画像通信

Vol. 12 No. 2 (通巻23)

目　　次

1989年10月

社団法人　日本放射線技術学会
画　像　部　会

〒604　京都市中京区西ノ京北壷井町88
二条プラザ内

第27回　画像部会予告

日　　時　昭和64年4月4日（水）

場　　所　神戸ポートピアホテル

テ ー マ　画像の視覚評価について（予定）

画像部会役員氏名

○会　長	山　下　一　也	大阪大学医療技術短期大学部
	伊　藤　　博	兵庫医科大学病院中央放射線部
	稲　津　　博	宮崎医科大学附属病院放射線部
	大　塚　昭　義	山口大学医学部附属病院放射線部
	小　寺　吉　衛	広島大学歯学部歯科放射線学教室
○編　集	小　水　　満	滋賀医科大学附属病院放射線部
	佐々木　常　雄	名古屋大学医療技術短期大学部
○庶　務	滝　川　　厚	大阪大学医療技術短期大学部
	滝　沢　正　臣	信州大学医学部放射線医学教室
○庶　務	段　床　嘉　晴	大阪大学医学部附属病院
○企　画	津　田　元　久	島津科学技術振興財団
○財　政	畑　川　政　勝	大阪市大病院中央放射線部
	藤　田　広　志	岐阜工業高等専門学校電気工学科
○総　務	若　松　孝　司	国立循環器病センター

第26回　画像部会のご案内

日　　時：1989年10月13日(金)　10:00 ～ 15:45

場　　所：札幌市教育文化会館　第2会場(小ホール)

プログラム：

教育講演　10:00 ～ 11:30

座長　大阪市立大　畑　川　政　勝

「コンピュータ支援診断の実際：胸部間質性浸潤の
定量的特徴抽出と検出」

岩手医科大学　桂川 茂彦 先生

教育講演　13:00 ～ 14:15

座長　大阪医療短大　山　下　一　也

「ROC解析の基礎について」

シカゴ大学　Charles E. Metz　教授

画像について語ろう　14:30 ～

研究発表

1. 空間周波数領域における放射線画像処理
（画像端に生じる周期断裂が周波数フィルターに及ぼす影響）

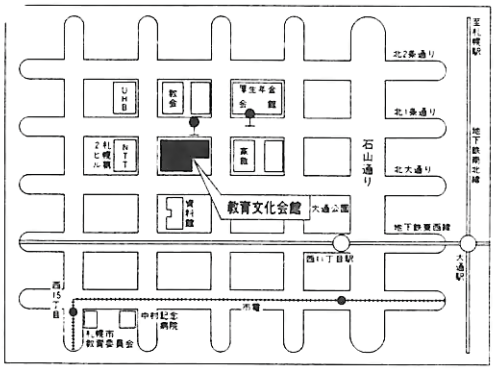
北大病院放射線部　仲　知　保

2. 骨梁像のディジタル解析

大阪中央病院　石　田　隆　行

大阪医療技術短大　山下一也・滝川　厚

会場案内



コンピュータ支援画像診断の初歩から理論、実際まで

岩手医科大学 放射線医学教室　桂　川　茂　彦

肺の間質性病変は、異常陰影を有する胸部写真の中で比較的良く見られる病変であるにもかかわらず、その診断は非常に困難であると言われている。その理由は、肺野テクスチャーが一般に不明瞭なパターンであること、そのパターンの変化が複雑であること、さらに、それらのパターンを記述する言葉と名称が主観的なものであり、客観的に定義されていないことなどである。したがって、間質性陰影と関係する肺野テクスチャーを定量的に分析し、放射線科医に客観的判断のための材料を提供すれば、診断の主観性が減少し診断精度が向上すると考えられる。

そこで、我々はディジタル胸部写真における間質性疾患の特徴抽出と検出を目的として、物理的尺度（texture measure）を用いて肺野テクスチャーを計測する方法を開発した。ここでは、コンピュータ支援診断の基本的な考え方、肺野テクスチャー解析法、正常肺と間質性陰影を持った異常肺の自動分類法などについて述べ、臨床応用上の有用性について解説する。

1。コンピュータ支援診断 (CAD) とは

コンピュータ支援診断
Computer-Aided Diagnosis (CAD)

ディジタル画像データを
定量化し、これを放射線診断へ
積極的に利用すること。

Fig. 1

CAD 開発の動機

開発の動機

見落としによる誤診を減少
させることと、
主観的判断による思い違いを
防止すること。

Fig. 2

CAD の目的

CAD の目的

放射線科医の診断を助けて、
診断の正確度と診断の
”便利さ”を向上させる。

Fig. 3

(注意) CAD は、放射線科医の役割をコンピュータで置き換えることではなく、放射線科医の診断をコンピュータを利用して援助することにある。したがって、CAD は、従来、自動診断と呼ばれていた方式や概念とは、根本的に異なる思想と方針に基づくものである。

CADの一般的手法

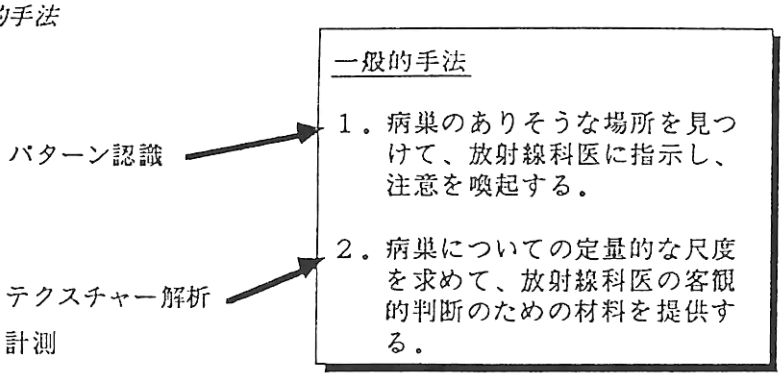


Fig. 4

対象画像

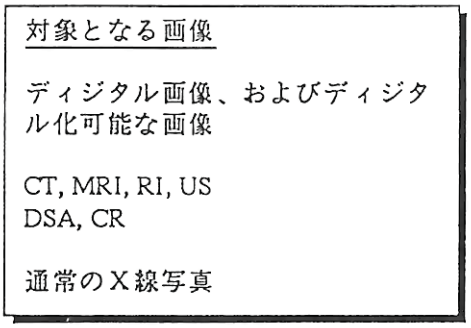


Fig. 5

情報の流れ

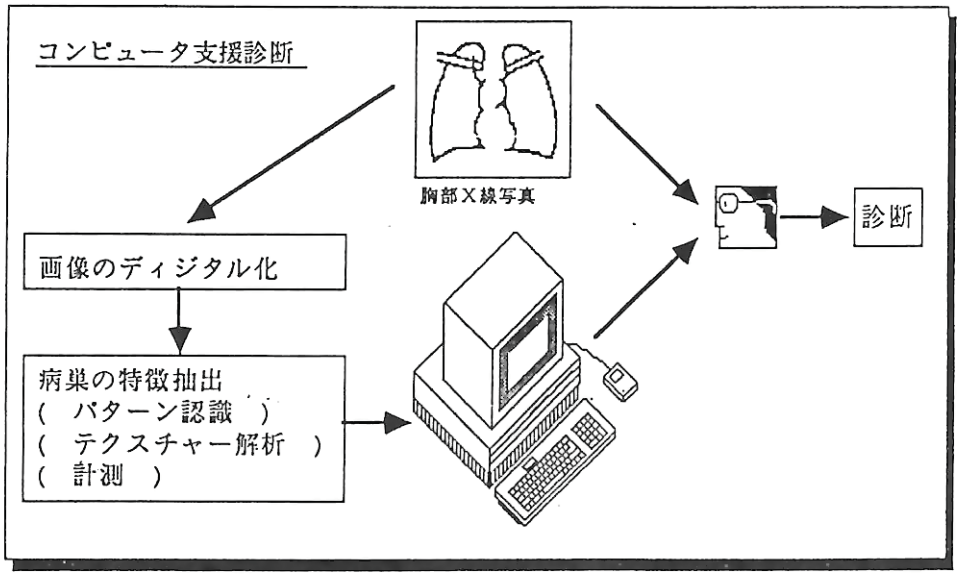


Fig. 6

CADの応用

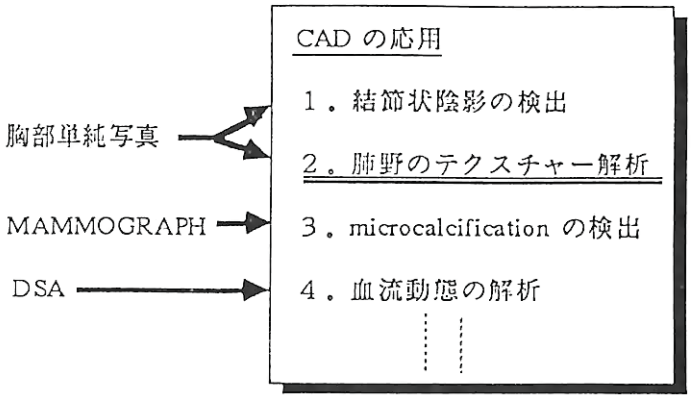


Fig. 7

2．胸部X線像における肺野テクスチャーの定量的解析

なぜ間質性陰影か?

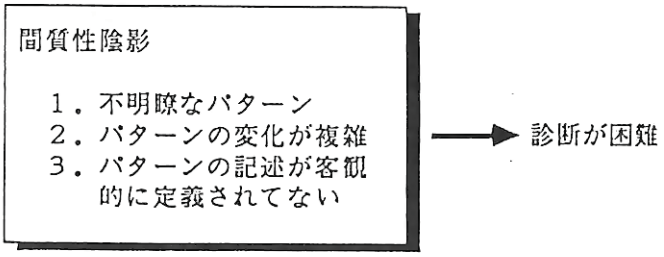


Fig. 8

客観的尺度

肺野テクスチャーを表現する物理量 (texture measure)

テクスチャー.....規則的あるいは不規則な模様

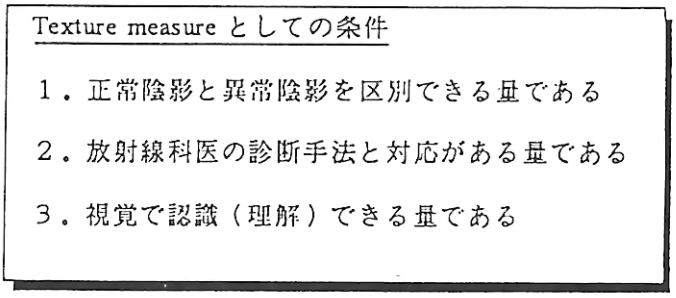


Fig. 9

X線写真上での間質性陰影の特徴

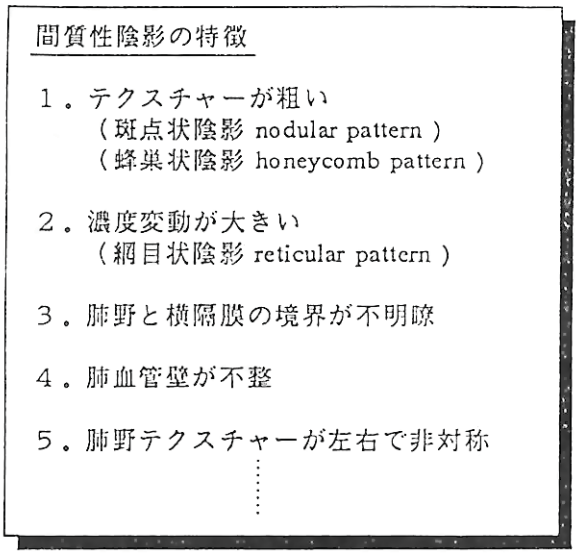


Fig. 10

定量化の例

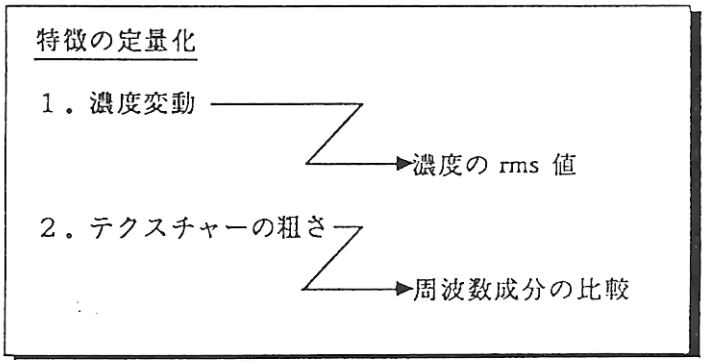


Fig. 11

CADの出力

肋骨、肺血管、心臓を検出して適当な部位の測定を行ない、肺野テクスチャー解析の結果と一緒に、エキスパート・システムなどの人工知能的手法を用いて、正常肺と異常肺の確率の推定を行なう。

３．肺野テクスチャー解析法

３．１　概要

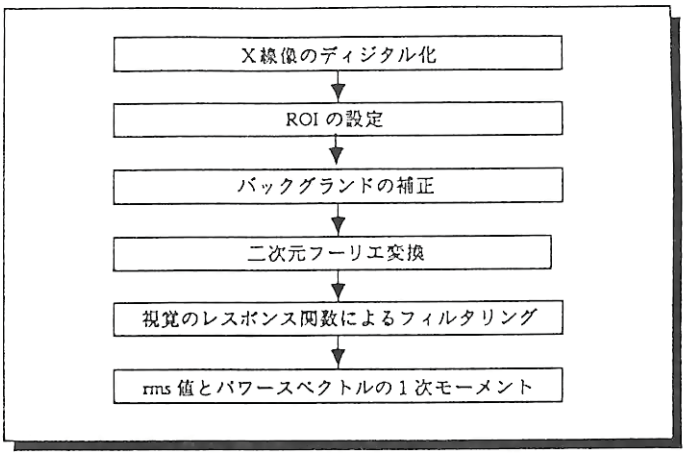


Fig. 12

３．２　テクスチャー解析に必要なシステム

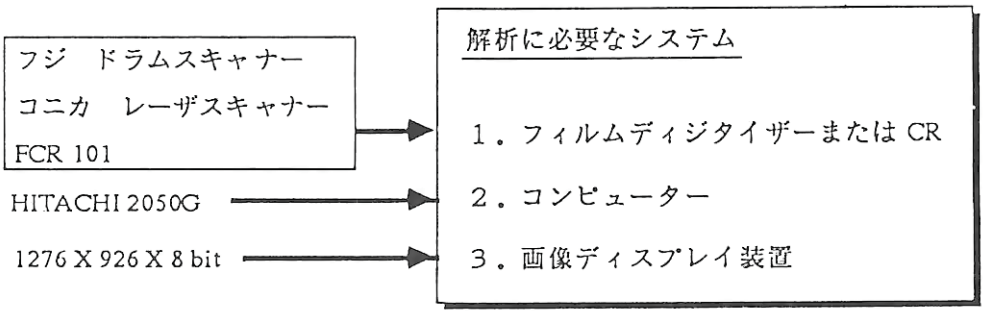


Fig. 13

コンピュータのメインメモリー.....1 MB 以上

ディスク容量.....40 MB 以上が望ましい。

その他、CT, MRI, DSA, RI などのコンピュータの利用が考えられる。

３．３　Ｘ線像のディジタル化

ディジタル化における因子

１．ピクセル（画素）寸法
（サンプリング・アパーチャー）

２．サンプリング間隔

３．グレイレベル数

Fig. 14

通常は、ピクセル寸法とサンプリング間隔は等しいことが多い。

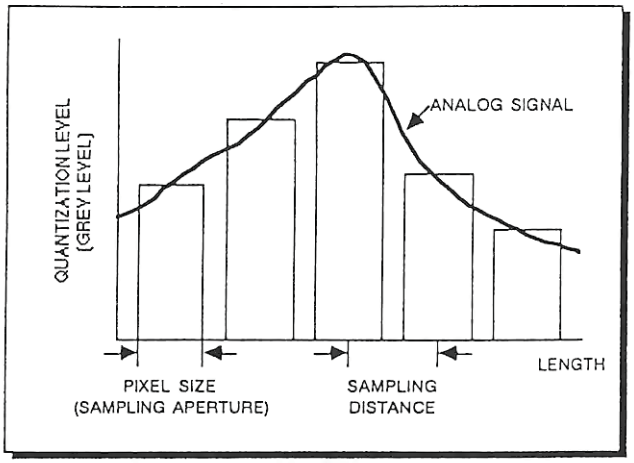


Fig. 15　土井邦雄：ディジタルラジオグラフィの基礎と将来の可能性．日本医放会誌、４９、１－１４、１９８９

	ピクセル寸法 (mm)	グレイレベル数 (bits)	画素数 (matrix size)	容量 (MB)
フジ　ドラムスキャナー	0.1	10	3550 X 3550	25.2
コニカ　レーザスキャナ	0.175	10	2000 X 2000	8.0
FCR 101	0.2	8	1760 X 1760	3.1

Fig. 16 大角サイズ・フィルムのディジタル化

FCR 磁気テープに対する、画像データの READ/WRITE 法

１．FCR 磁気テープフォーマットの理解（メーカーに問い合わせる）

２．システム・サブルーチンを利用して磁気テープ操作のプログラムを作る。

一般的な磁気テープ・フォーマット

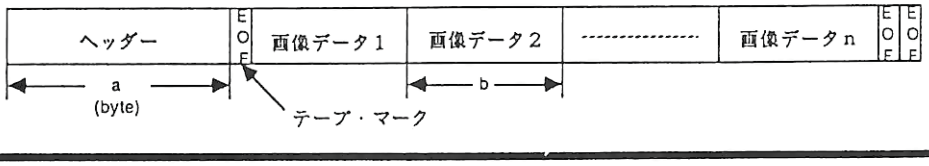


Fig. 17

使用しているOSによって、プログラムの形は異なるが、ここには、最小限必要な機能だけを示した。

必要なサブプログラム

MT_OPEN (DEV) ; 磁気テープ装置のオープン
磁気テープ装置番号

MT_READ (DEV, MAXBUF, BUF, N) ; データの読み込み
実際に読んだバイト数
（０は、テープ・マークの検出）
データを格納する配列
配列の要素数

MT_WRITE (DEV, NWRT, BUF, N) ; データの書き込み
実際に書き込んだバイト数
書き込みたいバイト数

MT_WEOF (DEV) ; テープ・マークの書き込み

MT_CLOSE (DEV) ; 磁気テープ装置のクローズ

Fig. 18

３．４　ROIの自動設定法

肋間の検出

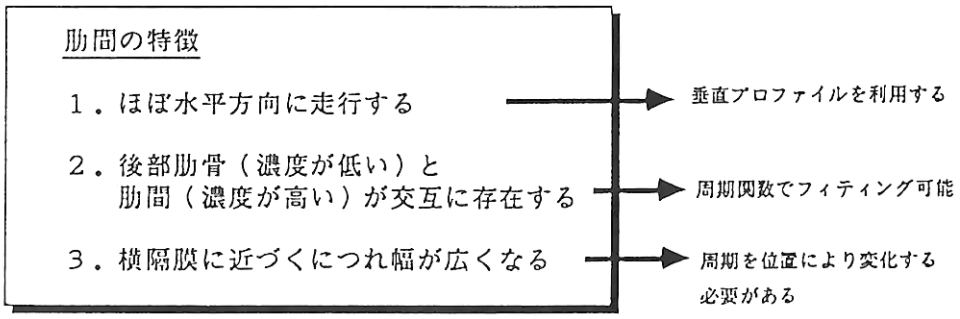


Fig. 19

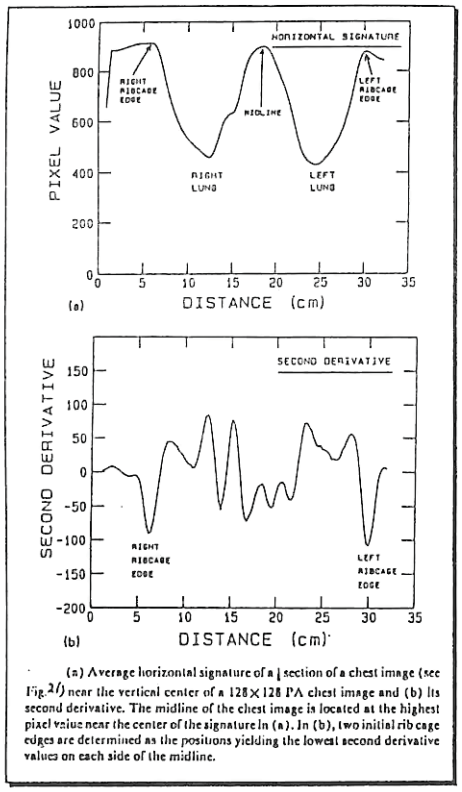
位置によって周期（周波数）が変化する関数

Shift -Variant Sinusoidal Function

$f(x) = A \cos [2\pi u(x)\, x + \phi]$ ；$u(x)$ は位置ｘにおける周波数

肋間の検出に用いる画像の画素数は１２８ｘ１２８でも十分である．

計算時間の短縮



(a) Average horizontal signature of a $\frac{1}{4}$ section of a chest image (see Fig. 21) near the vertical center of a 128×128 PA chest image and (b) its second derivative. The midline of the chest image is located at the highest pixel value near the center of the signature in (a). In (b), two initial rib cage edges are determined as the positions yielding the lowest second derivative values on each side of the midline.

Fig. 20

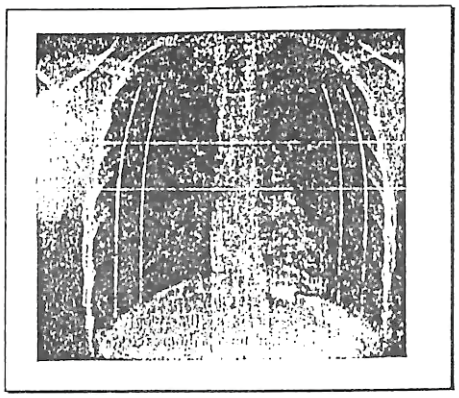
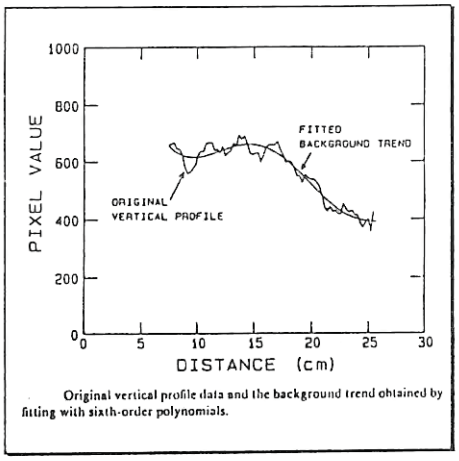
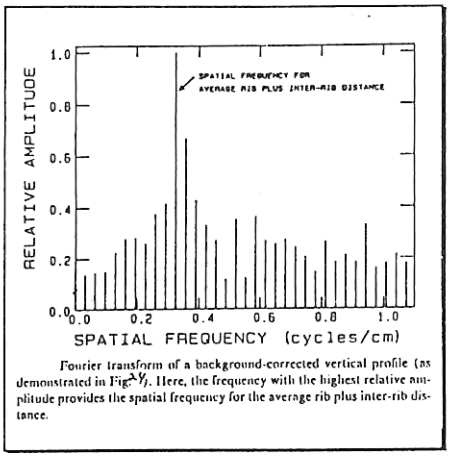


Fig. 21



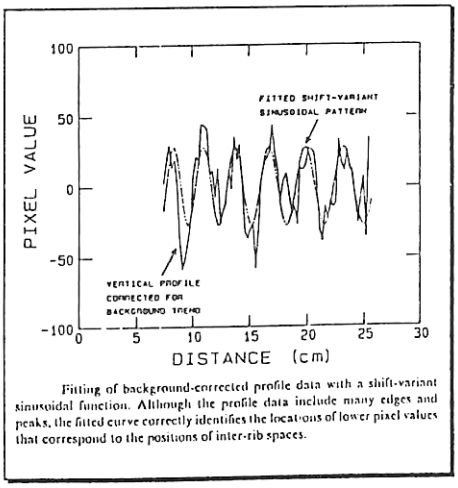
Original vertical profile data and the background trend obtained by fitting with sixth-order polynomials.

Fig. 22



Fourier transform of a background-corrected vertical profile (as demonstrated in Fig. 24). Here, the frequency with the highest relative amplitude provides the spatial frequency for the average rib plus inter-rib distance.

Fig. 23



Fitting of background-corrected profile data with a shift-variant sinusoidal function. Although the profile data include many edges and peaks, the fitted curve correctly identifies the locations of lower pixel values that correspond to the positions of inter-rib spaces.

Fig. 24

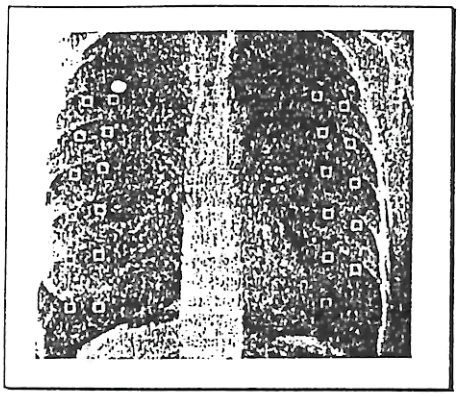


Fig. 25

$f(x) = A \cos [2\pi u(x)x + \phi]$

$u(x) = 1/d(x)$

$d(x)$；位置xでの、肋骨と肋間の幅を合わせた長さ

$d(x) = mx + b$

$b = 1/u_0$　；Fig. 23 から求まる

A；Fig. 22 から求まる

m, ø；少しづつ変化させ、近似式とプロファイルの差が最小の点を探す。

Fig. 26

3．5　ピクセル値の物理的意味

フィルム・スキャナーの場合

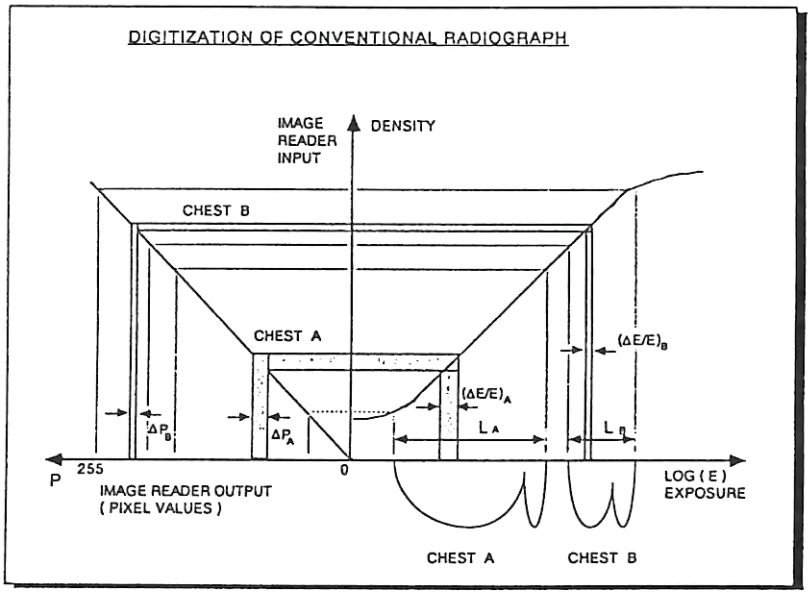


Fig. 27

ピクセル値から濃度への変換係数

特性曲線の傾き

ピクセル値の変動成分

$$\frac{\Delta E}{E} = \frac{a/\gamma}{0.434 \cdot N_{grey}} \Delta P$$

グレイレベル数

相対X線量の変動成分

FCR の場合

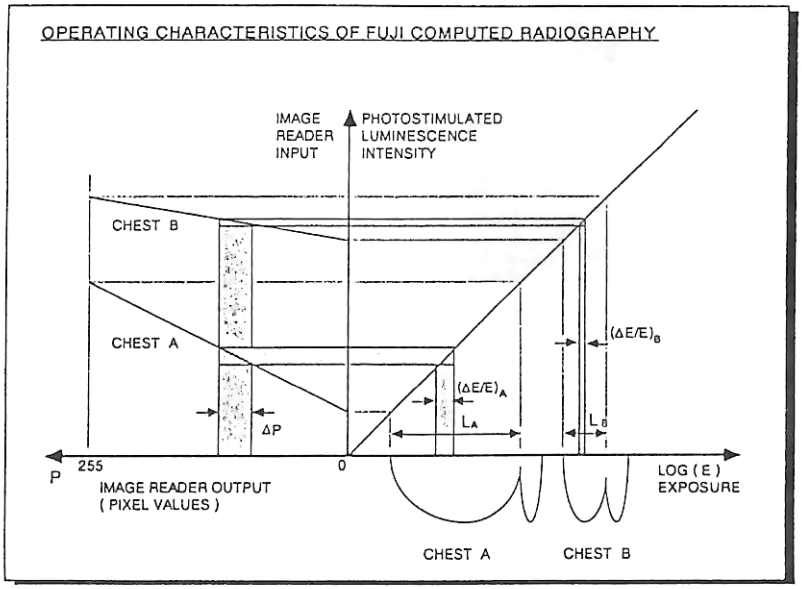


Fig. 28

$$\frac{\Delta E}{E} = \frac{L}{0.434 \cdot N_{grey}} \Delta P$$

L ← readout range （ L 値）

ピクセル値の変動成分は、相対X線量の変動成分に変換可能。

3．6　バックグランド補正

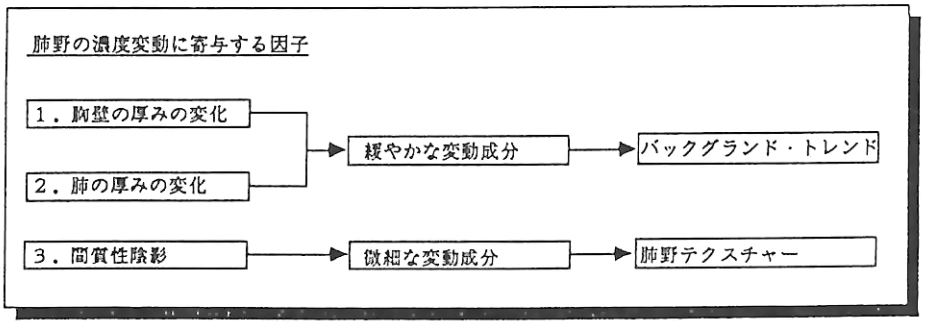


Fig. 29

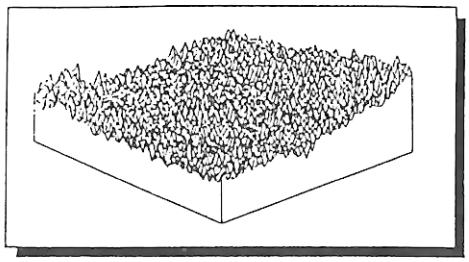


Fig. 30　オリジナル ROI

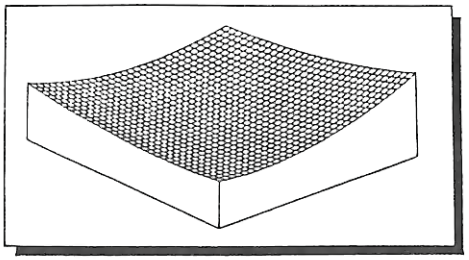


Fig. 31　バックグランド・トレンド

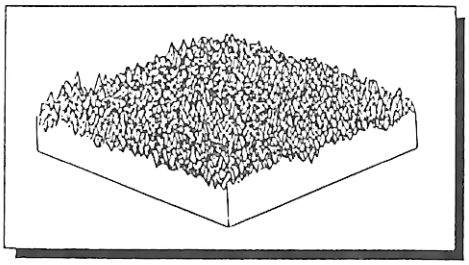
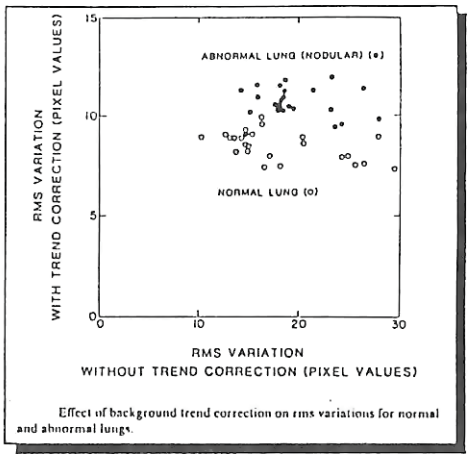


Fig. 32　補正後の ROI



Effect of background trend correction on rms variations for normal and abnormal lungs.

Fig. 33　バックグランド補正をしなければ、正常肺と間質性陰影を持った異常肺を、rms 値で区別することは不可能である。

バックグランド・トレンドを２次関数

$$f(x) = ax^2 + by^2 + cxy + dx + ey + k$$

で、最小二乗法を用いて近似する。

３．７　パワースペクトル

２次元フーリエ変換

$$F(u,v) = \iint f(x,y)\ \exp[\,-2\pi i(\,xu + yv\,)\,]\,dxdy$$

u,v；空間周波　←　バックグランド補正後の肺野テクスチャー

パワースペクトル

$$P(u,v) = |F(u,v)|^2$$

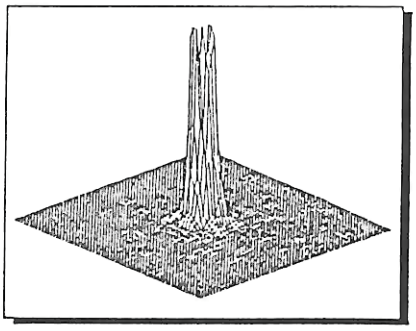


Fig. 34　肺野テクスチャーのパワースペクトル

3．8　視覚のレスポンス関数によるフィルタリング

高周波のX線写真モトル成分、および低周波におけるバックグランド補正後の残留成分を減少させる。

$$V(u,v) = \exp\left(-\frac{\left[\ln\sqrt{u^2+v^2} - \ln\left(25\sqrt{u_0^2+v_0^2}/D\right)\right]^2}{\left[2(0.973)^2\right]}\right)$$

V(u,v)；視覚のレスポンス関数　　u,v；空間周波数

$\sqrt{u_0^2+v_0^2}$ ；距離(D=25cm)でのピーク周波数

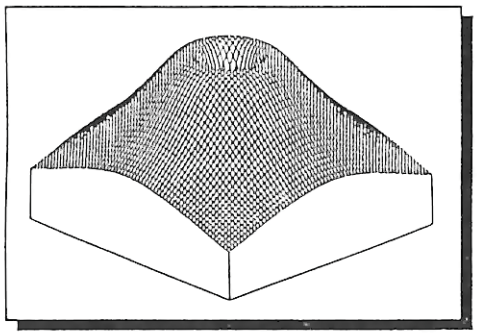


Fig. 35　視覚のレスポンス関数
ピーク周波数 1.5cycles/mm

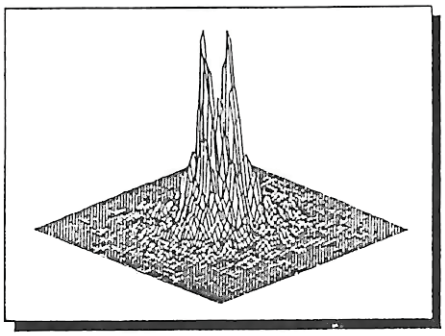


Fig. 36　フィルタリング後のパワースペクトル

フィルタリングの効果

正常肺のテクスチャ・メジャーと、
異常肺のそれとの差を大きくする。(強調効果)

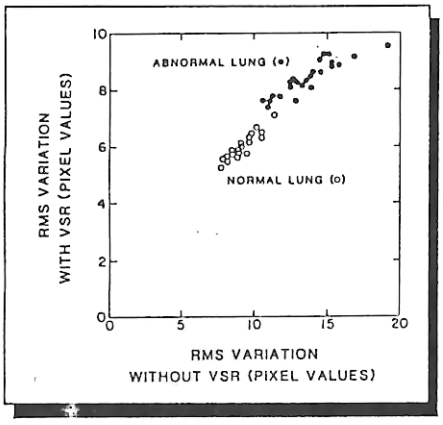


Fig. 37　rms 値に対する効果

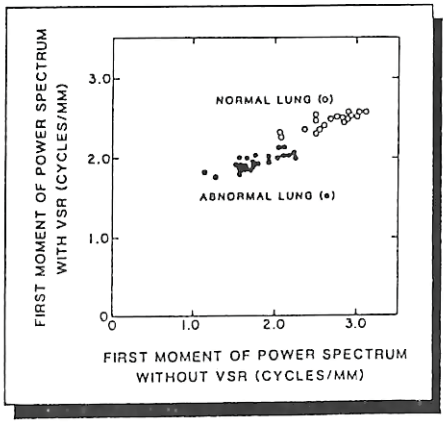


Fig. 38　パワースペクトルの
一次モーメントに対する効果

３．９　テクスチャ・メジャー

rms 値；R

$$R = \sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V^2(u,v)\,|F(u,v)|^2 du dv}$$

→ 濃度変動の大きさに対応

パワースペクトルの一次モーメント；M

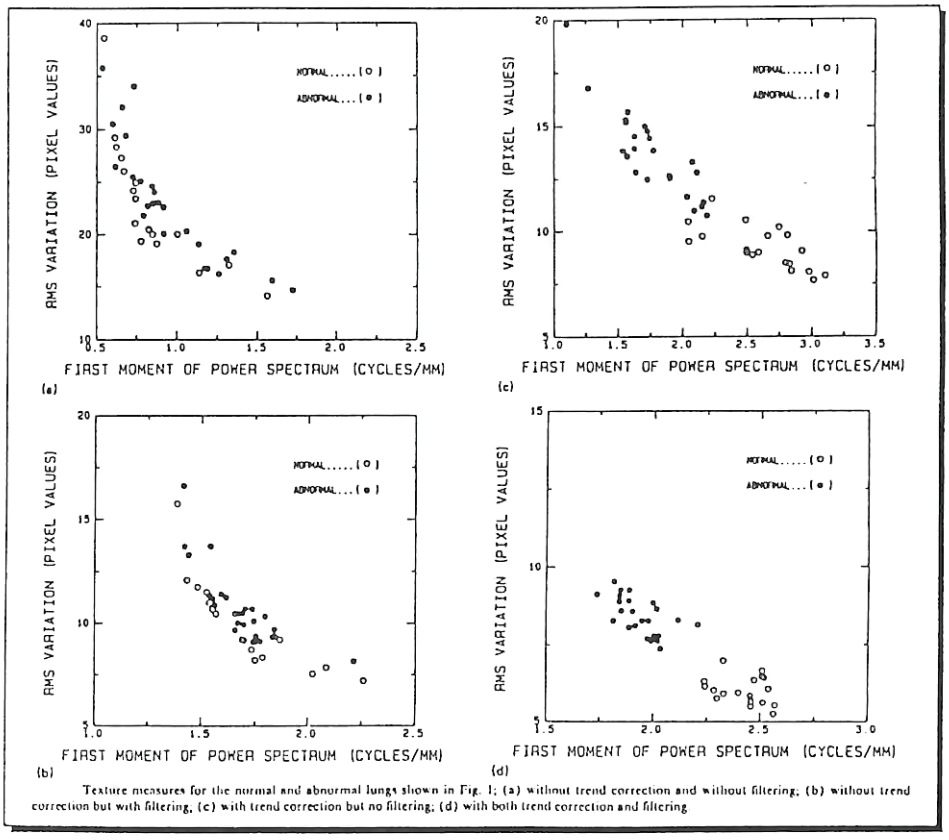
$$M = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sqrt{u^2+v^2}\, V^2(u,v)\,|F(u,v)|^2 du dv}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V^2(u,v)\,|F(u,v)|^2 du dv}$$

→ 肺野テクスチャの粗さ、または細かさに対応

V(u,v)　；視覚のレスポンス関数
F(u,v)　；肺野テクスチャーのフーリエ変換

Fig. 39

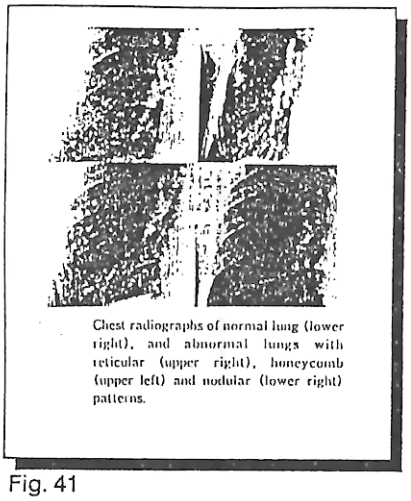
３．１０　バックグランド補正とフィルタリングの組み合わせ



Texture measures for the normal and abnormal lungs shown in Fig. 1; (a) without trend correction and without filtering; (b) without trend correction but with filtering; (c) with trend correction but no filtering; (d) with both trend correction and filtering.

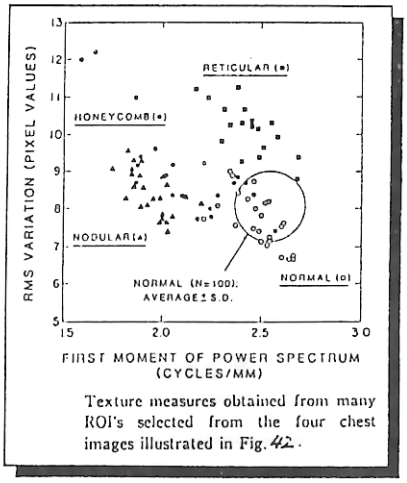
Fig. 40　バックグランド補正は、正常肺と異常肺を区別するためには、必ず必要な処理であり、また、視覚のレスポンス関数によるフィルタリングは、その効果を高める．

３．１１　臨床応用



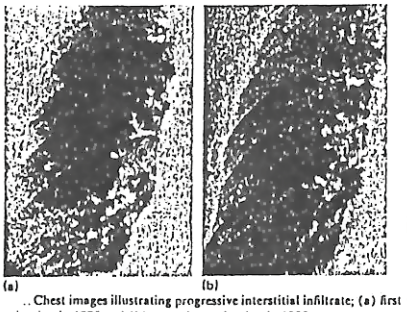
Chest radiographs of normal lung (lower right), and abnormal lungs with reticular (upper right), honeycomb (upper left) and nodular (lower right) patterns.

Fig. 41



Texture measures obtained from many ROI's selected from the four chest images illustrated in Fig. 42.

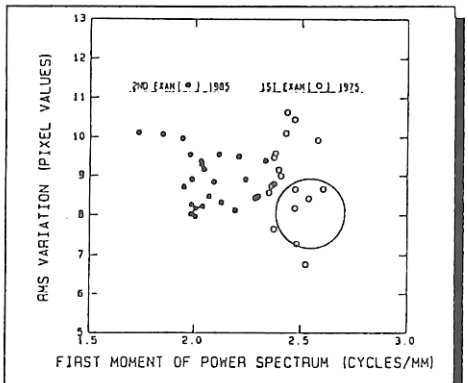
Fig. 42



(a) (b)

..Chest images illustrating progressive interstitial infiltrate; (a) first examination in 1975 and (b) second examination in 1985.

Fig. 43



Texture measures obtained from the two chest images showing progressive interstitial infiltrate (Fig. 43).

Fig. 44

４．正常肺と間質性陰影を持つ異常肺の自動分類

目的

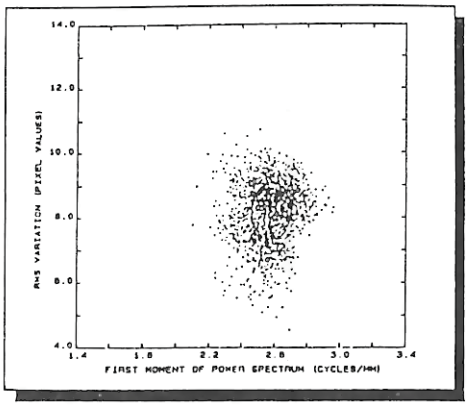
間質性陰影の検出に対する、使用した

客観的物理尺度（テクスチャ・メジャー）の

有用性の検討を行なう。

Fig. 45

4．1　テクスチャ・メジャーの正規化



正規化された rms 値
rms 値
正常肺の rms 値の平均
正常肺の rms 値の標準偏差

$$R_N = \frac{(R - \bar{R})}{\sigma_R}$$

正規化されたパワースペクトルの一次モーメント

$$M_N = \frac{(M - \bar{M})}{\sigma_M}$$

Fig. 46　データベースに含まれる正常肺１００例から求めたテクスチャ・メジャーの分布

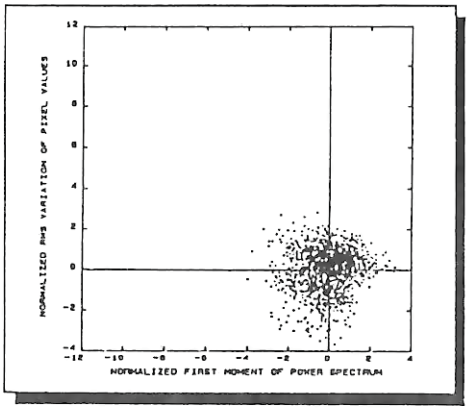


Fig. 47　正常肺１００例から求めた正規化されたテクスチャ・メジャーの分布

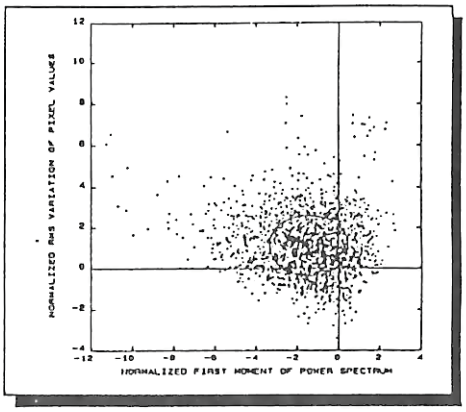
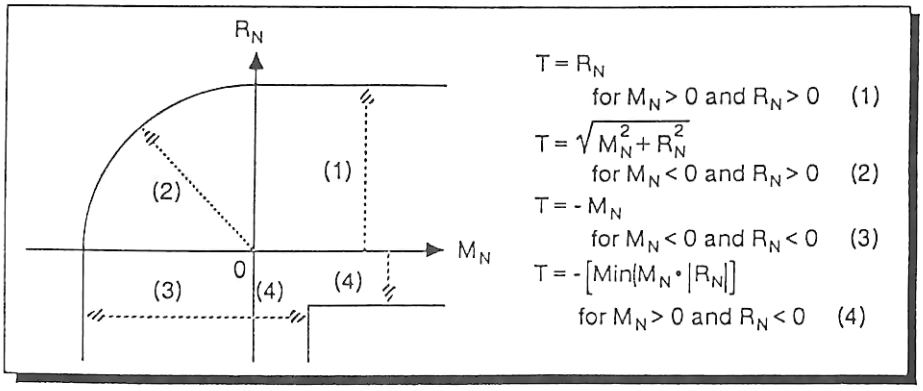


Fig. 48　異常肺１００例から求めた正規化されたテクスチャ・メジャーの分布

4．2　単一尺度（テクスチャ・インデックス）



$$T = R_N \quad \text{for } M_N > 0 \text{ and } R_N > 0 \quad (1)$$

$$T = \sqrt{M_N^2 + R_N^2} \quad \text{for } M_N < 0 \text{ and } R_N > 0 \quad (2)$$

$$T = -M_N \quad \text{for } M_N < 0 \text{ and } R_N < 0 \quad (3)$$

$$T = -\left[\text{Min}\{M_N \cdot |R_N|\}\right] \quad \text{for } M_N > 0 \text{ and } R_N < 0 \quad (4)$$

Fig. 49　テクスチャ・インデックス（T）

４．３　分類アルゴリズム

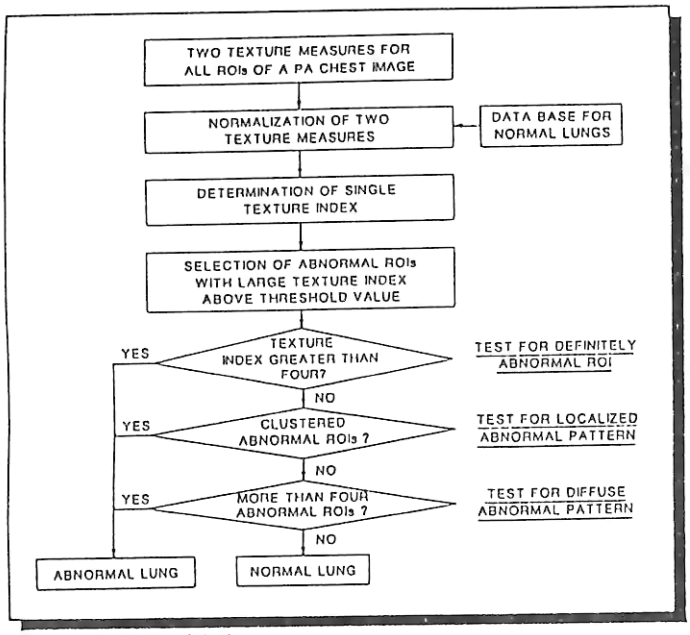


Fig. 50　分類アルゴリズム

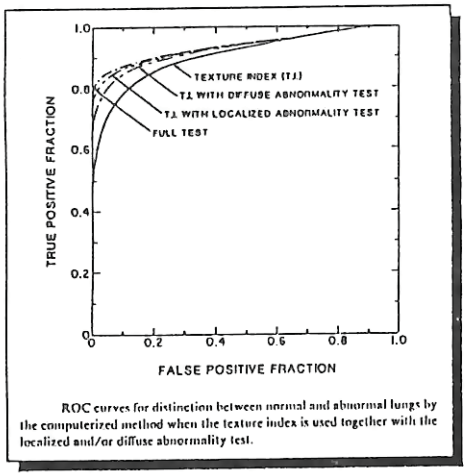


Fig. 51

４．４　放射線科医による診断結果との比較

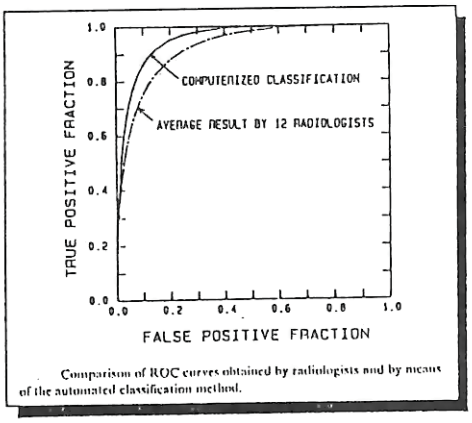


Fig. 52　対象；胸部単純写真　６０

２０　正常

２０　軽微な間質性疾患

２０　その他の病変

参考文献

1 R. N. Sutton, E. L. Hall ; Texture measures for automatic classification of pulmonary disease. IEEE Trans. Comput., **C21**, 667(1972).

2 R. P. Kruger, W. B. Thompson, A. F. Turner ; Computer diagnosis of pneumoconiosis. IEEE Trans. Systems, Man, and Cybernetics, **SMC4**, 40(1974).

3 J. R. Jagoe, K. A. Patton ; Reading chest radiographs for pneumoconiosis by computer. Br. J. Ind. Med., **32**, 267(1975).

4 S. Katsuragawa, K. Doi, H. MacMahon ; Image feature analysis and computer-aided diagnosis : Detection and characterization of interstitial lung disease in digital chest radiographs. Med. Phys., **15**, 311(1988).

5 G. Powell, K. Doi, S. Katsuragawa ; Localization of inter-rib spaces for lung texture analysis and computer-aided diagnosis in digital images. Med. Phys., **15**, 581(1988).

6 S. Katsuragawa, K. Doi, H. MacMahon ; Quantitative analysis of lung texture for computer-aided diagnosis of interstitial disease in digital chest radiographs. Proc. SPIE, **914**, 371(1988).

7 S. Katsuragawa, K. Doi, H. MacMahon ; Image feature analysis and computer-aided diagnosis in digital radiography : Classification of normal and abnormal lungs with interstitial disease in chest images. Med. Phys., **16**, 38(1989).

FUNDAMENTALS OF ROC ANALYSIS

Charles E. Metz, Ph.D.

Professor of Radiology

The University of Chicago

1. INTRODUCTION

From a practical point of view, an image is "good" only to the extent that it serves as a useful tool when it is applied to a particular task. Therefore, the quality of an imaging procedure in diagnostic medicine depends upon the quality of the medical decisions that can be made with it.

In the broadest sense, the value of a particular diagnostic imaging procedure depends upon the benefit of that procedure to society as a whole; not only the risks and financial costs of the procedure but also the availability of therapies and alternative diagnostic techniques must be taken into account. While this very general perspective addresses the ultimate concern, it does not provide an immediately practical basis for making measurements of image quality. Therefore, we must restrict our attention to a narrower definition of medical image quality.

Knowledge of each patient's state of health or disease is not a sufficient condition for effective medical care, but almost always it is a necessary condition for such care. Therefore, the quality of a medical imaging procedure can be defined and measured meaningfully in terms of the extent to which image-based diagnoses agree with the actual state of health or disease of each patient. At present, medical images must be interpreted by human beings, so an acceptable measure of image quality must take the abilities and limitations of human observers into account. Also, an acceptable measure of image quality must be *statistical:* it must describe both the typical characteristics and the variation of images and image readings. A committee of

the International Commission on Radiation Units and Measurements (ICRU) is currently drafting a report that will discusses this practical approach to image quality and will recommend specific techniques that can be used to implement it.

For simplicity, the evaluation of diagnostic decision performance usually is restricted to situations in which the truth concerning patients is divided into two states (such as "abnormal" versus "normal," "disease X present" versus "disease X absent," etc.) and in which two corresponding decisions can be made. The adequacy and limitations of this restriction have been discussed in the literature (Metz, 1986a). Often the two states are indicated by the abstract words "positive" and "negative" to denote a defined state of truth and its compliment. In studies that seek to measure medical image quality, these two states can be selected by the designer of the experiment to represent alternative diagnoses, the presence and absence of some diagnostically relevant image feature, or the presence and absence of an idealized geometric object. Images of actual clinical cases, images of phantoms, or images generated by computer can be used in the experiment, depending on the compromise between realism and convenience that is considered appropriate.

2. INADEQUATE MEASURES OF IMAGE QUALITY

The physical characteristics of many imaging systems can be described in terms of spatial resolution (by the Optical Transfer Function), noise magnitude and texture (by the Wiener spectrum), and contrast transfer (by the sensitometric curve). Knowledge of these specifications of the imaging system allow the physical properties of the image any object to be calculated. However, the quality of the medical decisions that an imaging procedure allows usually cannot be predicted from knowledge of the physical properties of the images, because the complex process of human visual perception is poorly understood. Someday in the future it may become possible to develop reliable models for visual detection and recognition that will

3. THE ROC CURVE

The fact that both Sensitivity and Specificity depend upon the observer's setting of his critical confidence level might seem to be a serious problem, because different observers usually set their critical confidence levels differently, and because the particular level-settings that an observer adopts cannot be measured or controlled. However, we can obtain a fully adequate measure of diagnostic performance if we acknowledgement this effect and use it to our advantage. In fact, all of the limitations of the various measures of diagnostic performance that I have mentioned up to this point are overcome if we measure all of the combinations of Sensitivity and Specificity that an observer can produce when he uses the set of images to be evaluated for particular two-alternative decision task. One of these indices can then be plotted as a function of the other. More commonly, "True Positive Fraction" (TPF), which is equivalent to Sensitivity, is plotted against "False Positive Fraction" (FPF), which is equal to 1.0 – Specificity, thereby producing a "Receiver (or Relative) Operating Characteristic" (ROC) curve (Green and Swets, 1966; Metz, 1978; Swets and Pickett, 1982).

ROC curves rise from the lower-left corner of the unit square, bend to the right with decreasing slope, and finally enter the upper-right corner of the square. High ROC curves represent better detection performance than low ROC curves, because appropriate settings of the decision criterion on a higher ROC produce a larger TPF for any given FPF, and a smaller FPF for a given TPF. Therefore, if ROC curves of interest do not cross, detection performance can be summarized by the "area index" A_z, which represents the area under an ROC inside the unit square (Swets and Pickett, 1982; Metz, 1986a). Alternatively, ROC curves can be compared in terms of the TPF values they provide at a particular FPF of practical interest (Swets and Pickett, 1982; McNeil and Hanley, 1984).

4. SELECTION OF CASES AND OBSERVERS FOR AN EVALUATION STUDY

The detectability of a lesion in an image obviously depends not only upon the physical properties of the imaging system, but also upon the size, contrast, etc. of the lesion in question and upon the characteristics of background structures that are (or may be) present in the image. More generally, the ability of an observer to discriminate visually between two classes of objects or "cases" depends upon the subtlety of the differences between the two classes of cases in question. Therefore the ROC curve that is measured with a set of images -- and so, according to our practical definition of image quality, the quality of those images -- depends not only upon the physical properties of the images, but also upon the particular actually positive and actually negative cases (in other words, phantoms, patients, etc.) from which the images were made.

At first, this dependence may seem confusing and undesirable, but in fact it is consistent with the common observation that different imaging systems may be best in depicting different classes of scenes and in different decision tasks. Similarly, image quality as we define it here depends upon the skill of the observer who reads the images, but this dependence also is appropriate, because different imaging systems may be best when used by observers with different training or experience. Ideally, all observers who perform a particular image-reading task would possess the highest level of skill that is humanly attainable; in reality, however, images are read by observers with diverse skills.

Because the results of an image-evaluation study generally depend upon the cases and observers that the study employs, care must be taken to ensure that those cases and observers are selected appropriately. It is important to recognize that sampling issues must be addressed in any evaluation study, and that ROC methodology is no more demanding in this regard than

other methods of analysis that provide less meaningful descriptions of detection performance (Metz, 1986a).

In designing a visual-detection experiment for the evaluation of image quality, one must decide first whether an absolute measure of the detectability of some class of diagnostic features or some particular disease is desired, or whether the goal is more simply to rank alternative imaging procedures (Metz, 1989). The sampling issues that must be confronted in these two kinds of experiments can be quite different.

4.1 Absolute measurements of detection performance

Reliable absolute measurements of disease detectability in a defined patient population are often extremely difficult to obtain, in part because the sample of patients included in the study must accurately reflect the population of patients at large about which conclusions are to be drawn. Therefore, many sources of potential bias must be taken into account (Begg and McNeil, 1989). For example, an experiment that attempts to measure the absolute detectability of lung nodules by chest radiography must ensure that the distribution of nodule sizes is the same in the study sample as in the defined population of interest, because the detectability of a nodule depends on its size. "Stratified sampling" techniques (Kendall and Stewart, 1976) have not yet been used formally in medical imaging but may prove useful for reliable absolute measurements of system performance. In the detection of lung nodules, for example, these techniques can help to ensure that an appropriate distribution of nodule sizes is used in a study (Metz, 1989).

The absolute detectability obtained in medical imaging depends not only upon the difficulty of the cases, but also upon the experience and skill of the observers who read the images; experienced mammographers have been shown to perform better than general radiologists in

using xeromammograms to discriminate between malignant and benign breast lesions (Getty et al, 1988), for example. Therefore, if reliable absolute measurements of detectability are to be obtained from a medical imaging study, the relevant population of observers must be defined, and the sample of observers employed in the study must accurately represent that population. Similarly, to measure accurately the absolute detectability of disease that would be obtained in routine clinical practice, the conditions under which images are read in the study (such as the amount of reading time, the ambient light level, etc.) must represent those that would be used in clinical practice.

4.2 System-ranking studies

Studies that attempt only to rank the quality of imaging systems are often much more straightforward. Sampling considerations still require attention in the study design, but now the only requirement is that these factors do not affect the ranking of the systems; their effects on absolute detectability are no longer of primary concern. The key need becomes one of ensuring that a system which would provide superior diagnostic performance in its real-world application is found better in the study, and that two systems providing equivalent performance in the real world are found equivalent by the study.

An individual who designs this second kind of study has greater freedom in choosing the cases it will employ. A basic question that often arises in designing a system-ranking study concerns the most appropriate level of case difficulty. Common sense suggests correctly that not all of the cases should be extremely easy, because then even a very poor imaging system would perform well. But should the cases cover a broad spectrum of difficulty to represent cases at large; should they be restricted to "subtle" cases; or should they be so challenging that even the best system finds them quite difficult? The answer depends in part on the statistical properties of ROC indices.

If the difficulty of cases in a particular diagnostic task depends primarily on only one factor, such as lesion size, and if we may assume that a system which is better in detecting small lesions will be better in detecting large lesions also, then the question at hand is reduced to one of choosing the best lesion size, or range thereof. If a single lesion size is to be employed in a system-ranking study, it should be chosen to maximize the difference between the expected values of the ROC index used to summarize performance, relative to the uncertainty in that difference -- in other words, to maximize the statistical power of the experiment. No formal guidance is available as yet concerning this choice, but a reasonable rule of thumb when the ROC area index, A_z, is used seems to be that the average of the A_z values of the two modalities should lie near 0.80 (Metz, 1989). For example, if lesion size is the primary determinant of detectability with each system, then a pilot experiment involving the two systems and several lesion sizes should be performed to determine approximately the lesion size that achieves this average level of performance. A range of lesion sizes distributed around this optimal size may be employed in the definitive experiment if that is considered desirable for "realism," but we must recognize that each lesion with a size substantially different from the optimal value will make a smaller contribution to statistical power.

Unfortunately, the task of selecting appropriate cases may not be so simple even for a system-ranking experiment if the detectability of the disease of interest depends strongly on two or more distinct image features, such as soft-tissue masses and microcalcifications in the detection of breast cancer by mammography. Some of the subtle issues that must be confronted in this situation are discussed elsewhere (Metz, 1988a).

4.3. Establishing diagnostic truth

All objective techniques for the evaluation of observer performance measure the agreement between the observers' decisions and some external standard of truth; therefore, the true state

of each case in an objective image evaluation study must be known. In evaluation studies that employ phantom images or computer-simulated images, the experimenter knows the actual state of each case because he/she controls it. The results of such studies may be reliable, in the sense that they can be reproduced, but often they are unrealistic, in the sense that the phantoms or computer simulations from which the study's images were made do not adequately represent the complexity of real patients and their images. Therefore, even the most carefully performed phantom and computer-simulation studies may leave doubt concerning whether similar results would have been obtained with real patients and images.

Unfortunately, the establishment of diagnostic truth in clinical studies is usually difficult, both in principle (because "truth" is ultimately a philosophical matter) and in practice (because great effort may be required to determine the actual state of health or disease in a particular patient at a particular point in time "beyond a reasonable doubt"). But despite these difficulties, carefully designed clinical ROC studies can be done, and useful conclusions can be drawn from them; for example, see reports listed by Swets and Pickett (1982) and by Metz (1986a).

Important issues that must be confronted in establishing truth in clinical evaluation studies have been reviewed by a number of authors (Ransohoff and Feinstein, 1978; Swets and Pickett, 1982; Metz, 1986a; Begg and McNeil, 1988). Particular attention must be focused on biases that may be caused by the omission of clinical cases for which truth is particularly difficult to establish (Ransohoff and Feinstein, 1978; Gray et al, 1984).

5. READING-ORDER EFFECTS

When two or more images of a particular patient are read by the same observer, the image read last will tend to be interpreted more accurately than the image read first if any relevant information is retained by the observer from a reading of one image of the patient to the next.

Therefore, if all of the images of a patient sample made with modality "A" are read before any of the images of the same patient sample made with modality "B," the results of the study will be potentially biased in favor of modality "B." Biases of this kind, called "reading-order effects," must be avoided in evaluation studies that use a single case sample to compare two or more imaging systems.

There are two methods by which biases due to reading-order effects can be reduced or eliminated; the common theme is to vary the order in which the modalities' images are read so that reading-order effects tend to cancel out. In the first method, the order of the modalities is varied across observers; thus, the ranking of the modalities remains potentially biased for individual observers, but any such bias tends to cancel across observers. The second method, which requires more complicated experimental designs, attempts to cancel the bias within each observer's results by breaking the case sample into several subsets and having each observer read the modalities in a different order for each subset (Metz, 1989).

6. ROC DATA COLLECTION

ROC curves that describe image quality can be measured in two ways. With the "Yes/No" method, the observer reads each image in a set of images as "positive" or "negative." He then rereads the set of images several times in different reading sessions, using different settings of his critical confidence level to distinguish "positive" image readings from "negative" readings. For example, the observer would read the images "conservatively" in one session and "aggressively" in another session. The data analyst can estimate one point on the ROC from the results of each reading session by comparing the image readings with the actual presence and absence of the disease or image feature in question. However, this method is extremely

inefficient, because it requires M readings of each image to estimate M points on each ROC curve.

In practice, visual-detection ROCs are almost always measured by obtaining "confidence-rating" data that represents the observer's relative confidence that each image was produced from an actually positive case. When these graded judgements are reported on a K-category scale, K-1 points on the ROC can be estimated from a single reading of each image (Green and Swets, 1966; Metz, 1978, 1979, 1986a). Usually a confidence-rating scale with five or six categories is employed, thereby yielding estimates of four or five points on the ROC in addition to the lower-left and upper-right corners of the unit square, into which every conventional ROC must pass. A larger number of confidence-rating categories would be desirable in principle, but in practice many observers find it difficult to grade subjective judgements on finer scales.

Discrete confidence-rating categories usually are assigned verbal labels such as "definitely or almost definitely positive," "probably positive," "possibly positive," etc. to ensure that each observer uses the categories in a strictly ordered fashion. The way in which these labels may be interpreted by an observer does not bias the measured ROC as long as the observer's relative confidence in the two states of truth varies monotonically across the categories (Metz, 1986a). However, for statistical efficiency and to guard against "degenerate" data sets (Metz, 1989), it is desirable for the confidence ratings to yield ROC points that are more-or-less uniformly spread along the curve. When approximately half of the images in an experiment are actually positive, this is accomplished if the observer uses the categories with roughly equal frequencies.

Clinical images can be read either with or without case-history information, and with or without the results of other tests that would precede the imaging test in real-world applications, depending on the goals of the evaluation study (Metz, 1986a). Similarly, clinical images can

be read either sequentially in controlled reading sessions or one at a time in the course of daily clinical practice. However, when the confidence-rating method is used in a controlled experiment, a brief training session should be held immediately before each image-reading session in which data will be acquired. In each of these training sessions, the observer should be shown a spectrum of cases that represents the range of decision difficulty that will be presented during data acquisition, and he/she should be motivated to use the confidence-rating categories in a way that will produce a roughly uniform spread of points along the ROC curve. Also, the observer must be made to understand that he/she should strive to use the rating scale in a constant way during data acquisition, because variation in use of the scale degrades decision performance (Goodenough and Metz, 1977; MacMillan and Kaplan, 1985). The cases that are presented during the training session should be typical of, but must not include, those that will be read during data acquisition.

7. ROC CURVE FITTING

After a set of confidence-rating data has been obtained, we are faced with the task of fitting a continuous ROC curve to those data. Objective curve-fitting methods require that some mathematical form be assumed for the ROC curve. Many functional forms have been proposed (Egan, 1975; Swets, 1986a), but the "binormal" form has been used most widely in medical imaging. According to the binormal model, which includes two adjustable parameters, each ROC is assumed to have the form that would be produced by two "normal" (Gaussian) decision-variable distributions with generally different means and standard deviations (Green and Swets, 1966; Egan, 1975; Swets, 1979). It is important to notice that any monotonic transformation of a decision variable changes the distributions that underlie the ROC but not the ROC itself; therefore, the binormal assumption concerns only the functional form of the ROC curve, which always can be examined empirically, and not the form of the underlying

distributions, which cannot be determined in many applications of ROC analysis (Metz, 1986a).

Empirically, the binormal form has been found to provide satisfactory fits to ROC data generated in a very broad variety of situations (Swets, 1986b), and it has the convenient property that all possible ROC curves are transformed into straight lines if they are plotted on "normal-deviate" axes (Green and Swets, 1966; Swets, 1979).

The two adjustable parameters of a binormal ROC can be taken to be the vertical-intercept and the slope of the straight line that represents the ROC when it is plotted on normal-deviate axes. These two parameters, usually denoted by "a" and "b", can be interpreted in terms of an effective pair of underlying Gaussian distributions as the distance between the means of the two distributions and the standard deviation of the "actually negative" distribution, respectively, with both expressed in units of the standard deviation of the "actually positive" distribution. Alternative ways of parameterizing a binormal ROC curve have been described by Swets (1979).

With the binormal model, the task of curve fitting becomes one of choosing numerical values for the {a,b} parameter pair to best represent the measured data. Conventional least-squares methods are not appropriate because the assumptions underlying those methods are not valid for ROC data; instead, maximum-likelihood estimation should be used (Metz, 1986b). Maximum likelihood algorithms for ROC analysis are readily available (Dorfman and Alf, 1969; Grey and Morgan, 1972; Metz et al, 1984, 1985; Swets and Pickett, 1982) and provide not only estimates of the parameters of the best-fit ROC curve, but also estimates of the uncertainties in those parameters (Metz, 1986b).

8. STATISTICAL TESTS FOR DIFFERENCES BETWEEN ROC CURVES

Now I shall mention briefly a variety of issues that arise in testing the statistical significance of differences between ROC curves estimated from confidence-rating data. The details of these issues have been discussed in publications by McNeil and Hanley (1984) and by Metz (1986b, 1989).

Before selecting a statistical test, we must answer at least three distinct questions :

(1) In what sense should a difference between two ROC curves be quantified in the particular situation at hand?

(2) Are the estimates of the ROC curves of interest statistically independent or are they potentially correlated?

and

(3) Was each ROC curve of interest measured from one or from more than one reading of each image?

The answers to these questions determine the test (or tests) that can be used to evaluate the statistical significance of an apparent difference between measured ROCs.

8.1. Ways in which differences between ROCs can be quantified

The null hypothesis that we must use to evaluate the statistical significance of a difference between two ROCs is dictated by the way in which we want to quantify the difference. In general, the most appropriate null hypothesis depends on the way in which the two imaging systems will be employed in practice.

One approach, which can be addressed with a Chi-square test (Metz and Kronman, 1980; Metz et al, 1984; Metz, 1986b), considers the null hypothesis that two sets of rating data arose from exactly the same binormal ROC curve. This null hypothesis of identical ROCs may be unnecessarily strict in some evaluation studies, however. For example, in practice we may not care whether the two imaging systems yield exactly the same TPF at all possible FPFs, but instead we may wish only to compare the TPFs of the two systems at a particular FPF of clinical interest -- FPF = 0.10, say. Then the null hypothesis of "no difference" would correspond to the condition that the two ROCs have equal TPFs at FPF = 0.10, and it could be addressed by an appropriate univariate z-score test (McNeil and Hanley, 1984; Metz et al, 1984; Metz, 1986b). Alternatively, we might be willing to consider two systems equivalent in clinical practice if they yield the same average value of TPF for FPFs ranging from 0 to 1 -- in other words, if the two systems' ROC curves have the same area (A_z) beneath them when they are plotted on conventional axes (Metz, 1986a). This null hypothesis can be addressed by other univariate z-score tests (Swets and Pickett, 1982; Hanley and McNeil, 1982, 1983; McNeil and Hanley, 1984; Metz et al, 1984; Metz, 1986b).

The three null hypotheses that I have mentioned are not equivalent, because ROCs can cross. Thus, for example, the two true ROCs in question may have equal areas beneath them yet be different curves, with different TPFs at all FPFs except one; or the true ROCs may have the same TPF at the FPF of interest, but be different curves and have different areas. Perhaps surprisingly, the results of the three statistical tests can disagree fairly frequently even when the ROCs are identical, and the null hypothesis of the "A_z" test and/or of the "TPF" test can be rejected even when the null hypothesis of identical ROCs is accepted (Metz, 1989)! Therefore, it is important to select the most pertinent null hypothesis and a corresponding test before statistical testing is begun, and to then perform only that single test. If multiple tests are performed, conflicting conclusions may be drawn at best, and at worst a highly significant

difference will be found after enough tests have been done -- even when the ROCs of the two systems are, in fact, identical.

8.2. Independent versus correlated ROC estimates

If estimates of two ROC curves are statistically independent, then no matter how the difference between the ROCs is quantified, the statistical variability of the difference depends only on the variabilities of the individual ROC estimates themselves. Several relatively simple significance tests apply to this situation (Metz and Kronman, 1980; Hanley and McNeil, 1982; Metz, 1986b).

When the estimates of the two ROC curves are not independent but instead tend to vary above and below their means together -- for example, because a case sample that is atypically difficult for one imaging system tends to be atypically difficult for the other as well -- then the difference between the ROC estimates will vary less than it would if the estimates were independent. Therefore, the statistical power of a system-ranking study can be increased by applying both systems to the same case sample. Increased statistical power will be achieved only if the covariance of the two ROC estimates is estimated by some means and then factored into the statistical test, however.

For comparisons in which each ROC is estimated from a single reading of each image, this can be accomplished by basing statistical tests on a "bivariate binormal" model (Metz et al, 1984; Metz, 1986b) that generalizes the conventional binormal model which underlies conventional ROC curve-fitting algorithms such as RSCORE II and ROCFIT. An alternative approach (Hanley and McNeil, 1983) can be used when the trapezoidal-rule approximation of the ROC area index is considered adequate.

8.3. Differences between ROCs estimated from multiple readings of each image

For simplicity, my discussion up to this point has focused on ROC curves that are measured from a single reading of each image. However, in many experiments each image from each modality is interpreted by more than one observer, and perhaps more than once by each observer. The advantage of multiple readings is fairly obvious: the effect of "between-reader" differences in skill is "averaged out" across the observer sample, and the effect of "within-reader" variation is reduced by each reader's replication. Much less obvious, unfortunately, are ways in which the variation-reducing effects of multiple readings can be taken fully into account in statistical tests for differences in modality performance.

Swets and Pickett (1982) have described a general model for variation in the difference between two area indices, A_z, measured with multiple readings of each image; it applies equally to any univariate difference between ROC indices. This model includes terms that represent the "case-sample" variation, "between-reader" variation, and "within-reader" variation that I have mentioned, together with correlation coefficients that account for the effects of matching readers and cases across modalities. The model shows clearly how the power of a statistical test for ROC differences is affected by the numbers of readers and readings in a multiple-reading experiment, and it prescribes a procedure for performing such a test. Swets' and Pickett's approach is the the most comprehensive one formulated to date, and in principle it is adequate for statistical analyses of data from most multiple-reading ROC experiments. The method is somewhat complicated, but it should be used whenever possible.

Of alternative techniques, the conventional "t-test for paired data" is clearly the simplest. This method can be employed whenever several readers interpret the images from both of two modalities. The test involves: (i) fitting an ROC individually to each reader's confidence-rating data for each modality, and extracting a univariate index (such as A_z) from each fitted curve;

(ii) calculating, for each reader, the difference between the index values for the two modalities; and (iii) using Student's "t" statistic to test the null hypothesis that the population mean of the resulting list of differences is equal to zero. This statistical test accounts fully for the effects of "between-reader" and "within-reader" variation, but its statistical power is governed strongly by the number of readers in the experiment (which determines the number of degrees of freedom of the t statistic), and it does not account for "case-sample" variation at all. The latter limitation may be acceptable with large random case samples and/or with small but carefully selected case samples that can be assumed to be representative, but with small random samples or small loosely selected samples it should be accepted only with great caution.

9. GENERALIZATIONS OF ROC ANALYSIS

As I mentioned earlier, conventional ROC analysis is limited to situations in which possible truth can be divided into two states and two corresponding decisions are available to the observer. Several generalizations of conventional ROC analysis have been proposed to overcome this limitation.

"Location ROC" (LROC) analysis (Starr et al, 1975) measures the ability of an observer not only to decide between the presence or absence of a single abnormality in an image, but also to localize that abnormality when it is present. Swets and Pickett (1982) have suggested that LROC analysis can be used also to measure the ability of observers to detect and classify single abnormalities; the measured curve is then called a "joint" ROC. A generalization of the LROC approach called "Free-response operating characteristic" (FROC) analysis applies to situations in which a particular kind of abnormality may be present at more than one location in an image and the observer is required to detect the presence of the abnormality at each of its locations (IAEA, 1977; Bunch et al, 1977). For experiments that seek to measure the detectability of lesions, FROC analysis can be more efficient than ROC or LROC analysis

because it allows more than one lesion to be be included in each image, thereby reducing the number of images needed to estimate reliably the probability (at each decision criterion) of detecting a lesion when it is actually present.

Unfortunately, LROC analysis and FROC analysis share two substantial practical limitations: no formal curve-fitting procedures and no generally applicable statistical tests for differences in performance have been developed. Thus, at present, neither LROC nor FROC analysis can be recommended broadly. Hope does exist for overcoming these limitations, however (Metz, 1989). In particular, a method for fitting FROC curves on the basis of an ad hoc but apparently reasonable model was proposed recently by Chakraborty (1989); although at present this method has not been fully tested, it may prove useful.

10. CONCLUSIONS

A review of the advantages and limitations of the various techniques used to assess diagnostic performance suggests that ROC analysis provides the most meaningful approach in most situations. Only ROC analysis distinguishes between the inherent diagnostic capacity of radiologists' image interpretations, on one hand, and any tendencies that they may have to "under-read" or "over-read," on the other. ROC techniques have been used successfully to assess the performance of a broad variety of radiologic imaging procedures in diverse diagnostic tasks.

Although the practicality of ROC methodology sometimes has been questioned on the grounds that it requires diagnostic truth to be established in "large" numbers of patients, diagnostic truth and statistically significant sample sizes are requirements of any objective method. Recent research on the statistical properties of ROC indices indicates that essentially similar sample sizes are required to achieve adequate statistical power with most "traditional"

measures of diagnostic performance and with ROC analysis. Misconceptions regarding the relative difficulty of ROC methods may be due to the unfortunate fact that issues of diagnostic truth, potential biases, and statistical power have not always been given appropriate attention in more traditional approaches. The inherently statistical foundations of ROC analysis force us to face those issues, however -- not only in assessing ROC methodology, but also in reassessing more familiar techniques. A balanced overview suggests that ROC methodology is no more demanding in most situations than other methods of analysis that provide less complete descriptions of diagnostic performance.

ACKNOWLEDGEMENT

This work was supported by the U.S. Department of Energy under Grant DE-FG02-86ER60418. Portions of this lecture were published previously (Metz 1986a, 1988a, 1988b, 1989).

REFERENCES

Begg, C.B., and McNeil, B.J. (1988). Assessment of radiologic tests: control of bias and other design considerations. Radiology 167: 565-569.

Bunch, P.C., Hamilton, J.F., Sanderson, G.K., and Simmons, A.H. (1977). A free response approach to the measurement and characterization of radiographic observer performance. Proceedings of the SPIE 127: 124-135.

Chakraborty, D.P. (1989). Maximum likelihood analysis of free response receiver operating characteristic (FROC) data. Medical Physics (in press).

Dorfman, D.D., and Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals -- rating method data. Journal of Mathematical Psychology 6: 487-496.

Egan, J.P. (1975). Signal detection theory and ROC analysis. New York: Academic Press.

Getty, D.J., Pickett, R.M., D'Orsi, C.J., and Swets, J.A. (1988). Enhanced interpretation of diagnostic images. Investigative Radiology 23: 240-252.

Goodenough, D.J., and Metz, C.E. (1977). Implications of a "noisy" observer to data processing techniques. In: Information processing in medical imaging (C. Raynaud and A.E. Todd-Pokropek, eds.). Orsay, France: CEA, pp. 400-419.

Gray, R., Begg, C.B., and Greenes, R.A. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Medical Decision Making 4: 151-164.

Green, D.M., and Swets, J.A. (1966). Signal detection theory and psychophysics. New York: Wiley. Reprinted with corrections in 1974 (Huntington NY: Krieger) and in 1988 (Los Altos, California: Peninsula Publishing).

Grey, D.R., and Morgan, B.J.T. (1972). Some aspects of ROC curve-fitting: normal and logistic models, Journal of Mathematical Psychology 9: 128-139.

Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29-36.

Hanley, J.A., and McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148: 839-843.

International Atomic Energy Agency (1977). IAEA co-ordinated research programme on the intercomparison of computer-assisted scinigraphic techniques. In: Medical radionuclide imaging, Volume 1. Vienna: IAEA, pp. 585-615.

Kendall, M., and Stewart, A. (1976). The advanced theory of statistics, Volume 3. New York: Hafner, New York, pp. 182-243.

Loo, L.-N., Doi, K., Ishida, M., Metz, C.E., Chan, H.-P., Higashida, Y., and Kodera, Y. (1983). An empirical investigation of variability in contrast-detail diagram measurements. Proceedings of the SPIE 419: 68-76.

Lusted, L.B. (1978). General problems in medical decision making, with comments on ROC analysis. Seminars in Nuclear Medicine 8, 299-306.

Macmillan, N.A, and Kaplan, H.L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false alarm rates. Psychological Bulletin 98: 185-199.

McNeil, B.J., and Hanley, J.A. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Medical Decision Making 4: 137-150.

Metz, C.E. (1978). Basic principles of ROC analysis. Seminars in Nuclear Medicine 8: 283-298.

Metz, C.E. (1979). Applications of ROC analysis in diagnostic image evaluation. In: The physics of medical imaging: recording system measurements and techniques (A. Haus, ed.). New York: American Institute of Physics, pp. 546-572.

Metz, C.E. (1986a). ROC methodology in radiologic imaging. Investigative Radiology 21: 720-733.

Metz, C.E. (1986b). Statistical analysis of ROC data in evaluating diagnostic performance. In: Multiple regression analysis: applications in the health sciences (D. Herbert and R. Myers, eds.). New York: American Institute of Physics, pp. 365-384.

Metz,C.E.(1988a).Current problems in ROC analysis. In: Proceedings of the Chest Imaging Conference 1987 (W.W. Peppler and A.A. Alter, eds.). Madison, Wisconsin: Department of Medical Physics, University of Wisconsin-Madison, pp. 315-336.

Metz,C.E.(1988b). Evaluation of medical images. Invited for presentation at the NATO Advanced Study Institute "The Formation, Handling, and Evaluation of Medical Images;" Povoa de Varzim, Portugal, September 1988. To be published in the proceedings (A.E. Todd-Pokropek and M.A. Viergever, eds..; in press).

Metz, C.E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. Investigative Radiology 24: 234-245.

Metz, C.E., and Kronman, H.B. (1980). Statistical significance tests for binormal ROC curves. Journal of Mathematical Psychology 22: 218-243.

Metz, C.E., Kronman, H.B., Wang, P.-L., and Shen, J.-H. (1985). ROCFIT: a modified maximum likelihood algorithm for estimating a binormal ROC curve from confidence-rating data. Chicago: Department of Radiology, The University of Chicago. Mini- and micro-computer versions of this and other software for ROC analysis are available from the author at no cost.

Metz, C.E., Wang, P.-L., and Kronman, H.B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Information processing in medical imaging (F. Deconinck, ed.). The Hague: Nijhoff, pp. 432-445.

Ransohoff, D.F., and Feinstein, A.R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New England Journal of Medicine 299: 926-930.

Starr, S.J., Metz, C.E., Lusted, L.B., and Goodenough, D.J. (1975). Visual detection and localization of radiographic images. Radiology 116: 533-538.

Swets, J.A. (1979). ROC analysis applied to the evaluation of medical imaging techniques. Investigative Radiology 14: 109-121.

Swets, J.A. (1986a). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychological Bulletin 99: 100-117.

Swets, J.A. (1986b). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psychological Bulletin 99: 181-198.

Swets, J.A., and Pickett, R.M. (1982). Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press.

Weinstein, M.C., and Fineberg, H.V. (1980). Clinical decision analysis. Philadelphia: Saunders.

第４回リフレッシャースクールに参加して

大阪市立大学病院　井　上　　誠

私がこの第４回画像リフレッシャースクールに参加したのは、職場の先輩であり今回のスクールのチューターでもある畑川氏に歓められたのと、ディジタル画像について詳しい勉強をしたことがなかったのでアナログ系との違いを学びたいと思ったからです。

リフレッシャースクールという名のとおり参加者の方は皆病院に10年以上勤めている私にとっては先輩ばかりで、このスクールに参加することで心身ともにリフレッシュし、ここで学んだことを明日からの仕事に生かそうという方ばかりでした。しかし勤めはじめてまだ1年と数か月、定まった勤務も得手分野も持たない私にとってはほとんど零からのスタートといっても過言ではないと思います。

前回までのリフレッシャースクールでは実際にMTFやROC曲線などを測定・計算し、得られた結果について検討し、その間に講義をはさむというかたちをとっていたそうですが、今回は3日間ビッシリと議義のみということなので長すぎて退屈してしまうのではないかと心配でした。しかし実際に議義を受けてみるとどれも大変興味深い話であり3日間が非常に短かく感じられました。

4つあった講義の中で最も印象に残ったのは桂川先生の「コンピュータ支援画像診断の初歩から理論、実際まで」でした。よく技師の中には「画像などは実際に臨床には役立たない」という人がいますが、先生の講義に出てきたRMS値など画像が肺の間質性陰影を区別するのに役立つということから、前記のような考えは間違いであるということがよくわかりました。

講義が終ってからのビールを囲んでの夜遅くまでの懇親会も大変有意義であったと思います。チューターや講師の先生方の意外な本音が聞け、3日間共に学んだ方々と交流を持てたということは私自身にとっても貴重な体験になりました。

このスクールに参加することで零からのスタートであった私の知識と経験がどのくらいプラスされたかわかりませんが、これからも勉強会や研究会などというものにできるだけ参加して、知識を上積みして行きたいと思っております。どうもありがとうございました。

国立循環器病センター放射線診療部　大竹野　浩　史

診療放射線技師になって約1年半、日常業務の中にも増感紙／フィルム系による画像だけでなく、DSAなどディジタル画像を扱う機会も増えてきました。今のところは、とにかくデータを取り込むことに精いっぱいで、より良い画像を得るにはどうすれば良いのか、実際にディジタル画像装置を扱う上でのテクニックも未習熟なままで検査を行っております。もっとも検査

内容がルーチン化していますからマニュアル通りに進めていけば、それなりの画像は出てきます。しかし何故そのテクニックを使うのか、理論的裏づけは何かといったことを知らないままですませるわけにもいかず、もっと勉強する必要があるとは思っていました。

そんなおり、今回の画像リフレッシャースクールに参加したわけですが、講義のレベルは私のそれをずっと超えたものでした。聞いたことのない用語が次々と出てきて、私は何と場違いな所へ来てしまったかと悔やむこともたびたびでした。それでも講師の先生方の用意された資料とスライドやOHPなどを使用されての講義は、ディジタル画像に関する知識を広げるものとなりました。惜しむらくは事前に資料をいただけてたら自分なりに問題点（知識の不足点やら弱点やら）を明らかにして議義に臨めたものをと思いますが、こんなことを書くと、参加されている多くの方々はすでに問題点を持って参加されているのだから自分の不勉強をたなに上げて何を言うかと、お叱りを受けるだけでしょう。いただいた資料を自分なりにフォローアップしていきたいと思っております。

そんなわけで、技術的な知識に関してはリフレッシュではなく、フレッシュなスクールとなりましたが、日常業務のあわただしさから解放された3日間で気分的に大変リフレッシュしたものとなりました。そして日々ディジタル画像と取り組んでおられる方、これから取り組もうとされている方が、全国の施設からより良いディジタル画像のあり方、その画質の向上を目指して参加された中に加えられたことは大変な刺激となりました。夜遅くまでビールを飲みながらの四方山話では参加された方々の人柄が出て楽しいひとときとなりましたし、その話の中で画像を考え直すきっかけができました。出身学校の先輩方が多く参加されていて、思わぬところで人のつながりを知ることもできました。

夜学での「私　画像」をテーマにしての2分間スピーチで、私にとって画像はあこがれであると述べましたが、言葉たらずの感がありますので補足しますと、私は画像について多くの人と語り合える人にあこがれていて、自分もそうなりたいとの意味で「私　画像」あこがれと述べたしだいです。リフレッシャースクールには、そのような方々がきら星のごとくいらっしゃいました。今回の参加がこの思いへのステップアップになりました。これからも画像に取り組んでいきたいと思っております。

国立呉病院　　岡　　透

この度、このスクールに参加して感じたことは、一言で言えば参加して良かったという事であります。学校を卒業して9年が過ぎ、日頃机に長時間座る事など、ほとんどありませんが、3時間という講義が、以外と短く感じられました。

3日間通じて、話題の中心は、CRであった様です。講義、講演の内容については、私自身、画像評価に関し、日頃から知識も経験もほとんどなく、理解に苦しむことが多々ありましたが、

結果的には、自分のレベルでそれなりに、消化吸収できたと思います。

特に、【誰にもわかるNEQ】の講義は、大変、良かったと思います。いままで、技術学会雑誌などで、NEQとかDQEなどの文献は読み飛ばしていたが、これからは少しは、興味がもてそうです。

夜学の方も非常に楽しく過ごせました。昼間の講義疲れを夜学で、吹き飛ばした感もあった様です。

奈良県立五条病院放射線科　　福　神　　敏

この度、この講座を受講するきっかけとなったのは、「ディジコル画像」という言葉にひかれたためであった。最近、マスコミの間でも「ディジタル画像処理」という言葉が一般化し、特に医療部門では、画像処理や、FCRが身近になっているにもかかわらず、その評価方法や、最近の評価したデータなど、まったく知らなかったので受講しようと思った。又、PACSにも興味があったので、駄目かもしれないと思いながらも申し込んだのであった。

さて、プログラムがきてみると、夜まで講座があるとの事。その夜学では、僕にとって理解とはほど遠い「NEQ」を易しく説明していただけるとのことなので、期待と本当に解るのだろうかという不安が混じった複雑な気持ちになった。

講座が始まってみると、講師の先生方がやさしく説明してくださったので、たいへん有意義な時間を過ごせたと感謝しています。又、NEQを測定してみようという勇気を与えられましたので、トライしようと考えています。これからも、難しい、解りにくい画像解析にも基礎となる部分を復習しながらとりくんでいこうと思っています。以上の様に、私にとっては、画像に関しての気持ちの“リフレッシュ”ができました事に感謝しています。又、この気持ちを持ち続けようと思います。

最後に、今回のリフレッシャースクールでいろいろな施設の方と知り合うことが出来ました。本当に有意義な夏の一時を過すことができたと思います。

大阪府立病院画像診断科　　船　橋　正　夫

今回のサマースクールで最も印象に残ったのは、なんといってもスペクトル解析の広範囲な応用の実情でした。それ以外では、講義終了後に我々と同様参加者で同宿舎のE大のT氏に講演していただいた内容が秀逸のものでした。わたしなりに理論化し、「対人的自己相関関数」と名付けたいと考えています。どういうものかといいますと、看護婦と放射線技師の関係について論じたもので、新人の看護婦ほど技師との相関は高く、看護婦の年齢が増すにつれて相関は低くなるというものでした。わたくしはここで得た自己相関関数をフーリエ変換し、病院内

対人パワースペクトルを求めて今後応用することを提唱したいと考えております。どのように応用するかは各自の自由ですが、酒を飲みに行く場合などには比較的低周波領域に的を絞るわけです。高周波成分が混ざりますとエリアシングの影響が無視できなくなり、暗い酒場では高周波成分が美人にみえたりします。これは真実の姿ではありませんので問題ですし、その酒の席も大変ノイジィーなものになります。　　おわかりかな？

なにはともあれ、全国から参加された同業者と語り合えたこと、そして画像理論の大先生方の生の姿を拝見させていただいたことなど、楽しいひとときでした。

犬山中央病院放射線科　　水　野　　裕

平成元年夏、海部新首相が誕生しようとしていた頃、神戸市北部の山中にて、３日間の合宿生活が始まった。午前・午後、各々１テーマ、３時間ずつ、というスケジュールは、私にとっては実にしっくりと馴染むものであった。また講師の先生方には大変失礼で、恐縮せずにはいられない様なラフな服装で参加できた。私たち受講者が、合宿中を快適に過ごし、講義に専念できる様にというスタッフの配慮のお蔭である。

一般に短期間に集中して物事を会得するには、ハングリー精神の様なある種の緊張が必要であると言われる。一方能力開発の研究からは、物事に集中するには、α波が出る状態、つまり心身をリラックスした状態にすることが必要とも言われる。当スクールには、後者の環境が用意され、この選択は正しかった。何故なら、１テーマ３時間という長丁場を絶えずメモを取っていたにもかかわらず、難なく過ごすことができ、終了後も何の疲れも残さなかったからである。そればかりか、講義中は勿論、それ以外の時間、即ち、食前・食後の休憩時間、また深夜に至るまで、絶えず画像や仕事に関する意見の交換が行なわれた。

１日目の夜９時までの講義のあと、私たち受講生は、セミナーハウス内の自販機で仕入れた缶ビールを手土産に、チューター、講師陣の宿舎へ、表慶訪問に行った。そこには、なんとツマミがあるではないか。先生方は不謹慎（？）にも、最初から飲むつもりであったらしい。そして、その場が急きょ「画像について語ろう会・第一夜」になり、当然の様にして翌日も「第二夜」として深夜まで意見の交換が行われたのである。これによって、日頃から画像や仕事について如何に真剣に考えているかをお互いに伺い知ることができた。また気を引き締めて頑張らないと落ちこぼれてしまうのではないかという危機感さえ感じた。

講義内容に関しては、どのテーマも魅力のあるものであった。最も興味深かったのは、画像解析によって診断の一助とするという桂川先生の講義（コンピュータ支援診断システム…）であった。今後私たちの仕事に画像処理、解析技術がどんどん導入され、診断に役立つ情報を作り出す作業が増えていくと思われる。１枚のフィルムの中に病気や病態としての情報が写し込まれているにもかかわらず、人間の眼の識別能の限界の為、私たちが見逃している例が多くあ

るのではないだろうか。心理学の実験等で眼の錯覚とか視点の違いによって、本来の図形と異なった形に見えてしまうことが知られている。目の錯覚の中には、人間が後天的に身に付けた情報があればあるほど見間違いを起こす場合がある。こうなるとRCC解析結果が信用できなくなってしまう。私たち放射線技師は、患者さんの画像情報を誰よりも早く知ることができる。単に撮影するだけでなく、その情報をもとに、より診断に有益な情報を引き出せるような画像処理、解析技術について、今後も勉強を続けていきたい。

香川労災病院放射線科　　守　屋　雅　光

今回の画像リフレッシャースクールは、神戸市北部の山中にある関西地区の大学セミナーハウスで、アカデミックな雰囲気の中で行われました。

私にとって、このような技術学会主催の会に参加するのは始めてのことで、少し不安でありましたが、画像部会長の山下先生をはじめ皆さんの優しい人柄（？）のお蔭で、リラックスした気持ちで三日間を過ごす事が出来ました。

夜学で、山下先生から「私　画像」という題で二分間スピーチをさせられました。私は突然の事でうまく話せませんでした。

今、もう一度「私　画像」の関係を考えてみたいと思います。

放射線技師として、病院に勤務し、あっと言う間に数年が経過しました。私の勤務する労災病院は、大学附属の病院とは異なり、全くの臨床のみの場であります。こういった場で働いていると、臨床的知識はそれなりに経験的に学ぶことが出来ます。それはそれで良いことと思います。

しかし、私達の取り扱っている放射線画像の基本的性質について本当に正確に知っているでしょうか？　遠い昔の学生時代に少しかじってもう忘れ去ったというのが、多くの技師の方の現状ではないでしょうか。（私自身を含めて）

私は、放射線画像工学は、最も技師にとって重要な分野であると思っています。自分達の作る画像の、基本的性質は、我々技師が一番に考えねばなりません。

特に、今回のリフレッシャースクールの主題「アナログ画像からデジタル画像への旅立ち」で歌われているように、今後、放射線画像の多くはデジタル化されていくでしょう。その時、この画像を作る装置は、ブラックボックス化されてしまい、我々技師は単に装置を操作する「手」「足」に成ってしまう可能性が大きいと思えます。

こういった時に、今一度、基本を振り返ってみなければならないと思います。

デジタル系の画像工学は、私のように四国の片田舎に住む者にとってはなかなか触れる事が出来ません。しかしデジタル画像装置は、当院のような所にも数年前から導入されています。CTとかDSAは毎日の臨床の場で盛んに使われています。

今回のリフレッシャースクールで、こういったデジタル画像の基本的性質や将来について、講師の先生方と討論することができ、私にとっての画像を考える良い機会と成りました。

最後に、「私　画像」の関係ですが、まだよく分かりません。私にとって画像は、基本的かつ重要なもので、大変大きな物です。そしてまだ、ほんの一部しか知りません。従って「私⟨⟨画像」としておきます。いつかは「私＝画像」になりたいものです。

あ と が き

きびしい残暑も去り、朝夕に涼しさが感じられるようになりました。８月の中旬から 山下部会長の大阪医療短大にシカゴ大の Metz 教授が客員教授として来日されており、９月から、全国へ講演が続くそうです。今回の画像部会においても、Metz 先生の教育講演を予定しています。ROC の基礎から勉強できる良い機会です。

会費を納めて下さい。

1,０００円です。

学会事務局宛お願いします。

1989 年 10 月 1 日発行

(社)日本放射線技術学会

画像部会々長　山 下 一 也

〒604 京都市中京区西ノ京北壷井町８８

二条プラザ内

ＴＥＬ （０７５）８０１−２２３８