AI Magazine Volume 4 Number 2 (1983) (© AAAI)

ON EVALUATING AI SYSTEMS FOR MEDICAL DIAGNOSIS

B. Chandrasekaran

Artificial Intelligence Group
Department of Computer & Information Science
The Ohio State University, Columbus, Ohio 43210

Abstract

Among the difficulties in evaluating AI-type medical diagnosis systems are: the intermediate conclusions of the AI system need to be looked at in addition to the "final" answer; the "superhuman human" fallacy must be resisted; both pro- and anti-computer biases during evaluation must be guarded against; and methods for estimating how the approach will scale upwards to larger domains are needed We propose a type of Turing test for the evaluation problem, designed to provide some protection against the problems listed above We propose to measure both the accuracy of diagnosis and the structure of reasoning, the latter with a view to gauging how well the system will scale up

TESTING SYSTEMS WITH MULTIPLE DIMENSIONS to performance and possibly designed to meet multiple objectives, especially involving many subjective components, is inherently fraught with pitfalls. Evaluation of AI systems for medical decision-making (AIM systems) is no exception. The problem of evaluation of such systems is very important, however, since these sophisticated approaches should eventually result in systems with real benefits in clinical practice. Their widespread introduction is not possible until reliable measures of the cost/benefit ratios of their use are available.

A staple of many of the evaluations of AI systems that have so far been conducted (Colby, Hilf, Weber, & Kraemer, 1972; Yu et al, 1979) is a central idea from a well-known proposal to evaluate AI systems: The Turing Test (Turing, 1963) The meat of the idea is to see if a neutral observer, given a set of performances on a task, some by a machine and others by humans, but unlabelled as to authorship, could identify, better than chance, which were machine and which were human-produced. Note that this really attempts to answer the question, "Do we know how to design a machine to perform a task which until now required human intelligence?", but not the question, "Is the cost of introduction of a given machine for a task acceptable in comparison to the benefits?" The latter question subsumes the former in a sense, because the machine not performing well in comparison to a human would presumably increase the cost significantly.

In this paper I follow tradition and consider the evaluation of AI systems for medical diagnosis from the viewpoint of the first question above. In particular, I would like to outline some of the difficulties inherent in the evaluation of AIM systems, and to offer an evaluation procedure to be considered as a proposal and discussed by the community of AIM researchers.

The proposed procedure is also a variant of Turing's Test. I am aware that the procedure does not fully avoid all

© 1983 IEEE Reprinted, with permission, from MEDCOMP 1982 Proceedings, Proceedings First IEEE Computer Society International Conference on Medical Computer Science/Computational Medicine, September 23-25,1982, Philadelphia, PA

The preparation of this article was supported by NSF Grant MCS-8103480 I thank Ted Shortliffe for reading a draft of the paper and making a number of useful suggestions

the difficulties that I will be outlining, but it does respond to most of them. It is expensive in terms of physician time that is needed, but again, attempts have been made to reduce this component.

The problem of evaluation of performance is not unique to AIM systems of course; in fact, within AI, the issue is currently of considerable relevance to the broader area of expert systems. The discussion in (Gaschnig, et al., 1983) on evaluation of expert systems has many points of contact with what we say in the next section. Information about the evaluation status of some of the better-known expert systems is given in (Duda, Shortliffe, 1983). While the AIM system evaluation problem shares many aspects with the more general expert system evaluation problem, the proposal in this article is meant in particular for AIM systems.

Problems in Evaluation of Medical Diagnosis Systems

Success/Failure Dichotomy Insufficient. When evaluating performance of complex systems, especially at a development stage, simple "success" vs. "failure" evaluations based on the final answer may be insufficient because they do not take into account the possibility of very acceptable intermediate performance. In order to avoid this, a sophisticated evaluation technique needs to be developed for comparing complex symbol structures representing the important steps in reasoning. Such a technique is particularly needed if the objective is to evaluate the promise of an approach, rather than the performance of a deliverable system. As pointed out by Yu *et al* (1979), "A complex reasoning program must be judged by the accuracy of its intermediate conclusion as well as its final decision."

Superhuman Human Fallacy. When performance of computer-based consultation or diagnosis systems is evaluated against an absolute standard of correctness, such as confirmed autopsy findings, the ability of the computer system to agree with clinical judgments may be underestimated A comparative evaluation involving the computer system performance and that of a group of clinicians may be more revealing of the real ability of the system.

Anti– or Pro–computer Bias. The evaluators of computer performance may be biased either in favor or against the computer. A "blind" procedure in which the evaluating clinician is given both the computer performance and that of other clinicians, coded in such a way that it is not possible to guess the identity, would eliminate this danger. But this coding itself may introduce biases or errors, and careful procedures for this coding need to be developed.

"Correct" Answer May Be Unknown. Often, there are no "correct" answers, since expert clinicians may disagree among themselves. Techniques by which the computer is given exactly as much benefit of the doubt as the human experts during disagreement are necessary.

Scaling Upwards. Ability to scale upward to larger domains of medicine may not be evident from many evaluations. For instance, a program dealing with a relatively small knowledge-base may be based on an approach that is not easily extensible, but may still be tuned to yield great accuracy in performance over its small space of possibilities. Evaluations over a sequence of knowledge-bases, each a superset of its predecessor, can yield information about ease of expansion and consistency of performance across knowledge–bases. However, a careful examination of the structure of intermediate reasoning in the small domain may still yield clues about ease of expansion.

Small Sample Size Problems. Performance of the system in "rare" diseases or unusual situations cannot often be reliably evaluated due to the generally small sample size of the available cases involving these situations.

Matching Distribution of Clinical Practice. Without some knowledge about the distribution of types of cases that a system will need to confront, the results of evaluation cannot be suitably interpreted. For instance, suppose the computer system is very efficient in solving most of the "common" occurrences of diseases in an area of clinical medicine, and relatively poor in solving rare or "difficult" cases. If the difficult cases were to be chosen because they are "interesting" as test cases, the statistical evaluation of the system might not represent its performance in a real clinical setting. A solution to this situation is to require that cases be selected as representative of the target clinical setting. Where "interesting" cases are chosen for other reasons, the statistical evaluation will need to be suitably reinterpreted.

Objectives of Testing and Evaluation

Some of the purposes of testing and evaluation are:

a Guaranteeing satisfactory performance to somebody outside the development group—the user, the sponsor of the development effort, etc

b. Locating weaknesses in the system so further development can be made, e.g., to see whether the knowledge-base is adequately rich, whether problem solving is sufficiently powerful, etc As soon as errors are detected, further theoretical or system development will be called for.

c. Evaluating different functions of the same system For instance, a complex medical consultation system may be evaluated as a diagnostician, as an explanation facility, as an intelligent fact-retrieval facility, etc

d. Evaluating different dimensions of the same function, e g., correctness of diagnosis, habitability of the system in the diagnostic mode, response time, etc., are different dimensions of the diagnostic function

Experience in Evaluation of Other AI Medical Systems

Several AIM systems have been subjected to evaluation procedures of varying degrees of rigor and completeness. An

evaluation of MYCIN is reported in (Yu, et al., 1979), and (Miller, et al., 1982) describe an evaluation of INTERNIST-1. CASNET, a consultation program for glaucoma, has been evaluated in a formal way (Kulikowski, Weiss, 1982), as have the DIGITALIS ADVISOR program (Long, 1980) and PUFF, a program for interpretation of pulmonary function tests (Aikins et al., in press). MDX, a program developed in our research group, has undergone a preliminary evaluation (Tatman, Smith, 1982) Most of the above evaluations used randomized trials, case studies or a combination thereof.

In addition to these programs, diagnostic ECG programs that were developed in the 1960's have been the subjects of intense evaluation (Computers in Cardiology, 1976). Experience in evaluating ECG programs is of mixed relevance to the complex consultation programs of the type we are discussing. The separation of designer-clinical experts from the evaluating clinical experts, and coding to avoid bias associated with knowing the identity are two examples of useful ideas that are applicable to our current purpose. On the other hand, because of the limited knowledge–base, and the more statistical/numerical nature of the computations in ECG programs, the difficulties associated with evaluating symbolic processes are not present in their evaluation. Our own proposed evaluation experiences of MYCIN and CASNET designers, but go beyond them in some crucial aspects.

PROPOSED EVALUATION PROCEDURE

Focus of Evaluation. Our major focus in the evaluation procedure will be on the diagnostic performance of an AI system. The two major components that will be considered are:

a. the accuracy of diagnosis, and

b. the structure of reasoning.

In b, we shall attempt to evaluate the efficiency with which conclusions are reached, i.e., examine whether many irrelevant hypotheses are pursued, or focused, purposeful reasoning is displayed. The importance of testing for this component is to give evidence for how the approach will scale upward to larger domains (see Scaling Upwards). In a relatively small knowledge-base, blind or unintelligent search may result in the correct answer, but that approach will not work in larger domains. Examining the structure of analysis even in the small domain will give information about the "intelligence" of problem solving displayed by the system.

Steps in Evaluation. The proposed evaluation procedure consists of the following steps.

S1. In order to avoid constant tinkering with the system after each failure, versions of the system will be fixed The initial versions may be abandoned after a series of failures, but later ones will be stabilized, and used continuously for evaluation.

S2. A group of experts in the relevant clinical domain will be assembled. These will be experts who have not participated in the development of the system in any manner. This group will be subdivided into two groups, G1 and G2. G1 is a small group of experts who will generate the cases (step S4), and play the neutral role of coding the responses to suppress the identities of the diagnosticians (Steps S6 and S7) Experts in G2 will both generate their own diagnoses, and grade human vs machine performance (Step S8)

S3. A clear specification of the scope of the system will be written. While our earlier attempts at specification will most likely be inadequate, after a few iterations we should have a reasonably precise specification of what kinds of diseases and data the system is meant to deal with This specification will be the basis for deciding whether a failure of the system should be included in the tabulations If the case description and the actual diagnosis fit the specification, then the failure is a mark against that version of the diagnostic system If, on the other hand, the specification establishes that the case does not fit it, then the failure is not recorded Without such a clear specification, there will always be a temptation to explain away failures.

S4 Experts from G1 will be asked to compile two sets of cases: one set, a random selection of cases from their practice satisfying the specification; the second, a smaller set of "interesting" cases, either from journals or their own practice, to probe possible weaknesses in the knowledge–base or problem solving. The improvement for the next version

S5 These cases, as far as possible, will be complete cases, i e , in addition to all the clinical data (lab, symptoms, etc), confirmed diagnoses (autopsy, morphological confirmations, etc) will also be available Partial information cases may be included during the later phases of the evaluation, if time permits The confirmed diagnoses will be used during the "grading" phase for comparative evaluation of human diagnosticians and the machine. (See our earlier remarks regarding the superhuman human fallacy. To circumvent the pitfall discussed in that paragraph, the confirmed diagnosis is the yardstick against which both the human and the machine will be compared).

S6 Each of the case descriptions, excluding, of course, the confirmed diagnoses portion, will be distributed to *one* of the experts in group G2 For purposes of discussion, let us say case C_i is given to expert E_j from G2 For each case, the expert's "thinking-aloud" protocol will be recorded. These protocols will be used by experts in G1 (preferably working together as subgroups to reduce the subjectivity of the process) to produce two performance data structures: α_i and β_i. α_i displays, for case C_i, the final diagnosis and supporting positive and negative evidence as given by expert E_j. β_i, on the other hand, is a representation of E_j's diagnostic activity coded as a sequence of disease hypotheses considered, along with the data used by E_j for accepting or rejecting intermediate hypotheses. β_i will be a tree-like structure

representing the space over which the search ranged during problem solving

S7. The same case will then be run on the system From the output of the system, experts in G1 will again produce the two data structures mentioned in S6: α_i^*, β_i^* Here the asterisks are a notational device to refer to the AI system; the contents of the data structures themselves do not have any information regarding this identity.

S8 For each case, we now have performance data structures from an expert and the machine, but coded so as to suppress the identities. This is now distributed to the other experts, along with the original case description including the information about confirmed diagnoses. More precisely, if case C_i had gone to expert E_j in stage S6, all experts in G2 except E_j will receive the case and the performance data structures for this case. (They will not be told which data structures represent human performance, which the machine's, and they will not know which expert E_j was assigned to that case. In fact, there is even no need for them to know that *one* of the structures represents the human's and *the other* the machine's performance)

The "grading" procedure will be in two substages. The experts will first be given only the data structures α_i and α_i^*, which represent final diagnoses Each expert will be asked to grade the performance (on some suitable normalized scale). They will be asked to give partial credit for partial solution (e.g., deciding extra- vs. intra-hepatic cholestasis will be given partial credit, even if the particular cause of cholestasis, say, stone, has not been pinpointed). In the second substage, they will be given the data structures β_i and β_i^*, which stand for the reasoning efficiencies Grading similar to the first substage will be requested.

The reason for comparing α's first is that the accuracy of diagnosis should be judged unbiased by the perceived search efficiencies indicated in the β's If the size of G2 is n2, we will have for each case, $(n2-1)$ comparative evaluations of machine vs. human expert performance for two components of performance: diagnostic accuracy and search efficiency.

S9. Tables can now be prepared which display machine vs. human performance as average scores for sets of cases.

In the above procedure, there are several details that can only be decided upon during the conduct of the evaluation, since they depend upon the availability of sufficient numbers of experts to take part in the study In particular:

1 The size of group G1 need not be large: one might be sufficient, but we would prefer at least two Because of the subjectivity of the phase of translation into data structures, it would be useful if they can work together

2 The role played by experts in G1 can go either way. That is, instead of coding human and machine performance into machine-like data structures, they might instead translate the machine output into a coherent natural language narrative. The effects on the evaluation would remain the same, since the evaluators are still blind with respect to the identity of the diagnostician.

3. The number of evaluation experts in G2 and the number of test cases can be adjusted to produce a total number of evaluations sufficient for significant conclusions to be drawn.

Finally, a few remarks on the subjective aspects of the evaluation would be useful. Subjectivity is localized in two processes: the role played by experts in G1 in coding the human and machine response into structures α and β; and the comparative evaluation of these structures by experts in G2. The latter subjective process is, in our view, a strength of the procedure; as long as identities are suppressed, peer judgements are one of the best means of evaluation. The subjectivity implicit in the role of G1 is unavoidable, until AI systems with smooth natural language performance and general world knowledge are produced. Until then, there will always be extra medical competence components in narratives and protocols to give away the human vs. machine identities. However, more complex variants are possible in order to distribute the role played by members of G1 as "man-machine identity suppressors " Experts in G2 can rotate this role with those in G1, producing a more randomized design with respect to bias in the production of translations. Such variations will mute the subjectivity of this phase, and thus will be responsive to concern expressed for bias.

References

Colby, K.M., Hilf, F.D., Weber, S , & Kraemer, H. C 1972 Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes *Artificial Intelligence,* 3:199–222

Computers in Cardiology. 1976. Washington University, St. Louis, Missouri In particular, Session 3, Diagnostic ECG Systems has many papers on their evaluation.

Duda, R. O , & Shortliffe, E. H. (1983) Expert systems research *Science* (to appear)

Gaschnig J , Klahr, P., Pople, H., Shortliffe, E , & Terry, A. (1983) Evaluation of expert systems: issues and case studies In Waterman, D. A., Hayes-Roth, R , and Lenat, D , (Ed.) *Building Expert Systems*, Addison Wesley (forthcoming)

Kulikowski, C. A and Weiss, S. M. (1982) Representation of expert knowledge for consultation: The CASNET and EXPERT projects. In Szolovits, P. (Ed), *Artificial Intelligence in Medicine,* Westview Press, pp 21-55.

Long, W (1980) Criteria for computer-generated therapy advice in a clinical domain In Ripley, K. L and Ostrow, H. G (Ed) *Computers in Cardiology,* IEEE Catalog No 80-CH1606-3, IEEE Computer Society.

Miller, R., Pople, H , and Myers, J (1982) Internist-I, an experimental computer-based diagnostic consultant for general internal medicine, *New England Journal of Medicine,* 307:468-494

(Continued on page 48)

CMSL programmer, all data appears a graph of named vertices connected by edges. In reality, the graph is represented as structures of connection machine cells. Some CMSL operations correspond almost daily to a single connection machine instruction. Others involve complex patterns of messages passed between cells.

Another approach to creating parallel systems is that of Professor Carl E. Hewitt and his associates. They have developed actor theory, a rigorous abstract theory of parallel systems, that may provide a foundation for the construction and analysis of highly parallel problem-solving systems. Several important components have been developed including the Act 1 subsystem to model specialized communicating agents, the Omega subsystem for parallel semantic taxonomic relational networks, implemented by Dr. Gerald R. Barber and Mr. Eugene Ciccarelli, and the Sprites subsystem to model the communications activities involved in processing general goals and assertions, implemented by Dr. William A. Kornfeld.

Working with Professor Hewitt, Mr. Henry Lieberman has completed a preliminary version of the Tinker System for concrete programming Tinker enables a user at a work station to develop general procedures by abstracting from specific instances of steps performed on concrete examples at the work station. Mr. Lieberman and Professor Hewitt also have developed a real-time garbage collection algorithm based on the life time of objects.

Finally, Professor Sussman and his associates have completed the design of the SCHEME-81 chip, a VLSI device for running SCHEME, a dialect of the LISP programming language. In addition to testing the limits of our automated design aides, the SCHEME-81 chip may be a step toward better LISP-oriented personal computers: preliminary estimates are that SCHEME-81 will interpret simple LISP prorams about five times faster than our best current hardware. For large, complex programs, it will do even better.

A special-purpose silicon compiler, by Mr. Philip E. Agre, has been important to the development of SCHEME-81. Given a small program definition, Mr. Agre's compiler produces code specifying the layout of a SCHEME-81 bus-compatible chip to implement that function. The compiler uses traditional techniques and some novel heuristic methods to reason about the tradeoffs involved in writing highly parallel microcode.

Also, Mr. Jonathan D. Taft has been working on building a small SCHEME computer for testing our chips. It uses a Motorola 68000 design module as a front-end processor for performing I/O, for user-level arithmetic, for character manipulation, and for console control and debugging of the SCHEME system

The Computing Environment

The Laboratory's computing resources were improved by the installation of a large 20/60 system, a VAX 11/780, and a VAX 11/750, all manufactured by the Digital Equipment Corporation. The machines complement a variety of existing machines, including nearly two dozen LISP Machines, designed and built by the MIT Artificial Intelligence Laboratory.

All of the machines, together with terminal concentrators, are linked together with an eight-megabit packet-oriented cable system known as the CHAOSNET. The cable system can support as many as 100 communicating computers before reaching intolerable performance deterioration.

(Continued from page 37)

Tatman, J. L. & Smith, J. W , Jr. (1982) Preliminary evaluation of MDX, a medical diagnosis system *Proc 6th Annu. Symp Comput. Appl Med. Care*, IEEE Computer Society.

Turing, A.M 1963 Computing Machinery and Intelligence. In E. A. Feigenbaum and J. Feldman (Eds.) *Computers and Thought* New York: McGraw-Hill

Yu, V.L., Buchanan, B G., Shortliffe, E H., Wraith, S.M., Davis, R , Scott, A C., & Cohen, S.N. 1979 Evaluating the Performance of a Computer-based Consultant. *Computer Programs in Biomedicine,* 9:95–102.

APEX
Applied Expert Systems, Inc.

Providing knowledge-based expert systems to the financial services community

Dr. Fred L. Luconi
President

Dr. Norton R. Greenfeld
Director — AI Technology

Five Cambridge Center
Cambridge, MA 02142
(617) 492-7322