SPECIAL ARTICLE

PERFORMANCE OF FOUR COMPUTER-BASED DIAGNOSTIC SYSTEMS

ETA S. BERNER, ED.D., GEORGE D. WEBSTER, M.D., ALWYN A. SHUGERMAN, M.D., JAMES R. JACKSON, PH.D., JAMES ALGINA, PH.D., ALFRED L. BAKER, M.D., EUGENE V. BALL, M.D., C. GLENN COBBS, M.D., VINCENT W. DENNIS, M.D., EUGENE P. FRENKEL, M.D., LEONARD D. HUDSON, M.D., ELLIOTT L. MANCALL, M.D., CHARLES E. RACKLEY, M.D., AND O. DAVID TAUNTON, M.D.

Abstract *Background.* Computer-based diagnostic systems are available commercially, but there has been limited evaluation of their performance. We assessed the diagnostic capabilities of four internal medicine diagnostic systems: Dxplain, Iliad, Meditel, and QMR.

Methods. Ten expert clinicians created a set of 105 diagnostically challenging clinical case summaries involving actual patients. Clinical data were entered into each program with the vocabulary provided by the program's developer. Each of the systems produced a ranked list of possible diagnoses for each patient, as did the group of experts. We calculated scores on several performance measures for each computer program.

Results. No single computer program scored better than the others on all performance measures. Among all cases and all programs, the proportion of correct diagnoses ranged from 0.52 to 0.71, and the mean proportion of relevant diagnoses ranged from 0.19 to 0.37. On average, less than half the diagnoses on the experts' original list of reasonable diagnoses were suggested by any of the programs. However, each program suggested an average of approximately two additional diagnoses per case that the experts found relevant but had not originally considered.

Conclusions. The results provide a profile of the strengths and limitations of these computer programs. The programs should be used by physicians who can identify and use the relevant information and ignore the irrelevant information that can be produced. (N Engl J Med 1994; 330:1792-6.)

OVER the past 20 years, computer-based systems designed to support clinical decision making have evolved from prototypes to commercially available systems.[1-10] Although many of these systems address narrow areas of subject matter, such as electrolyte and acid–base disorders,[2] diagnostic computer-based systems intended to address the entire field of internal medicine have gained increasing visibility.[11-16] Although most of these systems are generally designed to provide efficient access to medical information, they also include mechanisms for the assessment of clinical and laboratory data and the provision of diagnostic advice. As such systems become more widespread, evaluation of their diagnostic accuracy and usefulness to physicians is necessary. Studies of accuracy whose results have been reported have generally involved individual programs, a limited number and type of cases, and varying criteria and measures of performance.[15,17-32]

This study evaluated the ability of four programs — Dxplain (PC version 4.5),[33] Iliad (version 4.0),[34] Meditel (version 2.0),[35] and QMR (version 2.03)[36] — to suggest appropriate diagnoses to account for a set of clinical data. We used the same diagnostically challenging cases for each of the four systems and developed a number of measures of performance. We incorporated principles used in the development of specialty-board certification examinations to provide reliable estimates of performance — namely, a prospectively determined set of test specifications and an adequate number of cases, with an appropriate range of content and difficulty.

The four programs we studied have all been the subject of published research on their development, evaluation, and application.[11,12,14,15,19-32,37-43] Although they all incorporate expert judgment, they differ in the data used to determine their probability estimates, the extent to which diseases and related clinical data are addressed in their knowledge bases, the particular vocabulary they require to describe clinical data, and the algorithms they use to combine and analyze data. Iliad[34] and Meditel[35] use Bayesian logic, but they differ in the assignment of prior probabilities, in specific decision rules, and in the use of expert judgment. Dxplain[33] and QMR[36] use non-Bayesian algorithms, but they incorporate semiquantitative scales to express the probabilistic association of findings (signs and symptoms) with particular diagnoses, and they use these scales to derive a weighted assessment of the patients' combined signs and symptoms. After the data are entered, each program produces a list of diagnostic possibilities, ranked in order of likelihood. In general, none of the programs include a time-dependent dimension with regard to the appearance, sequence, or duration of signs and symptoms.

METHODS

Construction of the Test

All the cases involved the entire field of general medicine, including neurology. They were selected to present a spectrum of diagnostic difficulty but were all considered to be cases in which a physician might be prompted to seek diagnostic help from a colleague, in that they included atypical presentations, rare diseases, multiple disorders presenting simultaneously, or elements sufficiently complex that the physician would be likely to request a diagnostic consulta-

From the University of Alabama at Birmingham (E.S.B., A.A.S., J.R.J., E.V.B., C.G.C.); InforMed, Inc., St. Davids, Pa. (G.D.W.); the University of Florida, Gainesville (J.A.); the University of Chicago, Chicago (A.L.B.); the Cleveland Clinic Foundation, Cleveland (V.W.D.); the University of Texas, Dallas (E.P.F.); the University of Washington, Seattle (L.D.H.); Hahnemann University, Philadelphia (E.L.M.); Georgetown University, Washington, D.C. (C.E.R.); and Baptist Medical Center Montclair, Birmingham, Ala. (O.D.T.).

This study was conducted by the Office of Educational Development, University of Alabama at Birmingham School of Medicine, 933 19th St. South, Birmingham, AL 35294-2041, where reprint requests should be addressed to Dr. Berner.

Supported by a grant (LM05125) from the National Library of Medicine.

tion. All the cases were based on real patients. Those in which the principal challenge involved a choice among therapeutic options were excluded.

Each member of a group of 10 nationally recognized consultants in the fields of general internal medicine, eight subspecialties of internal medicine, and neurology contributed 15 detailed clinical summaries describing patients who had been referred for diagnostic consultation. The summaries included data (history, findings of physical examination, and results of laboratory tests) that were available at the time of the initial consultation and that indicated both normal and abnormal conditions. We omitted data collected subsequently at the consultant's direction; these usually included the definitive test that confirmed the diagnosis. Because the clinical data pertained to real patients, a few cases included vague descriptions by patients of their symptoms, earlier diagnoses that may not have been accurate, or normal results of laboratory tests forwarded by the referring physician that, when the tests were repeated later, were found to be abnormal. The group of experts arrived at a consensus on the diagnoses that were appropriate to consider in each case. They categorized each case according to the organ system or systems involved, the cause of disease, and the diagnostic difficulty. The experts then reviewed the cases to ensure that the test had an appropriate range of difficulty, that the weight given to the major organ systems was approximately equal, and that there was an appropriate gold standard for the diagnosis designated as correct in each case (i.e., a definitive diagnostic test or finding at autopsy or a consensus of experts when no definitive test could confirm the diagnosis). After this review, 120 of the original 150 cases were selected for further consideration.

Analyses of Cases

We attempted to include all the data in the written case descriptions, not just the especially pertinent ones. To ensure that data entry was optimal, we asked the program developers to indicate how they would enter specific clinical data in their particular programs. Bias in vocabulary selection that might have occurred if the program developers had chosen the vocabulary used in a specific context was avoided by having them express in the language of their program a master list of discrete data, collected from all the cases and listed alphabetically under the general categories of history, physical examination, and laboratory assessment. We then entered the data from each case into each program, using the developers' terms for the clinical data on the master list. Because of the limitations of individual systems, some data could only be approximated in some programs, or could not be entered at all. The data were analyzed by each program, and each produced a list of possible diagnoses for the case, ranked according to likelihood. All the analyses were carried out with versions of the four programs available in 1992.

After the programs had generated lists of diagnoses for a case, the top 20 diagnoses on each list were combined in a master list. Without knowing which program had suggested which diagnosis, the group of experts reviewed the diagnoses on the master lists for appropriateness, attempting to determine whether the programs had suggested any additional diagnoses that were appropriate and whether any cases should be eliminated because of ambiguity other than that associated with the performance of an individual program. One hundred ten cases remained after this validation stage. An additional five cases were deleted from the final test because they contained too few items to be run on some of the programs. One hundred five cases remained, including diagnoses such as giant-cell arteritis, histiocytosis X, ankylosing spondylitis, distal renal tubular acidosis, dissecting aortic aneurysm with infarction of spinal cord, thyroid carcinoma, pneumococcal pneumonia and bacteremia, Hodgkin's disease, gastric ulcer, and pericardial constriction.

We next determined the percentage of the diagnoses arrived at for each case that were included in the knowledge base of each program and calculated five scores to characterize the program's performance. The first two scores were based on the entire list of diagnoses that the program generated. The score for Correct Diagnosis is the proportion of the diagnoses included on the diagnosis list generated by the computer that were correct or closely related to the diagnosis that was considered to be correct. This variable is analogous to the concept of sensitivity. The score for Rank is the average rank of the correct (or closely related) diagnosis as it appears on the computer-generated list. Three other scores were derived by reviewing the first 20 diagnoses listed by each program. Like the score for Correct Diagnosis, the Comprehensiveness score is based on the list of appropriate diagnoses originally developed by the group of experts. The Comprehensiveness score is the average proportion of the appropriate diagnoses agreed on by the experts that is included on a computer-generated list. It reflects the extent to which the computer suggested all the diagnoses that the experts originally thought should be suggested. In some instances, the programs proposed diagnoses that the experts had not originally listed but that in retrospect they agreed were reasonable to consider. These diagnoses were the basis of two more scores. The score for Relevance is the average proportion of computer-generated diagnoses that the experts found reasonable to consider, given the clinical data. These diagnoses included the correct one and others that reflected an appropriate integration of the data. This score is conceptually, but not computationally, related to the notion of specificity. Finally, the score for Additional Diagnoses reflects the average number of additional diagnoses suggested by the computer that the experts considered appropriate after their final review of the cases.

Statistical Analysis

For each program, we calculated means and 95 percent confidence intervals for each score on the basis of primary case diagnoses. These calculations were made for all 105 cases and also for the 63 cases whose correct diagnoses were contained in the knowledge bases of all four computer systems. The scores for Rank were based only on the cases for which the computer suggested the correct diagnosis; as a result, in that analysis the number of cases included varied according to the program.

The overall difference between program means on the performance scores was tested for statistical significance with a multivariate repeated-measures analysis of variance.[44] In the case of dichotomous case scores, the procedure described by Guthrie was used.[45] A separate analysis of variance was conducted for each score except the score for Rank, since rankings were not available for all cases. Statistically significant analyses of variance were followed with pairwise comparisons between systems.[46] As with the overall analysis of variance, the pairwise comparisons were also adjusted for dichotomous case scores, with use of the procedures described by Guthrie.[45] An alpha level of 0.05 was chosen to indicate statistical significance in all tests.

To study how the score for Correct Diagnosis would change with a more stringent cutoff point for the lists of diagnoses, the scores for Correct Diagnosis were examined at various cutoff points. A two-factor repeated-measures analysis of variance was used to test for a statistically significant interaction between program and cutoff point.

Results

Table 1 shows the proportion of the 105 cases for which the correct diagnosis was included in the knowledge bases of all four computer programs, as well as the scores obtained by each program on each performance variable. For each variable, results are shown both for the total number of cases and for the number of cases with diagnoses included in the knowledge bases of all four programs — 105 and 63 cases, respectively, except in the case of Rank, for which the number of cases used varied according to the program. The numbers of cases on which the scores for Rank were based are included in a footnote to the table.

Knowledge Base

The proportion of the primary case diagnoses included in the knowledge bases of the individual programs ranged from 0.73 to 0.91. This value was significantly higher for Dxplain than for Iliad and QMR,

Table 1. Performance Scores of the Computer-Based Diagnostic Systems.

Variable and Sample Used*	Dxplain	Iliad	Meditel	QMR	Overall Analysis of Variance	P Value	Significant Pairwise Comparisons†
	mean (95 percent confidence interval)						
Diagnosis in Knowledge Base	0.91 (0.86–0.97)	0.76 (0.68–0.85)	0.85 (0.78–0.92)	0.73 (0.65–0.82)	$\chi^2 = 20.32$	<0.001	D vs. I, D vs. Q, M vs. Q
Correct Diagnosis							
105 cases	0.69 (0.60–0.78)	0.61 (0.52–0.70)	0.71 (0.62–0.79)	0.52 (0.43–0.62)	$\chi^2 = 11.58$	0.009	D vs. Q, M vs. Q
63 cases	0.79 (0.69–0.90)	0.76 (0.65–0.87)	0.89 (0.81–0.97)	0.71 (0.60–0.83)	$\chi^2 = 7.06$	0.070	—
Rank‡							
Diagnosis in program studied§	12.4 (9.5–15.3)	10.4 (8.0–12.8)	13.3 (10.5–16.1)	6.6 (3.0–10.3)	—	—	—
Diagnosis in all four programs¶	11.7 (8.3–15.1)	10.2 (7.5–12.9)	12.0 (8.8–15.3)	5.4 (3.7–7.1)	—	—	—
Relevance							
105 cases	0.24 (0.21–0.26)	0.19 (0.16–0.21)	0.22 (0.20–0.24)	0.37 (0.31–0.42)	F = 15.80	<0.001	Q vs. D, Q vs. M, Q vs. I, D vs. I, M vs. I
63 cases	0.26 (0.23–0.29)	0.21 (0.17–0.24)	0.23 (0.20–0.26)	0.46 (0.39–0.54)	F = 16.45	<0.001	Q vs. D, Q vs. M, Q vs. I, D vs. I
Comprehensiveness							
105 cases	0.38 (0.34–0.43)	0.25 (0.21–0.29)	0.38 (0.33–0.43)	0.28 (0.23–0.32)	F = 13.99	<0.001	D vs. I, D vs. Q, M vs. I, M vs. Q
63 cases	0.38 (0.33–0.44)	0.27 (0.22–0.32)	0.39 (0.32–0.46)	0.30 (0.25–0.35)	F = 5.05	0.004	D vs. I, D vs. Q, M vs. I, M vs. Q
Additional Diagnoses							
105 cases	2.3 (1.8–2.7)	2.0 (1.6–2.4)	2.1 (1.8–2.4)	1.8 (1.4–2.2)	F = 1.65	0.182	—
63 cases	2.6 (2.0–3.1)	2.2 (1.7–2.8)	2.2 (1.8–2.5)	2.0 (1.4–2.5)	F = 1.02	0.392	—

*The analyses of 105 cases were based on all cases included in the test, whereas the analyses of 63 cases were limited to the cases whose diagnoses were included in the knowledge base of all four programs.

†D denotes Dxplain, I Iliad, Q QMR, and M Meditel.

‡This variable could not be tested for significance because the sample varied in size according to the program used.

§This analysis included variable numbers of cases (72 for Dxplain, 64 for Iliad, 74 for Meditel, and 55 for QMR).

¶This analysis included variable numbers of cases (50 for Dxplain, 48 for Iliad, 56 for Meditel, and 45 for QMR).

and it was significantly higher for Meditel than for QMR. Three diagnoses were not included in any of the knowledge bases.

Correct Diagnosis

When all the cases were considered, scores for Correct Diagnosis ranged from 0.52 to 0.71 among the four computer programs. The mean scores for Dxplain and Meditel were significantly higher than the score for QMR. For nine cases, none of the programs included the correct diagnosis.

Using the scores for Correct Diagnosis, Figure 1 shows the proportion of cases in which the correct diagnosis was the first diagnosis listed, the proportion in which it was listed as 1 of the top 5 diagnoses, as 1 of the top 10, and so forth. There was a significant interaction between the program and the cutoff point used (chi-square = 70.28, 21 df, $P<0.001$); QMR had the highest score for Correct Diagnosis when the top 10 diagnoses were studied but the lowest score when the entire list was used. The programs were least distinguishable from one another with regard to Correct Diagnosis when cutoff points of 15 and 20 diagnoses were used.

In the analysis of the 63 cases whose diagnoses were included in all four knowledge bases, the scores for Correct Diagnosis ranged from 0.71 to 0.89. As would be expected, the mean score for each program was higher when the sample studied was limited to cases with diagnoses in the knowledge base of the program. The differences between programs were not statistically significant. Among the 63 cases, there was only 1 for which none of the programs suggested the correct diagnosis.

Rank

Among the cases for which each system generated a correct diagnosis, the mean rank of that diagnosis on the computer-generated list ranged from 6.6 to 13.3. For cases whose diagnoses were contained in all four knowledge bases, the mean rank of the correct diagnosis ranged from 5.4 to 12.0. Because the samples varied in size, the significance of the differences could not be calculated.

Relevance

The mean scores for Relevance ranged from 0.19 to 0.37 when the entire sample was studied. The mean score for QMR was significantly higher than those for the other programs, and the mean score for Iliad significantly lower than those for the other programs. When the 63 cases whose diagnoses were included in all four knowledge bases were studied, the scores ranged from 0.21 to 0.46. The scores for QMR were still significantly higher than those for all the other systems, but the only other significant difference was that the score for Dxplain was significantly higher than that for Iliad.

Comprehensiveness

Among the four programs, the mean scores for Comprehensiveness ranged from 0.25 to 0.38 when all cases were studied and from 0.27 to 0.39 when the 63 cases whose diagnoses were included in all four knowl-

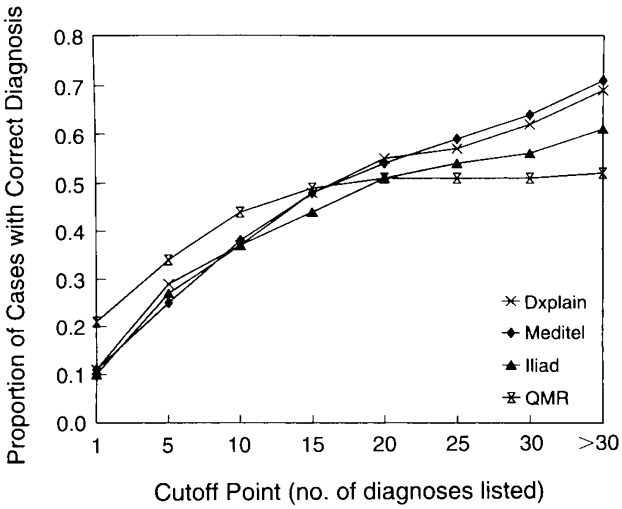

Figure 1. Proportion of Cases with a Correct Diagnosis in the Computer, According to the Cutoff Point Establishing the Numbers of Diagnoses Listed.

edge bases were studied. In both analyses, the mean scores for Dxplain and Meditel were significantly higher than those for Iliad and QMR.

Additional Diagnoses

Approximately six appropriate diagnoses per case appeared on the lists originally compiled by the experts. When either all 105 cases or the sample of 63 cases whose correct diagnoses were included in the knowledge base were studied, each computer program generated an average of approximately two appropriate diagnoses that had not originally been listed. There were no significant differences among the systems with regard to this variable.

Discussion

In the evaluation of computer-based diagnostic systems, two major issues need to be addressed: accuracy and usefulness. This study addresses only the first issue and is focused only on the ability of a system to generate diagnostic hypotheses from a set of data pertaining to a case. The study involved developing a set of cases that were real and diagnostically challenging. Although the programs are not expressly designed for any particular group of physicians, the experts considered that the use of diagnostic systems would probably be important for physicians presented with difficult clinical problems for which they might seek consultation. A problem might be challenging because it involved atypical findings for a common disease, a rare disease, or the interaction of multiple diseases. The most common seekers of such consultations were considered to be primary care physicians or subspecialists needing assistance outside their area of expertise. The clinical cases were chosen for a representative balance of these types of problems, as well as a balance of problems among organ systems. Patients being referred for diagnostic assistance to a broad spectrum of experts were considered an appropriate source of such cases. Although the resulting cases are likely to represent only a small portion of a generalist's normal case load, they are likely to represent a larger portion than a case sample that is limited to clinicopathological conferences and they may, in fact, reflect a large portion of the cases for which diagnostic help is sought.

The programs all produced moderately long lists of potential diagnoses. The lists included many diagnoses that a knowledgeable physician would regard as not being particularly helpful in explaining the case or guiding further studies. On the other hand, each program suggested some diagnoses, though not highly likely ones, that the experts later agreed were worthy of inclusion in the differential diagnosis.

Although each program performed better or worse than others on some of the performance measures, none performed consistently better or worse on all the measures. In many cases the differences, even when statistically significant, were not large. The relative importance of the measures is likely to depend on the individual user's preferences and needs. One of the greatest differences concerned the proportion of case diagnoses in the knowledge bases of the programs (range, 0.73 to 0.91). This variable may explain some of the differences in the overall scores for Correct Diagnosis and Comprehensiveness. The scores for Rank indicate where the diagnosis that was ultimately found to be correct appeared on the list of computer-generated diagnoses. For an atypical case, the correct diagnosis might appropriately be ranked fairly low if other diagnoses were more likely on the basis of the available data. For this reason, some system developers have emphasized that for the appropriate diagnoses to be included on the list at all is more important than their rank. It should also be remembered that the scores for Comprehensiveness and Additional Diagnoses both depend on the number of diagnoses in the initial expert consensus. Since the experts tried to list all the diagnoses that should be considered, the scores for Comprehensiveness and Additional Diagnoses are likely to be lower than they would be if the list had only included a few of the reasonable diagnoses.

Although the sensitivity and specificity of the programs tested in this highly focused study were not impressive, the programs have additional functions that we did not evaluate. These functions, many of which are interactive, include displaying the signs and symptoms associated with diseases, suggesting potentially relevant laboratory tests, and proposing alternative workup strategies. In addition, these programs provide scores that indicate the relative likelihood of each diagnosis. In this study, only the ranking on the diagnosis lists was used, rather than these likelihood scores.

The increasing popularity of computer-based diagnostic systems suggests that at least some physicians have found them helpful. However, such anecdotal data do not permit a systematic assessment of the clinical contexts in which these programs are most useful or of how they actually perform. Our study

arouses concern that important diagnostic considerations may be so obscured by other diagnoses that the value of the program may be significantly decreased, or that it could lead to excessive or costly interventions in inexperienced hands. However, results indicating low sensitivity and specificity do not in themselves show how these systems perform in a clinical setting. Although some clinicians may use one of these programs as described here, most would probably enter selected key findings and use some of the other functions of the system to refine the list of diagnoses. Medically knowledgeable persons would probably not only decide what data to enter, but also distinguish between diagnoses that are worthy of consideration and dismiss many of the poorly integrated diagnoses.[47] The developers of these systems intend these programs to serve a prompting function, reminding physicians of diagnoses they may not have considered or triggering their thinking about related diagnostic possibilities.[11,23] Clearly, as others have indicated, the next step in the evaluation of these programs will have to include examining the performance of the physician and the computer together.[48-50]

We are indebted to Faith Fitzgerald, M.D., for her contributions to the deliberations of the group of experts and her insightful comments on an earlier draft; to G. Octo Barnett, M.D., Randolph A. Miller, M.D., Homer Warner, Jr., Herbert S. Waxman, M.D., and William E. Worley, M.D., the developers of the diagnostic decision support systems, and their colleagues, Nuncia Giuse, M.D., Marvin Packer, M.D., and Hong Yu, M.D., for providing data; to Ms. Janice S. Pulliam for her diligent efforts as a research assistant; and to Ms. Mary Sue B. Pruett for her assistance in the preparation of the manuscript.

References

1. Barnett GO. The computer and clinical judgment. N Engl J Med 1982;307:493-4.
2. Bleich HL. The computer as a consultant. N Engl J Med 1971;284:141-7.
3. de Dombal FT. Computer-aided decision support in clinical medicine. Int J Biomed Comput 1989;24:9-16.
4. DeTore AW. Medical informatics: an introduction to computer technology in medicine. Am J Med 1988;85:399-403.
5. Miller RA. Medical diagnostic decision support systems — past, present, and future. J Am Med Informatics Assoc 1994;1:8-27.
6. Reggia JA, Tuhrim S, eds. Computer-assisted medical decision making. New York: Springer-Verlag, 1985.
7. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine: where do we stand? N Engl J Med 1987;316:685-8.
8. Shortliffe EH. Computer programs to support clinical decision making. JAMA 1987;258:61-6.
9. *Idem.* The adolescence of AI in medicine: will the field come of age in the '90s? Artif Intell Med 1993;5:93-106.
10. Shortliffe EH, Perreault LE, eds. Medical informatics: computer applications in healthcare. Reading, Mass.: Addison-Wesley, 1990.
11. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: an evolving diagnostic decision-support system. JAMA 1987;258:67-74.
12. Miller R, Masarie FE, Myers JD. Quick Medical Reference (QMR) for diagnostic assistance. MD Comput 1986;3(5):34-48.
13. Trace D, Evens M, Naeymi-Rad F, Carmony L. Medical information management: the MEDAS approach. In: Miller RA, ed. Proceedings: the Fourteenth Annual Symposium on Computer Applications in Medical Care. New York: IEEE Computer Society Press, 1990:635-9.
14. Warner HR Jr. Iliad: moving medical decision-making into new frontiers. Methods Inf Med 1989;28:370-2.
15. Waxman HS, Worley WE. Computer-assisted adult medical diagnosis: subject review and evaluation of a new microcomputer-based system. Medicine (Baltimore) 1990;69:125-36.
16. Weed LL. Knowledge coupling: new premises and new tools for medical care and education. New York: Springer-Verlag, 1991.
17. Georgakis DC, Trace DA, Naeymi-Rad F, Evens M. A statistical evaluation of the diagnostic performance of MEDAS — the Medical Emergency Decision Assistance System. In: Miller RA, ed. Proceedings: the Fourteenth Annual Symposium on Computer Applications in Medical Care. New York: IEEE Computer Society Press, 1990:815-9.
18. Nelson SJ, Blois MS, Tuttle MS, et al. Evaluating RECONSIDER: a computer program for diagnostic prompting. J Med Syst 1985;9:379-88.
19. Hammersley JR, Cooney K. Evaluating the utility of available differential diagnosis systems. In: Greenes RA, ed. Proceedings: the Twelfth Annual Symposium on Computer Applications in Medical Care. New York: IEEE Computer Society Press, 1988:229-31.
20. Feldman MJ, Barnett GO. An approach to evaluating the accuracy of DXplain. Comput Methods Programs Biomed 1991;35:261-6.
21. Heckerling PS, Elstein AS, Terzian CG, Kushner MS. The effect of incomplete knowledge on the diagnosis of a computer consultant system. Med Inf (Lond) 1991;16:363-70.
22. Lau LM, Warner HR. Performance of a diagnostic system (Iliad) as a tool for quality assurance. Comput Biomed Res 1992;25:314-23.
23. Barness LA, Tunnessen WW Jr, Worley WE, Simmons TL, Ringe TBK Jr. Computer-assisted diagnosis in pediatrics. Am J Dis Child 1974;127:852-8.
24. O'Shea JS. Computer-assisted pediatric diagnosis. Am J Dis Child 1975;129:199-202.
25. Swender PT, Tunnessen WW Jr, Oski FA. Computer-assisted diagnosis. Am J Dis Child 1974;127:859-61.
26. Wexler JR, Swender PT, Tunnessen WW Jr, Oski FA. Impact of a system of computer-assisted diagnosis: initial evaluation of the hospitalized patient. Am J Dis Child 1975;129:203-5.
27. Bankowitz RA, Lave JR, McNeil MA. A method for assessing the impact of a computer-based decision support system on health care outcomes. Methods Inf Med 1992;31:3-10.
28. Bankowitz RA, McNeil MA, Challinor SM, Parker RC, Kapoor WN, Miller RA. A computer-assisted medical diagnostic consultation service: implementation and prospective evaluation of a prototype. Ann Intern Med 1989;110:824-32.
29. Bankowitz RA, McNeil MA, Challinor SM, Miller RA. Effect of a computer-assisted general medicine diagnostic consultation service on housestaff diagnostic strategy. Methods Inf Med 1989;28:352-6.
30. Berman L, Miller RA. Problem area formation as an element of computer aided diagnosis: a comparison of two strategies within Quick Medical Reference (QMR). Methods Inf Med 1991;30:90-5.
31. Middleton B, Shwe MA, Heckerman DE, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. II. Evaluation of diagnostic performance. Methods Inf Med 1991;30:256-67.
32. Miller RA, Pople HE Jr, Myers ID. *Internist-I*, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982;307:468-76.
33. DXPLAIN. Boston: Massachusetts General Hospital, 1992.
34. ILIAD. Salt Lake City: Applied Informatics, 1992.
35. MEDITEL: computer assisted diagnosis. Devon, Pa.: Meditel, 1991.
36. QMR (Quick medical reference). Pittsburgh: CAMDAT, 1992.
37. Bankowitz RA, Blumenfeld BH, Giuse Bettinsoli N, et al. User variability in abstracting and entering printed case histories with QUICK MEDICAL REFERENCE (QMR). In: Stead WW, ed. Proceedings: the Eleventh Annual Symposium on Computer Applications in Medical Care. New York: IEEE Computer Society Press, 1987:68-73.
38. Bankowitz RA, Miller JK, Janosky J. A prospective analysis of inter-rater agreement between a physician and a physician's assistant in selecting QMR vocabulary terms. In: Clayton PD, ed. Proceedings: the Fifteenth Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1991:609-13.
39. First MB, Soffer LJ, Miller RA. QUICK (Quick Index to Caduceus Knowledge): using the INTERNIST-1/CADUCEUS knowledge base as an electronic textbook of medicine. Compute Biomed Res 1985;18:137-65.
40. Giuse DA, Giuse NB, Miller RA. Towards computer-assisted maintenance of medical knowledge bases. Artif Intell Med 1990;2:21-33.
41. Masarie FE Jr, Miller RA, Myers JD. INTERNIST-1 properties: representing common sense and good medical practice in a computerized medical knowledge base. Comput Biomed Res 1985;18:458-79.
42. Miller RA, Masarie FE Jr. The Quick Medical Reference (QMR) relationships function: description and evaluation of a simple, efficient "multiple diagnoses" algorithm. In: Lun KC, Degoulet P, Piemme T, Rienhoff O, eds. Medinfo 1992: proceedings of the Seventh World Congress on Medical Informatics. Amsterdam: Elsevier, 1992:512-8.
43. Miller RA, McNeil MA, Challinor SM, Masarie FE Jr, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project — status report. West J Med 1986;145:816-22.
44. Vonesh EF, Schork MA. Sample sizes in the multivariate analysis of repeated measurements. Biometrics 1986;42:601-10.
45. Guthrie D. Analysis of dichotomous variables in repeated measures experiments. Psychol Bull 1981;90:189-95.
46. Shaffer JP. Modified sequentially rejective multiple test procedures. J Am Stat Assoc 1986;81:826-31.
47. Rand TG. Medical knowledge bases free the mind for problem solving. ACP Obs 1992;12(11):10-1.
48. Salomon G, Perkins DN, Globerson T. Partners in cognition: extending human intelligence with intelligent technologies. Educ Res 1991;20(3):2-9.
49. Miller RA, Masarie FE Jr. The demise of the "Greek Oracle" model for medical diagnostic systems. Methods Inf Med 1990;29:1-2.
50. Miller RA. Why the standard view is standard: people, not machines, understand patients' problems. J Med Philos 1990;15:581-91.