How Can Computer Programs Reason?

Peter Szolovits, Ph.D.*
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts, 02139, USA

This paper was presented at CARDIOSTIM 88, Monaco.

June 1988

Abstract

Computer programs are increasingly being called on to suggest or to make decisions in medical applications. Traditional methods of decision making based on flowcharts and probabilistic classification have proven to be too cumbersome to apply to large domains. As a result, programs employing artificial intelligence methods were introduced in the 1970's. Deficiencies in those methods limited the capabilities of the resulting programs and suggested many research problems now being pursued in medical AI. Current research plans call for progress in research on representing the causal, pathophysiologic basis of disease, temporal evolution of cases, likelihood and utility, and the use of past experience in problem-solving. In addition, efforts to integrate progress in various areas and to demonstrate the applicability of the new methods are also key. A new generation of programs will result, with more sophisticated capabilities to bring to bear on the difficult problems of medical reasoning.

The computing elements of health care are becoming more and more significant, and increasing sophistication is demanded of the decision-making components of medical devices. Although this trend is generic to all of medicine, it is perhaps most advanced and clearly visible in the types of instruments that are the focus of the CARDIOSTIM meeting. This paper describes developments in the application of artificial intelligence (AI) methods to building programs that reason in a manner approximating the performance of human health practitioners. In addition, it points to some of the current outstanding research problems in medical AI and suggests some promising current research directions.

1 Background

Medical diagnosis has been one of those areas of computer application foreseen since the earliest days of practical computing, and indeed since the early 1960's there has been a literature describing a variety of approaches to how best to utilize the computer in medical reasoning. The motivations for such work are easily seen and compelling. Excellent health

*The work reported here has been supported (in part) by grants R01 LM 04493 from the National Library of Medicine and R01 HL 33041 from the National Heart, Lung and Blood Institute. Our research also benefits greatly from our long-term ongoing collaboration with the research group headed by Dr. Stephen G. Pauker at Tufts/New England Medical Center.

care is an important good for everyone, it is generally considered to be in short supply, there appear to be critical intellectual components to its performance, and some practitioners are acknowledged to be significantly better at it than others. These observations suggest that if we could capture the capabilities of the human expert in the form of a computer program, it would enable us to disseminate that expertise in readily-usable form and thus to improve the overall quality of health care. This basic idea has been proposed for application to a wide range of health-care settings, encompassing assistance for "barefoot doctors" (rural health-care workers in underdeveloped countries) to monitoring for occasional errors in practice at the most sophisticated tertiary-care medical centers. In addition, such computerized expertise is also seen as an important tool in medical education and training, and possibly as an aid in medical research.

Perhaps the simplest means of capturing medical reasoning in a computer, and the one first pursued by researchers, is just to try to write a program that reproduces the behavior of the practitioner. In principle, one could imagine being the proverbial "fly on the wall" in a doctor's office, watching the doctor interact with one patient after another. The first might enter complaining of long-lasting and severe headaches. We record the complaint, and then see what question the physician asks next. Once again, we record the question and its answer, and continue in this way until the practitioner reaches a conclusion.[1] Then if an identical patient were to appear, with the same presenting complaint and the same answers to questions, we might feel justified to reach the same conclusion as the doctor had earlier reached. Were we to record such investigative paths for many, many patients, we would have developed an elaborate *flow chart* describing the procedures that our exemplary physician follows.

The hypothetical approach described here is not really practical or desirable, at least in such a pure form. The most fundamental limitation is that it provides "answers" only for patients identical to ones it has previously encountered. Thus, any case presenting even the slightest degree of novelty would remain unclassified. This means that to build the program one would have to observe such a doctor/patient interaction for all possible patients of interest; if this were at all feasible, it would undoubtedly require the collaboration of vast numbers of physicians. The technique supposes, however, that the physician is completely consistent in approaching similar patients—an assumption particularly difficult to defend when numerous individual doctors are involved. In practice, of course, one can build some assumptions into the program—e.g., to ignore any data presented that it does not know how to interpret. This will allow it to deal somewhat with novelty, though at the risk of producing incorrect and unjustifiable results. The greatest problem with this approach, however, stems from a deeper source: it relies completely on a behaviorist description of the medical reasoning process, with no place for any of the conceptual structures that we commonly think underlie medical reasoning. As a result, its knowledge is acquired and stored piecemeal, with little chance that it will be consistent, and with corresponding difficulties in explaining it to users and in keeping it up to date and reliable.

The second major historical approach to medical reasoning rests on ideas of pattern matching and probabilistic classification. Though many related versions of this type of approach have been tried, the Bayesian inference method is most frequently in use today. In this approach, we assume that each patient belongs to one of a (typically small) number of categories (diseases), $D_1 \ldots D_n$, each describable by statistics describing a similar popula-

[1] We mean "question" generically, to include observations, tests, diagnostic procedures, etc., as well as literal questions.

tion. We then assume that among the total population of possible patients we can estimate the *a priori* likelihood, $P(D_i)$ of each category D_i. We further assume that there are available questions that help to (probabilistically) distinguish among the alternative categories. In particular, if S (a symptom) represents some particular answer to some question, then we must have available the *conditional probabilities*, $P(S|D_i)$ representing the fraction of patients with D_i who would exhibit S. By Bayes' rule,

$$P(D_i|S) = \frac{P(D_i)P(S|D_i)}{P(S)},$$

we can compute how likely each disease is after we have observed S from knowing the prior probability distribution, the conditional probabilities, and the prior likelihood of the observed symptom (which is also computable from the other known information by

$$P(S) = \sum_i P(D_i)P(S|D_i).$$

Assuming that the order of observations is immaterial (not always realistic, as some tests may preclude the use of or corrupt the results of others), this method may be used sequentially, to further update the distribution of probabilities as more and more evidence becomes available. Thus, a developing sense of the likelihood of each diagnosis may be maintained as new data are observed.

In the late 1960's, Gorry [9] observed that this method could also be extended to help suggest which question is most worth asking at each point in the process. The basic idea is that the question most worth asking is the one that will most dramatically reduce the remaining uncertainty among the diagnostic alternatives. Information theory defines this uncertainty as the *entropy*, E, of a probability distribution, which is $\sum_i P(D_i) \log P(D_i)$. Note that it has the desired character that $E = 0$ if the answer is certain (for some k, $P(D_k) = 1$ and $P(D_i) = 0$ for all $i \neq k$), and it reaches its maximum value when each of the D_i is equally likely. Thus the best question is that one which yield the lowest expected entropy.

If the available questions are $Q_1 \ldots Q_m$ and the possible answers for question Q_j are $S_{j,1} \ldots S_{j,l}$, then we may compute the expected entropy of the probability distribution resulting from asking question Q_j by

$$E(Q_j) = \sum_i P(S_{j,i}) \sum_k P(D_k|S_{j,i}) \log P(D_k|S_{j,i}),$$

where $P(S_{j,i})$ may be computed as suggested above.

In practice, the total of all possible observations is usually overkill to make a decision. The entropy heuristic suggests the most informative subset of observations sufficient to converge to an acceptably-decisive probability distribution. In one experiment, for example, about seven or eight (on average) of a possible thirty questions were sufficient to reach a diagnostic conclusion [10].

If the questions to be asked have significant associated risks, pain, or money costs (referred to in the aggregate as *costs*), then maximizing information with each question may be less desirable than obtaining the most available information for the least cost. For example, though an invasive surgical procedure leading to biopsy may be the most informative test for some condition, we may prefer to use a somewhat less conclusive combination of laboratory tests involving more easily-obtained measures, such as blood and urine samples,

radiologic tests, etc. This approach has been formalized in the discipline of *decision theory*, and a significant literature of its use in medicine has also developed[44].

Probabilistic computer diagnostic methods based on Bayes' rule have been successfully demonstrated over the past decade in real clinical settings, most effectively in the work of de Dombal and his collaborators in England. Their experiences, however, introduce a cautionary note. In the limited domain of diagnosing the acute abdomen, their first program relied on numerical estimates of prior and conditional probabilities given by human experts. In a study of this program, it was determined that although the program did reasonably well, the human experts were still more successful[5]. A later version of the program, based on actual clinically-gathered probabilities rather than human estimates, was in fact better at its task than the original experts[4]. This is an impressive demonstration of the value of such a program, and a real landmark in the clinical application of computers. A subsequent study[6], however, showed that the success of the program was very much dependent on having acquired accurate probabilities for its local population of patients. Moving it to another hospital again impaired its performance, demonstrating that the method is highly sensitive to the accuracy of its probabilities for the population being examined. Perhaps this sensitivity should not be surprising, because as in the case of flowcharts, such a program really incorporates no knowledge of medicine as we ordinarily conceive it—only large tables of associations. Thus, it has no basis for estimating which of its associations must be universal to the human population due to some underlying facts of physiology or pathology and which are expressions of more local conditions. The implications of this lack of knowledge for judging the adequacy of such a program or for maintaining it in the long term are also disquieting.

The greatest problem with the probabilistic methods appears to be their voracious appetite for data. Even if we assume (as above) that observations never interfere with each other and that a patient must have one and only one disease, and then we further assume that observations are statistically independent (i.e., that the likelihood of a particular observation depends only on the likelihoods of the possible diseases, not specifically on what other symptoms we have seen), the number of conditional probabilities needed by a program is still large. For one of Gorry's examples[10] containing 14 possible diseases, 31 possible questions, and (on average) three possible answers to each question, over 1300 conditional probabilities must be estimated or measured. The de Dombal studies suggest dangers in estimating the numbers, and the size of database needed suggests that many, many patient cases must be collected in a careful, consistent manner to yield an effective program for a particular setting. If we were to relax any of our assumptions, the corresponding need for additional data is devastating. For example, if we consider the possibility of multiple diseases co-occurring, or of symptoms influencing each other, the amount of required data becomes exponential.[2] Such difficulties limit the application of the probabilistic methods to rather narrow and well-constrained domains.

2 The Early AI Systems

In the early 1970's, several research groups became frustrated with the limitations of the flowchart and probabilistic methods we have outlined[8]. Each group noted that despite the

[2]One promising escape from this problem lies in better means to represent only those interactions among diseases or symptoms that truly matter. Recent work on probabilistic networks suggests that the explosion of required data can be contained somewhat[34, 35].

obvious difficulties of the formal methods we had been exploring, human experts seemed to be able to do medical reasoning quite effectively. Further, they could often easily change their approaches as new medical knowledge, tests and procedures became available. In addition, they appeared able to deal with problem domains much broader than those handled by the existing computer methods, and they could successfully deal with patients having complicated conditions in which several disorders were simultaneously present.

Using the then recently developed methods of protocol analysis and also applying a large dose of our own computational and psychological intuitions, we developed a number of techniques to represent fragments of medical knowledge in the computer and to combine those fragments in ways approximating human reasoning. The principal approaches and results of this early work are surveyed elsewhere[3, 41, 42], so the present discussion will simply illustrate a few of the interesting ideas in these programs.

Our own work was very much influenced by Minsky's theory of memory organization[29], which has come to be called the *frame* theory. It suggests that concepts in memory are directly linked to other concepts that play a role in the first, including mutual constraints that express what linkages "make sense", that computation may be initiated by the local propagation of the consequences of determining a new fact or association, and that concepts are also interconnected by a structure of similarities and significant differences, permitting the efficient exploration of a large space of possible alternatives. One program designed to take the present illness of a patient presenting with symptoms of renal disease (PIP)[33, 42] embodies these intuitions by organizing its knowledge around a set of frames, each of which represents a disease or a clinical or pathophysiologic state or syndrome. Each frame identifies those observations that are highly suggestive of its disorder and those to be strongly expected if the disorder is present. In addition, the frames are linked together via paths expressing the possible causal sequence of events that may lead from one condition to another, and by differential suggestions, indicating which disorders might often be confused with each other and how best to resolve those ambiguities. The program exploits the structure of the memory by causing evocative observations to activate their related hypotheses; it then follows causal and differential links to explore the extent of the patient's disease and the alternatives most important to exclude. The program reaches its conclusions based either on logical criteria that define certain conditions or on an approximate probabilistic scheme. It is the numerous types of links in the program that provide additional structure over the conditional probability tables in a typical Bayesian program. Their exploitation allows the program to behave reasonably, even though its database of probabilistic associations is far from complete. Furthermore, one can recognize in the behavior of the program "lines of investigation" that appear human-like and therefore explainable to and comprehensible by the intended user.

The INTERNIST-I program used a PIP-like frame structure, but with a more complete set of numerical measures related approximately to probabilistic links. This program relied on an interesting insight that if a patient indeed had two or more disorders simultaneously, the corresponding hypotheses should explain more of the actually-observed data than any hypothesis alone. Using this idea, it was the first program that demonstrated a reasonable approach to the diagnosis of simultaneous multiple disorders[28].

The CASNET/Glaucoma program[45] explicitly modeled the likelihood of one state of ocular disease causing another. It could thus use the predicted likelihoods of unobserved states to suggests which tests were likely to be useful. It could also recognize situations where its hypotheses were internally inconsistent; for example if condition A causes B, which in turn causes C, and no other causes are possible, then if we observe A and C but

not B, there must be some flaw in the model or the observations. A much more elaborate extension of these ideas has later been applied to the analysis of heart wall motion and ECG analysis by Tsotsos and his colleagues[43].

Our program for advising on the administration of digitalis[11] introduced an explicit notion of consultation as an ongoing process rather than a one-shot interaction. To support this, it included a model of the patient and the program's concerns about the patient, similar in spirit to Weed's problem-oriented medical record. Using this, the program could then focus on achieving an adequate combination of therapeutic effectiveness without too much expression of toxicity.

Probably the best-known of these early programs was MYCIN, providing consultation for the diagnosis and treatment of bacteremia[38, 2]. Its underlying structure was motivated by the observation that doctors appear to know many small chunks of specific knowledge and that they can flexibly chain these together to solve problems. Mycin's chunks consist of rules of the form "if A and B and C are all true, then conclude D." Correspondingly, MYCIN's control structure identifies a current goal—a fact whose truth or falsity is to be established (D above)—and searches among its set of rules for ones that might make a conclusion bearing on the current fact. If a rule might be applicable, each of its premises is in turn made the current goal and the whole process recurs. Ultimately, some fact may have no relevant rules for determining it; in that case, the program simply asks the user. The generality of this control structure assures that all rules that might be relevant are in fact tried. As a result, the program can be incrementally extended, and a new rule is automatically applied everywhere where it might be relevant. It is, therefore, as if the program were constructing a large flowchart, but doing it on the fly. This therefore addresses the most serious problem we identified with flowcharts earlier.

In retrospective analyses of these early programs, it in fact appears that they are often comparable to their flowchart or probabilistic forbears, but that additional data structures and more special-purpose computational methods are used to increase the flexibility of the initial approaches without requiring huge additional amounts of data. In addition, the richer data structures and strategies allow the programs to contain a much more explicit representation of the data and assumptions on which their behavior is based. Thus, the programs can explain to their users just how they have reached a particular conclusion in terms that are meaningful within their domain, not simply based on a "black box" calculation. Maintenance and upgrading of such systems should, as a result, also be correspondingly easier.

The techniques developed in these early systems have, in fact, enjoyed enormous practical success, as much of the commercial "expert systems" technology of the mid-1980's can directly trace its roots to these programs of the mid-1970's. Unfortunately, however, that success has been almost completely outside the medical arena. Thus, programs employing the techniques developed for medicine now assist with insurance claims adjustment, credit approvals, factory scheduling, capital financial analysis, process control, etc., but only a very small handful of AI programs have any clinical use in medicine. Why is this?

The limited use of medical AI programs might be attributed to many alternative explanations, among which are concerns about safety and liability, lack of uniform practice at different institutions, unavailability of much relevant clinical data in computer-processable form, the high cost of new technology, and a general conservatism among physicians. In the rest of this paper, we set aside these contextual concerns, to focus on the technical limitations of the early programs and promising research efforts now underway to overcome these limitations.

3 Current Technical Needs and Directions

A number of the early medical AI programs were shown, in careful studies done in the late 1970's, to exhibit a level of performance that was essentially comparable to the performance of expert human clinical consultants[49, 28]. Nevertheless, even the most enthusiastic researchers who built these systems continued to feel that there was something missing in the programs, that they were not ready for the challenge of actual clinical use. Indeed, many felt that a qualitative improvement was needed. This malaise has a number of technical roots.

- Despite the methodological advances in making more of the knowledge in systems explicit, much of the strategic knowledge embedded in programs remained implicit—thus hard to explain and nearly impossible to change.
- The programs really had no adequate theory of how disease manifestations overlap and interact in the common case of multiple co-occurring disorders. This meant that the programs were most at risk of failing exactly on those very complex cases where their advice would be most likely to be sought.
- Although medical reasoning is often deeply concerned with the timing and sequence of events, the programs lacked any reasonable models for representing temporal information, either for patient-specific observations or for describing the time-course of disease.
- The programs' knowledge was mostly based on associations between diseases and their observable consequences, and lacked any explicit pathophysiologic model in terms of which those consequences could be grouped and explained. An adequate medical explanation, however, often demands that the associations which may have suggested the right diagnosis then be backed up by a consistent account of how the patient's condition could have arisen from the suspected etiologies.
- Though almost all medical reasoning involves a great deal of uncertainty, the probabilistic models in these programs were usually rather *ad hoc* and poorly understood.
- New systems were enormously difficult and time-consuming to produce, because each system needed to be "hand-crafted."

Since the beginning of the 1980's, researchers have shared a set of intuitions about what technical advances are needed to make significant progress. Foremost among these is the belief that for decision-making in the most complex cases, "deep" theories of disease and treatment will be needed. As a result, not only must research be pursued on the range of topics suggested by the above listing of technical deficiencies, but it is critical that progress on all these topics be successfully integrated into coherent systems. Only by simultaneously exploiting such advances could we achieve the qualitative advance thought necessary. One further desideratum also became clear: ideally, simple cases should be solved simply, and the power of complex techniques should be brought to bear only on complex problems.

The next few pages present an overview of the research efforts currently underway in our group, highlighting some of the progress toward the above objectives. The work can be usefully organized into three categories:

1. fundamentals—developing the new methods needed to support complex reasoning about deep representations of medicine,

2. integration—finding means of bringing together progress in the fundamentals into a coherent system, and

3. applications—demonstrating the validity of the concepts and tools developed by showing their utility in some particular medical domain.

3.1 Fundamentals

Throughout the development of medical AI programs, we have always found it very helpful to study in some depth the clinical behavior of human physicians. Our original ideas about the nature of the diagnostic process were based on such investigations, and we continue to examine the problem-solving exhibited by experts and by less senior medical practitioners in order to understand what knowledge and strategies are brought to bear on various types of problems and how expertise develops in the human as he or she becomes more experienced. Recently, the focus of these studies has been on reasoning about uncertainty and risk[31], causality[22] and the creation and evaluation of plans[20].

The need for deep reasoning in complex cases suggests that models of human pathophysiology and the underlying models of physiology (and perhaps eventually biochemistry and genetics) become an important component of the knowledge of an expert medical system. Unfortunately, traditional models that have been built as the basis for simulation studies of the human organism are too large, turgid and complex for use in human-like reasoning. For example, Guyton's model of the human cardiovascular system[12] may seem like an appropriate basis for a deep analysis of complex heart problems in a patient. However, its hundreds of parameters and interrelationships make it impossible to tune the model to the specific case at hand and its complexity precludes a thorough exploration of all possibly-relevant configurations of the model for diagnostic or therapeutic purposes.[3]

[[Figure missing in this version of the paper.]]

Figure 1: Parameters and their relationships in the Heart Failure Program

It is possible instead to build a much less complex model that contains only states and parameters of clinical interest, as Long has done in the Heart Failure Program[25]. Figure 3.1 shows the parameters taken into consideration by this program, and the various influences among them. Some of these parameters are directly observable (e.g., `HEART RATE`), whereas others can only be estimated by use of the model (e.g., the degree of sympathetic stimulation, `SYMP STIM`). Some parameters represent controllable actions available to the physician (e.g., the amount of beta blockers present, `BETA BLOCK`), others correspond to disease conditions (e.g., `AORTIC STENOSIS`), while others change only as the body responds to disease and therapy. In such a model the specific relationships among adjacent nodes of the graph are

[3]It should be clear that this is not a direct criticism of that model, which was indeed constructed for a completely different purpose.

typically approximations of the physiological relationships that might be derived from a more complete differential equation model. In fact, because the complete model is often unavailable, the relationships will be based on empirical observations. In addition to the parameters shown here, the program includes a set of disease states and criteria that relate the states to the parameters.

A model such as this may be used for both diagnostic and therapeutic reasoning. In diagnosis, observable parameters are entered, the model predicts the consistent values of internal parameters, and the states consistent with all parameters are then identified. In therapy planning, therapeutic goals are specified, interventions are reflected in the controllable parameters of the model, and then consequences are checked both for conformance to the goals and for the possible occurrence of undesirable side-effects.

The heart failure model has been under development for several years, and we plan shortly to make it available for use to house staff and students in a cardiac intensive care unit.

Even in those circumstances where the precise differential equations controlling some physiologic process are known, it is typically impossible to solve the complex sets of equations analytically. Therefore, people have resorted to simulation methods to gain some insight into the behavior of such complex systems. We have been interested in extending analytical techniques to reason about the approximate behavior of such systems symbolically, where one can obtain directly the form of behavior of the system, not just its behavior under one particular set of numerical conditions. Earlier work in AI aimed at modeling the physical world has had some success abstracting from numeric data to a qualitative algebra of just *plus*, *zero* and *minus* values[1], but for reasoning with feedback systems such as those typical in medicine this approach loses too much information and predicts too many possible behaviors[21]. Sacks has recently shown that the application of classical engineering approximation methods can be largely automated to analyze the behavior of simple systems of nonlinear differential equations. His program[37] finds appropriate piecewise linear approximations, solves them analytically, and then assembles an overall approximate behavioral description by constraining the solutions to match at their common boundaries. These methods have been applied to equations describing relatively simple systems such as non-linear oscillators, and we hope to extend them to more realistically complicated medical examples.

As we have described above, probabilistic reasoning in medical programs has been problematic because the straightforward application of classical techniques appears to require too much data to be practical for large domains, whereas the invention of pseudo-probabilistic methods leaves doubts about their validity and robustness. We have sought to develop means of representing and reasoning with statements about absolute and relative likelihood that preserve the semantics of classical probability theory but permit a lower degree of precision than explicit numbers. Wellman has recently succeeded in defining a formalism of qualitative probabilistic networks[46], which are an abstraction of the Bayesian networks of Pearl. The abstraction is that links in Wellman's networks indicate the direction, but not the magnitude of influence of a probabilistic relationship. Thus, if the representation states that "A positively influences B," we know that raising the likelihood of A will cause an increase in the likelihood of B, all other things being equal; we do not know, however, the degree to which B's likelihood will change.[4] Wellman observes that this system answers

[4]An additional relationship, called *synergy*, is also necessary to model and reason about the joint effects of multiple influences[48]. Thus, if A and C both positively influence B, it is often necessary to be able to

some of the charges of lack of robustness of probabilistic representations. Although particular numerical relations may change from one population to another (as de Dombal found, for example), the direction of those relations is most likely determined by some universal underlying reality that will be consistent across populations. Clearly, one cannot use such a qualitative representation to make close tradeoff decisions where the specific numbers actually matter. Its utility is greatest in computer programs that need to exclude many nonsensical possible diagnoses or plans before commencing a detailed investigation of the remaining ones. It is being applied in the domain of diagnostic and therapeutic planning[48].

When one builds a complex program to deal with difficult cases, it is typical for it to handle even simple cases or ones whose solution it should already simply remember in the same complex manner as it deals with the most complex novel cases. This presents a serious practical problem, because if a program is to be used it must respond in a timely manner, and it is also disturbing to the goal that programs should behave somewhat like the expert human physicians that they are modeled after. We know that a human consultant often recognizes a case as being essentially like another whose solution is known; it would be highly desirable for a program to do so as well. Koton has made a major step in this direction recently by building a memory-based program, CASEY, that solves problems by retrieving similar candidate cases and solutions and then adapting them to the particular facts at hand[19, 18]. CASEY works with Long's Heart Failure program and can fall back to using it when no relevant cases can be retrieved and adapted. Unlike some earlier AI programs based on analogy, CASEY's adaptation of a previous solution rests not simply on gross similarity of features but on verifying that the causal explanations for the original case are in fact applicable to the new one (or that they can be altered to become applicable). This method is surprisingly effective in reaching the right solution for cases that have dissimilar but reasonably-close precedents, and at times it requires the examination of almost 100-times fewer nodes in the causal network than the original program. In principle, also, as the program's database of past cases and solutions grows, it may be able to solve new problems for which the causal models in the original program provide no basis for solution.

We have remarked earlier that most programs to date have had inadequate or nonexistent models of time and temporal progression in their representations. Remedying this is a difficult open problem, and our approach had taken many different routes. Perhaps the simplest is to add to all facts about the patient and to all stages in a disease model an explicit timestamp, reflecting either the actual calendar date of an observation or the relative time (during the course of a disease) when an event or condition can be expected. Such a scheme must be augmented with some notion of temporal spread or uncertainty[16], and careful attention must be paid to reducing the exponential amount of calculation possibly required to keep straight the timing relationships in such a system[17]. Nevertheless, even a simple scheme will help avoid silly errors that can arise by ignoring time altogether[40, 17]. Similarly, representing processes with associated temporal dependencies on other processes can support the reconstruction of a plausible history showing how some complex state of disorder could have come about[24]. Focusing on process also suggests an appropriate control structure for ongoing decision-making, such as arises in a patient-monitoring setting. Here, new information about past states of the world (e.g., the correction of an earlier error in the data, or the arrival of a measurement from the lab hours after the sample was taken) must be incorporated without forcing a recalculation of every conclusion ever made by the

say, for example, that an increased likelihood of A makes the influence of C on B stronger than otherwise. This is called (positive) synergy. Without this added capability, the representation would give ambiguous results when several influences might interact.

system. Russ has developed a temporal control structure that successfully abstracts out the time and update-dependent component of reasoning from the component that draws conclusions based on assumed current data[26]. This model has been applied to aspects of the problem of ventricular arrhythmia management[27, 32] and is now being used in a program for the management of ketoacidosis.

3.2 Integration

The earlier discussion makes clear that there is much room for technical improvement in the methods used in medical expert systems, and that there are many interesting technical advances that are in fact exploiting that room. How is it possible, however, to take advantage of such advances to build new programs? This is in fact a very serious problem. It is still the case that new systems are hand-crafted for each new application, and as the complexity of the techniques used by the systems grows, that hand-crafting becomes more and more difficult and time-consuming. We are at the stage where it is not unusual to see five years elapse between the formulation of a new technical idea and its incorporation even in a very limited experimental program. To build a program that at the same time takes advantage of several newly-developed methods might take much longer. For example, although many clinical decision problems involve sophisticated reasoning about both likelihood and time, there is no obvious simple way to combine the recent advances in these two areas to yield a program with strong capabilities in both. Nevertheless, for the field to make progress toward practically-buildable programs, we must be able to pursue research in combining such representations[39].

The ability to integrate aspects of new representations and reasoning methods in a common system is, then, crucial. The approach we have taken toward this problem is a linguistic one. We assume that it will require many experiments to figure out how best (and most efficiently) to combine probabilistic and temporal reasoning (to stay with that example). To perform those experiments, however, we must be able to express the representation and the computations of both within a single language. Our work thus far has concentrated on the use of the NIKL language[30, 15], which provides a clean definitional component and the possibility of connecting assertions to it. Unfortunately, our preliminary impression is that it is very difficult to represent the full complexity and richness of medicine within a language that is designed to have a very precise and simple semantics supported by very efficient deductive algorithms[13].

Similar linguistic approaches are also being pursued by researchers engaged in the Universal Medical Language System project[7] and by others whose goal is to create encyclopedic knowledge within the computer[23].

3.3 Applications to Decision Analysis

In describing our work on fundamentals, we have already pointed to several programs that are intended to test our new ideas and to demonstrate their feasibility in realistic medical settings. In addition to those mentioned above, we have also had a longstanding interest in combining the methods of decision analysis and artificial intelligence in common programs. Whenever an AI program is called-upon to make a decision, either internally to its operation or as its final conclusion, it is reasonable to apply decision analytic reasoning, trading off the expected costs and benefits of each alternative decision. The work described above for planning, for example, follows just such an approach. The transfer of ideas can flow in the

opposite direction as well, however, bringing new power to decision analysis by applying AI methods to the construction and critiquing of decision trees.

When an analyst is faced with a blank sheet of paper and a difficult medical decision, he or she typically draws on a body of knowledge to help determine what factors in the decision should be explicitly represented and what possible actions should be explicitly considered. These choices come well before the detailed search for objective or subjective probability and utility estimates that are considered the hallmark of decision analytic problem-solving. Knowledge-based models of typical decision settings and of the factors typically influencing the availability and choice of possible actions can help constrain the alternatives that should be considered. In fact, one can build AI-based programs to help formulate decision trees[14].

Just as the formulation of a new tree is a knowledge-intensive task, so is the critiquing and refinement of an existing tree to become a better basis for making a decision. In the decision making consult service at Tufts/New England Medical Center[36], a fellowship training program has given us the opportunity to observe fellows formulate decision analyses, both at the beginning of their fellowship and near its end, when they are much more experienced. In addition, we have seen the active give-and-take that occurs among the fellows and the senior staff in the regular weekly review of the decision-making consults. It is apparent from these observations that the skill of good decision analysis is not easy, is greater in more experienced practitioners, and can be taught by critiquing analyses in progress. Responding to this opportunity, we have undertaken studies of the types of errors that arise in decision analyses and have begun to build a computer program, BUNYAN, that critiques partially-specified decision trees[47]. Its criticisms are based on structural features of the decision tree, not the particular numbers used in the analyses. Thus, just like the qualitative probabilistic models described above, the criticisms from the program should be robust to variations in the particular probability and utility estimates used. BUNYAN critiques not the particular decision reached or the numbers it is based on but the form of the tree considered. A simple example of one of its critiquing rules is:

- Any test performed must result in some difference of subsequent actions depending on its outcome. Otherwise, a similar decision tree without the test would be strictly better, by avoiding the cost of the test.

A more complex rule says approximately the following:

- In a fair analysis, the possible outcomes along different branches of a decision must be explored to the same depth and breadth.

We believe that this critiquing program will be of real use to less-experienced decision analysts, and that its construction will also teach us more about how good decision analysis is done. In addition, it provides an excellent testbed for our developing ideas on how to represent partial and abstract knowledge of likelihoods and utilities.

4 Conclusions

In this paper we have surveyed the background of medical decision-making methodology from which medical AI research arose, we then described some of the fundamental ideas introduced by early AI programs, and we pointed out a number of deficiencies that appeared to stand in the way of building more capable programs. We then summarized a group of investigations being carried out by our research group. These include mainly specific

research efforts aimed at overcoming the noted deficiencies, but also include significant efforts to integrate the results of these individual projects and to demonstrate the utility of these ideas in applications. The applications focus particularly on reasoning about patients with complex conditions related to heart failure and on the aiding the clinical applications of decision analysis.

Will the advances anticipated from the current round of research efforts bring us above the threshold of practical utility for very complex clinical programs? It is difficult to tell without of course actually doing to research and evaluating its outcome. It is clear from earlier advances that techniques now in existence are already sufficiently good to enable us (with great effort) to build successful programs for narrow domains of application. As the methods currently being developed become available for practical use, they will certainly make it possible to broaden the domains of applicability of the clinical programs. Ultimately, it will be a combination of circumstances, including the contextual questions raised above, that will determine the impact of AI technology on clinical practice.

References

[1] Daniel G. Bobrow, editor. *Qualitative Reasoning about Physical Systems*. MIT Press, 1985. Reprinted from *Artificial Intelligence*, vol. 24, 1984.

[2] Bruce G. Buchanan and Edward H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, Massachusetts, 1984.

[3] William J. Clancey and Edward H. Shortliffe, editors. *Readings in Medical Artificial Intelligence: The First Decade*. Addison Wesley, Reading, Mass., 1984.

[4] F. T. de Dombal, D. J. Leaper, Jane C. Horrocks, John R. Staniland, et al. Human and computer-aided diagnosis of abdominal pain: Further report with emphasis on performance of clinicians. *British Medical Journal*, 1:376–380, 1974.

[5] F. T. de Dombal, D. J. Leaper, J. R. Staniland, A. P. McCann, and Jane C. Horrocks. Computer-aided diagnosis of abdominal pain. *British Medical Journal*, 2:9–13, 1972.

[6] F. T. de Dombal, J. R. Staniland, and Susan E. Clamp. Geographical variation in disease presentation. *Medical Decision Making*, 1:59–69, 1981.

[7] David A. Evans and Randolph A. Miller. Final task report (task 2)—unified medical language system (UMLS) project: Initial phase in developing representations for mapping medical knowledge: INTERNIST-I/QMR, HELP, and MeSH. Technical Report CMU-LCL-87-1, Laboratory for Computational Linguistics, Carnegie Mellon University, Pittsburgh, PA, 1987.

[8] G. A. Gorry. Computer-assisted clinical decision making. *Methods of Information in Medicine*, 12:45–51, 1973.

[9] G. A. Gorry and G. O. Barnett. Sequential diagnosis by computer. *Journal of the American Medical Association*, 205(12):849–854, 1968.

[10] G. Anthony Gorry, Jerome P. Kassirer, Alvin Essig, and William B. Schwartz. Decision analysis as the basis for computer-aided management of acute renal failure. *American Journal of Medicine*, 55:473–484, 1973.

[11] G. Anthony Gorry, Howard Silverman, and Stephen G. Pauker. Capturing clinical expertise: A computer program that considers clinical responses to digitalis. *American Journal of Medicine*, 64:452–460, March 1978.

[12] A. Guyton, C. Jones, and T. Coleman. *Circulatory Physiology: Cardiac Output and its Regulation*. W.B. Saunders Company, Philadelphia, 1973.

[13] Ira J. Haimowitz, Ramesh S. Patil, and Peter Szolovits. Representing medical knowledge in a terminological language is difficult. In *Symposium on Computer Applications in Medical Care*, pages 101–105, 1988.

[14] J. P. Hollenberg. The decision tree builder: An expert system to simulate medical prognosis and management. *Medical Decision Making*, 4(4), 1984. Abstract from the Sixth Annual Meeting of the Society for Medical Decision Making.

[15] Thomas S. Kaczmarek, Raymond Bates, and Gabriel Robins. Recent developments in NIKL. In *Proceedings of the National Conference on Artificial Intelligence*, pages 978–985. American Association for Artificial Intelligence, 1986.

[16] Kenneth Kahn and G. Anthony Gorry. Mechanizing temporal knowledge. *Artificial Intelligence*, 9(1):87–108, 1975.

[17] Isaac S. Kohane. Temporal reasoning in medical expert systems. In R. Salamon, B. Blum, and M. Jørgensen, editors, *MEDINFO 86: Proceedings of the Fifth Conference on Medical Informatics*, pages 170–174, Washington, October 1986. North-Holland.

[18] Phyllis A. Koton. Reasoning about evidence in causal explanations. In *Proceedings of the National Conference on Artificial Intelligence*. American Association for Artificial Intelligence, 1988.

[19] Phyllis A. Koton. *Using Experience in Learning and Problem Solving*. PhD thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, Cambridge, Massachusetts, 02139, 1988. Also appeared as MIT LCS TR-441, March 1989.

[20] B. Kuipers, A. J. Moskowitz, and J. P. Kassirer. Decisions in medicine: Representation and structure. *Cognitive Science*, 12:177–210, 1988.

[21] Benjamin Kuipers. Commonsense reasoning about causality: Deriving behavior from structure. *Artificial Intelligence*, 24:169–204, 1984.

[22] Benjamin Kuipers and Jerome P. Kassirer. Causal reasoning in medicine: Analysis of a protocol. *Cognitive Science*, 8:363–385, 1984.

[23] Doug Lenat, Mayank Prakash, and Mary Shepherd. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine*, 6(4):65–85, 1986.

[24] William J. Long. Reasoning about state from causation and time in a medical domain. In *Proceedings of the National Conference on Artificial Intelligence*, pages 251–254, 1983.

[25] William J. Long, Shapur Naimi, M. G. Criscitiello, and Robert Jayes. The development and use of a causal model for reasoning about heart failure. In *Symposium on Computer Applications in Medical Care*, pages 30–36. IEEE, November 1987.

[26] William J. Long and Thomas A. Russ. A control structure for time dependent reasoning. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pages 230–232, 1983.

[27] William J. Long, Thomas A. Russ, and W. Buck Locke. Reasoning from multiple information sources in arrhythmia management. In *Proceedings of the Conference on Frontiers of Engineering in Health Care*, pages 640–643. IEEE, 1983.

[28] Randolph A. Miller, Harry E. Pople, Jr., and Jack D. Myers. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307:468–476, 1982.

[29] Marvin Minsky. A framework for representing knowledge. In Patrick Henry Winston, editor, *The Psychology of Computer Vision*, chapter 6, pages 211–277. McGraw-Hill, 1975.

[30] M. G. Moser. An overview of NIKL, the New Implementation of KL-ONE. Technical Report 5421, Bolt, Beranek and Newman, Inc., 1983.

[31] A. J. Moskowitz, B. Kuipers, and J. P. Kassirer. Dealing with uncertainty, risk and tradeoffs: A cognitive science approach. *Annals of Internal Medicine*, 108(3):435–449, 1988.

[32] Steven L. Novick. Reasoning over time about the causes of arrhythmias. Master's thesis, Massachusetts Institute of Technology, 1987.

[33] Stephen G. Pauker, G. Anthony Gorry, Jerome P. Kassirer, and William B. Schwartz. Towards the simulation of clinical cognition: Taking a present illness by computer. *American Journal of Medicine*, 60:981–996, 1976.

[34] Judea Pearl. How to do with probabilities what people say you can't. Technical Report CSD-850031, UCLA Computer Science Department, September 1985.

[35] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.

[36] D. A. Plante, Jerome P. Kassirer, D. A. Zarin, and Stephen G. Pauker. A clinical decision consultation service. *American Journal of Medicine*, 80:1169–1176, 1986.

[37] Elisha P. Sacks. *Automatic Qualitative Analysis of Ordinary Differential Equations Using Piecewise Linear Approximations*. PhD thesis, Massachusetts Institute of Technology, February 1988.

[38] Edward H. Shortliffe. MYCIN: *Computer-based Medical Consultations*. American Elsevier, New York, 1976.

[39] P. Szolovits, J. P. Kassirer, W. J. Long, A. J. Moskowitz, S. G. Pauker, R. S. Patil, and M. P. Wellman. An artificial intelligence approach to clinical decision making. TM 310, Massachusetts Institute of Technology, Laboratory for Computer Science, 545 Technology Square, Cambridge, MA, 02139, September 1986.

[40] P. Szolovits and S. G. Pauker. Research on a medical consultation system for taking the present illness. In *Proceedings of the Third Illinois Conference on Medical Information Systems*. University of Illinois at Chicago Circle, November 1976.

[41] Peter Szolovits, editor. *Artificial Intelligence in Medicine*, volume 51 of *AAAS Selected Symposium Series*. Westview Press, Boulder, Colorado, 1982.

[42] Peter Szolovits and Stephen G. Pauker. Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11:115–144, 1978.

[43] J. K. Tsotsos. Computer assessment of left ventricular wall motion: The ALVEN expert system. *Computers in Biomedical Research*, 18(3):254–277, June 1985.

[44] Milton C. Weinstein and Harvey V. Fineberg. *Clinical Decision Analysis*. W. B. Saunders Co., Philadelphia, 1980.

[45] Sholom M. Weiss, Casimir A. Kulikowski, Saul Amarel, and Aaron Safir. A model-based method for computer-aided medical decision making. *Artificial Intelligence*, 11:145–172, 1978.

[46] Michael P. Wellman. Probabilistic semantics for qualitative influences. In *Proceedings of the National Conference on Artificial Intelligence*, pages 660–664. American Association for Artificial Intelligence, 1987.

[47] Michael P. Wellman, Mark H. Eckman, Craig Fleming, Sharon L. Marshall, Frank A. Sonnenberg, and Stephen G. Pauker. Automated critiquing of medical decision trees. *Medical Decision Making*, 9(4):272–284, 1989.

[48] Michael Paul Wellman. *Formulation of Tradeoffs in Planning under Uncertainty*. PhD thesis, Massachusetts Institute of Technology, July 1988.

[49] V. L. Yu, B. G. Buchanan, E. H. Shortliffe, S. M. Wraith, R. Davis, A. C. Scott, and S. N. Cohen. An evaluation of the performance of a computer-based consultant. *Computer Programs in Biomedicine*, 9:95–102, 1979.