

②比例系統誤差と偶然誤差：①と③の混合であり、その対策としては①の対策と③の対策の両方を行う必要がある。

③偶然誤差：測定の際のばらつきに起因する誤差で、本来各機関内における内部精度管理で管理・維持する部分であるが、日常検査が自動分析装置を用いて測定されている現在、測定値のばらつきはきわめて小さいのが通常である。したがって測定値のばらつきが大きくなる原因は多くの場合測定装置のメンテナンスが不十分であるために生じることが多い。対策としては測定装置の定期メンテナンス及び予防的保守管理プログラムを確立し、それを遵守する必要がある。

④一定系統誤差：試料濃度に関係なく一定の大きさ（濃度）で生じる誤差で、代表的な原因の一つに測定に使用する試薬盲検又は検体盲検が不適切な場合及び測定装置の0点調整が不備な場合に起こることが多い。この対策としてはそのかたよりの原因を究明した上で、それを是正する必要がある。

⑤記入ミス又は試料の取り違い：この問題の原因は確認ミスによるものが大半である。各作業ステップにおける確認マニュアルを作成して確認を励行する

ことにより回避できる問題である。この問題は精度管理調査のみでなく、日常検査においても生じる可能性があり、被検者に与える影響は予想以上に大きい。

⑥直線性の欠如：この問題の代表的な原因は、測定法又は測定装置の性能の限界によるもので、その他の原因としては自動分析装置へのアプリケーションの不適切などが考えられる。先ず高値試料を段階的に希釈することにより直線性の確認をするとともに測定法のアプリケーションが適切であるかチェックし、適正な測定法の導入及びアプリケーションを実施する必要がある。

⑦送付における問題点：この問題は参加機関ではなく調査実施組織（機関）に属する問題である。原因としては送付試料へのラベル貼りにおけるミスが考えられ、送付試料の準備段階における確認を励行することにより回避できるミスである。

過去5年間に全衛連調査で生じた無視できない問題として、複数の機関から受託している外部機関の成績が60点未満を示した例があげられる。外部機関(A)は、第11回調査で15機関から受注した中性脂肪の結果がすべて60点未満の成績を示した。生じている誤差として試

料濃度に関係なく参考値に対して各試料とも約10mg/dl高値に測定されており一定系統誤差の存在が示唆された(図1及び表)。このかたよりは盲検及び0点調整の不備により生じる誤差よりはるかに大きく、その原因を詳細に解析したところ、調査試料中に存在する遊離グリセロールが測定過程で完全に消去されていないことがわかり、測定試薬の不備が原因であることが判明した。この原因は直ちに是正され、次年度の第12回調査ではこのような問題は生じなかった。外部機関(B)は、第12回調査で8機関から受注したAST、4機関から受注したALT、7機関から受注した γ -GTの3項目が60点未満の成績を示した。生じている誤差としてAST、ALTについては試料濃度に関係なく参考値に対して+1.96SD程度高値に測定されており一定系統誤差の存在が示唆された(図2、3及び表)。このかたよりはるかにその大きさの程度から盲検及び0点調整の不備の可能性が考えられた。 γ -GTについては典型的な比例系統誤差の存在が示唆された(図4及び表)。これらの誤差の可能性については外部機関(B)にコメントを送付し改善を求めた。外部機関(I)は第

10回調査で2機関から受託している中性脂肪の結果が60点未満の成績を示した。生じている誤差として参考値に対して約6~7%低値に測定されており、その解析から比例系統誤差の存在が示唆された(図5及び表)。この外部機関の結果は第11回、第12回調査ともに良好な成績を示しており明らかに改善されている。

測定誤差は以上述べた種々の原因により生じるが、各機関内における内部精度管理の普及と実践により測定の精密さすなわち偶然誤差はきわめて小さい範囲内に管理・維持できるようになり²⁾、測定値の施設間差は、確認ミスのような偶発的な誤差を除けば、測定値の正確さに関連するかたよりが誤差原因の主流を占めるようになってきている。調査結果からもかたよりの中でも比例系統誤差が大部分を占め、その原因が主として測定の標準となる検量用試料の正確さを含むキャリブレーション時に生じる誤差であることが判明している。これらのかたよりの原因となる比例系統誤差を縮小するには、正確さが明確な高位標準物質を日常一般法で繰り返し測定を行い、その平均値とこれら高位標準物質の標準値の一致性を

確認するトレーサビリティの確認を実施し、もしかたよりが存在したときにはそれを校正する作業を行うことにより達成できる。しかし、現状では高位標準物質を用いてトレーサビリティの確認を実施している機関は著しく少なく、全く実施していない機関が70%も存在することは憂うべき状況である(図6)。この状況は日本医師会が年1回実施している臨床検査精度管理調査の結果においても同様である⁴⁾。測定値の施設間差の原因の大部分が校正可能な比例系統誤差であることは、一方では測定値の信頼性の保証は容易な状態にあることを意味している。現在、定量検査が行われている全ての検査項目について高位標準物質の開発が実現しているわけではないが、比較的件数の多い日常的な検査項目については既にこれらの高位標準物質が存在するので、これらの項目については各機関において少なくとも年1、2回の頻度でトレーサビリティの確認と校正を実施すべきである。それが実現したときには、検査がどこで実施されても同一レベルで比較可能になるだけでなく、個人の検査データの信頼性の高い経年的管理が可能になるであろう。

E. 結語

過去5年間の全衛連臨床検査精度管理調査の結果の中で60点未満の成績を示した項目及び機関について調査し、その原因を解析した結果以下の結論を得た。

1. 5年間に60点未満の成績が1項目以上存在した機関数の割合は、自機関で9.1%、外部機関で7.6%であり、自機関に比べ外部機関は比率が低かったが、外部機関は受託した複数の健診機関に影響を与えている例がありその影響の重大さは計り知ることが出来ない。

2. 誤差原因は種々存在するが、試料濃度に関係なく一定の比率で生じる比例系統誤差が、自機関、外部機関ともに大部分を占めていた。外部機関の一部に、特定の項目において一定系統誤差が見られ、複数の機関に影響を与えていた。

3. 比例系統誤差は、測定値のかたよりが原因であるが、正確さの明確な高位標準物質を用いて確認と校正が可能である。しかしそれを実施している機関は著しく少なく、約70%の機関が全く実施していない。

4. 比例系統誤差は容易に校正可能な誤差であり、少なくとも年 1、2 回はトレーサビリティの確認と校正を行うべきである。以上を実践することにより臨床検査の信頼性向上が果たせるとともに個人の検査データの信頼性の高い経年的管理が可能になる。

参考文献

- 1) 全国労働衛生団体連合会：総合性小戸管理事業に基づく全衛連臨床検査精度管理調査結果報告書平成 11 年度（第 8 回）～平成 15 年度（第 12 回）
- 2) 厚生労働科学研究費補助金労働安全衛生総合研究事業「職域における健康診断と精度管理のあり方に関する研究」平成 14 年度 総括・分担研究報告書（主任研究者 櫻井治彦）平成 15 年 3 月
- 3) 厚生労働科学研究費補助金労働安全衛生総合研究事業「職域における健康診断と精度管理のあり方に関する研究」平成 15 年度 総括・分担研究報告書（主任研究者 櫻井治彦）平成 16 年 3 月
- 4) 日本医師会：平成 15 年度 第 37 回 臨床検査精度管理調査結果報告書 平成 16 年 3 月

表1 過去5年間の自機関における60点未満の項目数および機関数(全衛連精度管理調査)

機関数	1項目のみ	2項目		3項目		合計機関数
		同一調査	異なる調査	同一調査	異なる調査	
	56	6	12	2	3	79

1項目のみ:過去5年間で1項目のみ60点未満であった機関数

2項目:過去5年間で2項目が60点未満であった機関数。同一調査で2項目、異なる調査で合計2項目が60点未満であった機関の2種類が存在する。

3項目:過去5年間で3項目が60点未満であった機関数。同一調査で3項目、異なる調査で合計3項目が60点未満であった機関の2種類が存在する。

4項目以上:過去5年間で4項目以上が60点未満であった機関は存在しなかった。

表2 過去5年間の外部機関における60点未満の項目と影響を与えた機関数(全衛連精度管理調査)

番号	外部機関コード	項目名	影響を与えた機関数	備考
1	A	TG	15	第11回調査
		Glu	1	第11回調査
2	B	AST	8	第12回調査
		ALT	4	第12回調査
		γ -GT	7	第12回調査
3	C	TC、TG	1	第11回調査 2項目とも同一機関
		γ -GT	1	第11回調査
4	D	TG、 γ -GT	1	第11、12回調査 2項目とも同一機関
5	E	AST、 γ -GT、Glu	1	第10回調査 3項目とも同一機関
6	F	Glu	1	第8回調査
7	G	ALT	1	第11回調査
8	H	TC	1	第11回調査
8	I	TG	2	第10回調査
10	J	ALT、Glu	1	第10、11回調査 2項目とも同一機関
		γ -GT、Glu	1	第10、11回調査 2項目とも同一機関
		Glu	1	第11回調査
11	K	TC	1	第11回調査
		Glu	1	第10回調査
12	L	γ -GT	1	第8回調査
13	M	AST	1	第11回調査
14	N	Glu	1	第9回調査
15	O	TC、 γ -GT	1	第9、11回調査 2項目とも同一機関
16	P	TC、TG、AST、ALT、 γ -GT	1	第8、10、11回調査 5項目とも同一機関
17	Q	TG	1	第10回調査
18	R	AST、Glu	1	第12回調査 2項目とも同一機関
19	S	Glu	1	第12回調査
20	T	AST	1	第12回調査
合計			58	

表3 過去5年間の機関別、項目別誤差原因の分類および合計件数(全衛連精度管理調査)

自機関測定

	PSE	PSE&RE	RE	CSE	転記/試料	直線性	送付	合計件数
総コレステロール	20	3	0	0	3	0	0	26
中性脂肪	14	0	0	5	1	0	0	20
AST	10	1	0	0	1	1	0	13
ALT	3	0	0	1	1	1	0	6
γ-GT	14	0	0	0	2	0	1	17
血糖	20	0	0	1	2	1	0	24

外部機関測定

	PSE	PSE&RE	RE	CSE	転記/試料	直線性	送付	合計
総コレステロール	2	0	2	0	1	0	0	5
中性脂肪	4	0	1	15	1	0	0	21
AST	4	0	1	8	0	0	0	13
ALT	2	0	0	4	1	0	0	7
γ-GT	9	0	2	0	1	1	1	14
血糖	7	0	1	1	1	0	0	10

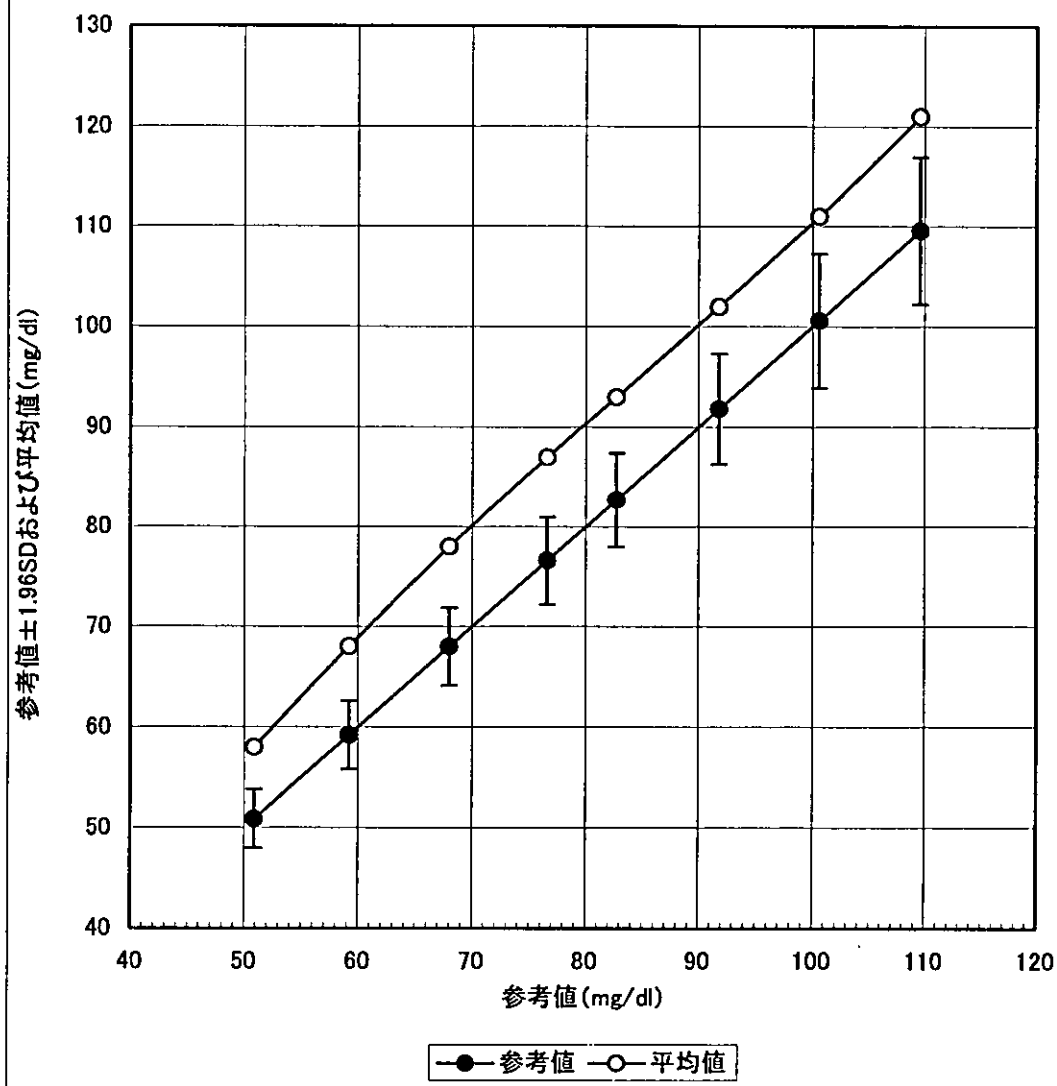
PSE: 比例系統誤差、PSE&RE: 比例系統誤差&偶然誤差、RE: 偶然誤差、CSE: 一定系統誤差、
 転記/試料: 転記ミスまたは試料の取り違い、直線性: 直線性の欠如、送付: 送付における問題

表4 定量検査における誤差の原因と対策

	誤差の種類	主な原因	対策
①	比例系統誤差(PSE)	使用する検量用試料の正確さのかたより	認証標準物質を用いてトレーサビリティの確認と校正
②	比例系統誤差と偶然誤差(PSE & RE)	①と③の混合	①の対策と③の対策の両方を行う
③	偶然誤差(RE)	測定装置のメンテナンス不良	定期及び予防的保守管理法の確立と遵守
④	一定系統誤差(CSE)	不適切な盲検又は0点調整の不備	原因の究明と是正措置
⑤	記入ミス又は試料の取り違い	確認ミス	確認の励行
⑥	直線性の欠如	測定法(装置)の性能の限界又は他の原因	高値試料の段階希釈による確認及び測定法の検討と適正化

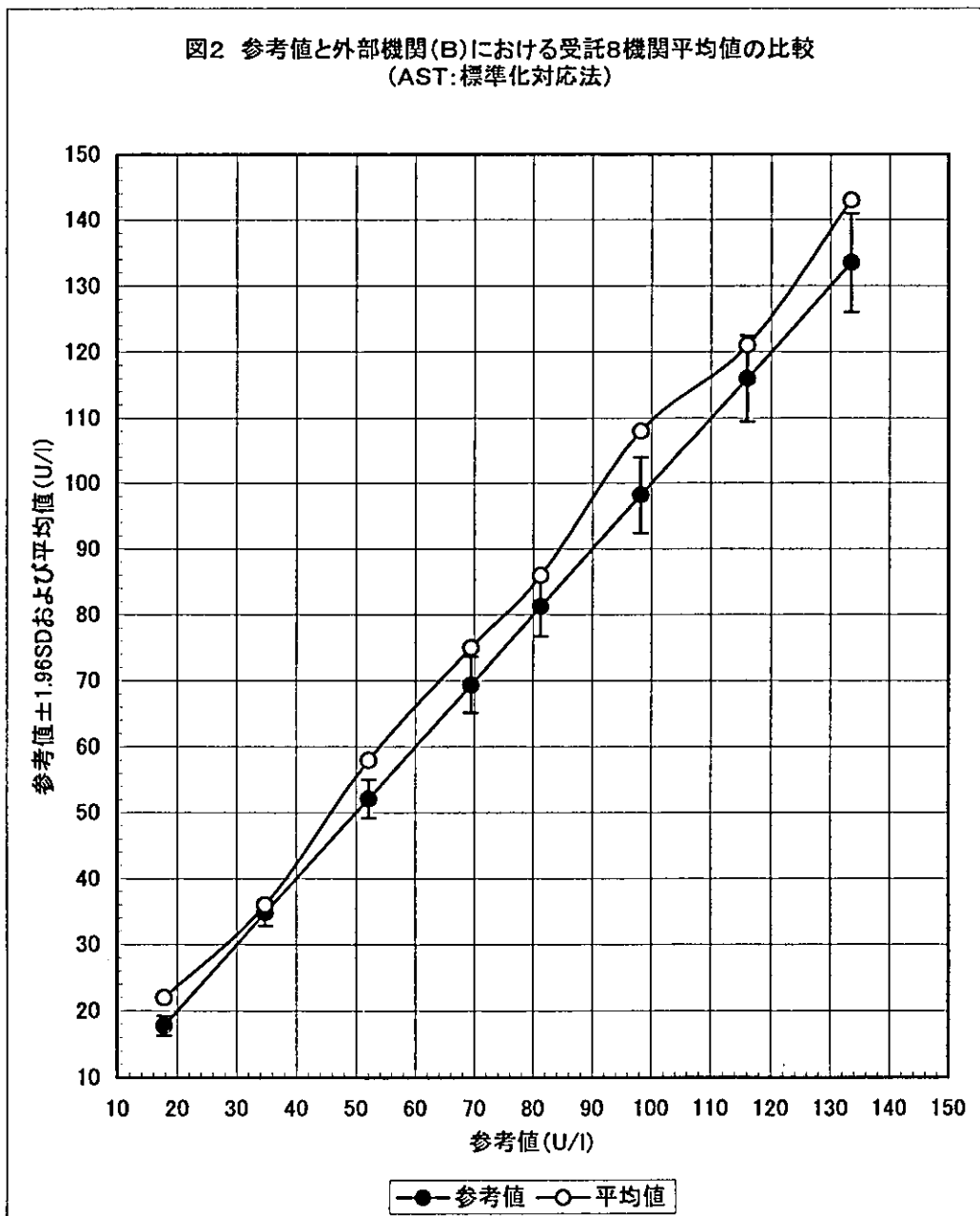
第 11 回 調査		試料 1	試料 2	試料 3	試料 4	試料 5	試料 6	試料 7	試料 8
外部機関(A)	参考値	50.9	59.2	68	76.6	82.7	91.8	100.6	109.6
	標準偏差	1.49	1.73	1.98	2.24	2.41	2.81	3.4	3.75
受託 15 機関	平均値	58	68	78	87	93	102	111	121
	SDI	4.8	5.1	5.1	4.6	4.3	3.6	3.1	3.0
中性脂肪	比率	113.9	114.9	114.7	113.6	112.5	111.1	110.3	110.4
	差	7.1	8.8	10.0	10.4	10.3	10.2	10.4	11.4

図1 参考値と外部機関(A)における受託15機関平均値の比較
(中性脂肪:グリセロール消去酵素比色法)



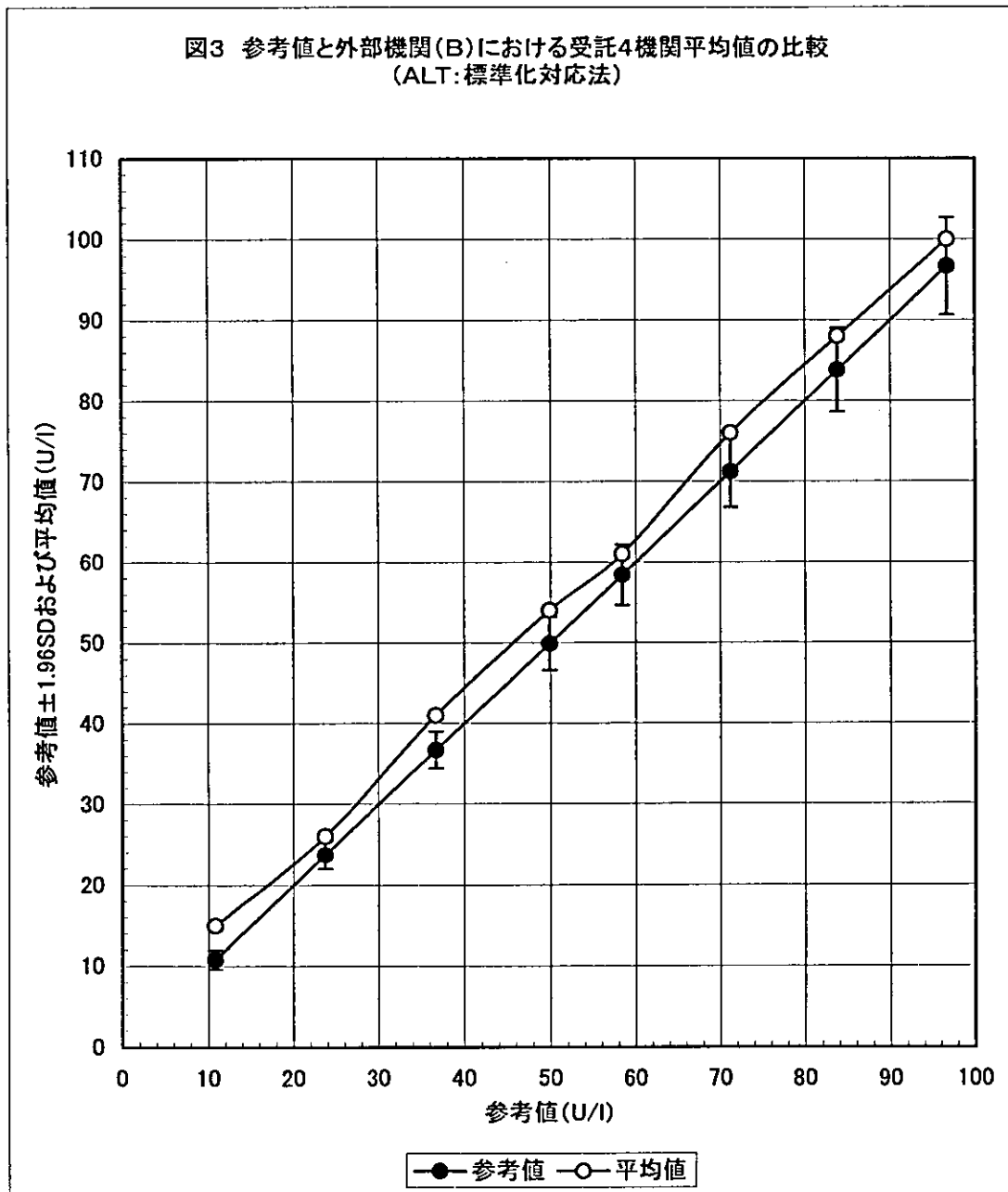
第12回調査		試料1	試料2	試料3	試料4	試料5	試料6	試料7	試料8
外部機関(B)	参考値	17.8	34.8	52.1	69.4	81.3	98.2	116	133.5
	標準偏差	0.75	0.99	1.49	2.17	2.32	2.95	3.32	3.82
受託8機関	平均値	22	36	58	75	86	108	121	143
	SDI	5.6	1.2	4.0	2.6	2.0	3.3	1.5	2.5
AST	比率	123.6	103.4	111.3	108.1	105.8	110.0	104.3	107.1
	差	4.2	1.2	5.9	5.6	4.7	9.8	5.0	9.5

図2 参考値と外部機関(B)における受託8機関平均値の比較
(AST:標準化対応法)



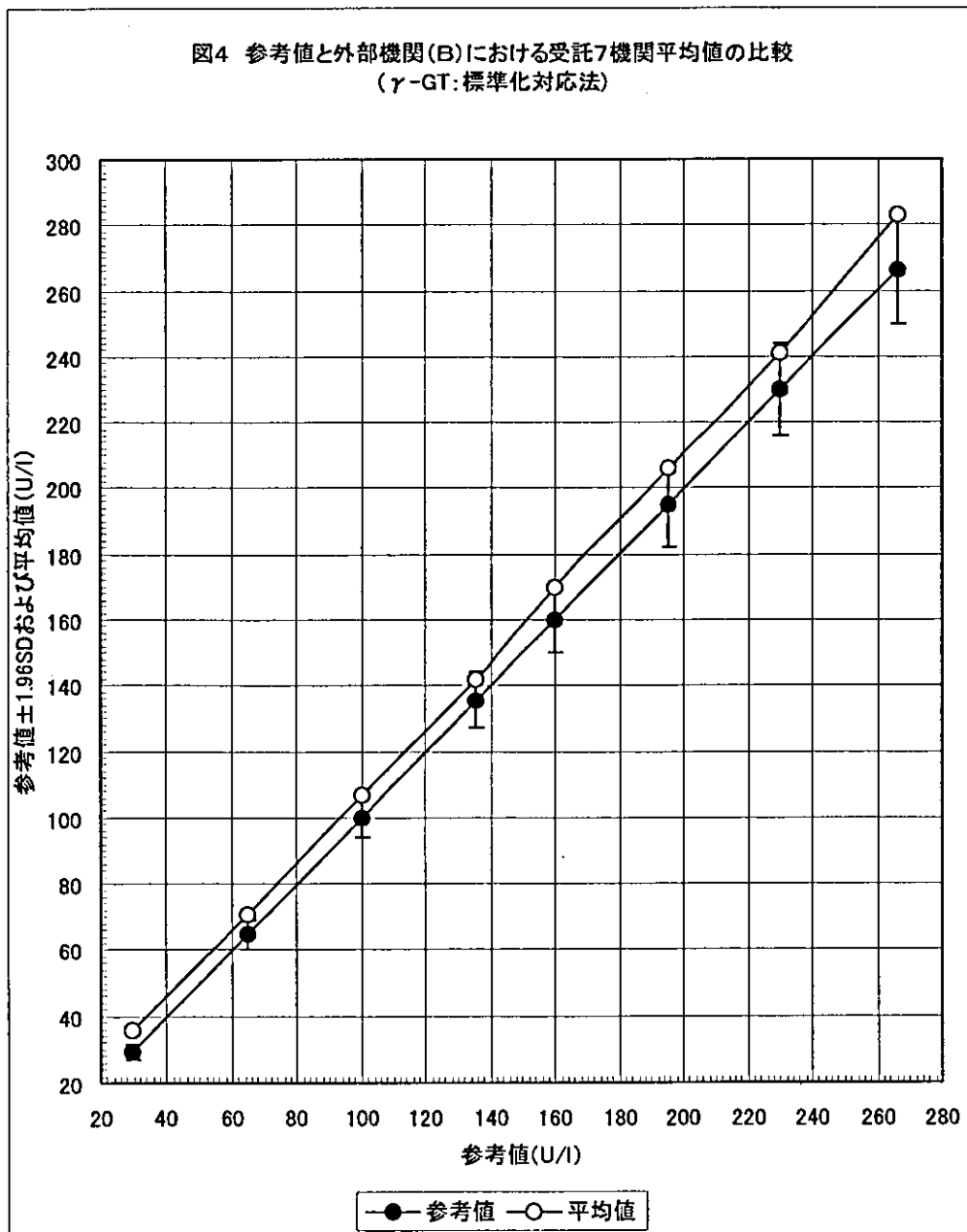
第12回調査		試料1	試料2	試料3	試料4	試料5	試料6	試料7	試料8
外部機関(B)	参考値	10.8	23.7	36.7	49.9	58.4	71.2	83.8	96.7
	標準偏差	0.59	0.87	1.16	1.67	1.93	2.26	2.66	3.07
受託4機関	平均値	15	26	41	54	61	76	88	100
	SDI	7.1	2.6	3.7	2.5	1.3	2.1	1.6	1.1
ALT	比率	138.9	109.7	111.7	108.2	104.5	106.7	105.0	103.4
	差	4.2	2.3	4.3	4.1	2.6	4.8	4.2	3.3

図3 参考値と外部機関(B)における受託4機関平均値の比較
(ALT:標準化対応法)



第12回調査		試料1	試料2	試料3	試料4	試料5	試料6	試料7	試料8
外部機関(B)	参考値	29.4	64.7	100	135.6	160.1	195.1	230	266.1
	標準偏差	1.26	2.25	3.13	4.36	5.02	6.49	7.21	8.34
受託7機関	平均値	36	71	107	142	170	206	241	283
	SDI	5.2	2.8	2.2	1.5	2.0	1.7	1.5	2.0
γ -GT	比率	122.4	109.7	107.0	104.7	106.2	105.6	104.8	106.4
	差	6.6	6.3	7.0	6.4	9.9	10.9	11.0	16.9

図4 参考値と外部機関(B)における受託7機関平均値の比較
(γ -GT:標準化対応法)



第 10 回調査		試料1	試料2	試料3	試料4	試料5	試料6	試料7	試料8
外部機関(I)	参考値	45.3	55.9	67	77.7	85.1	95.7	106.9	118
	標準偏差	1.25	1.54	1.89	2.28	2.41	2.64	2.94	3.25
受託2機関	平均値	41	52	63	73	80	89	101	110
中性脂肪	SDI	-3.4	-2.5	-2.1	-2.1	-2.1	-2.5	-2.0	-2.5
	比率	90.5	93.0	94.0	94.0	94.0	93.0	94.5	93.2
	差	-4.3	-3.9	-4.0	-4.7	-5.1	-6.7	-5.9	-8.0

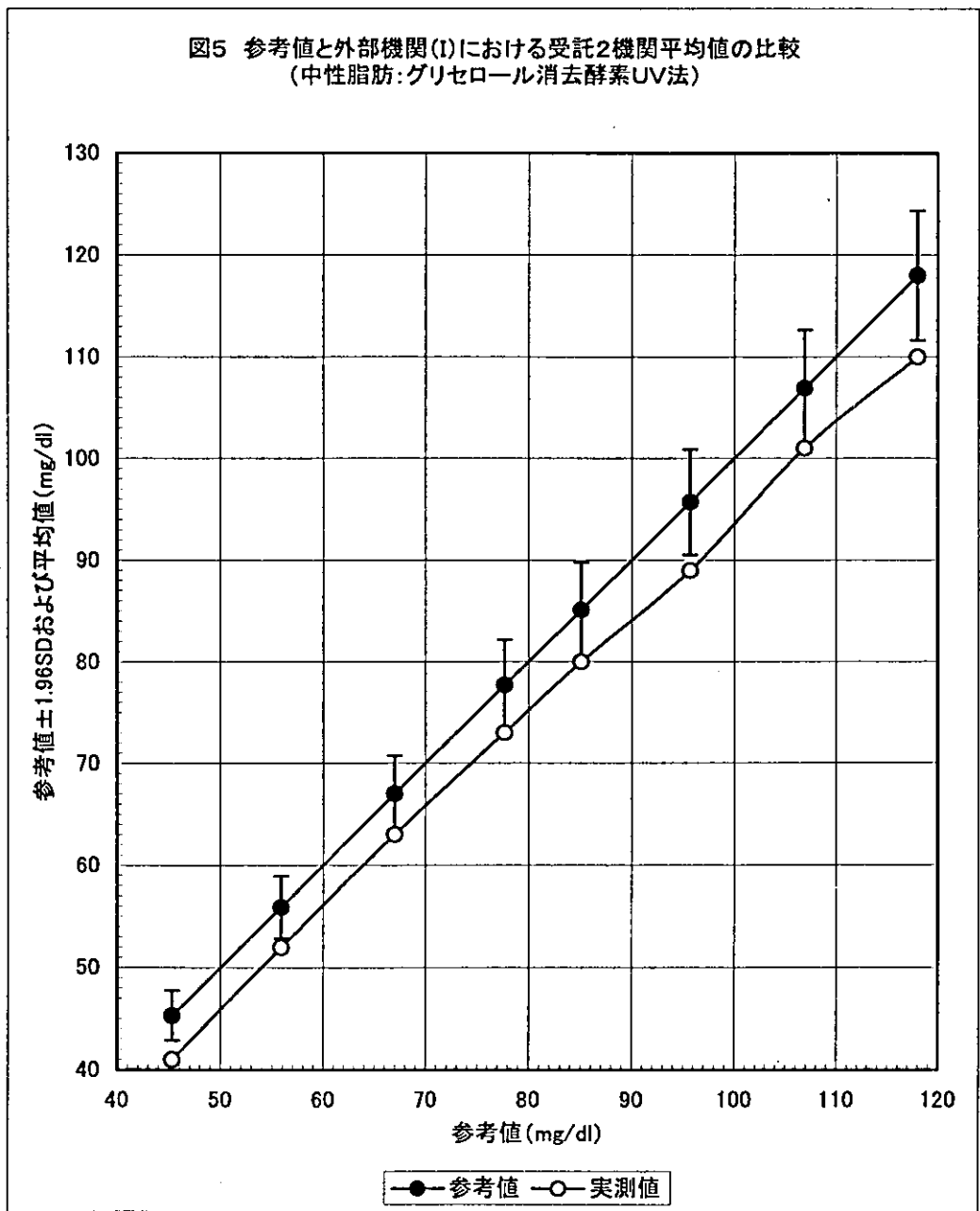
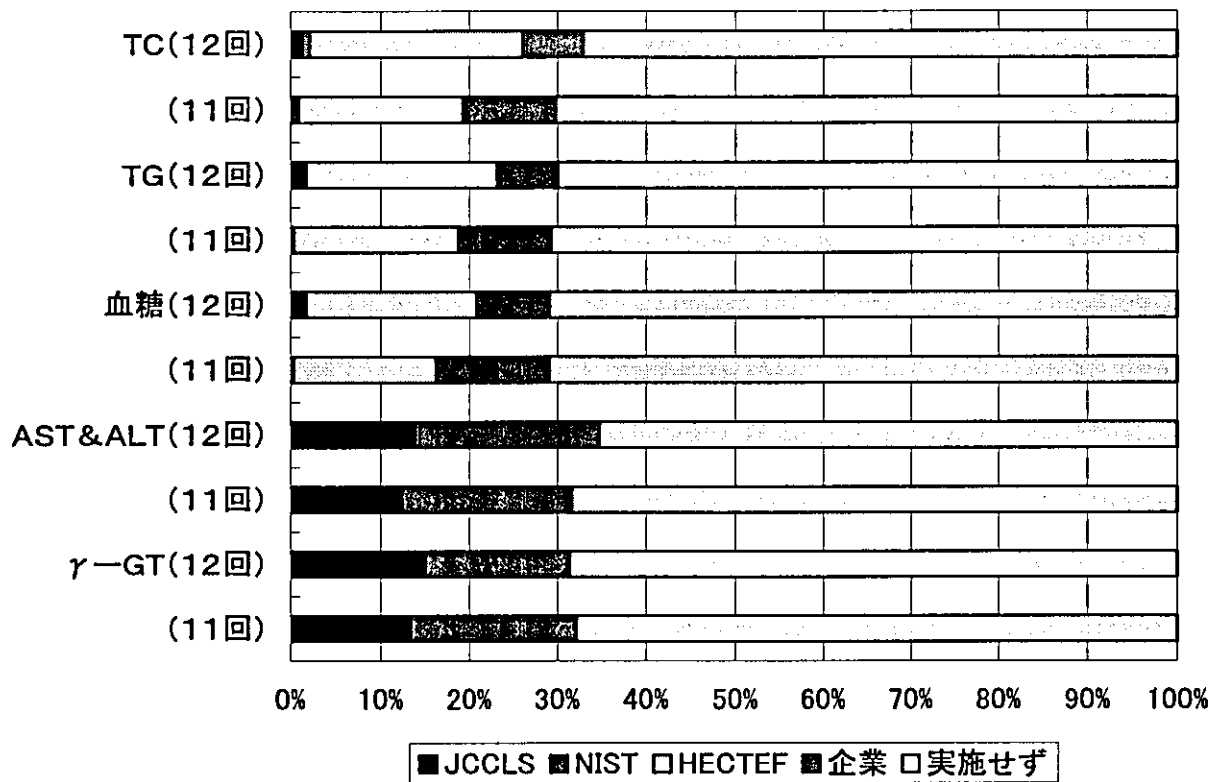


図6 トレーサビリティの確認における高位標準物質の使用比率
(全衛連精度管理調査)



個別研究

胸部 CT 画像読影の能率と精度および学習効果の評価法に関する研究

安曇総合病院院長 曾根脩輔

はじめに

医療技術革新の代表ともいえるべき CT 装置が胸部疾患診断の能率や精度の向上に寄与していることは疑問の余地がないように思われる。しかし、せっかくの技術もこれを活用する医師の読影技術が未熟であれば、患者や医療経営者に対するその恩恵は X 線被曝のリスクに比し不十分なものとなる。最近の研修医制度の導入とも関連して、CT 検査の安全と質を一層高めるため、研修医が医療現場で日常業務をこなしつつ、CT 画像読影技術をいかに習得し、その効果のほどをいかに確認するか、CT 画像診断の能率と精度を的確に評価できる方法論の確立が重要課題となっている。

このような現状に鑑み本報では、安曇総合病院に研修目的で来院した一外国人医師を対象に、CT 画像読影の能率と精度を評価する方法を検討、それを胸部 CT 画像読影の修練を目的にした読影実験に応用し学習効果を評価した。その結果、興味ある知見が得られたので報告する。

1. 材料および方法

胸部 CT 画像読影の学習効果を評価するために考案した方法論とそれを用いて行った読影実験の手順および材料を以下に示す。

1) 読影実験—1

研修目的で中国（上海）から安曇総合病院に来院した医師 1 名が勤務に入る直前に確定診断付き胸部 CT 画像データベースを読影、存在診断を行う読影実験—1 を行った。読影対象となった CT は 1 症例中の 1 スライスを延べ 270 例分集めたものからなる。胸部 CT 画像データベースの異常所見の有無および疾患の内訳を表 1 に示す。異常所見が有る場合は 1 画像中、重要な所見 1 個が定義された。研修医にも読影前、その旨が知らされた。

研修医は上海大学医学部卒、31 歳、女性、呼吸器内科医として 7～8 年の経験を有する。検診目的の CT 画像読影は今回が初めてであった。

読影対象の CT 画像は、実験担当者が操作する PC 画面(12 インチ)上に表示されたものをコネクタ接続された別の CRT 装置画面(19 インチ)上に表示した。

読影医は、その画像を読影、異常所見の有無判定を行い、その結果を、所見有りの場合は気管支分岐部辺の CT 画像シェーマ上に局在位置を示す印をつけ、かつ、その存在確信度をシェーマ右端に記された 0 から 100%表示のメジャー上にチェックした。その際、大別して 0~50%のチェックは所見無し判断、50~100%のチェックは所見有り判断であること、従ってそのどちらにも属さない 50%の確信度はチェックしないこと、所見無しまたは有りの確信の差がわずかな場合でも 50%を境にそのどちらかに決断することを申し合わせた。異常所見が存在しないと判断された場合は、0~50%の確信度のみがチェックされた。

2) 読影実験—2

実験—1 に引き続き、研修医は存在診断を再度行った後、質的診断も行う実験—2 を行った。CT 画像シェーマ上に、所見の局在および存在確信度を実験—1 と同様に記録するとともに、肺がん所見：1、肺がん以外の病変：2、正常：3 の鑑別結果を記録した。症例の疾患分布は知らせなかった。

3) 存在診断の学習—1

実験—2 に引き続き、No.137-270CT 画像（画像データベースの後半）に対する実験—2 の読影結果（存在診断、質的診断の結果）が研修医に開示され、自身の読影結果と照合するという存在診断の学習が行われた。

4) 実験—3

その後、研修医は No.1-136（画像データベースの前半）に対して再度、存在診断と質的診断を行った。

なお、以上の実験—1, 2, 3 は同日中の午後 1:00~4:00 の間に、随時休憩を挟んで実施した。読影は時間制限せず行ったが、1 スライス画像当たりの所要時間は、所見位置、確信度入力、質的診断結果入力を含めて、最大でも 30 秒以内、全体では休憩時間も含めて約 3 時間を要した。

5) CT 画像読影診断全般に関する学習—2

来日直後の実験—1, 2, 3 および学習—1 に引き続き、研修医は研修担当医師（曾根）と共に約 3 週間にわたり画像データベースの各例に対して詳細な症例検討を行った。その一方で、研修医は当院における日常業務の一部に参加、胸部 CT 読影に関する技能の研鑽に励んだ。なお、当該画像データベースに関する学習は滞在期限一週間前に終了した。

6) 実験—4

研修一ヶ月後の最終日、研修医は画像データベースの No.1-136 例（画像データベース前半）に対して実験—2 と同様の方法で存在診断と質的診断を行った。

次に、画像データベース後半；No.137-270 の読影に対して、CRT モニター上で、CT 画像表示、検出所見位置入力 (ROI 表示)、確信度入力、次画面表示、以下繰り返す一連の作業を全て PC 画面上で行うことができるビューアシステムを用いて読影を行った。

実験前、読影済みの CT 画像データベース前半の数症例を対象に、所見検出時の所見位置入力法、確信度入力法を練習してから本番の読影に臨んだ。

270 スライス画像全体の読影所要時間は、実験—1 または 2 と同程度(約 1 時間)であった。ビューアシシステムを用いて読影した画像データベース後半の 134 例では全体の読影時間と各例の読影所要時間がビューアシシステムにより自動記録された。

7) 読影結果の評価法

上記実験において、研修医が決断した存在診断および質的診断の結果を確定診断の結果と比較した。これより読影者が下した存在診断および質的診断の正診、誤診および読影診断の能率を以下の方法により評価した。

① 存在診断の精度

画像データベースの異常所見の有無は、確定診断の結果、1 スライス画像中に異常所見は 1 個またはなしとされている。それぞれに対して研修医が下した判断が確定診断と一致した場合は TP または TN となる。確定診断と不一致の場合は次の 3 種類が有りうる。

確定異常所見見逃し；FN,正常例に異常所見の存在を指摘；FP,確定診断所見を見逃したがそれ以外の別な箇所異常所見を指摘した場合；FN&FP である。

このようなデータを基に通常の ROC 解析を適用するため確定診断；異常所見あり例を異常所見有り領域となし領域からなると考えた。確定診断；正常例は異常所見なし領域である。各領域に対する研修医の判断の正誤 (TP,TN,FN,FP) を、回答された確信度が 50%より上か下で分類する。確定診断；異常領域の個数は異常所見有りの例数と同じ (s) である。確定診断；異常所見なし領域は正常例の領域 n と異常例の異常所見なし領域 s の合計 (n+s) である。

確定診断；正常例 (異常所見なし領域) に対して読影者が回答した確信度 p が 50%以上の時、0 か 1 かの判定結果(0,1 判定)は FP、50%以下の時は TN に分類される。各領域の FP,TN に対する確信度は p である。確定診断；異常例の異常あり領域に当てられた確信度 (p) が 50%以上の時、異常所見あり領域に対する 0,1 判定は TP、確信度は p、異常所見無し領域の判定は TN、確信度は $100\% - p$ (50%以下となる) であるとした。異常所見あり領域に対する確信度 p が 50%

以下の時、読影者の 0,1 判定は FN、異常所見無し領域に対する 0,1 判定は TN、確信度はともに p を当てた。確定診断：異常所見あり例で、異常所見なし領域に確信度 50%以上の異常所見が判定された時は FP と分類され、その時の確信度は p、異常所見あり領域の判定は FN に分類され、確信度は 100% - p を当てた。

以上の異常所見あり領域群 (s 個) に下された確信度 (0~100%) の頻度分布と異常所見なし領域群 (s+n 個) に下された確信度 (0~100%) の頻度分布を基に、 $TPR = \text{異常所見あり領域の TP 数} \div s$ 、 $FPR = \text{異常所見なし領域の FP 数} \div (s+n)$ を ROC 解析の通常の方法に従って算出、ROC 曲線を描き、その曲線下面積 Az を求めた。ROC 曲線の描出および Az の計算は SPSS で行った。

② 質的診断の精度

読影結果と確定診断 (肺がん、肺がん以外の病変、正常) のクロス集計表を求めた。

③ 読影診断の能率

CT 画像読影の能率を評価するため CT 画像ビューアシステムに、実験開始時間、画像表示から所見有りを判断して所見位置を記録する直前までの時間、所見位置を記録し存在確信度を入力するまでの時間および次画像表示までの時間 (所見無し判断の場合は画像表示一次画面表示までの時間)、実験終了時間を区別して記録できる機能を付加した。実験—4 に本ビューアシステムを応用し上記作業時間を測定・集計した。

2. 結果

1) 存在診断

画像データベースの前半 No.1-136 画像に対する実験—1, 2, 3, 4 の結果を図 1 に示す。実験—1 (存在診断のみ実施) に比較して、その後、存在診断のほか質的診断を行った実験—2 の結果は、ROC 曲線がやや下方に位置し、Az もやや低下した。さらにその後、実験—2 の No.137-270 に対して真の答えが公開され、自身の読影結果と照合する学習が行われた後の実験—3 の ROC 曲線と Az は実験—1, 2 に比しやや上昇した。

一ヵ月後 (研修終了間際) 再度読影した No.1-136 画像に対する実験—4 の成績は一ヶ月前の実験—1, 2, 3 に比し、ROC 曲線と Az は明らかに上昇した。

異常所見の存在確信度分布の形状は、実験—1 では 0%~100%まで連続分布したが、実験—2 以降は 0%か 100%のどちらかで回答されたことが明らかであ

る。

2) 質的診断

質的診断を行った実験—2 と 4 の結果と質的確定診断結果を比較したクロス集計を表 2 に示す。実験—2 (研修前) の肺がん、肺がん以外の病変、正常の正答率はそれぞれ、45.8%、24.3%、95.7%、確定診断との一致率 $k=0.284$ であったが、実験—4 (研修後) では、75.6%、60%、91.3%、 $k=0.614$ と上昇した。

3) 読影所要時間の定量

①存在診断

実験—4 で画像データベース後半 137-270 の画像に対してビューアシステムで自動記録した存在診断の正誤と判断所要時間の結果を図 2 に示す。画像が CRT 画面に表示されてから異常所見有り と判断して所見位置を記録する直前までの時間 (TP ; 99 例) は約 5 秒であった。正常と判断され、次画面を表示するまでの判断時間 (TN ; 30 例) は約 12 秒を要した。結果的に FP となった (3 例)、異常所見ありと判断してその局在位置を記録する直前までの時間は約 20 秒であった。異常所見無しと判断し次画面を表示するまで、結果的に FN となった判断 (2 例) に要した時間は約 25 秒であった。

②質的診断

肺がん病変を他病変に、またはその逆に鑑別するのに要した平均時間は 28.5 ± 9.7 秒、肺がんを肺がん、それ以外の病変をそれ以外と鑑別するのに要した平均時間は 22.7 ± 8.9 秒であり、両者の間に有意の差 ($p < .005$) が検出された。

3. 考察

1) 存在診断の評価

実験—2 および 3 により行われた研修開始直後の学習—1 は、読影者が判定したばかりの自分の結果 (画像データベースの後半画像 No.137-270 で) と真の答えを照合するという単純なものであった。その後、真の答えが未公開であった画像データベース前半 No.1-136 の CT 画像をもう一度読影して ROC 曲線 (Az) で学習前後を比較した。結果は図 1(a),(b) に示されるごとく ROC 曲線 (Az) の向上によりその効果が示された。図 1(c)~(f) に示すごとく、初めて CT 画像を読影した時の確信度の回答 (c),(d) は 0~100% の間に連続分布した。一方、2 回目以降の読影実験—2,3,4(e),(f) では、確信度はすべて 0 か 100% のどちらかで回答された。これは学習効果というよりは学習の影響というべきかもしれない。

異常所見の存在を確信度で回答するのを読影者に求めたのは診断精度を ROC

曲線で定量評価するのに必要なデータを取得するためであった。この種の読影実験では通常、実験開始前、読影者にその旨要請するのが常であり、今回も実験—1の直前そのように指示した。しかし、実験—2以降は新たにそのような指示はせず読影者の自然の行動に任せた。

胸部 CT 検診では大量の受診者の CT 画像読影が対象となり、異常所見の有無判定に基づく要精検または精検不要の判定が読影医の主な仕事である。連続確信度で回答する習慣はない。ROC 解析できるようにするため、異常所見の有無を確信度で回答させた今回の実験は、臨床現場と異なる読影の仕方を読影医に強要した可能性がある。

研修医は、日ごろ見慣れない検診目的の条件で撮影された画質の CT 画像を来院早々読影するよう指示された。従って、初回の読影実験—1では、自信の動揺や実験者の指示に忠実になろうと努力した結果が確信度に反映された。しかし、同時に確信度の回答という不慣れな行為が精神的負担を生じ、それが 2 度目以降の実験で 0,1 判断につながったと推測される。さらに根本の原因は、課せられた仕事が「存在診断」だった可能性もある。自分が異常と思って見えているものや異常所見は見えていないと思った判断に、敢えて 100%または 0%以外の確信度を与える必要はないと結論した研修医の態度は、実験—1の経験を通して検診目的の読影行動の基本を自己学習した当然の結果とも考えられる。(注；質的診断では、存在診断の結果、異常所見が歴然と見えている場合であっても癌かそれ以外の病変かの確信が 50%前後になることは有り得る。また、今回の存在診断の実験で禁止した 50%の回答も禁止できない。)

問題は、このような医師の行動を是認した場合、その結果を定量的、客観的に評価する方法はあるかということである。現在、世界に流布する Metz の ROC 解析プログラムは実験—1 図 1(a),(b)のデータは解析できるが、0,1 判断を行った実験—2以降のデータでは結果が出力されない。(台形公式による SPSS の方法は Az を算出できるが、過小評価された結果となる。)かといって ROC 解析プログラムによる定量評価のため、現実と異なる確信度の回答を求めた実験による ROC 解析の結果は現状を反映していないかもしれず、そもそも確信度を読影医が自ら申告することができるのかという意見もある。その解決法を暗示する結果が図 2 (実験—4)に示されている。所見の有無を判断し、結果的に TN,TP であった判断所要時間に比し、結果的に FP,FN であった判断所要時間は明らかに延長した。このような事実は他の画像診断においても多く観察されており、一般性のある現象である。これより、異常所見の有無判断に要する時間が確信

度の強さと比例することを根拠に読影者の無意識的確信度を推定し、ROC 曲線を描く方法が検討されている¹⁾。本法によれば、0,1 判断（胸部 CT 検診による要精検か精検不要かの判断）の結果からでも、読影医の判断所要時間から確信度が推定され、存在診断の精度を定量評価する従来の ROC 解析（Metz のプログラム）が適用できる。但し、その方法はアイデアの提案の段階にある。0,1 判断に対しても ROC 解析が適用可能な方法の開発は今後の重要な課題である。

図 1(a)に示されるごとく、実験—2 以降の ROC 曲線は 0,1 判断によるものであり、FPR、TPR の原点(0,0)と到達点(100,100)との間に動作点は 1 点しか存在しない。従って、台形公式から算出された A_z は真の（類推するしかない ROC 曲線の）値より過小評価されている。それにもかかわらず実験—1 の連続確信度に基づく ROC 曲線に比較して、実験—2 以降の結果は学習後のそれぞれの特徴を反映していると思われる。すなわち、実験—2 は実験—1 の直後、続けて存在診断のほか、質的診断も行うよう指示されたものであり、読影者のモチベーションの低下がそのまま A_z の若干の低下となって現れたと推測される。しかし、存在診断の再現性は表 2 に示されるごとく十分確保されている。

実験—3 は画像データベースの後半(No.137-270)の自身の読影結果を真の答えと比較する学習を行った直後、画像データベース前半(No.1-136)を再読影したものであり、 A_z の若干の上昇と確信度の回答の仕方に大きな変化が生じたことは前述した通りである。このような傾向は、将来、CT 画像読影スクリーナーを務める可能性を秘めた医療技術系大学生を対象に今回と同じ画像データベースを用いて行った別の読影実験でも再現された。図 3 にその結果を示す。CT 画像読影が初体験の学生 42 名に対して今回と同じく、自身の読影結果と真の答えを照合するという学習を行い、再度同症例画像を読影した結果である。研修医と同様の効果が観測されている。

学生および研修医で示されたこれらの事実は、実験—2 の後試みたような単純な学習でも効果が期待できることを示したと考えられる。但し、学習効果には個人差がある。研修医は当院で初めての画質に接したとはいえ、呼吸器内科医として CT 画像読影の現場の経験が 7~8 年ある。そのようなベースの基で、今回行ったような単純な学習後の(実験—3)の A_z の上昇率は CT 画像読影がまったく未経験の学生の平均より相対的に小さかったように思われる。

研修医の存在診断能の絶対値を日本の医師と比較した場合はどうか。研修医とほぼ同様の方法で同画像データベースを CRT 読影した日本の医師 5 名の読影結果を図 4 に示す。このときの A_z は 0.864~0.945 の範囲に分布した。研修医

の学習前の Az (実験—1, 2) は図 1 に示すごとく 0.880, 0.843、学習後 (実験—3) は 0.896 であったから、日本人医師と遜色ない。特に、研修 1 ヶ月後 (実験—4) の Az は 0.972 であり効果は歴然である。但し、実験—1, 2, 3 の後、読影医は同画像データベースを用いて研修担当医と一対一の詳細な症例検討を行った。従って、帰国前 1 週間の冷却期間をおいた後の実験—4 の時点で真の回答が研修医の記憶に残っていた可能性が無視できない。実験—4 の Az は、実験—1, 2, 3 の経験、日本の CT 肺癌検診で使われた画質への順応、当院における CT 画像読影に関する学習および現場での研修、自己の研鑽等が総合的に盛り込まれた結果と思われる。

いずれにしても、今回の一連の実験で得られた研修医の Az 向上の推移は、CT 画像診断の基本である存在診断に対して行われた当院での学習および研修が医師にとって意義あるものであったことを示していると考ええる。

2) 質的診断の評価

最近改定された肺癌取り扱い規約では肺癌疑い (E 判定) に対して確信度に近い概念が導入されている。しかし、今回の読影実験では、研修医は肺癌疑いの確信度は回答せず、所見なしの正常か、異常所見を検出した場合は肺癌か肺癌以外の病変かを分類した。その結果 (表 2) 研修前の実験—2 より研修一ヵ月後の実験—4 の質的診断能は明らかに向上した。しかし、向上後の肺癌 75.6%、肺癌以外の病変 60% の正診率は研修担当医と約 3 週間にわたり詳細に症例検討を行った割には決して高いとはいえない。ちなみに図 4 の存在診断能を示した日本人医師の肺癌正診率も 80% 前後であり、研修医と同等であった。以上のことより CT 画像診断のみから肺疾患を的確に鑑別診断できるようになるには、存在診断の場合より一層の研鑽が必要であることが示唆された。

3) 今後の課題

CT 画像読影の能率を表す指標のひとつに読影所要時間がある。モニター画像診断が次第に主流になりつつある昨今、CRT 読影に伴う医師の行動や CAD 参照効果を評価するための研究上、VDT 作業による医療従事者の健康管理上、HIS/PACS の管理上、CT 画像読影にどれだけ時間を要するか知ることが意味があると考えられる。しかし、いざ定量する段になり採られる通常の方法は、当事者の自己申告か手作業による測定・集計である。前者のデータは主観に基づく回答の集積であり妥当性に難があり、後者は莫大な手間と費用を要するため継続が困難という欠点がある。

今回、CT 画像をモニター診断する時、読影の開始から終了の間、CT 画像読