Report 83-01
Stanford -- KSL

Scientific DataLink

Hypothesis Generation in Medical Consultation Systems: Artificial Intelligence Approaches.
Edward H. Shortliffe,
Aug 1983

card 1 of 1

HPP Memo 83-01

HYPOTHESIS GENERATION IN MEDICAL CONSULTATION SYSTEMS: ARTIFICIAL INTELLIGENCE APPROACHES

Edward H. Shortliffe

Heuristic Programming Project
Departments of Medicine and Computer Science
Room TC-117, Stanford University School of Medicine
Stanford, California 94305

To appear in MEDINFO 83
(J.H. van Bemmel, M. Ball, and O. Wigertz, eds.)
North Holland, Amsterdam, 1983

(Presented at MEDINFO 83, Amsterdam, August 1983)

Keywords: AI, diagnosis, therapy, psychology

Dr. Shortliffe is recipient of Research Career Development Award LM-00048 from the National Library of Medicine.

Abstract

It has been argued that the problem of medical diagnosis is fundamentally ill-structured. This paper discusses the process of clinical hypothesis evocation and briefly surveys artificial intelligence (AI) methods that may help make computer-based solutions to the problem more tenable. Some example systems are briefly discussed to illustrate the key concepts.

1 INTRODUCTION

Medical diagnosis is inherently a classification task, one that involves sorting out the most likely hypothesis or hypotheses from what is often a wide range of possibilities. As is reflected in the title of this workshop[1], the generation of such hypotheses typically depends upon indirect observations -- the symptoms, signs, patient history, and test results that characterize the patient under consideration. Although medical computer scientists have approached the problem using a variety of techniques [15], the most common methods have been statistical, notably Bayes Theorem. Practical application of statistical methods requires that the range of possible diagnoses be defined in advance. In addition, assumptions are generally made about the mutual exclusivity of the competing hypotheses and the conditional independence of the disease manifestations that are observed.

A statistical technique such as Bayes Theorem does not actually "generate" hypotheses. High level hypotheses about the range of possible diagnoses are evoked by the physician based on the patient encounter before he or she decides to use the diagnostic program in the first place. Thus, for example, a Bayesian program designed to help a physician select among 8 or 10 causes of jaundice will naturally be useless if the patient does not have jaundice -- that initial "hypothesis" must be correct. Once the patient has been accurately classified at the general level, however, the program may accurately weight the competing diagnostic options and be helpful in defining the most likely cause of the abnormality.

[1] "The Formulation of Medical Hypotheses from Indirect Clinical Observations"

Blois has argued [2] that computer-based medical decision aids are inherently limited in their ability to assist the physician in reaching decisions about undifferentiated patients -- ones for whom initial screening and high-level generalized classification have not already been done. The argument is made that the diagnostic problem may be too ill-structured at the time of initial case presentation for a useful program to be developed to deal with the wide range of diagnostic options and potentially important disease manifestations that could be observed.

Certainly it is true that traditional statistical approaches cannot deal well with ill-structured tasks of this kind. The assumptions of Bayesian statistics, particularly mutual exclusivity, are simply invalid for patients with combinations of complicated disease processes, and access to the necessary conditional probabilities and interdependencies is totally unrealistic if one considers, for example, a Bayesian program to handle all diseases in internal medicine.

It must be acknowledged, however, that expert clinicians <u>do</u> somehow manage to meet a new patient, ask a few questions, generate diagnostic hypotheses worthy of further consideration, and, often within a few minutes, generate both a set of diagnostic options and a plan of action (diagnostic or therapeutic). This fascinating and elusive process is increasingly a subject of study. If we understand better how expert clinicians generate and refine hypotheses, it will help us to better teach the process to medical students and house officers [10]. It will also help provide design criteria for computer programs that can help in the diagnosis and management of undifferentiated patient cases.

This paper briefly describes some of the techniques developed in recent years for handling ill-structured diagnostic problems. The methods are drawn from the field of artificial intelligence (AI), the computer science speciality in which the focus is on the manipulation of symbolically represented concepts rather than on numerical calculations. The survey is necessarily superficial and incomplete, but it and the accompanying reference list may help the interested reader who wishes to learn more about these symbolic techniques for generating hypotheses from indirect observations.

2 REPRESENTATION OF INFERENCE CRITERIA

A variety of representation techniques have been used for encoding the kinds of inference steps used by a clinician in reaching a conclusion about diagnosis or therapy. For the purposes of the discussion here, we will call such inference steps <u>rules</u>, but it should be emphasized that several medical AI systems have encoded inferential knowledge using approaches other than formal production rules [5]. A recent book on artificial intelligence in medicine [17] provides an excellent introduction to the range of representation techniques that have been used.

Rules typically provide a statement of how strongly a given inference can be drawn when specific indirect observations are made. In a Bayesian system, each conditional probability P(D|E) (where <u>D</u> is the disease under consideration and <u>E</u> is the evidence that has been observed) can be viewed as a specialized kind of rule.

When problems are ill-structured, however, the conclusion and premise portions of a pertinent rule may not correspond cleanly to a disease and one of its manifestations[2]. We have found it useful to classify rules into one of several categories[3]: (a) definitional rules (e.g., "if the patient is male, then he is not pregnant or lactating" or "if the patient is immunosuppressed, then he is a compromised host"); (b) cause-to-effect rules (e.g., "if there is an elevation in parathyroid hormone level and the patient has normal renal function, then anticipate that urinary calcium is depressed and phosphate is increased"); (c) effect-to-cause rules (e.g., "if the patient has recurrent calcium oxalate kidney stones, then consider the diagnosis of hyperparathyroidism"); and (d) associational rules (e.g., "if the patient has gram negative sepsis and has been seriously burned, there is evidence that the offending organism may be pseudomonas aeruginosa").

Consider, for example, the problem of antimicrobial selection which is the task of the MYCIN consultation system [14]. Most of the knowledge in that system is encoded in production rules, but only a small subset of these corresponds to the specific diagnostic problem which lies at the heart of the system (i.e., the identification of the most likely bacteria causing disease, particularly those that need to be considered in arriving at an optimal therapy recommendation). Because therapy decisions motivate all questioning, it is not enough to infer the single most likely bacterial pathogen affecting the patient. Additional issues affect this ill-structured task; for example:

[2]Pople has discussed the nature of ill-structured problems in some detail in a chapter in the Szolovits book [13].

[3]See [18] for a more detailed discussion of these and related rule classes.

(a) determining how sick the patient is, (b) identifying immunological factors that may affect the diagnosis, the therapy, and the decision whether to treat at all, (c) assessing renal and hepatic status in case they are relevant in selecting drugs or adjusting doses, and (d) inquiring about recent or concurrent therapy with antibiotics that could affect test interpretation or drug selection.

When knowledge regarding these diverse tasks and problems is represented in inference statements such as rules, how can a computer program deal with the complex interrelationships that exist? How can pertinent hypotheses about the patient be generated and, in turn, used to guide the reasoning and data gathering process that ensues? A variety of control techniques have been developed by AI researchers, some of which are summarized in the following sections. Each has certain advantages, and one of the ongoing challenges in the field is to improve our understanding of the techniques so that we can better define how to match a given problem with an optimal representation and control strategy [6].

3 GOAL-DIRECTED REASONING

We will begin by considering the MYCIN system mentioned above. Although it deals with an ill-structured task, the range of possible diagnostic hypotheses with which it must deal is relatively constrained. The narrowness of its problem domain thus makes it similar to the kinds of tasks that Blois has argued are more amenable to computer-based approaches. In addition, because the program assumes that the user has already determined that the patient may have an infection and may require antimicrobial therapy, hypothesis generation is not a major element in the system. The program knows it has certain goals for the consultation session, and these can be used to guide the data collection and reasoning process that ensues.

MYCIN knows that its goal is to undertake a set of tasks: (a) to determine if there is a significant infection in the patient, (b) if so to determine the most likely identity of the infecting organism, (c) to consider the possibility that other organisms are present for which there is no direct (e.g., gram stain) evidence, and (d) to decide whether therapy is appropriate or whether further observation or data collection is warranted before therapy is begun. Only when these goals have been achieved does the system actually proceed to the generation of a therapy recommendation.

Reasoning of this kind is called "goal-oriented" because predefined goals provide the initial focus, and rules are invoked or questions asked in an effort to achieve a specific goal. Rules are selected from the 500 in the knowledge base in accordance with their ability to achieve a goal that has been established. MYCIN attempts to achieve any goal by applying all of the directly relevant rules. The need to establish the premises (i.e., conditional portions) of those rules sets up new subgoals that are treated in the same way. When the program eventually requests some factual information from the user, the rule that prompted the request may become applicable and, if so, is applied. The application of a rule enters a new fact into the database. This in turn is available when attempting to apply other rules.

Thus the line of questioning, the rules that are applied, and the conclusions that are reached are determined by the data obtained for the particular patient. Hypotheses are not "evoked" by this process, as they could be if users simply entered whatever data they felt were pertinent and the system then identified and invoked applicable rules ("data-driven" or "forward-chained" reasoning as described in the next section). However, a focussed interactive consultation session does result from the goal-directed strategy. Whenever a question is asked of the user, MYCIN already knows the reason that the information may be useful. This permits explanations of questions and may help educate the user. In addition, related questions are automatically grouped together by the control strategy, and this can be more pleasing to the physician than an interaction in which questions appear to be haphazardly selected.

It should be clear, however, that goal-directed reasoning of this kind is different from the style of problem solving used by an expert clinician who sees a patient for the first time and must ask questions before making any initial diagnostic hypotheses. The process of hypothesis evocation is inherently data-driven, and other medical AI programs have used that approach.

4 DATA-DRIVEN REASONING

A "data-driven" or "forward-chained" control mechanism is used in those settings where patient descriptors are entered without direct guidance by the computer program. As data are entered, the observations "evoke" certain inference rules which are tested and, if they succeed, add their conclusions to the database. These new inferred data can then be used to invoke additional rules that use the new conclusions in their premise conditions. No specific goal guides the reasoning process; the data that are entered "drive" the rules and, at completion, the system can make available whatever conclusions have been reached.

A pure Bayesian system that uses a fixed data set of patient observations can be viewed as a data-driven program, although the concept of forward chaining is of course meaningless in the statistical context. The similarity, rather, is that all relevant data are entered and then the computer program runs to completion, reaching what conclusions it can[4]. Data driven programs that use large rule sets in ill-structured domains typically <u>do</u> generate hypotheses from a large set of possible solutions. The hypotheses may be weighted relative to one another, or an optimal solution or course of action may be identified.

One setting in which the data-driven approach is particularly appropriate is for patient monitoring. Consider, for example, the Ventilator Manager program (VM) that was developed to assist physicians charged with weaning patients from ventilators after they had undergone open-heart surgery [8]. Here a rule-based program took as its inputs a large number of digitized physiological parameters that were acquired from the patient at regular

[4]There are variations on the Bayesian approach that permit sequential questioning and updating, but most programs have used a static data set. See [15] for further discussion and references on this point.

intervals. Patient data included monitored heart rate, mean arterial blood pressure, expired CO2 concentrations, respiratory rate, and information about ventilator settings and arterial blood gases that were occasionally entered by hand. As data became available that indicated that the patient satisfied the conditions in the premise of a rule, that rule would be invoked and executed, thereby updating the patient data base and establishing expectations about what should happen subsequently. An additional challenge in this system was the need to keep track of temporal trends and to use these in determining when it was appropriate to recommend that the physicians advance the patient to the next stage in the weaning process. In this system the hypotheses that were generated were not diagnostic categories (which were seldom at issue), but, for example, assessments of whether a patient's hemodynamic status was stable or expectations of what would happen to the patient's respiratory rate if the ventilator settings were adjusted downwards.

Another example of this type of system is our current work on a program called ONCOCIN [16]. This program is an experimental consultation system used by oncologists to assist in the treatment of patients enrolled in cancer chemotherapy protocols. The knowledge of the formal protocols, as well as the judgmental expertise of experts, is encoded in production rules. The system is also used for entering and maintaining the patient data base. Thus the physician uses the system to fill out flow sheet information on the patient (using a special high speed display which simulates the paper flow sheet previously filled out manually). As the data are entered, the reasoning component of the consultation system "monitors" the patient information and inference rules are invoked in accordance with the patient-specific information. Thus most of the reasoning is data-driven, and the system's advice is generated as a by-product of the data entry process. As with VM, however, the hypotheses generated are not diagnostic but deal with expectations and alternate therapeutic options.

5 HYPOTHESIS-DIRECTED REASONING

Studies in recent years have demonstrated that expert clinicians who are solving difficult problems tend to use a hypothetico-deductive strategy [7],[9],[10]. This term refers to the use of early hypotheses, generated from initial data, to guide further data collection as the hypotheses are refined. The refinement process includes discriminating between close competitors, pursuing highly likely but unproven possibilities, ruling out less likely competitors, and, occasionally, invoking new hypotheses as additional unexpected findings are obtained. It should be clear that neither the pure goal-directed nor the pure data-driven techniques capture this highly flexible approach to problem solving, and the developers of systems such as MYCIN, VM, and ONCOCIN make no claim to have based their approaches on formal psychological models of experts.

Systems have been developed, however, in which early diagnostic hypotheses have been used to guide the refinement process. An early example was the PIP program [12] in which the computer gathered information about the present illness and generated diagnostic hypotheses for patients with edema. In a non-medical application, a geologic system named PROSPECTOR uses a combination of data-directed and hypothesis-directed reasoning and has successfully predicted the location of a major ore deposit [3].

Perhaps the best known medical system of this class is the INTERNIST program [11], recently revised and reincarnated under the name CADUCEUS [13]. This system undertakes the ambitious task of diagnosing all major disorders in internal medicine, and its knowledge base currently contains over 600 disease entities. Its reasoning strategy is also designed to deal with multiple coexistent disease entities and to ignore "red herrings" if most of the evidence in a case strongly supports a given disorder. Most of the knowledge in this system is contained in data structures that encode the manifestations associated with diseases and two measures of the association: (1) a frequency weight (corresponding to sensitivity) which indicates the frequency with which the given manifestation is observed in patients with the disease, and (2) an evoking strength (corresponding to predictive value) which reflects the weight with which the manifestation, when observed, should "evoke" the disease hypothesis in question. These weights are scaled and combined in an ad hoc fashion, but the performance of the program has been impressive because of the richness of the knowledge encoded [11].

The physician begins using CADUCEUS by entering a list of pertinent patient descriptors (manifestations that are present or absent and that the physician feels ought to be mentioned). This initial data set "evokes" a set of disease hypotheses which are partitioned into subsets of competitors in accordance with a clever partitioning algorithm. The set of most highly supported hypotheses then becomes the focus of attention, and the program enters a questioning mode in which manifestations are requested in accordance with their ability to help sort out the best hypothesis among the competing set. Scores are recalculated as the physician enters these additional data, and the focus can shift as hypotheses are rejected or confirmed by the new information. Although CADUCEUS does not attempt to model formally the psychology of expert problem solving, it has clearly been influenced by observations about the hypothetico-deductive approach that is typically used by experts.

6 SUMMARY

We have described only a few of the techniques used by researchers who are building medical consultation systems based on representation and control methods drawn from the field of artificial intelligence. Detailed descriptions of this field have recently become available [1],[4],[17] for those readers who may wish more information about AI and its relevance to medical decision making research. Much of the field's appeal lies in its ability to allow the encoding and manipulation of the knowledge needed for dealing with the ill-structured problems of medicine -- those that defy formal analysis yet characterize the expertise of skilled clinicians.

References

1. Barr, A. and Feigenbaum, E.A. The Handbook of Artificial Intelligence, Vols. 1 and 2. Palo Alto, CA: Heuristech Press, 1981 (Vol 1) and 1982 (Vol 2).

2. Blois, M.S. Clinical judgment and computers. New Eng. J. Med. 303:192-197 (1980).

3. Campbell, A.N., Hollister, V.F., Duda, R.O., and Hart, P.E. Recognition of a hidden mineral deposit by an artificial intelligence program. Science 217:927-929 (1982).

4. Cohen, P.R. and Feigenbaum, E.A. The Handbook of Artificial Intelligence, Vol. 3. Palo Alto, CA: Heuristech Press, 1982.

5. Davis, R. and King, J.J. An overview of production systems. In Machine Intelligence 8: Machine Representations of Knowledge (E.W. Elcock and D. Michie, eds.), New York: John Wiley, 1977.

6. Duda, R.O. and Shortliffe, E.H. Expert systems research. Science 1983 (in press).

7. Elstein, A.S., Shulman, L.A., and Sprafka, S.A. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.

8. Fagan, L.M., Shortliffe, E.H., and Buchanan, B.G. Computer-based medical decision making: from MYCIN to VM. Automedica 3:97-106 (1980).

9. Johnson, P.E., Duran, A.S., Hassebrock. F., et al. Expertise and error in diagnostic reasoning. Cognitive Science 5:235-282 (1981).

10. Kassirer, J.P., Kuipers, B.J., and Gorry, G.A. Toward a theory of clinical expertise. Amer. J. Med. 73:251-259 (1982).

11. Miller, R.A., Pople, H.E., and Myers, J.D. INTERNIST-1: An experimental computer-based diagnostic consultant for general internal medicine. New Eng. J. Med. 307:468- (1982).

12. Pauker, S.G., Gorry, G.A., Kassirer, J.P., and Schwartz, W.B. Toward the simulation of clinical cognition: taking a present illness by computer. Amer. J. Med. 60:981-995 (1976).

13. Pople, H.E. Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnostics. In Artificial Intelligence in Medicine (P. Szolovits, ed.), Boulder, CO: Westview Press, Inc., 1982, pp. 119-190.

14. Shortliffe, E.H. Computer-Based Medical Consultations: MYCIN. New York: Elsevier/North Holland, 1976.

15. Shortliffe, E.H., Buchanan, B.G., and Feigenbaum, E.A. Knowledge engineering for medical decision making: a review of computer-based clinical decision aids. Proceedings of the IEEE 67:1207-1224 (1979).

16. Shortliffe, E.H., Scott, A.C., Bischoff, M.B., et al. ONCOCIN: An expert system for oncology protocol management. Proceedings of the 7th Intl. Joint Conf. on Artificial Intelligence, pp. 876-881, Vancouver, British Columbia, August 1981.

17. Szolovits, P., (ed). Artificial Intelligence in Medicine, Boulder, CO: Westview Press, Inc., 1982.

18. Wallis, J.W. and Shortliffe, E.H. Explanatory power for medical expert systems: studies in the representation of causal relationships for clinical consultations. Meth. Info. Med. 21:127-136 (1982).