原 著

カテゴリ分類を行わない評価方法
（連続確信度法）を用いた ROC 解析の有用性

白 石 順 二・山 崎 勝・田 中 克 尚

岸 本 健 治・畑 川 政 勝

大阪市立大学医学部附属病院中央放射線部

（論文受理 1993年10月25日）

（最終論文受理 1994年 3 月18日）

（Code No. 520）

Key words: Receiver operating characteristic (ROC), Rating method, Evaluation of radiological image, Categorized data, Non-categorized data

ADVANTAGE OF ROC ANALYSIS USING THE METHOD FOR CONTINUOUSLY-
DISTRIBUTED (NON CATEGORIZED) TEST RESULTS

JUNJI SHIRAISHI, MASARU YAMASAKI, KATSUHISA TANAKA

KENJI KISHIMOTO, MASAKATSU HATAGAWA

Department of Radiology, Osaka City University Hospital

Summary

In the past years, receiver operating characteristic (ROC) studies that has been used in medical imaging system have involved data collected on a discrete confidence-rating scale with five categories, though in principle ROC studies which are called method for continuously-distributed test results can be conducted with continuous scale. In this study, we investigate the advantage of this new rating method. Advantages were obtained from our experiments that used both discrete and continuous rating scale with same sample sets.

From the results of ROC analysis using the CR mammography with simulated mamma phantom, the method for continuously-distributed test results performs identical estimation of area under the ROC curve (A_z) compared with the discrete confidence-rating method, but performs good (1/3) standard deviation of A_z between readers.

When the range of A_z lies 0.65-0.70 which are "difficult" performance of ROC experiment, we concluded that the method for continuously-distributed test results has the advantages of good statistical power and

reliability of results compared with the discrete confidence-rating method.

1. 緒 言

信号検出理論を背景にした ROC 解析は、画像診断の有用性や画像システムの性能を評価できる主観的評価法として、これまで広く用いられてきた¹⁻⁶⁾。ROC 解析に関する方法論や実験時における諸問題についても多くの討論がなされており⁷⁻²⁰⁾、評価法としての信頼度の高さは他の主観的評価法に類を見ない。

画像評価における ROC 解析のデータは、画像の中に存在する微弱な信号（病変）の有無の程度を、観察者が画像（観察試料）を評定することによって得られる。ROC 解析では、この観察者の評定基準の分布は正規分布になると仮定した上で、信号検出理論を展開させ、信号が存在するものを正しく陽性と判断した確率（true positive fraction：TPF）と信号は存在しないのに誤って陽性と判断した確率（false positive fraction：FPF）を統計的に推定する^{1,2,5,17)}。

画像の評定方法としては、これまでは一般に評定確信度法（method for discrete confidence-rating test results）、とくに 5 段階評定法が広く用いられてきた^{2,5,6)}。この方法では、カテゴリ分類されたデータが、評定基準の正規分布の、ある範囲における積分値となるので^{1,2,5,6,17)}、TPF や FPF の算出も容易に行うことが可能

であり、そのための ROC 曲線推定も最尤度比検定により曲線を推定することが可能なプログラムが開発され、広く用いられている^{7,11,13,20)}。しかし、評定確信度法では、実験において、スムーズに ROC 曲線の推定を行うために観察者に対して適当なカテゴリ配分が要求されたり、また、偏ったカテゴリ配分のためにそのデータが無効となる場合が多い¹³⁾。こういった問題を解決するため、また、観察実験以外のデータ（例えば血液検査やコンピュータ支援診断（CAD）システムなどのデータ）にも ROC 解析の応用が可能となるように、カテゴリ分類されていない連続的に分布した実験結果を用いて ROC 解析を行う方法（連続確信度法：method for continuously-distributed test results）が報告された^{20,21)}。

評定確信度法と連続確信度法で与えられた結果に関する評価はこれまでも報告があるが²²⁻²⁵⁾、いずれも比較的に検出率が高い場合の ROC 解析で、得られた結果には評定確信度法と連続確信度法の差が見られない。本論文において、われわれは評定方法だけを評定確信度法と連続確信度法の 2 種類にして行った ROC 実験から、特に検出率が低い場合の ROC 解析の場合において、連続確信度法の有用性に関する知見が得られたので報告する。

2. 評定確信度法と連続確信度法

Fig. 1 Probability density distributions of an observer's confidence rating scale.

Fig. 2 ROC binormal distribution and categorized data

Fig. 1 は、ROC 解析で観察者が評価実験を行った場合に与えられる信号像に対する評価基準の正規分布 ($f(x|s)$) と、雑音像に対する評価基準の正規分布 ($f(x|n)$) を示している。5段階評価による評価確信度法では、各カテゴリごとに観察者が個々に設定する評価基準 (X_c) によって画像を評価する基準が5つに分割される (Fig. 2)。評価確信度法の評価実験によってカテゴリ分類された結果は、この5つに分割された個々のカテゴリ領域の積分値に等しい。そして、TPF と FPF は両正規分布の全体の積分値と、個々のカテゴリ領域の積分値の和との比から与えられる¹⁾。ここで、ROC 曲線はある評価基準における FPF に対応する TPF をプロットすることにより描かれるので、観察者の信号像および雑音像に対する評価基準の分布の形と位置が一定であると仮定した場合、カテゴリ領域を決定する個々のカテゴリにおける評価基準を変化させても、算出される TPF および FPF は、同一の ROC 曲線を描く点となることがわかる²¹⁾。

連続確信度法では評価実験においてカテゴリ分類を行わずに、単に観察試料の順位づけを行うために得点を配点する。これは、前述のとおり、ROC 曲線がカテゴリの領域範囲には関係なく、観察者の評価基準の両正規分布の形と位置にだけ依存するため、順位づけされたデータからその両正規分布が推測できれば ROC 解析を行うことができるという理論に基づいている²¹⁾。連続的に分布したデータから両正規分布を推測する方法には、Metz らによって LABROC1 と LABROC4 の2種類のアルゴ

リズムが開発されたが、膨大な計算を必要とする LABROC4 とパソコンで計算実行の可能な LABROC1 との計算結果に大きな差がないので、現在では LABROC1 が一般的に実用化されている。

LABROC1 では、連続的に分布した評価結果（得点づけされた結果）から評価基準の両正規分布を推定するために、まず、信号像および雑音像の得点を一緒にして、得点順に並べる。例えば、100点満点で評価した信号像5枚と雑音像5枚の試料があって、信号像5枚の得点が86, 74, 45, 68, 32で、雑音像5枚の得点が47, 23, 61, 15, 9である場合には、86, 74, 68, 61, 47, 45, 32, 23, 15, 9の順となる。次に、これを得点の大きさは関係なしに、信号像の得点をS、雑音像の得点をNと表示し直すと、この配列はSSSNSSNNNとなる。最後はこの配列を試料枚数に応じて分割する。この分割された数が評価確信度法で言うところのカテゴリ数で、それぞれの分割された1単位に含まれる信号像および雑音像の枚数がカテゴリデータとなり、最終的にはデータはROCFITに入力され^{13,21)}、評価確信度法と同様の統計的検定が行われる。つまり、前述の例では、2つに分割された場合、前半のカテゴリ1では信号像3枚、雑音像2枚であり、後半のカテゴリ2では信号像2枚、雑音像3枚となる。実際の実験では試料枚数が増えるので、分割数は10または20となる²¹⁾。

3. 実験方法

Fig. 3 Test object of simulated CR mammography

観察試料には疑似ファントムを撮像したCRマンモグラフィを用いた。疑似ファントムによる観察試料はFig. 3に示すように、模擬乳腺として線状に切断したフィルムと模擬腫瘍として1.5～2.5 cm ϕ の不均一円状に切断したフィルムを構造ノイズにして、模擬微小石灰化として20～200 μ m ϕ の大きさに砂を砕いたものを信号とした。信号の大きさは臨床的に存在する微小石灰化の大きさ80～200 μ m ϕ とほぼ一致するようにして、光学顕微鏡にて大きさの確認を行った。信号像には模擬微小石灰化を含む全てのファントムを存在させ、雑音像には模擬乳腺と模擬腫瘍だけで模擬微小石灰化は存在させなかった。散乱体にはアクリル板4 cmを用い、これで疑似ファントムを挟むようにしてフィルム上に置いた。試料サイズは6 cm $\times$ 8 cmで、拡大率1:1の6切CRで鉛マスクを使用して4分割し、1度に4枚の試料を同時に作成した¹⁶⁾。

使用装置はPHILIPS社MammoDIAGNOST UCで、IPにはST-IIIを使用した。IPにマンモグラフィ用のHRではなくST-IIIを用いた理由は、HRを用いた予備実験で、信号の検出率が非常に高くなってしまい、今回の実験の目的に適合しないと考えられたからである。

撮影条件は25 kV 32 mAsでFixモード、周波数強調度の設定は、同一の生データから画像を再構成しているので、周波数強調が観察者に分かる程大きいと、読み取り順序で観察者に予備知識を与えることになるので¹³⁾、通常使用している周波数ランク3、周波数強調度1.0および2.0の差の少ない2種類について比較実験を行った。

試料枚数は、一種類の周波数強調度について信号像50枚、雑音像50枚の100枚、全部で200枚を作成した。

評価方法は従来からの評価確信度法（5段階評価法）と試料1枚1枚について得点を配点する連続確信度法（100点満点法）で、それぞれについて、評価実験前に十分なトレーニングを行った。100点満点法については、一応、50点を中間として一番信号の存在がわかると思われるものが100点、絶対に信号がないと思われるものが0点となるような目安で評価するようにトレーニングしたが、基本的に得点の順位だけが問題となるので、特に配点の上限や下限、配点の間隔に関しては、5段階評価法の場合ほど厳しくはしなかった。

観察者は経験年数1～27年の放射線技師6名で、読み取り順序効果を考慮して観察者群を2つのグループに分けた。200枚の試料は2種類の周波数強調度が均等に混在するように、しかし、同一の生データからの画像が同一セットとならないように、2つの試料セットに組み替えた。各試料セットの中の試料の配列はランダムにして評価順序を変化させ、2つの観察者グループがそれぞれに異なった評価方法で評価実験を始め、次の評価実験までに2週間以上の間隔を開けることで、読み取り順序効果を排除した¹³⁾。

2つの評価実験から得られたデータを用いてROC曲線およびその曲線下の面積Azを推定する方法として、5段階評価法にはROCFIT¹³⁾、連続確信度法にはLABROC1²¹⁾を用いた。

また、各観察者から得られたROCデータの平均は各FRFにおけるTPFの平均値から求めるアベレージ法を用いた¹³⁾。

4. 結 果

Fig. 4 4 observer's ROC curves obtained from 5-category discrete scale.

Fig. 5 4 observer's ROC curves obtained from the method for continuously-distributed test results.

Fig. 4 は5段階評価法で求められた2つの周波数強度における ROC 曲線を示している。6名の観察者のうち、平均値より大きく値の良かった観察者および ROC 曲線を推定するために不適当な評価結果を示した観察者計2名のデータは、今回の実験結果には含めなかった。周波数強度度を1.0から2.0にすることで観察者の半分は ROC 曲線が良くなったが残りの半分は逆に悪くなった。

Fig. 5 は100点満点法の結果を LABROC1 に入力して求められた2つの周波数強度度における ROC 曲線を示している。5段階評価から得られた結果と、周波数強度度による影響で同じ傾向となったが、観察者間の変動が小さくなっているのがわかる。

Fig. 6 に示す5-1.0、5-2.0は5段階評価法から得られた ROC 曲線であり、Free-1.0、Free-2.0 が連続確信度法から得られたもので、それぞれ、2種類の周波数強度度を用いている。各曲線はそれぞれ4名の観察者からアベ

Fig. 6 ROC curves of each rating method 5-category and method for continuously-distributed test results and two enhancement factors of frequency power obtained with the averaged data of 4 observers.

Table 1 ROC areas (Az) of each rating method 5-category and method for continuously-distributed test results and two enhancement factors of frequency power obtained with 4 observers and the average of them.

	5-category		Continuous scale	
RE=1.0	Az	SD	Az	SD
Observer A	0.700	0.054	0.661	0.053
Observer B	0.648	0.063	0.613	0.057
Observer C	0.559	0.059	0.621	0.055
Observer D	0.560	0.059	0.639	0.054
Ave	0.617	0.059	0.633	0.055
SD	0.070	0.004	0.021	0.001
RE=2.0				
Observer A	0.633	0.057	0.610	0.056
Observer B	0.712	0.070	0.659	0.054
Observer C	0.562	0.059	0.596	0.056
Observer D	0.600	0.060	0.605	0.056
Ave	0.627	0.061	0.617	0.055
SD	0.064	0.006	0.028	0.001

レンジ法で平均して求めたものである。観察者の平均で見ると、個々の ROC 曲線に周波数強調度、評定方法の違いによる相関性は認められなかった。

Table 1 は Fig. 6 の ROC 曲線から求めた各周波数強調度における ROC 曲線下の面積 (Az) である。Az の値に関しても周波数強調度、評定方法の違いによる相関性は認められなかったが、各観察者の平均を求めた場合の Az のばらつきを示す標準偏差が、2 つの周波数強調度共に、5 段階評定法では連続確信度法の約 3 倍の値を示した。

5. 考 察

周波数強調度を 1.0 から 2.0 に変化させた場合の ROC 曲線は、観察者間で評価が分かれた。これは、周波数強調度が大きくなることによりエッジ強調が行われ、信号が模擬乳腺と重ならず模擬腫瘍部だけに存在する場合には検出率は上がるけれども、逆に模擬腫瘍部と模擬乳腺とが重なった部分には、信号がないのに偽の信号像があるようにも見られるため、この真陽性と偽陽性を生じさせる両者の影響が観察者によって違う傾向をもたらしたと考えられる。

CR マンモグラフィでの周波数強調度に関するこの考察は、使用している IP が HR に比べて解像力の低い ST-III であり、強調度の変化も少ないので、臨床的な評価と直接的に関連づけることはできない。しかし、人間の視

覚は潜在的に、微小石灰化のような小さい高信号と、乳腺や腫瘍のように比較的大きな低信号のものに対して、異なった特性を持つので¹⁾、今回の実験結果は周波数強調の臨床応用に関しての基礎的なデータになると考えられる。

以上のことから、CR マンモグラフィの周波数強調を大きくすることは、間違って信号のないものを信号ありと判断する偽陽性の割合を大きくするので、特に腫瘍があつて微小石灰化だけの検出を目的とする精査の場合には不適當であるということが示唆される。

評定確信度法と連続確信度法とでは、観察者の平均値による結果では有意差が認められなかったが、観察者間のばらつきにおいて大きな差が見られた。**Table 2** は本論文と同様に、評定確信度法と連続確信度法との評定方法の差を実験により求めたもので、Rockette らによって報告された結果である²⁵⁾。この報告では観察試料に 95 症例の CT 画像を用いている。そして、Az の平均値は、われわれの実験と異なり、0.88 と高い値を示している。この値はこの実験が病変の検出率の高いシステムについて行われたことを示しており、得られた Az の変動は統計的に少ない方に偏っていると考えられる。この文献の中で、Rockette らは連続確信度法の利点について、評定確信度と比べた場合の観察者の学習の容易さと、不適當な評定実験結果の減少を挙げている。われわれの実験の結果からは、それに加えて明らかに観察者間の変動の減少

Table 2 Comparison of reader-specific ROC areas (Az) for the detection of abdominal abnormalities obtained with discrete and continuous scales.²⁵⁾

	5-category discrete scale		Continuous scale	
	Az	SD	Az	SD
Observer A	0.90	0.03	0.88	0.04
Observer B	0.92	0.03	0.92	0.04
Observer C	0.86	0.04	0.87	0.04
Observer D	0.85	0.04	0.89	0.03
Observer E	0.88	0.04	0.86	0.02
Average	0.882	0.036	0.884	0.034
SD	0.029	0.005	0.023	0.009

5-category discrete scale

	1	2	3	4	5
TP	1	17	2	23	7
FP	0	26	8	16	0

Continuous scale

	1	2	3	4	5
TP	8	8	10	11	13
FP	12	12	10	9	7

Fig. 7 Category data and ROC curves of both rating methods discrete and continuous scale obtained from same observer.

という利点が挙げられる。

ROC 実験における観察者間の変動が、検出率が高い場合よりも低い場合の方が大きくなることは、Metz らからも指摘されており²¹⁾、このことが2つの実験結果の差となっていると考えられる。

Fig. 7 に、今回の実験に参加した一人の観察者の評価確信度法のカテゴリデータと、100点満点法から手計算で連続確信度法のアルゴリズムに従って5つに分割して求めたカテゴリデータ、およびそれぞれの ROC 曲線を示す。この観察者は評価確信度法では、カテゴリ3：〈わからない〉の回答が少な過ぎたために、信号像と雑音像に対する両正規分布が重複する形となってしまう、ROC 曲線では逆S字状を描くような結果となってしまう。ところが、連続確信度法（手計算で求めた5段階評価）で

はたとえ50点前後の値が少なかったとしても、得点の順位だけで各カテゴリには均等に評価結果が振り分けられるので、有用な ROC 曲線が与えられている。このことは、連続確信度法が多数の観察者の変動を減少させるだけでなく、これまでトレーニング不足や、評価基準の不適正により除外されていた評価実験の結果を有効に生かせることができると考える。

6. 結 論

連続確信度法による ROC 解析は、これまで ROC 解析に用いられなかったカテゴリ分類ができない血液検査や CAD システムのデータにも有効であり、すでにそれらに関連した研究も報告されている^{26,27)}。また、評価実験の手続きの簡略化にも有用であるのはもちろんのこと、今回

の実験結果から、特に低い検出率を対象とする ROC 解析では、観察者間の変動を少なくし、より有用性の高い実験結果が得られるという利点をもち、これまでの 5 段階評定に比べて、より有効な方法であることが分かった。

7. 謝 辞

本論文をまとめるにあたり、実験に参加し、また多くの援助を頂いた大阪市立大学医学部付属病院中央放射線部の技師諸兄に感謝します。また、連続確信度法が未公開であった時期に、理論および実験方法について検討する機会を与えて頂いた岐阜大学の藤田広志助教授および大阪大学の山下一也助教授に感謝いたします。

最後に、本論文で用いた ROC 曲線推定プログラムである ROCFIT および LABROC1 の使用に関して快く承諾を頂き、また、実験に関しても多くの助言や指導を頂いたシカゴ大学の Metz 教授に感謝いたします。

The authors wish to acknowledge Dr. Charles E. Metz at the university of Chicago for his assistance in using his software and for his technical advise.

本論文の要約は第 21 回秋季学術大会にて発表した。

また、ROC 解析用 ROCFIT は、日本放射線技術学会画像部会、Metz's ROC Software Users Group より配布されたものを用いた。

参考文献

- Goodeough, D.J, Rossmann K, Lusted LB.: Radiographic applications of signal detection theory. *Radiology*, 105, 199-200, (1972).
- Metz CE, Goodeough, D.J, Rossmann K.: Evaluation of receiver operating characteristic curve data in terms of infomation theory, with applications in radiography. *Radiology*, 109, 297-303, (1973).
- Goodeough, D.J, Metz CE, Lusted LB.: Caveat on use of the parameter d' for evaluation of observer performance. *Radiology*, 106, 565-566, (1973).
- Goodeough, D.J, Rossmann K, Lusted LB.: Radiographic applications of receiver operating characteristic (ROC) curves. *Radiology*, 110, 89-95, (1974).
- 山下一也, 若松孝司: 信号検出理論の X 線撮影系への適用. *日放技学誌*, 33(2), 111-117, (1977).
- 山下一也, 若松孝司: ROC 曲線による増感紙-フィルム系の評価. *日放技学誌*, 33(2), 156-160, (1977).
- Swets, J.A, Pickett, R.M.: Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press. (1982).
- Hanley JA, McNeil BJ.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29-36, (1982).
- 若松孝司: 放射線医療画像の解析に関する問題点とその展望 (ROC 曲線による画像評価). *日放技学誌*, 40(2), 259-265, (1984).
- Metz CE: Evaluation of image quality by ROC analysis: Concepts, techniques, and future possibilities. *日放技学誌*, 41(6), 990-1002, (1985).
- Metz CE.: ROC methodology in radiologic imaging. *Invest. Radiol.*, 21: 720-733, (1986).
- Hanley JA.: Receiver operating characteristic (ROC) methodology: the state of the art. *Crit. Rev. Diagn. Imaging*, 29: 307-335, (1989).
- Metz CE.: Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest. Radiol.*, 24: 234-245, (1989).
- Gur D, King JL, Rockette HE, et al.: Practical Issues of experimental ROC analysis—Selection of controls—. *Invest. Radiol.*, 25: 583-586, (1990).
- Gur D, Rockette HE, Good WF, et al.: Effect of observer instruction on ROC study of chest images. *Invest. Radiol.*, 25: 230-234, (1990).
- 白石順二, 山下一也: ROC 評定における観察距離と観察時間—単純な信号の場合—. *日放技学誌*, 47(11), 1937-1942, (1991).
- Chesters MS.: Human visual perception and ROC methodology in medical imaging. *Phys. Med. Biol.*, 37(7): 1433-1476, (1992).
- 山下一也: 画像評価法[III]信号検出論による画質評価. *日放技学誌*, 48(12), 2081-2092, (1992).
- Metz CE.: Quantification of failure to demonstrate statistical significance. *Invest. Radiol.*, 28: 59-63, (1993).
- Herrmann C, Buhr E, Hoeschen D, Fan SY.: Comparison of ROC and AFC methods in a visual detection task. *Med. Phys.*, 20: 805-811, (1993).
- ディジタル画像の ROC 解析検討班: ROC 解析の基礎と最近の進歩. *日放技学誌*, 49(9), 1685-1703, (1993).

カテゴリ分類を行わない評価方法（連続確信度法）を用いた ROC 解析の有用性（白石・他）

- 21) Metz CE, Shen JH, Herman BA. : New methods for estimating a binormal ROC curve from continuously-distributed test results. Presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim CA, August 7, (1990).
- 22) Rockette HE, Obuchowski NA, Gur D. : Nonparametric estimation of degenerate ROC data sets used for comparison of imaging systems. Invest. Radiol., 25 : 835-837, (1990).
- 23) Rockette HE, Gur D, Cooperstein LA, et al. : Effect of two rating formats in multi-disease ROC study of chest images. Invest. Radiol., 25 : 225-229, (1990).
- 24) Rockette HE, Obuchowski NA, Gur D. : Nonparametric estimation of degenerate ROC data sets used for comparison of imaging systems. Invest. Radiol., 25 : 835-837, (1990).
- 25) Rockette HE, Gur D, Metz CE. : The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. Invest. Radiol., 27 : 169-172, (1992).
- 26) Asada N, Doi K, MacMahon H, et al. : Neural network approach for differential diagnosis of interstitial lung diseases. Proc. of SPIE (Medical Imaging IV : Image Processing), 1233, 45-50, (1990).
- 27) Asada N, Doi K, MacMahon H, et al. : Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases : Pilot study. Radiology, 177, 857-860, (1990).