A COMPUTER-BASED EXAMINATION TO EVALUATE CLINICAL COMPETENCE

by National Board of Medical Examiners*

3930 Chestnut Street
Philadelphia, Penna. 19104

Abstract

For the past 12 years efforts have been underway to design and test a computer-based simulation model that can be used to assess the problem-solving and decision-making skills of physicians during the licensing and certifying process. The model has been designed to assess these skills as they relate to the selection of diagnostic tests, procedures, and therapeutic and surgical interventions over the course of a health care episode. It can be used to track the multiple decisions a physician examinee makes as a clinical case unfolds. Because the model simulates time, it is also possible to monitor the physician's response to changes in the patient's status. The findings from field tests of the model provide encouraging evidence for the reliability, validity, and practicality of administering computer-based simulations as a component of the total evaluation program.

Introduction

As an evaluation agency involved in the development of examinations used in the licensure and certification of physicians and other health professionals, the National Board of Medical Examiners has long recognized the need for testing methods that would enable responsible agencies to evaluate problem-solving and decision-making skills. Problem-solving in medicine is a multi-stage process involving multiple iterations of data collection, data assessment, problem definition, and action steps. The process often involves recurring communication between the physician and the patient. The interactive, computational and branching capabilities associated with computer technology can be expected to provide a superior tool for the assessment of these abilities.

During the past 12 years, the National Board of Medical Examiners (with the American Board of Internal Medicine) has been sponsoring the development and experimental testing of a number of computer simulation models in clinical medicine. The goal has been to develop a computer-based, interactive system that simulates real life clinical situations and is capable of assessing the problem-solving and decision-making behaviors of physicians. During this period, there have been two developmental phases of the Computer-Based Examination (CBX) Project. The first phase from 1971 to 1974, under the direction of Dr. John Senior, utilized models focused exclusively on the diagnostic aspects of clinical decision-making without the capabilities of implementing therapies, monitoring the patient's course over time, or adjusting therapies as needed. The second phase of CBX development, conducted by Dr. Richard Friedman and his staff at the University of Wisconsin School of Medicine, Madison from 1977-1980, produced a CBX model that allows the assessment of management skills and clinical judgment used in the selection of diagnostic tests and procedures as well as drug and surgical therapy. This is the model that is now being refined by the National Board of Medical Examiners.

The Model

The present model presents a clinical case to a physician, provides historic and physical examination data at the request of the physician, allows the physician to order diagnostic tests and procedures at any time and in any sequence desired, and permits the physician to select management strategies and treatment. The CBX model simulates clinical time as well as drug and test interactions, resulting in a test patient whose clinical status changes in response to disease progression over time and to the physician's diagnostic and therapeutic interventions. The model not only provides clinical data in response to the physician's requests, but keeps the physician advised of the patient's changing clinical status and records the multiple decisions the physician makes.

A CBX case is initiated by entering a specific code number at the computer terminal. A brief introductory statement of the presenting problem is then displayed on the screen (e.g., A 35-year-old man comes to the emergency room because of chest pain), along with the current simulated time and day of the week. At this point, the physician is given control of the case and can order tests (including a history and physical examination) and therapies by utilizing code numbers assigned to

* CBX staff, in alphabetical order:
Barbara J. Andrew, Ph.D., Mary Jo Ballman, Robert J. Campbell, Vivian F. Erviti, Ph.D., James Galbraith, M.S., Nathan Holtzman, Paul R. Kelley, Jr., Ph.D., Eliot Nierman, M.D., Jennifer S. Pappas, M.A., Anne Strain, Victor C. Vaughan, III, M.D.

0195-4210/82/0000/0048$00.75 © 1982 IEEE

each of the tests and therapies, typing in the test or therapy name, or, for common tests, by using special dedicated keys on the computer terminal's keyboard. Specific command keys are available to cancel tests, to move the patient to a different location (inpatient, outpatient, intensive care unit), to discontinue therapies previously ordered, to request the results of the tests that are due, and to advance the simulated clock.

A critical aspect of the simulation is the dynamic nature of the cases. As in real life, when time is moved forward by the physician-examinee, the disease course progresses and the interactions between therapies and the clinical problem are affected. Test results are not reported to the physician as soon as they are ordered, but are deferred until the simulated time when they would normally be available. The simulated clock does not move until the physician explicitly requests that it be moved. It can be advanced either a specific amount of time (e.g., 24 hours) or to the time the next test result is available. When critical events occur in the case or when there is a major change in the patient's condition, there will be an interruption in the advance of the clock in order to report this occurrence. A report, called a "Nurse's note," is provided for the physician (e.g., "patient is complaining of marked shortness of breath") at this time. On receipt of such a note, the physician may order additional tests and therapies or may alter the management strategy used to this point before advancing the clock again.

Test results are either calculated or textual. A test that is reported in a numerical format is calculated. Examples of calculated tests include most clinical chemistry studies, blood pressure and pulse rate. A test that is reported in a narrative format is textual. Examples of textual tests are x-rays, EKGs, physical examination findings. Test results involving pictorial or graphic material are presented to the physician examinee by referencing a set of slides and microfiches. It is anticipated that a computer-controlled video recorder eventually will present this type of data.

The CBX model is based upon the premise that a physician, given access via a computer terminal to a "patient" with a specific clinical problem, a data base that contains the major findings from the history and physical examination, and the opportunity to access a virtually unlimited number of investigative and therapeutic choices, will engage in behaviors that reflect clinical skills in decision-making he or she uses with real patients. The simulated clinical cases are developed to present reasonably complex clinical problems that the physician must solve and manage effectively within expected time frames.

The Computer Programs

The CBX model as first developed was comprised of approximately 255 computer programs written in the MIIS language. The National Board of Medical Examiners is working toward the implementation of the CBX model in licensing and certifying examinations. To achieve optimum computing efficiency and to assure appropriateness for administration to thousands of examinees, the current developmental version of the CBX model now consists of over 50 different computer programs written in PL/1. These programs can operate on a wide range of IBM computers. The programs can be divided into four categories: (1) Model Definition/Maintenance, (2) Script Definition/Maintenance, (3) CBX Simulation Program, (4) Scoring/Output Programs.

The current model contains over 500 tests and 200 therapies that can be ordered at any simulated time. Results of each test for a normal healthy individual, as well as the effect of therapies on test results for this individual, are stored in default data files and can be retrieved whenever necessary. Project staff can add new tests and therapies to the system, alter the results of specific tests and test-therapy interactions, and produce reports on the present status of tests included in the model. These changes can be made without altering the actual simulation program.

Script entry programs permit a clinical case to be entered into the system via a CRT either by a physician or by an editor working under the direction of a physician. The programs prompt the script author to enter the specific clinical and/or descriptive information required. Script entry programs permit easy modification of information entered and all entries are edited subsequently for completeness, accuracy and credibility. The data in the script entry files act as overrides to the data base of normal results.

Information on the effect of each of the therapies on a healthy 31-year-old white man was collected from the literature and from expert judgments of consultant pharmacologists and is in the model for access whenever necessary. Data on the cost, risk and availability of all tests and therapies that are available in the model for both inpatient and outpatient settings was gathered from physicians at the University of Wisconsin and from practitioners in private practice in Madison, Wisconsin. Data on therapies by route of administration were collected from the literature and from the studies published by the Boston Drug Surveillance Group. For each new test or therapy to be added to the system, the cost, risk, time of availability of result, normal results, therapy-test relationships, and test-test relationships in a healthy 31-year-old white man must be added to the model. Case-specific risks are also added as appropriate. Default (normal) results for all tests are used when a particular CBX case has not otherwise defined a specific laboratory result. The model also contains reasonable effects of therapies on test results in normal people. These default therapy-test interactions are used whenever the CBX case does not specify interactions. These default therapy-test interactions will act on the test values that were provided in the case or, when no test value is provided, the interactions will act on the default test values in the model. This design permits case authors to focus on the underlying disease and treatment of a specific case and lets the model handle non-case-specific results.

Case analysis programs automatically analyze scripts to monitor consistency in physiologic relationships. For example, these programs will calculate anion gap, blood gas relationships and leukocyte differential at every moment for a particular case and identify and correct unacceptable relationships.

Script search programs will permit the physician to search existing cases for particular types of data that are required in cases under development. When data required for a new case are found, they can be copied to the new case, thus eliminating the tedious task of duplicating existing data.

The CBX simulation program maintains a complete record of all user commands. These data are utilized by the scoring/output programs to provide reports of physicians' actions which can be reviewed manually and automated scoring based on algorithmic criteria. The output programs print all types of data requested from the system in a legible and analyzable format. These programs list default data, test or therapy specifications, case scripts, physicians' performance records, etc.

Results of Field Studies

In order to investigate the reliability and validity of CBX and to determine its potential usefulness as a testing methodology in licensing and certifying examinations, field studies were conducted. The studies were designed to determine the relationship between scores on CBX cases and scores on other indices of knowledge and clinical competence such as the Part I, II, and III Examinations of the National Board of Medical Examiners. The objective was also to investigate whether CBX cases would discriminate among physicians at different levels of training and experience. In light of the competencies CBX was designed to assess, it was important to investigate whether more experienced and presumably more highly skilled physicians would score significantly higher on CBX cases. Of particular interest was a comparison of the performance of first- and third-year residents.

A prototype scoring strategy was developed such that performance on CBX cases would be evaluated in relation to the patient outcomes achieved, the amount of unnecessary cost generated during the simulation, the degree of unnecessary risk to which the patient was exposed, the appropriateness of the decisions made at various nodal points in the evolution of each clinical case, the degree of harm incurred from diagnostic and therapeutic omissions, and the degree of benefit accrued from diagnostic and therapeutic decisions made.

In the field study, six simulated cases were administered to first- and third-year residents in internal medicine. Reliability estimates of the individual CBX scoring variables ranged from .63 to .79, with the exception of the patient outcome variable where the reliability estimate equaled .31 (very likely due to the small number of items used to score this variable). When diagnostic and therapeutic benefit scores were combined and normalized, the reliability estimate increased to .83. When cost and risk scores were combined and normalized, the reliability estimate increased to .82. The reliability estimates for the combined scores compare favorably with those generally obtained on examinations composed of patient management problems, although the estimates for the individual scoring variables are lower than those usually obtained. Efforts will be directed in the coming years to identifying mechanisms for improving these reliability estimates without compromising the fundamental nature of CBX.

To evaluate one aspect of the redundancy of CBX, the residents' CBX scores were compared with their scores on the National Board examinations. It was hypothesized that CBX scores would be moderately correlated with the NBME scores, which would suggest that the tests measured some of the same aspects but not to the extent that one of the measures was superfluous. CBX scores were expected to be moderately correlated with scores on Part III, which is designed to evaluate aspects of clinical competence related to those assessed by CBX, and less correlated with Parts I and II, which primarily assess knowledge of the basic medical and clinical sciences, respectively. The field trial data confirm these hypotheses. The clinical knowledge and skills assessed by the NBME examinations may be prerequisites for effective management of simulated patients, but they are not sufficient for predicting performance on the CBX cases.

A short written examination was also constructed specifically for the field studies. The examination consisted of both multiple-choice questions and patient management problems and the content areas covered were matched as closely as possible to the content areas covered by the CBX cases. The MCQ scores for all residents were significantly correlated with the CBX scores, but these correlations were modest. Similarly, the residents' PMP scores showed significant but low correlations with the CBX scores. The results suggest that there is very little overlapping variance between what CBX measures and what the more traditional forms of evaluation are assessing.

The data analyses showed that third-year residents demonstrated significantly better performance on all CBX scoring variables with the exception of scores on diagnostic benefit and diagnostic omission. Third-year residents achieved better patient outcomes at less unnecessary cost and risk to the patient. They made fewer therapeutic omissions, achieved greater therapeutic benefit, and demonstrated better decision logic.

The National Board is now conducting further analyses to refine the scoring strategy for CBX. Other measurement issues are being addressed, such as the problems of determining pass/fail standards, of sampling, and of equating of cases as to difficulty. It is also anticipated that the process of developing CBX cases can be streamlined and systematized. Finally, efforts are underway to identify an appropriate delivery system which will permit the National Board to introduce CBX testing into its examination programs within the next several years.