DEVELOPMENT OF THE NEW BEST INFORMATION ALGORITHM FOR A MEDICAL EXPERT SYSTEM (ILIAD)

By

Di Guo

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medical Informatics

The University of Utah

August 1993

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Di Guo

This dissertation has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

Chair: Homer R. Warner

Reed M. Gardner

1993

J. Lincoln

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

FINAL READING APPROVAL

To the Graduate Council of The University of Utah:

I have read the thesis of _____ in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the Supervisory Committee and is ready for submission to The Graduate School.

Date

Homer R. Warner
Chair, Supervisory Committee

Approved for the Major Department

Homer R. Warner
Chair/Dean

Approved for the Graduate Council

B. Gale Dick
Dean of The Graduate School

ABSTRACT

Iliad is a diagnostic expert system for internal medicine. One important feature that Iliad offers is the ability to analyze a particular patient case and to determine the most cost-effective findings to pursue next at any stage of a work-up. The "best information" algorithm combines an information content calculation together with a cost factor. The calculations then provide a rank-ordering of the alternative patient findings according to cost-effectiveness.

This dissertation presents a three-part study to evaluate the performance of different best information algorithms. In the first two parts of the study the suggestions about the next best data elements to pursue from different algorithms were collected for different vignettes. The performance of different algorithms was compared based on the judgments provided by expert clinicians. The results indicated that the current Iliad information content model could be improved by using a version of Shannon information content model.

The third part of the study evaluated different best information algorithms by a simulation approach. The results indicated that two types of diagnostic behaviors could be simulated. The first type of behavior was characterized by pursuing more history and physical examination findings, less laboratory tests, less expensive work-ups, and more steps to solve a patient case. The second type of behavior was characterized by pursuing less history and physical examination findings, more laboratory tests, more expensive work-ups, and less steps to solve a patient case. The Shannon information content model accomplished work-ups that were significantly less costly than work-ups performed by the current LR (likelihood ratio) information content model. However, the Shannon model required additional computational resources and more history and physical examination

steps than the LR model. Decisions regarding the implementation of alternative models require a balance of the relative merits of cost, steps, expert preference, and other important factors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my advisor Dr. Homer R. Warner for his support and counsel and also my advisory committee members Dr. Reed M. Gardner, Dr. Peter J. Haug, Dr. Charles W. Turner, and Dr. Michael J. Lincoln for their invaluable guidance.

Dr. Warner has shared with me his vast knowledge and experience, the sense of confidence, the pioneer spirit, the work ethics, and his generosity. He has given me enormous freedom to proceed with this study. Dr. Gardner provided me the motivation to pursue excellence and showed me the importance of looking at the problem from a different angle. Dr. Haug provided me with deep thoughts and helped me to build solid foundations for the research project. Dr. Turner provided me the power to deal with experimental design issues and data analysis. His expertise in psychology also helped me to discover my personal weaknesses and the strengths. Dr. Lincoln provided his medical expertise, numerous ideas, wise advice, and powerful computer.

I am very grateful for the financial supports from the National Library of Medicine (supported in part by NLM Grants 1R01-LM-04604 and 1R01-LM-052020). I am also very grateful to the financial support from the John Morgan Fellowship.

I wish to extend special appreciation to my parents for their constant support and encouragement throughout my entire educational life including the completion of this dissertation.

Special thanks go to my wife, Fuxiu. During the days when I worked like a restless robot, she provided the vital energy that kept the robot going. Her understanding, the comfort provided at home, her endurance of my late night work, and the sacrifice of her own study will always be a special memory in our life.

INTRODUCTION

Iliad System

Iliad is a diagnostic expert system for internal medicine, which represents the culmination of over 2 decades of expert systems research at the University of Utah. The system provides decision support and may be used as a teaching tool for medical students and practitioners. Iliad can run both under Macintosh systems and MS DOS-Windows. Iliad requires a 68030 Macintosh or 80386 SX DOS (or higher) processor with two megabytes of RAM memory. The system currently recognizes over 6300 disease manifestations and covers 1350 diseases and intermediate diagnoses in internal medicine.

The Iliad knowledge base is constructed using knowledge frames. Iliad uses both Bayesean and Boolean frames to describe diseases encountered in internal medicine. These frames permit use of sensitivities, specificities (in Bayesean frames) and rules (in Boolean frames) to describe the relationship of a disease to its manifestations and provide a basis for explaining Iliad's conclusions (1-2). The most common form of frames is the Bayesean frame. In a Bayesean frame, the relationship between a disease and related medical findings is represented by a True Positive Rate (TPR) and a False Positive Rate (FPR). In Boolean frames, Boolean logic is used to determine how close a disease is to true and how close the disease is to false. A heuristic algorithm allows Iliad to process incomplete information and to assign a pseudo-probability to a Boolean disease frame (3).

Iliad can be used either in consultation mode or simulation mode:

(1) Consultation mode. Iliad is designed to behave like an expert medical consultant. Iliad produces diagnostic suggestions based on medical findings presented by a particular patient. Iliad not only suggests all likely diagnoses at each stage of the work-up but also provides suggestions for further

requests for cost-effective data.

(2) Simulation mode. For this mode Iliad is designed to behave like a patient. A user is expected to ask Iliad relevant questions and order appropriate tests. Iliad can function as a testing simulator in which specific scoring algorithms have been designed to measure a user's performance in terms of pursuing a cost-effective work-up (4). The performance scores for inquiry skills are measured by comparing the questions asked by students to a best question calculated by Iliad's best information model.

Best Information Algorithm in Iliad

The basis for Iliad to generate cost-effective work-up advice in the consultation mode and to measure the appropriateness of a user's request for data in the simulation testing mode is the program's "best information" algorithm. The "best information" component in Iliad can be traced to a diagnostic system first tested in the early 1970s as a computerized interview tool. This system was further tested in the form of an interactive history collection program used directly by patients (5-7). The current system, Iliad, has been dramatically extended in terms of inference mechanism and knowledge base compared to the early system, and Iliad is still evolving. The best information algorithm evaluates the information content expected per dollar for uncollected data and selects the finding with the maximum information at the least cost. Iliad's dictionary stores the "cost" by using the actual dollar amount charged for each test at the University of Utah medical center; other medical centers may modify the charges in Iliad as needed. History findings are arbitrarily set to cost $1 and physical exam items $2. The procedure to calculate work-up suggestions can be summarized as below:

(1) Iliad selects a subset of the diagnostic hypotheses. If the user does not select a diagnosis to pursue, Iliad automatically selects a work-up suggestion for the most likely diseases, beginning with the top three diseases in the

differential diagnosis list.

(2) Iliad combines an information content calculation together with a cost factor to rank-order medical findings not yet known according to a cost-effectiveness algorithm in which cost-effectiveness is defined as information content/cost.

(3) Iliad then suggests the finding that is most informative at the least cost.

Enhancing the performance of Iliad's best information mode has been a continuous effort during the development of Iliad. The current algorithm was developed and refined to provide adequate results with reasonable computational speed. Currently, disagreements exist between Iliad and our medical experts concerning optimal strategies for data collection. In the past, the computational burden discouraged use of more complex cost-optimization algorithms. However, with the rapid development of more powerful hardware configurations on both PC and Macintosh machines, more computationally intensive algorithms were investigated and compared with the current algorithm. In this dissertation several utility models for determining cost-effectiveness of a medical finding and different ways of applying these models were studied. The research was intended to improve the current best information algorithm and thereby ensure that students who use Iliad receive accurate training. The research was also intended to explore techniques for analyzing the quality of expert advice suggested by Iliad.

Formal Decision Analysis

Decision making can be considered as a choice between actions or a choice of a course of action. Decision theory provides a formal, prescriptive framework for making logical choices under uncertainty. The main objectives of a decision analysis are, first, to provide models for describing our desires and our beliefs and, second, to use developed models to make rational decisions. Given a condition of a patient, there are different sequences of findings that might be pursued to reach a final diagnosis. The question then

is which sequence of findings to pursue to arrive at the correct diagnosis in the most cost-effective way. Formal decision analysis seems to be a natural choice to solve this problem. Formal decision analysis involves the following key steps:

(1) Generate a list of the possible actions or events. A decision tree is usually used to represent the course of actions. In the context of this study, possible actions were potential sequences of findings to reach a final diagnosis.

(2) Assign probabilities (objective or subjective) to each node of the action course. The probability assignments must be checked for internal consistency.

(3) Generate a utility model to describe each outcome state. The model can be represented by one single attribute, such as cost of reaching the final diagnosis, but it is usually described in terms of multiple attributes (e.g., time, risk, and cost) that are condensed into a single scale. The measurement of an attribute can be a hard measure, which is the result of controlled experiments, or a soft measure, which is the result of beliefs and personal experience about relevant events.

(4) Calculate the expected utility for each possible outcome. The objective is to choose an action which has the highest utility score. Usually a sensitivity analysis is performed to determine whether the best choice is robust or sensitive to reasonable variations.

Formal decision techniques have been applied in areas such as air-quality control, airport location, environmental and urban design, strategic business problems, and many others (8-9). However there are practical problems associated with applying decision trees and the expected value of utility algorithm. Frequently not every course of actions can be isolated and identified. The decision analysis may never be finished if every possible course of action must be analyzed. Quantitative knowledge of the utility of each correct and incorrect decision is often not available. It is also difficult to estimate utilities and backtrack

from all the final positions in a very large decision tree. In medicine, it is usually difficult for both physicians and patients to define what they want.

In formal decision analysis maximizing expected utility is the most common criterion used (10-12). However the expected utility criterion is not always appropriate to apply. A famous example is the St. Petersburg Paradox (12). Suppose there is game to play. A fair penny will be repeatedly tossed until the head appears. If the head occurs on the nth toss, the player will receive $\$2^n$. The question is that how much would the player be prepared to pay to play the game once? The probability of a fair penny first landing 'heads' on the nth toss is 0.5^n, so the expected payoff is

$$\sum_{n=1}^{\infty} (\frac{1}{2})^n . 2^n = \sum_{n=1}^{\infty} 1 = \infty \qquad (1\text{-}1)$$

The calculation means that the player is expected to win more no matter how much the player pays. Thus the player should be willing to risk everything to play the game just once. Yet, no one would consider such an action to be rational. This paradox can be solved by choosing a different utility measure, the logarithm of money and limiting a bound for the utility function. Actually one of the axioms in utility theory is that a utility function should be bounded so that no decision maker guided by the expected utility rule will reach the paradox conclusion. There are still debates about the validity of using expected utility criterion in decision analysis (11, 13-15). However, whether the use of formal decision analysis is valid or not is not the subject of this study. The goal of this study is to investigate different practical alternatives for measuring cost-effectiveness of a medical finding at a given stage in solving a diagnostic problem.

Quasi-Utility Concept

The traditional approach to analyzing a decision problem has been a foldback analysis in a decision tree. This starts with the end nodes of a decision tree and compares utility values at each decision node when folding back recursively toward the root. In terms of

medical diagnosis problems, there are many potential medical findings to confirm and exclude hypotheses being considered. The decision tree to depict all the possible sequence of medical findings to make a final diagnosis would be very large. It would become very difficult to obtain utility values for each scenario in a large decision tree from patients and doctors. Thus the traditional approach is computationally expensive and hard to manage when dealing with a medical diagnostic problem. To obtain a practical solution to this complex problem, certain assumptions have to be made. The most popular ones follow (16):

(1) No competition: Each information source (a medical finding in the context of this study) is evaluated in isolation, as if it were the only source available for the entire decision.

(2) One-step horizon: After obtaining the source with the highest utility value, a terminal decision is made. To make further decisions, the process must be repeated.

These assumptions result in a suboptimal myopic analysis; i.e., the analysis only looks one step ahead. In the context of this study, the physician is also assumed to be in the middle of a decision tree search and is not about to make an immediate decision on treatment. To overcome the difficulty in obtaining utility values for a decision tree, a substitute for utility can be used. If one assumes that the utility of a diagnostic decision can be measured in terms of reduction in uncertainty of a disease given a medical finding, then one has the substitutes for utility, called quasi-utilities to compare different medical findings. Quasi-utilities are introduced as compromises to evaluate potential actions one at a time instead of going through all possible combination of actions. Whenever a particular decision stage is analyzed using quasi-utilities, that stage acts approximately as if it is the final stage. This method makes it practical to make decisions by cutting down on the size of the search.

Different models to measure the reduction in uncertainty, also called information

content in the theoretic sense, can be derived from information theory and Bayes theorem. The derivation of these models will be discussed later in this chapter.

In this study, the quasi-utility for expressing cost-effectiveness is defined as the ratio of the quantity of uncertainty reduction (i.e., information content) and the charge for obtaining a medical finding. It always costs something to acquire a medical finding, such as time to wait for the result, effort, money, or risk. Because the purpose of this study was to investigate different characteristics of information content models, only the direct charges of laboratory test and other procedures were considered.

The charge for a laboratory test procedure and the cost of a laboratory test procedure are different. Costs are generated based on what hospitals pay to deliver the services; charges are what the patients, insurance companies, and Medicare pay for the services. The difference results from the complex reimbursement schemes that frequently underpay for some services and will overpay for others (17-18). There are other cost factors associated with obtaining a laboratory test or procedure result such as patient discomfort, test-related morbidity and test-related death, hospital waiting time for patient, etc. Obtaining the true costs of different services is not a trivial task. It is difficult to allocate some costs such as utility bills, ancillary services for each item or service, costs of test-related morbidity, and test-related death. Because there are hundreds of laboratory test procedures contained in Iliad, it would be extremely difficult and time consuming to have every procedure assigned a "true" cost value. Alternatively charges for services can be easily obtained from hospitals. The distinction between costs and charges is important, but quasi-utility models in this study only take into account actual charges for medical services. These models can be easily modified appropriately when the real cost data are available.

Information Theory

Information theory is the science of quantification, coding, and communication of information (19). The basic ideas were formulated back in the 1920s. Since the Second

World War modern information theory has had a considerable development due to the early work of Claude Shannon (20). The fundamental theorems from information theory are considered as the basis for information technology. The quantification of information in information theory is used to quantify the transmission errors due to accidental errors or random noises across communication channels, such as telephone, television, radio, and so on. Because information theory provides a formal mathematical basis for quantifying information and uncertainty, it is logical to use the method to measure diagnostic information provided by a medical finding.

The basic idea behind measuring the quantity of information in a signal or message is to measure the reduction of uncertainty. If a person is a male, and somebody tells him that he is a male, the message does not convey any information. If a message confirms a previously unlikely event, it contains information. A natural choice for quantifying information conveyed by a message is to take the difference of uncertainty before and after a message is received. One property of information suggested by Shannon is that information is additive. The total information conveyed by several messages is equal to the sum of the information conveyed by each of them. A natural choice to quantify uncertainty is the probability of an event. To satisfy the requirement of additivity, the logarithm of the probability was used by Shannon to quantify uncertainty. Quantity of information was expressed as $I = -\log_2 P$. The logarithm to the base two of the probability is commonly used to allow measurement of information in unit of "bit." The measure of average uncertainty is described as "entropy," which was chosen for its similarity to entropy in thermodynamics. The amount of heat from a chemical reaction can be measured by the change of thermo-entropy. The amount of information can be measured by the change of information-entropy. Entropy is defined as the average amount of uncertainty. In the context of medicine, the entropy can be used to represent the average amount of uncertainty as to whether a patient does or does not have a disease. The mathematical equation for the

entropy is

$$H(D) = -P(D)\log_2 P(D) - P(D-)\log_2 P(-D) \qquad (1\text{-}2)$$

Here, H(D) is the Shannon's uncertainty or entropy measured in "bits." P(D) is the probability that a patient has a disease D; P(D-) is the probability a patient does not have the disease D. If a medical finding's result is known as F, the information content provided by F is the difference in entropy before and after the finding F is known (based on the prior probability and posterior probability). If there is a perfect medical test that can change the probability of a disease to unity no matter what the prior probability of the disease and if the prior probability of the disease is 0.5, then the information content of the perfect test is 1 bit. This is because the uncertainty before the perfect test is 1 bit and the uncertainty after the test is 0 bit.

Quantity versus Quality about Information Theory

One shortcoming of information theory is that it only considers the quantity of a message without any regard for its meaning. For example, two messages may convey the same amount of information, but they can have significantly different meanings from a human point of view. If there are two tests, one test shows the diagnosis of lung cancer for patient A and another one shows the diagnosis of pneumonia for patient B, and if patient A and patient B have same prior probabilities of lung cancer and pneumonia respectively, the information contents of these two test are the same, but they really have different impact on patient A and patient B. In addition, the pain and risk associated with each test can not be reflected by the amount of information itself. The information value of a test does not carry semantic meaning of the circumstances being applied, because it is the same for every patient. This shortcoming may cause criticism about the validity of quasi-utility approach. However, information content value is not the only component to describe an objective or a goal. Multiple factors or attributes are needed in a model to reflect the real meaning of the quasi-utility (8-9, 21).

Derivation of Quasi-Utilities

The central feature of the quasi-utility models used in this study combined an information content calculation with a cost factor and then provided a rank-ordering of the alternative patient findings according to cost-effectiveness. The "information content" model in the current version of Iliad is

$$\text{Utility} = \frac{\text{prob} \times \text{information gain}}{\text{dollar charge of a finding}} \tag{1-3}$$

$$\text{information gain} = \max \{ (\frac{\text{sen}}{1 - \text{spec}}), (\frac{1 - \text{sen}}{\text{spec}}) \} \tag{1-4}$$

where prob = the prior probability of the frame being true (i.e., before obtaining the item of information), sen = true positive rate, spec = true negative rate. In terms of dollar charge of a finding, for history items the charge was set to an arbitrarily low value as $1, physical exam items $2; lab test procedures were the actual charges at the University of Utah Hospital Center.

Information gain is used interchangeably with information content in this dissertation. However, the author wants to make a finer point about these two terms. Information content in information theoretic sense is the uncertainty difference before and after a message is obtained. In the case of the current Iliad model, the information measurement is not based on the uncertainty difference. For the simplicity of discussion, the term "information content" will be used for all kinds of information measurement models.

The current Iliad model selects the finding with the maximum utility based on likelihood ratios and identifies the corresponding hypothesis for which the finding is relevant. The current algorithm was developed and refined to provide adequate results with reasonable computational speed. It has the advantage that the parameters required are easily accessible from Iliad's knowledge base. In addition, a minimum of calculation is required.

In this study, four new information content models are introduced (22-24). Two models are derived based on information theory, and two are based on ad hoc approaches. The derivation of the four new information content models is introduced next.

Two Information Content Models from Information Theory

One key assumption of information theory is that a message is not significant by itself. Rather the information in a message depends on the extent to which it resolves uncertainty. Another key assumption is that information conveyed by a series of messages is additive. Iliad's approach to the process of pursuing a group of diagnostic hypotheses is based on the assumption that the amount of diagnostic uncertainty in a case can be reduced by obtaining additional patient findings. The information provided by the patient findings can be measured quantitatively as the change in the level of uncertainty associated with a particular disease. One equation for measuring the uncertainty, H in "bits," is given by

$$H = -\log_2 P \tag{1-5}$$

This equation implies that uncertainty of disease is a function of the probability that the disease is present. Johnson used -H to represent the deficit of information (25). If the initial probability of a disease is P, the associated uncertainty is H bits of information. In order to conclude a disease, most of those H bits of uncertainty have to be removed. The information provided by a diagnostic observation or test is the difference between the diagnostic uncertainty on hand before the finding is known and the uncertainty after the finding is known. This difference (ΔH) can be expressed as

$$\Delta H = \log_2 P(D|F) - \log_2 P(D) \tag{1-6}$$

where P(D|F) is the posterior probability of the disease after the test is performed, and P(D) is the prior probability of the disease before the test is performed. If information content provided by a finding is always a positive number, then the information content I(F) provided by a finding whose result is F (either positive or negative) can be expressed as

$$I(F) = \text{abs}\ (\ \Delta H\ ) \tag{1-7}$$

Eq. [1-7] represents the "logP2-logP1" model, and the algorithm derived from it will be called the logP2-logP1 algorithm. This model calculates information contribution from the posterior disease probability and prior disease probability only. It does not take into account of the nondisease probabilities before and after a finding is known.

Another model from the information theory can be derived by using the average amount of uncertainty or "entropy" defined by Shannon. If there are a set of mutually exclusive probabilistic events (such as the presence and absence of a disease, or a group of mutually exclusive diseases) and if the events are denoted as D_i and the prior probability by $P(D_i)$, the uncertainty can be described by the following equation:

$$H(D) = - \sum_i P(D_i) \log_2 P(D_i) \qquad (1\text{-}8)$$

where H(D) represents the uncertainty or entropy and D is the set of D_i's. In this study D_i represents either the presence or absence of a disease. Eq. [1-8] becomes the same as eq. [1-2]. The information conveyed about D by F (either positive or negative):

$$I(F) = \text{abs}\ (H(D) - H(D|F)) \qquad (1\text{-}9)$$

The use of eq. [1-9] requires an appreciation of uncertainty as the function of the prior probability of disease. The standard Shannon model fails to capture reasonable intuitions about the quantity of information provided by a diagnostic finding (26). For example, when the prior and the posterior probabilities are complementary (e.g., the prior is 0.1, the posterior is 0.9), the finding provides no change in uncertainty, and thus no information has been conveyed. To overcome the limitations of eq. [1-8], the modified Shannon's information content model is used (27). It is known that the maximum uncertainty in a system exists when the probabilities of each possible event are equal. If the function H(D) passes through a maximum when the hypothesis moves from the prior state to the posterior state, the information contributed by finding F (either positive or negative) can be measured by

$$I(F) = (H_{max} - H(D)) + (H_{max} - H(D|F)) \qquad (1\text{-}10)$$

where H_{max} represents the value of H(D) at the maximum between the two states. If D represent the mutually exclusive probabilistic events: presence and absence of a disease and substituting disease status P(D) = 0.5 and nondisease status P(D-) = 0.5 in eq. [1-8], it can be seen that H_{max} = 1 (bit).

If the function H(D) does not pass through the maximum H_{max} due to the change of disease status (e.g., the prior and the posterior of the disease are both either greater than 0.5 or less than 0.5), I(F) still follows eq. [1-9]. Eq. [1-9] and [1-10] together serve as the second model of measuring information resulted from a finding F (either positive or negative). This approach is the "modified Shannon" model, and the corresponding algorithm will be called the modified Shannon algorithm.

Two Information Content Models from Ad Hoc Approach

The model called "linear information theory" measures information by linear change of the probability of disease (26). Warner used this approach to measure information content provided by different medical findings and to help selecting the next questions to present (28). The uncertainty representation here is the probability of a disease. The information content of a finding F (either positive or negative) can be expressed as

$$I(F) = \text{abs}\,(P(D|F) - P(D)) \qquad (1\text{-}11)$$

where P(D|F) is the posterior probability of disease D, P(D) is the prior probability of disease D. This third model for measuring information content is called "P2-P1" model, and the corresponding algorithm will be called the P2-P1 algorithm.

Another model is called "weight of evidence" (29-30). Medical experts do not always seek clinical data based on a global view of the differential diagnostic set. Under certain circumstances, they focus on a single diagnostic possibility and choose which information to seek in this context. Weight of evidence considers information in terms of its effect on the likelihood of a specific disease. The weight of evidence contributed by a finding F (either positive or negative) for a disease D is:

$$W(D|F) = \log_2 \frac{P(F|D)}{P(F|\overline{D})} \qquad (1\text{-}12)$$

where W(D|F) measures the contribution of a finding result F to the diagnosis of a specific disease D as opposed to the alternate condition $\overline{D}$. This model is called "LogLR" model, where LR stands for the likelihood ratio. The current Iliad model is actually a LR model. These two models ignore the prior probability of the disease and depend only on the sensitivity and specificity of a medical finding. Whether one uses the LogLR or the LR model should not make any difference in terms of ranking information value per se. However, the ranking in Iliad's best information algorithm is based on the information per dollar charge. The LogLR model is more sensitive to dollar charge because of the logarithm function. In addition, it should be noted that LogLR is additive and LR is not.

Expected Information Content Provided by a Finding

The derivation of different information content models above is all based on the assumption that the result of a finding F is known. A known result finding means (1) the result of presence or absence of a finding is known, such as cough. The result "cough present" or "cough absent" is a known result, (2) the value result of a finding is known such as CBC. The value result can be either normal or abnormal. The expected information content provided by a medical finding can be defined as the weighted average information content contributed by two possible states (e.g., present and absent). If the information provided by a positive finding is expressed as I(F+) and the information provided by a negative finding is I(F-), then the expected information provided by the finding is

$$\text{Expected } I(F) = P(F+)\, I(F+) + P(F-)\, I(F-) \qquad (1\text{-}13)$$

where P(F+) and P(F-) are the frequencies of the positive finding F+ and the negative finding F- respectively. I(F+) and I(F-) can be obtained by applying one of these information content models discussed above. The frequency of the positive finding F+ is

calculated as

$$P(F+) = P(F+|D)\,P(D) + P(F+|\overline{D})\,P(\overline{D}) \qquad (1\text{-}14)$$

P(F+) depends on the probability of the disease D, true positive rate of the finding F for D and the false positive rate of the finding F for D. Because the sum of P(F+) and P(F-) is unity, P(F-) = 1 - P(F+).

Strategies to Make Diagnostic Decisions Based on Quasi-Utilities

Decision Criteria

Selecting a most cost-effective medical finding to ask at any stage of work-up is the process of ranking each finding's cost-effectiveness based on their utility values. How are alternative findings evaluated? There are several methods to assign utility values to medical findings. The fundamental decision rule of formal decision analysis is the principle of maximizing expected utility. If the utility score of a finding's positive result is U(F+) and the utility score of a negative result is U(F-), the expected utility score of the finding is the weighted average of U(F+) and U(F-) based on frequency of the finding. There are other variations of expected utility criterion (31-33).

Maximax

This is an optimistic decision rule which uses maximum utility that can possibly be obtained. In the context of this study, the larger utility score between U(F+) and U(F-) will be assigned to represent the finding's cost-effectiveness to be compared with other findings.

Maximin

This criterion, in contrast, is a pessimistic rule which uses the utility value that can be assured first and then selects the maximum utility value among those assured values. This means that the less utility value between U(F+) and U(F-) will be assigned to represent the finding's cost-effectiveness to be compared with other findings.

Hruwicz Criterion

This decision rule is the combination of maximax and maximin criteria. It requires the decision maker to choose a "pessimism-optimism coefficient" and to compute a weighted combination of worst and best outcomes. The approach is not widely used because it is difficult to implement.

There are other criteria that can be used to describe decision making under risk such as minimax, minimin. In this study only two criteria were used and compared. One was the aggressive approach, maximax, which is currently being used by Iliad; another was the conservative approach, maximizing expected utility. The expected approach could be considered as a normative model, and the maximax approach could be considered as a descriptive model. There is a distinction between a descriptive and normative approach (32-33). The normative approach describes what a rational decision maker should do most of time. The descriptive approach describes how people in fact (rationally or not) make decisions. The two models are different, but they can agree with each other. For example, the normative model is widely accepted as a descriptive model to describe economic behavior (34-35).

Two Strategies

There are two ways to link medical findings and diseases:

(1) Single-frame strategy: consider findings within the context of each individual disease frame.

(2) Across-frame strategy: consider the information that a single finding could provide across several diseases.

Currently Iliad uses the single-frame strategy. One of this study's objectives was to consider a new strategy, the across-frame strategy, and compare the performance of different information content models under the two strategies.

Single-frame Strategy

This strategy evaluates the relative cost-effectiveness of each medical finding within each separate disease frame. Iliad is a frame-based expert system. Suppose there are findings F_1, F_2, F_3, ..., F_n contained in a disease frame shown in Table 1. The actual frame is usually more complex than one shown here. An Iliad's frame can contain findings, clusters, and other frames. The nested frame can be more than 10 levels deep. Assume all the findings listed in Table 1 are not yet answered. Quasi-utility values can be calculated for each finding's positive and negative results respectively based on cost factors and different information content models. Iliad currently uses the optimistic approach, maximax, to rank cost-effectiveness of these findings. Under the maximax approach, Iliad first compares the utility scores between the positive and negative results for each finding and chooses the larger one to represent the utility score for that particular finding; second Iliad compares all the better findings and selects the finding with the maximum utility score. If there are multiple diseases being considered, Iliad compares the best from each disease frame and chooses the best among the best ones. The procedure of applying the single-frame strategy under the expected utility rule is identical to the one described above.

Table 1. Summary of the Single-frame Strategy.
(Quasi-utility values, information content/cost, for
all findings in a disease frame)

Finding	positive result	negative result
F_1	u_{1+}	u_{1-}
F_2	u_{2+}	u_{2-}
.	.	.
.	.	.
.	.	.
F_n	u_{n+}	u_{n-}

Across-frame Strategy.

This strategy reflects the combined information provided by a medical finding to a group of hypotheses (see Table 2). It considers the information that a single finding could provide across several diseases. Suppose there are several hypotheses being considered, D_1, D_2, ..., D_m. All the medical findings not yet known in those disease frames are represented by F_1, F_2, ..., F_k. Positive results are represented by F_1^+, F_2^+,..., F_k^+. Negative results are represented by F_1^-, F_2^-,..., F_k^-. The ranking of cost-effectiveness of each finding is fundamentally different here. Suppose the decision rule is maximax. Take finding F_1 as an example, the utility value of F_1 to D_1 is represented by the larger value of the positive and negative utility scores, $\max(u_{11}, u_{11}')$. The total utility score of F_1 is the summation of utility values to D_1, D_2, ..., D_m. If F_1 has nothing to do with a particular disease, the score will be 0. Under this strategy, certain findings may not be cost-effective within a context of a single disease frame, but they may be most-effective overall in the context of all the hypotheses being considered.

Table 2. Summary of the Across-frame Strategy.
(Quasi-utility values, information content/cost, for all findings in multiple disease frames)

positive result					negative result				
	D_1	D_2	...	D_m		D_1	D_2	...	D_m
F_1^+	u_{11}	u_{12}	...	u_{1m}	F_1^-	u'_{11}	u'_{12}	...	u'_{1m}
F_2^+	u_{21}	u_{22}	...	u_{2m}	F_2^-	u'_{21}	u'_{22}	...	u'_{2m}
. . .		. . .		. . .	. . .				. . .
F_k^+	u_{k1}	u_{k2}	...	u_{km}	F_k^-	u'_{k1}	u'_{k2}	...	u'_{km}

An Example to Apply Two Strategies

Suppose hypotheses under consideration are pulmonary embolus and atypical pneumonia. Also suppose three unanswered questions exist for each disease frame, as shown in Table 3.

The approach for this scenario by the single-frame strategy is as follows:

(1) Calculate information content per dollar of each finding within each disease frame.

(2) Rank the cost-effectiveness of each finding in the Atypical Pneumonia frame and the Pulmonary Embolus frame separately.

(3) Select the finding that receives the highest of the six scores.

The approach by the across-frame strategy is as follows:

(1) Calculate information content per dollar of each finding for each disease frame.

(2) Sum across the information per dollar for the "common" findings, respiratory rate, and the chest x-ray, across the two hypotheses.

(3) Select the finding that receives the highest of four scores.

Table 3. A Scenario of Considering Two Hypotheses under Two Strategies

Atypical Pneumonia	Pulmonary Embolus
present history: cough with purulent sputum	present history: cough with gross hemoptysis
vital signs: respiratory rate	vital signs: respiratory rate
chest x-ray shows alveolar infiltrate	chest x-ray shows alveolar infiltrate

Significance of the Study

The quasi-utility model in this study represents an estimate of the information per dollar available for each medical finding. The critical part of the model is the measure of information contribution provided by a finding. By investigating the performance of different information content models, it was expected to find a best information content model that would improve the behavior of Iliad's consultation mode and make Iliad a more accurate model of expert behavior. Application of a well-developed best information algorithm has at least the following potentials:

(1) allow Iliad's test mode to better discriminate among students and physicians with different levels of expertise,

(2) improve the quality of education with the use of Iliad through providing high quality advice,

(3) serve as a diagnostic assistant to physicians,

(4) serve as a tool to measure cost-effectiveness of physician's work-up retrospectively.

The study of characteristics of different information content models will help to design a model that can incorporate specific needs of Iliad at different stages of work-up. Knowing why different information content models are different at different stages of a work-up will serve as a guide to identify advantages and disadvantages of potential information content models and therefore to predict behaviors of those models. Early rejections of inferior approaches could save time and effort for later improvement of Iliad's best information algorithm. The investigation of strategies to apply the quasi-utility model in diagnostic problems can further create a new dimension to improve the performance of Iliad's best information algorithm.

This study will also help to identify the relationship between different information content models and the cost factor. Some quasi-utility models may be more sensitive to the cost factor, and some may be less sensitive.

The performance of an information content model depends not only on the characteristics of the model itself but also on Iliad's ability to generate accurate differential diagnoses. The demonstration of usefulness of Iliad's best information mode may help drive further efforts to improve Iliad's diagnostic engine.

MATHEMATICAL METHODS

Probabilistic Inference by Iliad

The data elements in an Iliad frame are divided into the following categories: (1) findings with yes or no answers ("shortness of breath," "chest pain") or with value answers (e.g., heart rate, white blood count), (2) intermediate concepts that are themselves represented by frames ("sympathetic reaction to stress," "hypovolemia"), (3) a disease frame that can help rule in or rule out the top level disease. There are three kinds of frames in Iliad knowledge base, Bayesean frame, Boolean frame, and Value frame. Figure 1, 2 and 3 show examples of each these frames. A top-level diagnostic disease is represented by either a Bayesean frame or a Boolean (also called deterministic or rule-based) frames. The value frame is specifically designed to represent a certain concept or a value that requires multiple elements to determine. A value frame has only two states, true or false (see examples in Figure 3). A value frame can contribute information to the upper level frame only if each element in the frame is known. In contrast, both Bayesean and Boolean frames can be partially true or false depending available information. The posterior probability is used to represent the likelihood status of a Bayesean frame. A pair of pseudo-probabilities generated by a heuristic algorithm is used to represent the likelihood status of a Boolean frame. Iliad's ability to generate a differential diagnosis list (a probability list of diseases) given patient findings is the basis for generating cost-effective suggestions, because the best information algorithm requires the updated differential diagnoses to identify most likely diseases and also to determine partial information provided by a finding to a disease through one or more intermediate disease frames.Understanding the probability inference mechanism of Iliad is the first step to

Disease: ARDS (adult respiratory distress syndrome)
• Prevalence: 1 in 1,000 for statistics in "Iliad's inpatient" a priori table.
• Posterior Probability: 0.0

FINDINGS:	•Status •Cost	•TPR • FPR	•LR+ •LR-
• Blood gas pattern of ARDS		0.99 0.001	989 (99.9)
--- or ---			
• Hypoxemia		0.99 0.10	9.90 (90.0)
Vital signs: respiratory rate is === per minute (12 - 20)			
0 -- 21		0.10 0.90	(90.0) (n/a)
21 -- 41		0.24 0.05	4.80 (n/a)
> 41		0.75 0.001	749 (n/a)
Chest auscultation: crackles that are fine rales (velcro) diffuse and bilateral (9.90)		0.90	90.0 0.01
• Minimally scattered consolidation by CXR		0.10 0.005	20.0 (1.10)
--- or ---			
• Diffuse bilateral consolidation by CXR		0.90 0.04	22.5 (9.60)
• Risk factors for ARDS		0.99 0.05	19.8 (95.0)
History of present illness: shortness of breath (dyspnea) at rest with recent onset		0.95 0.05	19.0 (19.0)
Swan-Ganz (right heart catheterization)		0.001	(25.0)
pulmonary capillary wedge pressure is === mmHg, mean (5 - 13) if > 15	$650	0.025	1.02

TPR (True Positive Rate) = Sensitivity; FPR (False Positive Rate) = 1 - Specificity;
Cost = hospital charge (in most cases); (n/a) = not applicable;
LR+ = likelihood ratio for the positive result; LR- = likelihood ratio for the negative result;
• = the following item is a nested frame.

Figure 1. A sample Bayesean frame.

Disease: Blood gas pattern of ARDS
• Closeness to true: 0.0; •Closeness to false: 0.0

FINDINGS:	•Status	•Frequency	•Cost
A. Blood gases on high-flow (> 5 L/min) supplemental oxygen arterial pO2 is === mmHg (>80)		0.25	$163
B. Blood gases arterial pCO2 is === mmHg (35 - 42)		0.025	$51
C. Blood gases arterial pO2 is === mmHg (60 - 80)		0.025	$51

True if: A <= 50 and B < 40 and C < 50

Figure 2. A sample Boolean frame.

Disease: Predicted minus measured DLCO (mL/min/mmHg)
Value is: 0.0

FINDINGS:	•Status	•Frequency	•Cost
A. •Predicted DLCO (mL/min/mmHg)		0.025	
B. Lung diffusion testing single breath DLCO is === mL/min/mmHg (> 30)		0.10	$28

Value if: (A > 0 and B > 0) then A - B

Disease: Predicted DLCO (mL/min/mmHg)
Value is: 0.0

FINDINGS:	•Status	•Frequency	•Cost
A. General appearance: Height is === inches (59 - 79)		0.025	
B. General information: Age is === years		0.026	
C. General information: Sex: male		0.542	

Value if: (C) then 1.04 * A - 0.21 * B - 26.31

Figure 3. Sample value frames.

implement different quasi-utility models into Iliad.

Bayesean Frame Inference

If a test finding is denoted as F (presence or absence), the presence of disease is D; the absence of disease is D-. The joint probability of disease present and test is

$$P[F, D] = P[F|D]\ P[D] \tag{2-1}$$

P[F|D] is the probability of F given D. The joint probability of disease absent and test is

$$P[F, D-] = P[F|D-]\ P[D-] \tag{2-2}$$

P[F|D-] is the probability of F given D-. The probability of F can be expressed as

$$P[F] = P[F, D] + P[F, D-] \tag{2-3}$$

Let P[D|F] denote the conditional probability of disease given the test finding F as P[D|F], the Bayes' Theorem is expressed as

$$P[D|F] = \frac{P[D]\ P[F|D]}{P[D]\ P[F|D] + P[D-]\ P[F|D-]} \tag{2-4}$$

If F is positive, P[F|D] is the true positive rate, sensitivity; P[F|-D] is the false positive rate, 1- specificity. If F is negative, P[F|D] is the false negative rate, 1-sensitivity; P[F|-D] is the true negative rate, specificity. Based on eq. [2-4], the likelihood ratio LR for F is

$$LR = \frac{P[F|D]}{P[F|D-]} \tag{2-5}$$

If F is positive, LR is the ratio of true positive rate to the false positive rate, LR+ = sensitivity / (1 - specificity). If F is negative, LR is the ratio of false negative rate to the true negative rate, LR- = (1 - sensitivity) / specificity.

Eq. [2-4] is used to revise the posterior probability of disease given a finding. Iliad uses the equation sequentially to update the posterior probability of disease given multiple findings.

Data elements in a Bayesean frame consist of not only findings but also nested frames. Their status are represented by probability (Bayesean frames) or a pair of pseudo-probability (Boolean frames). To update the top-level disease probability given partially

true or false status of a nested frame, a modified version of Bayes' formula is used by Iliad:

$$P[D|\text{Frame Status } (a, b)] = \frac{P[D]\,(sen)^a\,(1\text{-}sen)^b}{P[D]\,(sen)^a\,(1\text{-}sen)^b + (1 - P[D])\,(1\text{-}spec)^a\,(spec)^b} \qquad (2\text{-}6)$$

Frame status (a, b) denotes that the positive status of the frame is "a" and the negative status of the frame is "b." For a nested Bayesean frame, if AP is the a priori probability of the frame and P is the posterior probability of the frame after a finding is known, then a and b are calculated as follows (3):

$$\text{if } P > AP \qquad a = (P - AP) / (1 - AP),\ b = 0$$

$$\text{if } P < AP \qquad a = 0,\ b = (AP - P) / AP$$

Note that AP is the a priori probability of a Bayesean frame; in almost all Bayesean frames, AP is close to 0. Other notations are sen = sensitivity of the nested frame to the upper level frame and spec = specificity of the nested frame to the upper level frame. For example, the answer of a finding in a nested frame revises the status of the nested frame to (a, b); the top level disease probability is updated by eq. [2-6]. This revision of probability is usually done sequentially because disease frames in Iliad knowledge base often have multiple levels. Currently Iliad assigns the square root of a and the square root of b as the exponents to adjust the dependence of P[D|Frame Status] on sensitivity and specificity to behave closer to expert performance as new information is propagated to a top-level frame.

Boolean Frame Inference

A Boolean frame is designed as a decision module built around a Boolean relationship among its findings. Any one or some combination of findings in the frame may be sufficient for the frame to come true or false. When there is not enough information to make the frame true, a pair of pseudo-probability numbers, close_true and close_false, is used to express the true state and false state of the Boolean frame respectively. For example, if the logic of a Boolean frame is "true if 3 of (A, B, C, D) are true," and assume A and B are true, C is false, D is unknown, and each item in the logic has the same

frequency. If only one item is true and other items are unknown in the logic, there is 1/3 of what is needed to be true. Since A and B are true, close_true = 0.67 (2/3). The negative logic (derived from the "true" logic) in this example is "false if 2 of (A, B, C, D) are false." The calculation of close_false is similar to close_true. Here C is false; there is 1/2 of what is needed to be false, so close_false = 0.50.

When items in the logic have different frequencies, normalization is needed to calculate close_true and close_false. For example, if the frequency of A and B is 0.5 and the frequency of C and D is 0.25, A and B both carry weighting factor of 2 (1/0.5), and C and D both carry weighting factor of 4 (1/0.25). If only A and B are true, the positive status of the Boolean frame is 0.50 (4/8).

A more detailed discussion about Boolean frame inference was given by Sorenson (3).

The Current Best Information Searching Algorithm

The critical part of searching the most cost-effective finding to acquire next is to determine the information content of the potential findings. The cost factor of each finding is fixed and readily available from the Iliad knowledge base. The current best information searching algorithm is a fairly complex recursive process. The idea is to find the most cost-effective finding based on the maximax criterion. The algorithm is described as follows:

(1) Identify the hypothesis being pursued and determine the type of the disease frame being considered. If the disease frame is a Baysean frame, Iliad compares the positive likelihood ratio to the negative likelihood ratio for each unanswered item contained in the top-level frame and uses the larger likelihood ratio value as the information content of that item. The word "item" here represents either a direct observation or test result or the result of processing a nested frame. If an item is a nested frame, the probability or pseudo-probability of 0.99 is the threshold to determine whether the item is answered or not and the charge of the item is set to be $1 temporarily. No calculation

is done for items contained in nested frames at this point. If the disease frame being considered is a Boolean frame, Iliad determines the information for each item as follows:

$$\text{Info\_true} = (\text{close\_true after item} - \text{close\_true before item})/(\text{close\_true before item}) \quad (2\text{-}7)$$

or

$$\text{Info\_false} = (\text{close\_false after item} - \text{close\_false before item})/(\text{close\_false before item}) \quad (2\text{-}8)$$

Info_true and Info_false are information contents provided by true and false result of an item respectively; close_true and close_false are the positive and negative status of the Boolean frame, and they can be generated by Iliad based on the Boolean frame inference. Again Iliad uses the larger value of Info_true and Info_false as the information content for each item in a Boolean frame; no calculation is done for items contained in nested frames at this point.

(2) Determine the utility scores of each item in the top-level disease frame and rank all items in the top-level disease frame. Identify the best item in the top-level disease frame. If the top item is a normal finding (not a nested frame), then stop the search. If the top item is a nested frame, then go down the ranking list until a normal finding is found and record the utility score of the finding. This finding will be referred as a "recorded finding." Each nested frame on the top of the ranking list (higher than the recorded finding) is referred as "recorded nested frame." If every item in the ranking list is a nested frame, then record nothing.

(3) Find the "best" item in the first recorded nested frame by repeating step 1 and determine the utility score of this "best" item to the top-level disease frame. The utility score is calculated based on partial information provided by the item to the top-level disease frame and the charge of the item. The method of determining partial information will be discussed in the next section.

(4) Compare the best item from the first nested frame to the recorded finding in the top-

level frame. If the best item in the first nested frame is better than the recorded finding, then check whether this best item is still a nested frame. If this best item is a normal finding, then compare it to the recorded finding and select a better one as the new recorded finding. If this best item is still a nested frame, then decide whether to continue the calculation. To justify further calculation, the utility score of this best item has to be greater than the score of recorded finding; otherwise just stop the calculation, go to the second recorded nested frame and repeat step 3 and 4.

(5) Go through the above steps recursively until all the recorded nested frames are processed. Because the recorded finding is always replaced by the best finding available, the final recorded finding is the most cost-effective finding in the top-level disease frame.

(6) Repeat step 1 to step 5 for each hypothesis being considered. The utility score of the best finding in each hypothesis frame is adjusted by multiplying the probability or pseudo-probability of that disease. After adjusting utility scores for each "best" finding for each hypothesis being considered, the finding with the maximum adjusted utility score is chosen as the best information to acquire next. The corresponding hypothesis is also identified.

(7) Suggest the hypothesis to pursue and the best finding to acquire next. Iliad usually suggests multiple "best findings" from the identified hypothesis disease frame. This is because that any finding whose utility score is within certain range (current range: 50%) of the maximum score is suggested together with the best item to make more alternatives available.

The advantage of the current searching method is its fast computational speed, because the method eliminates unnecessary calculations for findings whose utility scores are less competitive. The current searching method is specifically designed for the LR (likelihood ratio) model under the single-frame strategy. However there is a danger to ignore a potentially overall best finding when the objective is to find a best finding that

provides the maximum information to multiple diseases at the least cost (the across-frame strategy). The overall best finding for a group of diseases may not be the best finding for each individual disease. To overcome the limitation of the current searching method, a more generic searching method was designed and will be discussed later in this chapter.

Partial Information Passing Methods

Iliad uses frames to represent the relationship among findings and diseases. The linkage between a finding and a disease is usually through one or more intermediate frames. To obtain information content provided by a finding in a intermediate frame, Iliad currently uses the following method to achieve hierarchical propagation of information content.

If a top level decision frame is a Bayesean frame, information content is measured by likelihood ratio. Assume the disease frame is "Frame A." A finding "F" is in a intermediate frame "Frame B." The linkage between F and Frame A is F -> FrameB -> Frame A. The linkage between Frame B and Frame A in terms of likelihood ratio is already known by Iliad; let the positive likelihood ratio (given Frame B is true) be $LR^{+}_{B->A}$, and let the negative likelihood ratio (given Frame B is false) be $LR^{-}_{B->A}$. The question is how much likelihood ratio value finding F contributes to Frame A. The following rules are currently used to determine the partial information $LR^{+}_{F->A}$ (if F is positive) and $LR^{-}_{F->A}$ (if F is negative),

If Frame B is a Baysean frame:

if (close_true[B|F+] > close_true[B])

$$LR^{+}_{F->A} = (close\_true[B|F+] - close\_true[B]) \times LR^{+}_{B->A}/(1 - close\_true[B])$$

else

$$LR^{+}_{F->A} = 0$$

if (close_true[B] > close_true [B|F-])

$$LR^{-}_{F->A} = (close\_true[B] - close\_true\ [B|F-]) \times LR^{-}_{B->A}/(close\_true[B]\ )$$

else

$LR^{-}_{F->A} = 0$

If Frame B is a Boolean frame:

if (close_true[B|F+] > close_true[B])

$LR^{+}_{F->A}$ = (close_true[B|F+] - close_true[B]) × $LR^{+}_{B->A}$/(1 - close_true[B])

else

$LR^{+}_{F->A} = 0$

if (close_false[B|F-] > close_false [B])

$LR^{-}_{F->A}$ = (close_false[B|F-] - close_false [B]) × $LR^{-}_{B->A}$/(1 - close_false[B])

else

$LR^{-}_{F->A} = 0$

Note that only close_true is used to describe a Bayesean frame status, and close_false is not needed; close_true for a Bayesean frame is defined as

If (P > AP) close_true = (P - AP) / (1 - AP)

If (P < AP) close_true = 0

P represents the posterior probability of a Bayesean frame after a finding is known; AP represents the a priori probability. Note AP is not the prevalent probability of the Bayesean frame before the finding is known. For Boolean frames, close_true and close_false are generated by Iliad Boolean logic inference. In the rules of partial information passing described above, close_true[B|F+] and close_true [B|F-] represent the true status of Frame B given F is positive and negative respectively; close_true[B] is the true status of Frame B before the result of F (either positive or negative) is known; close_false[B|F-] represents the false status of Frame B(in case Frame B is a Boolean frame) given F is negative; close_false[B] represents the false status of Frame B (in case Frame B is a Boolean frame) before the result of F is obtained.

If a top-level decision frame is a Boolean frame, partial information content is

measured by eq. [2-7] and [2-8].

This current method of partial information passing was developed empirically. In the case that a finding links to a Bayesean frame through an intermediate frame, the assumption is that "equivalent likelihood ratio" (partial information) provided by the finding is proportional to the finding's ability to make the intermediate frame closer to be true or closer to be false. In the case that a finding links to a Boolean frame through an intermediate frame, the partial information is measured by the finding's ability making the top-level disease frame closer to be true or closer to be false. If there are multiple intermediate frames, Iliad determines partial information by using the modified Bayes formula (eq. [2-6]) or the Boolean inference logic inference sequentially and then uses the rules of partial information passing method described above.

The current partial information passing method is designed for pursuing best information under single-frame strategy. It is not suitable for pursuing best information under across-frame strategy, because the current method uses two utility scales to measure information (likelihood ratio in a Bayesean frame, how close to being true or false in a Boolean frame) and it is counter intuitive to sum across two different information measurement scales together. Even though there are few top level Boolean disease frames in the Iliad knowledge base, a more generic method of passing information is needed. The new method to pass information to both Baysean and Boolean frames is therefore developed and will be described next. For the convenience of discussion, the method described above will be referred as "partial information passing method I."

The new method of passing information is based on the mathematical definition of likelihood ratio. Given any finding in a disease frame, Iliad can generate a posterior probability or a pseudo-probability for the disease. Based on Bayes equation, the likelihood ratio can be calculated from the prior probability before the finding is obtained and the posterior probability after the finding is obtained. If a finding does not link to the top disease frame directly, the calculated likelihood ratios (LR+ and LR-) can be considered

as "equivalent" likelihood values. The new method of determining partial information provided by a finding to a top disease frame through one or more intermediate frames can be described as the following equations:

If top disease frame is a Bayesean frame,

$$LR+ = \frac{1 - P(D)}{P(D)} \frac{P(D|F+)}{1 - P(D|F+)} \tag{2-9}$$

$$LR- = \frac{1 - P(D)}{P(D)} \frac{P(D|F-)}{1 - P(D|F-)} \tag{2-10}$$

If top disease frame is a Boolean frame,

if [close_true (D|F+) > close_true (D)] or [close_true (D|F+) < close_true (D)]

$$LR+ = \frac{1 - close\_true\ (D)}{close\_true\ (D)} \frac{close\_true\ (D|F+)}{1 - close\_true\ (D|F+)} \tag{2-11}$$

$$LR- = \frac{1 - close\_false\ (D)}{close\_false\ (D)} \frac{close\_false(D|F-)}{1 - close\_false\ (D|F-)} \tag{2-12}$$

else if [close_true (D|F-) > close_true (D)] or [close_true (D|F-) < close_true (D)]

$$LR+ = \frac{1 - close\_true\ (D)}{close\_true\ (D)} \frac{close\_true\ (D|F-)}{1 - close\_true\ (D|F-)} \tag{2-13}$$

$$LR- = \frac{1 - close\_false\ (D)}{close\_false\ (D)} \frac{close\_false\ (D|F+)}{1 - close\_false\ (D|F+)} \tag{2-14}$$

else LR+ = LR- = 0

When a top-level disease frame is a Baysean frame, P(D) is the prior probability of the top-level disease before finding F is obtained. P(D|F+) and P(D|F-) are the posterior probabilities of the top-level disease given finding F is positive and negative respectively.

When a top-level disease frame is a Boolean frame, a pair of pseudoprobability (close_true(D), close_false(D)) is used to describe the prior status of the disease. If the positive status of the disease D changes given F is positive, eq. [2-11] and [2-12] indicate that the pseudolikelihood ratio can be calculated by close_true(D) and close_true(D|F+) and the pseudolikelihood ratio for the negative finding can be calculated by close_false(D) and close_false(D|F-). If the true status of the disease D changes given F is negative, the

derivation is similar to the one if F is positive. If the true status of the disease D does not change regardless of positive F and negative F, pseudolikelihood ratio values for both F+ and F- are zero.

This new partial information passing method will be referred as "partial information passing mechanism II." Both partial information passing mechanisms are designed for implementing LR and Log LR information content models and both mechanisms use approximate measures to estimated likelihood ratio values for findings that are not directly linked to the top disease frame. The partial information passing mechanism I is acceptable for LR and LogLR models under the single-frame strategy, but partial information passing mechanism II is more intuitive to be used for LR and LogLR models under the across-frame strategy. In the next chapter (EXPERIMENTS AND RESULTS) performance of different information content models by using both information passing methods will be presented.

Expected Quasi-Utility Value from a Finding

Let F represent the finding, F+ and F- represent the positive and negative result of F respectively. Let D represent the top-level disease. A general procedure of calculating expected quasi-utility value provided by F to D (either a Bayesean or Boolean disease frame) can be described as follows,

if D is a Bayesean frame

(1) calculate P(D|F+) and P(D|F-),

(2) back calculate P(F+|D) and P(F+|D-) by using Bayes' equation,

(3) calculate P(F+) and P(F-) by using P(F+|D), P(F+|D-) and P(D),

(4) calculate quasi-utility values, U(F+) and U(F-), provided by both F+ and F-,

(5) calculate expected quasi-utility value for F: P(F+)U(F+) + P(F-)U(F-).

if D is a Boolean frame

(1) calculate quasi-utility values, U(F+) and U(F-), provided by both F+ and F-,

(2) calculate expected quasi-utility value for F: (U(F+) + U(F-))/2.

When D is a Bayesean frame, P(F+) and P(F-) are calculated based on back-calculated sensitivity and specificity data plus the probability of D. This back-calculation can provide frequency estimates for findings that link either directly to D or through one or more intermediate frames. When D is a Boolean frame, proper values for P(F+) and P(F-) are very difficult to obtain, P(F+) and P(F-) are assumed to be 0.5 each. The expected quasi-utility value for F is the average of U(F+) and U(F-). The methods of obtaining quasi-utility values will be discussed next.

Methods of Calculating Quasi-utility Values from Different Information Content Models

The quasi-utility to measure cost-effectiveness of a medical finding in this study is the information content per dollar charge provided by the finding. There are different models of measuring information content. The following example will demonstrate methods of calculating quasi-utility values based on different information content models.

Take two findings, F1 and F2, in Aortic Dissection Frame, for example:

F1: Aortic angiography shows signs of aortal dissection or aneurysm. F2: Echocardiography shows aortic dissection. If D represents Aortic Dissection, the data below are obtained from the current Iliad knowledge base,

Finding	Sensitivity (F+|D)	1 - Specificity (F+|D-)	Charge($)
F1	0.95	0.01	1500
F2	0.70	0.03	390

Assume prior probability of D is 0.4, P(D) = 0.4. The question is what quasi-utility value each finding should be assigned at this particular stage of work-up.

Quasi-utility Values Based on LR and LogLR Models

For finding F1, the positive likelihood ratio LR+ = 0.95/0.01 = 95; the negative likelihood ratio LR- = (1-0.95)/(1-0.01) = 0.0505. For finding F2, the positive likelihood

ratio LR+ = 0.70/0.03 = 23.3; the negative likelihood ratio LR- = (1-0.7)/(1-0.03) = 0.309. Let I(F+) and I(F-) represent information content provided by a positive and a negative finding respectively. If LR model is used, I(F1+) is 95, I(F1-) is 1/0.505 = 19.8 (Iliad uses the reciprocal value if a likelihood ratio is less than one), I(F2+) is 23.3, I(F2-) is 1/0.309 = 3.24. If LogLR model is used, the information content is the logarithm (base 2) of likelihood ratio. Table 4 gives the quasi-utility (information content/cost) values.

Under the maximax decision criterion, if LR model is used, Iliad assigns quasi-utility value 0.0633 to F1 and assigns 0.0597 to F2; if LogLR model is used, Iliad assigns quasi-utility value 0.00438 to F1 and assigns 0.0116 to F2. Note that F1 is considered to be more cost-effective than F2 under LR model and F2 is considered to be more cost-effective than F1 under LogLR model. This is because LogLR model is more sensitive to the cost factor.

Under the expected utility decision criterion, the quasi-utility value of a finding is the weighted average of quasi-utility values from both the positive finding and the negative finding. Therefore a positive finding and a negative finding frequencies are needed to calculate the expected quasi-utility of a finding. In this example, let P(F1+) and P(F1-) be the frequency of finding F1 given the positive result and the negative result respectively,

$$P(F1+) = P(F1+|D)\,P(D) + P(F1+|D-)\,P(D-) = 0.95 \times 0.4 + 0.01 \times 0.6 = 0.386$$

$$P(F1-) = 1 - P(F1+) = 1 - 0.386 = 0.614$$

Table 4. Quasi-Utility Values Calculated Based on LR and LogLR Models

Finding Result	LR model	LogLR model
F1 positive	0.0633	0.00438
F2 positive	0.0597	0.0116
F1 negative	0.0132	0.00287
F2 negative	0.00831	0.00435

The expected quasi-utility value of F1 based on LR model is: 0.386×0.0633 + 0.614×0.0132 = 0.0325. The expected quasi-utility of F1 based on LogLR model is 0.00345. By similar calculations, the expected quasi-utility value of F2 based on LR model is 0.0281 and the expected quasi-utility value of F2 based on LogLR model is 0.00715. Note that under expected utility decision criterion, F1 is considered to be more cost-effective than F2 under the LR model, F2 is considered to be more cost-effective than F1 under the LogLR model.

Quasi-Utility Values Based on Modified Shannon, P2-P1 and LogP2-LogP1 Models

Modified Shannon, P2-P1 and LogP2-LogP1 models measure information provided by a finding based on different models of representing uncertainty. All these three models depend on prior probability and posterior probability of the disease being considered. Because the charge for each finding is readily available in Iliad knowledge base, for the simplicity of demonstration, only information content values are obtained in the following calculations. Take the same example described above; the prior probability of the disease is 0.4, the posterior probabilities for various results F1+, F1-, F2+, and F2- can be calculated by the Bayes' equation. P(D|F1+) = 0.98, P(D|F1-) = 0.033; P(D|F2+) = 0.94, P(D|F2-) = 0.17.

Modified Shannon Model

The equation for uncertainty representation can be expressed by the following equation:

if $P(D) \leq 0.5$

$$H(D) = -P(D)\log_2 P(D) - P(D-)\log_2 P(D-) \quad (2\text{-}15)$$

if $P(D) > 0.5$

$$H(D) = 2 + P(D)\log_2 P(D) + P(D-)\log_2 P(D-) \quad (2\text{-}16)$$

The above equations are equivalent to the equations introduced in the first chapter. Note that eq. [2-16] gives the entropy value if the entropy curve described by eq. [2-15] flips upward when P(D) > 0.5. Here, P(D) = 0.4, H(D) = 0.971; P(D|F1+) = 0.98, H(D|F1+) = 1.859; P(D|F1-) = 0.033, H(D|F1-) = 0.209. For F1, the information content of positive F1 is H(D|F1+) - H(D) = 1.859 - 0.971 = 0.888, the information content of negative F1 is |H(D|F1-) - H(D)| = |0.209 - 0.971| = 0.762. The frequencies of F1+ and F1- are 0.386 and 0.614. The expected information content of F1 is 0.386×0.888 + 0.614×0.762 = 0.811. By similar calculations, P(D|F2+) = 0.940, H(D|F2+) = 1.673; P(D|F2-) = 0.171, H(D|F2-) = 0.660; P(F2+) = 0.298, P(F2-) = 0.702. The information content of positive F2 is 0.702, the information content of negative F2 is 0.311, and the expected information content of F2 is 0.428.

P2-P1 Model

For finding F1, the information content of positive F1 is |P(D|F+) - P(D)| = |0.98 - 0.4| = 0.580, the information content of negative F1 is |P(D|F-) - P(D)| = |0.033 - 0.4| = 0.367, and the expected information content of F1 is 0.386×0.580 + 0.614× 0.367 = 0.449. For finding F2, the information content of positive F2 is |P(D|F2+) - P(D)| = | 0.94 - 0.4 | = 0.54, the information content of negative F2 is |P(D|F2-) - P(D)| = |0.171 - 0.4| = 0.229, and the expected information content of F2 is 0.298×0.54 + 0.702×0.229 = 0.322.

LogP2-LogP1 Model

For finding F1, the information content of positive F1 is $|\log_2 P(D|F1+) - \log_2 P(D)| = |\log_2 0.98 - \log_2 0.4| = 1.293$, the information content of negative F1 is $|\log_2 P(D|F1-) - \log_2 P(D)| = |\log_2 0.033 - \log_2 0.4| = 3.599$, and the expected information content of F1 is 0.386×1.293 + 0.614× 3.599 = 2.709. For finding F2, the information content of positive F2 is $|\log_2 P(D|F2+) - \log_2 P(D)| = |\log_2 0.94 - \log_2 0.4| = 1.233$, the information content of negative F2 is $|\log_2 P(D|F2-) - \log_2 P(D)| = |\log_2 0.171 - \log_2 0.4| = 1.226$, and the expected information content of F2 is 0.298×1.233+ 0.702×1.226 = 1.228.

Information Provided by a Finding to Multiple Diseases

There are two strategies to calculate information provided by a finding, a single-frame strategy and an across-frame strategy. The difference between the two strategies is whether the link between a finding and several hypotheses being considered is treated separately or simultaneously. Assume there are two hypotheses being considered, pneumonia and acute bronchitis. The probability of pneumonia is 0.4, the probability of acute bronchitis is 0.3. How much information does the finding "chest auscultation: crackles" provide according to single-frame strategy and across-frame strategy? The relationship among the finding and diseases is shown in Figure 4. For the purpose of demonstration, only the P2-P1 information content model and the expected utility decision criterion are applied to demonstrate the method of obtaining information content values for the finding F (crackles) under both strategies. Quasi-utility values will not be calculated here because they are easily obtained after information values are determined. Note the charge for "chest auscultation: crackles" is $2. Let D1 represent pneumonia, D2 represent acute bronchitis, and F represent crackles. Here, P(D1) = 0.4; P(D2) = 0.3.

Single-frame Strategy

Assume no findings have been answered in the "lung consolidation by PE" frame. According to Iliad, if chest auscultation with crackles presence, the status of "lung consolidation by PE" frame is (close_true = 0.183, close_false = 0); if chest auscultation with crackles is absent, the frame status becomes (close_true = 0, close_false = 0.3391). Based on the modified Bayes' formula, eq. [2-6], P(D1|F+) = 0.76, P(D1|F-) = 0.21; P(D2|F+) = 0.0082, P(D2|F-) = 0.31. Equivalent likelihood ratios of F to D1 and F to D2 can be back calculated by Bayes' equation, LR+(F->D1) = 4.75, LR-(F->D1) = 0.399, LR+(F->D2) = 0.0193, LR-(F->D2) = 1.048. Therefore equivalent sensitivities and specificities can be back calculated as P(F+|D1) = 0.656, P(F-|D1-) = 0.86, P(F+|D2) = 0.0009, P(F-|D2-) = 0.953. Given P(D1) = 0.4, the frequency of positive F is P(F+) =

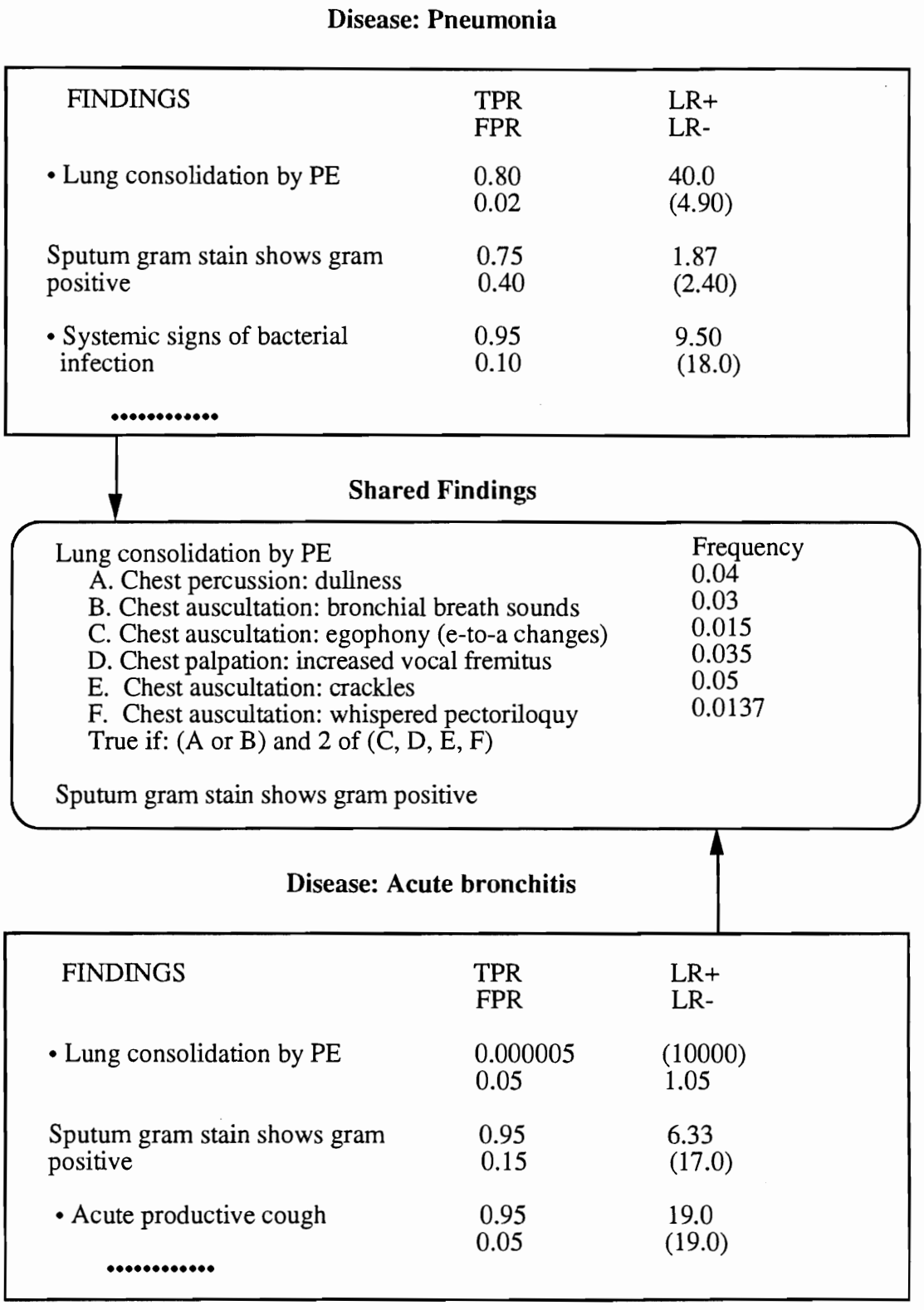

Figure 4. Example of calculating information provided by a finding to multiple diseases.

P(F+|D1)× P(D1) + P(F+|D1-)× P(D1-) = 0.656× 0.4 + (1-0.86)× 0.6 = 0.346; the frequency of negative F is P(F-) = 1 - 0.346 = 0.654. The expected information content provided by F to D1 is |P(D1|F+) - P(D1)| × P(F+) + |P(D1|F-) - P(D1)| × P(F-) = |0.76 - 0.4| × 0.346 + |0.21–0.4| × 0.654 = 0.249. By similar calculations, given P(D2) = 0.3, the expected information content provided by F to D2 is 0.0194. Iliad adjusts information content by multiplying the information content by the probability of disease. In this example, adjusted information content of F to D1 is 0.249 × 0.4 = 0.0996, and adjusted information content of F to D2 is 0.0194 × 0.3 = 0.00582. Under the single-frame strategy, the information content of F is max(0.0996, 0.00582) = 0.0996.

Across-frame Strategy

Under the across-frame strategy, the information content of F is the summation of adjusted information contribution to D1 and D2, that is 0.0996 + 0.00582 = 0.105.

Note that all the information content values have been weighted by the probability of the disease. The higher the probability of the disease, the larger the weighting factor; therefore Iliad tends to pursue findings in more probable disease frames. Actually the weighting factor is not limited to the probability of the disease; it can be the square of the probability of the disease or other values depending on the design of the best information algorithm.

The New Best Information Searching Algorithm

The new best information searching algorithm considers every finding related to the hypotheses being considered. The algorithm is designed to accommodate different combinations of information content models, decision criteria, and strategies. The algorithm is described as follows:

(1) Identify the disease being pursued. List all the unanswered findings including those in nested frames in the disease frame. Store all the intermediate frames between each finding and the top level disease frame.

(2) Identify the strategy being used. If the single-frame strategy is being used, determine each finding's quasi-utility value within each disease frame being considered, adjust each finding's quasi-utility value within each corresponding disease frame by multiplying a weighting factor (e.g., prior probability of the disease), rank findings within each disease frame being considered and choose the finding with the highest utility score. After ranking each disease frame being considered, choose the best finding among those "best findings" for each disease frame. If the across-frame strategy is being used, determine each finding's quasi-utility value. Generate a utility score list for each pursued disease frame and adjust each utility score list by multiplying a weighting factor (e.g., the prior probability of the corresponding disease). Generate a final utility score list that includes all the findings across all pursued hypotheses frames. Rank utility scores based on the final list, and select the finding with the highest score.

(3) Suggest the hypothesis to pursue and the best finding to acquire next. In the consultation mode, any finding whose utility score is within 95% range of the maximum score is suggested together with the best finding. In the simulation mode, any finding whose utility score is tied to the maximum score is requested.

Characteristics of Different Information Content Models

Different information content models measure information content differently. To show characteristics of different information content models, the following data are used to generate information content values at different prior probability of disease:

Sensitivity = 0.9	Specificity = 0.3
Sensitivity = 0.3	Specificity = 0.9
Sensitivity = 0.9	Specificity = 0.9
Sensitivity = 0.9	Specificity = 0.999
Sensitivity = 1.0	Specificity = 1.0

The prior probability changes from 0.0 to 1.0 with increment of 0.1 at each step. Two approaches of calculating information contents are used; the first approach is called expected information content approach, and the second approach is called one-way approach. The expected information content approach and the one-way approach are used for the expected utility decision criterion and the maximax utility decision criterion respectively. When prior probability is 0 or 1, mathematical calculations for some information content models can not be carried out. Therefore, numbers 0.000001 and 0.999999 are used to replace 0 and 1 respectively. Ten figures (Figure 5 to Figure 14) will show information content values at different prior probability of disease given a set of sensitivity and specificity values of a finding based on different information content models and different approaches of calculating information.

Figure 5 shows expected information content values calculated by the modified Shannon model. The figure shows that a perfect test (sensitivity = 1, specificity = 1) provides the highest information content when the prior probability of disease is 0.5. Note that test (0.9, 0.3) has the highest information content when prior probability is low (about 0.2), the test (0.3, 0.9) has the highest information content when prior probability is high (about 0.8), and the test (0.9, 0.999) has the highest information content when prior probability is about 0.35. The figure indicates that if all the tests cost the same, findings with high sensitivity and low specificity should be pursued during the early stage of the work-up, and findings with low sensitivity and high specificity should be pursued during the late stage of work-up.

Figure 6 shows one-way information content values calculated by the modified Shannon model. The figure shows that a perfect test provides the maximum information content when the prior probability is 0 or 1, which means that the positive result of the perfect test provides the most information when nothing is known about a patient and the negative result of the perfect test provides the most information when the patient's disease is known. Note that information contents provided by test (0.9, 0.3) and test (0.3, 0.9)

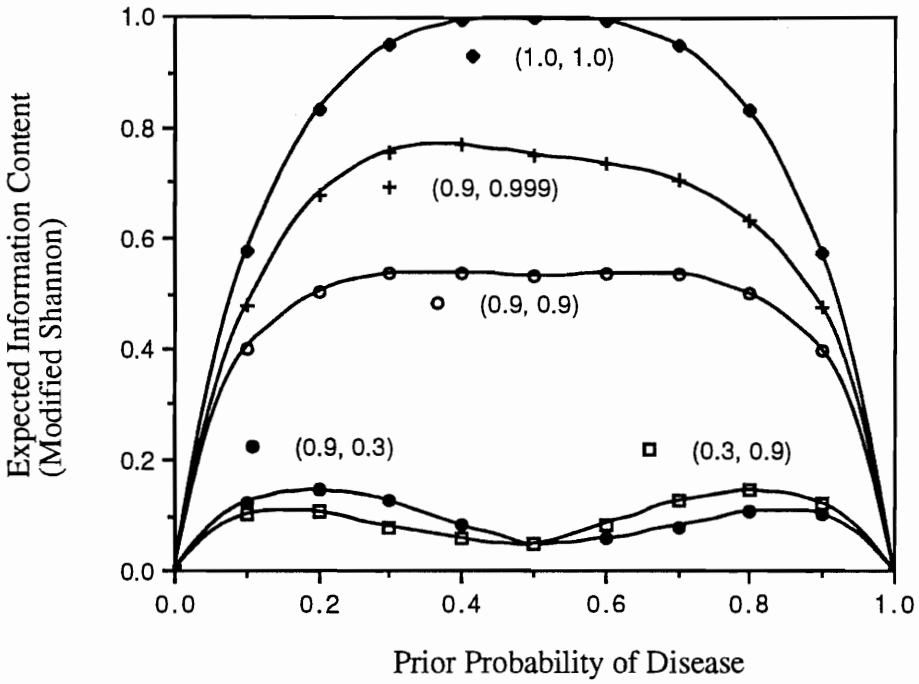

Figure 5. The expected information content calculated by the modified Shannon model provided by different tests that are indicated by "(sensitivity, specificity)."

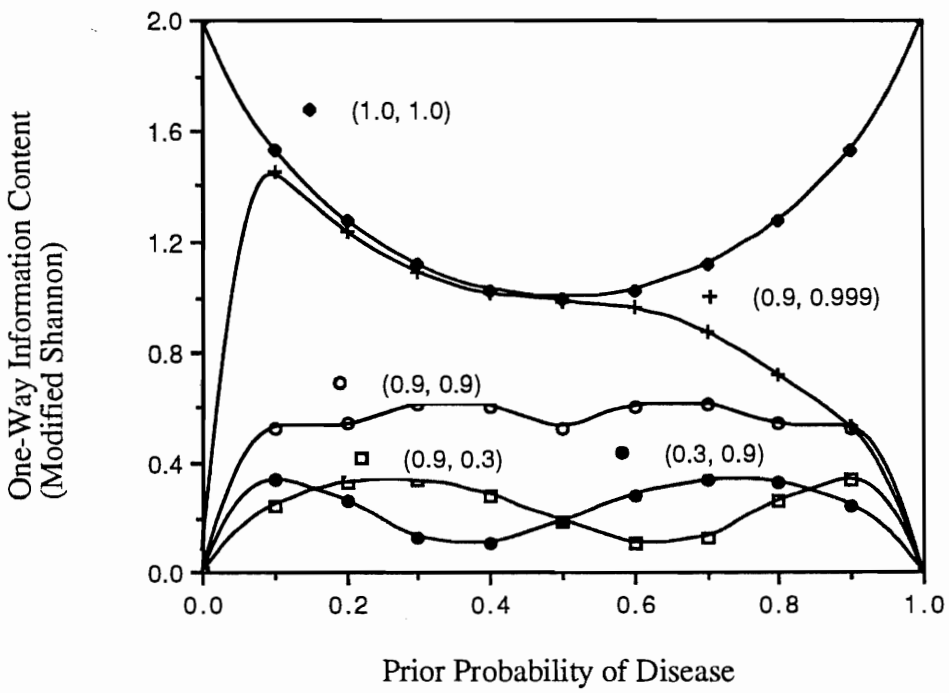

Figure 6. The one-way information content calculated by the modified Shannon model provided by different tests that are indicated by "(sensitivity, specificity)."

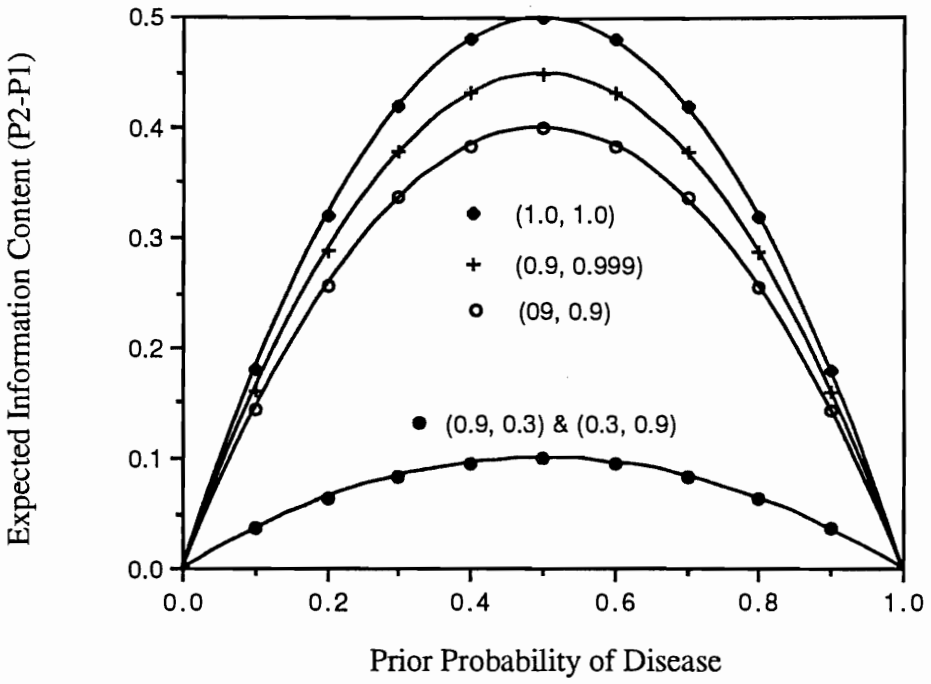

Figure 7. The expected information content calculated by the P2-P1 model provided by different tests that are indicated by "(sensitivity, specificity)."

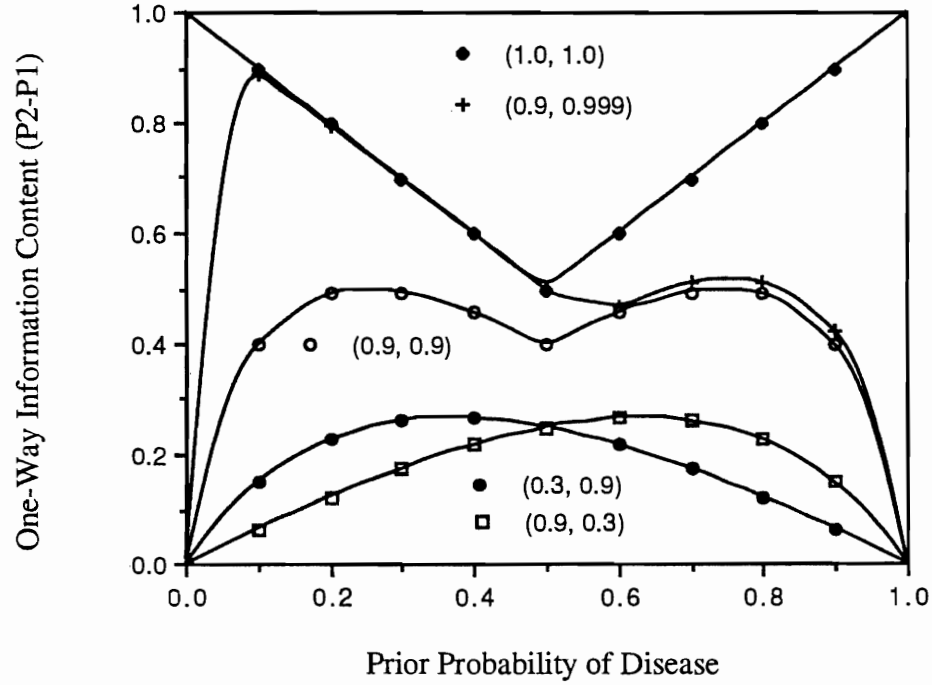

Figure 8. The one-way information content calculated by the P2-P1 model provided by different tests that are indicated by "(sensitivity, specificity)."

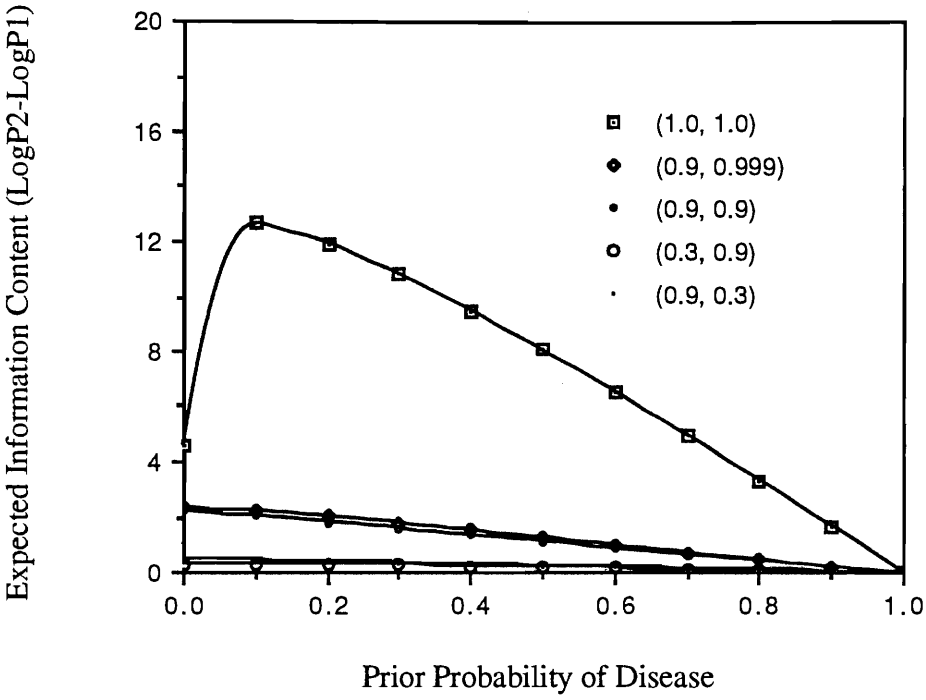

Figure 9. The expected information content calculated by the LogP2-LogP1 model by different tests that are indicated by "(sensitivity, specificity)."

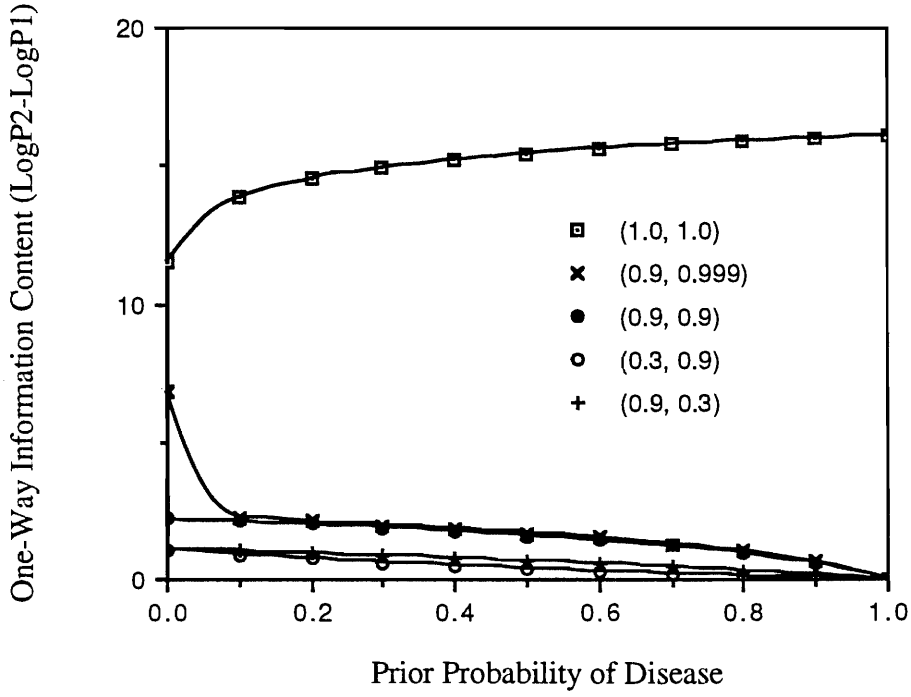

Figure 10. The one-way information content calculated by the LogP2-LogP1 model provided by different tests that are indicated by "(sensitivity, specificity)."

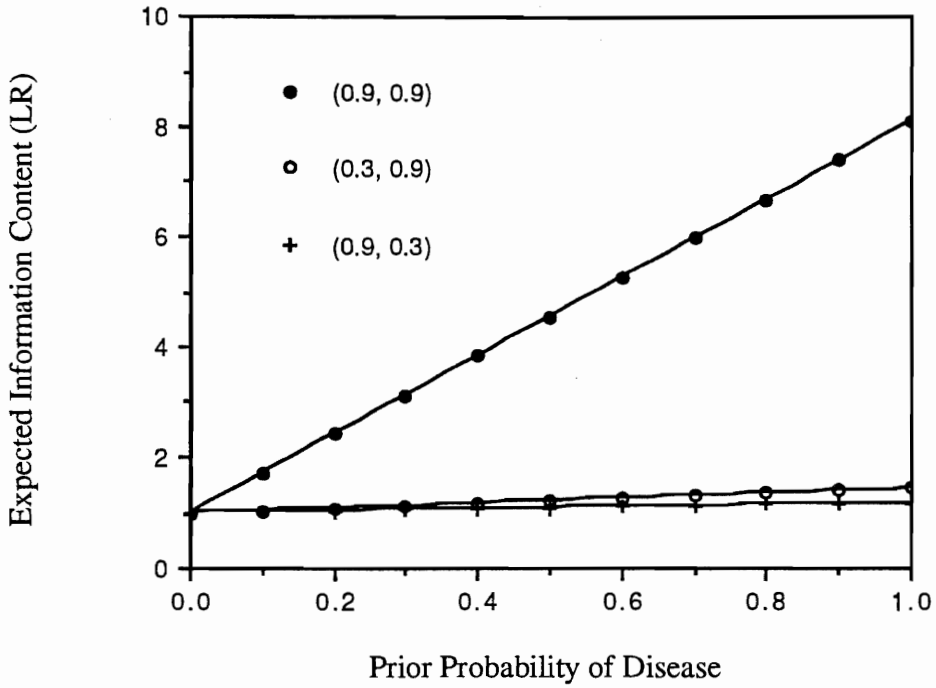

Figure 11. The expected information content calculated by the LR Model provided by different tests that are indicated by "(sensitivity, specificity)."

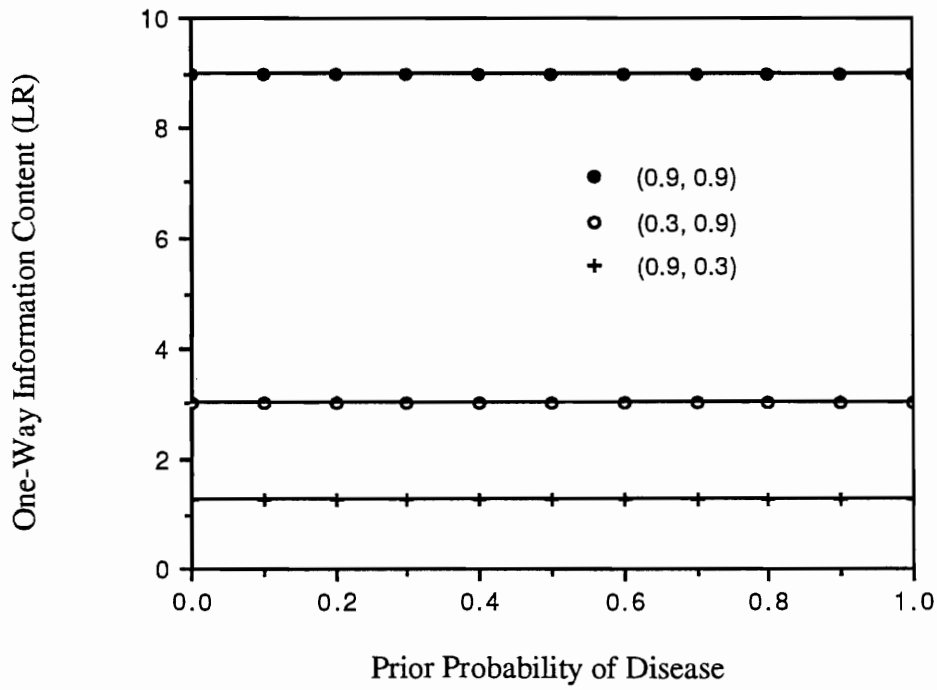

Figure 12. The one-way information content calculated by the LR model provided by different tests that are indicated by "(sensitivity, specificity)."

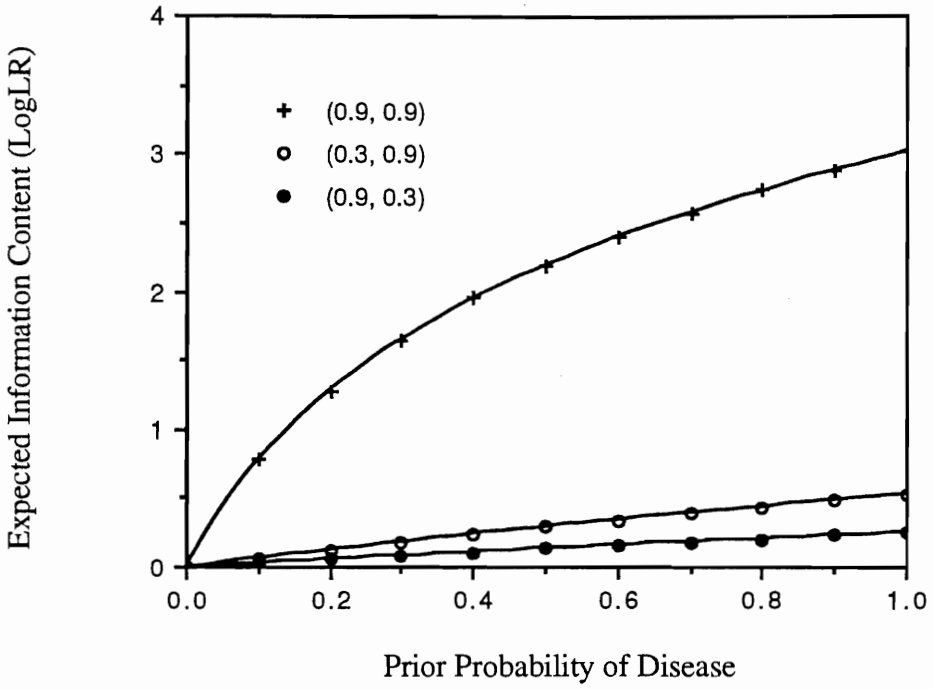

Figure 13. The expected information content calculated by LogLR model provided by different tests that are indicated by "(sensitivity, specificity)."

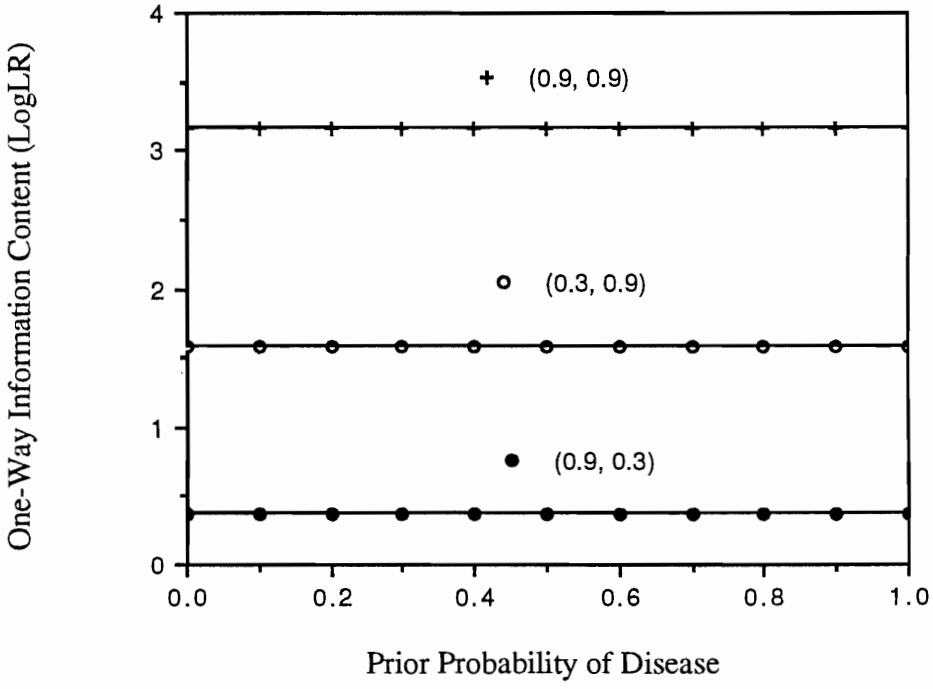

Figure 14. The one-way information content calculated by the LogLR model provided by different tests that are indicated by "(sensitivity, specificity)."

have two "maximum" information content points: one maximum point occurs in the early stage of a work-up, and the other maximum point occurs in the late stage of the work-up. The first maximum point is resulted from the positive finding, and the second maximum point is resulted from the negative finding.

Figure 7 shows expected information content values calculated by the P2-P1 model. The figure shows that each test provides the maximum information content at prior probability of 0.5. Each test provides the least information content during the early and late stage of a work-up.

Figure 8 shows one-way information content values calculated by the P2-P1 model. The figure shows a similar pattern of information content values versus different prior probability of disease to that based on the modified Shannon model (Figure 6).

Figure 9 shows expected information contents calculated by the LogP2-LogP1 model. A perfect test provides the maximum information content during the early stage of work-up and less and less information content as prior probability of disease decreases. Other tests shown in the figure provide less and less information content during the late stage of work-up (prior probability > 0.5).

Figure 10 shows one-way information contents calculated by the LogP2-LogP1 model. The figure shows that a perfect test provides high information contents throughout the change of prior probability of disease. Other tests shown in the figure have lower and lower information content as prior probability of disease decreases.

Figure 11 shows expected information contents calculated by the LR model. Information content values have linear relationship with the prior probability of disease. Because a perfect test has infinite likelihood ratio, the figure only shows three tests with limited likelihood ratio values.

Figure 12 shows one-way information content values calculated by the LR model. The figure shows that information content values are independent of the prior probability of disease.

Figure 13 shows expected information contents calculated by the LogLR model. Note the information content values do not have linear relationship with the prior probability of disease. The higher the information content the higher the prior probability of disease.

Figure 14 shows one-way information contents calculated by the LogLR model. Again the information content values are independent of the prior probability of disease.

Even though there are limited test conditions shown in the above figures, the changing pattern of information content values with the prior probability of disease reflect characteristics of different information content models under different decision criteria. Under different information content models given a decision criterion, a finding has maximum information content at different stages of a work-up. For example, a perfect test has the highest expected information content under the modified Shannon model when the prior probability of disease is 0.5; however a perfect test has the highest expected information content under the LogP2-LogP1 model when the prior probability of disease is small (about 0.1).

EXPERIMENTAL DESIGNS AND RESULTS

A best information algorithm is made of four important components. These components are (1) information content model, which reflects intrinsic disease-finding relation (e.g., likelihood ratio, prior probability and posterior probability of a disease); (2) strategy, which deals with number of diseases for which a finding provides information (single-frame strategy and across-frame strategy); (3) decision criterion, which deals with assignment of the information value when a finding has more than one possible outcome such as positive and negative results (expected criterion and maximax criterion); (4) partial information passing method, which determines the information propagation of a finding to a top disease frame through several intermediate disease frames (simplified method and complete method). The current Iliad best information algorithm is based on the combination of LR (likelihood ratio) information content model, the single-frame strategy, the maximax decision criterion, and the simplified partial information passing method.

Two evaluation approaches were utilized to evaluate the performance of different best information algorithms, which were different combinations of an information content model, a strategy, a criterion, and a partial information passing method. The first was the vignette approach; the second was the simulation approach.

A vignette is a snapshot of a particular stage during a diagnostic work-up. A real disease case can be divided into several diagnostically interesting stages; each stage was considered to be a vignette. In the vignette approach, each of the generated vignettes was solved by a particular best information algorithm. The suggestions about the next best data elements to pursue from each algorithm were collected for different vignettes. Expert physicians were asked to rank the suggestions by different algorithms. The frequency with

which each algorithm was chosen "the best" was calculated and the different algorithms were compared.

In the simulation approach, the overall work-up performance for an algorithm in solving a simulated case was evaluated by objective measures such as computational time, number of steps, charges, number of questions categorized by history, physical examination, and laboratory tests. A simulated case consisted of collecting sufficient medical findings to reach the final diagnosis for a patient. Iliad solved a simulated case automatically by using the best information mode sequentially. The details of the simulation approach will be discussed later in this chapter.

The central goal of this research was to evaluate different alternative algorithms as candidates for improving the current Iliad's best information algorithm. The hypothesis for the vignette approach was that a new algorithm would be found to provide work-up suggestions more like an expert. The hypothesis for the simulation study was that a new algorithm would be found to provide a more cost-effective work-up.

This study can be divided into three experiments. The vignette approach was used for the first and the second experiments. The first experiment was to evaluate the five information content models (LR, LogLR, Modified Shannon, P2-P1 and LogP2-LogP1) under the single-frame strategy only. The second experiment was designed to evaluate the five information content models under two different strategies: the single-frame strategy (the current strategy) and the across-frame strategy (the new strategy).

The simulation approach was used for the third experiment in which two decision criteria (maximax and expected) were combined with the five different information content models and the two strategies . This brought to the total of 20 possible algorithms to be evaluated ($5 \times 2 \times 2$). If the third experiment had been done as a vignette study with 20 vignette cases to be evaluated, it would have required evaluation of 600 scenarios. The time required for expert clinicians to evaluate six hundred scenarios would be prohibitive. It was also realized that there would be more and more factors to be included in Iliad's best

information algorithm as the study progressed, making the vignette approach become more and more time consuming and costly. Therefore, the simulation approach was used to study characteristics of different possible algorithms, and this allowed a relatively large number of cases to be analyzed at a minimal cost for expert time.

The simulation approach was not intended to replace the vignette approach. Rather, this approach was designed to evaluate different algorithms in a different way. The vignette approach was based on subjective measures of expert physicians at a particular decision point. The simulation approach was based on overall objective measures such as computational time, charges, number of steps, in terms of solving simulated cases. The subjective measures had more "face validity" than the objective measures because the subjective measures were based on the judgments of clinical experts. However, the vignette approach requires experts to build vignette cases and to rate the suggestions provided by different algorithms for those vignette cases. Experts have to be involved whenever a new algorithm needs to be evaluated. The simulation approach requires experts to enter data from real cases from which simulated cases can be easily created. Once the simulated cases are built, these simulated cases can be repeatedly used for the evaluation of new algorithms without further expenditure experts' time.

Vignette Approach

Experiment I

Implementation

The most useful information option identifies the most cost-effective findings to pursue next at any stage of a work-up. Figures 15 and 16 show how the current most useful information algorithm works. A patient presented with chest pain and shortness of breath, and Iliad generated a differential diagnosis list. At this stage of the work-up, a user wanted to know what the next most useful findings would be. Figure 15 shows the user selecting the most useful information mode, and Figure 16 shows the most useful

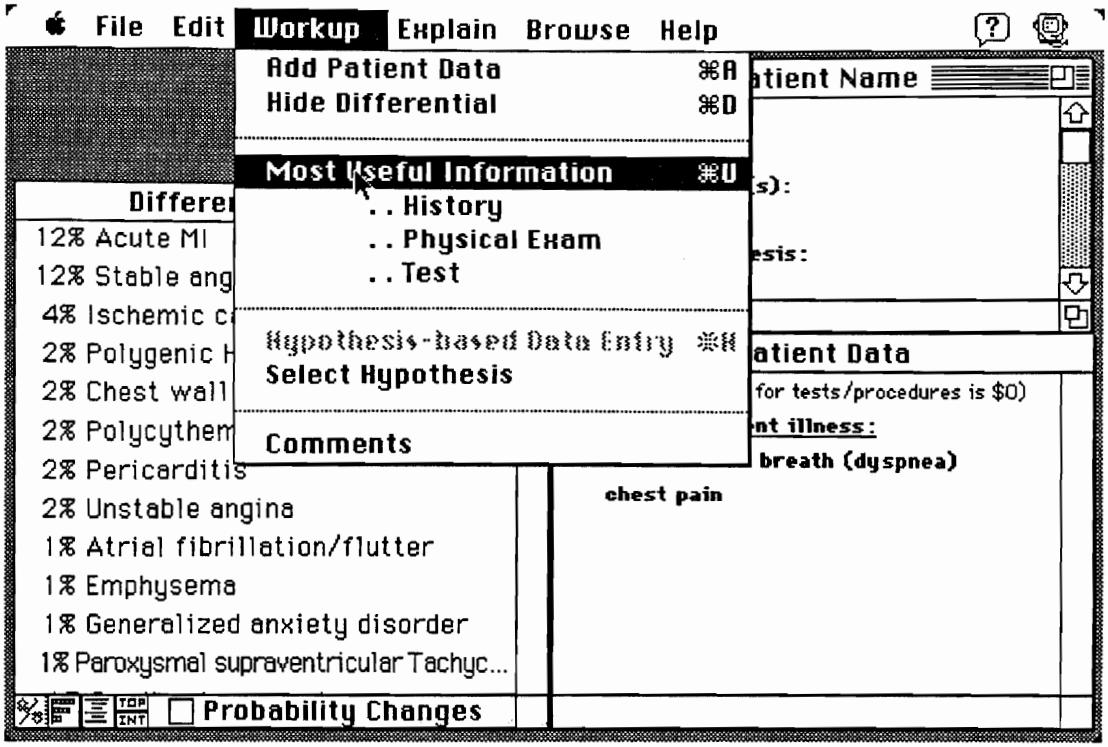

Figure 15. An example that a user selects the most useful information function during a work-up.

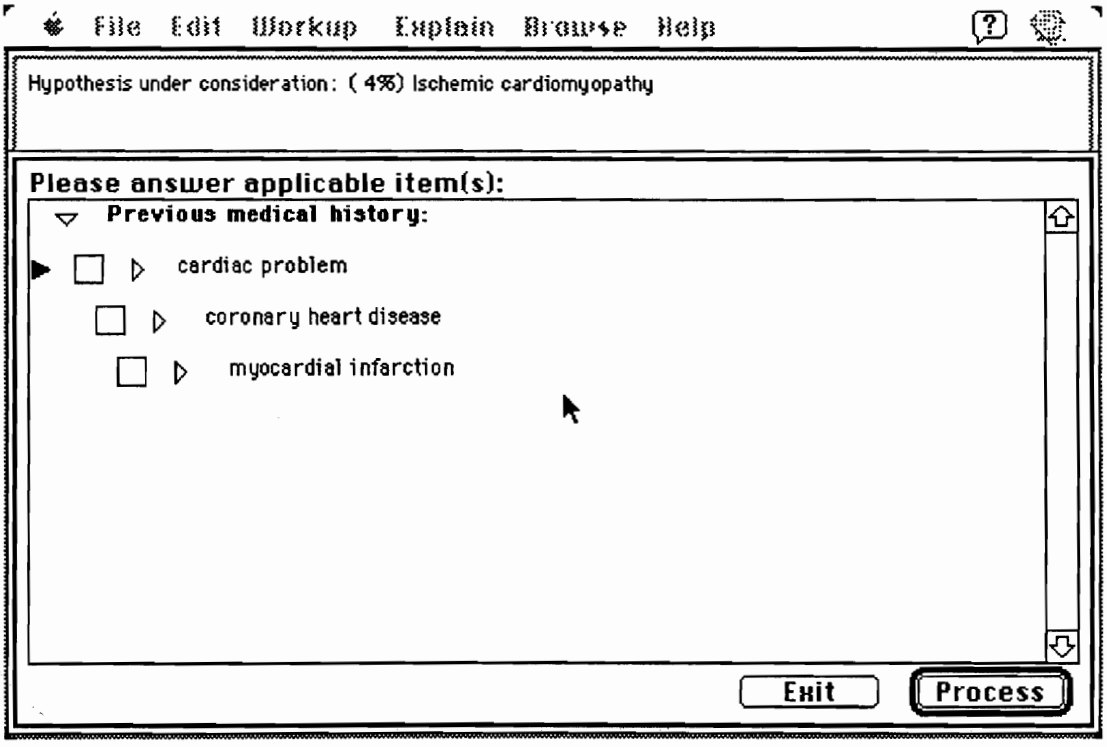

Figure 16. An example that a user obtains the feedback from the best information mode.

information feedback from Iliad. To evaluate the performance of five different information content models (the current one and four proposed ones), the four proposed models (Shannon, P2-P1, LogP2-LogP1, LogLR) were implemented in four additional versions of Iliad. Note that all these new models were implemented in Iliad by following the same most useful information searching method (partial information passing method I, the single-frame strategy) as the current Iliad.

Subjects

Six physicians specializing in internal medicine served as the subject-judges in the experiment. These physicians were all faculty members of the University of Utah School of Medicine. They were experienced in the use of computerized expert systems for medical decision making.

Experimental Design

The experiment was a $3 \times 5 \times 3$ (Case $\times$ Model $\times$ Work-up Stage) factorial design. All independent variables are within subjects factors. The two dependent variables were measures (1) of the probability of being chosen as the best model and (2) of the expert's judgments about the appropriateness of the findings selected by the models.

Procedure

Three actual pulmonary disease cases were selected by a pulmonary expert to provide the case material. Each case was divided into six vignettes, so there were total 18 vignettes. Those vignettes for each case were grouped into three diagnostic stages: the early stage (vignette 1 and 2) denotes the preliminary steps in the case work-up, when the diagnostic certainty was less and there were many diagnostic competitors. The middle stage (vignette 3 and 4) and the late stage (vignette 5 and 6) considered later steps in the work-up when major diagnostic competitors were considered and finally eliminated. Each version of Iliad suggested the best data elements to seek next in each vignette. It was found

that each of these different models often pursued different findings for the same vignette. The different strategies occurred because each model provided a different evaluation for information content of the alternative findings. The sample rating form is shown in Appendix A. Each expert first rated the suggestions from each version of Iliad. The ratings, "finding scores," were on a scale of 1 to 5 (1 = least cost-effective, 5 = most cost-effective). The finding scores reflected cost-effectiveness of the findings proposed by that model's hypothesis and were not based on the appropriateness of the model's hypothesis. Each expert then chose the best overall combination of the hypothesis and suggested work-up. The choice of the best overall model was based on the combination of the disease to be pursued and the cost-effectiveness of the question for the disease. Several models might be regarded as being equally effective and might be simultaneously chosen as the best model. Figure 17 shows a scenario of the evaluation procedure. Vignette 3 from case 2 was submitted to the five versions of Iliad; expert 3 rated the findings suggested by the five models and picked the best overall.

Results

There were two outcome variables generated for each model. The first outcome variable was the probability of each model being chosen as the best model. This measure represents the proportion of experts who chose the result of that model as the best of the diagnostic approaches. If three of the six physicians indicated that the work-up plan suggested by the first model for a vignette was the best, then this model was assigned a score of 50% for the vignette. For the statistical analysis, the scores for two vignettes at each work-up stage were averaged. There were ties where two or more of the models proposed the same work-up plan for the same hypothesis. In these cases, the tied models were given the same score. The judges' ratings of the best model dependent variable were analyzed using a 3 x 5 x 3 (Case x Model x Work-up Stage) factorial analysis of variance. The analysis indicated that the main effects of the Work-up Stage [$F(2,425) = 3.08$, $p <$

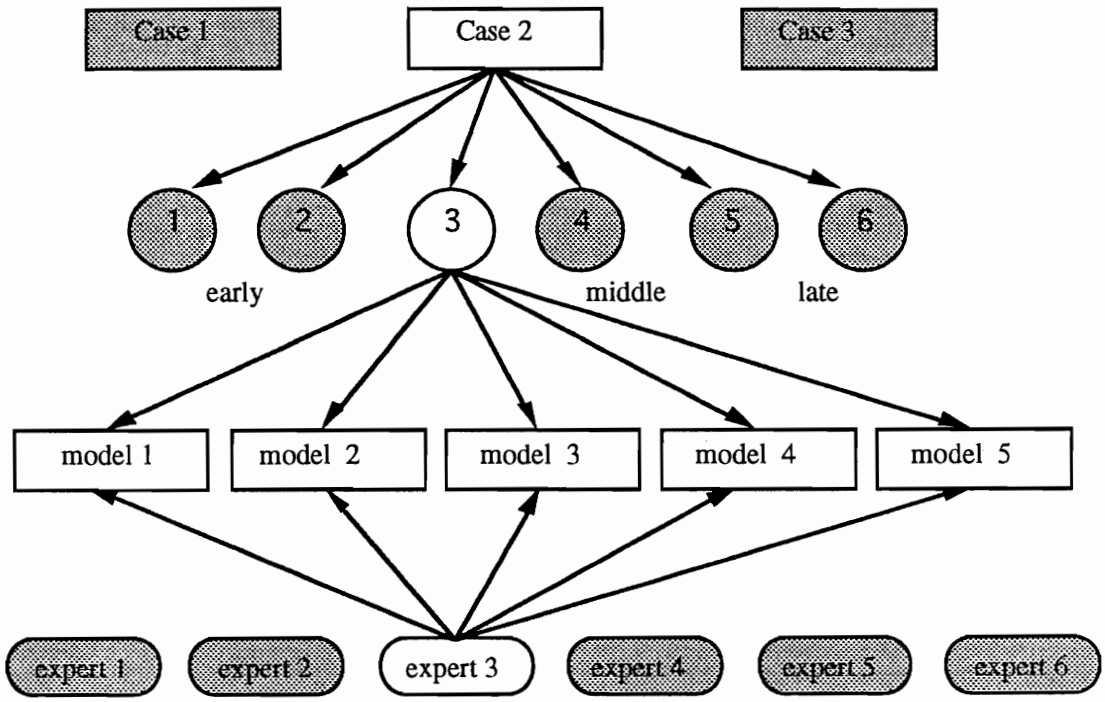

Figure 17. The evaluation procedure for experiment I.

.01] and the Model [$F(4,85) = 3.86$, $p < .006$] were statistically significant. The primary hypothesis was that the Model x Work-up Stage interaction would be significant and results supported this hypothesis, $F(8,425) = 2.23$, $p < .002$. All reported statistically significant comparisons among means were based on the Newman-Keuls procedures (36).

Figure 18 shows the overall frequency with which each model was chosen as the best at different stages of the work-up (early, middle, and late). Shannon and P_2 - P_1 models both performed better than the current LR (likelihood ratio) model in the early and middle stages. In the late stage, the relative performance of those three models was closer to each other. The logLR model and the LR model did not differ significantly and were chosen as the best overall model equally often for the three vignette stages. The $logP_2$-$logP_1$ model is better than the LR model in the early stage but is worse than the LR model in the middle and late stages.

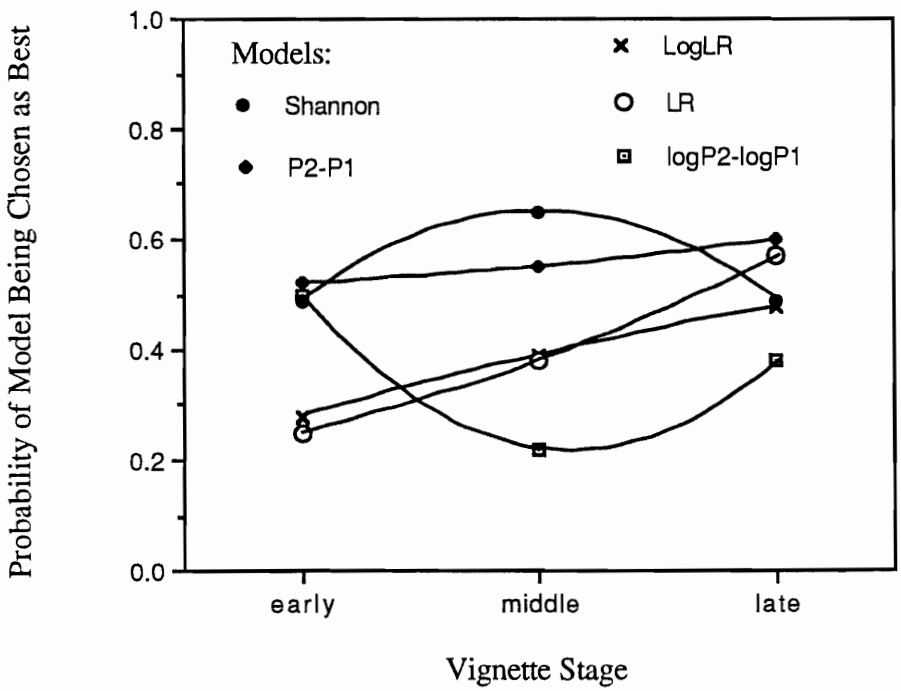

Figure 18. Probability of each model being chosen by experts as the best in different stage of work-up.

Figure 19 shows the overall (across stages) percentage of time each model was chosen as the best. Because of frequent ties the total is greater than 1.0. Newman-Keuls procedures demonstrated that the Shannon and P_2-P_1 models were not significantly different in performance, and both of them were significantly better than the LR model. The performance of the logP_2-logP_1, logLR, and LR models was not significantly different.

The second outcome variable was the "finding scores" (scaled from 1 to 5, 1 = least cost-effective, 5 = most cost-effective) given by the experts for each model. A 3 x 5 x 3 (Case x Model x Work-up Stage) factorial analysis of variance was performed on the expert judges' ratings of the findings scores. The results indicated that the Work-up Stage main effect [$F(2,425) = 9.98$, $p < .001$] and the Work-up Stage x Model interaction [$F(8,425) =$

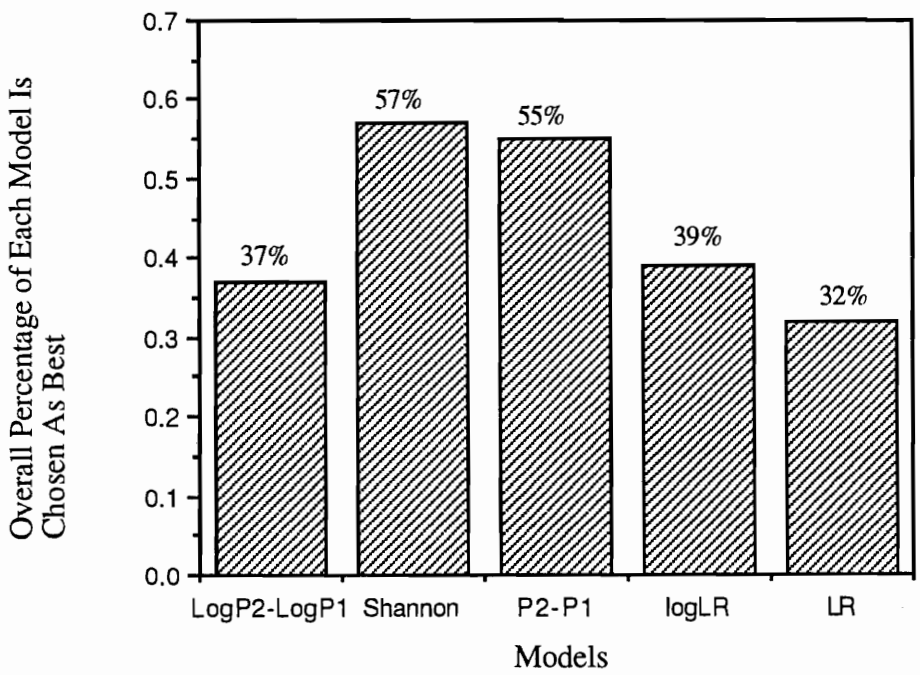

Figure 19. Overall (all stages) probability of time each model being chosen as the best by experts.

1.74, $p < .025$] were statistically significant. The model main effect was not statistically significant, $F = 1.69$. Figure 20 shows average scores given by the experts for the work-up, independent of the hypothesis proposed by each model. These scores represented the appropriateness of the finding to the hypothesis that the model pursued; the scores were unaffected by how good the suggested working hypothesis was. Again, Newman-Keuls procedures indicated that the Shannon and P_2-P_1 models were significantly better than the current model (LR) at each work-up stage (early, middle, and late). The logLR model was also significantly better than the LR model, but the LogP2-LogP1 model showed unstable performance compared to the LR model. It was significantly better than the LR model at the early work-up stage but was significantly worse at the middle work-up stage and was as good as the LR model at the late work-up stage.

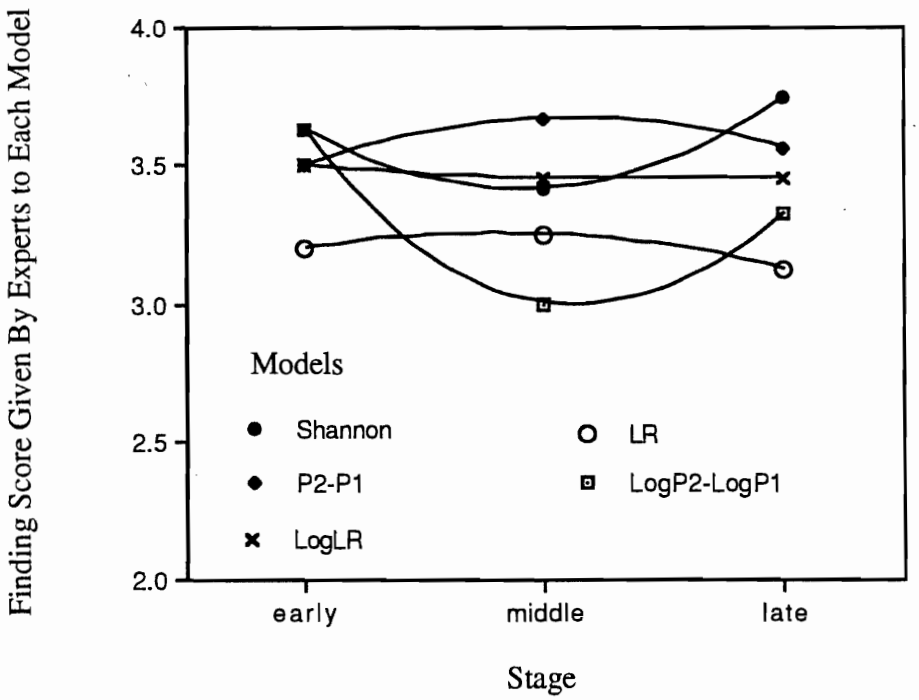

Figure 20. Average finding scores given by experts to each of five models at different stages of work-up.

Experiment II

Implementation

This experiment was designed to compare the performance of the five information content models and two strategies, the single-frame strategy and the across-frame strategy, to the judgments provided by expert clinicians. Five information content models and two strategies were implemented in an experimental version of Iliad, so that any combination of an information content model and a strategy could be used to pursue the most cost-effective work-up. Note that the partial information passing method II (new method) and the maximax criterion were used in the implementation.

Single diagnostic procedures, such as chest X-ray examinations or batteries of laboratory tests (e.g., Chem-20) can produce multiple findings. It was assumed that these findings should be evaluated together to give a total value for the information from the procedure. Thus, in this implementation of the across-frame strategy, all information from the same lab test was summed across together to represent the information content contributed by the test. Although history and physical examination findings are usually collected systematically in real life, they were treated individually in this implementation of the across-frame strategy.

Figure 21 and Figure 22 show how the experimental version of Iliad works. Figure 21 shows that an information content model can be selected from a hierarchical menu item "Algorithm." Figure 22 shows that a strategy can be selected from a hierarchical menu item "Strategy." After the selections, go to the menu item "Best Information," the system will provide most cost-effective suggestions based on the selected combination of the information content models and the strategy. Note that menu item "Modes" contains selections of "history," "physical exam," "lab test," and "overall." "Overall" is the default selection, which means suggested findings will not be limited to a particular category of history findings, physical examination findings, or laboratory test procedures; they will be cost-effective over all the categories.

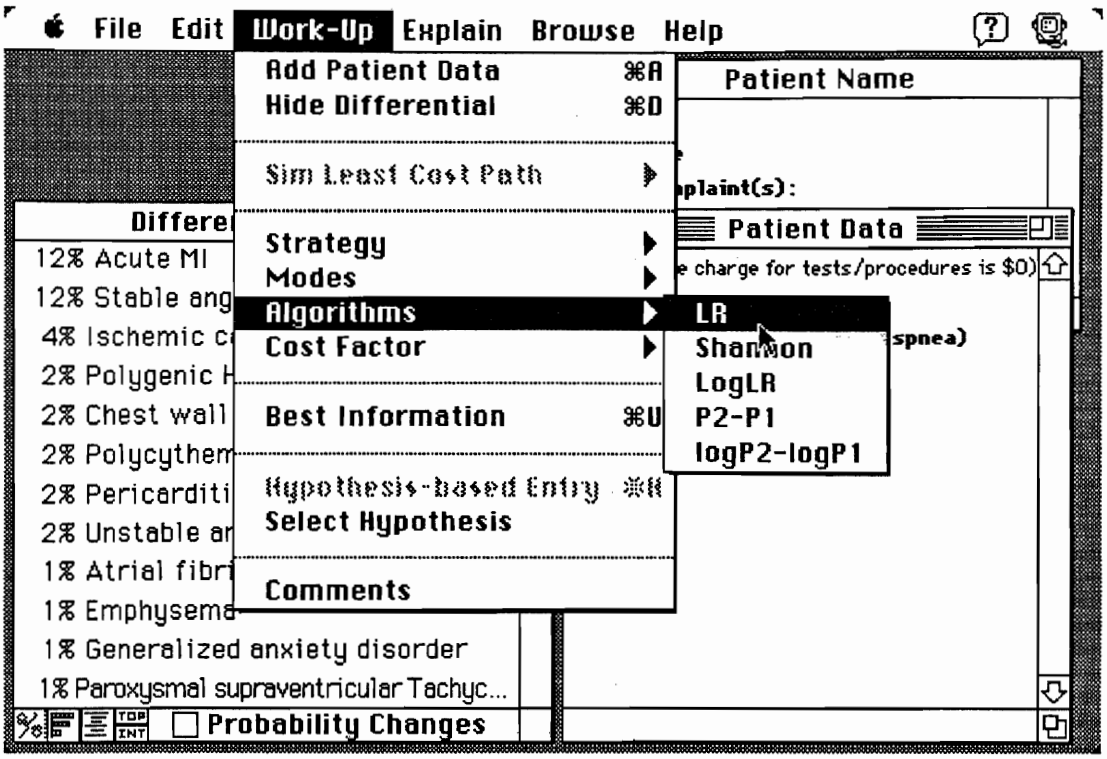

Figure 21. Selection of an information content model in experiment II.

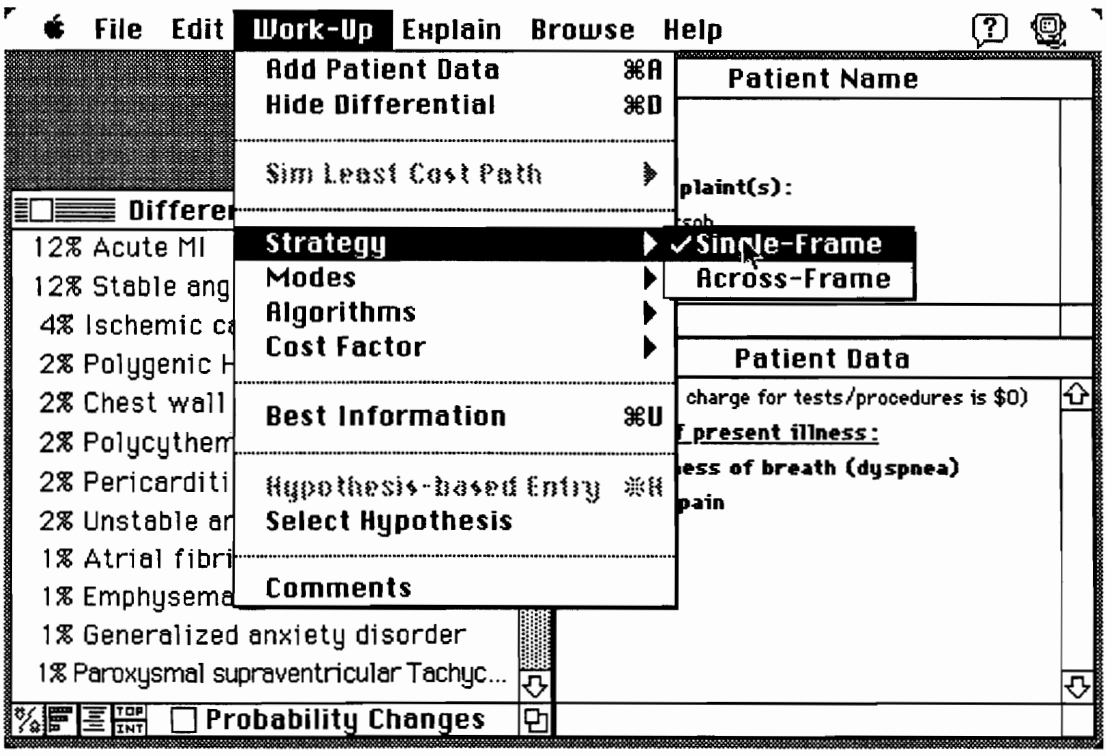

Figure 22. Selection of a strategy in experiment II.

Subjects

Six academic internists specializing in internal medicine served as the subject judges in the experiment. These physicians were all faculty members of the University of Utah School of Medicine. They were all experienced in the use of computerized expert system for medical decision making.

Experimental Design

The independent variables are: Case (six cases), Stage (three stages in each case), Information Content Model (five models), and Strategy (two strategies). All raters had substantially the same level of training and experience. Hence, no effort was made to classify the experts by level or type of expertise. The experiment was a $6 \times 3 \times 5 \times 2$ (Case $\times$ Stage $\times$ Information Content Model $\times$ Strategy) factorial design. All independent variables are within subjects factors. There were three dependent variables. The first dependent variable was the frequency of being chosen as the best information content model under the single-frame strategy. The second dependent variable was the frequency of being chosen as the best information content model under the across-frame strategy. The third dependent variable was the frequency of being chosen as the best strategy overall. Each dependent variable represents the proportion of experts who chose the result of that outcome as the best.

Procedure

Six pulmonary cases were randomly selected from real patient cases at the University of Utah Medical Center. Each case was divided into three stages. The first two stages denoted stages in the work-up when history and physical examination findings were acquired. The third stage denoted a later stage in the work-up when major competing diagnostic hypotheses were considered and finally eliminated through laboratory tests or other procedures. Thus there were 18 medical decision points, or vignettes, represented. Each information content model was applied under two strategies. Therefore, 10 best

work-up suggestions (5 algorithms × 2 strategies) were generated for each vignette. Each expert was provided with a copy of the vignette containing a subset of patient findings and also the hypotheses Iliad considered, the 10 work-up suggestions, and a simple rating form. A sample vignette rating form is given in Appendix B. Figure 23 shows the rating instructions given to experts. In the rating instructions, the term "student" instead of "algorithm" was used to make experts feel closer to real-life ratings.

Based on Iliad's suggested work-up items, the experts were instructed to choose (1) the best strategy for each information content model and (2) the best information content model for the single-frame and the across-frame strategies. "Ties" were allowed only when several information content models or two strategies produced the same work-up suggestion. A nominal scale was used to assess the information models and strategies. Whenever the strategy or the information model was chosen as the best, the score was assigned to 1; otherwise the score was 0. All the experts completed their entire set of evaluation forms.

Results

The judges' ratings of the best information model under the single-frame strategy were analyzed by using a $6 \times 5 \times 3$ (Case × Information Content Model × Stage) factorial analysis of variance. Comparisons among cell means were based upon a Bonferonni adjusted confidence interval (36). In the Bonferroni procedure, the desired significance level (α) is divided by the number of comparisons to be performed (k). In the study, four comparisons were to be made among the means so that the adjusted significance level was 0.0125 (0.05/4). The analysis of variance indicated that the main effect for Information Content Model was statistically significant, $F(4,360) = 10.72$, $p < 0.0001$. The interaction between Stage and Information Content Model was also statistically significant, $F(8,360) = 4.83$, $p < 0.0001$. Average scores (frequency of being chosen as the best) of each information content model at stage 1 and stage 2 were used to represent the

Rating Instructions

Five medical students were asked to work up several real patient cases. Students have reviewed each case history and independently proposed two work-up plans for each case, which we call Strategy I and Strategy II. Each type of strategy was based on slightly different sets of instructions and goals given to the students. We would now like you to rate each student's performance.

FIRST, please pick the best strategy for each student. Make this rating in the right-most column, labeled "Best Strategy." For instance, suppose student one's suggested work-up plan under Strategy I is superior to his plan under Strategy II. You would then write a "I" in the Best Strategy column for that student.

SECOND, please pick the best student for each type of strategy. Make this rating in the bottom row, labeled "Best Student". For instance, suppose student 4's Strategy I was the best of all the Strategy I suggestions. You would then write a "4" in the "Best Student" row under that strategy.

NOTE: How to handle identical work-up strategies. If a particular student suggests the same work-up plan for both Strategy I and Strategy II, write an "X" in the "Best Strategy" column. If two or more students each suggest the same work-up plan and they tie each other for the "Best Student" rating, write each student's number in the "Best Student" row.

Here is a sample rating sheet:

Student	Strategy I	Strategy II	Best Strategy
1			
2			
3			
4			
5			
Best Student			

Figure 23. Rating instructions for experiment II.

effectiveness of the model in suggesting history and physical exam findings, and the score at stage 3 was used to reflect the effectiveness of the model in suggesting laboratory test procedures. Comparisons among the mean finding scores for the algorithms indicated that the Shannon model was significantly better than the rest of models ($\alpha = 0.0125$) in terms of suggesting history and physical exam findings. However, the results revealed no significant differences between Shannon's model and the LR model (current model) in terms of suggesting laboratory test procedures during the late stage of work-up. No other models performed better than the LR model at the late stage. The results also indicated that the Shannon's model was the best in terms of overall finding scores across all stages ($\alpha = 0.0125$). The overall performance of five information models is shown in Figure 24.

Under the across-frame strategy, the results about the performance of different information content models were similar to the previous findings when information content was summed across frames. The analysis of variance showed that the main effect for Information Content Models was statistically significant, $F(4,360) = 3.20$, $p < 0.015$. The interaction between Stage and Information Content Model was also statistically significant, $F(8,360) = 2.30$, $p < 0.02$. The Shannon's model and the P2-P1 model were not significantly different in terms of suggesting history and physical exam questions, but they were all significantly better than the LR model ($\alpha = 0.0125$). During the late stage of the work-up, there were no significant difference among the Shannon, P2-P1 and LR models. No models performed better than the LR model in terms of suggesting laboratory test procedures. The Shannon's model and the P2-P1 model were the best overall among the five models across all stages ($\alpha = 0.01$). The overall performance of five information models under the across-frame strategy is shown in Figure 25.

The frequencies of each strategy being chosen as the best based on the grand average scores for all the information models were calculated. Each strategy was evaluated five times in each of the 18 vignettes. The five times represent implementation of five information models under each strategy for each vignette. The best strategy scores were

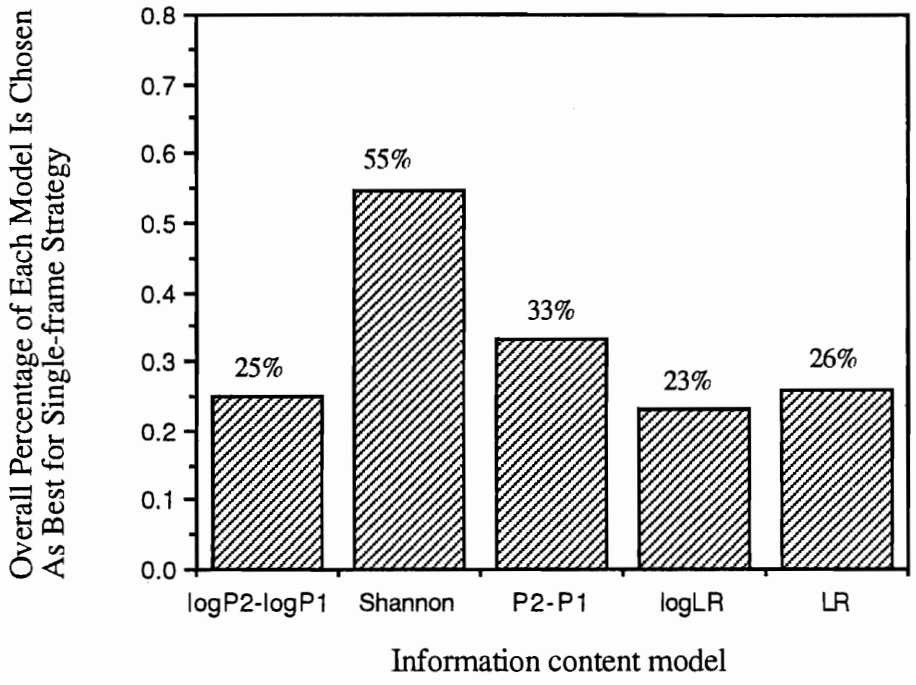

Figure 24. Overall (all stages) frequency that each information content model was chosen as the best by experts under the single-frame strategy.

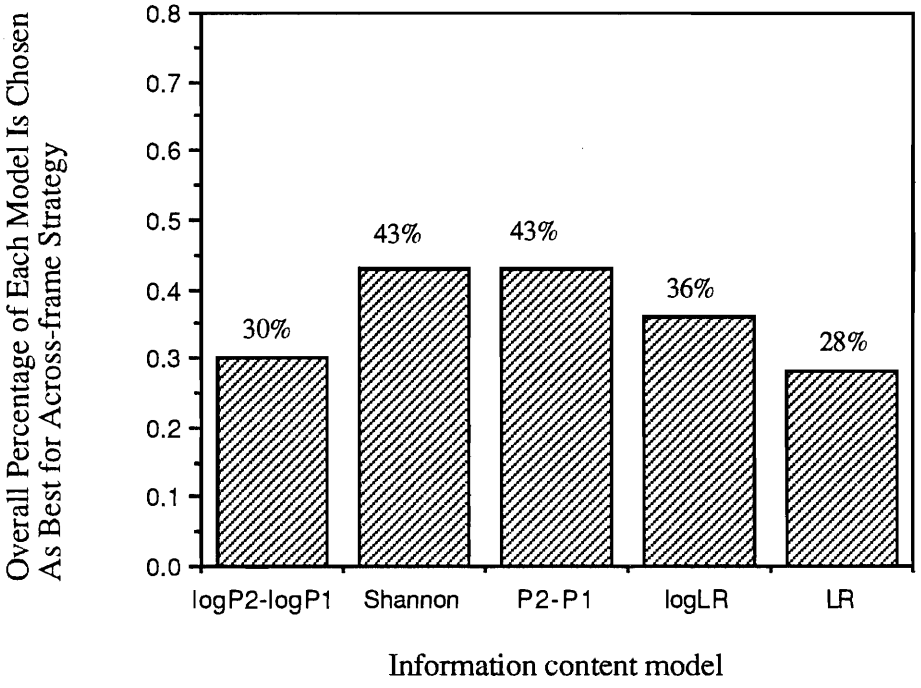

Figure 25. Overall (all stages) frequency that each information content model was chosen as the best by experts under the across-frame strategy.

analyzed by ANOVA using repeated measures. The results indicated that experts preferred the across-frame strategy to the single-frame strategy, as shown in Figure 26.

Simulation Approach

Experiment III

Implementation

This experiment was designed to demonstrate simulation techniques to compare the performance of different algorithms. Five information content models, two strategies, and two decision criteria were implemented into a special version of Iliad, so any combination of an information content model, a strategy, and a decision criterion could be used to solve a simulated case automatically. A simulated case is formed based on a actual patient case. The disease manifestations from the patient are stored as the simulated case. Starting from the chief complains of the patient, Iliad's best information algorithm generates the next best work-up findings to pursue next; the answers of the requested findings are based on the record in the simulated case. Iliad updates the differential diagnosis list given the answers of these findings and repeats the process (requesting the next most cost-effective finding to pursue) again until a stopping rule is met. There are three stopping rules: the first is that the probability of the simulated patient's disease(s) is greater than 0.95, the second is that all disease probabilities in the differential diagnosis list are less than the apriori probabilities, and the third is that Iliad runs out of questions to ask for the hypotheses being pursued. The first stopping rule is used when Iliad can successfully solve a case; the second or third stopping rules is used when Iliad can not solve a case. Figure 27 and Figure 28 show an example of a simulation study that is being solved by Iliad. In Figure 27, under the menu item "Algorithms," different information content models can be selected. A decision criterion can then be selected after an information content model is selected. Note that the item "Current Iliad" represents the exact current algorithm, which uses the LR information content model, the single-frame strategy, and partial information passing method I. The

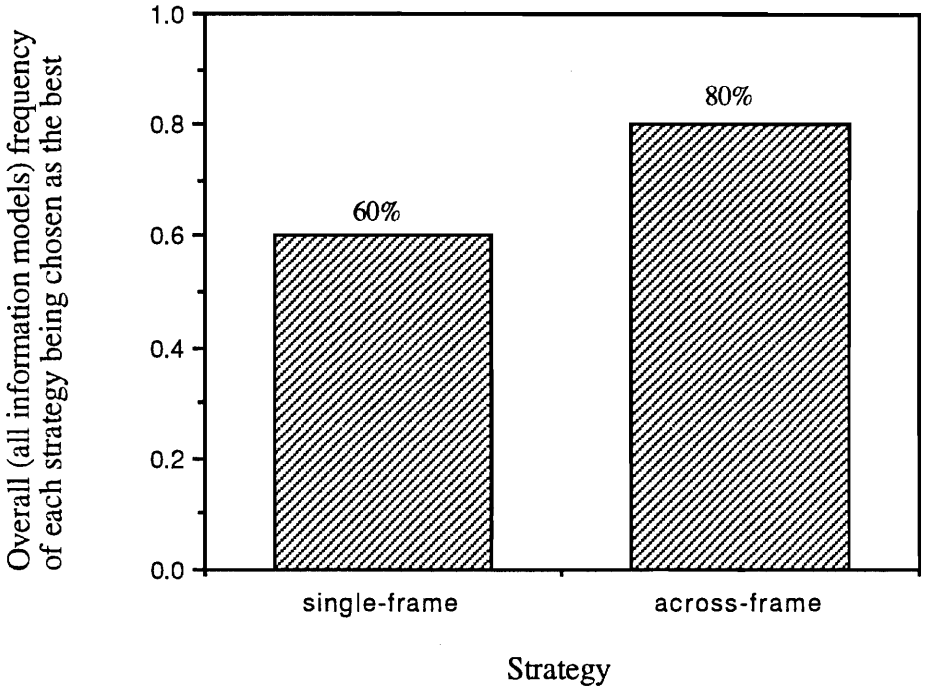

Figure 26. Overall (all information content models) frequency that each strategy was chosen as the best

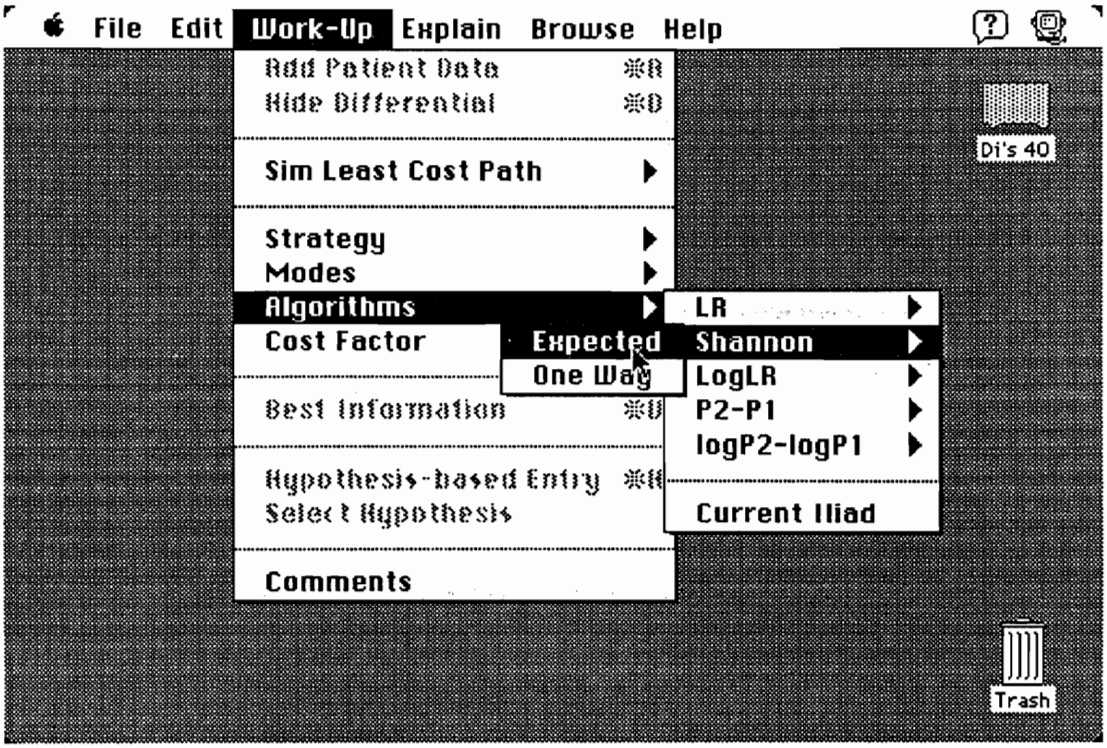

Figure 27. Menu from which an algorithm is selected to solve a simulated case.

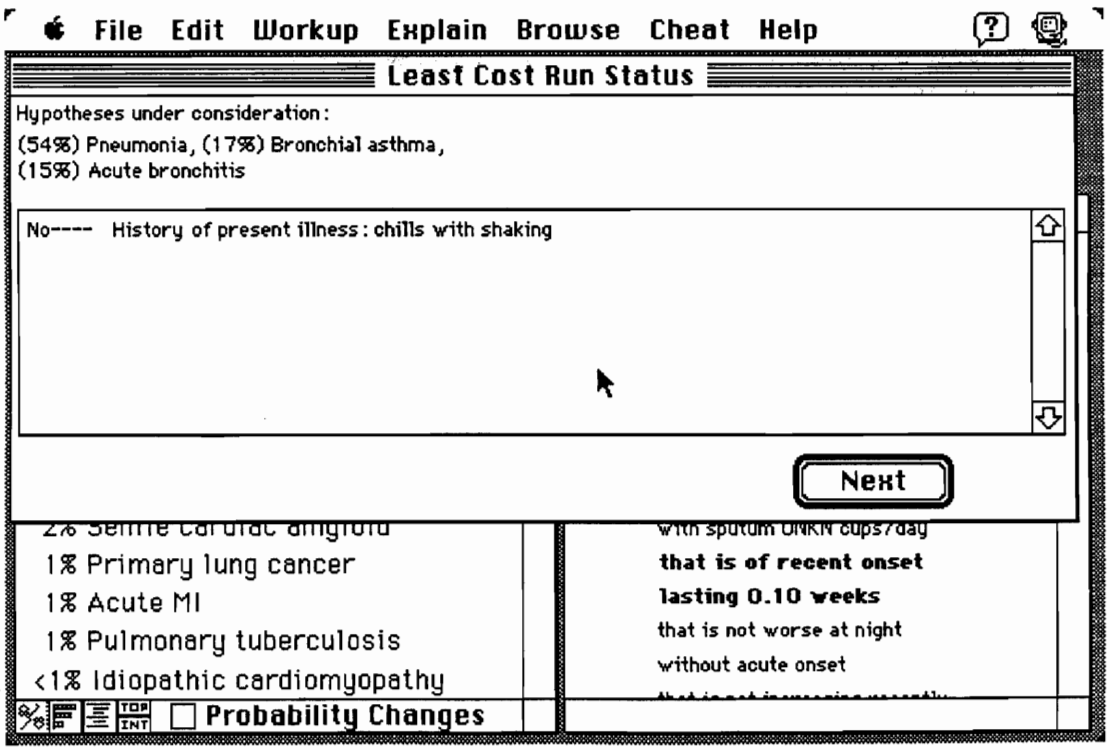

Figure 28. An example of a simulated case being solved by Iliad.

item "LR" represents the modified current algorithm, which uses the LR information content model and partial information passing method II (the new partial information passing method). Under the menu item "Strategy," the single-frame strategy and the across-frame strategy can be selected. Under the menu item "Modes," the overall best information was the default. After all the selections, go to the menu item "Simulation Least Cost Path," Iliad will start to solve a simulated case either "step by step" or "automatically." If step-by-step mode is selected, Iliad stops at each step. If automatic mode is selected, Iliad does not stop until a stopping rule is met. The Figure 28 shows an example that Iliad gets the answer about a question "chills with shaking" at a particular step. Iliad updates the differential diagnosis list sequentially based on answers to questions at each step until the stopping rule is met. Note that if an algorithm pursued history and physical exam findings whose results were not recorded in the case, those findings were set to negative answers. If an algorithm pursued laboratory test or procedures whose results were not recorded, those tests were set to unknown answers.

Case selection

Seventy simulated cases were randomly selected from test cases used for medical student training at University of Utah School of Medicine. These cases were based on real charts. Being actual cases, they could be solved by means of more than one pathway. The distribution of cases in different domains is shown in Table 5.

Table 5. Number of Simulated Cases in Different Domains

Cardiology	8	GI	14
Hematology	5	Infectious Diseases	15
Metabolic and Endocrine Diseases	2	Pulmonary Diseases	7
Renal Diseases	11	Rheumatology	8

Simulation Procedure

To avoid repetitive selections, a special version of Iliad, which runs each combination of an information content model, a strategy, and a decision criterion, was created. After the selection of an algorithm, the special version of Iliad did work-ups for 70 simulated cases sequentially until all the cases were processed. All the dependent variables were recorded for each simulated case.

Experimental Design

The independent variables were: Information Content Model (five models), Strategy (two strategies), and Decision Criterion (two criteria). The experiment was a $5 \times 2 \times 2$ (Information Content Model $\times$ Strategy $\times$ Decision Criterion) factorial design. All independent variables were within subjects factors. There were nine dependent variables. Each variable describes a performance measure for an algorithm to solve a simulated case. These measures were (1) time (seconds), (2) number of steps, (3) cumulative charges ($) for "known laboratory tests" ordered, (4) cumulative charges ($) for "all laboratory tests" ordered ordered, (5) number of known laboratory tests ordered, (6) number of total laboratory tests ordered, (7) number of history questions requested, (8) number of physical exam questions requested, (9) total number of history and physical exam questions requested. Note that "known laboratory tests" were those tests whose results could be found in the simulated case. "All laboratory tests" included known laboratory tests and those tests whose results could not be found in the simulated case.

Results

There were 20 combinations of algorithms used to solve 70 simulated cases. Because certain simulated cases could be solved by some algorithms but not by others, 44 cases (62%) were solved by all 20 algorithms. The statistical analysis of the dependent variables was based on the complete data sets of these 44 cases.

Each of nine dependent variables were analyzed by using a $5 \times 2 \times 2$ (Information Content Model $\times$ Strategy $\times$ Decision Criterion) factorial analysis with repeated measures. The comparisons between the means were based on paired t-test with the significant level at $\alpha = 0.05$.

Computational Time to Solve a Simulated Case

The effects of different factors on the computational time are listed in Table 6. The computational time was based on Macintosh Quadra 700 with 25 MHz 68040 CPU with math coprocessor. The Information Content Model main effect, Strategy main effect, and Decision Criterion main effect were all statistically significant. There was no interaction effect among the independent variables.

Note that the LR information content model was implemented according to the partial information passing method II (new method), which was much more computationally extensive than the partial information passing method 1 (simplified method used in current Iliad). The computational time for the LR model here did not reflect the current Iliad computational speed. The comparison between two partial information passing methods will be given later in this chapter.

Table 6. The Effects of Information Content Model, Strategy, and Decision Criterion on Computational Time (Seconds)

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 3.79	0.006
Strategy (Strat)	F(1. 42) = 10.61	0.002
Decision Criterion (Rule)	F(1, 42) = 10.12	0.003
Infomod by Strat	F(4, 168) = 1.05	0.385
Infomod by Rule	F(4, 168) = 0.27	0.898
Strat by Rule	F(1, 42) = 0.01	0.938
Infomod by Strat by Rule	F(4,168) = 2.19	0.072

The means of the computational time from different algorithms are given in Figure 29. The average time for probability dependent information content models (Shannon, P2-P1, LogP2-LogP1, $\underline{M}$ = 590.4 seconds) was significantly longer than for probability independent information content models (LR, LogLR, $\underline{M}$ = 474.7 seconds). The average time among probability dependent models was not significant different. The average time among probability independent models was also not significant different. The average time for the across-frame strategy ($\underline{M}$ = 586.4 seconds) was significantly longer than for the single-frame strategy ($\underline{M}$ = 501.7 seconds). The average time for the expected decision criterion ($\underline{M}$ = 607.3 seconds) was significantly longer than for the maximax decision criterion ($\underline{M}$ = 480.7 seconds).

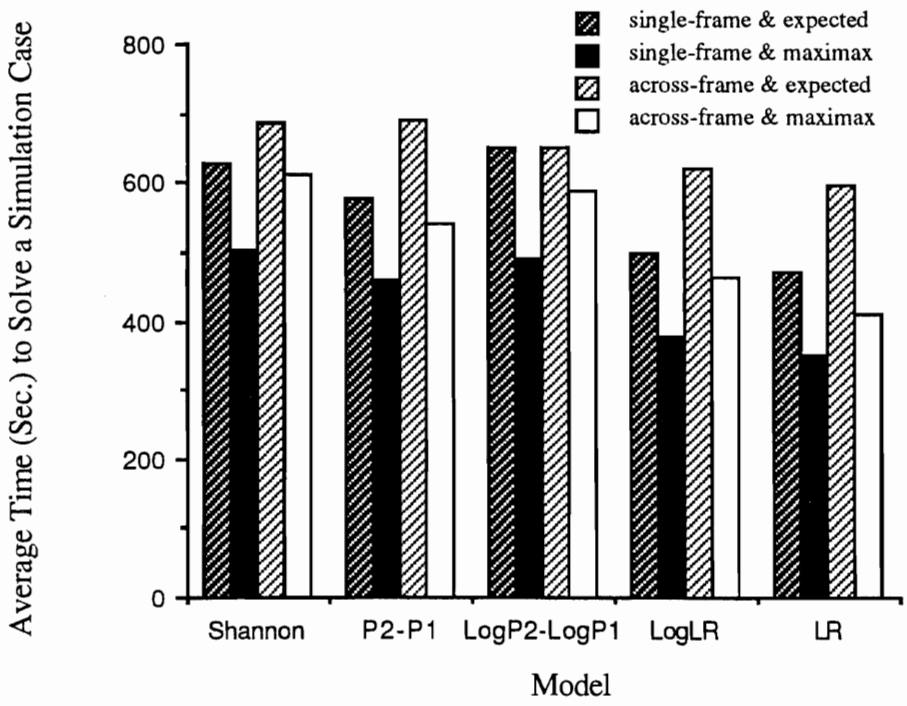

Figure 29. Average computational time for each algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and"maximax." Two strategies are indicated by "single-frame" and "across-frame."

Number of Steps to Solve a Simulated Case

The effects of different factors on the number of steps are listed in Table 7. The Information Content Model main effect, Strategy main effect, and Decision Criterion main effect were all statistically significant. There was no interaction effect among the independent variables. The means from different algorithms are indicated in Figure 30. The average number of steps for probability dependent models ($\underline{M}$ = 30) was significantly more than that for probability independent models ($\underline{M}$ = 22). There was no performance difference in terms of number of steps among the probability dependent models, and there was no difference among the probability independent models. It took fewer steps for the single-frame strategy ($\underline{M}$ = 25) to solve a simulated case than the across-frame strategy ($\underline{M}$ = 28). It took fewer steps for the maximax decision criterion ($\underline{M}$ = 24) to solve a simulated case than for the expected criterion ($\underline{M}$ = 29).

Table 7. The Effects of Information Content Model, Strategy, and Decision Criterion on Number of Steps to Solve a Simulated Case

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 11.83	< 0.001
Strategy (Strat)	F(1. 42) = 13.24	0.001
Decision Criterion (Rule)	F(1, 42) = 7.55	0.009
Infomod by Strat	F(4, 168) = 0.37	0.83
Infomod by Rule	F(4, 168) = 0.7	0.596
Strat by Rule	F(1, 42) = 0.13	0.725
Infomod by Strat by Rule	F(4,168) = 1.66	0.161

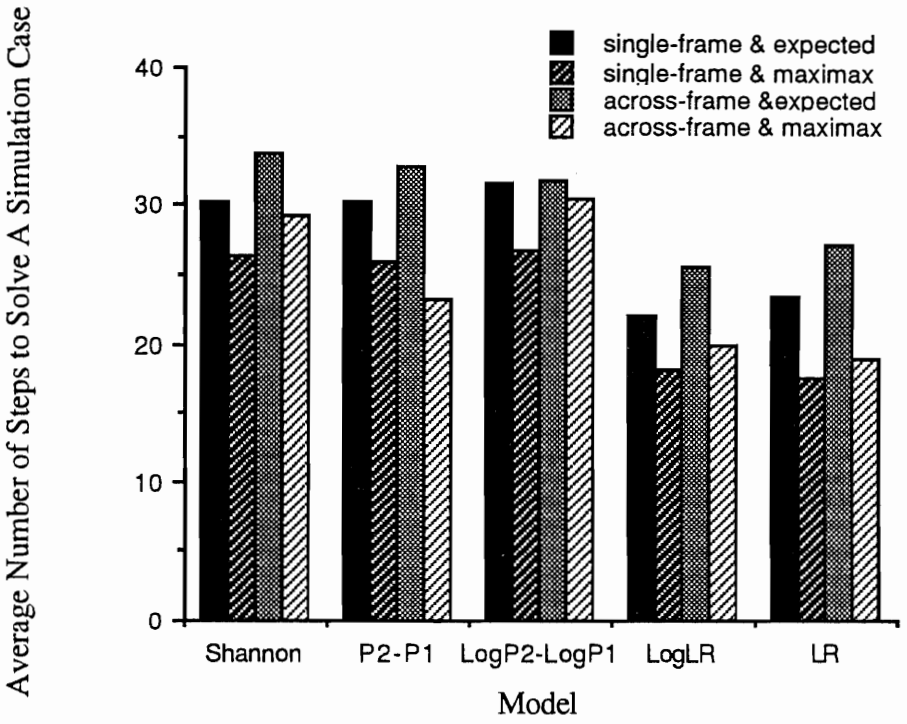

Figure 30. Average number of steps for each algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and"maximax." Two strategies are indicated by "single-frame" and "across-frame."

Cumulative Charges ($) for Results-Known Laboratory Tests Ordered

During the process of solving a simulated case, different laboratory tests were requested by the algorithms. Some tests were recorded in the simulated case, but some were not. Results of those tests recorded in the case are available to Iliad and are called results-known laboratory tests. The effects of different factors on the cumulative charges for results-known laboratory tests ordered are listed in Table 8. The Information Content Model main effect was statistically significant. The strategy main effect and the Decision Criterion main effect were not statistically significant. There was no interaction effect among the independent variables. The mean cumulative charges for results-known laboratory tests ordered from different algorithms are indicated in Figure 31. There was no significant difference among algorithms derived from probability dependent models. There was also no significant difference among algorithms derived from probability independent models.

Note that mathematically LogLR model is more sensitive to the cost factor than the LR model because of the logarithm of the numerator. However the average charges for results-known laboratory tests by the LR model ($\underline{M}$ = $422) were not significantly different from the charges by the LogLR model ($\underline{M}$ = $393).

Table 8. The Effects of Information Content Model, Strategy, and Decision Criterion on Average Cumulative Charges ($) for Results-Known Laboratory Tests to Solve a Simulated Case

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 3.52	0.009
Strategy (Strat)	F(1. 42) = 0.34	0.561
Decision Criterion (Rule)	F(1, 42) = 0.02	0.888
Infomod by Strat	F(4, 168) = 1.38	0.242
Infomod by Rule	F(4, 168) = 1.25	0.29
Strat by Rule	F(1, 42) = 2.65	0.111
Infomod by Strat by Rule	F(4,168) = 0.67	0.614

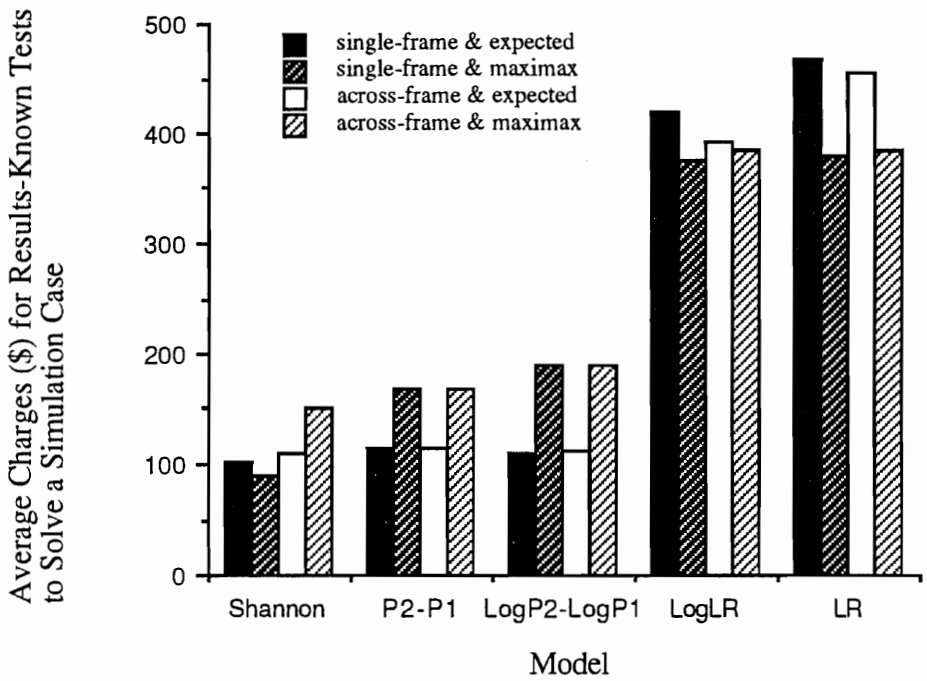

Figure 31. Average cumulative charges ($) for results-known tests asked by Iliad for each algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and "maximax." Two strategies are indicated by "single-frame" and "across-frame."

On average, the probability dependent information content models (M = $135) accrued fewer charges than the probability independent information content models (M = $408). Cost for laboratory charges was not significantly different between the single-frame strategy (M = $242) and the across-frame strategy (M = $247). Charges were also not significantly different between the expected decision criterion (M = $240) and the maximax decision criterion (M = $248).

Cumulative Charges ($) for All Laboratory Tests Ordered

The effects of different factors on the cumulative charges for all laboratory tests ordered to solve a simulated case are listed in Table 9. The Information Content Model main effect was statistically significant. The Decision Criterion main effect was also statistically significant. Strategy main effect was not statistically significant. There was no interaction effect among the independent variables. The mean cumulative charges for all laboratory tests ordered from different algorithms are indicated in Figure 32.

There was no significant difference among algorithms derived from probability dependent models. However there was a significant difference among algorithms derived from probability independent models. On average (across the single-frame and across-frame strategies), the charges by LogLR model with the maximax decision criterion ($\underline{M}$ = $980) were significantly more than those with the expected decision criterion ($\underline{M}$ = $645). There was no significant difference among algorithms derived from LR models. The average cumulative charges for all laboratory tests ordered by algorithms derived from LR model ($\underline{M}$ = $768) were not significantly different from the charges by algorithms derived from LogLR model ($\underline{M}$ = $812).

Table 9. The Effects of Information Content Model, Strategy, and Decision Criterion on Average Cumulative Charges ($) for All Laboratory Tests Ordered to Solve a Simulated Case

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 6.03	< 0.001
Strategy (Strat)	F(1. 42) = 0.35	0.559
Decision Criterion (Rule)	F(1, 42) = 4.71	0.036
Infomod by Strat	F(4, 168) = 0.25	0.907
Infomod by Rule	F(4, 168) = 1.84	0.123
Strat by Rule	F(1, 42) = 0.4	0.529
Infomod by Strat by Rule	F(4,168) = 0.7	0.592

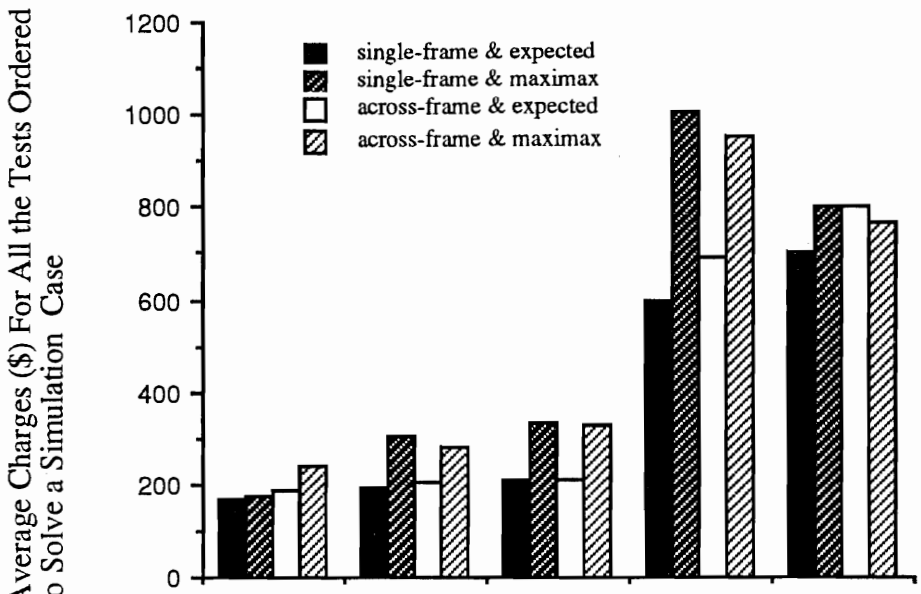

Figure 32. Average cumulative charges ($) for all the tests asked by Iliad for each algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and "maximax." Two strategies are indicated by "single-frame" and "across-frame."

On average, the charges by algorithms derived from probability dependent models (M = $236) were significantly less than those from probability independent models (M = $790). The charges by algorithms derived from the across-frame strategy (M = $465) were not significantly different from those by the the single-frame strategy (M = $450). Finally, the charges by algorithms derived from the expected decision criterion (M = $397) were significantly less than those by the maximax decision criterion (M = $519).

Number of Results-Known Laboratory Tests Ordered

The effects of different factors on the number of results-known laboratory tests ordered to solve a simulated case are listed in Table 10. The Information Content Model main effect, the Decision Criterion main effect and Strategy main effect were all not statistically significant. There was no any interaction effect among the independent variables. The means from different algorithms are indicated in Figure 33.

There was no significant difference between probability dependent information content models (M = 1.58) and probability independent information content models (M = 1.53). Also, there was no significant difference between the across-frame strategy (M = 1.57) and the single-frame strategy (M = 1.55). Finally, there was also no significant difference between the expected decision criterion (M = 1.61) and the maximax decision criterion (M = 1.51).

Table 10. The Effects of Information Content Model, Strategy, and Decision Criterion on Number of Results-Known Laboratory Tests Ordered to Solve a Simulated Case

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 1.32	0.265
Strategy (Strat)	F(1. 42) = 1.1	0.299
Decision Criterion (Rule)	F(1, 42) = 0.75	0.391
Infomod by Strat	F(4, 168) = 0.56	0.695
Infomod by Rule	F(4, 168) = 0.34	0.853
Strat by Rule	F(1, 42) = 3.01	0.09
Infomod by Strat by Rule	F(4,168) = 0.57	0.686

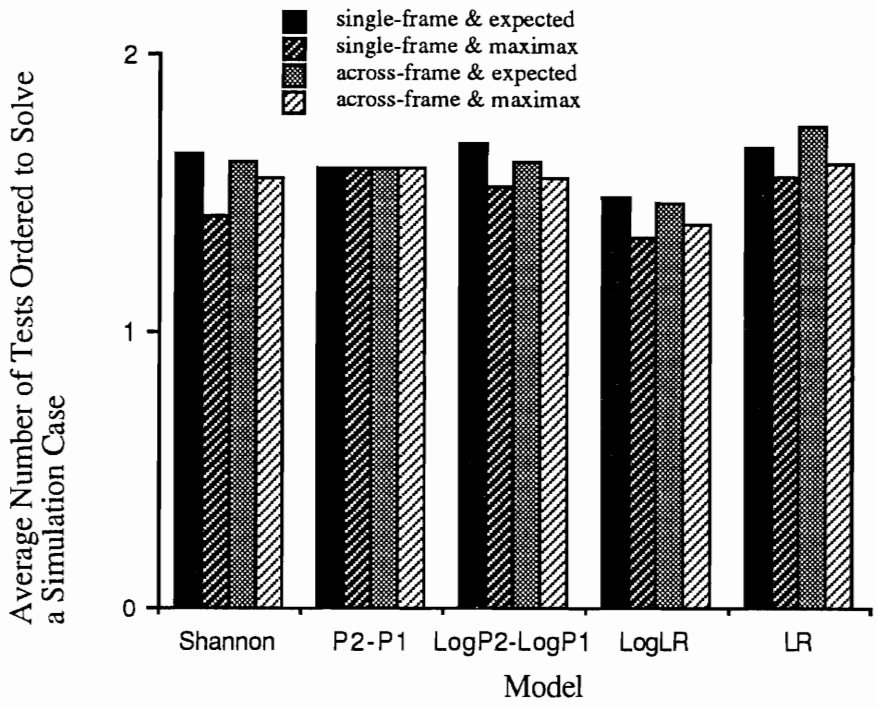

Figure 33. Average number of results-known tests for each algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and "maximax." Two strategies are indicated by "single-frame" and "across-frame."

Number of All Laboratory Tests Ordered

The effects of different factors on the number of all laboratory tests ordered to solve a simulated case are listed in Table 11. The Information Content Model main effect was statistically significant. The Decision Criterion main effect and Strategy main effect were not statistically significant. There was no any interaction effect among the independent variables. The means from different algorithms are indicated in Figure 34.

The number of all laboratory tests ordered to solve a simulated case by algorithms derived from probability dependent models ($\underline{M}$ = 2.64) was significantly less than the number of ordered by algorithms from probability independent models ($\underline{M}$ = 3.21). There was no significant difference between the across-frame strategy ($\underline{M}$ = 2.86) and the single-frame strategy ($\underline{M}$ = 2.87). There was also no significant difference between the expected decision criterion ($\underline{M}$ = 2.71) and the maximax decision criterion ($\underline{M}$ = 3.03).

Table 11. The Effects of Information Content Model, Strategy, and Decision Criterion on Number of All Laboratory Tests Ordered to Solve a Simulated Case

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 2.43	0.05
Strategy (Strat)	F(1. 42) = 0.04	0.849
Decision Criterion (Rule)	F(1, 42) = 1.04	0.313
Infomod by Strat	F(4, 168) = 0.72	0.581
Infomod by Rule	F(4, 168) = 2.08	0.086
Strat by Rule	F(1, 42) = 0.01	0.916
Infomod by Strat by Rule	F(4,168) = 0.24	0.914

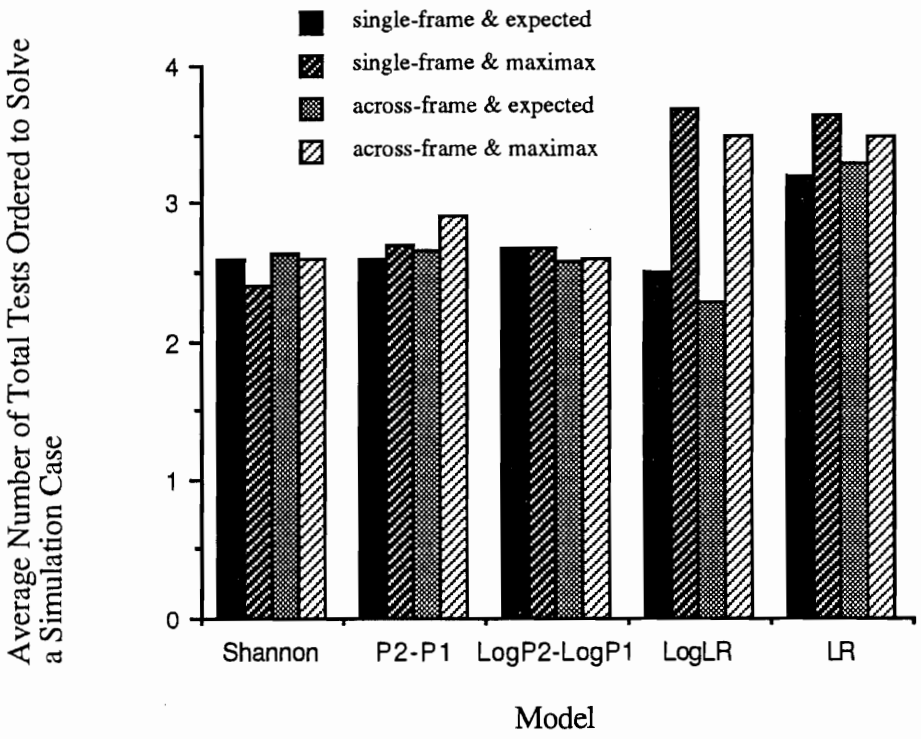

Figure 34. Average number of all tests ordered for each algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and"maximax." Two strategies are indicated by "single-frame" and "across-frame."

Number of History Questions Asked to Solve a Simulated Case

The effects of different factors on the number of history questions asked to solve a simulated case are listed in Table 12. The Information Content Model main effect was statistically significant. The Decision Criterion main effect and Strategy main effect were not statistically significant. The means from different algorithms are indicated in Figure 35. The average number of history questions requested to solve a simulated case by algorithms derived from probability dependent information content models (M = 34.0) was significantly more than that by probability independent information content models (M = 25.0). There was no significant difference between the across-frame strategy (M = 29.5) and the single-frame strategy (M = 30.6). There was also no significant difference between the expected decision criterion (M = 30.2) and the maximax decision criterion (M = 29.8).

A 3 × 2 × 2 (Information Content Model × Strategy Decision × Criterion) factorial analysis of variance among algorithms derived from probability dependent models (Shannon, P2-P1, LogP2-LogP1) indicated that there were no main effects by independent variables of Information Content Model, Strategy, and Decision Criterion. There were no interaction effects among the independent variables. A 2 × 2 × 2 (Information Content Model × Strategy Decision × Criterion) factorial analysis of variance among algorithms derived from probability independent models (LogLR, LR) indicated again there are no main effects by independent variables of Information Content Model, Strategy, and Decision Criterion. There are no interaction effects among those variables.

The interaction between Information Content Model and Decision Criterion was significant. For LogP2-LogP1 model, the average number of history questions under the expected decision criterion (M = 29) was significantly less than that under the maximax decision criterion (M = 36). In contrast, for LogLR model, the average number of history questions under the expected decision criterion (M = 29) was significantly more than that under the maximax decision criterion (M = 23). For models of Shannon, P2-P1 and LR models, the expected criterion and the maximax criterion were not significantly different.

Table 12. The Effects of Information Content Model, Strategy, and Decision Criterion on Number of History Questions Requested to Solve a Simulated Case

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 5.02	0.001
Strategy (Strat)	F(1. 42) = 1.11	0.299
Decision Criterion (Rule)	F(1, 42) = 0.04	0.84
Infomod by Strat	F(4, 168) = 1.86	0.12
Infomod by Rule	F(4, 168) = 3.97	0.004
Strat by Rule	F(1, 42) = 0.02	0.88
Infomod by Strat by Rule	F(4,168) = 1.61	0.173

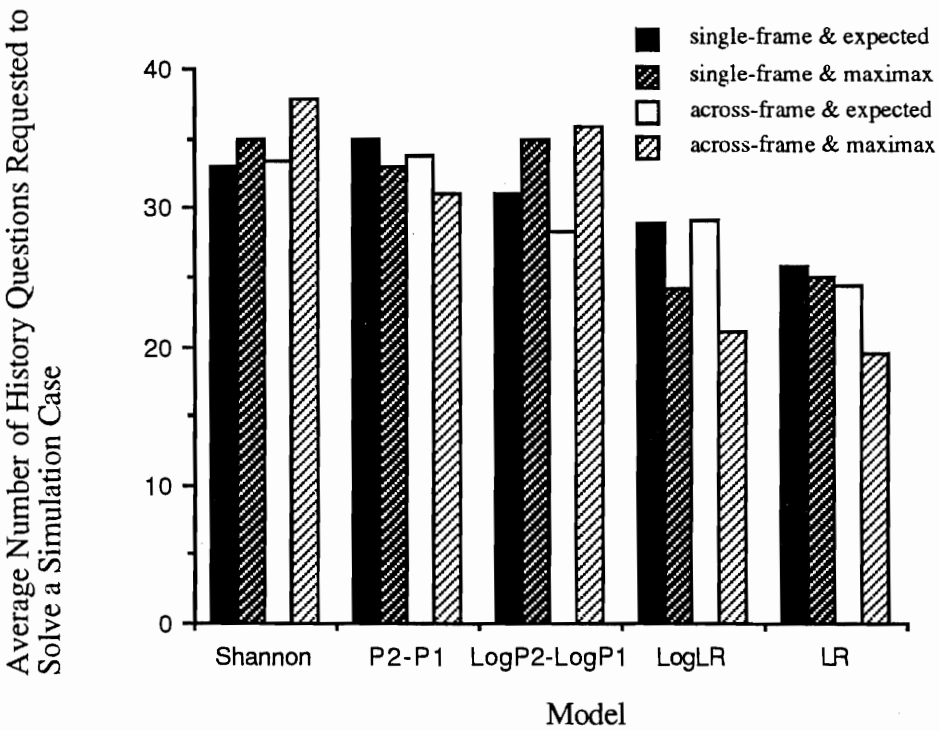

Figure 35. Average number of history questions requested for each algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and "maximax." Two strategies are indicated by "single-frame" and "across-frame."

Number of Physical Exam Questions Asked to Solve a Simulated Case

The effects of different factors on the number of physical exam questions asked to solve a simulated case are listed in Table 13. The Information Content Model main effect was nearly but not statistically significant. The Decision Criterion main effect and Strategy main effect were not statistically significant. The means from different algorithms are indicated in Figure 36.

A $3 \times 2 \times 2$ (Information Content Model $\times$ Strategy Decision $\times$ Criterion) factorial analysis of variance among algorithms derived from probability dependent models (Shannon, P2-P1, LogP2-LogP1) indicated that there were no main effects by independent variables of Information Content Model, Strategy, and Decision Criterion. There were no interaction effects among the independent variables. A $2 \times 2 \times 2$ (Information Content Model $\times$ Strategy Decision $\times$ Criterion) factorial analysis of variance among algorithms derived from probability independent models (LogLR, LR) indicated again there are no main effects by independent variables of Information Content Model, Strategy, and Decision Criterion. There are no interaction effects among those variables.

Table 13. The Effects of Information Content Model, Strategy, and Decision Criterion on Number of Physical Exam Findings Requested to Solve a Simulated Case

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 2.32	0.059
Strategy (Strat)	F(1. 42) = 0.47	0.496
Decision Criterion (Rule)	F(1, 42) = 0.25	0.619
Infomod by Strat	F(4, 168) = 2.56	0.04
Infomod by Rule	F(4, 168) = 1.27	0.285
Strat by Rule	F(1, 42) = 0.25	0.619
Infomod by Strat by Rule	F(4,168) = 1.67	0.16

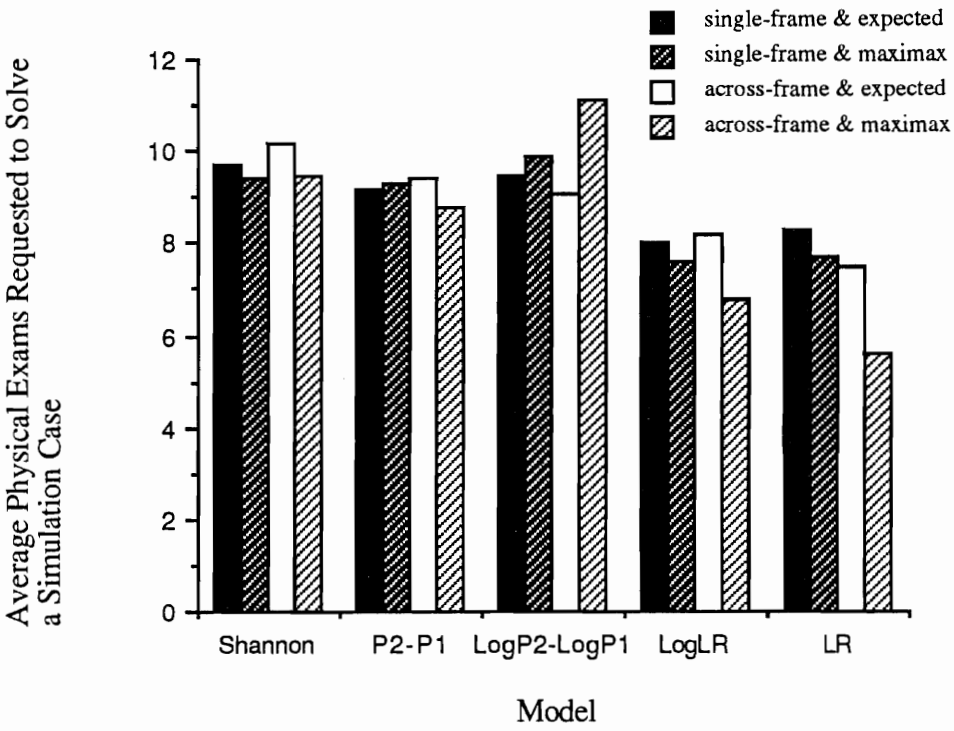

Figure 36. Average number of physical exam findings requested for each algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and"maximax." Two strategies are indicated by "single-frame" and "across-frame."

On average, there was no significant difference between the probability dependent information content models ($\underline{M}$ = 9.6) and the probability independent information content models ($\underline{M}$ = 7.5). There was no significant difference between the across-frame strategy ($\underline{M}$ = 8.6) and the single-frame strategy ($\underline{M}$ = 8.9). There was also no significant difference between the expected decision criterion ($\underline{M}$ = 8.9) and the maximax decision criterion ($\underline{M}$ = 8.5).

Number of History and Physical Exam Questions Asked to Solve a Simulated Case

The effects of different factors on the number of history and physical exam questions asked to solve a simulated case are listed in Table 14. The Information Content Model main effect was statistically significant. The Decision Criterion main effect and Strategy main effect were not statistically significant. The means from different algorithms are indicated in Figure 37. There was significant difference between the probability dependent information content models (M = 43.6) and probability independent information content models (M = 32.5). There was no significant difference between the across-frame strategy (M = 38.1) and the single-frame strategy (M = 39.5). There was no significant difference between the expected decision criterion (M = 39.1) and the maximax decision criterion (M = 38.3).

A $3 \times 2 \times 2$ (Information Content Model × Strategy Decision × Criterion) factorial analysis of variance among algorithms derived from probability dependent models (Shannon, P2-P1, LogP2-LogP1) indicated that there were no main effects by independent variables of Information Content Model, Strategy, and Decision Criterion. There were no interaction effects among the independent variables. A $2 \times 2 \times 2$ (Information Content Model × Strategy Decision × Criterion) factorial analysis of variance among algorithms derived from probability independent models (LogLR, LR) indicated again there were no main effects by independent variables of Information Content Model, Strategy, and Decision Criterion. There were no interaction effects among those variables.

The interaction between Information Content Model and Decision Criterion was significant. There were no other interaction effects among the independent variables. For LogP2-LogP1 model, the average number of history and physical exam questions under the expected decision criterion (M = 39.0) was significantly less than that under the maximax decision criterion (M = 46.0). In contrast, for LogLR model, the average number of history questions under the expected decision criterion (M = 37.0) was significantly

Table 14. The Effects of Information Content Model, Strategy, and Decision Criterion on Number of History Questions and Physical Exam Findings Requested to Solve a Simulated Case

Effect of Factor	F Value	p Value
Information Content Model (Infomod)	F(4, 168) = 4.47	0.002
Strategy (Strat)	F(1. 42) = 1.05	0.312
Decision Criterion (Rule)	F(1, 42) = 0.08	0.78
Infomod by Strat	F(4, 168) = 2.18	0.074
Infomod by Rule	F(4, 168) = 3.65	0.007
Strat by Rule	F(1, 42) = 0.08	0.784
Infomod by Strat by Rule	F(4,168) = 1.67	0.16

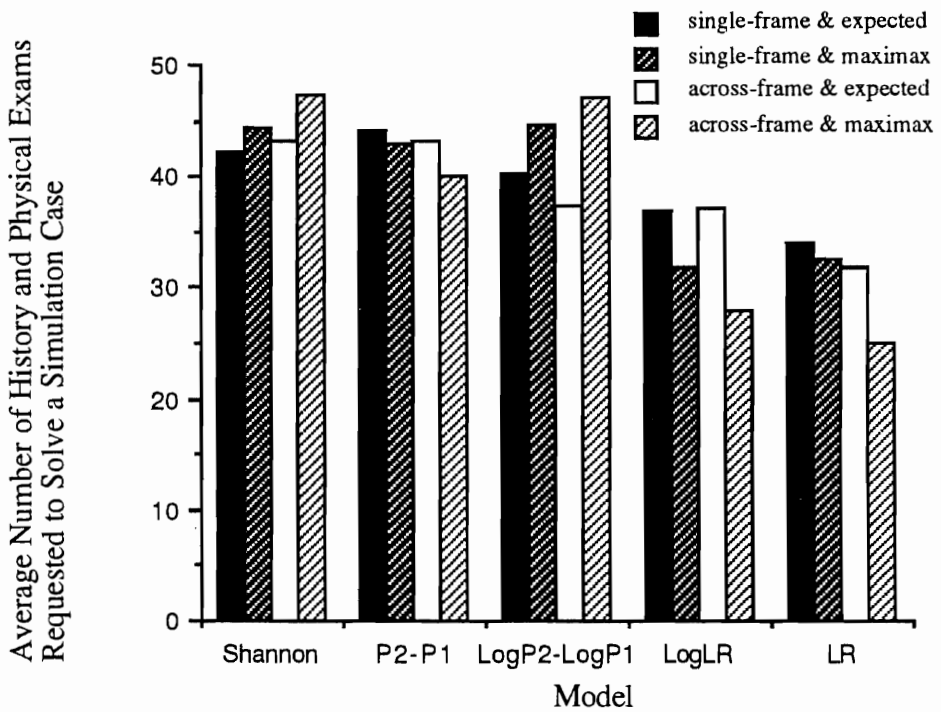

Figure 37. Average number of history questions and physical exam findings requested for an algorithm to solve a simulated case. Two decision criteria are indicated by "expected" and "maximax." Two strategies are indicated by "single-frame" and "across-frame."

more than that under the maximax decision criterion ($\underline{M}$ = 30.0). For models of Shannon, P2-P1 and LR models, the expected criterion and the maximax criterion were not significantly different.

Comparison of Two Information Passing Method by Simulated Study

Objective of the Comparison

There are two partial information passing methods for the LR model. Partial information passing method I represents the method currently used in Iliad. Partial information passing method II represents a new method that takes advantage of Iliad's inference engine and back calculates the "equivalent" likelihood ratios based on posterior probabilities. The Iliad with the LR model under the single-frame strategy and maximax criterion in Experiment III was based on the partial information passing method II. The two methods represent two mechanisms of implementing the LR model into Iliad. The two mechanisms have not been previously compared and evaluated. The objective of this comparison was to evaluate the performance of two partial information passing methods by the simulated study technique. Two corresponding versions of Iliad were used to solve the same 70 simulated cases in Experiment III.

Experimental Design

The independent variables are Partial Information Passing Method I (Method I) and Partial Information Passing Method II (Method II). The independent variables are the same as the nine measures used in the Experiment III. The comparisons were based on paired t-tests.

Results

Fifty-seven cases (81%) were solved by Iliad version with the Method I (current version of Iliad). Fifty-nine cases (84%) were solved by Iliad version the Method II.

There were 54 simulated cases (77%), which were solved by both versions of Iliad. The paired t-test comparisons were all based on the complete sets of data from those 54 cases.

The average time for Iliad with the Method I to solve a simulated case was 92 (seconds), and the time for Iliad with the Method II was 333 (seconds). The computational time was based on a Macintosh Quadra 700 with 25 MHz 68040 CPU with math coprocessor. Paired t-test indicated that Iliad with Method I was significantly faster than Iliad with the Method II in terms of the computational time to solve a simulated case.

The average number of steps for Iliad with Method I to solve a simulated case was 26, as compared to 18 with Method II. Paired t-test indicated that Iliad with Method I used significantly more steps to solve a simulated case than Iliad with Method II.

To solve a simulated case, Iliad with Method I and Iliad with Method II were not significantly different in terms of the charges for results-known laboratory tests, charges for all laboratory tests, the number of results-known laboratory tests ordered, the number of all laboratory tests ordered, the number of history questions requested, and the number of physical exam questions requested.

DISCUSSION

Summary of Results

Information Content Model

Experts preferred the Shannon's information content model (vignette study I and II). If this model was implemented in Iliad, the work-up suggestions would be more like the suggestions of experts. The Shannon's model accomplished work-ups that were significantly less costly than work-ups performed by the LR (likelihood ratio) model (simulation study). Implementing the Shannon's model would increase the cost-effectiveness of Iliad's suggestions. However, the Shannon's model required additional computational resources and more history and physical examination steps than the LR model. Decisions regarding implementing alternative models require Iliad researchers to balance the relative merits of cost, steps, expert preference, and other important factors.

Strategy

Physicians preferred algorithms that summed information across the relevant diseases being considered (vignette study II). It appears that the across-frame strategy may improve Iliad's ability to model diagnostic behavior of expert physicians. If the result can be replicated and as microprocessors improve the across-frame strategy is recommended to be adopted into Iliad.

In the implementation of the across-frame strategy (vignette study II), multiple findings produced by single diagnostic procedures, such as chest X-ray examinations or batteries of laboratory tests (e.g., Chem-20), were evaluated together to give a total value for the information from the procedure. However, history and physical examination findings were treated individually. Physicians tend to collect data in "chunks," which are

related groups of findings such as questions concerning shortness of breath. Future study is needed to group history and physical examination findings in a way preferred by expert physicians.

Decision Criterion

The expected decision criterion is recommended to be used at the early stage of a work-up to avoid expensive test ordering (simulation study). The maximax decision criterion is recommended to be used at the late stage of a work-up when confirmatory tests needed to be performed. The "cut-point" between the early and late stage of a work-up can not be specified from the current experiments.

Vignette Studies

In the first vignette study, five information content models were implemented into Iliad based on the current searching method. Algorithms derived from the modified Shannon model and P2-P1 model performed better than the current algorithm derived from LR model in terms of experts' judgments about the appropriateness of the findings selected and the best overall combination of the hypothesis and suggested findings. The LogLR model performed better than the LR model (current model), but the LogLR model was not as good as the Shannon and P2-P1 models. The LogP2-LogP1 model performed as well as the modified Shannon and P2-P1 models initially but worsened in the middle and late work-up stages.

In the second vignette study, the five information content models and two strategies were implemented into Iliad based on a new searching method. This implementation was to ensure any combination of an information content model and a strategy could be used to pursue the most cost-effective work-up. The modified Shannon information model was the best model overall, regardless of strategy. The Shannon's model was significantly better than the LR model during the initial stage of a work-up when history and physical exam findings were the major items to acquire. During the late stage of work-up when the

patient's major history and physical exam features were known, the LR model was just as good as the Shannon's model. These results also confirmed the first vignette study findings, which suggested that Shannon's model was preferable to other models especially during the initial phase of a patient case.

Physicians typically generate a differential diagnosis early in the work-up of a patient case. They then pursue findings that allow them to separate potential diagnostic competitors (37). The single-frame and the across-frame strategies have been used to model this process. The single-frame strategy, which is present in the current version of Iliad, evaluates the relative cost-effectiveness that each diagnostic finding has in relation to each hypothesis on the differential. This strategy allows Iliad to rank-order each possible diagnostic finding and select the best one. However, this strategy treats each finding and disease link independently. In some cases, obtaining one finding may provide positive information for one hypothesis and negative information for another. For instance, a chest X-ray may be ordered to work up a patient with shortness of breath when the physician is considering pneumonia versus pneumothorax. If the chest X-ray shows a pneumothorax, and not an infiltrate, information accrues (positive and negative) for both diagnostic hypotheses. The single-frame strategy does not reflect the combined information available in a group of diseases to which a particular finding may be relevant. The result is an underestimation of the total information value of findings that contribute to multiple diagnoses. This may explain why Iliad sometimes delays obtaining tests such as chest X-rays even when experts feel they are indicated. It appears that the across-frame strategy may improve Iliad's performance. The relative performance of different information models was closer to each other when the across-frame strategy was used as compared to the single-frame strategy. This finding indicates that the performance of best information algorithm in Iliad depends not only on the information content model, but also on the strategy with which the model is implemented.

The two vignette studies have indicated that (1) history and physical exam findings suggested by Shannon's information content model were chosen most frequently by expert physicians regardless of two strategies Iliad used; (2) in terms of suggesting laboratory test procedures, Shannon's model and the LR model (current model) were not significantly different; no models performed better than the LR model regardless of two strategies; (3) Suggestions by the across-frame strategy were chosen more frequently by expert physicians than ones by the single-frame strategy.

The vignette approach was designed to have experts' feedback about the performance of different algorithms. This feedback was essential to the development of best information algorithm for Iliad. However there are some limitations with this approach. This approach takes lots of experts' time to evaluate different algorithms for a large number of vignettes. The cost for experts' time is large. Because of the time and budget concerns, the number of vignettes had to be limited and the number of algorithms also had to be limited. Because many potential algorithms need to be investigated, it is hard to have every one of them evaluated by expert physicians.

Simulation Study

A simulation study is inexpensive and fast, and the process of data collection can be automated. The number of algorithms and simulation cases used in an experiment can be relatively large. Both vignette and simulation approaches are necessary to evaluate the performance of different algorithms. The vignette approach provides the experts' judgments about the performance of an algorithm at different stages of a work-up. The simulation approach provides objective measures about the overall performance of an algorithm when solving a patient case. The simulation approach can serve as a screening method to evaluate potential algorithms; the vignette approach can then be used to validate the results from the simulation study.

The simulation study showed that the performance of Iliad's best information algorithm depends on the information content model, the strategy, and the decision criterion with which the model is implemented.

The simulation study has shown the characteristics of two categories of information content models. One category is represented by probability dependent information content models, Shannon, P2-P1 and LogP2-LogP1. The other category is represented by probability independent information content models, LogLR and LR.

In terms of computational time to solve a simulation case, algorithms with probability dependent models were slower than those with probability independent models. This resulted from the fact that algorithms with probability dependent models took more steps and requested more questions to solve a simulated case. Algorithms with the across-frame strategy were slower than those with the single-frame strategy because of the extra considerations of multiple linkage between findings and diseases. Algorithms with the expected decision criterion were slower than those with the maximax decision criterion because the expected decision criterion required the frequency of a finding to calculate the weighted average information content provided by the finding; the maximax decision criterion did not require the frequency value to obtain the information content provided by the finding.

The algorithms with Shannon, P2-P1 and LogP2-LogP1 models took more steps than those with LogLR and LR models to solve a simulated case, but the former type of algorithms use much less amount of total charges (for both known and unknown tests) to solve a simulated case than the latter type of algorithms. Algorithms with Shannon, P2-P1 and LogP2-LogP1 models simulated one type of diagnostic work-up behavior with more emphasis on history and physical examination questions. The work-up by these algorithms was cheaper but more conservative. Algorithms with LogLR and LR models simulated another type of diagnostic work-up behavior with more attention on the expensive tests.

The more expensive algorithms ignore the prior probability of disease being considered before ordering a laboratory test.

Mathematically LogLR model is more sensitive to the cost factor than the LR model. However the simulation study did not indicate cheaper work-ups by algorithms with LogLR model than those by LR model. This finding suggests that taking the logarithm of likelihood ratio value can not make LR model more sensitive to work-up cost in terms of solving simulated cases.

Algorithms with the across-frame strategy took more steps than those with the single-frame strategy to solve a simulated case. There were no other differences between the two strategies. Note that only exactly the same findings in different disease frames were evaluated together to give a total value of information in the across-frame strategy implementation for the simulation study. In other words, medical findings were grouped according to the lowest level of findings hierarchy.

Iliad represents medical findings with six hierarchical levels. The first level indicates general categories of findings (e.g., Medical History, Physical Examination, Chemistry, Blood Bank, Radiology, etc.). The second level indicates the type of examination performed, name of the test, or procedure. For example, with history data the level 2 terms include "present history," "previous history"; with physical examination data the level 2 terms include "vital signs," "chest auscultation," etc; with laboratory tests or procedures the level 2 terms include "CBC," "Chest x-ray," etc. Meaningful clinical terms are found from level 3 to level 6. If findings at higher level of the hierarchy other than the lowest level were evaluated together, there might be different kinds of performance from two strategies. For example, present history findings of shortness of breath at rest (F1) and short of breath on exertion (F2) are two different findings. For the across-frame strategy implementation in the simulation study, information provided by F1 and F2 were not grouped together because they were not exactly the same findings. However if F1 and F2 were grouped at the upper level of the hierarchy, shortness of breath (noun level), the information content

value contributed by either F1 or F2 would be the total information value contributed by the two findings. Different implementations of the across-frame strategy might affect the performance of the best information algorithm of Iliad. Further investigation is needed to study characteristics of different across-frame strategy implementation methods.

Algorithms with expected decision criterion took more steps to solve a simulated case than those with maximax decision criterion. However the former type of algorithms consumed more charges (for known and unknown tests) than the latter type of algorithms. This may result from the fact that the maximax decision criterion is relatively more "aggressive" than the expected decision criterion.

The simulation study has indicated that two types of behaviors can be simulated. Algorithms with probability dependent information content models (Shannon, P2-P1, LogP2-LogP1) simulated the first type of diagnostic behavior, which was characterized by pursuing more history and physical examination findings, less laboratory tests, less expensive work-ups, and more steps to solve a patient case. Algorithms with probability independent models (LogLR and LR) simulated the second type of behavior, which was characterized by pursuing less history and physical examination findings, more laboratory tests, more expensive work-ups, and less steps to solve a patient case. Algorithms with expected decision criterion took more steps to solve a patient case than those with maximax decision criterion. However the expected decision criterion consumed less charges than the maximax decision criterion.

The simulation study has shown that the partial information passing method I and partial information passing method II were different in terms of the computational time and number of steps to solve a simulated case. The method I (currently used in Iliad) was faster even though it took more steps to solve a simulated case. This indicated that the method I was very computationally efficient compared to the method II. However method II was more "goal" oriented than the method I. This is because the likelihood ratio values are directly back calculated from the posterior probabilities by method II.

Future Studies

There are a variety of information content measures. These measures could be evaluated by the simulation approach first and a few selected measures can be further evaluated by the expert physicians. Renyi (38) suggests entropy of order α, which is defined by

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{m} p_i^\alpha \right), \quad \alpha > 0, \alpha \neq 1 \tag{4-1}$$

$$H_1(P) = -\sum_{i=1}^{m} p_i \log p_i \tag{4-2}$$

where $P = (p_1, p_2, ..., p_m)$, $\sum_{i=1}^{m} p_i = 1$ and $p_i \geq 0$ $i = 1, 2, ..., m$.

For $\alpha = 1$, Reni's entropy is the same as Shannon's entropy. The variation of α values provides a continuum of uncertainty measures from Renyi's entropy. Ben-Bassat suggests a possible information measure by the divergence measure of order α between P and Q (39):

$$D_\alpha (Q|P) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{m} q_i^\alpha \, p_i^{1-\alpha} \quad \alpha > 0 \quad \alpha \neq 1 \tag{4-3}$$

$$D_1(Q|P) = \sum_{i=1}^{m} q_i \log \frac{q_i}{p_i} \tag{4-4}$$

where the p_i represent the a priori probabilities and the q_i represent the corresponding posterior probabilities. The properties of Renyi's entropy are discussed by Aczel and Daroczy (40) and Ben-Bassat (41). Another possible information measure that is suggested by Ben-Bassat can be expressed as

$$V(Q|P) = \sum_{i=1}^{m} p_i \, (q_i - p_i)^2 \tag{4-5}$$

Eq. [4-5] is the expected weighted variance of the a posterior probability vector.

Patton and Woolfenden defined diagnostic utility (DU) of a finding as follows (42):

$$DU = 2\,P(D)\,(Se - Sp) + 2\,Sp - 1 \qquad (4 - 6)$$

where P(D) is the probability of the disease being considered, Se is the sensitivity of the finding for the disease, and Sp is the specificity of the finding for the disease. As stated by Patton and Woolfenden, DU is not an inherent property of a diagnostic test but of test-observer interactions in a clinical setting. The proposed diagnostic utility model is simple but incorporates important clinical decision analytic variables sensitivity, specificity, and disease probability.

In the current implementation of quasi-utility models, only dollar charges of clinical findings are considered. If data about the real costs of obtaining clinical findings were available, further studies could be done to investigate the effect of charges and true cost values on the performance of Iliad's best information algorithm. The costs of medical tests as a whole should include the cost of personnel, equipment, indirect overhead, test-related morbidity and mortality, the waste due to imperfect test performance, patient discomfort, waiting time for test results, etc. The estimation of different cost factors and how to include cost factors properly into the quasi-utility model present a challenge for the future research. Even though the task is not easy, these potential additions to the best information algorithm should improve the ability of Iliad to simulate the multifaceted environment in which real data collection decisions are made.

Clute (43) and Peterson (44) showed that general practitioners did far less history taking and physical examination than was deemed appropriate. If scoring options were based on data gathering, Marshall (45) found that experts gathered less data and as a result scored lower than relative juniors when patient management problems were presented to people with different level of expertise. The simulation study of different best information algorithms has shown two types of behaviors from algorithms with probability dependent information content models (Shannon, P2-P1, and LogP2-LogP1) and algorithms with

probability independent information content models (LR and LogLR). Algorithms with probability dependent models tend to pursue more history and physical examination questions than algorithms with probability independent models. A further study could be done on Iliad's test mode. The hypotheses are that one type of scoring algorithm based on probability dependent models such as Shannon model will score medical students higher than experts and another type of scoring algorithm based on probability independent models such as LR model will score medical students lower than experts. To examine the hypotheses, test cases could be administered to medical students and physicians at different expertise levels. Test scores could then be analyzed as an index to expertise.

Gorry and Barnett (46) applied a "real utility" approach to pursue additional tests. They represented utility values by dollar amount in terms of costs of tests (e.g., patient discomfort, time of skilled persons, etc.) and costs of possible misdiagnoses. A matrix was built to describe the costs associated with misdiagnoses. In the following matrix, TD_1, TD_2, ..., TD_n represent the actual diseases being covered. D_1, D_2, ..., D_n represent

	Actual Disease					
Diagnosis	TD_1	TD_2	...	TD_j	...	TD_n
D_1	u_{11}	u_{12}		u_{1j}		u_{1n}
D_2	u_{21}	u_{22}		u_{2j}		u_{2n}
.						
.						
.						
D_i	u_{i1}	u_{i2}		u_{ij}		u_{in}
.						
.						
.						
D_n	u_{n1}	u_{n2}		u_{nj}		u_{nn}

the diagnosis list. The utility for each pair (D_i, TD_j) u_{ij} represents the cost of misdiagnosis of TD_j with D_i. The cost for having the "right" diagnosis is 0 (e.g., $u_{11} = 0$, ... , $u_{nn} = 0$).

The assignment of utility values in the study was arbitrary. However the method presented provided an interesting approach to searching for cost-effective tests. Given a test result, the posterior probabilities at a particular stage of a work-up can be calculated. The expected utility values resulted from possible misdiagnosis given a test can be calculated according to the posterior probabilities and the utility values in the matrix. The utility value of a test was the summation of expected misdiagnosis cost resulting from performing the test and the cost of the test itself. The test with the lowest utility value is considered to be the best. If a similar approach could be taken, it would be very desirable to compare a real utility model to a quasi-utility model in terms of suggesting cost-effective findings.

The quasi-utility model was implemented in Iliad based on the sequential searching strategy. To save computational time, the sequences of multiple tests were not included in the implementation. There is a possibility that the sequential searching method may overlook a very cost-effective pathway that consists of several tests, none of which is particularly effective alone. As the speed of microprocessors continues to improve, it will become practical to consider multiple sequences of tests.

Finally, with the improvements in the best information algorithm of Iliad, a study involving a clinical trial to measure the impact of Iliad on the cost-effectiveness of the patient work-up can be performed.

APPENDIX A

VIGNETTE RATING FORM FOR EXPERIMENT I

In the first vignette study, three real pulmonary disease cases were selected. Each case was divided into 6 vignettes, so there were total 18 vignettes. Two sample vignettes from one case are given below.

Vignette 1

Patient Findings

Five medical students were asked to work-up a real patient case. These students had obtained the following findings part of the way through their initial work-up:

Age: 85
Sex: Female
Patient has the following findings:
Present history:
chest pain
that is sharp or stabbing
that is radiating
to the shoulder, neck, arm or jaw

Rating Section

The five students have proposed the following work-up plans:

Student 1, 4 & 5: Pursue Unstable Chest Pain Pattern -> from Stable Angina by asking:
"Present history of chest pain with a specific time pattern, with a specific aggravating factors, at rest?"
"Previous medical history of cardiac problem?"

Student 2: Pursue Acute Productive Cough -> from Pneumonia by asking:
"Present history of cough?"

Student 3: Pursue Stable Angina by asking:
"Previous lab and other tests: positive test for coronary artery disease?"

Section 1: Findings Score

Please score the cost-effectiveness of the question(s) proposed by each student for that student's diagnosis. Please do not score based on the appropriateness of the student's hypothesis. The score should range from 5 (similar to the cost-effectiveness of questions typically proposed by the best students you have known) to 1 (similar to the cost-effectiveness of questions typically proposed by the worst students you have known).

Student 1, 4 & 5: _____ (1 to 5)

Student 2: _____ (1 to 5)

Student 3:_____ (1 to 5)

Section 2: Best Overall Score

Please pick the best overall strategy (combination of disease to be pursued and cost-effectiveness of question for the disease). Put the number(s) of the student(s) who selected the best strategy in the blank line below:

Number(s) of the best student(s) _____

Vignette 2

Patient Findings

Now, we continue the work-up of the case. Despite your previous rating of the "best" student to be allowed to pursue the work-up, a single student was arbitrarily selected to continue. This student collected the following additional findings:

Present history:
- **chest pain**
  - **that did begin recently**
  - **that does come on suddenly**
  - that is not recurrent
  - **that is increased by breathing deeply**
  - **that is increased by coughing**
  - **that is at rest**
- No recent unusual or severe physical activity
- No major injury or trauma

Previous Medical History:
- **Cardiac problem**

Physical Exam:
- No tenderness to pressure or palpation

The complete case history has now been updated and is printed below. Please note that the underlined terms are the newly acquired findings.

Age: 85
Sex: Female
Patient has the following findings:

Present history:
- **chest pain**
  - **that did begin recently**
  - **that does come on suddenly**
  - that is not recurrent
  - **that is sharp or stabbing**
  - **that is radiating**
    - **to the shoulder, neck, arm or jaw**
  - **that is increased by breathing deeply**
  - **that is increased by coughing**
  - **that is at rest**
- No recent unusual or severe physical activity
- No major injury or trauma

Previous Medical History:
Cardiac problem
Physical Exam:
No tenderness to pressure or palpation

Rating Section

The five students have proposed the following work-up plans:

Student 1, 2 & 3: Pursue Acute Productive Cough -> from Pneumonia by asking: "Present history of cough?"

Student 4: Pursue Risk of Atherosclerosis -> from Acute MI by asking:
"Previous medical history of hypertension, peripheral vascular disease, coronary heart disease, diabetes?"
"Family history of cardiovascular disease?"

Student 5: Pursue Risk of Atherosclerosis -> from Acute MI by asking:
"Previous medical history of coronary heart disease?"

Section 1: Findings Score

Please score the cost-effectiveness of the question(s) proposed by each student for that student's diagnosis. Please do not score based on the appropriateness of the student's hypothesis. The score should range from 5 (similar to the cost-effectiveness of questions typically proposed by the best students you have known) to 1 (similar to the cost-effectiveness of questions typically proposed by the worst students you have known).

Student 1, 2 & 3: _____ (1 to 5)

Student 4: _____ (1 to 5)

Student 5: _____ (1 to 5)

Section 2: Best Overall Score

Please pick the best overall strategy (combination of disease to be pursued and cost-effectiveness of question for the disease). Put the number(s) of the student(s) who selected the best strategy in the blank line below:

Number(s) of the best student(s) _____

APPENDIX B

VIGNETTE RATING FORM FOR EXPERIMENT II

Age: 52
Sex: Male
Chief Complaint(s):
Dyspnea: fatigue;

Patient has the following findings:

Present history:
shortness of breath (dyspnea)
on exertion
that is worsened recently
not at rest
that does not cause the patient to awaken at night (PND)
easy fatigue (reduced exercise capacity)
palpitations (sensation of heart beat)
Previous medical history:
no neurologic or nervous problems
no chronic lung disease
no embolus or embolism

Hypotheses under consideration:

Ischemic Cardiomyopathy
Heart Failure
Idiopathic Cardiomyopathy
Hypertensive Heart Disease
Pneumonia

Rating Sheet

Student	Strategy I	Strategy II	Best Strategy
1	vital signs: systolic blood pressure, diastolic blood pressure, respiratory rate, heart rate, oral temperature	present history: cough, shortness of breath that is worsened when lying flat, diminished urine (oligurial), nocturia, decreased mental abilities	
2	vital signs: systolic blood pressure, diastolic blood pressure, respiratory rate, heart rate, oral temperature	present history: cough that is of recent onset, with gross hemoptysis, with blood streaked sputum shortness of breath that is worsened when lying flat, diminished urine (oliguria), nocturia, decreased mental abilities previous medical history coronary heart disease (myocardial infarction)	
3	vital signs: systolic blood pressure, diastolic blood pressure, respiratory rate, heart rate, oral temperature previous medical history: coronary heart disease (myocardial infarction)	previous medical history coronary heart disease (myocardial infarction)	
4	vital signs: systolic blood pressure, diastolic blood pressure, respiratory rate, heart rate, oral temperature	previous medical history coronary heart disease (myocardial infarction)	
5	vital signs: systolic blood pressure, diastolic blood pressure, respiratory rate, heart rate, oral temperature	present history cough, shortness of breath that is worsened when lying flat, diminished urine (oligurial), nocturia, decreased mental abilities	
Best Student			

REFERENCES

[1] WARNER, H.R., HAUG, P., BOUHADDOU, O., LINCOLN, M., WARNER, H.R. JR, SORENSON, D., WILLIAMSON, J.W., AND FAN, C.L. Iliad as an expert consultant to teach differential diagnosis. "Proceedings of the Symposium on Computer Applications in Medical Care." **12**, 371, IEEE Computer Society Press, Washington, DC (1988).

[2] WARNER, H.R. The Iliad program: An expert computer diagnostic program. *Med. Pract. Management* 123 (Fall 1992).

[3] SORENSON, D.K., CUNDICK, R.M., FAN, C., AND WARNER, H.R. Passing partial information among Bayesean and Boolean frames. "Proceedings of the Symposium on Computer Applications in Medical Care." **13**, 50, IEEE Computer Society Press, Washington, DC (1989).

[4] LINCOLN, M.J., TURNER, C.W., HAUG, P.J., WARNER, H.R., WILLIAMSON, J.W., BOUHADDOU, O., JESSEN, S.G., SORENSON, D., CUNDICK, R.C., AND GRANT, M. Iliad training enhances medical students' diagnostic skills. *J. Med. Systems* **15** (1), 93 (1991).

[5] WARNER, H.R., RUTHERFORD, B.D., AND HOUTCHENS, B. A sequential Bayesean approach to history taking and diagnosis. *Comput. Biomed. Res.* **5**, 256 (1972).

[6] HAUG, P.J., WARNER, H.R., CLAYTON, P.D., SCHMIDT, C.D., PEARL, J.E., FARNEY, R.J., CRAPO, R.O., TOCINO, I., MORRISON, W.J., AND FREDRICK, P.R. A decision-driven system to collect the patient history. *Comput. Biomed. Res.* **20**, 193 (1987).

[7] HAUG, P.J., ROWE, K.G., RICH, T., FARNEY, R., PEARL, J., SCHMIDT, C.D., AND CLAYTON, P.D. A comparison of computer-administered histories. *Proceedings AAMSI Congress* **21** (1988).

[8] BELL, D.E., KEENEY, R.L., AND RAIFFA, H. "Conflicting Objectives in Decisions." John Wiley & Sons, New York, 1977.

[9] KEENEY, R.L., AND RAIFFA, H. "Decisions with Multiple Objectives: Preferences and Value Tradeoffs." John Wiley & Sons, New York, 1976.

[10] von NEUMANN, J., AND MORGENSTERN, O. "Theory of Games and Economic Behavior," 2nd ed. John Wiley and Sons, Princeton University Press, Princeton, 1953.

[11] SAVAGE, L. J. "The Foundations of Statistics," 2nd ed. Dover, New York, 1972.

[12] FRENCH, S. "Decision Theory, An Introduction to The Mathematics of Rationality." Ellis Horwood Limited, New York, 1986.

[13] ALLAIS, M., AND HAGEN, O. "Expected Utility Hypotheses and the Allais Parasox." D. Reidel, Dordrecht, 1979.

[14] STIGUM, B.P., AND WENSTΔP, F. "Foundations of Utility and Risk Theory with Applications." D. Reidel, Dordrecht, 1983.

[15] SCHOEMAKER, P.J.H. The expected utility model: Its variants, purposes, evidence and limitations. *J. Econ. Lit.* **20**, 529 (1982).

[16] PEARL, J. "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference," 2nd ed. Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[17] HALEY, R.W. Measuring the costs of nosocomial infections: Methods for estimating economic burden on the hospital. *Am. J. Med.* **Suppl 3B**, 3B-32S (1991)

[18] WEINSTEIN, M.C., AND STASON, W.B. Foundations of cost-effectiveness analysis for health and medical practices. *N. Engl. J. Med.* **296**, 716 (1977)

[19] USHER, M. J. "Information Theory for Technologists." Macmillan Publishers, London, 1984.

[20] SHANNON, C. E., AND WEAVER, W. "The Mathematical Theory of Communication." University of Illinois Press, Urbana, IL, 1964.

[21] DEBER, R.B., AND GOEL, V. Using explicit decision rules to manage issues of justice, risk, and ethics in decision analysis: When is it not rational to maximize expected utility. Med. Dec. Making **10**, 181 (1990).

[22] GUO, D., LINCOLN, M.J., HAUG, P.J., TURNER, C.W., WARNER, H.R. Exploring a new best information algorithm for Iliad. "Proceedings of the Symposium on Computer Applications in Medical Care." **15**, 624, McGRAW-HILL, INC, Washington, DC (1991).

[23] BHARATH, R. Information theory. *BYTE*. December, 291 (1987).

[24] GUO, D., LINCOLN, M.J., HAUG, P.J., TURNER, C.W., AND WARNER, H.R. Comparison of different information content models by using two strategies: Development of the best information algorithm for Iliad. "Proceedings of the Symposium on Computer Applications in Medical Care." **16**, 465, McGRAW-HILL, INC, Baltimore, Maryland (1991).

[25] JOHNSON, H.A. Diagnosis by the bit: A method for evaluating the diagnostic process. *Ann. Clin. Lab. Sci.* **19** (5), 323 (1989).

[26] ASCH, D.A., PATTON, J.P., AND HERSHEY, J.C. Knowing for the sake of knowing: The value of prognostic information. *Med. Dec. Making* **10** (1), 47 (Jan-Mar 1990).

[27] PITKEATHLY, D.A., EVANS, A.L., AND HAMES, W.B. The use of information theory in evaluating the contribution of radiological and laboratory investigations to diagnosis and management. *Clin. Rad.* **30**, 643 (1979).

[28] WARNER, H.R. "Computer-Assisted Medical Decision-Making." Academic Press, New York, 1979.

[29] GOOD, I.J., AND CARD, W.I. The diagnostic process with special reference to errors. *Meth. Inform. Med.* **29** (1), 61 (1990).

[30] THORNBURY, J.R., FRYBACK, D.G., AND EDWARDS, W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. *Diagn. Rad.* **114**, 561 (1975).

[31] DUNCAN, L.R., AND HOWARD, R. "Games and Decisions." John Wiley & Sons, Inc., New York, 1957.

[32] DANIEL, K., AND AMOS, T. Prospect theory: An analysis of decision under risk. *Econometrica.* **47**, 263 (1979).

[33] GARDENFORS, P., AND NILS-ERIC, S. "Decision, Probability, and Utility Selected Readings." Cambridge University Press, Cambridge, 1988.

[34] FRIEDMAN, M., AND SAVAGE, L. J. The expected-utility hypothesis and the measurability of utility. *J. Pol. Econ.* **60**, 463 (1952).

[35] ARROW, K. J. "Essays in the Theory of Risk-Bearing." Markham, Chicago, 1971.

[36] Hays, W. L. "Statistics," 4th ed. Holt, Rinehart and Winston, Inc., Fort Worth, 1988.

[37] ELSTEIN, A.S., SHULMAN, L.S., AND SPRAFKA, S. A. Medical problem solving -- a ten year retrospective. *Eval. Health Prof.* **13**, 5 (1990).

[38] RENYI, A. On measures of entropy and information. *Proc. 4th Berkley Symp. Math. Stat. Prob.* **1**, 547 (1960).

[39] BEN-BASSAT, M. Myopic policies in sequential classification. *IEEE Trans. Comput.* **C-27** (2), 170 (1978).

[40] ACZEL, J., AND DAROCZY, Z. "On measures of information and their characterization." Academic Press, New York, 1975.

[41] BEN-BASSAT, M. Renyi's entropy and the probability of error. *IEEE Trans. Inform. Theory* **IT-24** (3), 324 (1978).

[42] PATTON, D. D., WOOLFENDEN, J. M. A utility-based model for comparing the cost-effectiveness of diagnostic studies. *Invest. Rad.* **24** (4), 263 (1989).

[43] CLUTE, K. F. "The General Practitioner." University of Toronto Press, Toronto, Ontario, Canada, 1963.

[44] PETERSON, O. L., ANDREWS, L. P., SPAIN, R. S., AND GREENBERG, B. G. An analytical study of North Carolina general practice: 1953-1954. *J. Med. Educ.* **31**, Part 2, December (1956).

[45] MARSHALL, J. Assessment of problem solving ability. *Med. Educ.* **11**, 329 (1977).

[46] GORRY, G. A., AND BARNETTE, G. O. Experience with a model of sequential diagnosis. *Comput. Biomed. Res.* **1**, 490 (1968).