Reprinted from *Journal of the American Medical Informatics Association,* **1:** 127-141, 1994

Evaluation of a New Method for Cardiovascular Reasoning

William J. Long, PhD
MIT Laboratory for Computer Science, Cambridge, MA, USA
Shapur Naimi, MD and M. G. Criscitiello, MD
New England Medical Center and
Tufts University School of Medicine, Boston, MA, USA

Abstract

Objective: Evaluate the accuracy of the detailed diagnostic reasoning of the Heart Failure Program incorporating a new mechanism to handle temporal relationships and severity constraints.

Design: Tools were developed to summarize diagnoses and automatically generate evaluation forms. Five expert cardiologists were asked to review the reasoning of the program, with two analyzing each case. Cases were gathered retrospectively for diversity and difficulty and 26 randomly selected cases were evaluated. The underlying issues were identified and classified.

Results: Both reviewers rated the first diagnosis correct in 25% of the cases and at least one rated it wrong in 10%. Analyzing the detailed reasoning, 137 issues were raised, about 5.3 per case. 53% of these were possible concerns raised by one reviewer. Of the 5.3 issues per case, 2.5 were attributable to controversies, misunderstandings, or mistakes; 1 was due to the overly simplistic representation of the summaries; and 1.8 were issues related to the program.

Conclusion: Overall, the program is capable of providing high quality detailed diagnostic hypotheses for complex cardiovascular cases. The results highlight several issues: 1) the difficulty of effectively summarizing hypotheses, 2) the nature of a physician's causal explanation, and 3) some problems in evaluating detailed diagnostic reasoning. The mistakes the program made imply that some additional refinement is needed but that the reasoning mechanisms developed can support the appropriate reasoning. The appropriate next step is a prospective evaluation addressing the program's usefulness.

$$
\begin{tikzcd}
\text{FIXED CORONARY OBSTRUCTION} \arrow[d] & & \\
\text{REGIONAL FLOW DEFICIT} \arrow[d] & & \text{IHSS} \cdots \arrow[d] \\
\text{MYOCARDIAL ISCHEMIA ACUTE} \arrow[d] \arrow[drr] & & \text{GENERAL FLOW DEFICIT} \arrow[ll] \arrow[d] \\
\text{MYOCARDIAL INFARCTION} & & \text{UNSTABLE ANGINA}
\end{tikzcd}
$$

Figure 1: A portion of the HFP knowledge base

1 Introduction

The Heart Failure Program (HFP) is a computer program designed to assist physicians in reasoning about patients with cardiovascular disease[1, 2, 3]. It takes as input the description of the patient's history, physical examination findings, and test results at the level of detail one would find in the physician's description in the patient record. This information is used by the program to generate a differential diagnosis for the case. It attempts to generate a hypothesis corresponding to each likely combination of diseases that could account for the findings. The program uses a pseudo-Bayesian network as a knowledge base representing the physiologic causality of the cardiovascular hemodynamics from which it generates hypotheses that explain the findings. A small portion of the network is sketched in figure 1. Each arrow between nodes represents a causal link with a conditional probability. Each hypothesis is a subnetwork of the knowledge base instantiated with the findings of the case. This subnetwork is a complete explanation from primary causes through the pathophysiologic mechanisms to the findings. The corresponding portion of a hypothesis is shown on the left side of figure 2. The only findings not in the hypothesis causal network are those considered normal or better explained by something outside of the domain. For each hypothesis generated, the program can compute a probability by combining the probabilities in the network. The differential diagnosis consists of these hypotheses ordered by the computed probability. To limit the differential to likely hypotheses, the list is cut off when the probabilities fall below 1% of the best hypothesis. Thus, the differential may consist of one or many hypotheses. This type of diagnostic hypothesis is much more informative than those of earlier programs, such as AI/RHEUM[4], QMR[5], or DXplain[6], which provide hypotheses consisting of a single word or phrase. The structure of the HFP hypothesis explains the findings by showing the causal mechanisms producing them and thus provides a justification that the physician can evaluate to decide whether the conclusions are reasonable. The causal mechanisms are modeled at a level of detail consistent with that at which a physician might explain the findings to a colleague. This enhances the usefulness of the hypotheses for justification because the concepts fit naturally with the understanding of the physician. It is not a handicap for reasoning because this is presumably the level at which the human physician does similar reasoning and the physician is the best model we have for this kind of reasoning.

We previously conducted a formative evaluation of the program[7]. This evaluation used 242 cases collected from discharge summaries and compared the program diagnoses to diagnoses collected from cardiologists using the same information. We examined the maximum potential accuracy of the program by iteratively revising the knowledge base and rerunning the cases. Ultimately, the program was able correctly diagnose 90% of the cases. The main reasons that the remaining 10% could not be diagnosed correctly were the lack of reasoning about temporal relationships, the distinctions of different severities of diseases, and inappropriate combining of probabilities of multiple diseases. As a result of this evaluation we have developed methods for reasoning with the additional constraints provided by temporal and severity relationships to deal with some of the limitations encountered. We also developed an appropriate framework for computing the probability of multiple diseases[8].

The additions to the program include both enhancements of the knowledge base and changes in the reasoning methods. In the knowledge base the diseases and pathophysiologic states (all referred to as nodes) are subdivided into levels of severity and subtypes with additional constraints on the causal links. The severity and subtypes distinguish qualitatively different hemodynamic consequences. The link constraints reflect what severities and time bounds are required for the cause to produce the effect. For example, aortic regurgitation has three levels of severity reflecting situations in which there is a murmur without any hemodynamic consequences, situations in which the heart has compensated by dilating, and those in which the aortic regurgitation has caused deterioration of systolic function. These differences are reflected in the different probabilities and constraints on the links conditioned on the severity. Aortic regurgitation is also divided into two subtypes: primary valvular regurgitation and secondary regurgitation due to dilation of the aortic root. This distinction allows the effects to be different or have different probabilities for different subtypes. General temporal constraints on the nodes reflect limitations on causality, such as how long it takes for the nodes to become true (eg, minutes for an acute myocardial infarction (MI) to years for aortic stenosis), how long the node will remain true after the cause ends, and how long a patient might have the state (eg, an "acute" MI only exists for two days by definition before it becomes a "recent" infarct). The causal links include statements relating the severities and temporal constraints. For example:

```
(defnode (high pulmonary-vascular-resistance)
        severities ((1+ reversible)(2+ irreversible))
        causes (...
                (pulmonary-embolism :prob (3+ 1.0 => 1+ .3 => 2+))
                ((low arterial-po2) :prob ((< 1wk) .3 => 1+))
                ((high la-press) :prob (3+ (> 1yr) .2 => 1+ .2 => 2+))))
```

Thus, high pulmonary vascular resistance (PVR) has two severities, distinguished by being reversible and irreversible, rather than a particular measured resistance. A number of possible criteria could be used for defining severity, such as right ventricular dilatation, but reversibility was chosen because in the case of PVR it is correlated with severity and has important implications for management. The possible causes include pulmonary embolism, which if severe (3+) causes high PVR immediately and is irreversible 30% of the time. Low arterial oxygen levels for a limited time (a week or less) can cause reversible high PVR 30% of the time. Finally, chronic (a year or more) severe (3+) high left atrial pressure (as mitral stenosis might cause) can cause either severity of high PVR. The severities of pulmonary embolism and high left atrial pressure are also defined in terms of qualitative differences in their effects.

These constraints are enforced by the reasoning mechanisms at two levels: 1) The causal pathways that are computed when the knowledge base is loaded are pruned using the constraints, and

2) the patient specific nodes that are generated from the findings to build the hypotheses carry the constraints which guide the process of building the hypotheses. In the completed hypothesis, each node has one or more data structures called *severities* that represent a constraint on disease severity and subtype true over a specified temporal interval. For example, a *severity* for high PVR includes the following:

[severity node: (high PVR) min: 1+ max: 1+ type: none
justification: *path to supporting findings*
time: *time interval with ranges for beginning and ending*
causes: *list of causing severities*
effects: *list of severities this causes*]

Each of these structures is linked to the corresponding *severity* structures of the causes and effects. More than one *severity* structure is necessary when the evidence for a node supports different constraints on severity over different time intervals, such as a chronic disease with acute worsening.

This work of incorporating temporal relations and severity constraints had progressed to the point where it was appropriate to evaluate the performance of the program.

2 Methodology

A number of evaluations of medical expert systems have been reported in the literature, mostly focusing on the accuracy of diagnostic information[4, 5, 6, 9, 10]. Evaluating the accuracy the program is the first step toward establishing the clinical usefulness of the program. The essential difference between these evaluations and the evaluation needed to assess the accuracy of the HFP stems from the difference between the diagnoses — a fixed list of atomic diagnoses in the case of the other diagnostic programs, versus a network of causal relations for each diagnostic hypothesis for the HFP. To some extent, this more extensive set of possible conclusions was also faced in the evaluation of therapeutic decision-support systems[11, 12]. However, here too there was a single decision for the evaluators to judge. The hypothesis structure presents a much more difficult challenge for evaluation. With atomic diagnoses it is relatively clear what is correct or incorrect and it is easy to compare expert diagnoses to the program's diagnosis. For the HFP we needed to break the hypotheses down into manageable units and develop a method for evaluating the hypotheses.

While we were developing the temporal and severity reasoning and knowledge base, we used a learning set of 92 cases for testing and refining the algorithms. These were used to insure that the additions to the knowledge base were consistent and captured the distinctions pertinent to the domain.

To evaluate the program, 60 cases not previously seen by the program were gathered independently. The cases were selected to have as much variety as possible to present the program with a broad challenge. The cases included patients under care at the time of collection and as well as records of previously treated patients not used for program development. To evaluate the program at the time point likely to be the most challenging for diagnosis, each case was entered using data available at a time after the initial physical exam was completed but before new echocardiographic or catheterization data were available. Any echocardiographic or catheterization data obtained earlier were entered, because the admitting physician would have had them available. Thus, the

Hypothesis fragment	Summary fragment
FIXED CORONARY OBSTRUCTION ↓ REGIONAL FLOW DEFICIT ↓ MYOCARDIAL ISCHEMIA ACUTE → NITRO-GLYCERIN ↓ MYOCARDIAL INFARCTION ↓ increased cpk-mb	CORONARY ARTERY DISEASE → NITRO-GLYCERIN ↓ MYOCARDIAL INFARCTION ↓ increased cpk-mb

Figure 2: Summarization of part of a hypothesis

evaluation focused on the more difficult issues of reasoning from physical findings. Once the cases were collected and entered, the files were set aside until we were ready to conduct the evaluation.

We developed a format for evaluating the cases by using the learning set. Because each complete hypothesis consists of 20-40 nodes plus the findings and each case has one to several hypotheses in the differential, it was necessary to distill the information into a format that would be more manageable for the reviewers. To accomplish this we developed an algorithm for automatically summarizing the hypotheses. Summarization is possible because the nodes in the hypothesis play different roles. While some characterize the state of the patient, others give details of the causal mechanisms. For justification and explanation the mechanism nodes are useful, but when conveying the essence of the diagnosis is important, they can obscure the more important nodes and increase the size and apparent complexity of the hypothesis.

The summarization algorithm uses node attributes (recently added to the knowledge base) that identify their roles relative to other nodes. Summarization retracts the mechanisms into the nodes that classify the process. That is, the classifying node covers any nodes that represent mechanisms of it. Part of the summarization of the case in the appendix is illustrated in figure 2. The node labeled *regional flow deficit* has the attribute *mechanism* so it is clustered under its cause (*fixed coronary obstruction*). The myocardial infarction node has an attribute indicating that its cause, myocardial ischemia, is a mechanism. Next the definitions of *summary concepts* are checked to see if there is any further clustering to do in the hypothesis. Here, *coronary artery disease* is defined to cover *fixed coronary obstruction*, *old MI*, and *exertional angina*. Multiple applications of the summary concepts are possible. For example, if *low cardiac output* and *pulmonary congestion* are present, they are clustered together as *left heart failure* and if both *left heart failure* and *right heart failure* are present, they are clustered together as *congestive failure*, as in the summary in the appendix. Once the nodes are clustered, they are given the name of the summary concept or the primary node in the cluster. The causal links between the clusters are formed from the links that existed between nodes in the different clusters. Some heuristics are invoked to link the clusters appropriately. For example, since *nitroglycerin* is an effect prior to the primary node in the MI cluster, it is linked to the *coronary artery disease* cluster. This summarization mechanism greatly reduces the size of the hypotheses.

The alternative hypotheses are just like the primary hypothesis, but with lower computed probability. Thus, they are summarized in the same way. However, because they usually contain many of the same nodes as the primary hypothesis, only differing in a few places, it is more compact to describe them by telling how they differ from the primary hypothesis. This is described in terms of the differences between the summaries rather than the differences between the underlying nodes to maintain a consistent vocabulary for the evaluators. For the evaluation, we took one further step and merged the difference lists of the alternative hypotheses into a single list. This step may lose some information about what combination of additions and removals go together, but still gives the evaluator the range of alternatives suggested by the program.

The forms given to the reviewers were automatically generated by the HFP using the summarization mechanism (see appendix). These forms were organized as follows: The first page of the form only has the input description given to the program, allowing the reviewer to consider the findings before seeing the program's conclusions. The second page had the graphical display of the summary of the best hypothesis of the program followed by a textual description of the summary, outlining the causal relationships, so the reviewer could use whichever form was easier. Whenever there were alternative hypotheses generated by the program, the features of these were summarized as a list of what they added, changed, or removed from the best hypothesis.

Starting with the next page, the reviewers were asked to judge the results. The summaries of the best hypothesis and of the alternative hypotheses were melded together and then divided into a statement about each node in the summary. These statements included what was coming in to the node (causes or worsening factors) and what was going out (findings and other nodes caused, possibly caused, or worsened). Because these statements cover multiple hypotheses, they include phrases such as "possibly caused by". The alternative hypotheses were combined with the primary hypothesis in this way to give the reviewer the range of possible causes and effects for each node. This gave a more detailed examination of the alternatives than reviewing the difference list. After each statement the reviewer had 5 choices with the following meaning:

Correct This is a reasonable accounting of the findings and nodes.

Possible This is a possible interpretation of the findings and nodes, but not the best.

Partly correct This is mostly reasonable, but has some small subset of the findings or nodes wrongly attributed.

Wrong This node in the hypothesis is wrong.

Seriously wrong This node in the hypotheses is wrong and could lead to inappropriate care for the patient.

For those statements that the reviewers determined to be less than correct, they were asked to either circle the offending part or make a short comment to help us analyze the problem.

Other decision classifications have been used in evaluations, such as the *ideal, acceptable, sub-optimal, unacceptable* scale used by Hickam, et al[12]. Our scale is similar, but does not try to differentiate between acceptable and sub-optimal. Instead, we attempt to capture the difference between a statement that is sub-optimal because it has some part (a cause or effect) that the reviewer disagrees with (*partly correct*) versus the statement itself not being the best explanation (*possible*). The difference between *wrong* and *seriously wrong* was intended to highlight errors that could have consequences if acted on, because the importance of the statements varies widely.

The first statements to be reviewed were those for which the summary node was in all hypotheses. These were followed by statements that were only in some of the hypotheses and thus represented aspects of the alternatives. Following these statements was space to indicate any diagnostic or pathophysiologic determinations that the reviewers felt were left out. Finally, the textual statement of the best hypothesis and the summary of alternatives were presented with the same choices. These provided the overall assessment of the diagnostic performance of the program on the case.

Because we recognized some problems with the summaries, we also included as a last page the detailed graphical display of the hypothesis to help resolve questions (not included in the appendix), but these were only consulted in rare instances.

The reviewers were five senior cardiologists with no connection or previous exposure to the HFP. All are in active practice at teaching hospitals and represent the four medical schools in Massachusetts. They were sent a set of instructions and a sample evaluation form from the learning set a few days prior to familiarize themselves with the task. The evaluation was conducted as a single three hour session with all reviewers together, but working individually. That format provided us a way of answering any questions they might have while critiquing the cases.

The cases were arranged in random order and given one at a time to the cardiologists. Each case was reviewed independently by two different cardiologists. Each cardiologist reviewed as many cases as possible in the three hours. The number completed by the cardiologists varied widely, ranging from 7 to 17. The second reviews of the first 10 cases were held until the second hour to balance out any differences due to initial adjustment to the format. A total of 26 cases were completed with two reviews each. The number of statements in each case reviewed by the cardiologists, including the statements about the leading hypothesis and about alternative diagnoses varied from 6 to 15 and averaged about 11.

The 26 cases reviewed included a wide variety of diseases and most had multiple diseases. There were 5 acute MIs, 2 recent MIs, 2 unstable angina, and 9 others with stable coronary artery disease. Valvular disorders included 9 with mitral regurgitation, 4 each with aortic stenosis, mitral stenosis, and tricuspid regurgitation, 3 with aortic regurgitation, and one with aortic and mitral valve replacements. Five patients had cardiomyopathy. Two had significant hypertension and two had pulmonary hypertension. There were also cases of endocarditis, pericarditis, and possible aortic dissection. Renal insufficiency and pulmonary disease were also well represented. The first hypotheses for the cases had an average of 3.2 disease nodes. These were all complex cases, chosen to be difficult, but otherwise representative of the case load of a tertiary care hospital.

The reviewers were not asked to record or critique their own diagnoses. Having the reviewer's diagnosis evaluated as well, as has been done in earlier evaluations[11], would make it possible to rank the program relative to the evaluators. However, the additional work required of the reviewers would have greatly increased the time required per case, but more importantly, the nature of the human differential diagnoses is sufficiently different from the computer generated diagnoses to make it very difficult to compare them. Physicians vary the specificity of the diagnosis depending on factors such as certainty and they tend to focus on the acute problem, sometimes not mentioning chronic problems. Additionally, because the source of a diagnosis is easily recognized, the bias of

Judgment 1	Judgment 2	Hypothesis 1	Alternatives
Correct	Correct	7	3
Correct	Possible/Partly correct	4	13
Possible/Partly correct	Possible/Partly correct	12	5
Seriously wrong	Possible/Partly correct	2	0
Wrong	Wrong	1	0

Table 1: Overall judgments of the first hypothesis and alternatives

knowing whether the reviewer is critiquing a physician or the computer would further obscure the comparison. As a result, how the reviewers would judge each other will remain unknown, although we can infer from the explicit disagreements in evaluating the cases that there would be significant differences among the reviewers.

Because the evaluation was held as a group session, we were able to answer many of the reviewer's questions and avoid a number of misunderstandings. They were also encouraged to write comments and circle inappropriate parts of statements. As a result, it is possible to analyze the evaluation to determine the reasons for dissatisfaction with statements they found less than correct. The environment also generated some conversation that disclosed some concepts as controversial that we previously thought were widely accepted.

3 Results

The statements that were judged can be divided into those about the overall diagnosis and those about the details. Overall rating was done on the first hypothesis in each case and on the summary of the alternatives in the 21 cases that had additional hypotheses. The ratings are given in table 1. There are three cases in which at least one reviewer considered the first hypothesis wrong. In two of these, the program missed diagnoses of mitral stenosis, once because the location of the murmur only matched tricuspid stenosis and once because the program considered the murmur to be a flow murmur of the mitral regurgitation that was present. In the third case, the program left pleuritic chest pain unaccounted. The reviewer criticized it for not suggesting pulmonary embolism, although it was actually pericarditis. The two *seriously wrong* judgments were made by different reviewers. The judgments of the alternatives included fewer in complete agreement and none considered wrong. With only about 10% of first hypotheses judged by any reviewer to be wrong there seems to be general agreement, but with only 25% of first hypotheses judged by both to be completely correct and fewer of the alternatives, there is considerable disagreement in the details.

In this analysis and in the analysis of the detailed statements, it became clear that some of the distinctions in the rating scale had no practical significance. The reviewers seemed to use *possible* and *partly correct* interchangeably, so in the following analysis they are combined and called *possible*. There were no statements that both reviewers rated as *seriously wrong*, in fact there were no statements that one reviewer rated *seriously wrong* that the other reviewer even rated *wrong*, so *wrong* and *seriously wrong* are combined as *wrong*.

To determine whether there was any systematic bias among the reviewers, we compared the

Reviewer	1	2	3		4		5		Definite		Possible	
% Correct	48	42	66	(.001)	46		30	(.01)	54	(.001)	37	(.001)
% Possible	40	42	22	(.001)	47		54	(.05)	38	(.05)	48	(.05)
% Wrong	12	15	11		7	(.05)	15		8	(.01)	16	(.01)
% Total	16	23	15		34		12		45		29	

Table 2: Distribution of ratings by reviewer and type of statement. Statements true of all hypotheses are definite, those true of some hypotheses are possible.

fraction of the statements each rated as correct, possible, or wrong to the fraction with that rating as judged by the rest of the reviewers. We also compared the ratings of statements about disease nodes in all hypotheses versus those in some hypotheses. Those fractions with their significance and the fraction of the total statements they represent are given in table 2. This analysis indicates that the reviewers were well balanced. The two reviewers that differed significantly from the rest contributed the smallest number of ratings and their differences balanced each other. Comparing definite nodes to possible nodes, the possible nodes were judged more harshly.

Analysis of the detailed statements requires caution because each statement may overlap with one to several other statements in the same case. In relating a node to its definite and possible causes and effects, each statement touches on several nodes. If there is a node that the reviewer feels should not be in the hypothesis, that affects his judgment of statements about each of the causes and effects as well as the statement about the node itself. For that reason we have analyzed the critiques to determine the source of negative judgments and clustered the affected statements together. This process required some interpretation in many cases, but fortunately the reviewers comments were extensive enough that it was always possible to have a good idea what they were concerned about.

In the 285 detailed questions there were 137 issues raised by one or both of the reviewers. One might infer that both reviewers were in complete agreement with 148 of the statements, but because of the influence of the issues on the statements about causes and effects, there were only 92 statements that were considered correct by both reviewers. Each issue was assigned for each reviewer the most serious rating of any of the statements in which that issue arose. Some issues were comments from the reviewers about relations and hypotheses that they considered missing. Because these did not have ratings, they were all rated as *possible*. Thus, if *possible* $= P$ and *wrong* $= W$, the issues that concerned both reviewers could be rated WW, WP, or PP, and those that were of concern to only one would be rated W or P. The 137 issues that arose in the cases were rated as follows: WW 11, WP 10, PP 16, W 27, P 73. The majority (53%) were possible changes that were of concern to only one reviewer. Still, there were 11 issues that the reviewers agreed were wrong and another 10 that one considered wrong and the other reviewer thought were not the best choice.

To determine the source of these issues, we analyzed and classified them. Classifying the issues is open to some interpretation, but it is very useful in determining whether the issues imply that refinement of the program is needed, the method of summarization and presentation misrepresented the conclusions of the program, there was some misunderstanding or mistake on the part of the reviewer, or that there is an underlying difference of opinion among cardiologists. To avoid missing problems with the program, we have classified issues as relating to the program if there is any doubt.

The classifications we used are as follows:

Controversy These are issues in which there is clearly a difference of opinion among cardiologists. Several of these came to light during the evaluation in discussions in which the reviewers disagreed with each other. Others are disagreements with carefully considered representations in the program, that is, disagreements between a reviewer and the developers. These were reviewed and classified by the cardiologist developers.

Reviewer wrong or inconsistent These are issues where the criticism is in conflict with the actual patient state as indicated in the patient record or, in two cases, the reviewer's ratings of different statements in the same case are mutually inconsistent. Given the limited information in the input, several diseases may be appropriate to consider in addition to what the patient actually had. Therefore, the judgment was only classified as wrong when the program's statement corresponded to the actual situation and reviewer rated that as incorrect.

Misunderstanding These are issues in which the reviewer probably overlooked part of the information or misunderstood some of the (occasionally convoluted) automatically generated text. That is, the intended meaning of the statement is consistent with the reviewer's objection.

Summarization These are issues in which the program that summarized the hypotheses and put the information from multiple hypotheses together into single statements about nodes obscured the relationships that exist among the nodes or inappropriately labeled node clusters.

Program Relationships in the program that need to be reexamined. (These will be further classified.)

As an example of how the cases were analyzed, consider the evaluation form in the appendix. The first reviewer marked all statements *correct* except two: one about atrial fibrillation and one about atrial septal defect (ASD), which were marked *possible*. The second reviewer marked the atrial fibrillation and mitral regurgitation statements *partly correct* with a note that atrial fibrillation is caused by dilated cardiomyopathy not mitral regurgitation. He marked the statement about dopamine *wrong* and those about acute MI, coronary artery disease, and ASD as *possible*. Thus, these reviews resulted in one W: dopamine contributing to heart failure; two PPs: atrial fibrillation caused by dilated cardiomyopathy instead of mitral regurgitation and ASD causing fixed splitting S2; and one P: acute MI as evidenced by elevated CPK-MB (with the coronary artery disease statement considered to be the same issue). These were classified as follows: ASD, because it is a weak alternative hypothesis, was attributed to the program; the dopamine and the cause of atrial fibrillation were considered controversies, because the cardiologist developers consider them correct; and the acute MI as a misunderstanding. The acute MI is an example of the kind of misunderstanding that can arise from an automatically generated evaluation form. The acute MI was only included in the first of the two hypotheses and therefore listed with the nodes only present in some hypotheses. The reviewer considered acute MI a likely possibility, and therefore was in agreement with the program, but marked the statement *possible* rather than *correct*, confusing the correctness of the statement with the likelihood of acute MI.

Given this scheme, the 137 issues identified by the reviewers were classified as shown in table 3. One controversy that accounted for six of the issues was whether diastolic dysfunction causes

	WW	WP	PP	W	P	Total
Controversy	1	4	3	6	27	41
Reviewer wrong	0	0	0	2	4	6
Misunderstanding	0	0	1	5	10	16
Summarization	2	0	3	6	15	26
Program	8	6	9	8	17	48
Total	11	10	16	27	73	137

Key: WW both wrong, WP wrong and possible,
PP both possible, W one wrong, P one possible

Table 3: Classification of the issued raised by the reviewers

low cardiac output and left heart failure. That is, whether a patient with LV hypertrophy, a normal ejection fraction, fatigue and pulmonary congestion should be described as having diastolic dysfunction causing the findings. It is clear that patients, especially older patients, present with those findings but it depends on how one defines "diastolic dysfunction" whether that is the cause or not. This difference generated discussion among the reviewers. Several other controversies were probably issues of definition as well — whether mitral stenosis causes anginal chest pain, what constitutes a left atrial abnormality on electrocardiogram, or how broadly one may define COPD (chronic obstructive pulmonary disease). Some of the disagreements were whether a particular disease was adequately supported. For example, whether murmurs as the only direct evidence were sufficient to suggest aortic stenosis or tricuspid regurgitation. Others were findings that the program left unaccounted that the reviewers felt should be accounted for, such as cough and non-specific ST and T wave changes.

Assuming that different reviewers take different sides in controversies, make different mistakes, and have different misunderstandings, the first three categories in the table are issues that would have a high rate of disagreement between reviewers. Because these are only a sample of what might contribute to inter-reviewer disagreement and they are 45% of the issues, the differences among the reviewers are likely to be comparable to the differences the reviewers had with the program.

Most of the summarization problems resulted from using the phrases *caused by* and *accounted for* in the statements to represent all of the linkages between summary nodes. Once clusters of nodes are abstracted to summary nodes, many of these become influences not normally considered causality (eg, the summary saying that aortic stenosis is causing hypertension because the hypertension cluster includes the high LV pressure node which is also caused by aortic stenosis) or a cluster having an inappropriate label given the severity or what it was influencing (eg, a cluster labeled high blood volume causing elevated liver function tests, skipping the intermediate splanchnic congestion). Another summarization problem was grouping acute and chronic manifestations of nodes together. For example, the program hypothesized that one patient had chronic mitral regurgitation worsened by an acute MI. In summarizing the mitral regurgitation it listed the acute MI as a cause and included both acute and chronic findings as effects. Another problem was not having different names for different severities of the clusters. For example, situations ranging from mild tachypnea to frank pulmonary edema were all labeled *left heart failure*, whereas treated failure without symptoms or with minor symptoms should be called *compensated left failure*.

	WW	WP	PP	W	P	Total
Needs more evidence	2	1	4	3	2	12
Other causes for finding	2	1	1	0	4	8
Missing important disease	1	0	2	0	0	3
Missing concept	0	0	1	0	5	6
Severity definition	0	2	0	4	2	8
Suspect probabilities	3	2	1	1	4	11
Total (Program)	8	6	9	8	17	48

Key: WW both wrong, WP wrong and possible,
PP both possible, W one wrong, P one possible

Table 4: Classification of the criticisms caused by the program

3.1 Criticisms Related to the Program

The remaining 35% of the issues were related to the functioning of the program. These can be further classified as shown in table 4. The largest source of these stem from a difference between the nature of hypotheses from a Bayesian network and hypotheses that a human expert would generate. The single criterion for a good hypothesis from a Bayesian network is high probability. Our heuristic approach also tends to find single causal pathways unless there is separate evidence for more than one. The criteria implicitly used by experts in generating hypotheses require adequate evidence for diseases, inclusion of all important diagnoses, and inclusion of all plausible explanations for findings. They handle side-effects of therapies in a special way, liberally considering the possibility but being conservative about definite attribution. These differences come in conflict in a number of ways. First, the program does not require any particular evidence to invoke a disease. In several instances the program included diseases with inadequate evidence (from the reviewer's perspective) because that yielded a higher probability hypothesis, eg, concluding low renal perfusion on the basis of the renal function tests without evidence of pedal edema, suggesting anemia from findings without having a hematocrit, or using functional tricuspid regurgitation to account for a murmur when there is cardiomegaly. Second, the program will eliminate important diseases if the probability is too low. The program left out pulmonary embolism because the pleuritic chest pain has a high false positive rate and the shortness of breath could be otherwise explained, not taking into account the importance of diagnosing pulmonary embolism, which is potentially serious and very treatable. Third, the program also left out possible causal pathways for findings such as cough, LVH on electrocardiogram, and pulmonary hypertension because one reasonable explanation had been found. In one case, poor ventricular function was attributed to existing hypertensive heart disease and not to an acute MI because all of the findings were consistent with chronic disease. The reviewers disagreed because the acute MI will certainly worsen the ventricular function even if it did not cause it and even though there is no actual evidence of acute worsening. Finally, in a few instances the program attributed findings to drug side-effects when there were other possibilities, eg, steroids to account for pedal edema when there were possible cardiac causes.

These differences do not imply that the Bayesian network is incompatible with expert-like hypotheses, but it does imply that some additional processing is needed. First, a utility model is needed to avoid pruning out important hypotheses with probabilities below the threshold. This could be implemented by extending the reasoning mechanism to include value nodes as in an influence diagram or by doing some post-processing of the hypotheses to allow the appropriate reasoning about utilities. Second, additional pathways to findings need to be added to reflect alternative causes within the hypothesis. Third, diseases that are not well justified usually represent the most likely of several possible causes for a set of findings. These should be presented to the user in a more general form, using names such as "poor ventricular function" that leave the actual cause unspecified. Finally, therapy side-effects should be handled in a way that considers the additional means the physician has of determining the causes by changing or adjusting the therapy.

A few of the alternative diagnoses suggested by the reviewers involved concepts that we have not included in the knowledge base, although we have considered adding them. These were thyroid therapy causing hyperthyroidism, high pCO_2 from inappropriate ventilator setting, and left ventricular aneurysm. These are concepts that can be added to the knowledge base. There were other suggested alternatives that still seem inappropriate, such as using diabetes as an explicit cause (beyond changing some of the probabilities, which it currently does) and cerebral embolism.

Another source of criticisms were inappropriate severity constraints. These included the level of hematocrit needed to account for high cardiac output, the severity of renal insufficiency that causes nausea, and the degree of pulmonary hypertension that causes RV hypertrophy. These can all be corrected in the knowledge base.

The rest of the issues are attributable to the probabilities in the knowledge base. There were diseases or pathophysiological states in hypotheses deemed too unlikely to be considered (but not ruled out) and diseases and states that the reviewers felt were should have been included. One problem that occurred twice was determining the source of murmurs from their location. The worst example was concluding the rare disease, isolated tricuspid stenosis, because the murmur description did not fit anything else. We have done considerable work on murmur description, but the variability of descriptions encountered in cases continues to be a challenge. One solution is to give the locations a high error rate. In two situations unlikely causal mechanisms were proposed by the program. One was essentially overcompensation: high vagal stimulation (evidenced by diaphoresis) causing low vascular resistance (with low blood pressure) causing high cardiac output (with a flow murmur). Another was letting the wrong causal pathway dominate in determining the effects of septic shock. The program concluded that the decreased filling pressure would cause low cardiac output rather than the decreased systemic resistance causing increased (or normal) cardiac output. Both of these problems can be corrected by adding causal constraints. Still, they provide interesting examples of how the underlying assumption of conditional independence in Bayesian networks can misrepresent causal models if secondary dependencies are overlooked.

One very general problem is the use of pertinent negatives to justify conclusions. The program uses negatives to find appropriate causal pathways in building hypotheses and to determine their probabilities, but only the abnormal findings are included on the displays and used to justify the nodes. It was clear both from written and oral comments that in many cases the negatives were crucial to making the statements convincing. In one case, the program had coronary spasm as the primary cause for an MI because a catheterization done six months prior did not mention the coronaries, which the program interpreted to mean that they were clear. Thus, coronary spasm was the most likely cause. Without that assumption, coronary artery disease would be a more likely cause. In another case, one reviewer judged the lack of aortic stenosis in the hypothesis to be a serious error and the other reviewer judged the same statement to be correct and commented that it should be justified by the lack of LVH on the electrocardiogram.

Thus, we have determined that the first hypotheses were rated correct by both reviewers in 25% of the cases and wrong by at least one in 10%. Analysis of the detailed reasoning identified 137 issues, about 5.3 per case. Of these 53% were possible concerns raised by one reviewer. Of the 5.3 issues per case, 2.5 were attributable to controversies, misunderstandings, or mistakes; 1 was due to the overly simplistic representation of the summaries; and 1.8 were issues related to the program. All of the program issues are ones that can be handled with refinement of the knowledge base and some additional processing to incorporate concerns of utility and evidence.

4 Discussion

This evaluation provides us with a reasonable picture of the diagnostic performance of the Heart Failure Program. The number of cases in the evaluation was too small to measure the performance relative to physicians. In any case, that would be missing the point. The performance of the program is at a level such that the experts can relate to the explanations and their criticisms are mostly in the details of the diagnoses. There are still occasional serious errors made by the program, but they imply that further refinement of the knowledge base is needed rather than fundamental changes to the reasoning mechanisms. The severity and temporal constraints give the program the tools necessary to keep from generating impossible causal pathways in the hypotheses and allowing better discrimination of the likelihood of the pathways it does generate. Some additional reasoning is needed to create a better fit with the human expectations of good hypotheses.

The evaluation also illuminated a number of difficulties faced in evaluating a program that provides highly detailed assessments. The first issue the evaluators commented on was the artificiality of the case descriptions with most of the echocardiographic findings left out. We chose to rely mostly on physical examination, electrocardiographic, and X-ray findings to get a more extensive test of the program's reasoning, but cardiology in the United States has come to a point where in complicated cases cardiologists are reluctant to form conclusions without seeing the echocardiographic findings.

The summarization of the differential diagnoses was not part of the program's diagnostic reasoning and was developed for the evaluation to reduce the work load for the reviewers. Indeed, it decreased the number of nodes in the hypotheses to a manageable number, making an evaluation of the detailed reasoning feasible. Because it proved to be a useful tool to convey the important conclusions in the diagnosis, we intend to incorporate it into the program. At present, the summarizer can also obscure the nature of relationships present in the hypothesis. The main problem is that not all of the summary links have the same meaning. The labels need to reflect important distinctions such as *causing* versus *possibly contributing to* or *chronic with acute worsening*. Using the word *causes* for all links often misrepresents the kind of relationship that exists. There are also a number of conventions used by people in conveying degrees of uncertainty that need to be incorporated in the summaries, such as saying *pulmonary disease* rather than *COPD* when the findings are non-specific. Thus, summarization of the differential diagnoses is a difficult problem that will take more work to bring it to the competence necessary to be an effective utility.

4.1 Evaluation Methodology

Evaluation of expert systems is always difficult, but it is more difficult for the HFP because of the complex conclusions that were being evaluated. Conducting the evaluation as a single session with all reviewers was extremely beneficial. It enabled us to answer their questions as they arose, monitor their progress, make sure they understood the instructions as intended, and to encourage them to provide comments with their judgments. It also allowed for some discussions, which slowed the progress. Since three hours is about the maximum length of time people can effectively do such detailed work, a larger evaluation may require multiple sessions or the rest of the reviews to be done in a different context.

The use of evaluation forms proved to be much more practical than using the program interactively because of the logistics involved, the additional delays that the program would have introduced, and the ability of paper to take comments anywhere. The evaluation forms themselves left us with the difficult task of analyzing the judgments to determine the source of criticisms. Fortunately, the reviewers wrote enough comments to make that analysis possible. The problem is essentially one of designing a multiple question test when an essay is needed. The reviewers are being asked to critique the details, but there are too many details to consider them one by one. The node interconnections imply that a single issue changes the rating of multiple nodes. A possible improvement would be to organize the questions around paths through the summaries rather than single nodes, or organizing the questions around "diagnostic" nodes (the main disease and syndrome nodes).

The automatic generation of the evaluation forms utilized a limited vocabulary and "computer logic" that made some of the statements difficult to understand. Particularly difficult were statements about nodes appearing only in some of the hypotheses. The significantly poorer rating of such possible nodes and the smaller number of correct ratings of alternative hypotheses may be due to these problems of presentation or it may be that determining the range of a differential is a more difficult problem than determining the best hypothesis.

The final issue that makes an evaluation of diagnostic programs difficult is the lack of a gold standard. It is tempting to say that the final diagnosis of the patient is the gold standard, but the objective of the differential is to determine all of the hypotheses that are consistent with the patient presentation rather than only the diseases that the patient actually has. Because there is no objective way of obtaining such a list, the best that can be achieved is expert consensus. As a result, there will always be some level of controversy and misunderstanding in the critiques. A common strategy to control for these factors is to have the experts critique each other as well as the program. That is not feasible when the diagnoses are detailed because the expert diagnoses are given with much more selective details. An alternative would be to have the reviewers individually critique a case and then collectively agree on a final critique. The problem, of course, is that it would lengthen an already time consuming and somewhat tedious process.

Even with the difficulties of designing and analyzing this type of evaluation, the evaluation has been very helpful in determining the level of performance of the program. Overall, the program is capable of providing high quality detailed diagnostic hypotheses for complex cardiovascular cases. With some additional refinement of the knowledge base and processing of the hypotheses, the mistakes encountered by the reviewers should be eliminated and the error rate decreased significantly. Once these changes and the summarizer have been appropriately validated, the appropriate next step is a prospective evaluation to address the usefulness of the program.

This research was supported by National Institutes of Health Grant No. R01 HL33041 from the National Heart, Lung, and Blood Institute and No. R01 LM04493 from the National Library of Medicine.

References

[1] Long WJ, Naimi S, Criscitiello MG, Larsen G. Differential diagnosis generation from a causal network with probabilities. In: Proceedings, Computers in Cardiology Conference. Washington, DC, 1988;185-188.

[2] Long W. Medical diagnosis using a probabilistic causal network. Applied Artificial Intelligence. 1989;3:367-83.

[3] Long WJ. Flexible reasoning about patient management using multiple models. Artificial Intelligence in Medicine. 1991;3:3-20.

[4] Kingsland LC. The evaluation of medical expert systems: experience with the AI/RHEUM knowledge-based consultant system in rheumatology. In: Ninth Symposium on Computer Applications in Medical Care, Washington, DC, 1985;292-295.

[5] Miller RA, Pople HE, Myers JD. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. New England Journal of Medicine. 1982;307:468-476.

[6] Feldman MJ, Barnett GO. An approach to evaluating the accuracy of DXplain. In: Fourteenth Symposium on Computer Applications in Medical Care, Washington, DC, 1990;38-43.

[7] Long WJ, Naimi S, Criscitiello MG. Development of a knowledge base for diagnostic reasoning in cardiology. Computers in Biomedical Research. 1992;25:292-311.

[8] Long WJ. The probability of disease. In: Fifteenth Symposium on Computer Applications in Medical Care, Washington, DC, 1991;619-23.

[9] Kahn CE. Validation, clinical trial, and evaluation of a radiology expert system. Methods of Information in Medicine. 1991;30:268-74.

[10] van der Lei J, Musen MA, van der Does E, Manintveld AJ. Critiquing physician decision making using data from automated medical records: assessing the limitations. In: Fourteenth Symposium on Computer Applications in Medical Care, Washington, DC, 1990;559-63.

[11] Yu VL, Fagan LM, Wraith SM, et al. Antimicrobial selection by a computer: a blinded evaluation by infectious diseases experts. Journal of the American Medical Association. 1979;242:1279-82.

[12] Hickam DH, Shortliffe EH, Bischoff MB, Scott AC, Jacobs CD. The treatment advice of a computer-based cancer chemotherapy protocol advisor. Annals of Internal Medicine. 1985;103:928-36.

A Example Evaluation Form

Note: Annotations that were not part of the actual form are in bold.
The following is the input description given to the program, to allow the reviewers to consider their own diagnosis:

Patient: PT2006

HISTORY: 65 year old male with normal weight, orthopnea progressive, dyspnea on-exertion progressive and cough having known-diagnoses of congestive-cardiomyopathy, atrial-fibrillation, chronic-renal-insufficiency and diabetes and on furosemide, nitroglycerin, captopril, digitalis, dobutamine for 1wk and dopamine for 1wk

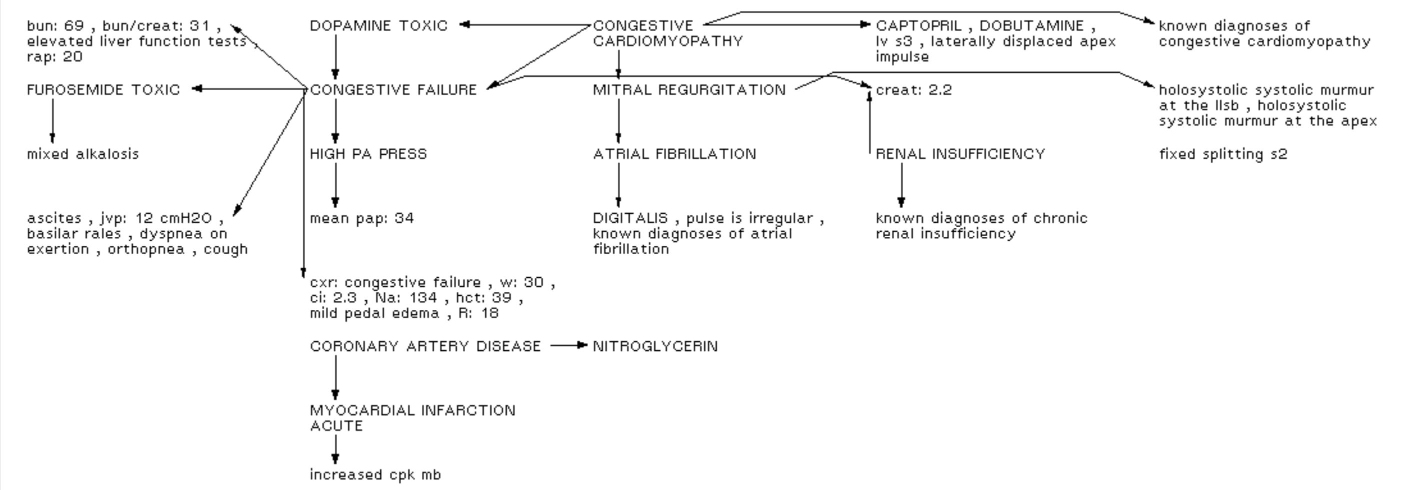
VITAL-SIGNS: bp: 114/76 hr: 77, R: 18 and T: 97.4

PHYSICAL-EXAM: appears in no-acute-distress, conscious, chest revealed basilar rales, jugular-pulse shows jvp: 12 cmH2O, pulse is irregular, auscultation revealed fixed-splitting s2, lv-s3, II/VI holosystolic systolic-murmur at the apex and III/VI holosystolic systolic-murmur at the llsb, laterally-displaced apex-impulse, ascites and mild pedal-edema

LABORATORY-FINDINGS: ekg (1hr): atrial-fibrillation, cxr (1hr): congestive-failure, electrolytes (1hr): Na: 134 k: 4.4 bun: 69 creat: 2.2 Ca: 10.0 serum-albumin: 4.4, blood-gases (1hr): po2: 83 ph: 7.49 pco2: 36, cbc (1hr): hct: 39 wbc: 9.3, increased-cpk-mb (1day) and elevated liver-function-tests (1hr)

HEMODYNAMIC-DATA: w: 30, ci: 2.3, pap: 54/24, rap: 20 and data from swan-line (now)

New page, so the reviewer is less influenced by the program's conclusions.
This is a graphical summary of the leading assessment generated by the program, followed by a textual description and the differences in other hypotheses. The following pages will ask for a critique of the parts of this and the rest of the hypotheses.



Patient: PT2006
HISTORY: 65 year old male with normal weight, orthopnea progressive, dyspnea on-exertion progressive and cough having known-diagnoses of congestive-cardiomyopathy, atrial-fibrillation, chronic-renal-insufficiency and diabetes and on furosemide, nitroglycerin, captopril, digitalis, dobutamine for 1wk and dopamine for 1wk
VITAL-SIGNS: bp: 114/76 hr: 77, R: 18 and T: 97.4
PHYSICAL-EXAM: appears in no-acute-distress, conscious, chest revealed basilar rales, jugular-pulse shows jvp: 12 cmH2O, pulse is irregular, auscultation revealed fixed-splitting s2, lv-s3, I-II/VI holosystolic systolic-murmur at the apex and III/VI holosystolic systolic-murmur at the llsb, laterally-displaced apex-impulse, ascites and mild pedal-edema
LABORATORY-FINDINGS: ekg (1hr): atrial-fibrillation, cxr (1hr): congestive-failure, electrolytes (1hr): Na: 134 k: 4.4 bun: 69 creat: 2.2 Ca: 10.0 serum-albumin: 4.4, blood-gases (1hr): po2: 83 ph: 7.49 pco2: 36, cbc (1hr): hct: 39 wbc: 9.3, increased-cpk-mb (1day) and elevated liver-function-tests (1hr)
HEMODYNAMIC-DATA: w: 30, ci: 2.3, pap: 54/24, rap: 20 and data from swan-line (now)

The same hypothesis in outline form:
Best Hypothesis:

- CONGESTIVE CARDIOMYOPATHY accounts for lv s3, laterally displaced apex impulse, known diagnoses of congestive cardiomyopathy, treated with CAPTOPRIL, DOPAMINE TOXIC, DOBUTAMINE, causing:
  1. MITRAL REGURGITATION accounts for holosystolic systolic murmur at the llsb, holosystolic systolic murmur at the apex, causing:
     - ATRIAL FIBRILLATION accounts for pulse is irregular, known diagnoses of atrial fibrillation, treated with DIGITALIS,
  2. CONGESTIVE FAILURE accounts for bun: 69, creat: 2.2, bun/creat: 31, elevated liver function tests, rap: 20, cxr: congestive failure, w: 30, ci: 2.3, Na: 134, hct: 39, mild pedal edema, R: 18, ascites, jvp: 12 cmH2O, basilar rales, dyspnea on exertion, orthopnea, cough, treated with FUROSEMIDE TOXIC, causing:
     - HIGH PA PRESS accounts for mean pap: 34,

- CORONARY ARTERY DISEASE treated with NITROGLYCERIN, causing:
  - MYOCARDIAL INFARCTION ACUTE accounts for increased cpk mb,
- RENAL INSUFFICIENCY accounts for creat: 2.2, known diagnoses of chronic renal insufficiency,
- left unaccounted: fixed splitting s2,

The following are the ways in which the other hypothesis generated for this case differs from the first hypothesis. If there were were more hypotheses, there would still be just one statement about differences. The changes are not justified by what they account for, but that will be examined in the detailed statements to follow.

Additional Hypotheses may add ATRIAL SEPTAL DEFECT exclude MYOCARDIAL INFARCTION ACUTE, CORONARY ARTERY DISEASE

New page, and the beginning of the questions to be answered by the reviewer.

Following is an explanation of how each of the concepts in the hypotheses became part of the differential. They are all of the form: [concept] caused by [concepts] accounts for [concepts and findings]. The findings from the input are given in lower (or mixed) case and the concepts inferred by the diagnostic computation are in upper case. The alternatives from different hypotheses are integrated in the explanations, indicated by *or* and *possibly*. Please critique each mechanism in the boxes provided.

The following list describes nodes that were in both hypotheses:

- CONGESTIVE FAILURE caused by ATRIAL FIBRILLATION, CONGESTIVE CARDIOMYOPATHY, DOPAMINE TOXIC accounts for cough, orthopnea, dyspnea on exertion, basilar rales, jvp: 12 cmH2O, ascites, R: 18, mild pedal edema, hct: 39, Na: 134, ci: 2.3, w: 30, cxr: congestive failure, rap: 20, elevated liver function tests, bun/creat: 31, creat: 2.2, bun: 69, HIGH PA PRESS, FUROSEMIDE TOXIC

correct	possible	partly correct	wrong	seriously wrong	comment:

- CONGESTIVE CARDIOMYOPATHY accounts for known diagnoses of congestive cardiomyopathy, laterally displaced apex impulse, lv s3, CONGESTIVE FAILURE, DOBUTAMINE, MITRAL REGURGITATION, DOPAMINE TOXIC, CAPTOPRIL, and possibly, DIGITALIS

correct	possible	partly correct	wrong	seriously wrong	comment:

"Possibly digitalis" means one hypothesis made this causal connection and the other did not.

- ATRIAL FIBRILLATION caused by MITRAL REGURGITATION accounts for known diagnoses of atrial fibrillation, pulse is irregular, DIGITALIS

correct	possible	partly correct	wrong	seriously wrong	comment:

- MITRAL REGURGITATION caused by CONGESTIVE CARDIOMYOPATHY accounts for holosystolic systolic murmur at the apex, holosystolic systolic murmur at the llsb, ATRIAL FIBRILLATION

correct	possible	partly correct	wrong	seriously wrong	comment:

- RENAL INSUFFICIENCY accounts for known diagnoses of chronic renal insufficiency, creat: 2.2

correct	possible	partly correct	wrong	seriously wrong	comment:

- HIGH PA PRESS caused by CONGESTIVE FAILURE accounts for mean pap: 34

correct	possible	partly correct	wrong	seriously wrong	comment:

- NITROGLYCERIN treating possibly, CORONARY ARTERY DISEASE

correct	possible	partly correct	wrong	seriously wrong	comment:

- FUROSEMIDE TOXIC treating CONGESTIVE FAILURE accounts for mixed alkalosis

correct	possible	partly correct	wrong	seriously wrong	comment:

"Furosemide toxic" just means that the hypothesis is attributing side-effects to the drug.

- DOPAMINE TOXIC treating CONGESTIVE CARDIOMYOPATHY accounts for CONGESTIVE FAILURE

correct	possible	partly correct	wrong	seriously wrong	comment:

- DIGITALIS treating ATRIAL FIBRILLATION, and possibly, CONGESTIVE CARDIOMYOPATHY

correct	possible	partly correct	wrong	seriously wrong	comment:

Both of the following are possibly unaccounted since each is accounted for in one of the two hypotheses.

- Unaccounted Findings: possibly, fixed splitting s2, increased cpk mb

correct	possible	partly correct	wrong	seriously wrong	comment:

The following are included in some hypotheses:
The first two were only in the first hypothesis, the third node was in the second hypothesis.

- MYOCARDIAL INFARCTION ACUTE caused by CORONARY ARTERY DISEASE accounts for increased cpk mb

correct	possible	partly correct	wrong	seriously wrong	comment:

- CORONARY ARTERY DISEASE accounts for MYOCARDIAL INFARCTION ACUTE, NITROGLYCERIN

correct	possible	partly correct	wrong	seriously wrong	comment:

- ATRIAL SEPTAL DEFECT accounts for fixed splitting s2

correct	possible	partly correct	wrong	seriously wrong	comment:

Relationships left out: ______________________________

____

Following is the differential diagnosis of the program. Please critique the hypothesis and summary of alternates in the boxes provided.
These are the same as before, only now the reviewer is asked to critique the best hypothesis and alternates.

- Best Hypothesis

  - CONGESTIVE CARDIOMYOPATHY accounts for lv s3, laterally displaced apex impulse, known diagnoses of congestive cardiomyopathy, treated with CAPTOPRIL, DOPAMINE TOXIC, DOBUTAMINE, causing:
    1. MITRAL REGURGITATION accounts for holosystolic systolic murmur at the llsb, holosystolic systolic murmur at the apex, causing:
       - ATRIAL FIBRILLATION accounts for pulse is irregular, known diagnoses of atrial fibrillation, treated with DIGITALIS,
    2. CONGESTIVE FAILURE accounts for bun: 69, creat: 2.2, bun/creat: 31, elevated liver function tests, rap: 20, cxr: congestive failure, w: 30, ci: 2.3, Na: 134, hct: 39, mild pedal edema, R: 18, ascites, jvp: 12 cmH2O, basilar rales, dyspnea on exertion, orthopnea, cough, treated with FUROSEMIDE TOXIC, causing:
       - HIGH PA PRESS accounts for mean pap: 34,
  - CORONARY ARTERY DISEASE treated with NITROGLYCERIN, causing:
    - MYOCARDIAL INFARCTION ACUTE accounts for increased cpk mb,
  - RENAL INSUFFICIENCY accounts for creat: 2.2, known diagnoses of chronic renal insufficiency,
  - left unaccounted: fixed splitting s2,

correct	possible	partly correct	wrong	seriously wrong	comment:

- Additional Hypotheses may add ATRIAL SEPTAL DEFECT exclude MYOCARDIAL INFARCTION ACUTE, CORONARY ARTERY DISEASE

correct	possible	partly correct	wrong	seriously wrong	comment:

The following was used by the reviewers to indicate alternatives that they felt were left out.

Missing hypotheses: ______________________________

____

Comments: ______________________________

____

In addition a final page was included with the detailed graphical representation of the best hypothesis, in case the summary representation was insufficient.