厚生労働科学研究費補助金

医療安全・医療技術評価総合　研究事業

医師国家試験のコンピューター化に関する研究

平成19年度　総括研究報告書

主任研究者　細田瑳一

平成20(2008)年 4 月

目　次

厚生労働科学研究費補助金(医療技術評価総合研究事業)
総括研究報告書

医師国家試験コンピューター化に関する研究

主任研究者　細田　瑳一　財団法人日本心臓血圧研究振興会附属榊原記念病院　最高顧問

研究要旨

昨年、一昨年と継続して開発してきたコンピューターによる症例シミュレーション試験の問題作成、実行可能性とその効果について調査し、実用化に当たっての問題を検証した。高林が開発したシュミレーション試験では、特に今年度は選択の順序により、より正確な採点ができるかどうかについて検討した。また、吉岡が開発した症例シュミレーション試験(Problem-solving ability test, P－SAT)では、学生の解答パターン分析、共用試験CBT成績との相関などを検討し、P-SATの信頼性・妥当性について検討を行った。それぞれの試験は千葉大、東京女子医大の学生の協力の下実証実験を行い、学生の回答パターンと専門医(熟練医)の判断との乖離で評価をおこなった。この結果から今後国家試験のコンピューター使用の具体的目標の設定、セキュリティの確認、紙ベースの試験との比較検討、また問題作成のプロセスと具体化の費用など残された問題はあるが、コンピューター導入について具体的検討に入ることを提言する。

分担研究者　高林克日己
千葉大学医学部附属病院
企画情報部　教授

吉岡俊正
東京女子医科大学教授
医学教育学　教授

A. 研究目的

高林は昨年までに作ったデータベースを利用して、新たな試験問題を作成するとともに、東京女子医大、千葉大学医学部の4年次の学生各24名および総合診療部の医師9名の順に同一施設では同時に試験問題を与え、試験の配点は、問診・所見、検査、診断、治療それぞれの正解と禁忌肢に対して計算し全体を100点満点とした。試験の採点は自動的に行なわれ集積された。試験結果の中で、中途で終了したものなどを排除し、得点、選択項目数、選択項目正解率、問診初期選択5項目の正解率を算出した。これら得点とともに選択した項目について比較検討することで専門医師と学生との差分を見出すこととした。

また、吉岡のP-SATは、臨床推論・判断に特化したコンピューター試験で、患者への医師の対応のシミュレーションのなかで臨床的能力を評価する試験システムである。今年は第4学年を対象に実証実験を行い、評価システムの妥当性、新しい出題フォーマットとして多選択肢問題、ショートエッセイ問題(短文記入型問題)、語句抽出型問題の3つの型による学生の解答パターン分析、共用試験CBT成績との相関などを検討する。

B. 研究方法

1)新システムによる問題の追加作成

試験問題は昨年までの5題のほかインフルエンザウ

イルスの問題を新たに作成した。これはcommon diseaseの中でも特に代表的な疾患であり、また診療方法も検査ではなく本来問診と所見だけで結論に導かれる点で、今までの他の問題とも大きく異なったものとなっている。この作成には高林が1日で作成することができた。問診選択項目498(うち今回追加分17項目)、検査選択項目540(うち今回追加分1項目)、診断選択項目(うち今回追加分1項目)、診断病名選択項目903(うち今回追加分2項目)、治療選択項目427(うち今回追加分5項目)だった。

2)模擬試験実施と評価

《出題形式》

・臨床事例について、医師が考え、判断する順序に従ってその分析・判断を回答する臨床シミュレーション型の試験を構築した。
・コンピューター化することにより、新たな情報が加わりながら判断・推論が深化する臨床の過程に沿って連続した設問を設定し、一旦回答すると元へ戻れない形式にした。
・学内試験・国家試験等で実施可能なウェッブブラウザを利用したコンピューター試験として構築した。
・3つの問題形式を設定した。

多選択肢問題:6から50個の選択肢から解答を選ぶ形式で、従来の五肢選択問題と比して類推、あるいは除外によって回答することが難しくなる。

語句抽出型問題:事例の中から問題発見・解決に必要な語句を選ぶ形式の問題。

短文記入型問題(ショートエッセイ問題):必要な情報、判断の根拠などを短い文(語句)で記入する問題。

《評価方法》

・臨床推論・判断では適切な推論・判断が一つとは限らないが、熟練した医師は、根拠の基づいた妥当性の高い判断をすることから、クリニカル・エビデンスなどに基づく専門医の判断(解答)との一致で解答を評価した。
・評価段階として以下の4段階を設定した。

A:専門医の判断(選択)と完全に一致する
B:専門医の判断と一部一致し、かつ不適切あるいは行ってはならない判断(選択)をしない
C:専門医の判断とは一致しないが、間違いではない判断を行い、かつ不適切・行ってはならない判断をしていない。
D:間違ったあるいは行ってはならない(禁忌)の判断をしている。

《評価システム》

・評価についても電子評価システムを構築した。

ただし、語句抽出、短文記入においては、回答者が必ずしも予想された解答を行うとは限らない。その場合は、出題者が妥当性を再評価し正誤を新たに決定する。これもウェッブ上で行えるので従来の紙による記述式問題よりも評価が行いやすく、受験者全員について回答が適応されるので評価の正確性・公平性が高まる。

《実証実験》

平成 20 年 2 月 6 日に医学部学生(第 4 学年)95 名について P-SAT を行なった。

《事後評価》

- 学生の解答パターンの評価
- 試験問題数の適正
- 識別指数(A評価を多く取っている回答者がAB評価で、D評価を多く取っている回答者がD評価となる割合
- 共用試験 CBT 成績との比較

(倫理面への配慮)

学生、および医師の受験者の個人情報の取り扱いに留意する。

C. 研究結果

1)新しく作成した問題

実得点と問診、検査の選択項目数を比較したところ、学生と医師では得点は医師の方が高く、また問診、検査選択項目は医師の方が少なく、また得点と選択項目数には負の相関が認められた。これは特に得点と検査選択項目において著明であった。しかし個々の学生間における得点と問診、検査項目の相関をみると、得点と選択項目数には必ずしも負の相関は認められなかった。また選択数の中でどれだけ選択すべき項目を選んだ率が高いか(正解選択率)では、いずれの問題でも医師の方が選択数の中の正解率が高かった一方で、問診の選択での正解数を比べるとむしろ学生の方が高かった。

2)問題解決能力の評価

①診療問題解決能力評価を実施するために、問題を作成する専用サーバーを構築した。本サーバーには問題作成者が離れた場所にいても問題作成ができるよう Web による問題作成機能を構築し、さらにセキュリティーを強化するため https プロトコールによる問題作成にも対応した。本機能により、問題作成者は自分のデスクから問題を作成でき、さらにそれらのデータを暗号化する事により、安全にデータを送信する事ができた。

②診療問題解決能力評価システムではテストを実施後、受験者の解答データを解析するため、解答データを別サーバーに蓄積した。データ解析プログラムは、HTML と親和性の高い PHP 言語を用い、ホームページ上から試験実施者が設定、解析できるよう作成した。また、問題作成サーバーと同様にセキュリティーを強化する目的で、SSL による暗号化にも

対応させた。本システムでは、試験管理、設問管理、試験データ取り込み、評価判定、集計、検索の 6 機能を装備した。

③ P-SAT 実証実験

・平成20年2月6日に95名の学生が受験した。
・2時間で58問を出題したが、全問題を解けない学生が約45%いた。
・前年度に実施し、設問・解答パターンが明らかになっている29問についてAからD評価を行った。
・95名中、A評価16-20が11名、11-15が59名であった。
・D評価については、前年度のトライアルであDを4つ以上とったものが7%であったが、本年度はゼロであった。3つとったものが2名あった。
・共用試験CBTとA評価の数、D評価の数の相関：同時に行われた共用試験CBTの成績とA評価の数、およびD評価の数の相関を検討した。どちらについても有意の相関は認めなかった。
・識別指数はD評価をとった受験生が少なかったため有意の所見を得なかった。

D. 考察

高林の今回の検討から、実際の解答結果を詳細に検討すると、それぞれの学生・医師がどのような意図で診断を進めているのかがよく理解でき、MCQの同一問題と比較して、その設問量や個々の設問自身が簡明すぎて試験に適さないことを考えると、common diseaseの管理能力の評価はシミュレーション試験にしかできないといえる。

しかしこうした解法の評価を実際にコンピュータ上で行なおうとなると、解法は一つではありえず、さまざまな正解順序が存在することから、なかなか困難であることが推察される。今回の結果で、正解選択率は医師の方が高かったのにもかかわらず、はじめの5つの問診選択の正解数はむしろ学生の方がよい結果であったことは、診断の浅いルーチンの検索では学生と医師の間に実力差が出ないものの、評価のためにはそれぞれの深いレベルの診断論理に入っていくことを示しており、そうしたことを評価することが全体の得点と比較して別の能力の検定として計算できるほど精緻なものとするためには、相当量の仕事が必要となることが考えられるためである。今後はベクトルの概念でこうしたグルーピング問題が解決できるかどうかを試みる方向もあるが、むしろ得点が低く不合格となる学生が具体的にどのような思考回路であったのかを検証するときに、生データの設問の具体的な選択順序をみることで正確に評価ができるという利用法もあるように思われる。

Webを用いた今回の試験において、トラブルはサーバ側には起こらず、かなりの人数を対象としてもサーバのスペックさえ十分であれば、トラフィック的にもネットワーク環境においても問題なく運営できることが確認できた。学生のコンピュータリテラシーの向上もあり、現在最大の課題としては試験場を通常のスペースより広く、かつ相互の端末画面を見ることができないような試験場の構造を考えなければならないことであろう。

昨年までの研究で国家試験基準に合わせた基本的な選択項目とその標準解のデータセットがほぼ出来上がっているため、新しい問題の作成は容易になり、大量の問題作成も可能となった。問題はそれを客観的、一般的にみても不整合のないものであるのかの細かい確認にあり、このためには多くの専門家や一般医の参加が不可欠である。現在必要とされている国家試験実施後の問題の開示についても、以降の問題のヒントとならないような方法を考える必要があろう。

吉岡の検討は、P-SATの臨床能力評価の新たなコンピューター試験としての実用化の可能性を示唆する。問題作成サーバーの構築により、問題作成者は自分のいる部署からオンラインで問題作成を行う事ができるとともに、2つの問題点、セキュリティーと通信経路上の盗聴も作成ソフトとサーバーの設置場所を工夫することによって解決出来、従来の問題作成システムに比べて利便性・堅牢性が高いものとなった。

また、時間と労力がかかるとされた診療問題解決能力の評価についても、自動的に採点し評価するシステムの構築によって評価時間の短縮、採点ミスの減少が実現した。

ただし、P-SATは臨床推論・臨床判断能力評価として開発しコンピューター試験システムであるため、医師国家試験のプラットフォームとして考えた場合は、その評価目標・特性が従来の医師国家試験と全く異なることを考慮しなくてはならない。P-SATとCBTでは別の能力を測定している可能性を検討しなくてはならず、また、P-SATの実証実験結果では、能力評価が一定の分布を示すことが明らかになったが、臨床医に必要とされる推論能力・判断能力は従来の評価では測定が難しく、今後P-SAT評価結果が、臨床実習あるいは卒後研修などでの臨床推論・判断能力と相関するかの検討を行わなくては最終的信頼性判定ができない。

E. 結論

・コンピューター試験は、経済性・効率性の利点だけでなく、従来の国家試験では評価できなかった医師としての能力特性を評価できる可能性がある。
・コンピュータによる症例シミュレーション問題はcommon diseaseに対する管理能力を検定するうえで非常に有用であると考えられた。
・しかし試験として導入するとなると、その採点が客観的で普遍的なものであるためには非常に複雑になることから、むしろ不合格者の再評価として用いることから進めるべきではないかと考えられた。

F. 健康危険情報

なし

G. 研究発表

1. 論文発表

Ishihara S, Matsui K, Sato Y, Tang AC, Suganuma T, Fukui Y, Yamaguchi N, Kawakami Y, Yoshioka T. Self-efficacy achieved through problem-based learning tutorial. 医学教育 2007 ; 38 : 391-397

2. 学会発表

選択項目による症例シミュレーション試験の評価

高林克日己 吉岡俊正 細田瑳一

第40回第日本医学教育学会 2008年

H. 知的財産権の出願・登録状況

1. 特許状況

なし

2. 実用新案登録

なし

3. その他

なし

平成１９年度　厚生労働科学研究費補助金（医療安全・医療技術評価総合研究事業）
医師国家試験のコンピューター化に関する研究（H17－医療－一般－016）
分担研究報告書

分担研究者　　千葉大学　高林克日己　千葉大学医学部附属病院　企画情報部教授

研究要旨

昨年の問題を改変し、また新作問題を追加して、東京女子医大と千葉大４年生の有志に試験を行ない、さらに千葉大学総合診療部の医師による結果と比較して、正解選択項目による得点に加え試験の解法に関する採点の可能性についての検討を行なった。その結果得点と選択項目数の間には概ね負の相関があり、項目数だけで評価をする意味はなく、また選択順序を評価するのは実際に複雑すぎることから、不合格者の評価確認のために生データとして用いるほうが有意義であると考えられた。

A．研究目的

コンピュータによる症例シミュレーション試験の問題作成、実行性、その効果について検証する。とくに今年度は選択の順序により、より正確な採点ができるかどうかについて検討する。

B．研究方法

昨年までに作ったデータベースを利用して、新たな試験問題を作成するとともに東京女子医大、千葉大学医学部の4年次の学生各24名および総合診療部の医師9名の順に同一施設では同時に試験問題を与え、試験の配点は、問診・所見、検査、診断、治療それぞれの正解と禁忌肢に対して計算し全体を１００点満点とした。試験の採点は自動的に行なわれ集積された。試験結果の中で、中途で終了したものなどを排除し、得点、選択項目数、選択項目正解率、問診初期選択５項目の正解率を算出した。これら得点とともに選択した項目について比較検討することで専門医師と学生との差分を見出すこととした。

（倫理面への配慮）

学生、および医師の受験者の個人情報の取り扱いに留意する。

C．研究結果

１）試験問題の作成

試験問題は昨年までの５題のほかインフルエンザウイルスの問題を新たに作成した（この作成の方法については検査、治療、症候、診断とに分けて考えた。サンプルとして、診断の一部を添付する）。これはcommon diseaseの中でも特に代表的な疾患であり、また診療方法も検査ではなく本来問診と所見だけで結論に導かれる点で、今までの他の問題とも大きく異なったものとなっている。この作成には高林があたり、１日で作成された。問診選択項目４９８（うち今回追加分１７項目）、検査選択項目540（うち今回追加分1項目）、診断選択項目（うち今回追加分１項目）、診断病名選択項目903（うち今回追加分2項目）、治療選択項目427（うち今回追加分5項目）だった。

２）試験結果

実得点と問診、検査の選択項目数を比較して表1にまとめた。ここで示すように学生と医師では得点は医師の方が高く、また問診、検査選択項目は医師の方が少なく、また得点と選択項目数には負の相関が認められた。これは特に得点と検査選択項目において著明（図１）であった。しかし図２に示すように個々の学生間における得点と問診、検査項目の相関をみると、得点と選択項目数には必ずしも負の相関は認められなかった。また選択数の中でどれだけ選択すべき項目を選んだ率が高いか（正解選択率）を表2に示したが、いずれの問題でも医師の方が選択数の中の正解率が高かった。一方問診の選択での正解数を比べると、むしろ学生の方が高かった。

D．考　察

１）問題の解法過程の調査による新たな採点法の検討

実際の解答結果を詳細に検討すると、それぞれの学生・医師がどのような意図で診断を進めているのかがよく理解できる。MCQ試験の同じ一問題とは比

較にならないほど、実際にここで測定できる評価項目数は極めて大きなものとなる。もしこれと同じことをMCQ方式の試験で行なうことにすると大量の設問になることや、common diseaseについてのMCQでは個々の設問自身が簡明すぎて試験に適さないことから、common diseaseの管理能力の評価はシミュレーション試験にしかできないといえる。しかしこうした解法の評価を実際にコンピュータ上で行なおうとなるとそれはなかなか困難である。なぜならばそれは解法が一つではありえず、さまざまな正解順序が存在するからである。一般に問診・診察や検査において選択する数は少ないほど効率的に探していることになる。しかし鑑別診断の否定用件として必要な項目も含めると、ただ少ないだけではなくてある一定範囲に存在すべきと考えられる。今回の結果で学生と医師では得点は医師の方が高く、また問診、検査選択項目は医師の方が少なく、また得点と選択項目数には負の相関が認められた。しかし図２に示すように個々の学生における得点と問診、検査項目間の相関をみると、得点と選択項目数には必ずしも負の相関は認められなかった。これは特に得点の低い群で選択数も少ない学生がいるためで、これらを除くと概ね、得点数と選択数には負の相関があるように思われる。また正解選択率はいずれの問題でも医師の方が高かったのに対し、はじめの５つでの問診の選択の正解数を比べると、むしろ学生の方がよい結果であった。このことはまず診断の浅いルーチンの検索では学生と医師の間に実力差が出ず、評価のためにはそれぞれの深いレベルの診断論理に入っていくところをチェックしなければならないことを示している。そうであるとまず想起すべき疾患とそのための鑑別診断としての質問群を整理して、それぞれの重み付けをしていけば評価はできるかもしれないが、そこまでしたときの評価が全体の得点と比較して別の能力の検定として計算できるほど精緻なものとするためには相当量の仕事が必要となる。例えばひとつには同等に並列に質問すべき内容があるときに、それぞれを選ぶことによりさらに次の深い設問になり、配点を比較することが難しくなることである。今後はベクトルの概念でこうしたグルーピング問題が解決できるか試みるとともに、少なくても選択数からの評価を総合評価に加える方式を継続して検討する価値はあるかもしれない。しかしこうしたベクトル化するなどして複雑な計算をすることではむしろ結果がブラックボックスになってしまう可能性もある。得点が低く不合格となる学生が具体的にどのような思考回路であったのかを検証するときに、生データの設問の具体的な選択順序をみることでより正確に評価ができるという点だけでも十分に利用法はあるように思われる。

またこうしたコンピュータ症例シミュレーション試験の結果は一概に正規分布にはなりえない。これは設問数、あるいは評価項目数にもよるが、プラス点のほかマイナス点も存在するからと考えられる。このことは国家試験の評価上問題になるかもしれない。しかしこの方式がその知識体系が不十分で医師として相応しくない受験者を篩いにかけるための、資格試験なのであれば、ここで得点が絶対的に低い受験者をスクリーニングし、上述のようにその内容を詳細に検討することだけで評価は十分可能であると思われる。

２）Webによる試験のシステムとしての可能性と限界

Webを用いた今回の試験において、トラブルはサーバ側には起こらなかった。同時に大量の受験者に対して試験を行なうときには、一定数の端末側の故障は不可避でありこの準備対策は常に必要であるが、8000名以上を対象としてもサーバのスペックさえ十分であれば、トラフィック的にも問題なくネットワーク環境においては問題なく運営できることが確認できた。以前より円滑にできるようになったもう一つの要因は学生のコンピュータリテラシーの向上であり、こうしたコンピュータ試験に対する拒絶反応はＣＢＴの影響もあるのか、既にほとんどなくなったといってよい。最大の課題としては試験場を通常のスペースより広く、かつ相互の端末画面を見ることができないような試験場の構造を考えなければならない。これを100名単位の規模で行なうにはまだ課題が残っている。

３）試験問題作成に関する課題

昨年までの研究で国家試験基準に合わせた基本的な選択項目とその標準解のデータセットがほぼ出来上がっている。したがって新しい問題の作成に当たってはこれにその問題に特異的な項目と鑑別のための設問を加えるだけで作成できるので、問題の素案の作成においてこの基本のデータベース上で1時間もあれば一題の問題を作ることは不可能ではなく、大量に問題を作成できるところまで来ている。問題はそれを客観的、一般的にみても不整合のないものであるのかの細かい確認にある。このためには一問題の作成には少なくても数名の専門家が担当する必要があるだろう。またこの問題が国家試験においてその対象をcommon diseaseを解くためのものであるとすると、subspecialtyの専門家だけでなく、一般医の参加が不可欠である。いわゆるcommon diseaseは30種類くらいに限られるが、より広い意味での一般の疾患の典型例まで加えるとその問題は100問以上作成可能であろう。またそうでないと、疾患が限定されてしまい、鑑別すべき疾患が少なくて試験が簡単になってしまう。現在必要とされている国家試験実施後の問題の開示についても、この内容をすべて露呈するわけにはいかないであろう。オープニン

グシーンと出題された疾患名と治療くらいでないと、以降の問題のヒントを与えてしまうことになるし、それだけでも翌年以降に残されたcommon diseaseの数を減らすことになってしまうからである。

E．結　論

コンピュータによる症例シミュレーション問題はcommon diseaseに対する管理能力を検定するうえで非常に有用であると考えられた。医師として求められるのは稀な疾患の診断能力以前に、日常遭遇するcommon diseaseに正しく対応できるかであるが、以前からこの領域をMCQで検定することは困難であった。それゆえコンピュータによる臨床シミュレーション試験は医師国家試験受験者のcommon diseaseの基本的な診療手順を見る上で、今後必修化すべきものであると考えられる。しかし試験として導入するとなると、その採点が客観的で普遍的なものでなければならない。単に適切な項目を選択するだけでなく、解法の順序に着目してみたが、結果としては得点と選択数にはある程度負の相関が認められ、また具体的に行なうには非常に複雑になることから、むしろ不合格者の再評価として用いることから進めるべきではないかと考えられた。

F．健康被害情報

とくになし

G．研究発表

1　学会発表
　選択項目による症例シミュレーション試験の評価
　高林克日己　吉岡俊正　細田瑳一
　第４０回第日本医学教育学会　２００８年

資料

表１　得点と選択項目の比較（千葉大学学生と総合診療部医師）

図１　学生と医師の得点、選択項目の平均値における相関図

図２　インフルエンザ、ＳＬＥ．腎盂腎炎の問題における個々の学生の得点と選択項目（問診と検査項目）の相関図

表２　得点と正解項目率の関係　　学生医師間の正解項目率とはじめの５選択の正解率の比較

問題作成のサンプル（インフルエンザ診断）

表1　　得点と選択項目の比較

	数		得点		問診数		検査数	
	学生	医師	学生	医師	学生	医師	学生	医師
インフルエンザ	24	9	39.7±31.9	66.7±14.2	22.4±12.8	12.4±4.7	30.7±17.7	4.8±6.8
SLE	22	3	16.2±16.1	52.0±4.4	20.0±13.0	18.0±8.1	30.3±17	27.0±26.0
腎盂腎炎	21	4	30.2±18.1	44.3±2.8	14.2±9.0	7.75±3.3	28.2±15.5	17.8±3.9
アニサキス	13	7	96.2±13.9	78.6±26.7	5.2±2.1	8.7±4.5	10.2±10.1	10.0±12.0
癒着性イレウス	8	4	51.3±27.5	55±26.5	16.1±7.3	13±4.7	17.8±11.0	3.8±4.3

図1　学生と医師の得点、選択項目の平均値における相関

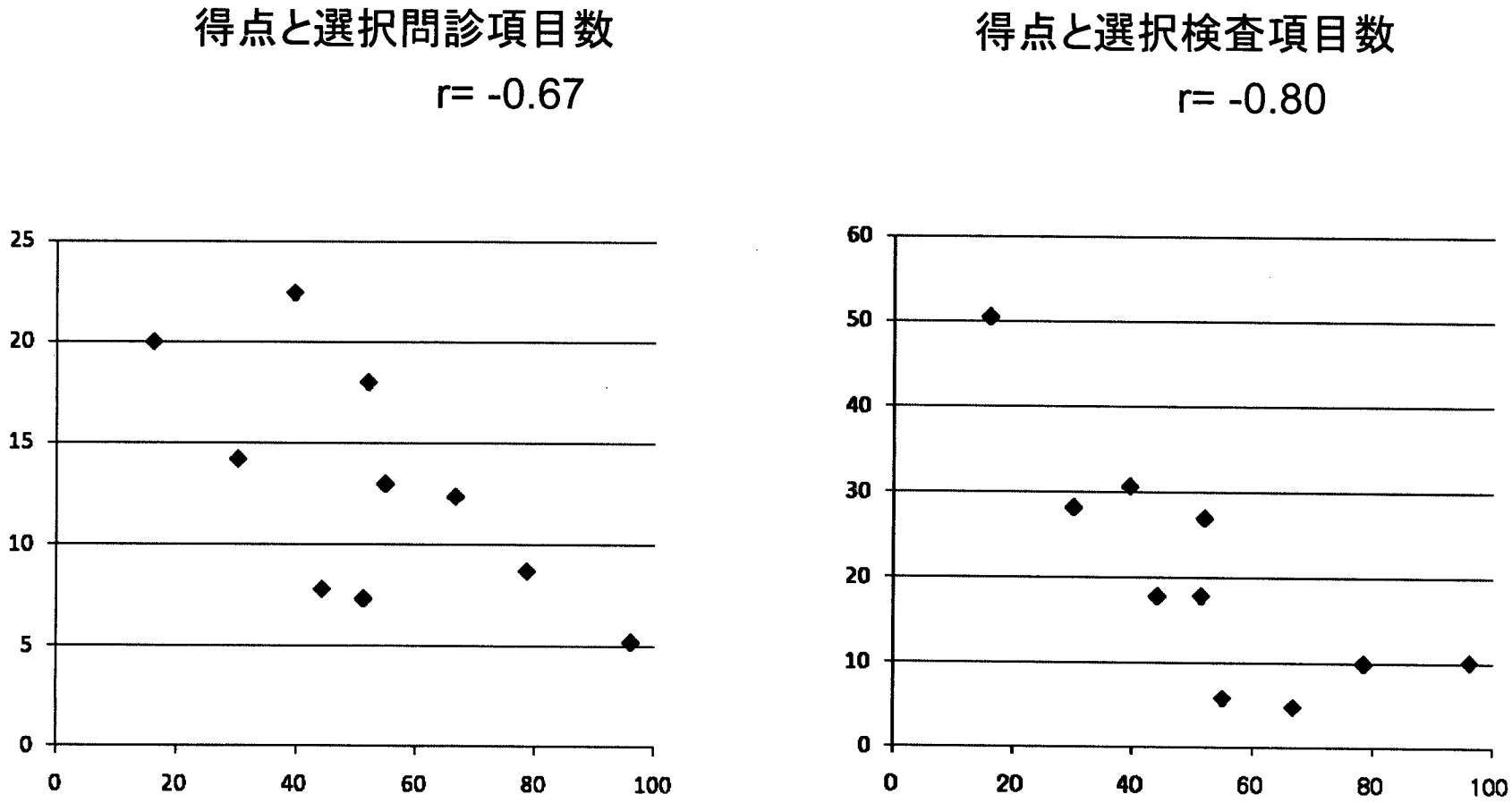
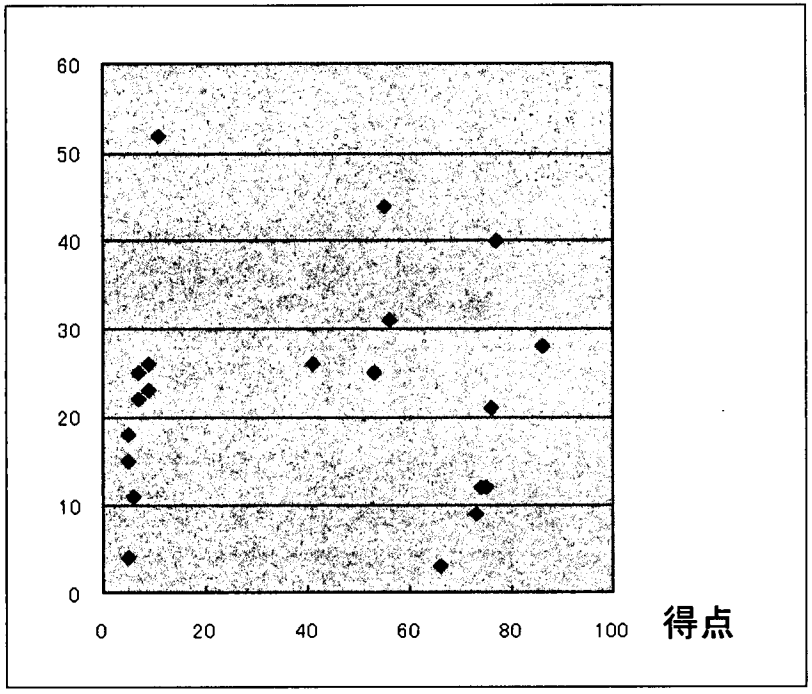


図2　得点と選択項目の相関　　インフルエンザ

得点と選択問診数



得点と選択検査数

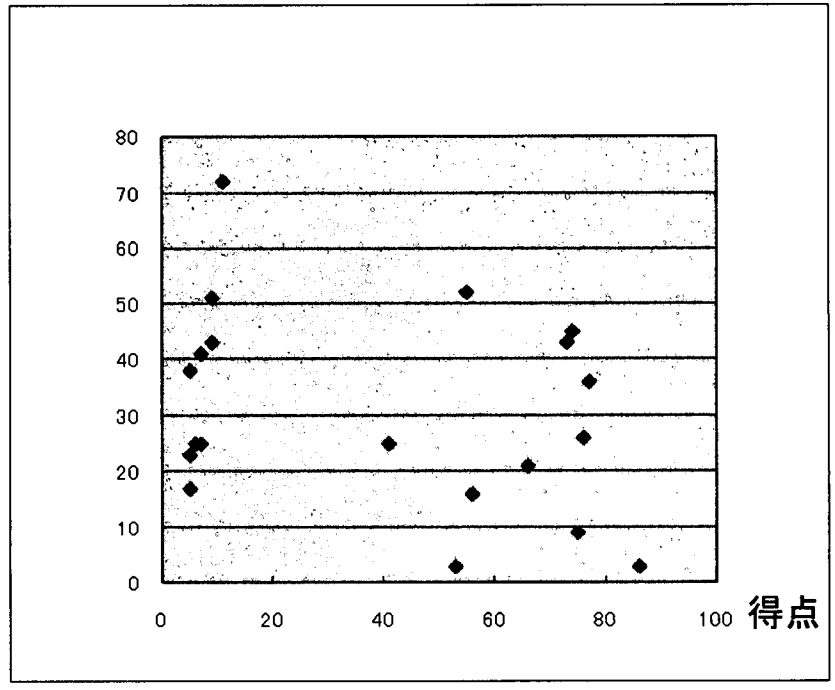
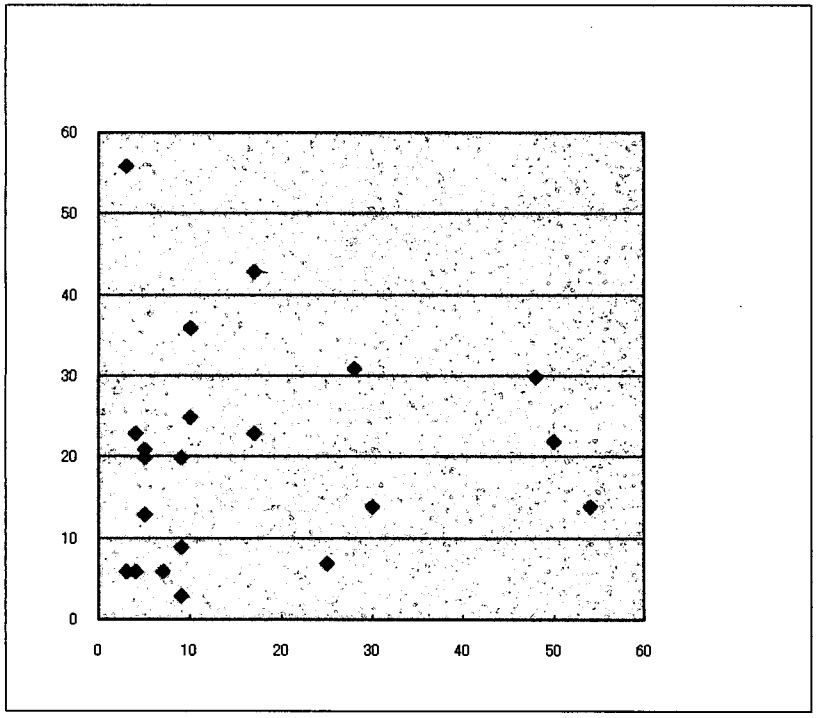


図2　得点と選択項目の相関　SLE

得点と選択問診数



得点と選択検査数

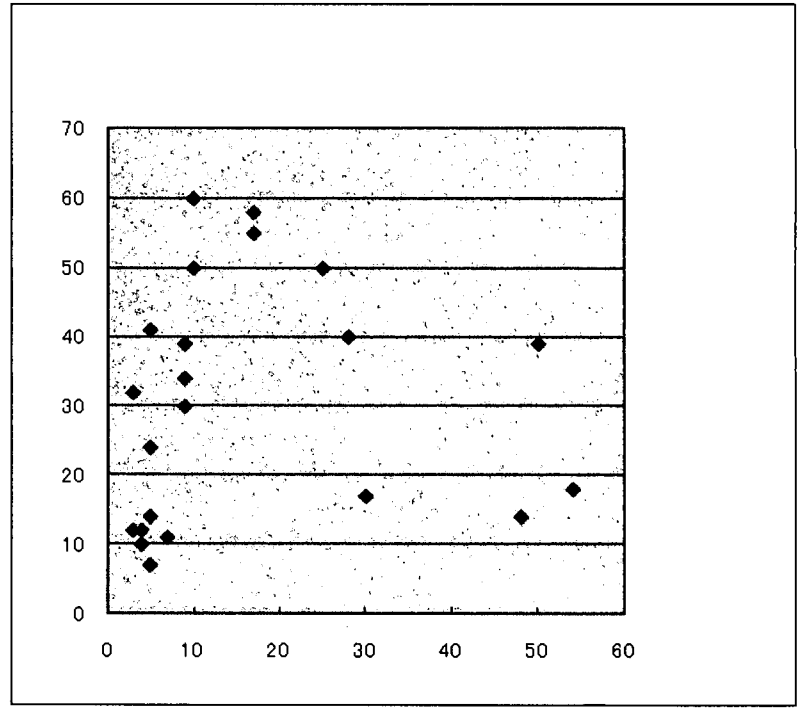
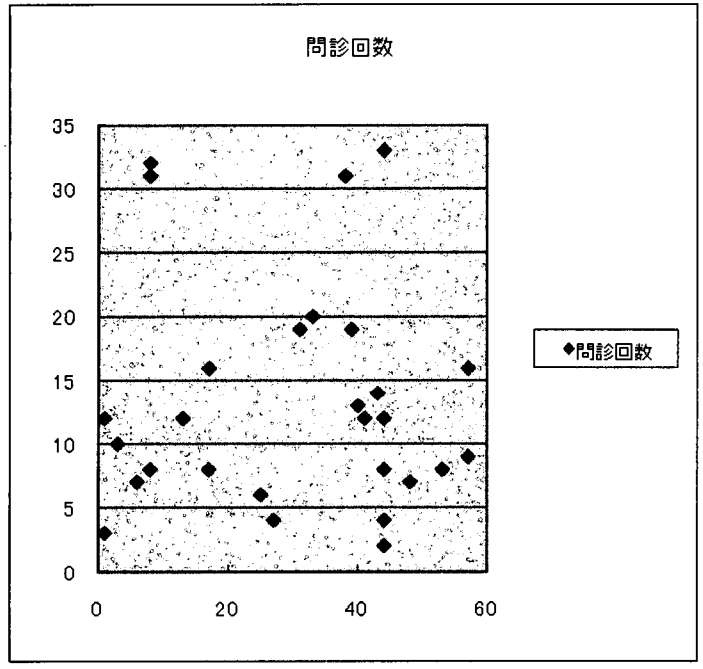


図2　得点と選択項目の相関　　腎盂腎炎

得点と選択問診数



得点と選択検査数

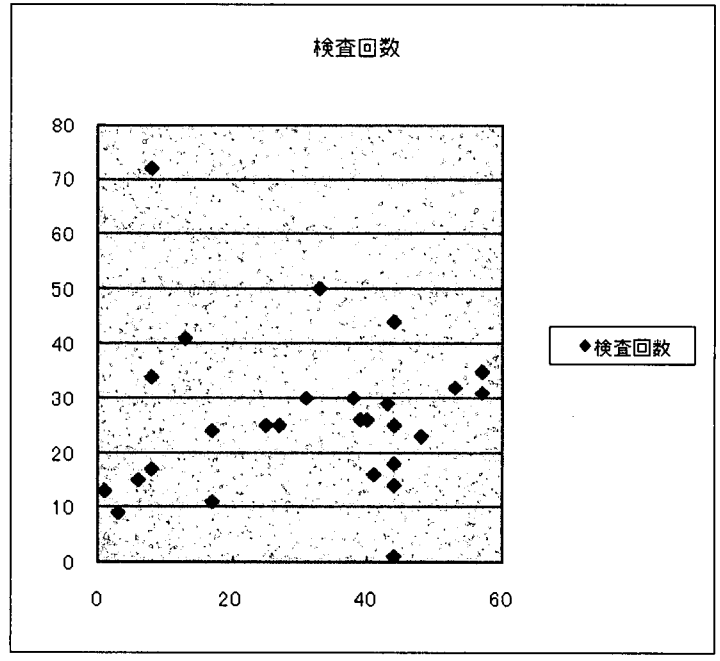


表2　　得点と正解項目率の関係

	問診正解率			初めの5問正解率		
疾患	インフルエンザ	SLE	腎盂腎炎	インフルエンザ	SLE	腎盂腎炎
学生	0.394±0.192	0.317±0.180	0.424±0.29	0.73±0.254	0.42±0.25	0.49±0.27
医師	0.456±0.158	0.360±0.095	0.486±0.244	0.511±0.176	0.27±0.11	0.5±0.38

問題作成のサンプル（インフルエンザ診断）

	A型インフルエンザ(追加)			
	B型インフルエンザ(追加)	ない		
	気管支喘息(追加)	ない		
	菌血症(追加)	ない		
99998	左心不全	ない		
99997	急性左心不全	ない		
97371	SAH	ない		
97291	RTA	ない		
97141	SIADH	ない		
97101	SIRS	ない		
33001	急性心筋梗塞	ない		
33000	うっ血性心不全	ない		
13132	腎障害	ない		
13118	滲出性胸膜炎	ない	正解　他に選べ	5(SLEを選んだときは0)
13104	心室瘤	ない		
13076	深在性真菌症	ない		
13063	進行麻痺	ない		
13004	進行性壊疽性鼻炎	ない		
12962	腎硬化症	ない		
12954	腎結石	ない		
12946	腎結核	ない		
12752	腎炎	ない		
12751	心炎	ない		
12743	腎盂腎炎	ない		
12738	腎盂拡張〔症〕	ない	問題から項目削除	
12732	心因性多飲症	ない		
12728	心因性うつ病	ない		
12722	心アミロイドーシス	ない		
12655	自律神経失調症	ない		
12636	初老期痴呆	ない		
12634	初老期うつ病	ない		
12604	ショック肺	ない		
12603	ショック腎	ない		
12588	女性仮性半陰陽	ない		
12532	食品アレルギー	ない		
12525	食道無弛緩症	ない		

厚生労働科学研究費補助金(医療技術評価総合研究事業)

医師国家試コンピューター化に関する研究

分担研究報告書

コンピューターで実施する診療問題解決型試験の検討

分担研究者　吉岡俊正　東京女子医科大学医学部医学教育学教授

研究要旨
コンピューター試験で可能となる臨床能力評価についての検討をおこなった。対象とした臨床能力は臨床推論と臨床判断であり、臨床シミュレーションする問題解決能力試験（Problem-solving ability test, P−SAT）を開発し評価した。新しい出題フォーマットとして多選択肢問題、ショートエッセイ問題（短文記入型問題）、語句抽出型問題の３つの型を確立した。学生で実証実験を行い、学生の回答パターンと専門医（熟練医）の判断との乖離で評価をおこなった。判断能力評価と共用試験CBTの成績と明らかな相関はなく、問題解決能力評価は通常のCBTとは違う能力を評価していることが示唆された。

② 研究目的

平成17-18年度の調査で、医師国家試験として臨床的問題発見解決能力を評価することを各国が検討し、米国では臨床事例について臨床的思考を評価するコンピューター試験を国家試験として導入していることが明らかになった。

本研究では従来の紙媒体で実施することのできない問題解決能力評価をコンピューター試験で実施することができるかについて、Problem-solving ability test (P-SAT)の実証実験を行った。P-ＳATは平成17年度から作成を開始した臨床推論・判断に特化したコンピューター試験で、患者への医師の対応のシミュレーションのなかで臨床的能力を評価する試験システムである。平成19年度は第4学年を対象に実証実験を行い、学生の解答パターン分析、共用試験ＣＢＴ成績との相関などを検討し、P-SATの信頼性・妥当性について検討を行った。

③ 研究方法

2) P-SAT 作成の背景

臨床的思考力として問題発見解決能力を開発することは医学教育で重要であるが、その評価法は確立していない。東京女子医科大学では1990年にテュートリアル教育を導入し、学部4年間の教育で臨床的問題発見解決能力を開発することを目的に教育を行っている。従来問題発見解決能力は、トリプルジャンプ法あるいは客観的臨床能力試験などで評価されたが、人的・時間的制約が高く、妥当性についてもハイステークスな評価に用いることができるか明確でなかった。

近年のe-ラーニング、ＣＢＴの発達とともに判断能力を評価する方法が開発されてきた。そこで、テュートリアル教育の中で開発する臨床推論・臨床判断能力についてコンピューター試験で評価できると考え、試験システムを開発した。

3) 出題形式

a. 臨床事例について、医師が考え、判断する順序に従ってその分析・判断を回答する臨床シミュレーション型の試験を構築した。
b. コンピューター化することにより、新たな情報が加わりながら判断・推論が深化する臨床の過程に沿って連続した設問を設定し、一旦回答すると元へ戻れない形式にした。
c. 学内試験・国家試験等で実施可能なウェッブブラウザを利用したコンピューター試験として構築した。
d. 3 つの問題形式を設定した。
   - 多選択肢問題：6 から 50 個の選択肢から

解答を選ぶ形式で、従来の五肢選択問題と比して類推、あるいは除外によって回答することが難しくなる。
- 語句抽出型問題：事例の中から問題発見・解決に必要な語句を選ぶ形式の問題。
- 短文記入型問題（ショートエッセイ問題）：必要な情報、判断の根拠などを短い文（語句）で記入する問題。

4） 評価方法

a. 臨床推論・判断では適切な推論・判断が一つとは限らないが、熟練した医師は、根拠の基づいた妥当性の高い判断をすることから、クリニカル・エビデンスなどに基づく専門医の判断（解答）との一致で解答を評価した。

b. 評価段階として以下の4段階を設定した。
- A：専門医の判断（選択）と完全に一致する
- B：専門医の判断と一部一致し、かつ不適切あるいは行ってはならない判断（選択）をしない
- C：専門医の判断とは一致しないが、間違いではない判断を行い、かつ不適切・行ってはならない判断をしていない。
- D：間違ったあるいは行ってはならない（禁忌）の判断をしている。

c. 評価システム
- 評価についても電子評価システムを構築した。
- ただし、語句抽出、短文記入においては、回答者が必ずしも予想された解答を行うとは限らない。その場合は、出題者が妥当性を再評価し正誤を新たに決定する。これもウェッブ上で行えるので従来の紙による記述式問題よりも評価が行いやすく、受験者全員について回答が適応されるので評価の正確性・公平性が高まる。

d. 実証実験
- 平成20年2月6日に医学部学生（第4学年）95名についてP-SATを行なった。

e. 事後評価
- 学生の解答パターンの評価
- 試験問題数の適正
- 識別指数（A評価を多く取っている回答者がAB評価で、D評価を多く取っている回答者がD評価となる割合
- 共用試験CBT成績との比較

④ 結果

(ア)診療問題解決能力評価における問題作成サーバーの構築
- 診療問題解決能力評価を実施するために、問題を作成する専用サーバーを構築した。サーバーには、日本HP ProLiant ML350を用い、OSとして、Microsoft Windows 2000 serverを使用した。Webサーバーは IISを使用した。本サーバーには問題作成者が離れた場所にいても問題作成ができるようWebによる問題作成機能を構築した。さらにセキュリティーを強化するためhttpsプロトコールによる問題作成にも対応した。本機能により、問題作成者は自分のデスクから問題を作成でき、さらにそれらのデータを暗号化する事により、安全にデータを送信する事ができた。

(イ)診療問題解決能力評価システムの構築
- 診療問題解決能力評価システムではテストを実施後、受験者の解答データを解析するため、解答データを別サーバーに蓄積した。データ解析プログラムは、HTMLと親和性の高いPHP言語を用い、ホームページ上から試験実施者が設定・解析できるよう作成した
- 問題作成サーバーと同様にセキュリティーを強化する目的で、SSLによる暗号化にも対応させた。本システムでは、試験管理、設問管理、試験データ取り込み、評価判定、集計、検索の6機能を装備した。

(ウ)P-SAT実証実験

① 試験実施結果

② 平成20年2月6日に95名の学生が受験した。

③ 2時間で58問を出題したが、全問題を解けない学生が約45%いた。

④ 前年度に実施し、設問・解答パターンが明らかになっている29問についてAからD評価を行った。

⑤ 95名中、A評価16-20が11名、11-15が59名であった。

⑥ D評価については、前年度のトライアルであDを4つ以上とったものが7%であったが、本年度はゼロであった。3つとったものが2名あった。

⑦ 共用試験CBTとA評価の数、D評価の数の相関：同時に行われた共用試験CBTの成績とA評価の数、およびD評価の数の相関を検討した。どち

らについても有意の相関は認めなかった。

⑧ 識別指数はD評価をとった受験生が少なかったため有意の所見を得なかった。

⑤ 考察

P-SATは臨床能力評価の新たなコンピューター試験として実用化できると考えられた。問題作成サーバーの構築の構築により、問題作成者は自分のいる部署からオンラインで問題作成を行う事ができた。本サーバーの構築において2つの問題点があった。一つはセキュリティーの問題である。問題を作成し、サーバー上に蓄積する事は、セキュリティーが低下し、問題漏出の危険が増すとされていた。本サーバーでは、作成ソフトにPerceptionを採用する事により、蓄積された問題を安全かつ長期的に保存する事が可能となった。2番目の問題点として、通信経路上での盗聴が考えられた。本システムでは、学内からのみアクセスできるネットワーク上に、本サーバーを設置し、ネットワーク経路を暗号化する事により、盗聴される可能性を低下させた。またハード面の対策として、スイッチの利用、ネットワーク監視システムなど既存のものを利用した。これらの問題点を解決する事によって本システムは、従来の問題作成システムに比べ、利便性・堅牢性が高いと考えられる。

診療問題解決能力を評価するためには、評価者が受験者の解答を吟味し、それらを客観的に評価する事が必要だった。従来評価者が受験者の診療問題解決能力を評価するには、時間と労力がかかった。この問題を解決するため、本研究では自動的に採点し評価するシステムを構築した。このシステムを利用することで、評価時間の短縮、採点ミスの減少が実現した。

P-SATは臨床推論・臨床判断能力評価として開発したコンピューター試験システムである。ただし、作問法によっては基礎医学のデータ解釈、臨床倫理判断などへの応用も可能であり平成19年度にはトライアル問題（採点対象外）として出題された。医師国家試験のプラットフォームとしてP-SATを考えた場合は、その評価目標・特性が従来の医師国家試験と全く異なることを考慮しなくてはならない。今回の実証実験と同じ時期（1日のずれ）で想起的知識評価として全国規模で行われ信頼度の高い共用試験CBTとの成績の相関を認めなかった。この結果は、ただちにP-SATの信頼性がないということではなく、P-SATとCBTでは別の能力を測定している可能性を検討しなくてはならない。臨床的な思考能力は医学教育モデル・コア・カリキュラムや医師国家試験出題基準にも明示されている、医師の基本的能力の一つであるが現在の医師国家試験では測定できない、P-SATは大学の教育理念に従った教育についての評価法という位置づけで開発されているが、これは医師の資質として一般に求められているものでもある。

P-SATの実証実験結果では、能力評価が一定の分布を示すことが明らかになった。臨床医に必要とされる推論能力・判断能力は従来の評価では測定が難しく、今後P-SAT評価結果が、臨床実習あるいは卒後研修などでの臨床推論・判断能力と相関するかの検討を行わなくては最終的信頼性判定ができない。

国際的に臨床能力（Clinical competency）が医師資格の目標となっており、カナダ、オーストラリアではObjective structured clinical examination、米国ではClinical skill assessmentが取り入れられている。これらの評価項目の一つが臨床推論・判断能力であるがP-SATはその評価を信頼度・再現性・妥当性・経済性・妥当性で上回るポテンシャルを持っている。今後国際的基準による評価を行うことにより、世界で医師資格の標準化が論じられている中で日本発信の解答となる可能性がある。

⑥ 結論

コンピューター試験は、経済性・効率性の利点だけでなく、従来の国家試験では評価できなかった医師としての能力特性を評価できる可能性を持ち、国家試験の評価内容と臨床および臨床研修で求められている医学の実践能力が乖離している現状を打破できる可能性がある。

今回のP-SATの結果は、臨床能力（臨床推論・臨床判断）評価の妥当性、再現性、識別性などについては更なる検討が必要であるが、日本独自のグローバルスタンダード評価としての可能性を持っている。

別紙４

研究成果の刊行に関する一覧表

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書　籍　名	出版社名	出版地	出版年	ページ

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Ishihara S, Matsui K, Sato Y, Tang AC, Suganuma T, Fukui Y, Yamaguchi N, Kawakami Y, Yoshioka T.	Self-efficacy achieved through problem-based learning tutorial.	医学教育	38	391-397	2007