《原　著》

肝シンチグラムの臨床的有効度の定量的評価

—— (2)　SOL 診断の医師間変動の解析 ——

松本　　徹*　　飯沼　　武*　　舘野　之男*　　町田喜久雄**

要旨　第1報では肝シンチグラムによる SOL 診断の正診率が種々の因子によりどのような影響をうけるかを検討したが，本報においては，読影結果の医師間変動を解析することにより，肝シンチグラム検査の信頼性を定量的に評価した．まず，ROC 曲線の医師間変動の程度や傾向を測定し，その結果より医師間変動が小さく，SOL 検出効率が最大になる最適動作点が ROC 曲線上のどの辺に存在するかを示した．次に各症例に対する読影結果のばらつきから，「診断の正確さ」や「診断のつけ易さ」が症例の質に依存することを明らかにした．また，2人の医師の読影のばらつきから読影の個人差の程度を計量し，最後に，読影に個人差がある2人の医師をうまく組み合わせると，SOL に対する診断能が向上する可能性があることを示した．

I. 緒　　言

人間を介した診断系の性能を評価するのに ROC (receiver operating characteristic) 解析が有用であることは一般によく知られている[1]．われわれは第1報[2]で肝シンチグラムによる SOL (space occupying lesion) 診断の ROC 解析を行ったが，ここでは医師11人の平均的な正診率のみを問題とし，医師間の正診率のばらつきやその原因などについては十分検討しなかった．

しかし，医師間の正診率のばらつきは各医師が本来もっている SOL 診断能の優劣によるばらつきのほか，シンチグラム読影時の医師の心理的状態や環境条件の変化などによって生じた誤差等いろいろなものを含んでおり，これらを測定することは肝シンチグラム検査の信頼度の目安を知る上で重要と考える．

本報ではまず，11人の医師による SOL 診断の結果のばらつきをいろいろな成分に分けて測定し，肝シンチグラムの臨床的有効度を定量的に評価した．次に読影結果にばらつきのある2人の医師を組み合わせたダブルチェック検査によって診断能の向上がはかれる可能性を検討し，興味ある知見を得たので報告する．

* 放射線医学総合研究所臨床研究部
** 東京大学医学部放射線科
受付：56年12月3日
最終稿受付：57年2月19日
別刷請求先：千葉市穴川4-9-1　(〒260)
放射線医学総合研究所臨床研究部
松　本　　徹

II. 方法および結果

1. 対象症例と読影データ

本報で解析の対象となった症例は，手術や剖検等の肝シンチグラム検査以外の方法により SOL の有無が確定された401例 (SOL 有り124例，無し277例) である．また，読影データは，これらの症例の肝シンチグラム (正面像，右側面像，後面像) を11人の医師がそれぞれ独立に読影して，SOLの有無を「有」，「有疑」，「無疑」，「無」の4段階の確信度で判定した結果を使用した．ただし読影時に臨床情報として患者の性別，年齢，体重，肝機能検査，触診所見が参考にされた．また，医師11人中8人の自施設症例に対する読影データは，解析からはずされた．(方法論の詳細は第1報参照)

2. ROC 曲線の医師間変動

Fig. 1 は SOL 有り群，無し群の症例について，11人の医師が SOL の有り無しを判定したデータから求めた ROC 曲線である．SOL 診断は4段

階の確信度でなされたので本報の ROC 曲線は3点であらわされている．網目の曲線は3点における11人の平均の有病正診率 (true positive) と，無病誤診率 (false positive) を計算し，これらの点を通る平均 ROC 曲線を無病誤診率 0～60% 付近まで外挿したものである．各曲線とも，SOL 診断を確信度の弱いところで行うと，有病正診率，無病誤診率が単調増加する傾向を示している．ROC 曲線の医師間変動を各動作点でしらべてみると，確信度の強いところ (動作点1) では無病誤診率は医師間でほとんど差がなく，有病正診率の差がいちじるしいのに対して，確信度の弱いところ (動作点3) では逆に，無病誤診率の差が医師間で大きく，有病正診率の差は小さい．また，この場合，有病正診率が低い医師は無病誤診率も低く，有病正診率が高い医師は無病誤診率も高い傾向を示している．

一方，Fig. 2 は ROC 曲線が医師間で，Fig. 1 のごとく変動した場合，有徴正診率と無徴誤診率の差が有病率 (prevalence) の関数としてどの程度変動するかを動作点1と3の場合について求めたものである．有徴正診率，無徴誤診率は，ベイズの定理に従って次式のごとく，各動作点ごとに，有病率の関数として計算される．これらの差が大きいほど SOL 診断の性能が良いことを表わす[3]．

有徴正診率

$$P(D+|T+)=\frac{P(D+)\cdot P(T+|D+)}{P(T+|D+)P(D+)+P(T+|D-)P(D-)}$$

無徴誤診率

$$P(D+|T-)=\frac{P(D+)\cdot P(T-|D+)}{P(T-|D+)P(D+)+P(T-|D-)P(D-)}$$

ただし，$P(D+)$: 有病率

$P(D-)=1-P(D+)$: 無病率

$P(T+|D+)$: 有病正診率

$P(T+|D-)$: 無病誤診率

$P(T-|D+)$: 有病誤診率

$P(T-|D-)$: 無病正診率

Fig. 2 に示された実線は11人の平均値である

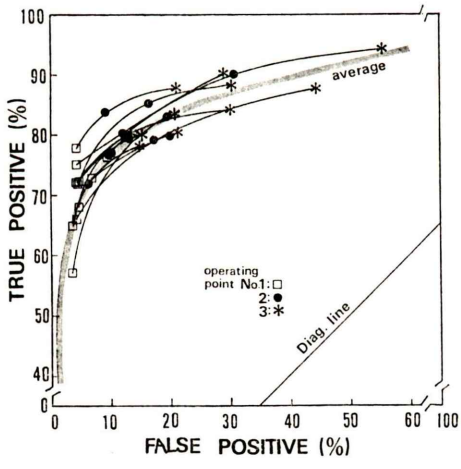

Fig. 1 Interobserver variations of the ROC curve for SOL-diagnosis with liver image.
Points with the symbol (□), (●) and (*) represent the true positive rate (TP) and false positive rate (FP) at the operating point 1 (=positive), 2 (=positive+probably positive) and 3 (=positive+probably positive+probably negative), respectively. The shadow line shows the average ROC curve.

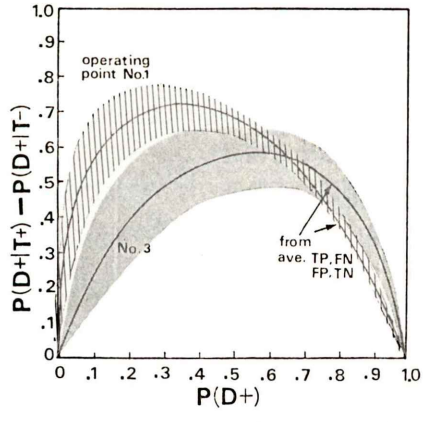

Fig. 2 Posttest probability differences as a function of the prevalence of disease in the case of the operating point 1 and 3. A solid line is a post-test probablity difference calculated from an average TP (No. 1=70.7%, No. 3=84.8%) and FP (No. 1=5.3%, No. 3=26.6%) shown in Fig. 1. A shadow zone shows the inter-observer variations which express the range (maximum-minimum) of the posttest probability differences calculated from the TP and FP of each doctor.

上述の4つのパラメータから〔P(D+|T+)−P(D+|T−)〕を計算したものである．また，斜線部分は11人の各医師についてそれぞれの有病正診率，無病誤診率を用いてこれを計算し，その最大値一最小値間の範囲で医師間変動の程度を有病率の関数としてあらわしたものである．確信度の強いところ（動作点1）では有病率が約0.33のところで11人の医師の平均的な有徴正診率と無徴誤診率の差が極大(0.7)となる．その時の医師間変動〔(最大値一最小値)/2であらわす〕は極大値の±10%程度である．一方，確信度が弱いところ(動作点3)で判定した場合動作点1に相当する縦軸の値は0.5と低く，医師間変動はその±25%と大きい．すなわちFig. 2は本報の有病率〔SOL有り例/(SOL有り例+無し例)〕が0.31であったことを考慮すれば，動作点1でSOL診断を行うと検出効率が高くて，しかも医師間変動の少い成績が得られることを示唆している．

3. 症例ごとのSOL診断の医師間変動

Fig. 3はSOL有り群，無し群に属する各症例について11人中10人の医師（各症例に対して必ず1名分の読影データが自施設データとして除かれるため[2]）によってなされた平均の判定(確信度)と医師間変動の関係を図示したものである．平均の判定はSOL診断の「有」，「有疑」，「無疑」，「無」に対して3, 2, 1, 0のスコアを与えたものか

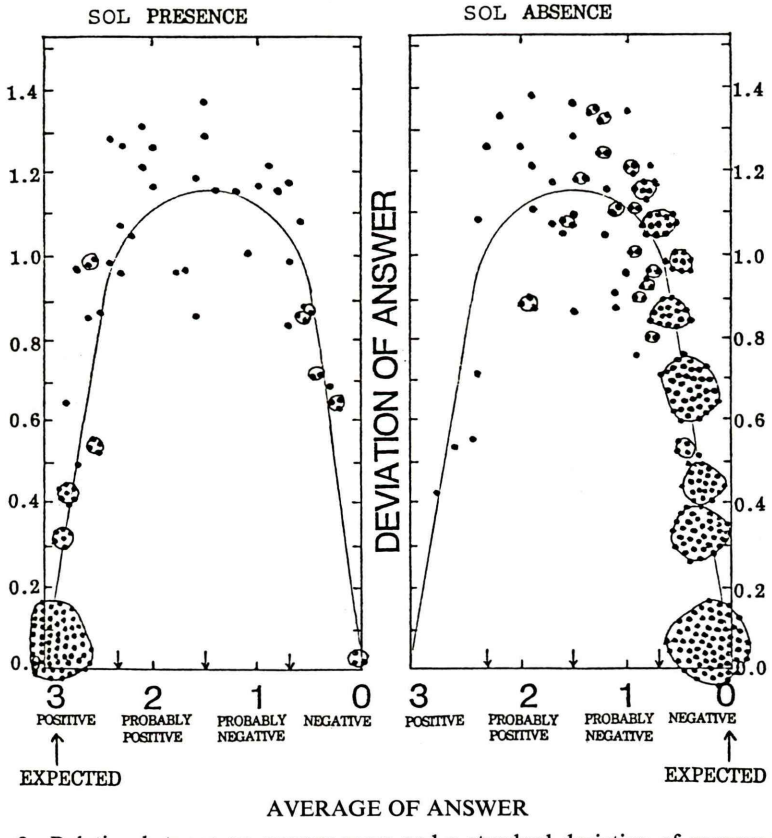

Fig. 3 Relation between an average score and a standard deviation of answers by 11 doctors for a SOL-presence and -absence case. An average score for a case represents an average confidence. A standard deviation correspond to the interobserver variations of score for a case. A solid line is fitted to the observed points by a manual method.

ら求め，症例ごとの診断の医師間変動はスコアの標準偏差であらわしている．この図より，平均の判定が 1.5 の場合，診断の医師間変動は最大となり，両端 (3,0) に近づくほど小さくなることがわかる．すなわち，平均の判定が 1.5 付近の症例は診断がつけにくく，3 か 0 に近いものほど診断がつけやすいといえる．ただし「診断のつけやすさ」と「診断の正確さ」とは一致しない．SOL 有り群に対しては平均の判定が 1.5 より小さく 0 に近い症例，SOL 無し群では 1.5 以上で 3 に近い症例ほど診断はつけやすいが不正確な診断となる．

4. 任意の 2 人の医師による SOL 診断の変動

Fig. 4 は 11 人の医師から任意に 2 人の医師を選びだし，この 2 人により，与えられた SOL 有り群，無し群の各症例のスコアから，SOL 診断の個人差を「繰り返しのない 2 元配置法」[4,5] により計算したものである．

医師 (α) により各症例 (β) について与えられたスコア x は症例の質の差と読影の個人差によって変動すると考える．

$$x=f(\alpha,\ \beta)+e$$

e は偶然誤差とする．規準値 (α_0, β_0) の近傍で $f(\alpha, \beta)$ をテイラー展開し，高次の項を省略すると，

$$f(\alpha,\beta)\fallingdotseq f(\alpha_0,\beta_0)+(\alpha-\alpha_0)\cdot\frac{\partial f}{\partial\alpha}+(\beta-\beta_0)\cdot\frac{\partial f}{\partial\beta}$$

第1，2，3 項をそれぞれ μ，α，β とおくと

$$x_{ij}=\mu+\alpha_i+\beta_j+e_{ij}$$

$$\begin{cases} i=1\sim M\ (M\text{: 医師数}) \\ j=1\sim N\ (N\text{: 症例数}) \end{cases}$$

上式はスコアのデータ構造をあらわす式である．x はスコアがばらつかない部分 (μ) とばらつく部

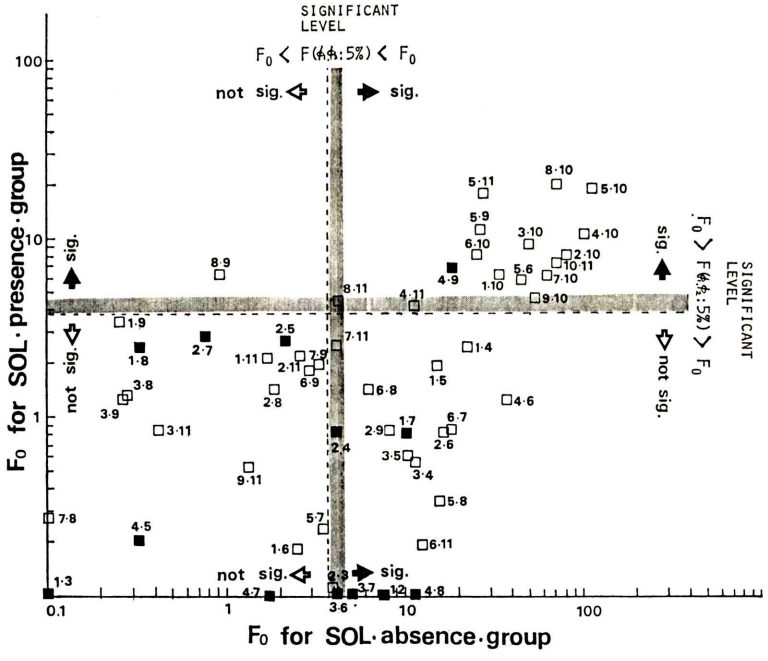

Fig. 4 Variance ratio F_0 calculated by applying the analysis of variance technique for answers (scores) obtained from 2 doctors. When variance ratio F_0 for a pair of any two doctors is greater than/equal to, the value of the shadow zone (significance level F), it is regarded as differences in personality with image reading between 2 doctors being significant for SOL-presence group (124 cases) and SOL-absence group (227 cases), respectively. (Refer to Fig. 6)

分から構成される．後者は読影の個人差 (α_i)．症例の質の差 (β_j)，偶然誤差 (e_{ij}) から成る．ばらつかない部分 μ はスコアの総平均であり α，β，e がないとした時の診断のつけやすさをあらわすものと考えられる．スコアがばらつく部分を解析することにより読影の個人差，症例差に相当するパラメータを計算することができる．具体的な手順は以下のとおりである．

$$w=\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{ij}-\bar{x})^2$$

$$w_1=N\cdot\sum_{i=1}^{M}(\bar{x}_i-\bar{x})^2$$

$$w_2=M\cdot\sum_{i=1}^{N}(\bar{x}_j-\bar{x})^2$$

$$w_3=\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{ij}-\bar{x}_i-\bar{x}_j+\bar{x})^2$$

$$\bar{x}=\frac{1}{M\cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N}x_{ij}$$

$$\bar{x}_i=\frac{1}{N}\sum_{i=1}^{N}x_{ij}\qquad(i=1\sim M)$$

$$\bar{x}_j=\frac{1}{M}\sum_{i=1}^{M}x_{ij}\qquad(j=1\sim N)$$

$$w=w_1+w_2+w_3$$

w はデータ全体のばらつき，w_1，w_2，w_3 は個人差，症例差，偶然誤差によるばらつき (偏差平方和) である．これより，w_1，w_2，w_3 に対応する不偏分散を求め，以下の分散比 (variance ratio) を計算する．

$F_0=(N-1)w_1/w_3$

$F_1=(M-1)w_2/w_3$

F_0 は個人差，F_1 は症例の質の差に原因するスコアのばらつきが偶然誤差のばらつきにくらべてどの程度あるかを表わすパラメータである．これらの分散比は F 分布するので F 検定で有意性を検定することができる．本報では有意水準を危険率 $\varepsilon=5\%$にとり，医師2人 ($M=2$, 自由度1)，症例数 N (自由度 $N-1$) の場合における個人差の有意性を以下により検定した．

$F_0\geqq F(1, N-1: 5\%)$ の時　個人差あり

$F_0<F(1, N-1: 5\%)$ の時　個人差なし

Fig. 4 は2人の医師の55とおりの組み合わせに対して SOL 有り群，無し群ごとに分散比 F_0 を求め，これを2次元的に表示したものである．ただし医師 No. 1, 4, 5, 6, 7, 9, 10, 11 の8名が2人のどちらか，または両方に選ばれた時，その医師 No. に相当する施設の読影データは解析からはずされた．その理由は，この8名が症例提出者であり，自施設症例の内容を記憶している恐れがあったためである[2]．したがって解析対象のデータ数 (N) および症例の内容は医師2人の組み合せごとに少しづつ異り，Fig. 4 に示した有意水準 (危険率) 5% に相当する $F(1, N-1: 5\%)$ 値には多少の幅がある．これより SOL 有り群，無し群とも読影の結果に個人差のある組み合せは29%(16/55) SOL 無し群のみ有意であったのは 35% (19/55) SOL 有り群のみ有意 1.8%(1/55)，SOL 有り群，無し群とも有意差なし，35% (19/55) であった．両群に対して個人差のあった組み合わせ，16組中10組は医師 No. 10 と他の医師の場合であった．

5. ダブルチェックによる SOL 診断の ROC 解析

前節までの検討により肝シンチグラムで SOL 診断を行う時には種々の医師間変動が観察されることを示した．ここでは個人差のある2人の医師による読影スコアを組み合わせた時の ROC 曲線を作成し，ダブルチェックによりどの程度診断能が向上するか，ダブルチェック検査時における医師の最適組み合わせが存在するかどうかを検討した．ただし，ここでいうダブルチェックとは，第1報で述べられたごとく，たがいに独立に (相談し合わないで) 読影されたスコア (「有」，「無疑」，「無疑」，「無」＝3, 2, 1, 0) を以下に示すいくつかの方式で組み合せたものに限る．

(1)　幾何平均型の診断方式[6]

SOL 有り群，無し群ごとに医師 A と B の読影スコアに関して Table 1 のごとく判定の2次元分布を求める．これより10段階の確信度で A と B を組み合わせた時のスコアの頻度分布を作り，確信度の順に累積して有病正診率，無病誤診率を計算し，ROC 曲線を作成する．

Table 1 The double check study by means of the geometric means method. The two dimensional histogram of score for image reading by two doctors is constructed for each group of SOL-presence and -absence. According to the order of 10 confidence levels shown in Table, the number of cases which correspond to the confidence level combined with the scores of both doctors, is calculated. On the the basis of this data, the double check ROC curve is formed.

histogram

Doctor-B	Doctor-A				
3	d	c	b	a
2	g	f	e	b′
1	i	h	f′	c′
0	j	i′	g′	d′
	⓪	①	②	③

confidence level
3 : positive
2 : probably positive
1 : probably negative
0 : negative

order	confidence level	NO.of sample	
		D+	D-
1	③.3	a	the same as left side
2	③.2,②.3	b′ + b	
3	③.1,①.3	c′ + c	
4	③.0,⓪.3	d′ + d	
5	②.2	e	
6	②.1,①.2	f′ + f	
7	②.0,⓪.2	g′ + g	
8	①.1	h	
9	①.0,⓪.1	i′ + i	
10	⓪.0	j	

(2) 算術平均型の診断方式

医師 A と B の，ある症例に対する読影スコアを x_A, x_B とする時，その症例のスコアとして算術平均値 $x=(x_A+x_B)/2$ を使用する．SOL 有り群，無し群ごとにこのスコアの頻度分布を求め，これをもとに ROC 曲線を作成する．

(3) 積極型の診断方式

x_A と x_B のうち確信度の強いスコアを採用し，これをもとに ROC 曲線を作成する．ただし，$x_A=x_B$ のときはそのどちらかをとる．
(If $x_A \geq x_B$, $x=x_A$)

(4) 慎重型の診断方式

x_A と x_B のうち確信度の弱い方を採用して ROC 曲線を作る．ただし，$x_A=x_B$ のときはそのどちらかをとる．(If $x_A \leq x_B$, $x=x_A$)

(5) (2) と (3), (4) を組み合せた方式

医師 2 人の判定がともに確信度 1.5 以上の時 3 に近い判定(積極型の診断)を採用し，医師 2 人の判定がともに確信度 1.5 以下の時 0 に近い判定(慎重型の診断)を採用する．確信度 1.5 を境に意見がわかれた時は (2) の算術平均したものを採用する．〔If $(x_A \,\&\, x_B > 1.5) \,\&\, x_A > x_B$, $x=x_A$. If $(x_A \,\&\, x_B < 1.5) \,\&\, x_A \leq x_B$, $x=x_A$. If $(x_A > 1.5 \,\&\, x_B < 1.5)$ or $(x_A < 1.5 \,\&\, x_B > 1.5)$, $x=(x_A+x_B)/2$.〕

Fig. 5 は医師 No. 1 と 7 のダブルチェックの効果を図示したものである．矢印⬇の曲線は (1), (2) (5) の平均法により作成されたもので結果はほとんど一致していた．この曲線は医師 No. 1, 7 単独のものよりも明らかに ROC 曲線全体が上方に位置しておりダブルチェックの効果が示されている．(3) の積極型の診断方式 (△) では，無病誤診率が高くなっているが，有病正診率も高く単独の ROC 曲線よりも成績が良くなっている．しかし (4) の慎重型診断方式 (○) では無病誤診率，有病正診率とも低く，効果は認められない．(1), (2), (5) は (3) と (4) を平均した成績を示している．な

お図中●の曲線はある症例が SOL 有り群に属していた場合，2人の判定のうち確信度が強いスコアを採用し，SOL 無し群の症例であったら確信度の弱い方を意識的に採用した時の ROC 曲線である．また，下方の曲線▲は SOL 有りの症例である場合，2人の判定のうち確信度の弱いスコアを採用し，SOL 無し例の場合は，確信度の強い方をとってみたものである．このようなことは現実にあり得ないが，前者はダブルチェックにより得られる効果の上限を，後者は最悪の場合(下限)をあらわしている．

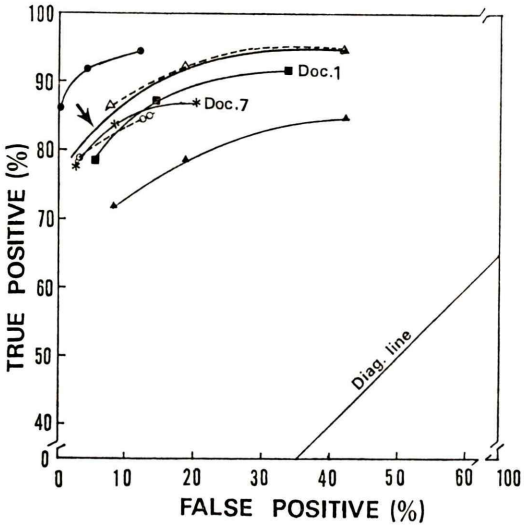

Fig. 5 A example of the results of the double check study. The ROC curve of the symbol (△) corresponds to a double check ROC curve formed by the method adopting a positive judgement out of two answers on a case and the curve of symbol (○) is a double check ROC curve by the method adopting a prudent judgement out of two answers. A solid line represented by the symbol (⬇) shows the double check ROC curve obtained from the average methods (for instance, geometric mean method shown in Table 1), which is equivalent to the average of curves of the symbol (△) and (○). The double check ROC curve is improved rather than that of both observers (No. 1 and 7). The curves of the symbol (●) and (▲) represent the double check ROC curve formed by the method adopting a positive judgement out of two answers for a SOL-presence case and a prudent for a SOL-absence case, and the method adopting a prudent judgement out of two answers for a SOL-presence case and a positive for a SOL-absence case, respectively. These methods are not realistic, but the former shows the best limit line which can be reached by the double check study and the latter shows the worst limit line.

Fig. 6 は55通りの医師 A·B の組み合わせについて (1), (2), (5) の平均法で得られた ROC 曲線のダブルチェック効果の有無を検討したものである．○印は医師 A, B のどちらからみてもダブルチェックの効果があったと判定された組み合わせを示す．ただし効果有りとは，単独の ROC 曲線がそれを構成する3点のうち1点はダブルチェックによる曲線の上にあるが，他の2点はその下方にある場合，または3点とも下方にある場合を指している．効果の程度は，良くても Fig. 5 くらいのものである．☆は医師 A に対して効果有り，★印は医師 B のみに効果があった場合である．他の組み合わせは単独の曲線とほとんど同じか，曲線が交差する場合であり，ダブルチェックの効果が認め

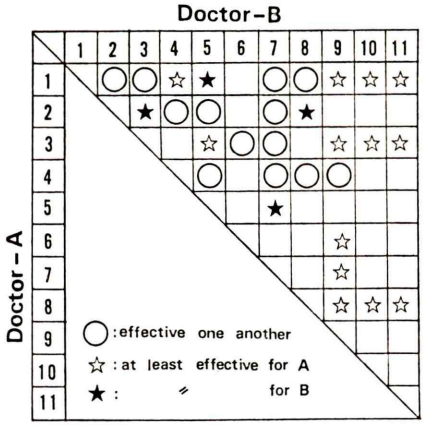

Fig. 6 Combinations of two doctors that were tested for the double check study by means of the average methods shown in Fig. 5 and Table 1. Pairs of the symbol (○) and (☆, ★) represents the case which the double check ROC curve was superior to the ROC curve of both observers and single observer only, respectively. In this study, we found no pairs for whom the double check ROC curve was inferior to the curves of both observers.

られなかったものである．

これより組み合わせて効果があったのは全部で13組 (24%) あり，そのうち医師 No. 4 と他の医師の組み合せが5つ，No. 1, 2 および 7 と他の組み合せがそれぞれ4つあるのが目立っている．また，ダブルチェックの効果と読影の個人差の関係を Fig. 4 でみてみると（■印に相当するもの）医師間の読影の個人差が比較的小さい組み合せに対して効果を生じていることがわかる．

III. 考　察

医用画像による検査の臨床的有効度は正診率と検査結果の信頼性の両面から測られるべきである．今まで医用画像の有効度を測定した論文は多数あるが[7]，その多くは正診率のみを強調し，後者の検討はまったくしていないか，または付加的にしか扱っていない．

検査の信頼性は次の3つの尺度で評価されると考える．

(1) 撮像法のちがいによる検査結果のばらつき，または施設間変動

(2) 読影者間変動

(3) 再現性

すなわち，検査結果の施設間変動や読影者間変動が小さく，再現性が良いほど，信頼性の高い臨床的に有効な検査であるといえる．

肝シンチグラムでは Friz et al,[8] 心筋シンチグラムでは Trobaugh[9]，永井等[10] が撮像法のちがいによる検査結果のばらつきを解析した．読影結果の再現性については pancreatogram で検討した Ruben et al[11] の報告がある．しかし，シンチグラム検査で，これを検討した報告はまだない．肝シンチグラムで Nishiyama et al,[12] 心筋シンチグラムで Curón et al[13] は専門医2人と他科の医師，または研修医や学生 (2, 3人) を対象として，正診率が経験や学習のちがいによりどの程度ばらつくかを検討した．本報では経験に余り差のない多数の核医学専医 (11人) による肝シンチグラム読影結果の医師間変動を解析し，検査法の臨床的有効度を定量的に評価した．

第1報では肝シンチグラムの SOL 検出能（正診率）を医師11人の平均 ROC 曲線として示したが，本報ではまず，ROC 曲線における医師間変動に注目した．検査法の信頼性という観点からいうと ROC 曲線の医師間変動はできるだけ小さいことが望ましいが，実際には Fig. 1 程度の多少のばらつきが観察された．このような変動を生じた原因としては，各医師の SOL 検出能の優劣や診断基準 (いかなる所見を positive，または negative と判定したかの基準) のちがい，等を挙げることができる．今後の検討課題であるが，各医師の診断基準がいかなるものであったかを追跡，調査し，その結果と ROC 曲線の優劣との関係を研究すれば肝シンチグラム読影の際にもっとも重要な診断基準を明らかにできるかもしれない．

Fig. 2 は SOL の有無を判定するのに用いた各動作点で有徴正診率と無徴誤診率の差が有病率の関数としてどう変化するかを推定したものである．ここに示された結果のばらつきは ROC 曲線の医師間変動を反映しており，Fig. 1 と同様の原因により生じたと思われる．また動作点1と3の結果を比較することにより SOL の有無を慎重に判定した方が（動作点1)，積極的に判定する場合（動作点3) よりも，読影実験時の有病率に対して各医師の SOL 検出能が高く，医師間のばらつきは小さいことが理解される．

次にこのような ROC 曲線の医師間変動を生じた原因をさらに詳しく知るため，症例ごとに各医師がどのような判定を下したかを検討した．

Fig. 3 は各症例の「診断のつけやすさ」=「読影結果のばらつき」と「診断の正確さ」=「読影結果の平均値」を示すと同時に，症例の質のちがいに依存して各医師の読影態度がどう変るかをあらわす一般的な傾向をも示唆していると思われる．すなわち，SOL 診断に関して SOL 有りの症例では SOL の大きさや個数が診断のつけやすさを左右する主な因子と考えられるが，それに依存して各医師の読影態度がどう変化するかを Fig. 3 から予測することができる．SOL が多数あるか十分大きければその症例に対する，各医師の判定は迷う

こと少なく，「有」(=3 のスコア) の側に下されるであろう．また，撮像法の検出限界以下の小さな SOL，または少数の SOL しかない症例では同様に迷うことなく「無」(=0 のスコア) の側に判定が下ると思われる．さらに SOL が「有」か「無」かもっとも迷う症例の判定は「有」と「無」の中間に来るものと予測され，その時，症例に存在した SOL の大きさや個数は検査に用いられた撮像法の検出限界をあらわすものと考えられる．

一方，SOL 無し例では診断のつけやすさを左右する因子として，SOL 以外の肝疾患の有無や変形，腫大，萎縮等で代表される肝異常および肝外性異常の有無，さらには生理的欠損やアーチファクトの存在等があり，それらの組み合わせから生じた複雑微妙な臨床的状況のもとで，各医師について SOL 有り例の場合と同様な読影傾向が観察される．また，この時も各医師がもっとも判定に苦しんだ症例の臨床的状態を検討すれば，撮像法の，SOL「無」の検出限界に関与した因子について情報を得ることができると思われる．

このように，医師の読影結果は症例の質に依存して変化することがわかったが，本報ではさらに医師間の読影の個人差についても検討した．Metz et al[14,15] によれば，現在，2 つ以上のROC 曲線の差の有意性を統計的に，厳密に，検定する方法はないといわれている．ROC 曲線の，もとの読影スコアから読影の個人差を計量化したわれわれの研究は，2 つの ROC 曲線のちがいを間接的に検定する 1 つの方法を提案したものと考える．

採点法で得られた評点データに分散分析法を適用し，対象の質の差や個人差を測る方法は工学(官能検査)[16] 分野では実用化されているが，医学分野[5,17] ではまだ余り使われていない．とくに医用画像の読影スコアから読影の個人差を計量化した報告は本報がはじめてである．

読影の個人差の測定は次のような根拠に基づいている．11人の医師の中から選ばれた55組の 2 人の医師について，SOL 有り，または無しの症例群に対する読影スコアのばらつきの大きさを測り，それが偶然誤差と比較して非常に大きければ読影に個人差があるとみなされる．したがって，この結果は直接 ROC 曲線のちがいを検定するものとはいえないが，Fig. 4 の結果を Fig. 1 の ROC 曲線と比較してみると，次のような傾向が観察される．すなわち，Fig. 4 において No. 10 は他のすべての医師との間で読影の個人差のパラメータ F_0 が SOL 有り，無し両群において有意であった．この ROC 曲線を他の医師と比較してみると，曲線全体の形は，他と優劣をつけ難いが，動作点 2, 3 における有病正診率，無病誤診率は医師群中もっとも大きかった．すなわち，No. 10 の医師は SOL を積極的に検出する傾向が強いのに対して他の医師は比較的慎重に読む傾向を示した．一方，SOL 有り，無し両群で読影の個人差が有意でない組み合わせ (1, 3)，(4, 5) 等の ROC 曲線はほとんど同じであった．

このように読影の個人差は ROC 曲線の医師間変動と密接に関係しているので，これを解析することは，2 つの ROC 曲線のちがいを分析し，その有意性を検定する上で非常に有用であると考える．また，Fig. 4～6 で示したとおり 2 人の医師によるダブルチェック検査時の効果を事前に推定するのに (十分ではないが) 必要な情報を提供してくれると考える．

ダブルチェック検査の有用性についてはすでに多くの論文で述べられており[18~20]，日常治療の中でも現実に行われることが多いと聞いている．しかし，本報のごとく，実際の読影データを使ってその効果を系統的に調べた報告はあまり知られていない．

ダブルチェック効果がなぜ生じたのか原因を明らかにすることは医師の最適組み合わせを意識的に作ることと結びつくので重要である．しかし，ここでは医師の読影の個人差やくせのちがいが原因の 1 つとして挙げられたにすぎず，今後さらに検討する必要があると思われる．

IV. まとめ

肝シンチグラムによる SOL 診断の臨床的有効度を定量的に評価するため (1) 11人の医師による

ROC曲線のばらつきを測定した．(2) ROC曲線上の3つの動作点のデータ（有病正診率と無病誤診率）から有徴正診率と無徴誤診率の差を計算し，動作点1（確信度「有」）でSOL有無の判定を行うと，SOL検出効率が大きく，かつ医師間変動の少ない成績が得られることを明らかにした．さらに，(3) 症例ごとの読影結果の医師間変動を解析すると，その症例に対して「診断のつけやすさ」と「診断の正確さ」に関する情報が得られる，(4) 2人の医師によって得られた同一症例群の読影スコアのばらつきから読影の個人差を計量化すれば，ROC曲線の構造を分析し，そのちがいを検定することができる，(5) 個人差のある2人の医師をうまく組み合わせるとダブルチェック検査により，SOL検出効率を向上させることができる，等を指摘した．

謝辞：本研究は日本アイソトープ協会核医学開発委員会（委員長・飯尾正宏），エフィカシー1小委員会（委員長・町田喜久雄）により行われたものである．症例を提出していただいた8施設と肝シンチグラム読影に貴重なお時間を割いて下さった先生方は次の通りである．（敬称略）

内山 暁，堀田とし子，宇野公一（千葉大学医学部），川上憲司（慈恵医大），久保敦司，高木八重子（慶応大学医学部），宍戸文男，舘野之男（放射線医学総合研究所），中島哲夫（埼玉がんセンター），村田 啓（東京都養育院附属病院），山崎統四郎（東京女子医大），町田喜久雄（東京大学医学部）

データ処理に際して，技術的ご援助を賜った放医研・福久健二郎電算機室長に厚く感謝致します．

文 献

1) 飯沼 武：医用画像における臨床的有効度の評価(I). 核医学 **17**: 639–646, 1980. (II) 核医学 **17**: 1035–1043, 1980
2) 松本 徹，飯沼 武，舘野之男，町田喜久雄：肝シンチグラムの臨床的有効度の定量的評価 (1) 方法論とSOL診断のROC解析を中心に．核医学 **19**: 51–65, 1982
3) Hamilton GW, Trobangh GB, Ritchie JL, et al: Myocardial imaging with ^{201}Thallium: An analysis of clinical usefulness based on Bayes' theorem. Seminars in Nuclear Medicine **8**: 358–365, 1978
4) 水野哲夫：臨床統計学，治療評価を中心として．医学書院，東京，1978
5) 河田敬義，丸山文行，鍋谷清治：数理統計，裳華房，東京，1974
6) 村田和彦，松下 哲，田中敏行：虚血性心疾患および心肥大の診断におけるベクトル心電図，直交3軸誘導心電図の価値．成人病の研究，**1**: 38–45, 1972
7) 永井輝夫，平敷淳子：総合画像診断学．丸善，東京，1981
8) Friz SL, Preston DF, Gallaghr JH: ROC analysis of diagnostic performance in liver scintigraphy. J Nucl Med **22**: 121–128, 1981
9) Trobauch GB, Wackers FJ Th, Sokole EB, et al: Thallium 201 myocardial imaging: An interinstitutional study of observer variability. J Nucl Med **19**: 359–363, 1978
10) 永井輝夫：放射性タリウム心筋梗塞イメージの客観的解析，第20回日本核医学会．会長講演要旨，群馬，1980
11) Reuben A, Johnson AJ, Cotton PB: Is pancreatogram interpretation reliable?—a study of observer variation and error. Brit J Radial **51**: 956–962, 1978
12) Nishiyama H, Lewis JT, Ashare AB, et al: Interpretation of radionuclide liver images: Do training and experience make a differences? J Nucl Med **16**: 11–16, 1975
13) Cuarón A, Acero AP, Cárdenas M, et al: Interobserver variability in the interpretation of myocardial images with ^{99m}Tc-labeled diphosphonate and pyrophosphate. J Nucl Med **21**: 1–9, 1980
14) Metz CE, Kronman HB: A test for the statistical significance of differences between ROC curves, Information processing in medical imaging **88**: 647–660, 1979
15) Grey DR, Morgan BJT: Some aspects of ROC curve-fitting: normal and logistic models. J Math Psych **9**: 128–139, 1972
16) 佐藤 信：官能検査入門．日科技連，東京，1978
17) 丹後俊郎：臨床検査値の個人差の計量化について．第9回日本行動計量学会大会発表論文抄録集，名古屋，p 140–141, 1981
18) Tateno Y: Basic problems in integrated body imaging. MEDIX **5**: 1–6, 1979
19) 舘野之男：検査の「診断能」と「有効性」の評価．サクラX-ray写真研究，**31**: 5–9, 1980
20) 岡村一博訳(Galen RS, Gambino SR 著)：正常値以前の問題——医学的診断の予測値と効率．宇宙堂八木書店，東京，1978

Summary

Estimation of Clinical Efficacy for Scintigraphic Images of Liver

(2) A study on Interobserver Variation of SOL Detection by Image Reading

Toru MATUMOTO*, Takeshi A. IINUMA*, Yukio TATENO and Kikuo MACHIDA**

*National Institute of Radiological Sciences
**Department of Radiology, Tokyo University

The purpose of this study is to investigate the clinical efficacy No. 1 (diagnostic accuracy) of liver images on the SOL-disease from a viewpoint of the reliability of diagnosis with liver image. The results of which 11 physicians read the liver images of 401 cases (SOL-presence=124 case, SOL-absence=227 cases) collected from the 8 situations, were analyzed.

The reliability of diagnosis in detecting the SOL by reading liver images may be evaluated from the measurements of variation of finding by 11 doctors. The results of analysis are as follows.

(1) The amount of interobserver variations of ROC curve in detecting the SOL is different at any operating point. An operating point at which the detection efficiency of SOL-disease is large and the interobserver variations is small, may be searched by calculating the posttest probability differences as a function of the prevalence of disease.

(2) By scoring the confidence level of SOL-presence judged by a doctor and calculating an average value and a standard deviation of the scores obtained from 11 doctors, the relation between an average confidence and an interobserver variations of SOL-diagnosis for each case may be estimated.

(3) A deviations of answer gained by any two observers out of 11 are calculated by the analysis of variance technique and differences in personality with image reading are detected.

(4) Finally, it is presented that the double check study by two doctors having the different personality with image reading improves the ROC curve of SOL-diagnosis.

Key words: interobserver variation, liver image, ROC analysis.