

For 849 people who lived in the local government, chest screening for lung cancer with mobile CT unit was performed. A participant recommended detailed examination underwent thin-section CT, and the participant recommended further examination underwent diagnostic examination.

Results

Detailed examination was recommended for 100 people, and 83 people underwent thin-section CT. Five lung cancer and one atypical adenomatous hyperplasia (AAH) were diagnosed by CT-guided biopsy, open lung biopsy or video-assisted thoracic surgery. 18 people underwent follow-up for less than 2 years. During follow-up, lesion appeared in other locus and reduced in one case, and lesion enlarged in another case. Between group C (6 lesion of six case in lung cancer and AAH) and group F (22 lesion in 18 follow-up case), there was a significant difference in the ratio of five diagnostic CT findings (ill-defined, irregular, ground-grass opacity, air bronchogram and venous involvement).

Conclusion

CT findings of ill-defined, irregular, ground-grass opacity, air bronchogram and venous involvement suggested malignant lesion.

Key words: Mobile CT unit, Thin-section CT, Lung cancer, Follow-up, Chest screening

表1 F群とC群でのthin-section CT所見の比較

		F群	C群	
症例数		18	6	
陰影の数		22	6	
長径平均	mm	8	13	
長径範囲	mm	5-32	6-22	
辺縁性状	不整	9	6	*
	不鮮明	5	5	*
	分葉	0	1	
	spiculation	2	3	
内部構造	すりガラス濃度	4	4	*
	不均一	9	4	
	空洞	0	0	
	air bronchogram	1	4	**
	石灰化	1	0	
既存構造との関係	血管気管支の集束	5	3	
	静脈関与	3	4	*
	胸膜陥入	2	2	
	胸膜陥凹	0	1	
	胸膜肥厚	1	0	
	satellite lesion	3	0	
形	円形, 楕円形	11	6	
	不整型	6	0	
	多角形	5	0	
胸膜との関係	胸膜に接する	6	0	
	胸膜直下	4	3	

*: $p < 0.05$, **: $p < 0.01$

トピックス

医療技術者によるCT画像読影の潜在性の評価

松本 徹・古川 章・遠藤真広・松本満臣・長尾啓一・柿沼龍太郎
曾根脩輔・藤野雄一・和田真一・山本眞司・村尾晃平

日本放射線技術学会雑誌第61巻第6号所載

別 刷

2005. 6. 20

日本放射線技術学会

Topics

医療技術者によるCT画像読影の潜在性の評価

放射線医学総合研究所 松本 徹・古川 章・遠藤真広
 東京都立保健科学大学 松本満臣
 千葉大学総合安全衛生管理機構 長尾啓一
 国立がんセンター・がん予防検診研究センター 柿沼龍太郎
 安曇総合病院 曾根脩輔
 NTTサービスインテグレーション基盤研究所 藤野雄一
 新潟大学 和田真一
 豊橋技術科学大学 山本眞司
 富士通株式会社バイオIT事業開発本部 村尾晃平

緒言

医療技術革新の代表ともいべきCT装置は胸部疾患診断の能率や精度の向上に多大に寄与している。しかし、これを活用する方法論(撮影法および読影法)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。

現在、日本で使用されているらせんCT装置はシングルスライスCTからマルチスライスCTへと移行しつつ、精密検査はもちろん、肺癌検診¹⁾や外来検査によるスクリーニングにも応用されている。その結果、CT画像情報の量的・質的拡大²⁾が進み、CT画像読影を担当する医師の負担は増大し³⁾、すべての画像を見ることができないことによる所見の見落しの可能性が懸念されている。また、処理可能なCT画像読影の量により検査数が限定される事態も生じている。

われわれは、このような現状を打破しCT技術の進歩の恩恵を多くの人を受けられるようにするため、高精細・大容量の胸部CT画像を能率よく高い精度で読影するのに必要と思われる、以下の三つのCT画像読影支援システムを開発した⁴⁾。1)CRTモニター上にCT画像を動画表示して読影、経年受診者のCT画像を過去画像と比較読影できるシステム、2)高速ネットワークを介して遠隔地間の各CRT読影支援システムにCT画像を送受信し、医師がCT画像読影するシステム、3)コンピュータにより肺癌所見を自動検出するコンピュータ支援画像診断システム(CAD)。また、これらのCRT読影支援システムを用いて所見の見逃しを防ぐ読影法(人と人または人とCADのダブルチェック法)についても検討した⁵⁾。一方、新しい社会システムを構築してCT画像読影の能率と精度の向上を図る試みも進行している。Friedenberg⁶⁾はCT画像読影を医師以外のものが務める可能性を検討した。これを受けて日本で

も、本学会スーパーテクノロジスト委員会⁷⁾、胸部CT検診研究会および厚生労働省土屋班⁸⁾が、診療放射線技師によるCT画像読影スクリーナー(以下、CTスクリーナーと略す)認定制度の確立を検討中である。

本研究の目的は、医師以外の者(医療技術者)が異常所見を効果的に検出することができるかどうかを明らかにし、CTスクリーナーの潜在性を確認することにある。本論文では、CTスクリーナー候補者が医師と同等以上の診断能を持つために必要と思われるCT画像読影の学習およびCAD併用の効果を測定した結果について報告する。

1. 方法および材料

1-1 CTスクリーナー読影演習の概要

胸部CT検診を行って収集・作成された確定診断付き胸部CT画像データベース(1スライスの270症例分⁹⁾)を使用して、将来、CTスクリーナーとなる可能性がある医療技術系学生(4年生43名)(以下、CTスクリーナー候補者:CTSと略す)を対象に、画像診断学の講義の一環としてCT画像読影演習を行った。演習には画像診断学担当教授(医師)1名も参加した。演習内容の計画・実行は1名の医学物理士が担当した。演習の目的は、CTSがCT画像に現れる胸部所見の検出を実地に体験することにより、近い将来CTスクリーナーを務めることの心構えを養うこと、自らの読影データを基にROC曲線を描く方法を学び、現在の自分の読影能力を自覚するとともにCT画像読影法を習得し、CAD活用の意義を理解することにあつた。なお、演習担当者はCTSおよび医師に上記目的ならびに今後の教育・研究に資するため、演習結果を公表することを説明し同意を得ている。

1-2 演習内容および手順

1-2-1 CT画像の表示方法

270スライスのCT画像はPowerPointを使って270枚のスライドに調整され、PCプロジェクターにより教壇前の白色スクリーン上に映写された。最初に、暗黒画面のスライドを表示、次にCT画像表示、暗黒画面表示、CT画像表示をスライドショーで繰り返した。現在読影している画像番号をCT画像の左上側に表示した。CT画像表示時間は一律約20秒とした。室内は窓にカーテンをかけ、画像表示スクリーンより最後方の蛍光灯のみを点灯し、レポート用紙への記録ができる程度の照明とした。

1-2-2 CT画像読影とレポート作成の事前練習

まず、本番演習用の270例とは別シリーズの画像16例を使ってCT画像の読影とレポートデータの記録の仕方を練習した。異常所見があると判断した場合は、重要なものを1個のみ検出すること、レポート用紙にコピーされている気管支分岐部辺りの胸部CTシェーマ上に局在位置とその存在確信度を答えること、確信度の記録は胸部CTシェーマ右側に用意した0~100%の物差し上のどこかをチェックして行うこと、その際、異常所見があると判断したときは51~100%の間の確信度をチェックし、異常なしと判断したときは0~49%の間をチェックすること、50%真上の確信度はチェックしないこと等を指示した。さらに、読影者はストップウォッチを使ってCT画像表示から確信度記入終了後までの時間を計測し、そのデータを当該CTシェーマ上に記載するよう指示された。

CTSは、CT画像読影、読影時間測定、某メーカーCADの、正解と告げられた処理結果の表示、それを自分の読影結果と比較、レポート作成という作業を16回繰り返した。次に演習で使用する予定の270例画像のなかから本番では使用しなかった10例(No.255~264)を対象に上記と同様、CT画像読影、読影時間測定、レポート作成の練習を行った。ただし、このとき、正解は公開されなかった(注:CT画像読影時間は、本報告とは別な目的のため次に示す演習-1に限り計測された。しかし、本論文ではこのデータの分析は行わなかったもので以下言及しない)。

1-2-3 CT画像読影演習-1

上記事前練習に引き続き、CT画像読影演習を実施、CTSは270スライスCT画像のうちの前半No.1~115例を1-2-1、1-2-2に示した方法で読影した。

1-2-4 CT画像読影演習-2

演習-1から1カ月後、CTSはCT画像データベースNo.136~185の50例を用いて以下のような学習を行った。CT画像を表示、読影、その結果(所見の局在位置と存在確信度)をレポート用紙に記入、真の答えを表

示、それを自分の判断と比較する作業を50回繰り返した。その後、No.1~115画像を1-2-1、1-2-2の方法で読影、所見位置の記録、確信度のチェックを行った。なお、CTSは演習-1から演習-2に至る1カ月の間に、CTSとともに演習に参加した画像診断学教授より、肺疾患の画像所見(間質性疾患、肺腫瘍:肺癌(肺門部肺癌、肺末梢性肺癌)、良性腫瘍、肺感染症(肺炎、肺結核)、びまん性肺疾患など)に関する講義を4回に分けて受講した。

1-2-5 CT画像読影演習-3

演習-2から3カ月後、CTSは演習-2で記録した自分のCT画像読影レポートを手渡され、仮想的CAD(1-5参照)の処理結果を参照して1-2-1、1-2-2の方法で各CT画像を読影するよう指示された。その際、演習前に次のような教示がなされた。まず、演習-2のCTS各人の読影成績(specificity, sensitivityの数値データ)と仮想的CADの成績および自分とCADの結果を組み合わせて期待される最良の読影成績(到達可能な最高値)が一覧できる図(後述するFig. 4)が各人に渡され、それらのデータの意味や図の見方が説明された。さらに、これから行われる演習の目的は、以前自分が下した判断とCADの判断を比べて正しいと思われる判断を自分の責任において決定し、診断精度を高めることにあると説明された。ただし、各人の読影成績は、演習-2で回答された確信度が0~49%のときは異常所見なし(0)、確信度が51~100%のときは異常所見あり(1)と判断したと見なして計算したspecificityとsensitivityで示された。その後、CTSは画像データベースNo.1~105例を読影、この各画像を縮小コピーして演習-1、2と同じ配置で並べたレポート用紙上に、異常所見の位置と確信度を1-2-1および1-2-2に示した方法で記録した。

1-3 医師グループ(以下、医師Gと略す)を対象とした読影実験

本演習に参加した1名の医師とは異なる医師5名(医師G)を対象に、CTSと同じCT画像データベースを用いて、CT画像読影時の医師の視線解析を目的とした読影実験を行った¹⁰⁾。視線データ収集用アイカメラはnac社製非接触型アイマークレコーダである。CT画像は19インチNANA O製CRTモニターに表示された。読影実験は、窓にカーテンをかけ、レポート用紙に読影結果を記録することができる程度の照明下で行った。読影対象は、CTSと同じ画像データベースのNo.1~100例である。読影医師は、胸部CT画像の読影経験数年~10年以上を持つ呼吸器内科医4名、放射線科医1名からなる。読影実験は2回に分けて行われた。1回目はNo.1~100例に対してCTSの演習-1と同じく、

異常所見の有無を検出する存在診断, 2 回目は同症例画像を読影して異常所見を検出, それが肺癌所見か, 肺癌以外の病変か, 正常か判断する質的診断であった. 本報告では, 存在診断の結果がCTSの結果と比較された.

1-4 読影対象となったCT画像データベースの症例分布

全体は, 確定診断のついた1スライス画像270例分であり, 確定診断: 有所見例201例(肺癌131例, 肺癌以外70例), 無所見例(正常)69例からなる. 本研究でCTSおよび医師Gにより読影されたCT画像数は, 実験時間の都合により各シリーズで異同がある. 演習-1, 2の読影対象は115例, うち有所見例83例(肺がん58例, 肺癌以外25例), 正常32例, 演習-3の読影対象は105例, うち有所見例75例(肺癌52例, 肺癌以外23例), 正常は30であった. 後述する医師Gと共通する読影対象100例の場合は, 有所見例73例(肺癌51例, 肺癌以外22例), 正常27例であった. 本報告では確定診断: 有所見例は有所見領域と無所見領域の二つからなり, 確定診断: 正常は, 無所見領域と見なされた(理由は1-6参照). したがって, 無所見領域の合計は読影対象総数に, 有所見領域数は有所見例数に等しい.

1-5 仮想的CADの性能

仮想的CADは, 1スライスのCT画像に対して, 異常所見なしまたはあり(1個のみ)の処理結果を出力する特長を持ち, 演習対象の105例全体ではFPR=8.3%, TPR=79.7%の性能(FPR, TPRの定義は2-6参照)を持つように調整された. また, CTSの所見有無判断(0, 1)とCADの処理結果(0, 1)との補完性を表す指標: $\phi^{(0)}$ が結果的に, 確定診断: 異常所見なし領域に対してCTSの平均 $\phi=0.17$, 異常所見あり領域に対してCTSの平均 $\phi=-0.02$ となるように, 各CT画像の異常所見の検出または未検出の結果を調整した. 補完性指標 ϕ は以下のようにして確認された. 演習-2におけるCTS各人の連続確信度の回答を, 0~49%のときは異常所見なし(0), 51~100%のときは異常所見あり(1)と判断されたと見なして0, 1データに変換し, これとCADの結果(0, 1)との ϕ 係数を確定診断: 有所見領域および無所見領域ごと, CTSごとに計算, CTS全体の平均を求めた.

1-6 読影結果の評価法

CTSの演習および医師Gの読影実験における各人の存在診断結果は以下の方法で確定診断の結果と比較・評価された.

各CT画像の異常所見は, 確定診断の結果, 1スライ

ス画像中異常所見は0または1個のどちらかであった. それに対する読影者の回答も異常所見が0または1個のどちらかであった. ただし, 読影者の判断には確信度が付与された. 本論文では, 確定診断: 有所見例は有所見領域と無所見領域からなり, 確定診断: 正常は無所見領域であると考え, それぞれの領域に対する読影者の判断を, 以下のごとくTP, FN, FP, TNに分類し, それらに確信度を割り振った. このようにした理由は, すべての読影者の, 読影結果の分母を一定にするためであった. したがって, 無所見領域総数は, 有所見例の無所見領域数(s)+無所見例数(n)に等しく, 有所見領域数(s)は確定診断: 有所見例数(s)に同じである.

確定診断: 有所見例では, 有所見領域または無所見領域のどちらか一方の判断(所見の有無とそれに対する確信度)が決定されれば, 自動的にもう片方の領域に対する判断も決定される. すなわち, 確定診断: 有所見例の真病変の位置が正しく検出され, その確信度 p が51%以上のとき, 有所見領域の判断はTP, その確信度は p であり, 無所見領域の判断はTN, その確信度は $100-p$ であると見なした. 確定診断: 有所見例の有所見領域の確信度 p が49%以下のときFN, その確信度は p であり無所見領域の判断はTN, 確信度は p と見なした. 真病変とは別の領域から所見を検出し, 確信度 p が51%以上のときFP, その確信度は p , 有所見領域の判断はFN, その確信度は $100-p$ であると見なした. 確定診断: 無所見例から異常所見を検出し, 確信度 p が51%以上のときFP, その確信度は p であり, 異常所見の指摘がなく確信度 p が49%以下のときTN, 確信度は p と見なした.

以上の有所見領域群(s)に下された確信度(0~100%)の頻度分布および無所見領域群(s+n)に下された確信度(0~100%)の頻度分布を基に, 通常ROC解析法に従ってTPR=有所見領域のTP数+s, FPR=無所見領域のFP数+(s+n)を求めてROC曲線を描き, その曲線下面積 A_z を求めた. なお, 統計解析はSPSSで行った.

2. 結果

Fig. 1に演習-1, 2, 3の結果得られたCTSの典型例4名のROC曲線と, 演習-1, 3に同席した医師1名(P1)および同症例を読影した医師G(PG: 5名)のROC曲線を比較した. CTS1にある菱形印は演習-2で参照された仮想的CADの結果(FPR=8.3%, TPR=79.7%)を示す.

初めて大量のCT画像を読影したとき(演習-1)より, 学習すると(演習-2)存在診断能は向上し, CADの結果を参照すると(演習-3)さらに向上した. ただ

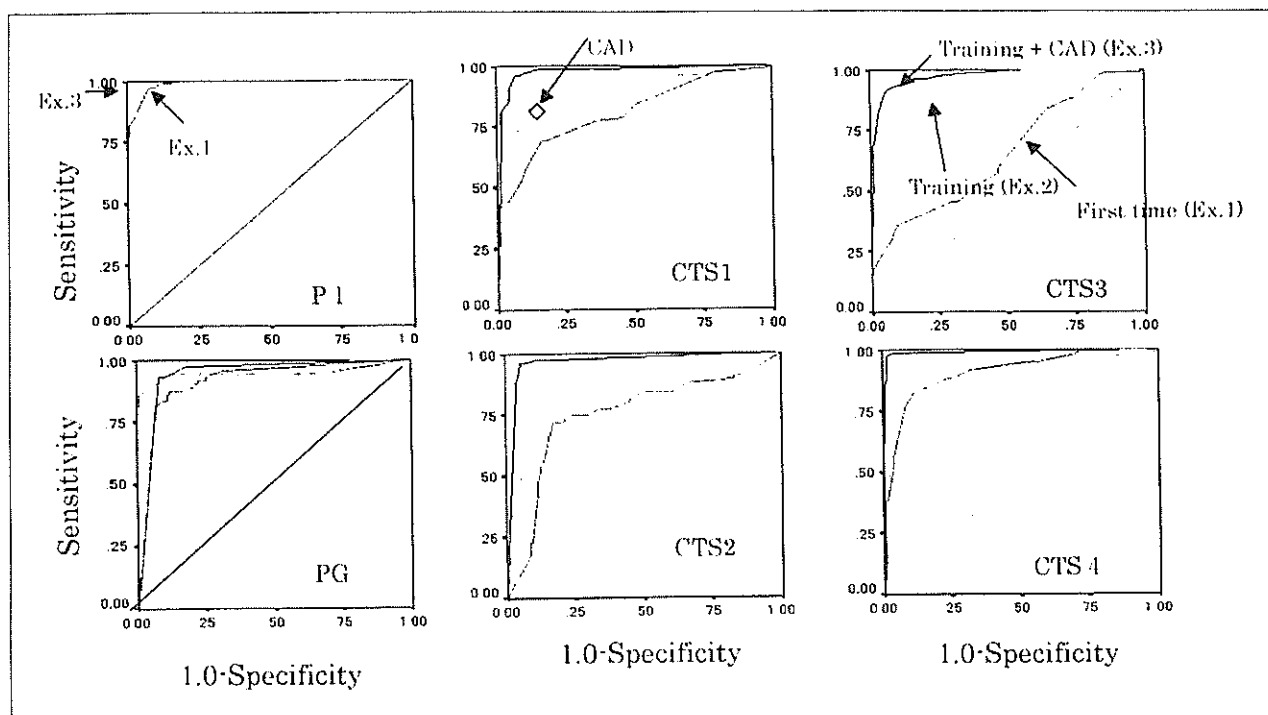


Fig. 1 Comparison of the performance of diagnosing the presence or absence of an abnormality in the case of the physicians (P1 & PG), the students (CTS1-4) for the Exercise 1, 2, 3 & the virtual CAD.

し、その効果の程度は読影者により異なる。例えば、CTS4の学習効果は少なくCTS3では大きかった。CAD参照効果はどのCTSでも高いようである。CTSと同席した医師でもCAD参照効果が認められた。CTS 4名のCAD参照時のROC曲線は医師Gと遜色ないように見える。

演習-1, 2, 3のCTS全員のROC曲線をFig. 2に示す。Fig. 3にCTSおよびCTSと一緒に演習に参加した1名の医師、医師GのAzを比較した。医師GとCTS演習-1(a)との間に有意の差($p < 0.05$)が認められたが、演習-2(b), 3(c)では有意でなくなった。演習-1のCTSの存在診断能は大きく変動し、演習-2, 3では小さくなった。CTS群のAzが医師Gの下限のAzと同等以上の割合は演習-1で40%、演習-2で86%、演習-3では95%を占めた。

3. 考察

Fig. 2, 3において演習-1のCTSのROC曲線・Azの変動が広がったのは、CTSにとってCT画像を大量に読影する仕事は初めてだったことによる。それでもCTSの40%は現役医師G下限Azと同等以上であった。この事実は、少なくとも存在診断に限れば、CTSの一部がCTスクリーナーを務めることの潜在性を示唆したといえる。また、CT画像読影法を学習すれば(演習-2)、その多く(CTSの86%)が医師G下限のAzと同等以

上となり、AzのCTS間変動も小さくなった。この事実はCTスクリーナーの潜在性を一層強化すると思われる。ただし、演習-2までに行われた学習の内容は、演習前に肺疾患の画像所見に関する講義を受講したことと、演習直前に本番とは別の50症例の画像を対象に各画像を読影、真の答えと自分の判断を比較するという作業を繰り返した、というものであった。さらに、CADの結果を参照すれば(演習-3)CTSの大部分(95%)が医師G下限Az以上になった。しかし、AzのCTS間の変動は演習-2(学習時)と同程度に大きかった。その原因は、CADの結果を効果的に使うことができなかったCTSが少数いたためである。

Fig. 1に示された仮想的CAD単独の性能は、処理対象全体に対するspecificity, sensitivityを示したものである。この数値のみから各読影者がCADの結果を参照する価値があるか否か推し量ることはできない。なぜなら、仮に、各症例画像に対するCADの処理結果が読影者のそれと全く同じであれば、その読影者にとってCADの結果は参考にならず、CAD参照に費やす労力は無駄となるからである。しかし、読影者の判断を補完するような処理結果をCADが出力すれば、読影者を支援する可能性が出てくる⁹⁾。すなわち、読影者が検出できない所見(FN)をCADが検出(TP)し、読影者が読みすぎた(FP)所見をCADが検出しない場合(TN)が多いほどCAD参照効果は大きくなる可能性が

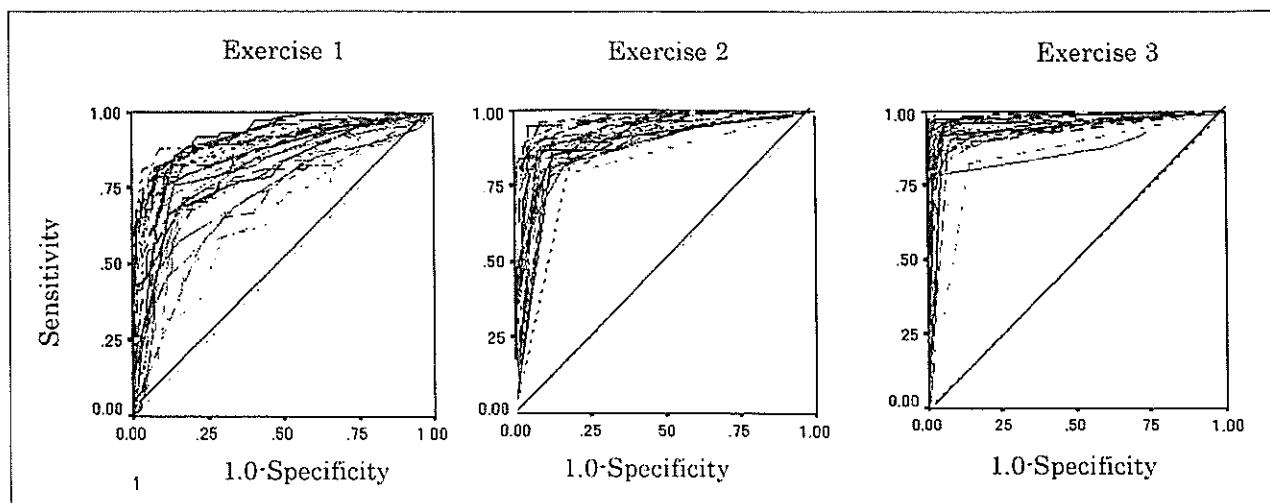


Fig. 2 Comparison of the ROC-curves of the Exercise 1 (first time), 2 (after training), 3 (training + virtual CAD) for all students (CTSs).

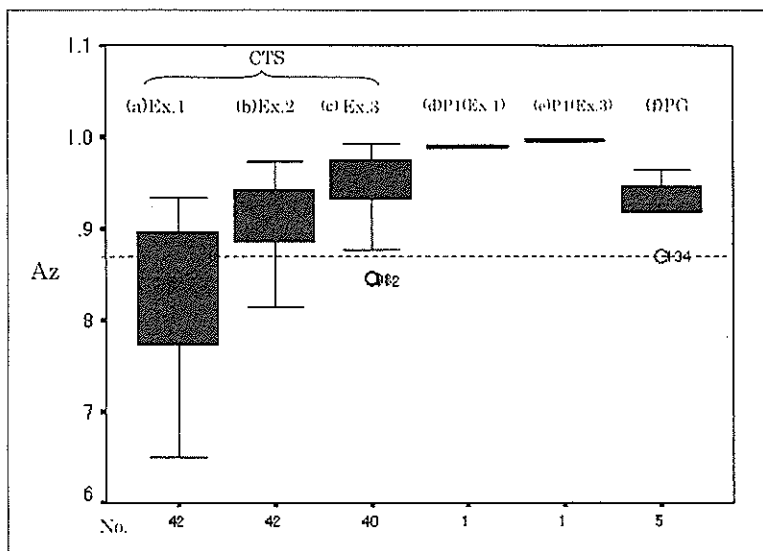


Fig. 3 Comparison of the Az in the case of (a); the students of the Exercise 1 (first time), (b) Exercise 2 (after training), (c) Exercise 3 (training + virtual CAD), (d) P1 corresponding to Ex.1, (e) P1 corresponding to Ex.3, (f) PG compared with the CTSs.

ある。CADが役立つ可能性は、specificity, sensitivityを示すのみでは不十分であることから、人との補完性がどの程度か評価する必要がある。われわれは以前の論文⁵⁾で、人とCADの補完性を表す指標として ϕ 係数を提案した。 ϕ 係数は-1~+1まで変化し、 ϕ が-1に近づくほどCADと読影者の補完性は強くなり、両者の0, 1判断を組み合わせて得られる正診率(specificity, sensitivity)は高くなるのが期待できる。今回は、読影者の0, 1判断と組み合わせて、結果的にFig. 4のような期待値(右上破線の塊)が得られるよう、各CT画像に対するCADの処理結果(0, 1)が演習企画者により意図的に作成された。結果的に、仮想的CAD(直線交点)と演習-2各CTSの読影結果(上記、期待値およびCAD以外)の補完性 ϕ は、無所見領域に対してCTS群全体で

平均 $\phi=0.17$ 、有所見領域に対して平均 $\phi=-0.02$ であるように調整された。ただし、各CTSが仮想CADの結果を参照したときの期待値が現実にも得られるかどうかは保証されていないことに注意すべきである。CAD参照の実を上げるには、どちらが正しいか不明なCADと読影者と同じまたは異なる二つの判断から、読影者は正しい判断を選択しなければならない。Fig. 3に示すごとく、演習-3のAzが演習-2より全体的に上昇したのは多くのCTS(95%)が二つの、同じまたは異なる判断から正しい方の判断を効果的に選択できたことを示唆する。その一方で、演習-1に同席して最高の診断能を示した医師1名に対してもなおCAD参照効果が認められたという事実は、仮想的CADの性能(補完性)が優れたものだったことを示している。さら

に、仮想的CADの検出所見が処理対象となった1CT画像当たり0または1個であり、それに対する読影者の存在診断結果も0または1個であった点が現実と異なることを強調しておかなければならない。演習-3は、CADと読影者の結果が0または1かに単純化された状態の下で、自分の意見とCADの結果を比べて正しい決断をする課題であった。一方、現在、開発中の現実のCADは読影者の判断よりFPを比較的多く出力する傾向が一般的である¹¹⁾。まず、現実のCADが読影者とどのような補完性を有するか確認する必要がある。また、読影者の検出すべき真病変が0かまたは必ずしも1個とは限らない(2個以上の)状況下で、CADがFPを含む異常所見候補を多数出力する場合¹²⁾でも、CTSが自分の判断と照らしてCADの正しい結果を選択し診断精度を向上させることができるかどうか、また、そのような場合のCAD参照法を研究する必要がある。演習-3の仮想的CAD参照実験は、このような性能(補完性)を有するCADであれば、CTSでも現役の医師と同等以上の診断ができるようになる、その目標をCAD開発者へ提示したといえる。

CTSおよび医師が参加したすべてのCT読影は室内を薄暗くしたなかで行われた。CTSと比較された医師Gは、視線データを収集するためアイカメラを装着されてCRTモニター上に表示された画像を読影した。一方、CTSはPCプロジェクターを使ってスクリーンに投影された画像を読影した。CTSの観察条件は、席がスクリーンに近すぎる、または遠すぎる、あるいはスクリーンの斜め方向から観察した、などさまざまであった。読影環境がCTSと医師Gどちらが有利だったかは断定できない。Fig.3で比較されたCTSと医師Gの診断精度は、異なる読影環境で得られたという限界がある。

CTSの演習に同席した1名の医師はCTSと同じ条件で読影に臨み、その結果得られた医師1名の成績Azは、CTSはもちろん医師G5名と比較しても最高となった。これは当該医師本来の読影能にプラスして、本演習でCTSと直接読影能が比較されるという厳しい環境に置かれた効果が反映したと推測される。演習-3の直前、各読影者は自分の性能(specificity, sensitivity)を併記されたFig.4のコピーを手渡され、自分の判断とCADの結果を上手く組み合わせれば図右上(破線の範囲)に示す高い成績が得られることを知らされた。したがって、演習-3の結果には到達目標を与えられた効果が反映されている。また、演習-1の事前の練習、演習-1と2の間に受講したCT画像診断学の授業、演習-2直前の読影法の学習、および同じ症例(No.1~100)を(演習-1と2の間に1カ月、演習-2と3の間に3カ月の期間をあけたとはいえ)3回

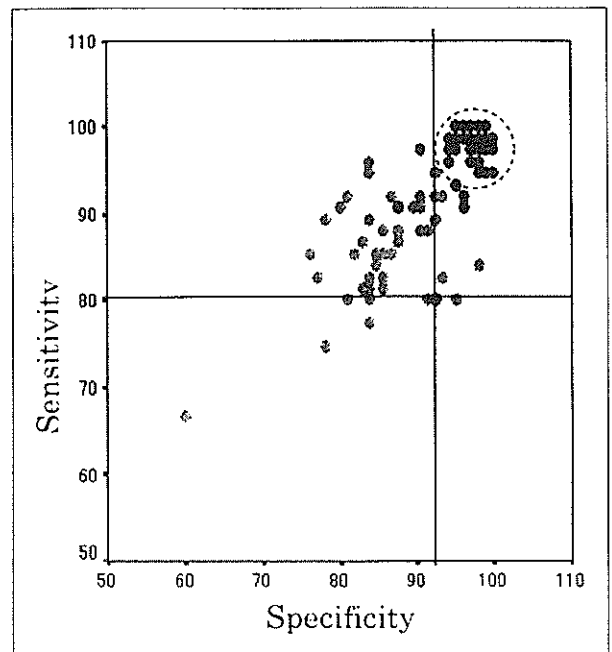


Fig. 4 The data exhibited for the students (CTSs) before the Exercise 2; Cross point of two lines; hypothetical CAD system, The cluster of circles at upper right; expected value from combining an each interpreter (CTS) with virtual CAD system, the others' points; the specificity & sensitivity data for each interpreter (CTSs and the physician; P1).

繰り返し読影したことによる自己学習効果も反映されたと推測される。したがって、これらの特殊な条件の下で得られたCTSの成績を、読影環境が異なった医師Gの成績と比較した結果(Fig.3)から、直ちに、CTSによるCTスクリーナーの潜在性を一般化することはできない。さらに、今回読影対象となったCT画像は、1スライス画像を集合したものであった。本来のCT画像は、シングルスライスCT画像なら1例当たり約30枚程度、マルチスライスCTならそれ以上の数のスライスからなるボリューム画像である。今後は、現実に即したCT画像データベースを対象に、CTスクリーナーの潜在性を検討する必要がある。

Fig.5に演習-1, 2で回答されたCTS5, 6の確信度分布を示す。CTS5の確信度は演習-1で連続分布したが、学習の結果、演習-2ではその分布は0%または100%方向へシフトし、ROC曲線は向上した。CTS6も演習-1で連続分布したが、演習-2の回答のほとんどは0%または100%で回答された。ROC曲線は演習-1, 2で変化しなかった。ここには示さなかったが、演習-3の結果はどちらも0%または100%の確信度で回答された。このような読影態度の変化は、学習の効果というよりは影響と表現すべきであろう。

演習-1で異常所見の存在確信度を0~100%の間で連続的に回答するのを読影者に指示したのは、診断精

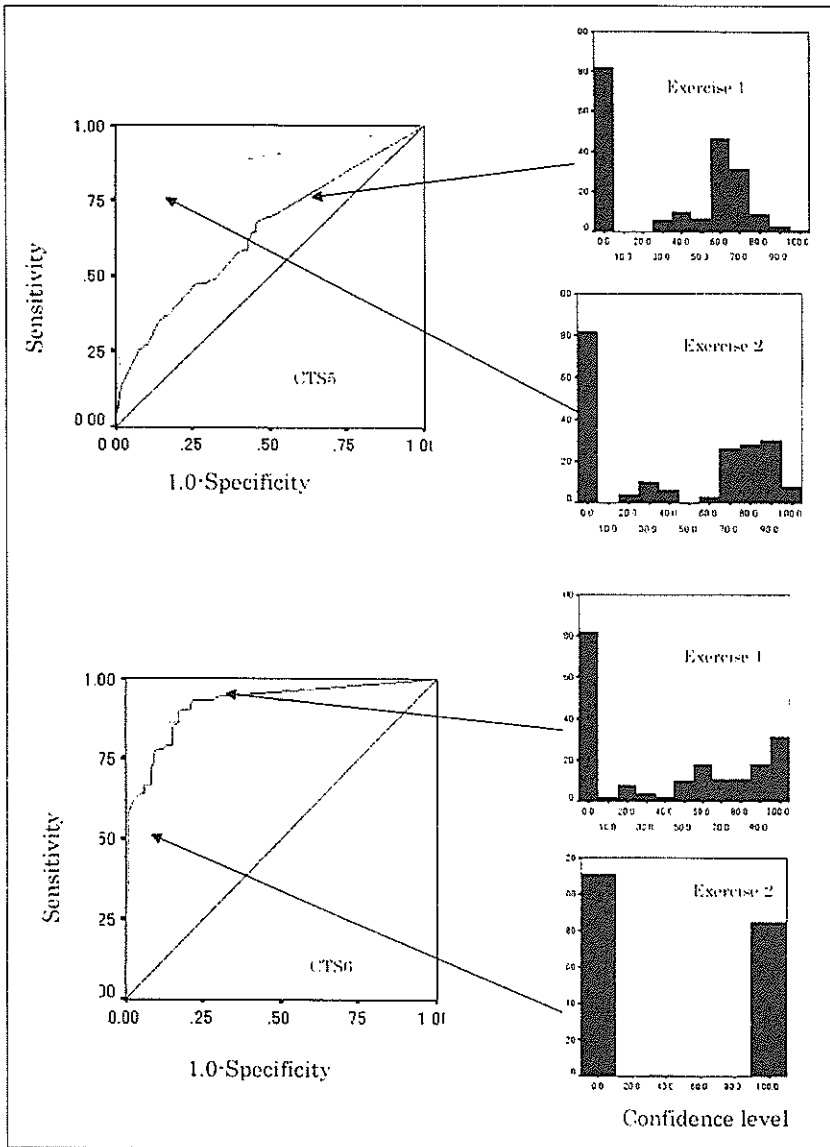


Fig. 5 Influence of training on confidence distribution.

度をROC曲線で定量評価するのに必要なデータを入力するためであった。演習-2以降は読影者の自然の行動に任せた。

胸部CT検診の読影は多数の受診者のCT画像を対象とする。読影作業の主たる内容は、異常所見の有無判定に基づく要精検または精検不要の決定である。連続確信度の回答は通常行われず。したがって、今回の演習で行った連続確信度記載の指示は臨床現場と異なる読影法を読影者に強要した可能性がある。

Fig. 5の演習-1で得られた確信度分布は、初めて経験した演習-1で、CTSが連続確信度回答の指示を忠実に守ったことを示している。しかし、演習-2以降は、不慣れた確信度の回答という行為に精神的負担を生じてそれを解消するためか、あるいは異常所見の有無判定に自信を持ったためか、異常と見える(存在す

る)または見えない(存在しない)ものに対して、100%または0%以外の確信度を与える必要はないとCTSが判断したと推測される。

この現象は、演習-1の経験を通して「存在診断」の基本を学習した当然の結果とも考えられる。ただし、Fig. 5下段のROC曲線に示されるごとく、この読影態度の変化は必ずしも診断精度の向上を伴わないことにも注意すべきである(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(100%)であっても例えば、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、このような読影者の0,1行動を受け入れた場合、その結果からROC曲線を描き、その性能を定量的に評価する方法はあるかということである。現在、

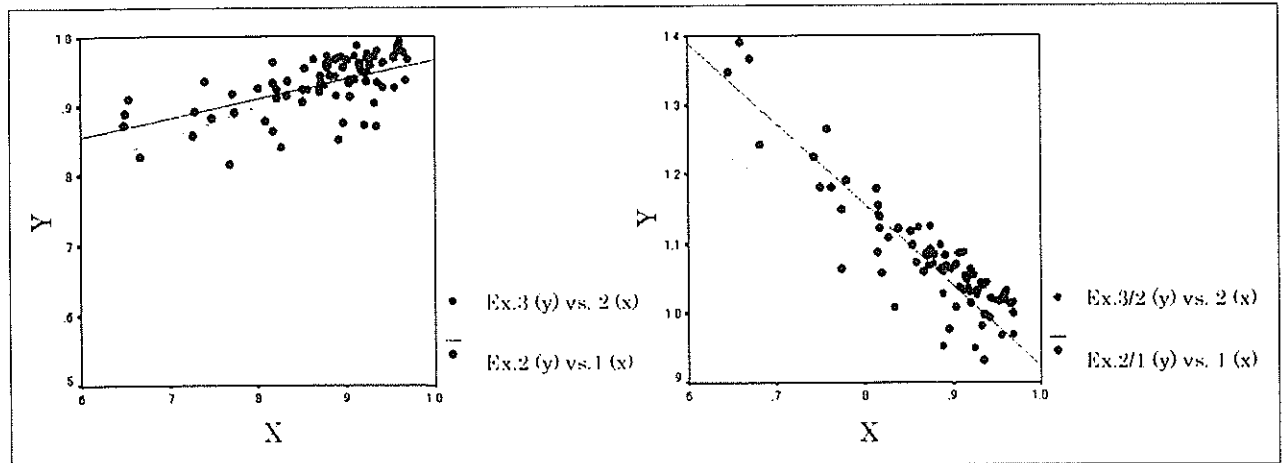


Fig. 6 Effect on CT image interpretation of training and use of CAD. See text about the meaning of figure (a) and (b).

a | b

世界に流布するROC解析プログラム¹³⁾は0, 1判断のデータに対応していない。台形公式に基づくROC曲線は過小評価されたAzを出力する。0, 1判断の結果からROC曲線を描き、Azを精度高く定量できる新しい方法論¹⁴⁾を開発する必要がある。

CTSは演習-1と2の1カ月の間に、肺疾患の画像所見の講義を4回に分けて受講し、演習-2の直前、50症例分のCT画像を対象に、自分の読影結果と真の答えを比較する学習を行った。その結果、Fig. 3に示したとおり、CTS全体では、学習後の診断精度は学習前より有意に向上した。CTS各人のAzの変化は、Fig. 6(a)に示すごとく、演習-1(学習前)の成績が高いCTSほど演習-2(学習後)の成績も高く、演習-2(学習後)の成績が高いほど、演習-3(CAD参照後)の成績も高い傾向があった($p < .00$)。一方、Fig. 6(b)に示すごとく、学習効果(演習-2のAz/演習-1のAz)は演習-1の成績が低いCTSほど高く、CAD参照効果(演習-3のAz/演習-2のAz)は学習後の成績が低いほど高い傾向が認められた($p < .00$)。後者の傾向は、今回と同じCT画像データベースを用いて行った以前の論文⁵⁾の結果を再現した。なお、CT画像読影を初めて経験した演習-1の成績は、以前⁵⁾も今回もほぼ同等であった。しかし、CAD参照後の成績Azの絶対値は、今回の方が以前の報告⁵⁾の結果より有意に高かった。その原因として、以前の報告の演習では仮想的CAD参照演習前までに肺疾患の画像所見に関する講義を受講しなかったことおよび演習-2の直前、今回のような「学習」をしなかったことが考えられる。この事実は、

CADの処理結果を効果的に使用するためには、事前にCT画像読影法を学習して、ある程度の診断能を身に付けておく必要があることを示唆していると考えられる。

4. 結論

医師以外の者がCTスクリーナーを務めることの潜在性を確認するため、医療技術系学生を対象にCT画像読影演習を行った。その結果、存在診断に限るが、適度の画像読影法を学習すれば、医師と同等程度の読影能に到達できる可能性がある。また、読影者とある程度の補完性と性能を持ったCADを活用すれば診断精度はさらに向上し、CTスクリーナーの潜在性は一層強化されることが示唆された。

謝辞

本研究の一部は新潟大学・富士通(株)バイオIT事業開発本部共同研究および厚生労働省がん研究助成金 縄野班(15-25)、科学研究費補助金 小畑班・池田G(15070205)の支援を受けた。読影実験にご協力いただいた医師、読影演習に参加された学生および関係者の皆様に感謝致します。

本稿は2005年2月12~17日に開催されたSPIE International Symposium; Medica ImagingのProceedingsに発表した「An assessment of the potential for interpretation of CT images by radiological technologists」の邦文原稿を一部改変したものからなる。

参考文献

- 1) Matsumoto T, Miyamoto T, Suzuki T, et al.: "Development of mobile CT unit for lung cancer screening", Elsevier Science B.V. *Advances in the Prevention of Occupational Respiratory Diseases*, 485-489, (1998).
- 2) 縄野 繁: 読影フィルムが津波のように押し寄せてくる. *CADM News Letter*, pp.28, 12-13, (2000).
- 3) 中村仁信: 放射線科医の疲弊—画像診断からの一部撤退も止むなし. *日本医放会誌*, 63, 131-132, (2003).
- 4) 松本 徹 編: らせんCT肺がん検診システムの研究開発. *放射線医学総合研究所, 実業広報社*, (2003).
- 5) Matsumoto T, Furukawa A, Machida K, et al.: Methods of evaluating the effectiveness of double-checking in interpreting mass screening images. *Proceedings of SPIE 2004*, 5372, 496-508, (2004).
- 6) Friedenbergrm RM: "The role of the Supertechnologist". *Radiology*, 215(3), 630-633, (2000).
- 7) 松本満臣, 木村千明, 栗井一夫, 他: スーパーテクノロジスト認定制度に関する検討委員会報告書. *日放技学誌*, 61(3), 359-371, (2005).
- 8) 花井耕造: 胸部CTスクリーナー制度の構築, 厚生労働省科学研究費補助金がん予防等健康科学総合研究事業「新しい検診モデルの構築と検診能率の向上に関する研究」土屋班資料, 2004.12, 私信.
- 9) Sone S, Takashima S, Li Z, Yung F, et al.: Mass screening for lung cancer with mobile spiral computed tomography scanner. *Lancet*, 351, 1242-1245, (1998).
- 10) Matsumoto T, Furukawa A, Tsuchikawa M, et al.: Relationship between changes in pupil size over time and diagnostic accuracy, *Proceedings of SPIE 2003*, 5034, 393-402, (2003).
- 11) Yamamoto S, Takizawa H, Jiang H, et al.: "A CAD system for lung cancer screening test by X-ray CT", *The Proceedings of the 15th International Congress and exhibition Computer Assisted Radiology and Surgery, Excerpta Medica International Congress Series 1230, CARS2001 Elsevier Science B.V.*, 605-610, (2001).
- 12) Kallergi M, Carney GM, and Gaviria J: Evaluating the performance of detection algorithms in digital mammography. *Med Phys*, 26(2), 267-275, (1999).
- 13) Metz CE: ROCKIT 0.9B Beta Version, IBM compatible ROCKIT User's guide, Dep. Radiology, Univ. of Chicago, (1999).
- 14) Matsumoto T, Fukuda N, Tsuchikawa M, et al.: Observer performance study for CT-image reading of one slice or multi-slice by the cine display mode of CRT system—An application of the Diagnosis-Dynamic Characteristic (DDC) model, *CARS 2001-Computer Assisted Radiology and Surgery*, 1190, Elsevier Science, Amsterdam, (2001).


An assessment of the potential for interpretation of CT images by radiological technologists

T. Matsumoto, M. Matsumoto, K. Nagao,
R. Kakinuma, S. Sone, A. Furukawa,
Y. Fujino, S. Wada, S. Yamamoto,
K. Murao, M. Endo

Reprinted from

Medical Imaging 2005
**Image Perception, Observer
Performance, and Technology
Assessment**

15–17 February 2005
San Diego, California, USA

 The International Society
for Optical Engineering
Volume 5749

SPIE paper # 5749-65

An assessment of the potential for interpretation of CT images by radiological technologists

Toru Matsumoto^{*a}, Mitsuomi Matsumoto^b, Keiichi Nagao^c, Ryutaro Kakinuma^d,
Shusuke Sone^e, Akira Furukawa^a, Yuichi Fujino^f, Shinichi Wada^g, Shinji Yamamoto^h,
Kohei Muraoⁱ and Masahiro Endo^a

^aNational Institute of Radiological Sciences,9-1,Anagawa-4,Inage-ku, Chiba,263-8555, JAPAN;

^bTokyo Metropolitan University of Health Sciences,
7-2-10, Higashiogu, Arakawa-ku, Tokyo-to, JAPAN 116-8551;

^cChiba Science Center, Chiba University,
1-33, Yayoi-cho, Inage-ku, Chiba-ken, JAPAN 263-8522;

^dResearch Center for Cancer Prevention and Screening, National Cancer Center,
5-1-1,Tukiji, Chuo-ku, Tokyo-to, JAPAN 104-0045;

^eJA Azumi General Hospital,
3207-1 Ikeda-ooaza, Ikeda-cho, Kitaazumi-gun, Nagano-ken, JAPAN 951-85
^fNTT Laboratory, 3-9-11 Midori-cho-3, Musashino-shi, Tokyo,180-8585 JAPAN;

^gNiigata University,746 2-bancho, Asahimachi-dori, Niigata-shi, JAPAN 951-8518

^hToyohashi University of Technology,
1-1,Hibarigaoka, Tenpaku-cho, Toyohashi-shi, JAPAN 441-8122;
ⁱBioIT Business Development Group, Fujitsu Ltd,
17-25, Shinkamata 1-chome, Ota-ku, Tokyo-to, JAPAN 144-8588

ABSTRACT

The increasing number of CT images to be interpreted in mass screening requires radiologists to interpret a huge number of CT images, and the capacity for screening has therefore been limited by the capacity to process images. To remedy this situation we considered paramedical staff, especially radiological technologists, as "potential screeners," and investigated their capacity to detect abnormalities in CT images of lung cancer screening with and without the assistance of a computer-aided diagnosis (CAD) system. We then compared their performances with those of physicians. A set of 100 slices of thoracic CT images from 100 cases (73 abnormal and 27 normal), one slice per case, was interpreted by 43 paramedical college students. A second interpretation by the students was performed after they had been instructed on how to interpret CT images, and a third interpretation was assisted by a virtual CAD system. We calculated the areas under the ROC curve (Az values) for both students and physicians. For the first set of interpretations, the Az values of 40% out of students placed the Az values within the range of Az values of the physicians, which varied from 0.870 to 0.964. For the second set of interpretations after the students had been instructed on CT image interpretation, the students' rate was 86%, and for the third set of virtual CAD-assisted interpretations it was 95%. The performance of paramedical college students in detecting abnormalities from thoracic CT images proved to be sufficient to qualify them as "potential screeners."

Keywords: CT, mass screening, image interpretation, paramedical staff, ROC curve, training, CAD, supertechnologist

1. INTRODUCTION

CT equipment, which may truly be said to represent a revolution in radiological capability, has made a substantial contribution to improving the efficiency and accuracy of diagnosing thoracic disorders. An insufficiently developed methodology (imaging and interpretation methods) for its use, however, would mean that its benefits for patients and medical administrators are inadequate in comparison with the risk posed by X-ray exposure.

*matsu@nirs.go.jp; phone 81 43 206-3240; fax 81 43 206-3246

The CT equipment currently in use in Japan is moving from single-slice CT to multi-slice CT (multidetector-row CT, MDCT), and the use of CT for diagnosing lung cancer¹ as well as in MDCT scans for outpatient screening is becoming widespread. As a result, both the volume and quality of CT images is increasing², the workload of physicians responsible for CT image interpretation is becoming heavier³, and concern is mounting for the possibility that some abnormalities may be missed owing to the impossibility of viewing every image. It also invites a situation in which the number of scans and outpatient screenings that can be carried out is limited by the volume of CT images that can be interpreted.

To remedy this situation and ensure that a large number of people can benefit from advances in CT technology, we have developed the following three CT-interpretation support systems necessary for the efficient and highly accurate interpretation of high-definition, large-volume thoracic CT images. (1) A system that displays CT images for interpretation as moving images on a CRT monitor, enabling interpretation in comparison with previous CT scans of the same patient. (2) A system using high-speed networks to transfer CT images between CT-interpretation support systems in remote locations so that physicians can interpret the CT images. (3) A computer-aided diagnosis (CAD) system that automatically detects signs of lung cancer. We have also investigated a method of interpretation using these CRT interpretation support systems to avoid overlooking abnormalities (the method of double checking by either two people or one person and a CAD system)⁵. Attempts are also in progress to construct new collaborative systems to improve the efficiency and accuracy of CT image interpretation. R. M. Friedenber⁶ has investigated the possibility of employing persons other than physicians to interpret CT images. In light of this, within Japan too the Supertechnologist Committee of the Japanese Society of Radiological Technology⁷, the Society of Thoracic CT Screening, and the Ministry of Health, Labor, and Welfare Tsuchiya Group⁸ are all in the process of investigating the establishment of a system of accreditation for medical radiological technologists as CT screeners.

The purpose of our research was to ascertain whether or not persons other than physicians, medical technologists, could effectively identify abnormal findings, and to ascertain their potential as CT screeners. In this paper, we report that we have determined the effectiveness of training in CT image interpretation and the simultaneous use of a CAD system, as candidate CT screeners have diagnostic abilities equal to or greater than those of physicians.

2. METHODOLOGY

2.1 Outline of CT interpretation exercises for CT screeners

We used a CT image database with final diagnosis attached (one slice each of 270 cases⁹) gathered and put together from thoracic CT screenings to give CT image interpretation exercises to 43 fourth-year students of medical technology who might become CT screeners in future, hereafter referred to as CT screener candidates, or CTSs, as one part of their courses in image diagnostics. A professor of image diagnostics (a physician) also participated in the exercises. A medical physicist was responsible for designing and administering the content of the exercises. The objective of the exercises was to give the CTSs hands-on experience of detecting thoracic findings visible on the CT images, thereby preparing them to work as CT screeners in the near future. They also learned the method of drawing ROC curves based on their own data interpretation, became aware of their own current interpreting abilities, learned how to interpret CT images, and understood the significance of the use of CAD. In addition, the person responsible for the exercises explained that the results of the exercises would be published in order to contribute to the objectives outlined above for CTSs and physicians, as well as to future education and research, and obtained consent for this procedure.

2.2 Content and procedure of exercises

2.2.1 Method of displaying CT images

The 270 CT image slices were prepared as 270 PowerPoint slides, and projected by a PC projector onto a white screen in front of the lecturer's podium. Initially a black screen was displayed, followed by a CT image display, with the alternation of a black screen and CT image display being repeated as a slide show. The number of the image currently under interpretation was displayed in its top left-hand corner. The CT images were displayed for approximately 20 seconds each. Curtains were drawn in the room, and only the fluorescent ceiling lights in the farthest row from the display screen were used to provide sufficient light to fill in report sheets.

2.2.2 Preliminary practice in CT image interpretation and completing reports

Initially, students practiced the interpretation of CT images and how to fill out reports by using another series of 16 images, separate from the 270 cases used in the actual exercises. They were instructed that if they judged that an abnormal finding was present, they should identify only one important finding, and indicate its localized position and the degree of certainty of its presence on the thoracic CT schema of the bronchial bifurcation illustrated on the report sheets.

They were also told to mark the degree of certainty on the 0%–100% scale at the right of the CT schema, checking a figure between 51% and 100% if they judged that an abnormal finding was present and between 0% and 49% if they determined there was no abnormality, and in neither case checking exactly 50%. In addition, interpreters used a stopwatch to measure the time that elapsed between the CT image being displayed and the completion of their filling in the degree of certainty, and were instructed to mark this data on the CT schema being used.

CTSs repeated the process of interpreting a CT image, completing a report, measuring the time for interpretation, displaying the results of the correct diagnosis given by a CAD from an undisclosed maker, and comparing it with the result of their own interpretation, sixteen times. They next practiced the same procedure of CT image interpretation, report completion, and recording of the time taken for interpretation on ten examples taken from the 270 images scheduled to be used for the exercises but not actually employed (nos. 255–264). This time, however, the correct answer was not revealed. Note: the time taken to interpret the CT images was measured for a purpose other than that of the present report in only Exercise 1 described below. As these data are not analyzed in this report, they are not further discussed here.

2.2.3 CT Image Interpreting Exercise 1

Following upon the preliminary practice described above, CT image interpreting exercises were administered. The CTSs interpreted the first half of the 270 CT image slices (nos. 1–115) according to the method described in sections 2.2.1 and 2.2.2.

2.2.4 CT Image Interpreting Exercise 2

One month after Exercise 1, the CTSs were given training using 50 cases from the CT image database (nos. 136–185) as follows. Students repeated 50 times the procedure of CT image display, interpretation, recording the result and indicating the localized position of the observation and the degree of certainty of its presence on a report form, revelation of the correct answer, and comparison with their own determination. They subsequently interpreted images nos. 1–115, recorded the position of the finding, and marked the degree of certainty, according to the method described in sections 2.2.1 and 2.2.2.

2.2.5 CT Image Interpreting Exercise 3

Three months after Exercise 2, CTSs were given their own CT image interpreting reports completed during Exercise 2, and were instructed to interpret each CT image by the method outlined in sections 2.2.1 and 2.2.2, using the result given by a hypothetical CAD system as a reference. At this point, the following instructions were given before the exercise commenced. First, each CTS was provided with his or her interpreting performance in Exercise 2 comprising numerical data on specificity and sensitivity and the performance of the hypothetical CAD, as well as the optimal performance to be expected by combining the results of the individual and the CAD, that is to say, the highest attainable figure, as a diagram readable at a glance (Fig. 4 below), and the meaning of these data and the way to read the diagram were explained. In addition, it was explained that the objective of the following exercise was to take responsibility for deciding on the correct judgment through comparing the judgment made previously by the individual with that of the CAD system, increasing diagnostic accuracy. Note that individual performance was indicated by specificity and sensitivity, calculated by treating an answer in Exercise 2 of a degree of certainty of 0%–49% as a judgment of no abnormal finding (0) and a degree of certainty of 51%–100% as a judgment of an abnormal finding (1). CTSs subsequently interpreted nos. 1–105 from the image database, recording the position of an abnormal finding and the degree of certainty on a report form with a reduced copy of each image arranged in the same position as in Exercises 1 and 2.

2.3 Interpreting experiment by Physicians' Group (PG)

We carried out an interpretation experiment on a group of five physicians (hereafter referred to as the Physicians' Group, PG) using the same CT image database as that used by the CTSs, with the objective of analyzing the line of sight of physicians when interpreting CT image data¹⁰. The physician who participated in the main exercises was not a member of the PG. The eye camera used to gather line-of-sight data was a non-contact EyeMark recorder manufactured by NAC Co. CT images were displayed on a 19-inch CRT monitor. The interpretation experiment was carried out under low ambient lighting sufficient to complete report sheets, with curtains drawn across the windows. The CT images used for interpretation were nos. 1–100 from the same image database as used by the CTSs. The PG that interpreted them consisted of four respiratory physicians and one radiologist, all with experience ranging from several to more than ten years in thoracic CT image interpretation. The interpretation experiment was carried out in two stages. During the first

stage the subjects diagnosed the presence or otherwise of abnormalities in cases nos. 1–100, in the same way as Exercise 1 for the CTSs. In the second stage, they interpreted the same case images to detect abnormalities and undertake a qualitative diagnosis to determine whether these represented findings of lung cancer, pathological changes resulting from a disorder other than lung cancer, or healthy individuals. In this report, we compare the results of the PG's diagnoses of the presence of abnormalities with those of the CTSs.

2.4 Distribution of cases in the CT image database used for interpretation

Out of a total of 270 image slices with a confirmed diagnosis, 201 had positive findings (131 cases of lung cancer, 70 cases other than lung cancer) and 69 were healthy and had negative findings. The number of CT images interpreted in this research by CTSs and the PG differed in each series according to the time taken to carry out the experiments. In Exercises 1 and 2, 115 slices were interpreted, out of which 83 included abnormalities (58 cases of lung cancer, 25 cases other than lung cancer) and 32 were healthy. In Exercise 3, 105 slices were interpreted, of which 75 included abnormalities (52 cases of lung cancer, 23 cases other than lung cancer) and 30 were healthy. In the 100 slices interpreted in common with the PG reported below, 73 included abnormalities (51 cases of lung cancer, 22 cases other than lung cancer) and 27 were healthy. In this report, cases with a confirmed diagnosis of an abnormality included both regions with and without an abnormality and, whereas a confirmed diagnosis of healthy was regarded as a region without abnormality, for the reason for this see section 2.6. Accordingly, the total number of regions without abnormality is identical with the overall number of slices for interpretation, and the number of regions with abnormalities is identical with the number of cases of abnormality.

2.5 Functions of the hypothetical CAD system

The feature of the hypothetical CAD system was the fact that it detected one abnormal finding only from one CT image slice. It was adjusted to have a performance of TPR = 80.0% and FPR = 7.6% for the total 105 images used in the exercise, for definitions of TPR and FPR see section 2.6. The system's detection or non-detection of an abnormality in each CT image was adjusted to result in a correlation (ϕ^{10}) between the CTSs determination of the absence or presence of an abnormality (0,1) and the result given by the CAD system (0,1) of $\phi = 0.17$ on average for CTSs for a confirmed diagnosis of regions without abnormality and $\phi = -0.02$ on average for CTSs for regions with abnormality. The correlation (ϕ) was confirmed as described below. The answers given during Exercise 2 by each CTS as a continuous range of degrees of certainty were transformed into binary (0,1) data by regarding an answer of 0%–49% as no abnormality (0) and one of 51%–100% as a judgment of the presence of an abnormality (1). The ϕ coefficient of complementarity with the (0,1) results of the CAD system was calculated for each confirmed diagnosis of regions with or without abnormality as well as for each CTS, and the overall average for the CTSs was derived.

2.6 Evaluation of results of interpretation

The results of the diagnoses of each CTS and physician during the exercise and interpreting experiment were compared with the confirmed diagnoses and evaluated according to the method below.

The confirmed diagnosis resulted in a finding for each CT image of either the presence or absence of one abnormality in a single image slice. The answer given by the interpreters was also whether or not an abnormality was present. However, each assessment by an interpreter had a confidence level attached. In this paper, a confirmed diagnosis of the presence of an abnormality included regions with and without abnormalities, and a confirmed diagnosis of healthy was regarded as a region without abnormalities. The assessment of the interpreters regarding each of these categories was classified as described below as TP, FN, FP or TN, and a confidence level allocated. The reason for adoption of this procedure was to provide a common denominator for the interpretation results of all the interpreters. Accordingly, the total number of regions without abnormality equals the sum of the number of regions without abnormality for cases with abnormality (s) and the number of cases without abnormality (n), and the number of regions with abnormality (s) is the same as the number of confirmed diagnoses of abnormality (s).

For cases with a confirmed diagnosis of abnormality, when each had been assessed as either a region with or a region without abnormality (the presence or absence of an abnormality and its confidence level), this automatically resulted in a decision regarding the opposite category of region. Specifically, when the location of a true pathological change was detected in a case with a confirmed diagnosis of the presence of an abnormality and a confidence p of 51% or greater, an assessment as a region with abnormality was TP with a confidence of p , whereas an assessment as a region without abnormality was TN with a confidence of $100 - p$. When a case with a confirmed diagnosis of abnormality had a confidence level p of 49% or less for an assessment as a region with abnormality, it was FN with a confidence of p , and an assessment as a region without abnormality was TN with a confidence of p . If an abnormality was detected in a region

differing from the true pathological change this was FP with a confidence of p if the confidence p was 51% or greater, and its assessment as a region with abnormality was FN with a confidence of $100 - p$. When an abnormality was detected in a case with a confirmed diagnosis of healthy with a confidence p of 51% or greater this was FP with a confidence of p , and when no abnormality was indicated with a confidence p of 49% or less this was TN with a confidence of p .

Based on the frequency distributions of the confidence levels (0%–100%) given for the groups of regions with abnormalities (s) and without abnormalities ($s + n$), we calculated the TPR according to the normal ROC method of analysis as the number of TPs for regions with abnormalities divided by s , and the FPR as the number of FPs for regions without abnormalities divided by $(s + n)$. We drew ROC curves and derived the areas under the curve A_z . Statistical analysis was carried out by using SPSS.

3. RESULTS

Figure 1 compares the ROC curves obtained from the results of four typical CTSs in Exercises 1, 2 and 3 with the ROC curves of the physician who sat in on Exercises 1 and 3 and of the five members of the PG who interpreted the same cases. The diamond in CTS1 indicates the results of the hypothetical CAD system used as a reference in Exercise 2 (FPR=7.6%, TPR=80.0%).

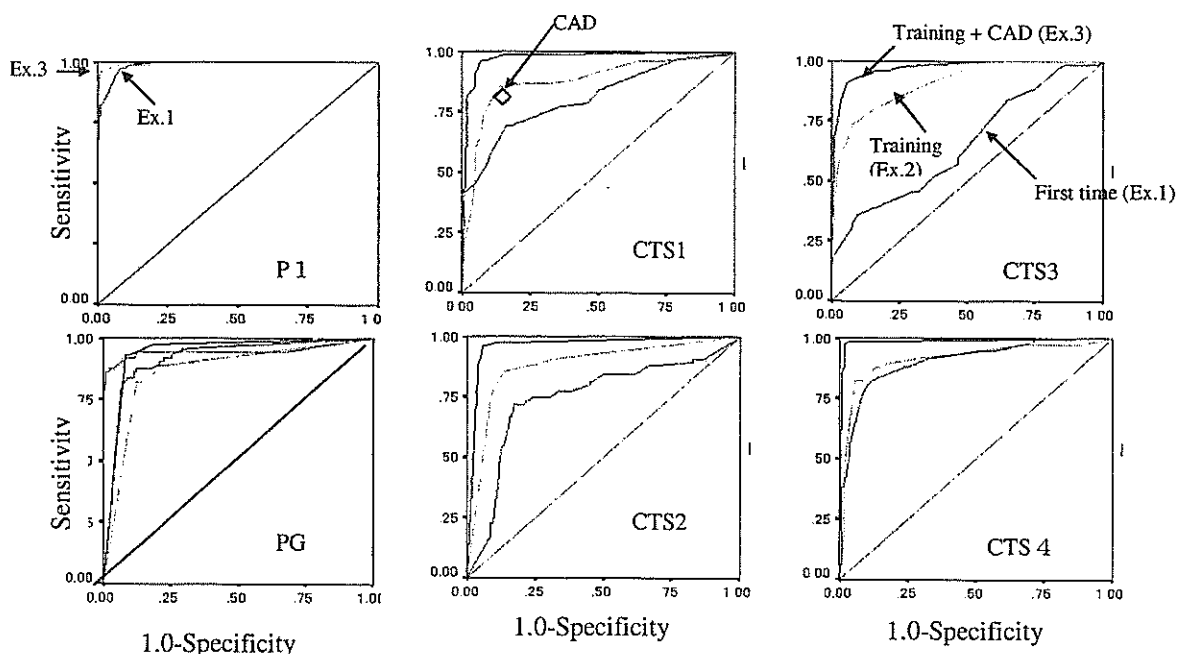


Figure.1 Comparison of the performance of diagnosing the presence or absence of an abnormality in the case of the physicians(P1 & PG), the students(CTS1-4) for the Exercise 1,2,3 & the CAD.

The diagnostic ability of students improved with training (Exercise 2) in comparison with their initial interpretations of a large number of CT images, and improved further when the results of a CAD system were used as references (Exercise 3). The level of effectiveness differed between interpreters, however. For example, the effect of training was slight for CTS4 but high for CTS3. The effectiveness of referring to a CAD system appears to be high for all CTSs. The effect of the CAD system was also observed for the physician (P1) who participated alongside the CTSs. The ROC curves of the four CTSs when using the CAD system for reference are equivalent to those of the Doctors' Group.

Figure 2 shows the ROC curves for all the CTSs in Exercises 1, 2 and 3. Figure 3 compares the A_z values for the CTSs, the one physician who participated in the exercises with the CTSs and the PG. A significant difference ($p < 0.05$) was observed between the PG and the CTSs in Exercise 1 (a), but this was not significant for Exercises 2 (b) and 3 (c). The ability of the CTSs to diagnose the presence of abnormalities varied considerably in Exercise 1, but this was reduced in

exercises 2 and 3. The proportion of the area under the curve Az for the CTS group that was at the same level or higher as the lower limit of the PG was 45% in Exercise 1 and 86% in Exercise 2, and it reached 95% in Exercise 3.

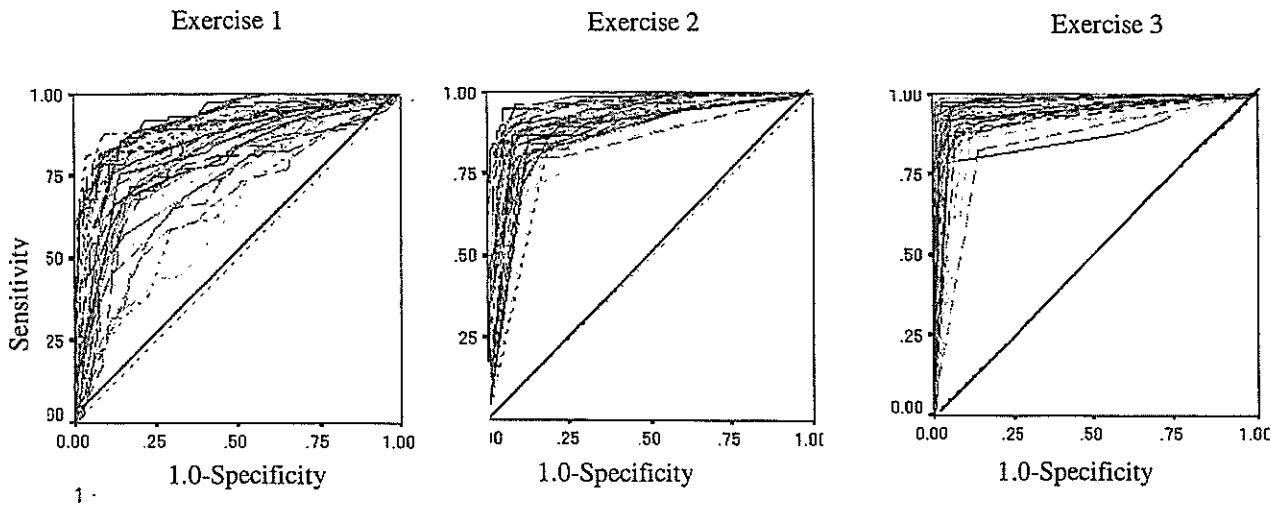


Figure.2 Comparison of the ROC-curves of the Exercise 1(first time), 2 (after training), 3 (training + CAD) for all students (CTSs)

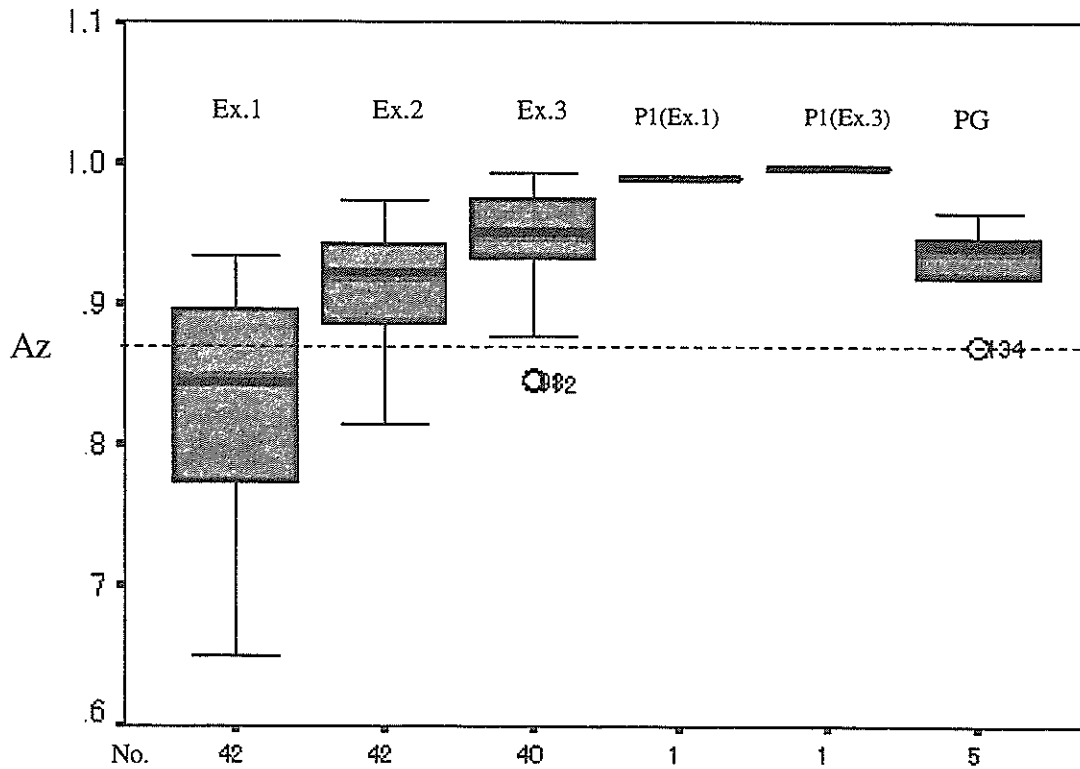


Figure.3 Comparison of the Az in the case of (a); the students of the Exercise 1(first time), (b) Exercise 2(after training), (c) Exercise 3 (training + CAD), (d) P1 corresponding to Ex.1, (e) P1 corresponding to Ex.3, (f) PG compared with the CTSs.

4. DISCUSSION

The wide variation in both ROC curves and Az between CTSs in Exercise 1 shown in Figures 2 and 3 is due to the fact that this was the first time the CTSs had interpreted a large volume of CT images. Even so, 40% of CTSs had an Az at least equal to that of the lower limit of the PG, for whom this is their normal occupation. This fact suggests that a proportion of CTSs possess the potential for employment as CT screeners, at least if this is limited to diagnosing the presence of abnormalities. Having received training in methods of interpreting CT images, the majority (86% in combination with the CTSs in Exercise 1) had an Az at least equal to the lower limit of the PG, and there was less variation in Az. This fact further strengthens their potential as CT screeners. The training carried out here, however, consisted only of interpreting 50 cases, not included in the main exercise, and comparing the true answers with the students' own assessments immediately before beginning the exercise. When the results of the CAD system were used as a further reference, the great majority of CTSs (95%) reached a level above the lower limit of the Az of the PG. There was little difference in the variation of Az, however, from Exercise 2 when training was given. The reason for this is that a few CTSs were unable to make effective use of the results of the CAD system.

Figure 1 shows the performance of the CAD system as its specificity and sensitivity with regard to the entire body of images for interpretation. It is impossible to measure from this figure alone whether or not there is any value to using the results of the CAD system as a reference. Hypothetically, if the result of running the CAD system for each image of a case were completely identical with that of the interpreter, the results of the CAD system would be of no use as a reference for that interpreter, and the labor involved in referring to the system would be a waste of time. If the CAD system has the ability to produce a result that complements the assessment of the interpreter, it may be of assistance⁵. In other words, the effectiveness of using the results of a CAD system for reference increases in proportion to the number of abnormalities the system detects (TP) that are missed by the interpreter (FN) and the number of abnormalities mistakenly identified by the interpreter (FP) that are not detected by the CAD system(TN).

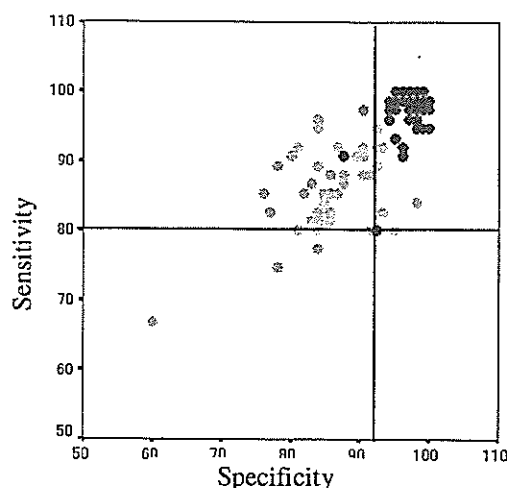


Figure.4 The data exhibited for the students (CTSs) before the Exercise 2 ; Cross point of two lines; hypothetical CAD system, The cluster of circles at upper right ; expected value from combining an each interpreter (CTS) with CAD system, the others' points; the specificity & sensitivity data for each interpreter (CTSs and the physician;P1).

The coefficient ϕ indicates the degree of complementarity between the interpreter and the CAD system varies between -1 and $+1$. The closer ϕ approaches to -1 the stronger the complementarity between the interpreter and the CAD system becomes, and combining the binary (0,1) assessments of both may be expected to raise their accuracy (specificity and sensitivity). For this research, the designers of the exercises intentionally produced the results given by the CAD system (0,1) for each CT image to obtain the expected values shown in Figure 4 (the cluster of circles at the upper right) in CTS in Exercise 2 had a mean value of 0.17 for the CTS group for regions without abnormality, and a mean value of -0.02 for regions with abnormality. This is, however, nothing more than a possibility. For a CAD system of reference to be useful