Despite their early successes, Diagnoser and Deducer did not have a clear, comprehensible structure that is required for the kind of experiments we wish to perform. Galen was built to remedy this problem, taking advantage of the experience gained in the design of Diagnoser and Deducer. Additional discussion of the structure of GALEN can be found in prior annual reports and in the relevant publications.

To determine the generality of our model of expertise in diagnostic reasoning, we are also investigating domains outside medicine. As with our work in congenital heart disease, we have concentrated on the design of mechanisms for structuring problem specific knowledge and for focusing limited computational resources.

One of the Principal Investigators has published results of a study in Expertise in Trial Advocacy, discussing the significance of current research in expertise in legal problem-solving. [Johnson, Johnson, and Little, 1985] Research on legal expertise in corporate acquisition problems has also been investigated. The results of that research suggest that expert corporate acquisition attorneys differ from novices in their greater reliance on internalized norms, prototypes and heuristics. Both expert and novice attorneys in the study went beyond the information provided in task cues in interpreting and predicting actions and situation scripts in the simulated problems. The subjects reasoned heuristically as well as logically. Differences between attorneys in different specialty areas were not large suggesting that the subjects within a domain of problem solving such as legal reasoning acquire meta level reasoning skills that apply to issues within and outside their areas of specialization.

Research is also being completed in a study of cognitive strategies used in making strategic decisions in business. Corporate acquisitions were again used as the context in which to examine expertise. Twenty-four executive subjects were asked to perform an experimental task in which they evaluate companies as candidates for acquisition. The goals of the research are to test for the existence of specialty-related reasoning strategies and to determine the importance of strategic and financial information in problem formulation, problem structuring and choice of strategies in problem solving.

Research in Progress --

Since human experts are notoriously poor at describing their own knowledge, our work requires the creation of problem solving tasks through which experts can reveal criteria for initiating specific hypotheses and methods for investigating those hypotheses.

Current techniques of representing hypotheses and their expectations for diagnosis do not, however, provide much detailed information about the control processes experts use to guide their reasoning. Such control processes typically incorporate highly refined heuristics about which the experts are almost wholly unaware. New research is being proposed to investigate these control structures in legal reasoning, specifically in reasoning by analogy in appellate decision making. Reasoning by analogy appears to be an important inference tool used by experts in many domains as a fundamental problem solving tool. The ability to form plausible analogies lies at the heart of much of the expert ability to be generative when faced with unfamiliar problems. This research will include the implementation of a cognitive simulation of the reasoning by analogy process based upon data obtained by observation of experts solving problems. The results of the simulation will be validated by comparison with human subject data.

We are also investigating several research questions relevant to the architecture of Galen. We have designed an interface to Galen so that users who are unfamiliar with the inner workings of the program can interactively enter case data. Designing the interface raised questions about what forms of data are necessary to adequately and completely represent all possible cases.

One project to test the extensibility of GALEN into other domains is being conducted

by a graduate student in the Graduate School of Management. His thesis, Auditing Internal Controls: A computational model of the review process, includes the construction of a working expert system using GALEN. The objective of this study is to formulate and test a model of the processes employed by audit managers and partners in reviewing and evaluating internal accounting controls.

Another project explores the extension of the GALEN architecture into a problem in plant pathology. The main purpose of this research is to find out how the basic postulates about expert reasoning made in Galen hold in a second diagnostic domain. The problem domain chosen for this purpose is Plant Pathology. In collaboration with Professor Paul Teng of the Plant Pathology Department of the University of Minnesota a prototype knowledge base has been implemented. Currently, the knowledge base can diagnose ten potato diseases and has 124 rules. The system is going through evaluation and fine tuning to bring it up to an expert performance level. This system will be useful in the Extension Service at the Plant Pathology department at the University of Minnesota, which provides diagnostic information to farmers over the phone lines.

Dr. Spackman's thesis is entitled "Induction of classification rules under the guidance of comprehensibility-enhancing logical structures and diagnostic performance goals." The purpose of this research is to study and implement methodologies for the automated generation of comprehensible decision rules from empiric data, with emphasis upon logic-based knowledge representation formats and upon problems drawn from the domain of medicine. This work builds upon some of the machine learning methodologies developed at the University of Illinois by R. S. Michalski and others.

This work addresses two shortcomings of previous work on induction of classification rules. These are, first, lack of comprehensibility of the induced rules, and second, lack of flexibility in specifying the diagnostic performance (sensitivity, specificity, or efficiency) desired for the rules that are to be derived.

Comprehensibility of the derived rules or descriptions can be enhanced by imposing restrictions upon the format which the rules may take. For example, the restriction of rules to a unate boolean function format allows the induction of rules that can often be simplified to a "criteria table" type of representation. The type of diagnostic performance a rule must have will depend upon its purpose, and specifying the purpose may allow inductive inference algorithms to trade off small decrements in diagnostic performance for large increments in comprehensibility, or to increase their robustness in the face of noisy or uncertain data.

Successful development of these techniques will lead to enhanced capabilities for deriving rule bases for expert classification systems from empiric data, and will provide new methods for the conceptual analysis of data.

Preliminary results have been obtained for the problem of deriving rules for the identification of bacteria based upon their biochemical profiles in the medical microbiology lab. Other problem domains under investigation are the analysis and interpretation of endocrine laboratory tests, and the induction of rules for the diagnosis of congenital heart disease, for comparison with the rules used in GALEN.

Research is also under way in methods of automating knowledge acquisition in pediatric cardiology. This is being done as thesis research by Paul Krueger. The objective of the research is to design, implement, and test a computerized procedure to derive from examples a nonmonotonic set of rules for an expert classification system. Systems using such rules are generally more efficient than those using monotonic classification processes and more closely approximate psychological models as well.

The research proposes a process for automated learning of preliminary rulebases subject to a set of efficiency constraints which are consistent with a formally defined,

psychologically plausible model of classification. The constraints include an upper bound on the amount of information required to explain observations not accounted for by the current set of beliefs, and a lower bound on the degree of inconsistency allowed in the knowledge base at any given time. It will be shown that these constraints can be used to guide the automated determination of both the content and organization of the rules of expert classification systems. The result is behavior that is more focused and efficient, and more closely duplicates the lines of reasoning of domain experts.

A representational formalism for classification knowledge bases based upon a nonmonotonic logic of belief called "autoepistemic logic" (Moore, 1985) is proposed. Having thus defined a representation for the knowledge base the research will propose a methodology for instantiating its concepts within a given application domain. The general approach is to use heuristics to identify from a set of input examples various contextual situations that occur and the types of rules to associate with them. The rule acquisition module (RAM) is then tested in two different application domains. The resulting expert systems will be evaluated for correctness of classification and similarity of their lines of reasoning with those of human experts.

The major conclusion of the research is that constraints similar to those observed in expert human classification processes can be used to guide the empirical induction of efficient expert system rulebases. Supporting this conclusion is the elucidation of a formal nonmonotonic model of classification, and the design and subsequent testing of the Rule Acquisition Module and expert systems derived by it.

D. List of Relevant Publications

1. Connelly, D. and Johnson, P.E.: *Medical problem solving*. Human Pathology, 11(5):412-419, 1980.

2. Elstein, A., Gorry, A., Johnson, P. and Kassirer, J.: *Proposed Research Efforts*. IN D.C. Connelly, E. Benson and D. Burke (Eds.), CLINICAL DECISION MAKING AND LABORATORY USE. University of Minnesota Press, 1982, pp. 327-334.

3. Feltovich, P.J.: *Knowledge based components of expertise in medical diagnosis*. Learning Research and Development Center Technical Report PDS-2, University of Pittsburgh, September, 1981.

4. Feltovich, P.J., Johnson, P.E., Moller, J.H. and Swanson, D.B.: *The Role and Development of Medical Knowledge in Diagnostic Expertise*. IN W. Clancey and E.H. Shortliffe (Eds.), READINGS IN MEDICAL AI, Addison-Wesley, 1984, pp. 275-319.

5. Johnson, P.E.: *Problem Solving*. IN ENCYCLOPEDIA OF SCIENCE AND TECHNOLOGY, McGraw-Hill (in press).

6. Johnson, P.E., Moen, J.B., and Thompson, W.B.: *Garden Path Errors in Medical Diagnosis*. IN Bloc, L. and Coombs, M.J. (Eds.), COMPUTER EXPERT SYSTEMS, Springer-Verlag (in press).

7. Johnson, P.E.: *Cognitive Models of Medical Problem Solvers*. IN D.C. Connelly, E. Benson, D. Burke (Eds.), CLINICAL DECISION MAKING AND LABORATORY USE. University of Minnesota Press, 1982, pp. 39-51.

8. Johnson, P.E.: *What kind of expert should a system be?* J. Medicine and Philosophy, 8:77-97, 1983.

9. Johnson, P.E., *The Expert Mind: A new Challenge for the Information*

Scientist IN Th. M. A. Bemelmans (Ed.), INFORMATION SYSTEM DEVELOPMENT FOR ORGANIZATIONAL EFFECTIVENESS, Elsevier Science Publishers B. V. (North-Holland), 1984.

10. Johnson, P.E., Severance, D.G. and Feltovich, P.J.: *Design of decision support systems in medicine: Rationale and principles from the analysis of physician expertise.* Proc. Twelfth Hawaii International Conference on System Science, Western Periodicals Co. 3:105-118, 1979.

11. Johnson, P.E., Duran, A., Hassebrock, F., Moller, J., Prietula, M., Feltovich, P. and Swanson, D.: *Expertise and error in diagnostic reasoning.* Cognitive Science 5:235-283, 1981.

12. Johnson, P.E. and Hassebrock, F.: *Validating Computer Simulation Models of Expert Reasoning.* IN R. Trappl (Ed.), CYBERNETICS AND SYSTEMS RESEARCH. North-Holland Publishing Co., 1982.

13. Johnson, P.E. and Thompson, W.B.: *Strolling down the garden path: Detection and recovery from error in expert problem solving.* Proc. Seventh IJCAI, Vancouver, B.C., August, 1981, pp. 214-217.

14. Johnson, P.E., Hassebrock, F. and Moller, J.H.: *Multimethod study of clinical judgement.* Organizational Behavior and Human Performance 30:201-230, 1982.

15. Moller, J.H., Bass, G.M., Jr. and Johnson, P.E.: *New techniques in the construction of patient management problems.* Medical Education 15:150-153, 1981.

16. Sedlmeyer, R.L., Thompson, W.B. and Johnson, P.E.: *Knowledge-based fault localization in debugging.* The Journal of Systems and Software, vol. 3, no. 4 (Dec 83) pp. 301-307, Elsevier.

17. Sedlmeyer, R.L., Thompson, W.B. and Johnson, P.E.: *Diagnostic reasoning in software fault localization.* Proc. Eighth IJCAI, Karlsruhe, West Germany, August, 1983.

18. Smith, K.A., Farm, B., Johnson, P.E.: *Surface: A prototype expert system for selecting surface analysis techniques.* Proceedings of IEEE Conference on Computers and Comm., 1985.

19. Swanson, D.B.: *Computer simulation of expert problem solving in medical diagnosis.* Unpublished Ph.D. dissertation, University of Minnesota, 1978.

20. Swanson, D.B., Feltovich, P.J. and Johnson, P.E.: *Psychological Analysis of Physician Expertise: Implications for The Design of Decision Support Systems.* In D.B. Shires and H. Wold (Eds.), MEDINFO77, North-Holland Publishing Co., Amsterdam, 1977, pp. 161-164.

21. Thompson, W.B., Johnson, P.E. and Moen, J.B.: *Recognition-based diagnostic reasoning.* Proc. Eighth IJCAI, Karlsruhe, West Germany, August, 1983.

E. Funding and Support

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A. Medical Collaborations and Program Dissemination via SUMEX

Work in medical diagnosis is carried out with the cooperation of faculty and students in the University of Minnesota Medical School and St. Paul Ramsey Medical Center.

B. Sharing and Interactions with Other SUMEX-AIM Projects

William Clancey, Stanford University, acted as a reviewer of the MEIS Intelligent Systems Project in September, 1984 at the University of Minnesota. The Principal Investigators in the SOLVER project are also principal investigators in that project.

Paul Johnson was a panel member at the SUMEX-AIM conference in Columbus, Ohio in 1984. Dr. Connelly and two graduate students associated with the SOLVER PROJECT also attended the conference.

III. RESEARCH PLANS

A. Project Goals and Plans

Near term -- Our research objectives in the near term can be divided in three parts. First, we are committed to the design, implementation, and evaluation of Galen, as described above. We have completed an interactive front end so that physicians can directly enter patient data, and Galen's knowledge base is currently being "tuned" with the help of Dr. James Moller, an expert physician collaborator from the University of Minnesota Pediatric Cardiology Clinic, the Diagnoser program, and with expert physicians. We believe that GALEN has passed through phases of expertise assessment and cognitive simulation and that it is now approaching a level of performance that will qualify it as a true expert system. An objective now is to extend the explanation capability of GALEN. We are initiating a new investigation into two aspects of expert problem solving that relate to the interaction between a problem solving system and its environment: "query generation" and explanation. Some simple expert systems proceed from a fixed set of input data to an evaluation of that data. For most problem domains, however, the space of possibly relevant information is large, and some or all of this information may have costs associated with its acquisition. Thus, computational and other costs can be reduced by some mechanism which intelligently selects appropriate queries designed to solicit information that is relevant and cost effective in terms of the problem being solved. Expert systems for complex problem domains must also be able to generate explanations for their actions. Unless the system operates in an entirely autonomous manner, users must be apprised of the rationale for system actions. There is a particular need for explanations tailored for system users rather than system designers.

Experienced experts are typically quite proficient at asking relevant questions, even when the criteria for relevance is difficult to specify. These experts use heuristics capable of keying on selected aspects of data already examined and on the current problem state in order to select the next needed query. We propose to incorporate these heuristics into a *"query generation knowledge base"* . This knowledge base can be thought of as a form of domain specific meta-knowledge. It contains rules by which the problem state can be efficiently evaluated in order to determine the next course of action. By basing these rules on actual expert knowledge and experience, it will often be possible to bypass the combinatorial complexity associated with either blind search or optimization techniques.

Our approach to explanation starts from the premise that substantially different forms of explanation are required within a single expert system. The type of explanation is distinguished both by the level of sophistication of the person receiving the explanation and by whether that person is principally interested in the specific problem being solved or in the internal working of the expert system. Less sophisticated users of the system are likely to have only a superficial understanding of the nature of the system being diagnosed and will require explanations in terms of simplified system properties with which they are familiar. Expert users will require information about significant details of the state of the system being diagnosed and the causal relationships that connect system state with observable symptoms. Designers and maintainers of the expert system require explanations in terms of the actual lines of reasoning used to arrive at a decision.

We will be focusing principally on providing explanations for system *users* rather than system *designers*. Explanations for users must be phrased in terms of the system being diagnosed. Descriptions of the system itself are more important that descriptions of the reasoning strategies used to understand the system. For example, many diagnostic tasks are efficiently approached utilizing recognition-based reasoning strategies using knowledge arising from empirical association. Experts (or possibly automatic learning systems) learn to associate particular interpretations with particular patterns in the data. For many problem domains, knowledge of this sort is quite powerful, providing accuracy without the complexity associated with causal reasoning. The user of such a system, however, requires explanations in terms of causality. This suggest a two-step process. Problem solving is done using a recognition-based strategy. Explanations are generated by combining the results of this process with additional, causally-based explanation knowledge.

Our second objective consists of making extensions to the knowledge capturing strategies developed in our original work in medical diagnosis. In the near term this work will examine descriptive strategies in which experts attempt to use a formalized language to express what they know (e.g. production rules), observational strategies in which experts perform tasks designed to reveal information from which a theory of task specific expertise can be built, and intuitive strategies in which either experts behave as knowledge engineers or knowledge engineers attempt to perform as pseudo experts. The research projects of Dr. Spackman and Paul Krueger which have been discussed previously are both directed toward this objective.

Our third near term objective will be to investigate one of the central problems of recognition based problem solving, how to classify problems when solving them. Questions related to problem classification which we will be examining include: What patterns do experts and novices detect in a problem that allows them to classify it as an instance of a problem type that is already known? How does an expert make an initial choice of the level of abstraction to be used in solving a problem? How can an expert recover from an initial incorrect choice of levels? How can the difference between causal and prototypic modes of reasoning be modeled as differences in levels of abstraction, and how can a common model for these two types of reasoning be

constructed? We will be pursuing these questions in the areas of problem solving like law, auditing, and management, as well as in medicine.

Long range -- Our long range objective is to improve the methodology of the "knowledge capturing" process that occurs in the early stages of the development of expert systems when problem decomposition and solution strategies are being specified. Several related questions of interest include: What are the performance consequences of different approaches, how can these consequences be evaluated, and what tools can assist in making the best choice? How can organizations be determined which not only perform well, but are structured so as to facilitate knowledge acquisition from human experts? In the coming year we will be exploring these questions in areas of design and management as well as in law, management and medicine.

B. Justification and Requirements for Continued SUMEX Use

Our current model development takes advantage of the sophisticated Lisp programming environment on SUMEX. Although much current work with Galen is done using a version running on a local VAX 11/780, we continue to benefit from the interaction with other researchers facilitated by the SUMEX system. We expect to use SUMEX to allow other groups access to the Galen program. We also plan to continue use of the knowledge engineering tools available on SUMEX.

We are working toward a Commonlisp implementation of the GALEN system and expect to rely heavily on Commonlisp for future projects.

One of our students implemented a demonstration legal expert system in EMYCIN using the SUMEX resource, and we still find that the resource is valuable for making available major systems which we do not have locally, such as EMYCIN.

C. Needs and Plans for Other Computing Resources Beyond SUMEX-AIM

Our current grant from MEIS has permitted us to purchase four Perq 2 AI workstations for our Artificial Intelligence laboratory. The availability of Commonlisp on these machines is one reason why we expect to make use of that language in the future.

SUMEX will continue to be used for collaborative activities and for program development requiring tools not available locally.

D. Recommendations for Future Community and Resource Development

As a remote site, we particularly appreciate the communications that the SUMEX facility provides our researchers with other members of the community. We, too, are moving toward a workstation based development environment, but we hope that SUMEX will continue to serve as a focal point for the medical AI community. In addition to communication and sharing of programs, we are interested in development of Commonlisp based knowledge engineering tools. The continued existence of the SUMEX resource is very important to us.

6.3. Stanford Pilot Projects

Following are descriptions of the informal pilot projects currently using the Stanford portion of the SUMEX-AIM resource, pending funding, full review, and authorization.

6.3.1. CAMDA Project

CAMDA Project

CAMDA Research Staff:

Prof. Samuel Holtzman, Co-PI	Engineering-Economic Systems
Prof. Ronald A. Howard, Co-PI	Engineering-Economic Systems
Prof. Ross Shachter	Engineering-Economic Systems
Leonard Bertrand	Engineering-Economic Systems
Jack Breese	Engineering-Economic Systems
Kazuo Ezawa	Engineering-Economic Systems
Keh-Shiou Leu	Engineering-Economic Systems
Seok Hui Ng	Engineering-Economic Systems
Emilio Navarro	Engineering-Economic Systems
Dr. Adam Seiver	Engineering-Economic Systems
Joseph Tatman	Engineering-Economic Systems
Dr. Emmet Lamb	School of Medicine
Dr. Robert Kessler	School of Medicine
Dr. Frank Polansky	School of Medicine

Associated faculty:

Prof. Edison Tse	Engineering-Economic Systems

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The Computer-Aided Medical Decision Analysis (CAMDA) project is an attempt to develop intelligent medical decision systems by combining the descriptive generality of expert-system technology with the normative power of decision analysis.

B. Medical Relevance and Collaboration

The primary effort of the CAMDA project during 1984 and early 1985 has been focused on the design and implementation of RACHEL, an intelligent decision system for infertile couples. This system is designed to help patients and physicians deal with difficult medical treatment choices. RACHEL is being developed in close cooperation with the Engineering-Economic Systems Department, the Obstetrics and Gynecology Department, and the Surgery Department (Urology Division), all at Stanford.

In addition to the development of RACHEL, there are several active research programs within the CAMDA project. One such program is aimed at developing a representation for dynamic decision processes (such as those faced by cancer patients) that do not necessarily satisfy the Markov assumption. Another is concentrating on the development of fast algorithms for the solution of general decision problems.

A recent addition to our research project is a program to design cost-effective strategies for monitoring the recurrence of bladder cancer.

C. Highlights of Research Progress

C.1 Accomplishments this past year

We have successfully implemented a pilot-level version of RACHEL. As we define it, a pilot system is one where the essential algorithms work individually as well as interactively with one another, operating with knowledge that is representative of the system's domain. Such a system lacks two important elements that must exist within a prototype-level implementation: an extensive knowledge base, and a front end usable by trained users who may not be familiar with the details of the system.

As part of the development of RACHEL, we have developed a facility to construct individualized models of the patient's preferences over the set of possible outcomes of an infertility therapy. This facility operates in two consecutive stages. The first stage constructs a parametric model from a library of plausible model elements. A typical consideration at this stage is whether to explicitly account for the patient's lifetime. For instance, a treatment strategy which involves surgery would warrant such explicit consideration, whereas a therapy consisting strictly of drugs would not. The second stage in the preference model development process involves the assessment of specific parametric values. These values are obtained directly from the patient to ensure that the overall preference model genuinely reflects his or her desires.

It is important to note that since the preference model is built to fit the specific needs of each case, the interaction between the patient and the system is short and well-focused. In particular, the patient is only asked to respond to a few (about five to ten) questions. These questions are selected so that their relevance to the case is intuitively obvious from the patient's point of view.

Also as a part of RACHEL, we have developed a knowledge base dealing with the decisions faced by the subset of infertile couples whose inability to conceive has been traced to a blockage of the Fallopian tubes of the female partner. In particular, the knowledge in RACHEL deals with the choice between two important procedures pertinent to this condition: laparotomy and in-vitro fertilization.

Another accomplishment during this past research year has been the improvement of our influence-diagram solution procedure. In its original form, this procedure essentially took a brute-force approach to the solution of well-formed influence diagrams. Although its solutions were mathematically correct, the program was inefficient in terms of both computational time and storage requirements. In its current implementation, the program is considerably more efficient and has an adequate front end which makes it accessible to a fairly wide class of users. Empirical results indicate that the size and complexity of problems that can be represented and solved with the system not only exceed the bounds of its original design, but are comparable and possibly superior to those of the best commercially available decision-analytic software.

Similarly, RACHEL's inference engine has been improved in several important ways. Prominent among these are a means for attaching general procedures at any point in the inference process, a variety of built-in procedures for the acquisition and display of information coupled with a facility for controlling these procedures (i.e., for the control of ASKability and TELLability), and a simple explanation mechanism.

C.2 Research in progress

The RACHEL system continues to be developed along four distinct directions: the efficiency and flexibility of RACHEL's inference engine are being improved, its explanation mechanism is being enhanced, RACHEL's facility for the development of patient preference models is being upgraded, and its knowledge base is being enlarged.

As it is currently implemented, the inference engine used by RACHEL is quite inefficient. This inefficiency is, to some extent, a deliberate design choice since the engine was designed to be very general and highly modular. Thus, there are many procedural redundancies and much unnecessary baggage in the programs that implement it. Now that we have a clearer idea of how the engine is to be used we have redesigned it by doing away with some of the original generality and modularity in favor of a more efficient process. Furthermore, the new design emphasizes and enhances particularly useful engine features such as its ASKability and its TELLability.

A further enhancement to RACHEL's inference engine concentrates on the system's ability to explain its line of reasoning. The original design only responds to online "why" queries by displaying its dynamic goal stack. In its new form, the engine allows offline as well as online queries in both "why" and "how" formats.

Beyond traditional explanation capabilities, we are exploring possible means to explain decision-theoretic inferences. In particular, we are trying to understand how to explain decision recommendations that are based on the maximization of expected utility to users unfamiliar with decision theory. Our current research indicates that a promising way to do this is to break down large decision problems into smaller, more manageable pieces whose formal solution can be checked against intuition. Although still at an early stage, this line of research seems to be on the path of eliminating an important barrier to the widespread use of normative decision techniques.

An exciting area of current interest is the improvement of RACHEL's facility for the creation and assessment of parametric models of patient preferences. In particular, we are trying to increase the generality of RACHEL's model library to account for acute as well as chronic conditions and to simplify the corresponding assessment process. This simplification is based on the notion that a better understanding of the major concerns of patients can help us redesign the questions asked by RACHEL so that they are closer to the specific experiences of individual patients. As part of this effort, we expect to have significant contact with actual patients to ensure the clinical relevance of our research.

A fourth area where RACHEL is being enhanced is the expansion of its medical and decision-analytic knowledge bases. Planned additions include further knowledge about the treatment of tubal blockage (including more data on in-vitro fertilization procedures and an ability to consider a wider class of patients) and a new packet of knowledge dealing with deterministic sensitivity analysis.

In addition to the development of RACHEL, there are several active research programs within the CAMDA project. One such program is aimed at developing a representation for dynamic decision processes (such as those faced by cancer patients) that do not necessarily satisfy the Markov assumption. This research has led to a generalization of influence diagrams which allows multiple value nodes. This generalization makes it possible for complex sequential decision processes (whose solution would otherwise be infeasible) to be efficiently solved.

Another research program within the CAMDA project is the development of fast algorithms for the solution of decision problems formulated as influence diagrams. In general, the solution of an influence diagram (i.e., the calculation of a recommended decision strategy) is obtained by the repeated application of an operation, known as "removal", to all nodes in the diagram other than the value node. The removal of a node in the diagram is a generalization of the foldback operation needed to solve a decision tree. With rare exceptions, the order in which nodes are removed from a diagram is not unique. Current results indicate that significant reductions in the computational burden of solution can be achieved by controlling the order in which diagram nodes are selected for removal.

At a more fundamental level, we are exploring the consolidation of the predicate calculus with probabilistic logic. Of particular interest is the design of an integrated inference engine that performs logical inferences within a probabilistic framework. A central problem in this research is the definition of universal and existential quantification in probabilistic terms.

A recent addition to our research project is a program to design cost-effective strategies for monitoring the recurrence of bladder cancer. We expect this research to interact with our ongoing search for more effective models of patient preferences.

D. Publications

1. Holtzman, S.:*A Model of the Decision Analysis Process*, Department of Engineering-Economic Systems, Stanford University, Stanford, California, 1981.

2. Holtzman, S.:*A Decision Aid for Patients with End-Stage Renal Disease*, Department of Engineering-Economic Systems, Stanford University, Stanford, California, 1983.

3. Holtzman, S.:*On the Use of Formal Models in Decision Making*, Proc. TIMS/ORSA Joint Nat. Mtg., San Francisco, May, 1984.

4. (*) Holtzman, S.: *Intelligent Decision Systems*, Ph.D. Dissertation, Department of Engineering-Economic Systems, Stanford University, Stanford, California, 1985.

5. Shachter, R.: *Evaluating Influence Diagrams*, Department of Engineering-Economic Systems, Stanford University, Stanford, California, 1984.

6. Shachter, R.: *Automating Probabilistic Inference*, Department of Engineering-Economic Systems, Stanford University, Stanford, California, 1984.

E. Funding Support

E.I Principal Funding Source

E.I.1. Title of gift

"Research on Intelligent Decision Systems".

E.I.2. Principal investigator

Samuel Holtzman, Ph.D.
Consulting Assistant Professor
Department of Engineering-Economic Systems
Stanford University

E.I.3. Funding source

Olivetti Advanced Technology Center, Inc.

E.I.5. Funding amount

$33,400 (Direct Costs), unrestricted.

E.II Additional Funding Source

E.I.1. Title of gift

"Cost-effective strategies in monitoring for recurrence of bladder cancer"

E.II.2. Principal Investigators

Ross Shachter, Ph.D. -- PI
Assistant Professor
Department of Engineering-Economic Systems
Stanford University

Linda Shortliffe, M.D. -- Co-PI
Palo Alto Veterans Administration Hospital

Dan Kent, M.D. -- Co-PI
Division of General Internal Medicine
Stanford University Medical Center

Samuel Holtzman, Ph.D. -- PI: CAMDA Project (SUMEX)
Consulting Assistant Professor
Department of Engineering-Economic Systems
Stanford University

E.II.3. Funding agency

Stanford's American Cancer Society Institutional Research Grant Committee

E.II.5. Total award

$4634 (Direct Costs), for the year starting April 1, 1985

E.III Other Funding

E.III.2 Donated Equipment

The CAMDA project has access to the facilities of the Decision Systems Laboratory (DSL) in the Department of Engineering-Economic Systems, and constitutes the laboratory's most active research project. The DSL maintains several terminals, printers and personal computers for research on the development of computer-based decision systems. The majority of the terminals and printers were donated to the DSL by Qume Corporation. Olivetti Advanced Technology Center, Inc., has made four M24 personal computers and two high-quality printers available to the DSL on a "Beta-test-site" basis. MAD Computer, Inc., has also contributed to the support of the CAMDA project through the consignment of a MAD-1 personal computer.

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

II.A Medical Collaborations and Program Dissemination Via SUMEX

Since its inception, the CAMDA project has benefited from an active relationship among decision analysts, computer scientists, and members of the Stanford medical community. In particular, RACHEL is being developed in close cooperation with physicians in the Infertility Clinic at Stanford. Other programs within the CAMDA project such as our research on the form and use of medical preference models are being done in cooperation with physicians at the Palo Alto Veterans Administration Hospital and at El Camino Hospital.

II.B. Sharing and Interactions with other SUMEX-AIM Projects

II.B.1 SUMEX-AIM 1984 Workshop:

Samuel Holtzman participated in the 1984 AIM workshop in Columbus, Ohio. In addition to the presentation of a summary of CAMDA research, he had many opportunities to interact with workshop participants on an informal basis. Of particular interest were several discussions with members of the MIT/TUFTS group interested in medical decision analysis which have led to an interchange of ideas that continues to this date.

II.B.2 Decision Systems Laboratory Research Meetings

As part of the CAMDA project, we have instituted a weekly research meeting for those interested in the design and implementation of computer-based decision systems. This weekly meeting has become a very active forum for the presentation of research results. The following topics of direct relevance to medical decision making were presented during the last two academic quarters.

Date	Speaker	Topic
03-OCT-84	Ross Shachter	Probabilistic Inference
17-OCT-84	Jack Breese	Dempster-Shafer Theory
24-OCT-84	Kazuo Ezawa	Efficiency in Solving Influence Diagrams
07-NOV-84	Majid Khorram	Fuzzy Sets and Decision Making
14-NOV-84	Dan Kent	Utility Theory Underlying Physicians' Treatment Thresholds: HELP!
21-NOV-84	Yann Bonduelle	Explanation in Decision Systems
09-JAN-85	Ross Shachter	What Do You Call the Offspring of SUPERID and INFLUENCE?
23-JAN-85	Doug Logan	The Value of Probability Assessment
06-FEB-85	Seok Hui Ng	Minimal Tumor Follow-up Examination Schedule for Recurrent Bladder Cancer Patients.
13-FEB-85	Keh-Shiou Leu	TEREISIAS' Explanation Facility
06-MAR-85	Joe Tatman	Algorithm for Decision Processes Optimization
13-MAR-85	Gerald Liu (UC)	Knowledge Structure in Evidential Reasoning

II.B.3 Course in Medical Decision Analysis

A new course in medical decision analysis, taught by Prof. Samuel Holtzman, is being offered for the first time during the Spring quarter of 1985. The course is offered jointly by the Engineering-Economic Systems Department, the Medical Information Sciences Program, and the Computer Science Department. The objective of the course is to expose students to the practice of decision analysis for clinical purposes and to introduce them to the design and use of computer-based medical decision tools.

II.C. *Critique of Resource Management*

The CAMDA project is heavily dependent upon the availability of the SUMEX computing resource. The physical facility as well as the staff of SUMEX-AIM are excellent. In particular, it has been a pleasure to deal with Ed Pattermann, who is invariably courteous, responsive to our needs, and effective in his actions. We will certainly miss him now that he has moved to industry. Pam Ryalls has also provided much needed help in managing the CAMDA project in a manner that is friendly and efficient.

As an update to last year's report, the previously reported Ethernet deficiencies have been corrected. This improvement was part of a campus-wide effort to improve Stanford's computer network which directly affected our campus connection to SUMEX. The system load on SUMEX continues to be heavy, although it appears to be somewhat lower than it was last year. The ability of the CAMDA project to use the DECSYSTEM-2020 machine operated by SUMEX (referred to as TINY) has had a significant effect on our ability to demonstrate our systems during normal business hours, further reducing our frustration with the main system's load.

III. RESEARCH PLANS

III.A *Project Goals and Plans*

During the upcoming year, we intend to enhance four specific elements of the RACHEL system: its inference mechanism, its explanation facility, its ability to model patient preferences, and its medical and decision-analytic knowledge bases. Furthermore, we intend to continue to improve our understanding of normative decision methodologies, with particular emphasis on the use of these methodologies for computer-based decision support. Section I.C.2 describes the near-term goals of the CAMDA project in more detail. Our long-term goal remains that of designing and implementing usable, fully-validated and documented systems for medical decision support.

III.B *Justification and Requirements for Continued SUMEX Use*

The CAMDA project is truly interdisciplinary. It draws on elements of decision analysis, artificial intelligence, and medical science. The project has the potential to contribute to each of these disciplines in important ways.

In particular, the CAMDA project is likely to lead to the development of tools and techniques that greatly improve the quality of decision making in medicine. For instance, RACHEL explicitly considers uncertainty, decision alternatives, and patient preferences in developing recommendations. In spite of its generality, RACHEL's interaction with the user is sufficiently terse and simple to support the claim that systems based on its methodology can be effective clinical decision tools. Much of the simplicity and terseness of RACHEL's operation is a direct consequence of the AI foundations of the system's design.

The heavy reliance of the CAMDA effort on artificial intelligence technology make SUMEX-AIM an ideal environment in which to pursue this research.

III.C *Needs and Plans for other Computing Resources beyond SUMEX-AIM*

The CAMDA project has access to four Olivetti M24 and one MAD-1 personal computers (IBM-PC type) as well as to one Apple Macintosh (128K) computer. In addition, we continue to search for funds to acquire one or more state-of-the-art LISP machines.

III.D *Recommendations for Future Community and Resource Development*

What would be the effect of imposing fees for using SUMEX resources (computing and communications) if NIH were to require this?

A major benefit provided by the existing SUMEX-AIM facility is the availability of very low-cost computing resources. Access to these resources is granted primarily on the basis of an assessment of the value of the proposed research to the overall goal of making artificial intelligence a useful medical tool. Imposing fees for using SUMEX would prevent users with modest means from obtaining access to the facility on the basis of merit alone.

Do you have plans to move your work to another machine workstation and if so, when and to what kind of system?

The CAMDA project has access to several personal computers for its research. These machines include Olivetti M24's (marketed as the A.T.&T. personal computer in the U.S.) and a MAD-1 personal computer -- all of which are compatible with the IBM-PC. In addition, the project has purchased an Apple Macintosh. These machines are used as a supplement to the SUMEX mainframe, and are not intended to replace it.

6.3.2. REFEREE Project

Bruce G. Buchanan, Ph.D.
Computer Science Department
Stanford University

Byron W. Brown, Ph.D.
Dept. of Biostatistics
Stanford University

Daniel E. Feldman, Ph.D., M.D.
Department of Medicine
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The goal of this project is two-fold: (a) use existing AI methods to implement an expert system that can critique medical journal articles on clinical trials, and (b) in the long term, develop new AI methods that extract new medical knowledge from the clinical trials literature. In order to accomplish (a) we are building the system in three stages.

1. System I will assist in the evaluation of the quality of a single clinical trial. The user will be imagined to be the editor of a journal reviewing a manuscript for publication, but the program will be tested on a variety of readers, including clinicians, medical scientists, medical and graduate students, and clerical help.

2. System II will assist in the evaluation of the effectiveness of the treatment or intervention examined in a single published clinical trial. The user will be imagined to be a clinician interested in judging the efficacy of the treatment being tested in the trial.

3. System III will assist in the evaluation of the effectiveness of a single treatment examined in a number of published clinical trials.

B. Medical Relevance

The burden of "keeping up with the literature" is particularly onerous in the practice of medicine and in medical research [62, 63]. Reading the abstracts in a few journals and selecting several key articles for a rapid survey are the best that most clinicians can hope to accomplish each week. The time and effort necessary for a thorough and critical reading of even a few research reports are not available.[1] Sackett reports that to keep up with the 10 leading journals in internal medicine a clinician must read 200 articles and 70 editorials per month [63]. It was also estimated that the biomedical

---

[1] In an informal check on this intuition two of us, with considerable training in analyzing clinical trials (BWB and DEF) timed critical readings of a five page article on a clinical trial in the New England Journal of Medicine [4]. Our times were 30 and 120 minutes.

literature is expanding at a compound rate of 6% to 7% per year, or doubling every 10 - 15 years [63, 59]. Furthermore, even if more time were available the statistical and epidemiological skills necessary for critical reading are not part of most clinicians' repertoires[1]; and yet decisions about which therapy to use, what intervention to adopt, or what advice to give patients must be based on a combination of clinical experience and published literature. But the existing literature is often confusing and contradictory [42]and publication in the most prestigious medical journals does not guarantee freedom from serious methodologic flaws and erroneous conclusions [44, 18]. Any assistance to the clinician must deal with both the problem of the vastness of the literature and the quality of the research report. Similar problems are faced by the editors of medical journals, swamped with manuscripts to review and evaluate, and by research scientists and academicians trying to stay abreast of the developments in their fields. How can they cover more and yet evaluate better and more consistently? Clearly any machine assistance would be welcome.

C. Highlights of Progress

This project is just getting started.

Preliminary work has been done on REFEREE [23], a prototype expert system for determining the quality of a clinical trial report, and the efficacy of the intervention evaluated in the trial. REFEREE is written in EMYCIN, a rule-based programming language which allows rapid prototyping of a consultation system that gives advice to a user. It presupposes that a knowledge base about the problem area has been constructed, which usually involves codifying an expert's knowledge.

The basic format of a REFEREE session is fairly simple. The reader is asked a series of questions pertaining to the paper and the study described. The answers given are used to rate the overall quality of the paper and the probable efficacy of the treatment described. (See sample dialogs below).

In the first version of REFEREE, after the program has finished with its chain of questions and deductions, the quality of the paper and the efficacy of the drug are given to the user as a "merit score", an integer between 0 and 10, with 10 indicating the highest quality. Additionally, the user is provided with a series of English language messages indicating the main flaws detected in the paper. The merit score was used because the expert system makes its judgements by using a weighted average of values assigned to each aspect of the paper being critiqued. As the user answers the consultant's questions, the answers are given individual merit scores. For example, if the user's answer indicate that experimental blinding was done correctly, the paper is given a high score in the blinding category. When all merit score assignments have been made, the total merit score is calculated as a weighted average of the categorical merit scores, with those categories that are more crucial to a good paper or clinical trial being given a higher weight.

The final result of this calculation is a number between 1 and 10 which serves as a quality measure for the paper or the treatment. A 1 indicates low quality; a 10 indicates the highest quality. An integer as a final result, however, can be very cryptic. It is usually quite difficult, given just an integer, to understand or believe the findings of the consultant. It was discovered quite early that users, when presented with just the bare merit score of the paper, would want to know why the paper was rated in the way it was. For this reason, English language statements are given to the user, indicating the nature of the main flaws of the paper. In each category, if the calculated merit score is

---

[1]A recent survey of the statistical methods used by authors in the New England Journal of Medicine indicated that 42 per cent of the articles surveyed relied on statistical analysis beyond descriptive statistics [15].

found to be less than an arbitrary minimum, this is noted in a sentence or two, and given to the user at the end of the consultation. In this way, the user not only gets an overall picture of the quality of the paper, but also an indication of the general areas in which the paper was found to be lacking.

Several problems were found in the original version of REFEREE. It was discovered that the use of a weighted average precluded the use of EMYCIN's certainty factors. Because of this, the user would often be forced to choose from a fairly limited set of possible answers to the consultant's questions. The lack of versatility implied by this constraint dictated that a new approach which could make full use of EMYCIN's certainty factors should be used.

In order to do this, the old rule base was scrapped, and a new one was written. Instead of deciding on a rating between one and ten to indicate quality, the new version simply decides whether or not the paper in question is of "high academic and scholarly quality", with an EMYCIN certainty factor modifying the conclusion. For example, in the case of a mediocre paper, the program would conclude that the paper was of "high quality", but only with a certainty of say, .5, on a scale between -1 and 1. Though the words "certainty factor" are used for historical reasons, our final number is the equivalent of a merit score.

While at first glance the two approaches seem similar, the second approach was found to be much more flexible and satisfying from the user's standpoint. Since the conclusion is in terms of the programs certainty that the paper's quality is good, the user may incorporate his or her own uncertainty into the dialogue with the program. This was accomplished by asking mainly yes/no questions, and at all times allowing the user to indicate his or her certainty in the answers given. Thus, if the program asks the user if the quality of the paper's literature review was high, he or she can answer simply "yes" or "no", indicating complete confidence in the answers, or modify a yes/no answer with a certainty factor, indicating that he or she is not completely certain. The user's answers, along with the uncertainty indicated by him or her, will be combined by EMYCIN to give a final conclusion on the paper's quality.

As an example, one of the old-style rules might have been something like this: If the user indicates that the literature review is of "poor quality", conclude that the merit of the paper is 3 with a (built-in) weight of 2. After all the merit values had been calculated, a weighted average, (using built-in weights) would be taken to come to the final merit score. In contrast, one of the new rules would be of the form: If the user gives a "yes" answer to the question "Is the literature review thorough and balanced?", conclude that the paper is of good quality with a certainty of .3. While in the first case the user was limited to a set of possible answers (e.g. excellent, good, poor), the second rule gives the user the opportunity to answer either yes or no, and qualify that answer with any degree of certainty desired. If, in the second rule, the user gives a certainty of less than 1 that the literature review was of good quality, the inferred conclusion about the quality of the paper will be automatically downgraded as well. In other words, if the user expresses uncertainty, the conclusion about the quality of the paper will be less certain.

The new approach, in addition to supplying the user with the ability to express varying degrees of uncertainty, also allows for a hierarchical question structure. At any point, if the user is unclear of the appropriate response, the program can prompt with further, more detailed questions, until a conclusion about the original question can be provided. Conversely, whenever a user is willing to give an answer, the program will refrain from dwelling on the issue and omit its long series of sub-questions. In this manner the amount of detail provided can be individualized.

This current version of REFEREE has two hundred rules and has been tested by the present research team on several papers. It is this program that will be expanded as described in Section III-A. Part of a sample consultation is shown below.

```
--------MEDICINE-1--------
The first paper of MEDICINE-1 will be referred to as:

    --------PAPER-1--------

        --------STATISTICS-1--------
        1) What is the size of the control sample?
        ** 25
        2) How many of the subjects in the control sample responded to
           treatment?
        ** 14
        3) What is the size of the test sample?
        ** 23
        4) How many of the subjects in the test sample responded to
           treatment?
        ** 23
...


        --------PLANNING-1--------
        9) Was there an explicit stopping rule defined before the experiment
           was run?
        ** N

        --------RANDOMIZATION-1--------
        10) Was there any mention of the use of randomization in patient
            assignment?
        ** Y
        11) Was the assignment of subjects in the experiment performed blindly?
        ** UNK
...


        --------BLINDING-1--------
        16) Was the experiment double blinded, or was any mention made of
            blinding in the experiment?
        ** Y
        17) Was there any mention of an effort to make the placebo and
            medication as similar as possible?
        ** N
...

The strength of the evidence indicating the efficacy of PAPER-1 is as
   follows:
        There is some evidence for efficacy, but further study is needed.

The general quality of the paper is as follows:
        The current paper is of poor quality.

The flaws of the current paper are as follows:
        A stopping rule was not defined or was not adhered to in the
           experiment.
        The measures taken to evaluate subject compliance were inadequate or
           non-existent.
        Subjects were not randomly assigned treatment groups, seriously
           weakening the validity of the conclusions.
        Though an effort was made to blind the experiment, the techniques
           used were not effective.

The final calculated efficacy of the drug as indicated by the given clinical
   trial (between 0 and 10, with a score of 10 being the highest) is as
   follows:
        5.

The final merit of the current paper is as follows:
        3.

   23) Are there any other papers on MEDICINE-1?
   ** N
   24) Do you want the results of this consultation output to a file?
   ** N
```

E. Funding Support

Grant applications submitted to the NLM:

Title: Understanding and Critiquing Clinical Trials Literature

PI's: Bruce G. Buchanan, Byron W. Brown

Agency: National Library of Medicine (Pending)

Total Amount: $178,923.

Dates: July 1, 1985 - June 30, 1988

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A. Medical Collaborations

Dr. D. Feldman is a physician and epidemiologist at the Stanford Center for Disease Prevention.   Prof. B. Brown is currently teaching a Medical School class on reading medical journal articles.

B. Interactions with other SUMEX-AIM projects

Our interactions have all been through the Knowledge Systems Laboratory where we have discussed design and implementation issues.

C. Critique of Resource Management

The SUMEX staff has been most cooperative in helping get this project started.   We have tried to place few demands on the SUMEX staff, but have received prompt answers to all questions.

III. RESEARCH PLANS

A. Goals & Plans

It is proposed to construct three computer-based expert systems to assist a variety of different readers in the evaluation of an extensive but well defined area of the medical literature, clinical trials.   It is further proposed to test the hypothesis that such programs will enable a variety of users to read the literature on clinical trials more more critically and more rapidly.

The expert systems will be developed using the EMYCIN programming environment and the production rule approach followed successfully in previous expert systems [24, 36, 43, 48, 6].

The three programs to be developed are separate, but closely related:

1. System I will assist in the evaluation of the quality of a single clinical trial. The user will be imagined to be the editor of a journal reviewing a manuscript for publication, but the program will be tested on a variety of readers, including clinicians, medical scientists, medical and graduate students, and clerical help.

2. System II will assist in the evaluation of the effectiveness of the treatment or intervention examined in a single published clinical trial. The user will

be imagined to be a clinician interested in judging the efficacy of the treatment being tested in the trial.

3. System III will assist in the evaluation of the <u>effectiveness</u> of a single treatment examined in a <u>number</u> of published clinical trials.

Within the duration of this research it is also proposed to test the first two systems against unassisted evaluations by the various categories of readers. The testing will include a formal testing of the programs by comparing the speed and number of flaws found in using the program with similar measurements on unassisted reading. In addition there will be a more informal evaluation by questionnaire of the subjective impressions of users of the program, ascertaining the likelihood of routine use and the value of such a program to the user.

This proposal with its concentration on clinical trials is regarded as the initial step in a more general research goal - building computer systems to help the clinician and medical scientist read the medical literature more critically.

B. Justification for continued SUMEX use

We will continue to use SUMEX for developing the AI methods. We need EMYCIN at the moment because it provides a good environment for building a rule-based system that may grow to many hundreds of rules. EMYCIN is not available on other machines without substantial cost.

C. Need for other computing resources

In the short term we will not need additional resources. Should we decide to implement a new system in a framework other than EMYCIN, we might seek funding to buy a LISP workstation.

D. Recommendations

Although our use has been small, we find the load average on SUMEX often precludes running test cases during the day. We have no specific recommendation, but would like to have access to small amounts of high quality computer time.

6.4. National AIM Pilot Projects

Following is a description of the informal pilot projects currently using the national AIM portion of the SUMEX-AIM resource, pending funding, full review, and authorization.

6.4.1. PATHFINDER Project

PATHFINDER Project

Bharat Nathwani, M.D.
Department of Pathology
University of Southern California

Lawrence M. Fagan, M.D., Ph.D.
Department of Medicine
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

Our project addresses difficulties in the diagnosis of lymph node pathology. Five studies from cooperative oncology groups have documented that, while experts show agreement with one another, the diagnosis made by practicing pathologists may have to be changed by expert hematopathologists in as many as 50% of the cases. Precise diagnoses are crucial for the determination of optimal treatment. To make the knowledge and diagnostic reasoning capabilities of experts available to the practicing pathologist, we have developed a pilot computer-based diagnostic program called PATHFINDER. The project is a collaborative effort of the University of Southern California and the Stanford University Medical Computer Science Group. A pilot version of the program provides diagnostic advice on 80 common benign and malignant diseases of the lymph node based on 150 histologic features. Our research plans are to develop a full-scale version of the computer program by substantially increasing the quantity and quality of knowledge and to develop techniques for knowledge representation and manipulation appropriate to this application area. The design of the program has been strongly influenced by the INTERNIST/CADUCEUS program developed on the SUMEX resource.

A group of expert pathologists from several centers in the U.S., have showed interest in the program and helped to provide the structure of the knowledge base for the PATHFINDER system.

B. Medical Relevance and Collaboration

One of the most difficult areas in surgical pathology is the microscopic interpretation of lymph node biopsies. Most pathologists have difficulty in accurately classifying lymphomas. Several cooperative oncology group studies have documented that while experts show agreement with one another, the diagnosis rendered by a "local" pathologist may have to be changed by expert lymph node pathologists (expert hematopathologists) in as many as 50% of the cases.

The National Cancer Institute recognized this problem in 1968 and created the Lymphoma Task Force which is now identified as the Repository Center and the Pathology Panel for Lymphoma Clinical Studies. The main function of this expert panel of pathologists is to confirm the diagnosis of the "local" pathologists and to ensure that the pathologic diagnosis is made uniform from one center to another so that the comparative results of clinical therapeutic trials on lymphoma patients are valid. An expert panel approach is only a partial answer to this problem. The panel is

useful in only a small percentage (3%) of cases; the Pathology Panel annually reviews only 1,000 cases whereas more than 30,000 new cases of lymphomas are reported each year. A Panel approach to diagnosis is not practical and lymph node pathology cannot be routinely practiced in this manner.

We believe that practicing pathologists do not see enough case material to maintain a high-level of diagnostic accuracy. The disparity between the experience of expert hematopathology teams and those in community hospitals is striking. An experienced hematopathology team may review thousands of cases per year. In contrast, in a community hospital, an average of only 10 new cases of malignant lymphomas are diagnosed each year. Even in a university hospital, only approximately 100 new patients are diagnosed every year.

Because of the limited numbers of cases seen, pathologists may not be conversant with the differential diagnoses consistent with each of the histologic features of the lymph node; they may lack familiarity with the complete spectrum of the histologic findings associated with a wide range of diseases. In addition, pathologists may be unable to fully comprehend the conflicting concepts and terminology of the different classifications of non-Hodgkin's lymphomas, and may not be cognizant of the significance of the immunologic, cell kinetic, cytogenetic, and immunogenetic data associated with each of the subtypes of the non-Hodgkin's lymphomas.

In order to promote the accuracy of the knowledge base development we will have participants for multiple institutions collaborating on the project. Dr. Nathwani will be joined by experts from Stanford (Dr. Dorfman), St. Jude's Children's Research Center -- Memphis (Dr. Berard) and City of Hope (Dr. Burke).

C. Highlights of Research Progress

C.1 Accomplishments This Past Year

Since the project's inception in September, 1983, we have constructed several versions of PATHFINDER. The first several versions of the program were *rule-based* systems like MYCIN and ONCOCIN which were developed earlier by the Stanford group. We soon discovered, however, that the large number of overlapping features in diseases of the lymph node would make a rule-based system cumbersome to implement. We next considered the construction of a *hybrid system*, consisting of a rule-based algorithm that would pass control to an INTERNIST-like scoring algorithm if it could not confirm the existence of classical sets of features. We finally decided that a modified form of the INTERNIST program would be most appropriate. The original version of PATHFINDER is written in the computer language Maclisp and runs on the SUMEX DEC-20. This was transferred to Portable Standard Lisp (PSL) on the DEC-20, and later transferred to PSL on the HP 9836 workstations. Two graduate students, David Heckerman and Eric Horvitz, designed and implemented the program.

C.1 The PATHFINDER knowledge base

The basic building block of the PATHFINDER knowledge base is the disease profile or *frame*. The disease frame consists of *features* useful for diagnosis of lymph node diseases. Currently these features include histopathological findings seen in both low- and high-power magnifications. Each feature is associated with a list of exhaustive and mutually exclusive *values*. For example, the feature *pseudofollicularity* can take on any one of the values *absent, slight, moderate,* or *prominent.* These lists of values give the program access to *severity* information. In addition, these lists eliminate obvious interdependencies among the values for a given feature. For example, if pseudofollicularity is *moderate,* it cannot also be *absent.*

Evoking strengths and frequencies are associated with each feature-value pair in a