the infinite time of inspection, the conflict reaches the equilibrium, which satisfies the equation, $k(1 + ap^*)n^* = k(1 + bn^*)p^*$. Here we define a parameter C, which is $2p^*/[p^*(1 - p^*)]$. After the infinite time of inspection, the conflict reaches the equilibrium, which satisfies t that changes in the confidence level with the time (*dp/dt*) is proportional to $[k(1+ap)n - k(1+bn)p]$, i.e. $k[-cp^2 + (c - 2)p + 1]$. Solving the differential equation, we derived the equation; t(p) and p(t) depending with the parameters; k, c, S. S (0-1) is the value arbitrary selected and related with probability of "abnormal" before the image inspection ($S = p(0)$).

Image reading studies were executed for CT images. ROC curves were generated both by the traditional 4-step score-based method and by the confidence level; p estimated from the equation t(p) of the DDC model using observed judgment time. It was concluded that ROC curves could be generated by measuring time for dichotomous judgment without the subjective scores of diagnostic confidence and applying the DDC model.

Keywords: DDC model, judgment time, confidence level, ROC curve, dichotomous judgment

1. INTRODUCTION

Medical judgments based on image diagnosis involve a degree of uncertainty, which is derived from various factors such as the characteristics of the image and the skill of the physician. When interpreting images, physicians are under pressure to make a decision in a state of uncertainty that varies in type and degree. In other words, when the physicians have to choose one diagnosis from two or more possible, yet contradictory diagnoses, they are either caught in a dilemma because the factors that support each possible diagnosis conflict with each other; or they are moving toward or away from a particular diagnosis[1].

In ROC analysis, the "confidence level" is used as a measurement for evaluating the physician's interpreting capability or accuracy in diagnosing various medical images. The confidence level of the physician's own evaluation of the accuracy of the response (here, the judgment made at the conclusion of image diagnosis) made by the physician. Here, the

Medical Imaging 2006: Image Perception, Observer Performance, and Technology Assessment, edited by Yulei Jiang, Miguel P. Eckstein, Proc. of SPIE Vol. 6146, 61460Y, (2006) · 1605-7422/06/$15 · doi: 10.1117/12.652933

confidence level is considered to be a data point that represents an introspective fact. In contrast, the time from presentation of the stimulus (the image) until the judgment is made (judgment time) is considered to be an objective indicator that corresponds to this level of confidence. In the image diagnosis scenario, the time to search the image required to make a decision corresponds to this judgment time. The fact that the judgment time and the confidence level are correlated has been known for a long time in the field of psychology[2-6].

In the present report, when the physician has responded at a constant confidence level, a constant relationship is established with the judgment time required to reach that certainty, via the "conflict" identified at the beginning of this paper. This is the psychological model (known as a diagnosis-dynamic characteristic (DDC) model[7]) proposed in this paper. This fact is also confirmed by an observer performance study. Moreover, assuming the generalizability of the judgment time-confidence level relationship inferred from a dichotomous judgment-based DDC model, the confidence level judged by the physician is inferred from the observed judgment time. From this, a method for generating ROC curves from the reader's 0 and 1 judgment is proposed.

2. DDC MODEL

We present an outline of the proposed model below. The basis of the image diagnosis made by the physician is a judgment of 0 or 1 (for example, normal or abnormal, benign or malignant). A slightly more detailed judgment may involve a response using a confidence level in the range between 0 and 1. This model presupposes that after the image is presented to the physician, an image-searching time required to respond with a confidence level between 0 and 1 will be observed. This time is assumed to reflect the intensity of the conflict experienced by the reader in reaching a judgment of 0 or 1, as a result of the image or the reader's skill, the difficulty of diagnosing the image, and the suppositions (or preconceptions) held by the reader beforehand. The state of conflict that incorporates these factors is shown by the 2-compartment model in Figure 1.

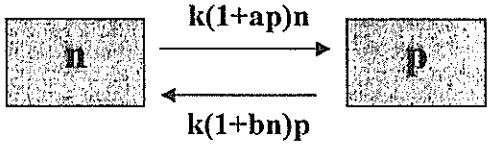


Fig.1 Two compartment DDC model

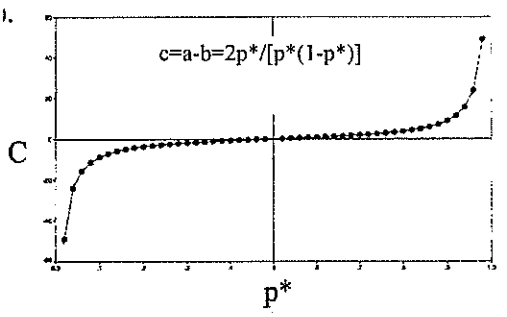


Fig.2 Relationship between the confidence level p*, reached after an infinite inspection time, and the induction coefficient c

When physicians inspect an image, or a set of images, they decide with a certain degree of confidence, or probability, that the image(s) are abnormal; $p(t)$, or normal; $n(t)$ [$n(t) = 1 - p(t)$].

After an infinite inspection time, they approach the equilibrium confidence levels p^*= p(∞) and n^* =n(∞) ($n^* = 1 - p^*$).

We assume that there are psychological conflicts between the decisions of normal and abnormal and the decision of "normal" is distracted by the decision of "abnormal" by a factor of $k(1 + ap)$, and in an inverse direction by a factor of

$k(1 + bn)$, where k (> 0) is a parameter that relates to image quality and skill of the physician, and parameters a and b are unknown constants.

Thus, the degree that the decision of "normal" is influenced by the decision of "abnormal" is $k(1 + ap)n$, and the degree in the inverse direction is $k(1 + bn)p$.

During the conflict between a decision on normal or abnormal,

$k(1 + ap)n \neq k(1 + bn)p$

After an infinite inspection time, the conflict will approach the equilibrium level, which satisfies the equation:

$k(1 + ap^*)n^* = k(1 + bn^*)p^*$.

Herewith, we define the parameter

$c \equiv a-b=2p^*/[p^*(1-p^*)$. $P^*=1/2$; $c=0$, $p^*>1/2$; $c>0$, $p^*<1/2$; $c<0$ (1)

Fig. 2 shows the relationship between the induction coefficient "c" and the final confidence level "p*". The induction coefficient c represents the ease (or difficulty) of diagnosis for each image. It shows that the closer c is to 0, the diagnosis of the image becomes more difficult, and the further from 0, the diagnosis becomes easier.

The variations in p as the conflict approaches equilibrium are expressed by the differential equation shown below. In brief, the degree of conflict of the "abnormal" confidence level with time (dp/dt) is proportional, as follows.

$dp/dt = k(1+ap)n - k(1+bn)p = k(-cp^2 + (c-2)p + 1)$

Solving the differential equation as follows,

$\int dp/[-cp^2 + (c-2)p + 1] = \int kdt + K$

and solving the indefinite integral yields the following.

$1/(\sqrt{(c^2+4)})[\ln((\sqrt{(c^2+4)})+2cp-(c-2))/((\sqrt{(c^2+4)})-2cp+(c-2))] = kt + K$ (2)

(NB: for the above indefinite integral, there is one other solution, but this is omitted, and in the above case, leads to the following.)

K is the integral constant for the initial condition t=0.

$K = 1/(\sqrt{(c^2+4)})[\ln(\sqrt{(c^2+4)}+2cS-(c-2))/((\sqrt{(c^2+4)})-2cS+(c-2))]$ (3)

The variable p in Equation (2) is substituted with S (supposition). S represents the baseline confidence level p (t=0) between 0 and 1 from which the reader commences searching the image (judgment). For example, S=0.5 when commencing judgment from a confidence level of 0.5, and S=0 when commencing judgment from a confidence level of 0. The value is arbitrarily selected in the range $0 \leq S \leq 1$, and relates to the supposition S=p(0), which is equivalent to the confidence level of the supposition before image inspection by a reader.

From Equation (2), we obtain t as a function of p.

$t(p) = (1/k)(1/(\sqrt{(c^2+4)}))[\ln((\sqrt{(c^2+4)})+2cp-(c-2))/((\sqrt{(c^2+4)})-2cp+(c-2))] - K$ (4)

where, p=hp, 0<h<1* (5)

Next, solve Equation (4) for p. In brief,

Substitute Equation (3) in (4). For simplicity, let $\gamma = \sqrt{(c^2+4)}$, and g=c-2,

$\gamma kt = ln(\gamma+2cp-g)/(\gamma-2cp+g) - ln(\gamma+2cS-g))/(\gamma-2cS+g)$

$= ln(\gamma+2cp-g)/(\gamma-2cp+g)/((\gamma+2cS-g))/(\gamma-2cS+g))$

$exp(\gamma kt) = (\gamma+2cp-g)/(\gamma-2cp+g)/((\gamma+2cS-g))/(\gamma-2cS+g))$

From this, for simplicity, let f=exp(γkt), and obtain p as a function of t.

$$p(t)=[-\gamma^2(1-f)+2\gamma cS(1+f)-2gcS(1-f)+g^2(1-f)] / [2\gamma c(1+f)-4c^2S(1-f)+2gc(1-f)] \quad (6)$$

3. INFERENCES BASED ON THE DDC MODEL

Fig. 3 shows examples of p-t and n(=1-p)-t curves based on Equation (6) of the DDC model.

Reading commences from S=p(0)=0 (supposition: normal) at starting time t=0, and the state of internal conflict as to whether the image is normal or abnormal is gradually settled, and over an infinite judgment time, the decisions of n and p reach equilibrium (k=1.0) when the confidence level p for an abnormal finding is 0.5 (the confidence level n of normal is 0.5).

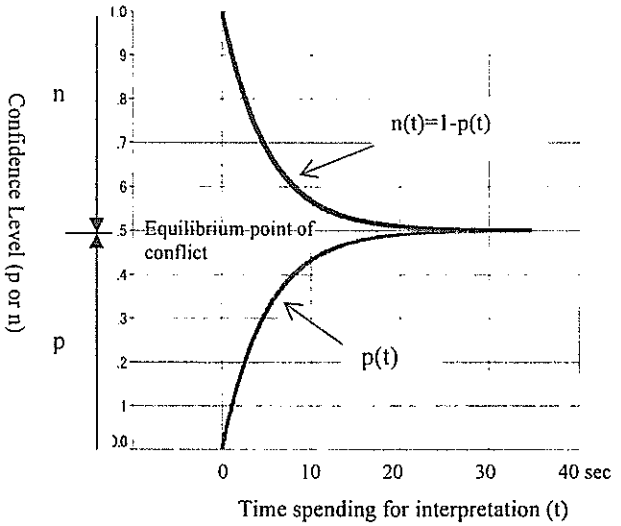


Fig.3 A diagnostic probability (p or n) – interpretation time (t) relation based on the formula (6) of the DDC model; Supposition S=p(0)=0.0 for "p" (for n, S=1.0) from t=0 to infinite, up to p*=0.49999

Fig. 4 shows the family of p-t curves generated to reach various confidence levels p while the reader is conflicted in the range between 0 and 1 over an unlimited judgment time, for initial conditions of time t=0 and S=0, with the results derived from Equation (6), similar to those in Fig. 3. The ratios of the confidence levels for p* and n* reach equilibrium over an unlimited judgment time. (The n-t curves are not displayed, k=1.0)

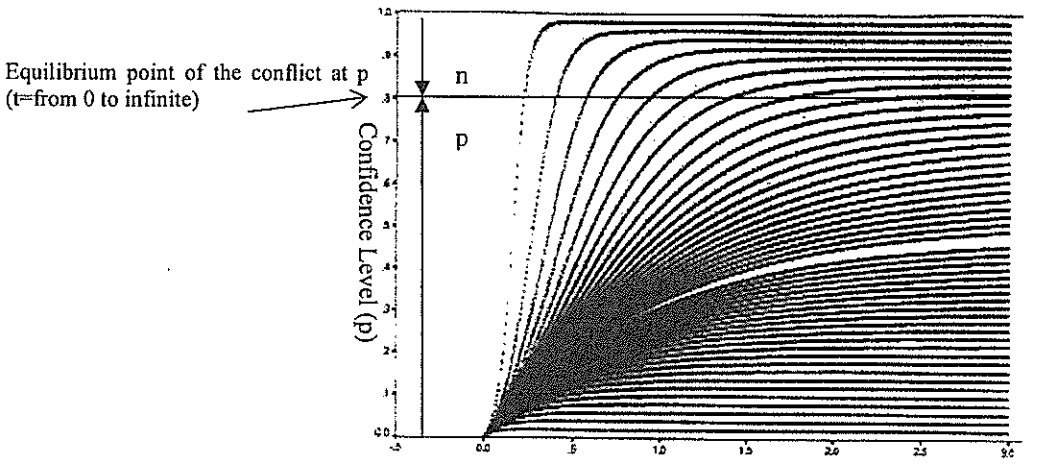


Fig.4 P-t curves from formula (6) in the case of the various p (t=0 to infinite) when interpretation time is unlimited; Supposition S=p (0)=0.0 for "p" (for n, S=1.0)

Fig. 5 shows the family of t-p curves corresponding to Fig. 4, derived from Equation (4) of the DDC model. In all p-t curves shown in Fig. 4, it is assumed that at the time at which the confidence level p* over unlimited judgment time (h in Equation (5) = 0.99), the reader suspends the psychological conflict and produces a medical decision. It can be inferred that the overall tendency of t for each of the curves is to have a bell-shaped relationship with p.

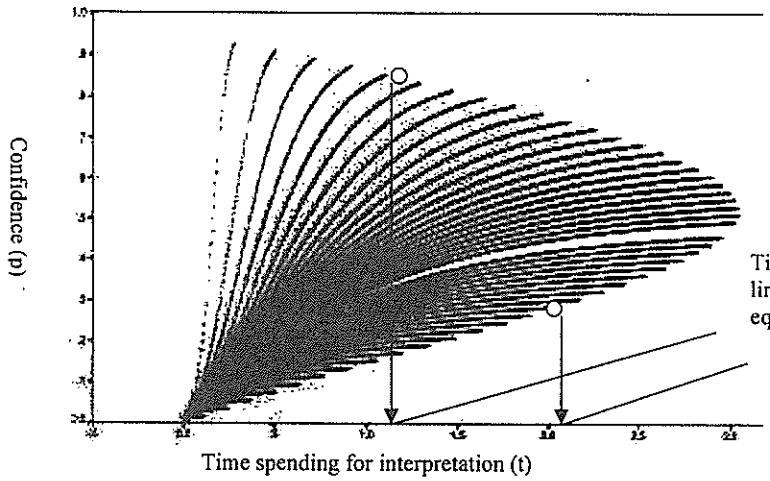


Fig.5 p-t curves based on the formula (4) when interpretation time is truncated and forced to report a diagnosis at the time when t reaches to p=0.99p* ,Supposition S=p(0)=0.0

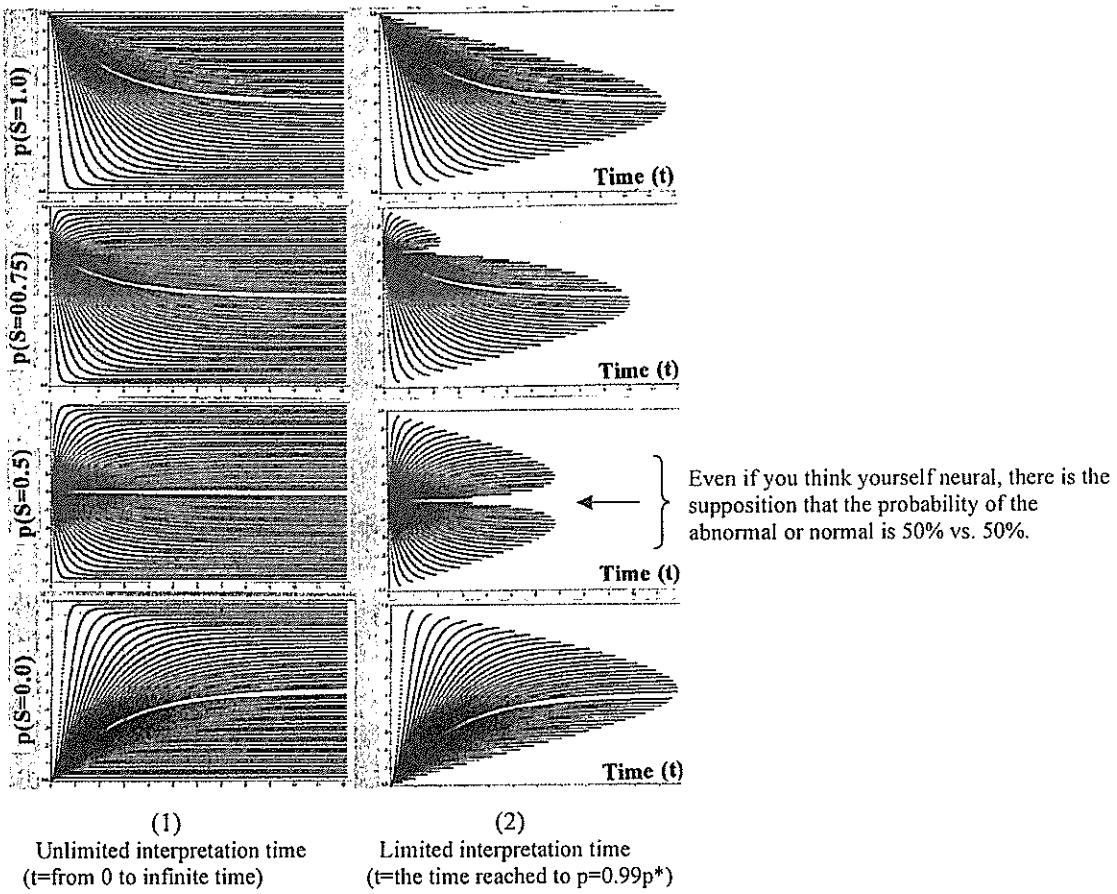


Fig.6 Aspect of p-t curves as a function of the supposition ; S , From these figures, we provide a rule!, "No judgment is free from supposition""

Fig. 6 shows the family of curves generated when the judgment commences from S at time t=0, and the corresponding family of t-p curves (t until p=0.99p* is reached), in addition to Fig. 5 inferred from the DDC model. The t of the t-p curve decreases in the vicinity of S. The tendency of the t-p curve for S=1 (time t=0, supposition abnormal) is the same as that for S=0 (time t=0, supposition normal; Fig. 5), and is inferred to show a bell-shaped relationship.

4. OBSERVER PERFORMANCE STUDY

4.1 Methods

The following observer performance study was conducted with the aim of verifying the DDC model. The image database consisted of CT films of the mandibular anterior dentition of 20 patients (10 with abnormal findings and 10 without). The digitized film was designated as the original image (Image quality 1). As described below, the quality was then successively artificially degraded to prepare four sets. Image quality 2 is a two-fold enlarged image after the pixels of image quality 1 were thinned to 1/2. Similarly, in image quality 3, 4, and 5, the pixels of image quality 1 were thinned to 1/4, 1/8, and 1/16 respectively, after which the images were enlarged 4, 8, and 16-fold respectively. An example is

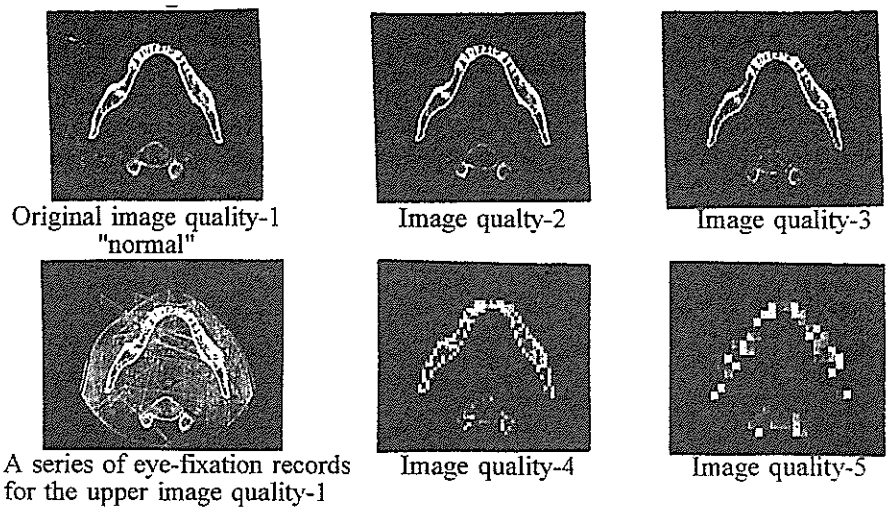


Fig.7 Example of images used in the observer performance study (Image quality 1-5)

shown in Fig. 7. These 100 images were presented randomly on a CRT monitor, one by one, and read by 8 dentists from a distance of 80.9±2.4 cm. During the reading, the dentist's line of gaze and pupil diameter were measured by a contactless eye tracking system (Model 504; ASL). The judgment time from the presentation of each image on the CRT monitor until the diagnostic decision by the reader was measured (image search time) from the collected line of gaze data. The readers had image reading experience that ranged from 1 to 22 years, and their professional discipline was either dental radiology or oral surgery. The readers reported the diagnostic results according to a 4-step scoring scale: Normal (1), Probably normal (2), Probably abnormal (3), and Abnormal (4)[8].

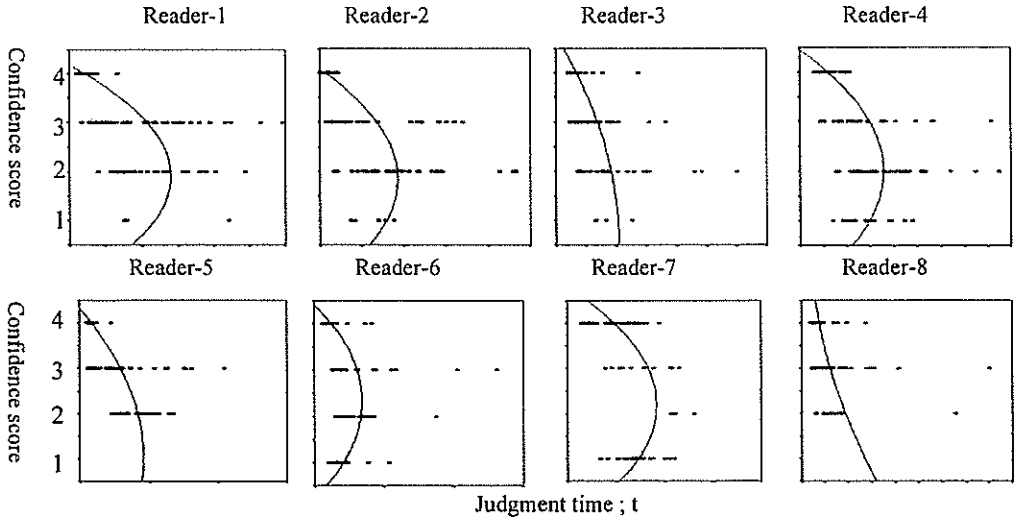


Fig.8 Observed data : Relation between 4-step confidence level score and judgment time t

4.2 Results

Fig. 8 shows the relationship between the judgment time (image-searching time) for each reader measured from the line of gaze data collected by the eye tracker and the 4-step confidence scores reported by each reader. A quadratic function has been fitted to the results for image quality 1 through 5 (20 normal and abnormal cases each, a total of 100 cases) to visualize a general trend. Excluding reader 8, the relationship between judgment time and confidence level describes a bell shape.

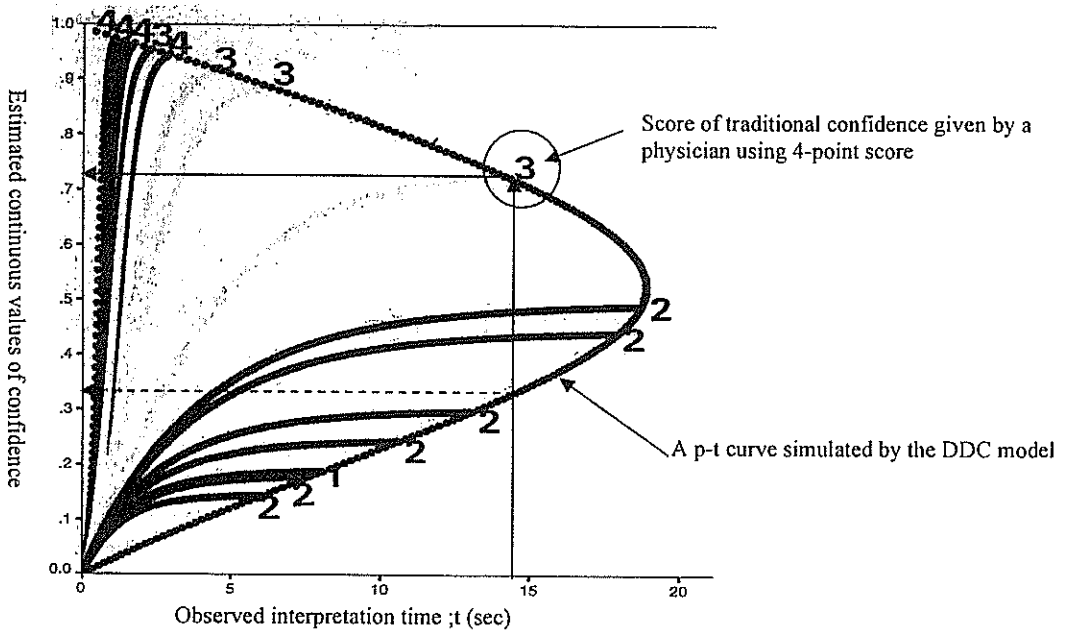


Fig.9 Method of estimation of the confidence level by the DDC model in the case of the reader-1

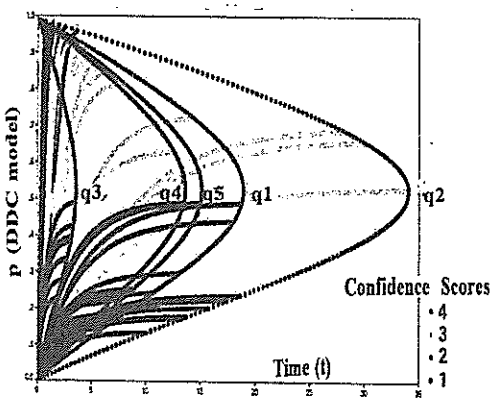
5. ESTIMATION OF CONFIDENCE LEVEL USING THE DDC MODEL AND ITS APPLICATION TO ROC CURVE GENERATION

Fig. 9 shows the result of fitting a t-p curve derived from the DDC model with initial conditions time t=0 and S=0 to the judgment time t observed for image quality 1 (10 normal and 10 abnormal cases) by reader 1. Here, the t-p curve of Fig. 9 is assumed to be generalizable. In other words, it is assumed that the bell-shaped curve relationship between the confidence level p and judgment time t for the overall image database for interpretation will be generated for each image quality. It is also assumed that the judgment time t is always measured accurately.

When fitting the t-p curve, the method for estimating the confidence level p from the observed interpretation time t is as follows.

First, when the maximum value of the judgment time t is h=0.99 and p*=0.5, the induction coefficient (k) of the t-p curve is adjusted so that the confidence level p=0.99×0.5=0.495. Next, as shown in Fig. 9, the confidence level p on the t-p curve corresponding to the observed interpretation time t is estimated. At this time, the estimated confidence level p corresponding to the same observed time t on the bell-shaped t-p curve, as can be seen in Fig. 5 or Fig. 9, may produce two values either side of 0.5. For an FP judgment in which definitive diagnosis of normal is diagnosed as abnormal and TP case in which an abnormal case is classified as abnormal, a confidence level p of >0.5 is assigned. For a TN case in which a definitive diagnosis of normal is assessed as normal, and a FN case in which abnormal is assessed as normal, a confidence level p of <0.5 is assigned.

(1) Application of DDC model to image quality 1-5 for reader 1



(2) Relation between estimated confidence level and observed confidence score

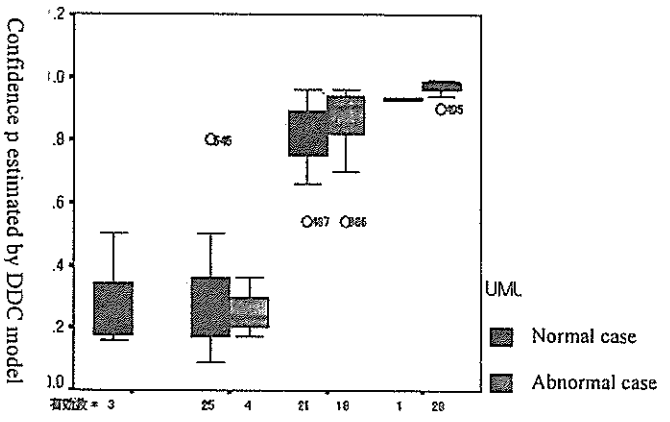


Image quality	1	2	3	4	5
Normal	10.4±5.5	10.7±6.3	8.9±3.0	10.8±5.7	21.3±17.1
Abnormal	2.1±1.6	2.6±3.2	3.4±3.9	6.7±4.6	14.7±8.0
Average	6.2±5.8	6.7±6.4	6.1±4.4	8.8±5.5	18.0±13.4

(3) Observed judgment time (sec)

Fig.10 Fitting of observed judgment time for the t-p curve by DDC model (1), relationship between the estimated confidence and the observed (2), and observed judgment times for the normal and the abnormal cases.

Figure 10 (1) shows the result of fitting observed judgment times t for reader 1 at image quality levels 1-5 (50 normal and 50 abnormal cases) to the t-p curves derived from the DDC model. It shows that the longest judgment time was observed for 1 case of image quality 2 and the shortest time on the whole was the judgment time for image quality 3. Fig. 10 (2) shows the correlation (r=0.865) between the estimated confidence levels p and the confidence levels reported by the reader according to the 4-step score. Fig. 10 (3) shows the judgment times t observed for reader 1 for normal and abnormal cases at each image quality. There was a tendency for the judgment time for normal cases to be longer than that for abnormal cases. The mean judgment times for image quality levels 1, 2, and 3 were approximately comparable, but the judgment times for image quality levels 4 and 5 were clearly longer.

Figure 11 (1) is an ROC curve drawn on the basis of the estimated confidence levels p for image quality levels 1-5 interpreted by reader 1. The symbol (dark black) represents the ROC curve drawn according to the traditional (normal) method (confidence level according to a 4-step score). Each operating point (cumulative number of FP, cumulative number of TP; dilute symbol) on the ROC curve estimated by this method has been interpolated between the on the normal ROC curve. The square symbols in Fig. 11 (2) represent operating points when normal or abnormal is judged as 0 or 1 (divided into two groups, a score of 2 or lower, and a score of 3 or higher).

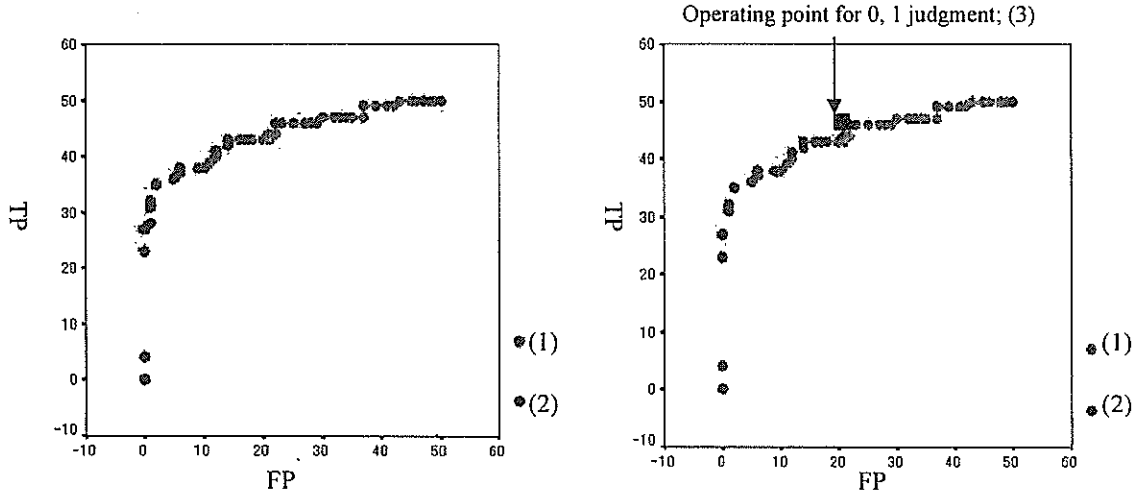


Fig.11 Estimated ROC curve; (1)drawn from the results of Fig. 10 (1) for Reader 1, the traditional ROC curve; (2) by 4-step confidence scores (Image quality 1-5; normal 50 cases, abnormal 50 cases) and (3) the case of 0,1 judgment

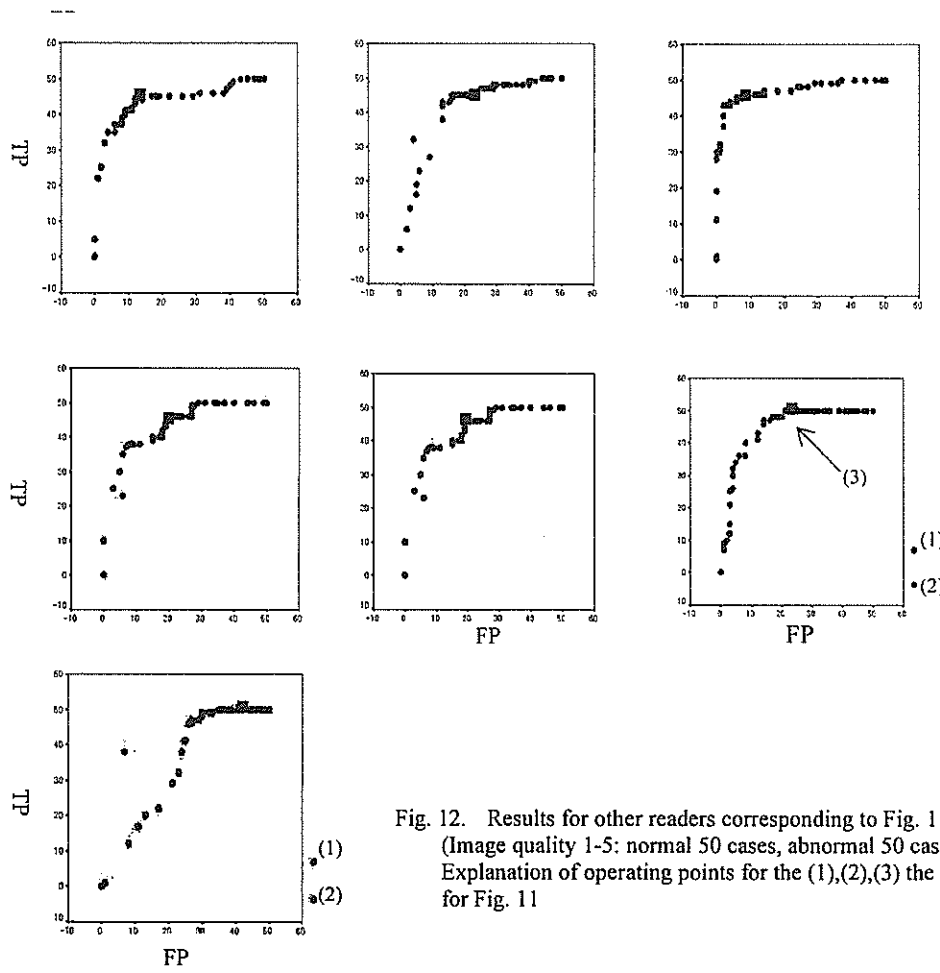


Fig. 12. Results for other readers corresponding to Fig. 11 (Image quality 1-5: normal 50 cases, abnormal 50 cases) Explanation of operating points for the (1),(2),(3) the same as for Fig. 11