An Evaluation of Explanations of Probabilistic Inference

H.J. Suermondt* and Gregory F. Cooper**

*Section on Medical Informatics
Stanford University, Stanford, CA

**Section of Medical Informatics
University of Pittsburgh, Pittsburgh, PA

ABSTRACT

Providing explanations of the conclusions of decision-support systems can be viewed as presenting inference results in a manner that enhances the user's insight into how these results were obtained. The ability to explain inferences has been demonstrated to be an important factor in making medical decision-support systems acceptable for clinical use. Although many researchers in artificial intelligence have explored the automatic generation of explanations for decision-support systems based on symbolic reasoning, research in automated explanation of probabilistic results has been limited.

We present the results of an an evaluation study of INSITE, a program that explains the reasoning of decision-support systems based on Bayesian belief networks. In the domain of anesthesia, we compared subjects who had access to a belief network with explanations of the inference results, to control subjects who used the same belief network without explanations. We show that, compared to control subjects, the explanation subjects demonstrated greater diagnostic accuracy, were more confident about their conclusions, were more critical of the belief network, and found the presentation of the inference results more clear.

INTRODUCTION

Computers have brought about numerous improvements in medical care, for example, through digital imaging, automatic patient monitoring, and on-line literature access (Shortliffe and Perreault, 1990). Medical decision-support systems provide a means by which computers can bring additional improvements to medical care, as additional information resources for physicians. Decision-support systems should not be seen as the doc-in-a-box, intended to replace the physician, but rather, as information tools (Shortliffe, 1982; Miller and Maserie, 1990). By providing and organizing information that would otherwise not be easily available, these systems may help medical personnel to make better patient-care decisions.

Although numerous medical decision-support systems have been developed to date, the clinical use of such systems has been minimal so far. Teach and Shortliffe (Teach and Shortliffe, 1981) performed a study to determine what a medical expert system should offer before physicians would consider using it in their practice. The most prominent requirement given by physicians was the ability of the system to explain its advice.

In this paper, we discuss a system designed to provide insight into the reasoning of decision-support systems based on Bayesian belief networks. We present the results of a study in which we tested the effects of the resulting explanations on users in the domain of anesthesia.

BACKGROUND

Bayesian belief networks can be used to frame probabilistic knowledge in a representation that is explicit about the conditional dependencies and independencies among variables. A belief network is a directed, acyclic graph in which nodes represent stochastic variables and arcs among nodes represent probabilistic dependencies. Typically, the variables are discrete-valued; the possible values of each variable are mutually exclusive and exhaustive. Arcs are represented as conditional-probability distributions. Belief networks, also known as *probabilistic influence diagrams, causal probabilistic networks,* and *Bayesian networks,* are described in more detail in (Cooper, 1989; Horvitz et al., 1988; Pearl, 1988).

Belief networks can be used to determine how a set of *findings* (the *evidence*) affects the probabilities of other, unknown, variables. We can observe, as a finding, any variable that is modeled in the network: symptoms (for example, if we want to determine the probability of one or more diagnoses), diseases (for example, if we want to predict the probability of observing a particular symptom), risk factors, intermediate nodes, and so on. Given a particular set of findings, we have, for each node in the network, a *prior* probability distribution (which reflects the baseline probabilities of the possible values of the node) and a *posterior* probability distribution (which contains the probabilities *given the observed findings*). We shall use the term *inference result* to refer to the changes from the prior to the posterior probability distribution for a particular variable of interest, given a set of findings.

Reasoning under uncertainty in belief networks takes place according to a paradigm consistent with probability theory (Horvitz et al., 1988). Therefore, the theoretical foundation for the conclusions of a system based on belief networks is strong, unlike the situation in systems that obtain results by heuristic methods (Cooper, 1989). Such a theoretical foundation has been demonstrated to improve the accuracy and consistency of systems' conclusions

0195-4210/92/$5.00

(Heckerman and Horvitz, 1987; Heckerman, 1990a), which may ultimately affect our ability to obtain user confidence in the system's conclusions.

Traditionally, the term *explanation*, in the context of automated reasoning, has had ambiguous meaning. In some expert systems—especially systems that reason in a goal-directed manner—the immediate goal for which a particular finding was needed served as the *explanation* of the system's request for that finding (Shortliffe, 1976). Thus, explanations for such systems focused primarily on intermediate reasoning steps. In some forward-chaining systems, on the other hand, the diagnostic conclusion of the system was presented as the *explanation* of the observed features (Miller et al., 1982). A comprehensive view of explanation is taken in the ABEL system (Patil et al., 1981). ABEL's explanations include the findings to be explained, the conclusion of the system, and the reasoning mechanism and intermediate steps leading to the conclusion. In this paper, we focus on the explanation of *inference results*; thus, we discuss the use of evidence, given a knowledge base, to reach conclusions.

Developers of decision-support systems that are based primarily on symbolic reasoning have often claimed that it is difficult, if not impossible, to explain the reasoning of systems that reason under uncertainty using probability theory (Davis, 1982; Clancey, 1983). Nonetheless, there have been several probabilistic decision-support systems that can explain their inferences to some degree. Among the most prominent is the Glasgow-Dyspepsia (GLADYS) system (Spiegelhalter and Knill-Jones, 1984). The explanation for this system consists of a table that allows the user to see which factors contribute to and which ones conflict with the conclusion. Other examples of probabilistic systems that can show the influence of the various findings on the inference results are discussed in (Cooper, 1984; Heckerman, 1990b; Henrion and Druzdzel, 1990). In this paper, we focus on a system called INSITE (Insight about Network Structure and Inference Through Explanation).

THE INSITE SYSTEM

We developed the INSITE system to provide users of belief networks with a means to dissect a belief-network inference problem, in order to answer the following question: *Why does the evidence E affect the marginal probability distribution of variable D in the way the system describes?* An intuitive reason for such a request for explanation is surprise. The user is confronted with the conclusions of the decision-support system (a set of probability distributions), and finds that these conclusions do not meet her expectations. Among the possible causes of such surprise are

- The set of findings is so large that the system user cannot determine properly the combined effect without analyzing which findings are most influential.
- There is conflict among the findings, so that the combined effect of the evidence is different from what the user would have expected.
- The user does not have an understanding of the chains of reasoning through which the findings affect the variable of interest.
- There is a difference in opinion between the user and the developers of the knowledge base regarding the conclusions that should be drawn from the findings.

The INSITE system enhances the user's insight into the inference results by highlighting the relationships between findings and conclusions, and by discussing the chains of reasoning through which the evidence affects the variable of interest. The resulting explanations allow system users to examine and evaluate the knowledge base modeled in the belief network, and to judge the appropriateness of the inferences based on that knowledge base.

INSITE runs on an Apple Macintosh II. The system has a graphical user interface that is standard for Macintosh applications; the graphical display is the primary focus of INSITE's explanations. For more detailed discussions (for example, of steps in chains of reasoning), these displays are supplemented with free text that is shown in a separate window. INSITE generates this text automatically by combining text fragments based on the results of its analyses, and by filling in the names of nodes and possible values of these nodes as needed.

INSITE can explain which node is affected most strongly by the evidence, which findings contribute to and which ones conflict with the inference result, which finding is most influential, how and why a particular finding affects some other variable, what arcs and chains of reasoning contribute to and conflict with the inference result, and why the overall evidence did or did not affect a particular variable of interest. For more details about INSITE, see (Suermondt, 1992). In the remainder of this paper, we shall discuss a study in which we evaluated the effects of INSITE's explanations on users of the system in the domain of anesthesia.

AN EVALUATION STUDY OF INSITE

Whenever a new methodology sees the light, we must determine whether the technique is merely *new*, or whether it is also useful; in medical informatics, we should determine whether the method has a potential for contributing to medical practice. In this section, we describe a preliminary study to determine the plausibility that the explanations provided by INSITE can have a beneficial effect on decisions made by users of belief-network–based decision-support systems in medical practice.

Due to pragmatic constraints on the scope of the study, we chose to investigate the effects of INSITE in one clinical domain, anesthesia, on cases for which we knew in advance that the belief network used by INSITE—the ALARM monitoring system—provided reasonable conclusions. We discuss our methods and procedures in more detail in the following subsection.

INSITE is a domain-independent program; the system is designed to be applicable to any belief network. To evaluate INSITE, however, it was necessary to select a particular belief network. We chose the ALARM belief network (Beinlich et al., 1989). ALARM was developed by Beinlich and associates as a research prototype of a system that aids anesthesiologists in the interpretation of monitor data during surgery. The network consists of 37 nodes that describe variables from cardiac and pulmonary physiology. The input to the network consists of findings that are monitored routinely during surgery. The output consists of the probabilities of several anesthetic emergency conditions. Among our reasons for selecting ALARM are (1) the system's diagnostic accuracy has been evaluated previously, in which study ALARM gave reasonable conclusions in a large percentage of cases (Beinlich and Gaba, 1989); (2) enthusiastic clinical collaborators were available in the Anesthesia Service at the Palo Alto VA Hospital; and (3) there is a relative abundance of subjects knowledgeable in anesthesia (in comparison with alternative domains for which we had readily available knowledge bases).

METHODS AND PROCEDURES

To evaluate INSITE's explanations, we presented 10 abstracted clinical cases to 6 residents and to 7 fourth-year medical students who had completed at least one clerkship in anesthesia; we compared the case assessments of subjects who used only ALARM to those of subjects who used ALARM plus INSITE's explanations of ALARM's conclusions.

The study had a case-by-case test-retest design; for each case, the subjects were asked first to give their clinical impression of the case (without use of the computer), including the differential diagnosis, the key abnormal findings (if any), and the action(s) to be taken next. After establishing their baseline assessments of a case, the subjects were given access to a computer interpretation of the case to aid in their analysis. For each case, one-half of the subjects were given access to ALARM only (without explanations); the remaining subjects not only could use ALARM, but also could use INSITE to generate explanations of ALARM's conclusions. After the subjects used the computer to interpret the case, they were asked once more to give their impression of the case.

The unit of measure for the evaluation study is the *subject case (SC)*, the interpretation of a single case by a single subject. We compared the baseline and follow-up assessments from the batch of SCs in which only ALARM was used (the *control batch, C*), to those from the batch of SCs in which subjects had access to the INSITE explanation facility (the *intervention batch, I*).

Cases Cases were designed with two criteria in mind: (1) ALARM has to be able to diagnose each case to a degree of accuracy that the staff anesthesiologists who were collaborators in the study found acceptable; and (2) the case has to be sufficiently difficult that decision support by the computer is potentially helpful. Each case consisted of a short vignette, describing a perioperative situation, and a snapshot of the monitor values at the time. The vignette contained information about the clinical history, the procedure being performed, and the time since surgery began. The monitor values consisted of a set of nine findings usually available to anesthesiologists during surgery.

After the set of 10 cases was generated, we tested the cases on ALARM—without the INSITE explanation facility—and verified, in each of the cases, that ALARM identified the "correct" diagnosis as the most probable one. We ensured that ALARM reached a reasonable conclusion for each case to prevent misleading the subjects with the computer's advice. Thus, in this study, improvement in user performance will correspond to enhanced agreement with the computer program's conclusions. As a result, any improvement in user performance can be used as a measure of the effect of the explanations: If performance improves with ALARM only (batch *C*), and improves more in ALARM plus INSITE (batch *I*), we can conclude that the difference is due to the addition of INSITE's explanations.

Assessments The availability of the explanation facility was determined by the *presentation mode* for a case. There are two presentation modes. In *control mode* (used for SCs in batch *C*), the subject saw the vignette and the monitor data, and was given use of ALARM (running under the INSITE interface, but with no explanations) to generate probabilistic conclusions about the case. In *intervention mode* (used for SCs in batch *I*), the subject had the same information as control-mode subjects, and in addition, was allowed to use INSITE to generate explanations for ALARM's advice.

For each case, subjects recorded a *baseline* and a *follow-up* assessment. First, the subject was given the vignette (describing the perioperative situation) and the monitor data, but was not allowed to apply ALARM or INSITE to the case. After the subject had been given sufficient time to interpret the case, she was asked to describe her baseline assessment of the clinical situation by means of the *baseline questionnaire*. The baseline questionnaire consists of three questions. The first question establishes the subject's initial impression of the perioperative situation. The subject either can mark that nothing appears to be abnormal, or can give a differential diagnosis. In the second question, the subject is asked to explain the answer to the first one: "If you suspect one or more problems, why?" Finally, in the third question, the subject is asked: "What would you do next?"; she is given a choice of four options, which correspond to the following four states of "confidence" about the case:

1. Keep monitoring the (normal) case
2. Seek information to try to determine whether there is a problem

3. There is a problem—try to differentiate among possibilities

4. Treat the problem—the subject is satisfied that she knows what is wrong

After completion of the baseline questionnaire, we entered the case into INSITE, and, depending on the presentation mode, the subject was allowed to obtain various pieces of information about the case from the computer. After the subject indicated that she had spent sufficient time exploring the case on the computer, she was asked once more to describe her clinical assessment of the case by means of the *follow-up questionnaire*. This second questionnaire was identical to the first one, except for the addition of three questions in which the subjects were asked to rate subjectively the computer's reasoning about the case on a seven-point scale: whether the information provided by the computer was useless or helpful, whether ALARM's model of the clinical situation was too simplistic or sufficiently complex, and whether—aside from ALARM's limitations—the information given by the computer was confusing or clear.

Design Details Subjects were selected as follows: We included *all* anesthesia residents at the Palo Alto VA Hospital; medical students were picked randomly from a list of fourth-year students who had completed at least one full clerkship in anesthesia. Subjects were matched pairwise for clinical experience in anesthesia, after which each pair was split randomly between two *groups* of subjects.

The cases were matched pairwise by difficulty (by the staff anesthesiologist who had generated the cases). Each pair was split randomly between two *sets* of cases, resulting in two randomized sets of cases that were matched in difficulty.

Once the groups of subjects and sets of cases were determined, we assigned (randomly) each group of subjects to control mode for one set of cases and to intervention mode for the other set. Each subject saw each case in only one mode; each subject saw one-half of the cases in control mode and the other half in intervention mode; and each case was seen by one-half of the subjects in control mode and by the other half in intervention mode.

To control for a possible bias generated by the order in which the cases were seen, we determined randomly for each subject which presentation mode the subject would see first. In this manner, we controlled for the fact that responses to INSITE's explanations might be influenced by previous experience with ALARM *without* explanations, or that judgments about ALARM (without explanations) might be affected by previous use of INSITE's explanation facility. In addition, we ordered the cases within each set randomly for each subject. Thus, we controlled for potential biases in subjects' responses due to exposure to other cases in the set.

Data Interpretation We analyze the results in three categories: diagnosis, actions, and opinions. We shall describe briefly the variables that we study in each category.

In the category *diagnosis*, we study the effects of computer advice on the user's differential diagnosis. For each SC, we have two differentials: one from the baseline assessment (written down by the subject before the computer was applied to the case), and the other from the follow-up assessment. For each diagnosis that appeared in any SC, we asked a staff anesthesiologist not involved in ALARM or INSITE to assess the correctness of the diagnosis, given the information available in the vignette and in the monitor data.

In the category *actions*, we look at the responses to the question regarding what the subject would do next. We study the conclusiveness of the responses on each SC. We have two measures for this conclusiveness: (1) the category of action, as described in the Assessments subsection; and (2) a subjective determination, from the changes in the list of actions, whether the subject has become more confident. The same anesthesiologist who determined the correctness of diagnoses assessed the relative confidence of actions informally, taking into account the difficulty and invasiveness of actions, as well as the degree to which these actions are performed routinely in the operating room.

In the third category, *opinions*, we study the answers to the subjective questions on each of the SCs. The subjective questions addressed the helpfulness of the computer's reasoning, the scope of ALARM's model, and the clarity of computer's presentation.

RESULTS

The results of the evaluation were encouraging. Among the striking conclusions of the study were that explanations by INSITE

- Prevented incorrect diagnoses from being added to the differential diagnosis
- Led to a more critical rating of ALARM's domain knowledge
- Increased the confidence with which users acted

Throughout this section, we shall use the terms *batch C* and *batch I* to refer to the SCs assessed in control mode and to those assessed in intervention mode, respectively.

In the category *diagnosis*, we found that the primary effect of INSITE was to *prevent* new incorrect diagnoses from being added to the differential. We show this result in Table 1. We can see that there are marked differences between *C* and *I* in terms of the changes from baseline to follow-up differential. The incorrect fraction of the differential shrank more often in batch *I* (16 SCs, versus 13 SCs in batch *C*); more significantly ($p = 0.01$ by Fisher's exact test), this incorrect fraction *grew* in 6 SCs

Table 1 Comparison of changes in differential diagnoses from baseline to follow-up assessments

Measure of effect	*C*	*I*	significance
number of SCs in which incorrect fraction shrank	13	16	
number of SCs in which incorrect fraction stayed the same	46	49	
number of SCs in which incorrect fraction grew	6	0	$p = 0.01$
number of SCs with new incorrect diagnoses	7	1	$p = 0.03$

Table 2 Comparison of changes in confidence from baseline to follow-up assessments

Measure of effect	*C*	*I*	significance
number of SCs with increased confidence (by action code)	20	24	
number of SCs with unchanged confidence (by action code)	41	37	n.s.
number of SCs with decreased confidence (by action code)	4	4	
number of SCs with increased confidence (by subjective assessment)	19	31	
number of SCs with unchanged confidence (by subjective assessment)	36	30	$p < 0.05$
number of SCs with decreased confidence (by subjective assessment)	10	4	

Table 3 Comparison of subjective assessments of the computer program

Measure of effect	*C*	*I*	significance
number of SCs where advice was rated worse than "somewhat helpful"	16	17	
number of SCs where advice was rated "somewhat helpful"	13	17	n.s.
number of SCs where advice was rated better than "somewhat helpful"	35	31	
number of SCs where ALARM's model was rated worse than "captures essence"	14	24	
number of SCs where ALARM's model was rated "captures essence"	41	30	$p = 0.10$
number of SCs where ALARM's model was rated better than "captures essence"	9	11	
number of SCs where presentation was rated worse than "clear"	20	21	
number of SCs where presentation was rated "clear"	44	36	$p = 0.01$
number of SCs where presentation was rated better than "clear"	0	8	

in batch *C*, whereas it *never* grew in batch *I*.* We obtain a different perspective on the same phenomenon by looking at the *number* of SCs in which the subject *added* incorrect diagnoses. In batch *C*, there were seven SCs in which there was at least one new incorrect diagnosis, compared to only one such SC in batch *I*. This difference is statistically significant ($p = 0.03$ by Fisher's exact test).

In the category *actions*, we found that subjects in batch *I* acted more confidently than did those in batch *C*. As mentioned before, we studied two variables: the action codes indicated by the subjects on their questionnaires, and the subjective determinations of confidence that were assessed by considering the subject's actions for the SC. We summarize the results in Table 2. Subject confidence increases more often in batch *I*, and decreases more often in batch *C*. This result is statistically significant only in the case of the *subjective* assessment of confidence: In batch *I*, 31 SCs showed *increased* subject confidence, versus 19 in batch *C*; on the other hand, 10 SCs in batch *C* showed *decreased* subject confidence, versus 4 in batch *I* ($\chi^2 = 6.0$, 2 d.f., $p < 0.05$).

In the category *opinions*, subjects rated three aspects of the computer program in their follow-up assessment of each case: helpfulness of the advice, degree to which ALARM's model captures the essence of the case, and clarity of presentation. In Table 3, we show the numerical results. From the results, it is clear that there were no significant differences in perceived "helpfulness" of the advice. Surprisingly, subjects found the advice more than "somewhat helpful" more frequently in batch *C* (35 SCs) than in batch *I* (31 SCs); however, this result is not statistically significant. The subjects found that ALARM's model was *insufficiently* detailed for the case in 24 SCs in batch *I*, versus 14 SCs in batch *C*. This difference in rating of ALARM's model—even though batches *C* and *I* used the same model—was statistically significant ($\chi^2 = 4.53$, 2 d.f., $p = 0.10$). On the other

* Of the six SCs in which the incorrect fraction grew, three were assessed by medical students and the other three by residents.

hand, the *clarity* of the advice was rated *better* than "clear" in 8 SCs in batch *I*, versus *no* SCs in batch *C*. The difference in clarity rating was highly significant ($\chi^2 = 8.82$, 2 d.f., $p = 0.01$).

DISCUSSION

The results of this evaluation study indicate that INSITE's explanations have a potential for improving user performance. The addition of explanations to ALARM's analyses improved subjects' diagnostic accuracy. The particularly telling results about diagnosis were that subjects who saw explanations were influenced to avoid adding incorrect diagnoses to their differential more often than did subjects who saw no explanations. The resulting differences in diagnostic accuracy between SCs assessed in control mode and those assessed in intervention (explanation) mode may be indicative of the use of the explanation for purposes of verification (Wick and Thompson, 1989). Verification can improve diagnostic accuracy primarily in cases where the user originally disagreed with ALARM's—correct—conclusions: Thanks to the explanations, the changes in probability indicated by ALARM become understandable—and therefore more credible—and are taken into account in the user's differential. Thus, subjects who have seen an explanation of a case may be more likely to include diseases for which the probability has increased, and less likely to include diseases for which the evidence led to a decrease in probability. The result is a more accurate user diagnosis.

The findings about user confidence indicate that subjects who are given explanations become *increasingly* confident more often than do subjects who are not given explanations, whereas subjects who are not given explanations become *less* confident more often than do those who receive explanations. The use of explanations for ratification, especially in cases where the user *agrees* with ALARM's conclusions, may explain why subjects who saw explanations became increasingly confident more often than did those who used ALARM alone (without explanations). On the other hand, we conjecture that a decrease in confidence due to the computer feedback takes place most often in cases where the user *disagrees* with the computer's analysis of the case. Since probabilistic reasoning does not always parallel the heuristics by which humans update their beliefs (Tversky and Kahneman, 1974), the results of probabilistic inference may be counterintuitive if they are not explained. If, through explanations, the user understands why ALARM reached its conclusion, the original disagreement may not result in a decrease in confidence; however, if it is *not* clear to the user why ALARM concluded what it did, the user may act less confidently.

Explanations did not affect substantially whether the users rated the computer's advice as "helpful" versus "useless." This is understandable if we take into account that the advice, regardless of the explanation, is the same in control mode as in intervention mode. However, we would have expected users to find advice complemented by explanations more helpful than advice without the explanations.

Interestingly, on SCs seen in intervention mode (that is, with explanations), subjects found ALARM's model overly simplistic more often than on SCs seen in control mode. This difference may indicate another aspect of the verification role of explanations: Even though ALARM diagnosed the cases in the study reasonably, the explanations demonstrated to the user how limited ALARM's model was, and how many alternative possibilities were not modeled in the belief network.

On the other hand, the *clarity* of the advice was rated higher on average in SCs where explanations were available. This result is intuitive; the explanations show the user what ALARM's advice means for the case at hand.

CONCLUSION

Due to pragmatic constraints on the study, we did not have an opportunity to determine fully the areas in which explanations of belief-network advice can have clinical utility. Rather, this study should be viewed as a pilot study, which gives us a first impression of the areas in which explanations may have an effect. Among the questions that remain for future study are the following:

1. What are the effects of INSITE in cases where the computer's conclusions are misleading? Does the explanation lead to false confidence in incorrect conclusions, or does it allow users to eliminate the incorrect advice?
2. Does the impact of explanations on diagnostic accuracy depend on the clinical domain? How do users of INSITE perform in domains other than anesthesia, as compared to control subjects?
3. In a real-world belief network in a clinical environment, would clinicians want to use the explanation facility (assuming that they would use the belief-network–based system), or would they take the computer's conclusions for granted?

Question 3 is particularly interesting, as generation of explanations increases the computational complexity significantly. Thus, when there is a cost to using the explanation facility—in terms of additional inference time, or an actual charge—users may find verification and ratification insufficiently valuable to justify asking for explanations.

In summary, the explanations by INSITE led to improved diagnostic accuracy and to increased user confidence. In addition, they helped the users to assess ALARM's scope more critically. The goal of the current evaluation study was to investigate whether INSITE can have a beneficial effect on decisions by users of medical decision-support systems. The results of this study—in particular, the beneficial effect of the explanations on diagnostic accuracy—support the hypothesis underlying INSITE, that explanations improve users' insight into probabilistic inference results, and that such enhanced insight can lead to improved decision making by medical practicioners.

Acknowledgments

Financial support for this work was provided by the National Science Foundation under grant IRI-8703710, by the U.S. Army Research Office under contract P-25514-EL, and by the National Institutes of Health under grants RR-00785 (Division of Research Resources) and LM-05208 (National Library of Medicine). This research benefited from many discussions with Edward Shortliffe, Ross Shachter, and Max Henrion. Jeremy Wyatt and Charles Friedman helped extensively with the design of the evaluation study and with the interpretation of results. David Gaba and Steven Howard provided the clinical expertise that was necessary for the evaluation study. Ingo Beinlich made the ALARM belief network available.

References

Beinlich, I.A., and Gaba, D.M. (1989). The ALARM monitoring system: Intelligent decision making under uncertainty. *Anesthesiology*, **71**(3A), A337.

Beinlich, I.A., Suermondt, H.J., Chavez, R.M., and Cooper, G.F. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medical Care*, London, 247–256. Springer Verlag, Berlin.

Clancey, W.J. (1983). The epistemology of a rule-based expert system: A framework for explanation. *Artificial Intelligence*, **20**(3), 215–251.

Cooper, G.F. (1984). *NESTOR: A Computer-Based Medical Diagnostic Aid that Integrates Causal and Diagnostic Knowledge*. Ph.D. Thesis, Program in Medical Information Sciences, Stanford University, Stanford, CA.

Cooper, G.F. (1989). Current research directions in the development of expert systems based on belief networks. *Applied Stochastic Models and Data Analysis*, **5**, 39–52.

Davis, R. (1982). Consultation, knowledge acquisition and instruction: A case study. In P. Szolovits (Ed.), *Artificial Intelligence in Medicine*, 57–78. Westview Press, Boulder, CO.

Heckerman, D.E. (1990a). An empirical comparison of three inference methods. In R.D. Shachter, T.S. Levitt, L.N. Kanal, and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 4*, 283–302. North-Holland, Amsterdam.

Heckerman, D.E. (1990b). *Probabilistic Similarity Networks*. Ph.D. Thesis, Program in Medical Information Sciences, Stanford University, Stanford, CA.

Heckerman, D.E., and Horvitz, E.J. (1987). On the expressiveness of rule-based systems for reasoning with uncertainty. In *Proceedings of the AAAI-87 Sixth National Conference on Artificial Intelligence*, Seattle, WA, 121–126. MIT Press, Cambridge, MA.

Henrion, M., and Druzdzel, M.J. (1990). Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, Cambridge, MA, 10–20. Association for Uncertainty in Artificial Intelligence, Mountain View, CA.

Horvitz, E.J., Breese, J.S., and Henrion, M. (1988). Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, **2**, 247–302.

Miller, R.A., and Maserie, F.E. (1990). The demise of the "Greek oracle" model for medical diagnostic systems. *Methods of Information in Medicine*, **29**, 1–2.

Miller, R.A., Pople, H.E., Jr., and Myers, J.D. (1982). INTERNIST-1: An experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, **307**, 468–476.

Patil, R.S., Szolovits, P., and Schwartz, W.B. (1981). Causal understanding of patient illness in medical diagnosis. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, British Columbia, 893–899. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1988). *Probabilistic Reasoning in Expert Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Shortliffe, E.H. (1976). *Computer-Based Medical Consultations: MYCIN*. Elsevier, New York.

Shortliffe, E.H. (1982). The computer and medical decision making: Good advice is not enough. *IEEE Engineering in Medicine and Biology Magazine*, **1**(2), 16–18.

Shortliffe, E.H., and Perreault, L.E. (Eds.). (1990). *Medical Informatics: Computer Applications in Health Care*. Addison-Wesley, Reading, MA.

Spiegelhalter, D.J., and Knill-Jones, R.P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society, Series A*, **147**, 35–77.

Suermondt, H.J. (1992). *Explanation in Bayesian Belief Networks*. Ph.D. Thesis, Program in Medical Information Sciences, Stanford University, Stanford, CA.

Teach, R.L., and Shortliffe, E.H. (1981). An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research*, **14**, 542–558.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124–1131.

Wick, M.R., and Thompson, W.B. (1989). Reconstructive explanation: Explanation as complex problem solving. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI, 135–140. Morgan Kaufmann, San Mateo, CA.