SOUNDING BOARD

ARTIFICIAL INTELLIGENCE IN MEDICINE
Where Do We Stand?

[Reprinted from the *New England Journal of Medicine* **316**:685-688 (March 12), 1987]

After hearing for several decades that computers will soon be able to assist with difficult diagnoses, the practicing physician may well wonder why the revolution has not occurred. Skepticism at this point is understandable. Few, if any, programs currently have active roles as consultants to physicians. The story behind these unfulfilled expectations is instructive and, we believe, offers hope for the future.

Research on computer-aided diagnosis began in the 1960s with high hopes that difficult clinical problems might yield to mathematical formalisms. Most work therefore centered on the application of flow charts Boolean algebra, pattern matching, and decision analysis to the diagnostic process. Except in extremely narrow clinical domains, each of these techniques proved to have little or no practical value. Most observers came to believe that for a program to have expert capability, it must in some fashion mimic the behavior of experts. Early work on computer-aided diagnosis was thus largely discarded, and in the early 1970s attention shifted to the study of the actual problem-solving behavior of experienced clinicians.[2,5] The resulting insights have subsequently been used to construct models of clinical problem solving that, in turn, have been converted into so-called artificial-intelligence programs or expert systems.[1,6,7]

The Evolution of New Systems

The new generation of programs designed to emulate clinical expertise has developed along two quite different paths. One, called rule-based systems, has incorrectly become almost synonymous with the term 'artificial intelligence" in the minds of most physicians. These systems, such as MYCIN,[8] are based on the hypotheses that expert knowledge consists of a large number of independent, situation-specific rules and that computers can simulate expert reasoning by stringing these rules together in chains of deduction.[9,10] Each rule consists of an *If* statement followed by a *Then* statement. The former identifies a situation in which the conclusion or action specified by the latter can be carried out. For example, in a patient who has an infection, the initial rule might be, *If* there is an organism that makes therapy necessary, *Then* determine the best recommendation for therapy. The system then seeks out other rules that will help it decide whether a pathogenic organism is present. If a pathogen is found, the system mobilizes other rules to arrive at a treatment recommendation. The conclusion reached by a group of rules is accepted when a numerical scoring factor exceeds some critical threshold.

The pioneering work on rule-based systems was designed to deal with clinical problems,[9] but ironically, nearly all successes with such systems have occurred outside medicine. Rule-based systems have proved useful in a variety of commercial tasks, such as evaluating trouble with telephone lines, laying out preventive maintenance programs for large power stations, and configuring computer systems. But it is clear that these successes have been possible primarily because the domains that have been addressed are limited and because the programs are valuable despite their inability to perform at a nearly perfect level; failure to identify a defect in a telephone network can be tolerated more readily than a misdiagnosis in a seriously ill patient.

In contrast to phone systems and power stations, the domain of even a single major field such as internal medicine is so broad and complex that it is difficult, if not impossible, to capture the relevant information in rules. Furthermore, other important difficulties in rule-based systems prevent even relatively small programs from performing effectively and with an acceptable degree of reliability. Although rules appear to be free-standing entities, their interactions with other rules are not always consistent or predictable. To achieve the desired overall behavior from a system, the author of the rules must anticipate the ways in which each rule will interact with every other. Moreover, as the domain encompassed by a rule-based system is expanded, new knowledge often interferes with information already available, in ways that are unexpected and difficult to remedy.'[1]"[2] Such problems are hardly surprising, given that there is no explicit overall diagnostic strategy governing the flow of reasoning by the program. Rules capture only the surface behavior of experts, not the reasons they behave as they do. The obvious shortcomings of rule-based systems have constrained their practical application to a few limited clinical situations, such as the evaluation of pulmonary-function tests[13] and the interpretation of findings obtained by electrophoresis.[14]

During the 1970s, in parallel with the work on rule-based systems, there evolved a second and very different approach to the modeling of human clinical expertise. This school of thought views diagnostic acumen as the ability to construct and evaluate hypotheses by matching a patient's characteristics with stored profiles of the findings in a given disease. A numerical value is used to indicate how often a particular finding is encountered in a given disease a second value indicates how strongly a particular finding should arouse suspicion that the given disease is present. The finding is also weighted according to its clinical importance: e.g., massive gastrointestinal bleeding or a high fever is assigned more importance than low back pain or mild leukocytosis. Programs relying on the approach just described produce a diagnostic ranking of the various hypotheses by using a scoring method that considers these three weights.[15]

In one such program, the Present Illness Program,[16] all disease entities are considered competitors—a weakness that leads to poor performance when two diseases coexist. But another program, INTERNIST, overcomes this problem by using a strategy that allows it to identify a set of competing hypotheses and to postpone consideration of other diseases that may be present.[17]

Programs that match clinical findings with stored profiles of diseases often perform in an impressive fashion but nevertheless demonstrate serious weaknesses, which preclude their practical application in consultation. First of all, they are virtually unable to cope with variations in the clinical picture. In particular, they have difficulty in recognizing variations in the way that a disease can present, in terms of both the spectrum of findings and severity. They are also unable to cope with the evolution of a disease over time, as in the case of acute glomerulonephritis. Furthermore, they cannot recognize how one disease may influence the presentation of a second, or how the effects of previous treatment can alter the patient's illness. Finally, programs based on simple matching strategies are unable to explain to the physician how they have reached their conclusions. Some of these deficiencies were noted in an editorial in the *Journal* several years ago.[18]

By the late 1970s, disappointments with both rule-based systems and matching strategies stimulated investigators to move in directions that have led to new insights, even if not to clinically useful programs. Studies of problem-solving strategies employed by experts made it ever more apparent that clinical expertise in difficult medical cases is to a considerable extent reliant on causal, pathophysiologic reasoning. Investigators became aware that the key deficiencies in most previous programs stemmed from their lack of pathophysiologic knowledge. Only programs relying on such reasoning would be able to cope with the enormous number of ways in which diseases can present, evolve, and interact with each other.

In an effort to simulate expert performance, new programs build specific models of a given patient's illness by linking clinical findings with pathophysiologic knowledge stored in the program's memory.[19-21] Such knowledge is organized as nodes of information connected by links that specify causal relationships. Nodes contain knowledge such as the range of clinical and laboratory findings that can be anticipated at the onset of an illness and during its evolution. One program also incorporates information on how the clinical picture of disease varies with its severity.[20] Some nodes deal with effects and others with causes, and the links between them allow the program to use patient-specific information about the severity and temporal stage of an illness to determine whether the findings in the nodes match the disease hypothesis. Links not only reason forward, from cause to effect, but backward, from observed effects to their expected causes. The program can thus use observed findings to reason about causation, and vice versa.

When confronted with a case, be it a chief complaint or a larger body of information, the program constructs a small set of hypotheses that is consistent with the available information.[19-21] Pathophysiologic knowledge provides a powerful mechanism for establishing constraints that allow identification of the most reasonable hypotheses. In one program, the chief organizing principle consists of linking all important abnormalities by a chain of causal reasoning. Only the limited set of hypotheses that is compatible with the resulting logical structure then need be considered.[21]

To further the process of differential diagnosis, each hypothesis, as embodied in the computer-generated model of the patient's illness, is expanded to create a scenario that projects the consequences to be expected if that particular disease is present. On the basis of these scenarios, the program identifies additional information that could differentiate among the various diagnostic possibilities.[20] For example, the urinary sodium concentration would be singled out as a feature that can help to distinguish between oliguria due to acute tubular necrosis and that due to volume depletion. Scenarios thus provide a powerful strategy for efficient acquisition of further information, be it historical data, laboratory findings, or other relevant data.

Although detailed pathophysiologic knowledge has greatly increased the ability of a program to handle complexity, it has also added enormously to the computational task.[20~22] When a program employing causal reasoning is asked to explore each case in great detail, including straightforward cases that do not merit such attention, the process is so slow that it is impractical even with modern high-speed computers. To deal with this difficulty, a strategy has been developed that allows reasoning at multiple levels of detail.[22] In the straightforward cases the program begins by simply looking at shallow associational information (e.g., that pulmonary insufficiency causes hypercapnia). But when such a relatively simple strategy fails to resolve the problem, the program moves to deeper levels of reasoning that allow detailed evaluation of each observed abnormality and its contribution to the clinical picture. For example, a reported loss of blood that is not sufficient to account for an observed degree of anemia will alert such a program to look for other causes of bleeding or, in the absence of such a cause, to consider the possibility of a laboratory error.

Programs based on causal, pathophysiologic reasoning also have the great virtue of leaving a trail that can be converted into an English-language explanation of their diagnostic activities.[23,24] Without such explanations, it is obviously unreasonable for the physician to rely on such programs; ultimately, a program, like any consultant, must justify its conclusions to the physician responsible for the patient's care.[25]

Linking the Old and the New

Ironically, now that much of the artificial-intelligence research community has turned to causal, pathophysiologic reasoning, it has become apparent that some of the earlier, discarded diagnostic strategies may have important value in enhancing the performance of new programs. Programs that use causal reasoning typically have only general-purpose strategies for exploring competing hypotheses and playing out scenarios. But such strategies are often quite inefficient because the exploration of the diagnostic possibilities in particular clinical situations can be embodied in rules or flow charts that are highly specific to the problem at hand, such as gastrointestinal bleeding or chest pain.[26,27] Moreover, if further diagnostic information can be obtained only at the cost of some risk or pain, a component of decision analysis will almost certainly be a useful addition to a program, allowing systematic balancing of medical costs against medical benefits.[28,29] An extensive research effort is required, however, before all these techniques can be incorporated into a single program.

Interim Payoffs from Artificial-Intelligence Research

It is obvious from the foregoing discussion that we have not reached the point at which artificial-intelligence programs can act as reliable consultants on a wide range of medical problems. But, quite logically, attempts are under way to use specific components of recent research in artificial intelligence to implement programs that can provide simple but potentially valuable diagnostic assistance to the physician. For example, the enormous data base of INTERNIST[17] may provide a considerable advantage over textbooks to the physician who is searching for facts about a particular illness.[30] The data base for a given disease in medical textbooks typically consists simply of a list of manifestations, accompanied by ambiguous and relatively unhelpful qualitative descriptions—e.g., that a given finding occurs "frequently" or "uncommonly." The INTERNIST data base has the advantage, as we have noted earlier, of including numerical information on the frequency of findings and on the diagnostic importance of each finding. Furthermore, by applying relatively simple computing strategies to such a data base, the program can generate a list of hypotheses that may deserve consideration. Unfortunately, such lists are usually quite long and thus do not greatly narrow the diagnostic focus; instead, they provide a checklist that helps the user make certain that no diagnostic possibility has been overlooked.

Relatively simple systems such as those just described can also be used for critiquing diagnostic hypotheses or plans for treatment.[31,32] It is far easier for a program to examine the reasonableness of a plan constructed by a physician than to create such a plan itself.

What Does the Future Hold?

In 1970 an article in the *Journal* predicted that by the year 2000 computers would have an entirely new role in medicine, acting as a powerful extension of the physician's intellect.[33] At the halfway point, how realistic does this projection seem? It is now clear that great progress has been made in understanding how physicians solve difficult clinical problems and in implementing experimental programs that capture at least a portion of human expertise. On the other hand, it has become increasingly apparent that major intellectual and technical problems must be solved before we can produce truly reliable consulting programs. Nevertheless, assuming continued research, it still seems possible that by the year 2000 a range of programs will be available that can greatly assist the physician. It seems highly unlikely that such a goal will be achieved much before that time.

Tufts University School of Medicine
Boston, MA 02111 — *WILLIAM B. SCHWARTZ, M.D.*

Massachusetts Institute of Technology
Cambridge, MA 02139 — RAMESH S. PATIL, PH.D.
PETER SZOLOVITS, PH.D.

Supported by grants from the Commonwealth Fund and the Robert Wood Johnson Foundation (Dr. Schwartz) and by grants from the Naijonal insijiutes of Health: R24-RR-01320 from the Division of Research Resources, RO1-RL-33041 from the National Heart, Lung, and Blood Inslilute, and RO1-LM-04493 from the National Library of Medicine (Drs. Patil and Szolovits). The views expressed are those of the authors and do not necessarily represent the views of any of the granting agencies.

References

1. Reggia JA, Tuhrim S, eds. Computer-assisted medical decision making. New York: Springer-Verlag, 1985.
2. Kassirer JP, Gorty GA. Clinical problem solving: a behavioral analysis. Ann Intern Med 1978; 89:245-55.
3. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge, Mass.: Harvard University Press, 1978.
4. Swanson DB, Feliovich PJ, Johnson PE. Psychological analysis of physician expertise: implications for design of decision support systems. In:Shires DB, Wolf H, eds. MEDINFO 77: Proceedings of the Second World Conference on Medical Informatics, Aug. 8-12, 1977. Amsterdam: North-Holland, 1977:161-4.
5. Kuipers BJ, Kassirer JP. Causal reasoning in medicine: analysis of a protocol. Cognitive Sci 1984; 8:363-85.
6. Szolovits P, ed. Artificial intelligence in medicine. Boulder. Colo.: Westview Press, 1982.
7. Clancey WJ, Shortliffe EH, eds. Readings in medical artificial intelligence:the first decade. Reading, Mass.: Addison-Wesley, 1984.
8. Shortliffe EH. Computer-based medical consultations: MYCIN. New York: Elsevier, 1976.
9. Buchanan BG, Shortliffe EH, eds. Rule-based expert systems: the MYCIN experiments of the Stanford heuristic programming project. Reading, Mass.: Addison-Wesley, 1984.
10. van Melle WJ. System aids in constructing consultation programs. Ann Arbor, Mich.: UMI Research Press, 1981.
11. Davis R. Expert systems: where are we? And where do we go from here? Al Mag 1982; 3(2):3-22.
12. Clancey WJ, Leisinger R. NEOMYCIN: reconfiguring a rule-based expert system for application to teaching. In: Proceedings of the 7th international Joint Conference on Artificial intelligence, Aug. 24-28, 1981. Los Altos, Calif.: Morgan Kaufmann, 1981:829-36.
13. Kunz JC, Fallat RJ, McClung DR, eta). A physiological rule-based system for interpreting pulmonary function test results. Proc Comp Crit Care Pulmonary Med 1979:375-9.
14. Weiss SM, Kulikowski CA, Galen RS. Developing microprocessor based expert models for instrument interpretation. In: Proceedings of the 7th international Joint Conference on Artificial Intelligence, Aug. 24-28, 1981. Los Altos, Calif.: Morgan Kaufmann. 1981:853-5.
15. Szolovits P, Pauker SG. Categorical and probabilistic reasoning in medical diagnosis. Artif Intell 1978; 11:115-44.
16. Pauker SG, Gorry GA, Kassirer JP, Schwartz WB. Towards the simulation of clinical cognition: taking

a present illness by computer. Am J Med 1976; 60:981-96.

17. Miller RA, Pople HE Jr, Myers JD. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982; 307:468-76.
18. Barneit GO. The computer and clinical judgment. N EngI 3 Med 1982; 307:493-4.
19. Pople HE Jr. The formation of composite hypotheses in diagnostic problem solving: an exercise in synthetic reasoning. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence, Aug. 22-25, 1977. Los Altos. Calif.: Morgan Kaufmann, 1977:1030-7.
20. Paul RS, Szolovits P, Schwartz WB. Causal understanding of patient illness tn medical diagnosis. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, Aug. 24-28, 1981. Los Altos, Calif.: Morgan Kaufrusen, 1981:893-9.
21. Pople HE Jr. Heuristic methods for imposing structure on ill-structured problems: the structuring of- medical diagnostics. In: Seolovits P, ed. Artificial intelligence in medicine. Boulder, Colo.: Westview Press, 1982:119-90.
22. Paul RS. Causal representation of patient illness for electrolyte and acid-base diagnosis (Technical rep. TR-267). Cambridge, Mass.: Laboratory for Computer Science, Massachusetts Institute of Technology, October 1981.
23. Swartout WR. XPLAIN: a system for creating and explaining expert consulting programs. Artif Intell 1983; 21:285-325.
24. Clancey WJ. The epistemology of a rule-based expert system: a framework for explanation. Artif Intell 1983; 20:215-51.
25. Teach RL. Shortliffe EH. An analysis of physicians' attitudes. In: Buchanan BG, Shortliffe EH, eds. Rule-based expert systems: the MYCIN experiments of the Stanford heuristic programming project. Reading, Mass.: Addison-Wesley, 1984:635-52.
26. Chandrasekaran B, Mittal S. Conceptual representation of medical knowledge for diagnosis by computer: MDX and related systems. In: Yovits MC, ed. Advances in computers. vol.22. New York: Academic Press, 1983:217-93.
27. Clancey WJ. The advantages of abstract control knowledge in expert system design. In: Proceedings of the 2nd National Conference on Artificial Intelligence, Aug. 22-26, 1983. Los Altos, Calif.: Morgan Kaufmann, 1983:74-8.
28. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N EngI J Med 1975; 293:229-34.
29. *Idem*. The threshold approach to clinical decision making. N Engl J Med 1980; 302:1109-17.
30. Miller RA, McNeil MA, Challinor SM, Masari FE Jr, Myers JD. The INTERNIST-I/QUICK MEDICAL REFERENCE project -- status report. West J Med 1986; 145:816-22.
31. Miller PL. Expert critiquing systems: practice-based medical consultation by computer. New York: Springer-Verlag, 1986.
32. Shortliffe EH, Scott AC, Bischoff MB, Campbell AB, van Melle W, Jacobs CD. ONCOCIN: an expert system for oncology protocol management. In: Proceedings of the 7th international Joint Conference on Artificial Intelligence, Aug. 24-28, 1981. Los Altos, Calif.: Morgan Kaufmann, 1981: 876-81.
33. Schwartz WB. Medicine and the computer: the promise and problems of change. N EngI J Med 1970; 283:1257-64.