Reprinted from *Computers in Biomedical Research,* **25:** 292-311, 1992

Development of a Knowledge Base for Diagnostic Reasoning in Cardiology*

William J. Long,
MIT Laboratory for Computer Science, Cambridge, MA, USA
Shapur Naimi and M. G. Criscitiello
New England Medical Center Hospital, Boston, MA, USA

April 13, 1994

Abstract

This paper reports on a formative evaluation of the diagnostic capabilities of the Heart Failure Program, which uses a probability network and a heuristic hypothesis generator. Using 242 cardiac cases collected from discharge summaries at a tertiary care hospital, we compared the diagnoses of the program to diagnoses collected from cardiologists using the same information as was available to the program. With some adjustments to the knowledge base, the Heart Failure Program produces appropriate diagnoses about 90% of the time on this training set. The main reasons for the inappropriate diagnoses of the remaining 10% include inadequate reasoning with temporal relations between cause and effect, severity relations, and independence of acute and chronic diseases.

[*] 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.

1 Introduction

Over the past several years we have been developing the Heart Failure Program to assist physicians in reasoning about patients with cardiovascular disease. The program takes a description of the case including information about the history, symptoms, physical examination, and test results, and generates a differential diagnosis that explains all of the findings that might indicate cardiovascular disease. The program can also suggest other measurements to refine the diagnosis and therapies to manage the problem, and can predict the hemodynamic effects of the therapies, but only the differential diagnosis is addressed by the experiments described in this paper (see other papers about other aspects of the system[1, 2, 3]).

This paper reports on a formative evaluation of the diagnostic capabilities of the Heart Failure Program (HFP). The process of formative evaluation combines aspects of system development with assessment of effectiveness and was undertaken with specific objectives in mind. The major part of the development effort on the basic diagnostic algorithms and diagnostic knowledge base of the program was completed and the program has been functioning in a reasonably stable way for a couple of years. In that time we identified two main circumstances that can lead to incorrect diagnoses: ones in which the temporal relationships among the diseases and findings determine the diagnosis, and ones in which the relationships between severities of findings are important. Both of these are problems that would require a major effort to solve in their full generality with the potential for greatly increased computational requirements, but it is possible to handle specific instances by making provision for them in the knowledge base. Since the frequency or extent of these problems in practice was unknown, we did not know their practical significance. Given this state of affairs, we conducted this development and assessment process to 1) determine the accuracy of the program with the present diagnostic algorithms, and 2) to determine the applicability of the system for diagnosis of patients typical of a tertiary care hospital.

To conduct the formative evaluation, we collected a set of 242 cases of patients classified by DRG (diagnosis related group) as falling within the domain of the program. On these cases we analyzed the performance of the program in its present state, refined the knowledge base to obtain the best performance achievable with the present algorithms, and used the results to focus our plans for further development. The cases were distilled from hospital discharge summaries and entered into the program. They were separately diagnosed by the cardiologists on the project from the program's case summary without seeing the computer generated diagnosis. Errors made by the program were classified into those correctable by refinements of the knowledge base and those that would require additional reasoning algorithms. The corrections to the knowledge base were made and the whole process repeated through a number of iterations until optimal correlation with the cardiologist's diagnoses was obtained. The program currently produces a first hypothesis which agrees with the diagnosis of the cardiologists in about 90% of the cases. We have analyzed the cases in which there remained disagreement with the clinical diagnoses after correcting the knowledge base. We used these analyses to categorize and determine the significance of the limitations of the current reasoning mechanisms.

2 Description of the Diagnostic Reasoning Mechanism

The Heart Failure Program (HFP) is a computer system which acts as an "intellectual sounding board", operating in the domain of cardiovascular disorders, assisting the physician by providing differential diagnoses for findings, predicting effects of therapy, and suggesting additional measurements, all with detailed graphical physiologic explanations. The mechanism for differential diagnosis consists of three parts, 1) an input interface that takes the findings about the patient in menu form, 2) a knowledge base in the form of a probability network of causal relationships between pathophysiologic states and findings, and 3) a reasoning mechanism designed to find likely explanations for the findings in terms of causal pathways through the pathophysiologic states. The result of differential diagnosis is an ordered list of complete explanations for the findings (called hypotheses) with relative probabilities.

The input interface is a dynamically expanding menu with entries divided into categorical and numeric values, with arbitrary constraints among categorical values and appropriate precision for numeric values along scales. The intention is to capture the information pertinent to the cardiovascular disease without requiring the system to do reasoning outside of the domain and to display the relevant patient information in an effective manner. It is assumed that the data has been interpreted and filtered by the user. For many test results an interpretation is entered rather than the test value, such as hypoxemia rather than a pO_2 value.

The HFP knowledge base (KB) is a clinically defined physiologic model of the cardiovascular system. From the perspective of diagnosis, the model consists of data structures representing physiologic states and measurement categories (the categories of patient information), constituting the general diagnostic knowledge about the domain and a template from which the more specific knowledge about a case is generated. Using the information from a case, the states and measurement categories are instantiated as nodes and findings representing the relationships that potentially exist in the case — essentially the superset of all possible diagnostic hypotheses for the patient. The states include diseases, qualitative states of physiologic parameters, and therapies. The measurement categories represent the observables entered in the input: the history items, symptoms, and lab results. The states are linked by probability relations, as are the relations between states and values in measurement categories. When a case is entered, the states and measurement values are instantiated as nodes and findings. The probability relations between them may be conditional on input values or on the nodes in a hypothesis. These probability relations are partially evaluated to provide the constraint implied by the input values.

There are two essential features of this knowledge representation that make the reasoning mechanisms tractable but also limit the expressive power of the KB. The first is the essentially binary nature of the physiologic states. For example, the *low cardiac output* node is either true or false. There are no degrees of severity. However, a parameter is not restricted to two states, so there is also a *high cardiac output* node with the constraint that high and low can not be simultaneously true. The probabilities between nodes can be adjusted for values in the input or even other nodes included in a hypothesis, overcoming some of this restriction, but essentially there is no representation of severity in the model. The second is lack of time relationships between nodes. For example, there is no way to represent and reason about a finding that was present yesterday but absent today. A hypothesis is a snapshot in time. This restriction is partially alleviated by having explicit nodes for some chronic states with different characteristics than their acute counterparts. For example, the KB has both *high left atrial pressure* and *chronic high left atrial pressure* represented. The only way to deal with effects that persist for some time after their causes cease is to make the cause node represent the average state of the cause over a longer time period and adjust the probabilities for findings that track the cause more closely. For example, the *high left atrial pressure* node represents an average value over hours so that it can act as a cause for pulmonary congestion, which takes hours to resolve. Otherwise, a single normal pressure reading would rule out a cardiac cause for pulmonary congestion. Without a time representation, the usual assumptions about the independence of diseases are also suspect. That is, the probability of disease combinations involving chronic diseases is actually much higher than simply multiplying the probabilities would indicate.

The KB covers the common and some not so common causes of heart failure or hemodynamic disturbance including myocardial ischemia and infarction, congestive, restrictive, and hypertrophic cardiomyopathy, the valvular diseases, atrial and ventricular septal defects, constrictive pericarditis and tamponade. It also has non-cardiac diseases that cause the same symptoms or complicate the hemodynamic situation such as pulmonary, renal, liver, or thyroid diseases, anemia and infection. The non-cardiac diseases are not further differentiated. That is, there is a node for *primary liver disease* but not for any specific types of liver disease.

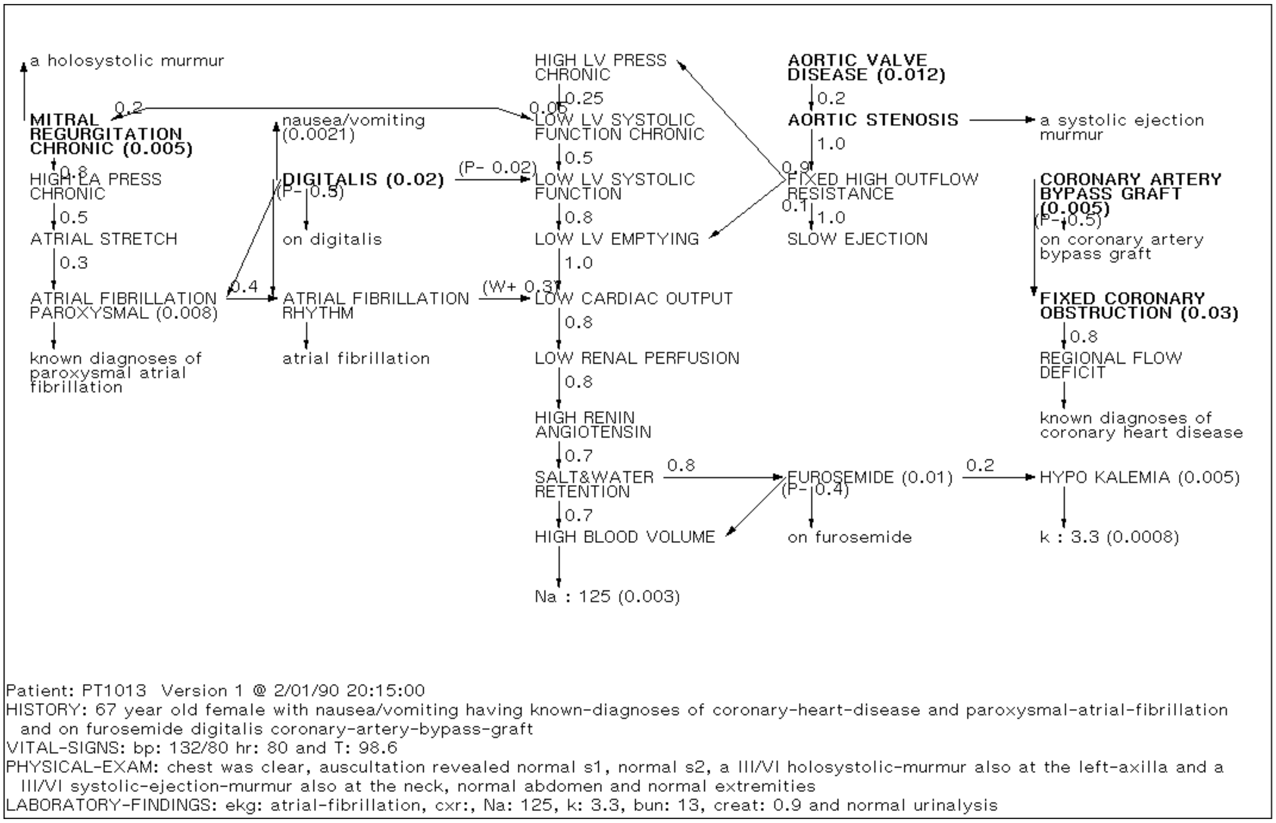
The probabilities on the links between nodes are combined using a "noisy-or" combination rule[4] except for special links called worsening factors, which increase the probability of another cause but are insufficient to produce the effect alone, and correcting factors, which decrease the probability. Thus, if the causes are P, the worsening factors W, the correcting factors C, and at least one of the causes in P is true, the probability of a node is $(1 - \prod_{i \in P,W}(1 - p_i)) \prod_{i \in C}(1 - p_i)$. Similarly, each finding has a probability of being produced by nodes. The model is similar to those investigated by Pearl[4] as Bayesian probability networks. However, this model has forward loops (excluded by Pearl), some probabilities that are conditional on other nodes in the hypothesis, and nodes with multiple paths between them (handled only in exponential time by Pearl's methods). Thus heuristic methods are necessary to reason about the model.

Our solution to the differential diagnosis problem is to generate complete hypotheses (causal paths from primary causes) for the findings and present the user with a list of hypotheses and their relative total probabilities for comparison. In comparing hypotheses we discovered that the natural notion of *different* hypotheses requires that they differ in some significant node, nodes which we have labeled *diagnostic*. The algorithm is as follows: 1) check the input for definite implications, findings that require nodes to be true or false (known diagnoses, therapies, or pathognomonic findings); 2) collect the abnormal findings from the input; 3) find all of the diagnostic or primary nodes that could account for each finding; 4) rank the diagnostic and primary nodes by the number of findings they account for; 5) use the better of these as the initial nodes for generating small covering sets of primary nodes; 6) for each covering set, order the findings by the difference between the first and second highest probability path to it (since the best path to those findings are most likely to remain best as the hypothesis evolves); 7) for each finding, find the best path from the partial hypothesis and add it; and 8) prune the hypothesis of unneeded primary nodes and extra paths that decrease the probability. Finally, the probabilities of the hypotheses are computed by multiplying the probabilities of the nodes given the other nodes in the hypothesis and they are rank ordered and presented to the user. It is unnecessary to normalize by the probability of the findings as long as we are only interested in the rank order and relative probabilities of the hypotheses.

Figure 1 shows a simple, but otherwise typical, best hypothesis from case 13 in the format HFP displays it on the screen. This hypothesis can be summarized as having causal pathways that show:

1. coronary heart disease with a coronary artery bypass graft

2. aortic stenosis causing low LV systolic function, low cardiac output, salt and water retention, and high blood volume



Patient: PT1013 Version 1 @ 2/01/90 20:15:00
HISTORY: 67 year old female with nausea/vomiting having known-diagnoses of coronary-heart-disease and paroxysmal-atrial-fibrillation and on furosemide digitalis coronary-artery-bypass-graft
VITAL-SIGNS: bp: 132/80 hr: 80 and T: 98.6
PHYSICAL-EXAM: chest was clear, auscultation revealed normal s1, normal s2, a III/VI holosystolic-murmur also at the left-axilla and a III/VI systolic-ejection-murmur also at the neck, normal abdomen and normal extremities
LABORATORY-FINDINGS: ekg: atrial-fibrillation, cxr:, Na: 125, k: 3.3, bun: 13, creat: 0.9 and normal urinalysis

Key:	number on link	causal probabilities (not displayed to findings)
	$(P - 0.5)$ on link	probability of correcting effect
	$(W + 0.3)$ on link	increase in causal probability
	(number) after node	probability of occurring without cause in model
	bold face node	diagnosis or primary node
	lower case node	finding from the input

Figure 1: First hypothesis for case 13

3. mitral regurgitation secondary to low LV systolic function (through a dilated LV) and causing atrial fibrillation

4. on digitalis which is causing nausea/vomiting

5. on furosemide which is causing hypokalemia

This summary mentions only selected nodes, but one could produce a text version of the detailed hypothesis, although that would be difficult to follow. For example, the text version of the third causal path, leaving out the findings would be:

> The patient has chronic mitral regurgitation which is either primary with p=0.005 or secondary to chronic low LV systolic function with p=0.05. The chronic mitral regurgitation causes chronic high LA press with p=0.8, which in turn causes atrial stretch with p=0.5. The patient has paroxysmal atrial fibrillation which is either primary with p=0.008 or caused by atrial stretch with p=0.3. However, the digitalis decreases the probability of paroxysmal atrial fibrillation by p=0.3. Paroxysmal atrial fibrillation causes atrial fibrillation rhythm with p=0.4 but the digitalis decreases the probability by p=0.5.

The hypothesis accounts for all of the abnormal findings and shows the causal paths that explain the hypothesized mechanisms.

3 Collection of Case Material

The cases used for testing and refining the program were gathered from discharge summaries at The New England Medical Center Hospital of patients hospitalized in 1988 and early 1989. These were all patients for whom discharge summaries were available, in Diagnosis Related Groups (DRG) 121, 123, 127, and 135. These DRGs include patients with acute myocardial infarctions (MIs) with cardiovascular complications (both discharged alive and expired), heart failure and shock, and valvular disorders with complications. Since these are general categories with many specific diseases included and additional diseases present as complications, these DRGs contained all of the available types of cases that are relevant to the domain of the Heart Failure Program. There are other DRGs that the program could handle, such as uncomplicated MIs, but most of these cases are relatively simple. The DRG for complicated MIs included enough variety to test the program more thoroughly for acute ischemia cases than for most other diseases. By choosing these DRGs the case set was assured of including the most complicated cases, providing the greatest challenge for the program.

The discharge summaries, usually two to three page long accounts of the patient admission, include the history of the present illness, medications, physical exam, and laboratory data on admission, short description of the hospital course, final diagnosis, and discharge information. These summaries are dictated by the house officers from the medical record after the patient has been discharged. Because these are usually highly condensed, they do not provide all of the data available in the patient record. However, they proved to be useful summaries of the patient data and not atypical of a case description that might be given by a house officer.

A total of 246 cases were collected. Of these four were eliminated because of insufficient data. We were very liberal in accepting cases, assuming that if the cardiologists examining the information could come to some reasonable conclusion from the data, the program should also be able to. Thus we included cases from which a considerable amount of data was missing. The average age of the patients was 67.5, including 8 from 24 to 40 years old with 120 females and 122 males.

The New England Medical Center Hospital is a tertiary care, teaching hospital, so the patient population includes a large number of complex cases. Of the 30 disease entities included in the model knowledge base, 19 are definitely true of patients in the sample according to the expert diagnoses and three more are possibly true. These 22 diseases are listed in table 1. The columns are the number of cases in which the disease was known to be present prior to admission, the number in which it was a definite part of the expert's diagnosis but not previously known, and the number in which it was mentioned as a possibility in the diagnosis.

	Previous	Definite	Possible
Disease	Known	New	New
adult respiratory distress syndrome	0	0	1
aortic regurgitation	12	6	2
aortic stenosis	18	5	36
chronic hypertension	109	3	8
congestive cardiomyopathy	104	23	9
COPD or chronic bronchitis	45	7	13
exertional angina	33	61	66
hypertrophic cardiomyopathy	2	0	1
mitral prolapse	3	0	0
mitral regurgitation	14	43	27
mitral stenosis	10	1	7
myocardial infarction	17	22	28
pericarditis	3	0	2
pneumonia	1	3	16
pulmonary embolism	1	2	22
pulmonary hypertension	2	3	2
pulmonic regurgitation	0	0	2
renal insufficiency	55	4	3
septic shock	0	0	1
tricuspid regurgitation	7	4	8
unstable angina	19	3	70
ventricular septal defect	1	0	2

Table 1: Case mix of data set according to expert diagnosis

The number of patients labeled as *congestive cardiomyopathy* is high because we considered this to be a physiologic diagnosis including a number of etiologies, such as ischemic, hypertensive, and other causes of a dilated heart. The high number of patients labeled as *renal insufficiency* results from a liberal definition of renal insufficiency on the discharge summaries and does not mean that the renal function tests were abnormal at the time of admission. The program distinguishes between acute and chronic variants of mitral regurgitation and renal insufficiency, but the expert diagnoses often did not, so they are included together here.

These cases were primarily ones in which multiple diseases were present. There were an average of 1.7 diseases known prior to admission and an additional average of 1.5 diseases definitely present as part of the expert diagnosis. There were only 5 cases in which the experts could not identify any diseases as definitely true. (If we had included more simple cases, there would have been many since all acute MIs or new unstable angina with no other diseases would have been in this category.) There were 6 cases in which there were 6 or more known or definite diseases.

4 Evaluation and Refinement Method

The discharge summaries were used to create worksheets with the information used by the program. Since the most completely described point in the discharge summaries is the initial examination, that point was used for determining a diagnosis. Thus, the presenting illness, initial examination data, and some laboratory findings were likely to be available. Anything measured or done after that time was excluded from the input.

Filling out the worksheets involved some interpretation since the terminology used in the discharge summaries was not always consistent with the measurement values used by the program. Some of the program inputs are interpreted values rather than raw test results. To translate test values such as pO_2, pCO_2, hematocrit, and white blood count into the qualitative values accepted by the input menu, a table was used to maintain consistency. Interpretation of chest pain and electrocardiogram (EKG) results was more difficult. For chest pain, the program has four descriptors: *anginal, atypical, pleuritic,* and *other non-ischemic chest pain.* Chest pain whose description was consistent with the characteristic attributes of anginal chest pain, was entered as *anginal.* If there were characteristics of the chest pain that were not typical, but it had some of the features of anginal chest pain, it was entered as *atypical.* To interpret the EKG description, it was necessary to decide whether the description was consistent with old, evolving, or acute MI, or with ischemia. Often the description was in these terms, but when it was in terms of changes in specific leads, we used a simple table to translate the description. When the description did not match any of the interpretation descriptions, but there were still changes in the ST segment or T wave, it was entered as non-specific ST and T changes. Times of tests are not presently included in the input to the program. This precludes distinguishing between tests done during the present admission and ones done in the past that still provide useful information. (This will be corrected in a future version of the program.) The murmur descriptions also presented problems. Many times the murmurs were only described as systolic or diastolic without specifying character. These were given default characteristics as appropriate. The location descriptions were often less specific than the options in the program, and these were translated using a simple table. Sometimes locations were omitted entirely. In case 13 only the secondary locations, the locations of radiation, were given.

Once the worksheets were completed, the patient information was entered and the program was used to print a textual version of the information. For example, the computer generated description of case 13 was:

Patient: PT1013 Version 1 @ 2/01/90 20:15:00

HISTORY: 67 year old female with nausea/vomiting having known-diagnoses of coronary-heart-disease and paroxysmal-atrial-fibrillation and on furosemide digitalis coronary-artery-bypass-graft

VITAL-SIGNS: bp: 132/80 hr: 80 and T: 98.6

PHYSICAL-EXAM: chest was clear, auscultation revealed normal s1, normal s2, a III/VI holosystolic-murmur also at the left-axilla and a III/VI systolic-ejection-murmur also at the neck, normal abdomen and normal extremities

LABORATORY-FINDINGS: ekg: atrial-fibrillation, cxr: no cardiac-enlargement, Na: 125, k: 3.3, bun: 13, creat: 0.9 and normal urinalysis

These program generated summaries were used by the cardiologists to determine the diagnoses. The final diagnoses given in the discharge summaries were not used because they were diagnoses based on more information than was available initially or even than was included in the summary. Often those diagnoses were not adequately supported by the information in the discharge summary. Using the program summaries means that the program diagnoses are determined from the same information as the expert diagnoses. The diagnoses were determined by agreement between the two cardiologists. In the process of examining the summaries a number of data inconsistencies were discovered which were corrected by a more careful reading of the discharge summary.

A typical expert diagnosis (the one for case 13) is:

> Coronary heart disease, atrial fibrillation, compensated heart failure, mitral regurgitation, aortic stenosis or aortic sclerosis, possible digitalis toxicity, and possible diuretic complications

This diagnosis is actually a differential in terms of the program because it admits a number of possible hypotheses expressed as nodes by the program. The hypothesis must have atrial fibrillation. It must have either aortic stenosis or aortic sclerosis. It may or may not have digitalis or furosemide (a diuretic) accounting for abnormal states. Coronary heart disease is not a single node but may be accounted for by either fixed coronary obstruction or an old MI. For such terms used in diagnosis we defined expansions in terms of the nodes in the KB. Compensated heart failure is part of the heart failure syndrome and can have several incarnations. There must be either low left ventricular (LV) systolic function, low LV compliance, or low right ventricular (RV) systolic function. These are the minimum systolic or diastolic manifestations that would count as heart failure. Since it is compensated, the left atrial pressure (LAP) is not high and there are no congestive findings on the left or right. This description does not include all of the intermediate pathophysiologic nodes that might be in a corresponding computer generated hypothesis, but we assumed that all of the diagnostic nodes in a computer hypothesis are either in the expert diagnosis or definitely implied by the input. That is, states listed as previously known diseases in the input, heart rhythm on the EKG, or other direct inferences are automatically included in the diagnosis. When the diagnosis included *unknown etiology*, we allowed the program to attach some plausible etiology, since HFP produces completely specified hypotheses.

The hypotheses implied by the expert diagnoses were considered unordered. There were times during the process of collecting the diagnoses when the cardiologists put some partial ordering on them, such as stating that the patient probably had unstable angina, but it could be an MI. Because this information is only available in a fraction of the cases and the difficulty in using it in the comparisons, we chose to consider all of the expert differential diagnoses as unordered.

Once the diagnoses were decided by the cardiologists, we used HFP to generate a differential diagnosis. The differential diagnoses produced by HFP are ordered lists of one or more completely specified hypotheses with relative probabilities, as discussed in section 2. The criterion we used for accepting the machine's diagnosis is that the top hypothesis in the differential list match one of the admissible diagnoses listed by the experts. That is, the top hypothesis must include all of the required entities in the diagnosis, may include any of the optional entities, and may not include any diagnostic node that is not part of the diagnosis or definitely implied by the input. More or less stringent criteria could have been specified, from all hypotheses acceptable and all acceptable hypotheses included in the differential to some acceptable hypothesis included in the differential. Any kind of comparison that considers how many of the alternatives in the expert diagnosis are included in the computer differential is difficult because it depends on what probability cutoff is chosen in selecting the differential and because the process of pruning the hypotheses eliminates some of the alternatives. The top hypothesis criterion seemed best for identifying the main issues for further research.

The matching process can be illustrated with the case in figure 1. Coronary heart disease is covered by fixed coronary obstruction; atrial fibrillation was explicit; compensated heart failure was manifest as low LV systolic function; mitral regurgitation is included as a chronic condition; aortic stenosis, rather than aortic sclerosis was used to account for the systolic ejection murmur; digitalis accounted for the nausea or vomiting, a sign of toxicity; and furosemide accounted for the low potassium. The hypothesis also suggests a mechanism for the low sodium level, but no diagnostic node is included in that causal chain so it is not evaluated. There are no diagnostic nodes in the hypothesis that are not accounted for by the diagnosis, so the match is successful.

The matching is done by a small program that takes the diagnosis and a table of the description translations and generates all of the allowed combinations of nodes. It compares that list to the top hypothesis. If there is a combination of nodes that all occur in the hypothesis and any other nodes in the hypothesis are non-diagnostic or definitely true from the input, the match succeeds.

If the top hypothesis was not acceptable, there were three possible explanations: 1) the hypothesis from HFP was wrong, 2) the expert diagnosis was wrong, or 3) the translation of the diagnosis into required nodes was wrong. We reviewed a sample of the unacceptable cases, analyzed the nature of the problems, corrected what was easy to correct (either the KB, the expert diagnosis, or the diagnosis translation, as appropriate), and repeated the process. Over the course of a dozen iterations 93 of the cases were analyzed in detail by the cardiologists, some more than once.

The analysis of the erroneous hypotheses and the corrections to the KB will be discussed in the next section. The kinds of corrections that were made to the expert diagnoses are shown in table 2.

Correction	Count
add disease	38
make definite disease possible	16
remove disease	2
make disease more specific	7
Total corrected diagnoses	61

Table 2: Corrections made to expert diagnoses

For the most part these are relatively minor changes in the diagnosis, for example, adding possible chronic obstructive pulmonary disease (COPD) that was overlooked or not requiring aortic stenosis that is only supported by a murmur that could be functional. Still, it is indicative of how difficult it is to specify a complete diagnosis.

The translation of the diagnosis into nodes in the model was also fairly difficult. For example, stating that the patient had left heart failure usually meant that there was the systolic causal chain of low LV emptying, low cardiac output, and high LAP. However, it also happened that the high LAP could be caused by diastolic dysfunction manifest as low LV compliance, LV hypertrophy, or some cause that produced a chronic state of high LAP (such as mitral stenosis) and those situations were also called left heart failure. Furthermore, specifying that the left heart failure was systolic, diastolic, or compensated changed the list of nodes that characterized the state. It took a number of iterations to get the matching program to accept all of the hypotheses that were in fact consistent with the diagnoses.

5 Analysis of Results

The process of revising the KB, expert diagnoses, and diagnosis interpretation was done iteratively. That is, after each run we reviewed a number of the cases in which the program did not have an acceptable hypothesis as the first hypothesis, determined the types of errors involved and made revisions. Because of the time involved in this process, we did not review all of the unacceptable cases until the numbers were manageable. However, one approximate way to measure the effect of the changes made in the KB separately from changes in the expert diagnoses and interpretation is to rerun the cases with the original KB. Doing this yields 141 of 242 cases correct or 58.3%. This is likely to be somewhat low since many of the failed cases have not been reviewed and may have acceptable top hypotheses. In the final run of the cases through the program, the program produced good first hypotheses for 216 of the 242 cases or 89.3%. The cases that failed are listed in table 3. The refinement process leading to this state has not been completely monotonic. Indeed, three of the cases that failed in the final run are done correctly using the original KB. Thus, at this point handling the existing problems in a more general way is more likely to produce a robust diagnostic program than further adjustments to the probabilities in the KB.

We analyzed the errors in the failed diagnoses and attempted to categorize the problems. Because of the interactions among different parts of the KB and the diagnosis process, the classification is somewhat subjective and for many of the cases multiple explanations are possible. The most illuminating classifications (those used in table 3) included the varying manifestations with different chronicities and severities of the causes, the independence assumption for diseases, incomplete or imprecise data, and the generation of causal pathways especially when there were conditional probabilities involved. In the following paragraphs we will discuss the nature of these problems in the final run plus illustrative examples from the previous runs.

5.1 Missing and Inaccurate Findings

The problems with data seem to be due either to the abstraction process involved in summarizing the essential information about a case or to lack of careful specification of findings. These are accentuated by the nature of discharge summaries, but are not unique to them. Summarizing an examination requires the physician to leave out the information that in his or her judgement is not pertinent, which is a very context sensitive process and may be biased (since the physician knows the diagnosis when the summary is generated). The diagnosis method requires knowing more of this information than is typically included in the summary, so the first problem is reconstituting the information that was summarized away.

Case	Classifications[a]						Summary
	C	S	I	M	R	F	
50			x				COPD and infection vs pneumonia
79		x					accounting for mild pedal edema
96					x		most probable vs most important
100				x			murmur with multiple locations
102					x		COPD incompatible with hypocapnia
107			x				cause for ventricular septal defect
116			x				mitral stenosis causing pulmonic regurgitation rare
121		x			x		probabilities dependent on hypothesis
125	x					x	chronic high LAP with few findings
128	x					x	high LAP with tachypnea but no dyspnea
133		x					accounting for mild pedal edema
139				x			bad murmur description
141	x	x					specialization of renal insufficiency
148					x		probabilities dependent on hypothesis
153			x	x			murmur with multiple locations
171			x				using aortic stenosis to account for unstable angina
180					x		missing best causal path
183			x				mitral regurgitation vs functional murmur
202					x		almost incompatible nodes
209		x			x		probabilities dependent on hypothesis
212						x	echo rules out cardiac causes of edema
213	x						chronic high LAP with few findings
214		x					accounting for mild pedal edema
231					x		missing best causal path

[a] Classifications: C: chronicity, S: severity, I: independence of causes, M: murmur interpretation, R: reasoning mechanism, F: findings not given

Table 3: Classification of errors made by the program

Missing Data The interpretation of missing data is a difficult issue. When a particular finding is not mentioned, it may be unknown because the value was never determined, false and not deemed a significant negative, or true but considered redundant in the context of the case. The general strategy in HFP is to handle the data by categories. If any value in a category is specified, it is assumed that all items in that category are known within a set of constraints about what findings can hide others. For example, if an EKG finding of LBBB (left bundle branch block) is specified, EKG findings of LV hypertrophy and LV strain are considered unknown because they can be obscured by LBBB and all other EKG findings, such as RV hypertrophy or long PR interval, are considered false because they would have been observed in reading the EKG. From examination of the assertions made in typical cases, this mechanism seems to work well for test results, but works less well for physical examination findings and history findings. For example, the jugular venous pressure was often noted but rarely were any other characteristics of the jugular pulse commented on, so they were treated as absent under the assumption they would have been noticed when determining the jugular pressure. However, these other findings require more careful observation than just determining the pressure. This means that the probabilities of diseases such as tricuspid regurgitation producing jugular findings appears low, while in more completely described physical exams, they might be higher. Such missing physical exam findings may be due to the summarization process or to increased reliance among physicians on test results such as echocardiography instead of the more subtle physical exam findings.

Some findings are related to one another in ways that are not readily captured by the probabilistic formalism. For example, in case 128 there is significant tachypnea (rapid breathing) but no dyspnea (difficulty catching breath) mentioned. Since tachypnea is easily and reliably measured by the observer and dyspnea is dependent on the patient's description and perceptions, the expert takes the more reliable evidence and attributes the lack of dyspnea to the process of data collection. For the program, the lack of dyspnea and lack of other evidence of pulmonary venous congestion (attributable in this case to the acuteness of the situation) ruled against high LAP as an explanation.

Symptoms such as chest pain presented a particular problem. On the program data collection menu, chest pain is a separate category, so initially if it was not marked as present or absent, the program treated it as unknown. Since it is reasonable to expect that chest pain would be mentioned if it had been present within the hours prior to the examination, the program was changed to assume that the values representing recent chest pain are false if not entered. There are other findings where such assumptions would be reasonable, such as extreme tachypnea or hypotension, but the program needs more capabilities of handling severity to take advantage of this information. In a number of cases no therapies were listed on admission and the program assumed all therapies were false. For most of these the patient was probably not taking any medications, but there were several cases where the experts suspected drug toxicity from chronic medications that were probably left out of the summary. For example, a chronic hypertensive patient with low blood pressure might have been receiving too much anti-hypertensive, even though it was not listed.

Incomplete Data Incomplete data was also suspected in some of the test values. For example, in case 125 the only X-ray finding mentioned was pleural effusion. This was a case of chronic heart failure, so there was probably evidence of pulmonary congestion in the lung fields as well. Since pleural effusion is a finding indicating more chronic disease, and in a sense more severe disease, only it was mentioned. However, there are other causes for pleural effusion, so with X-ray lung field indications false (assumption of completeness in a category), the program produced a hypothesis with pulmonary congestion false and a more unusual explanation for the pleural effusion.

The lack of times on test values, especially echocardiogram results, was a problem for the program. Often the echocardiogram results were years old, but they still provided useful information since many of the conditions only get worse. To account for this, initially the program considered any unspecified echocardiogram values as unknown rather than false. This proved to be a problem because the program would often propose a disease such as aortic stenosis as a cause, based primarily on its prevalence. Since aortic stenosis is a disease that progresses over years, even an old echocardiogram without aortic stenosis is good evidence against it. On the other hand, a disease such as mitral regurgitation can happen acutely. As a result, we developed a table of probabilities that the various disease states would appear as findings if the test were current, days, months, or years old. In the absence of information about the age of the test, months is assumed. Still, the normal ejection fraction (indicating normal systolic cardiac function) on an old echocardiogram in case 212 was enough to misdirect the program to account for the congestive findings with renal rather than cardiac causes.

Murmurs The findings that have proven the most difficult to interpret are murmurs. Murmurs present a particular challenge for the probability network formalism because there can be multiple murmurs and they have varying location and extent. For example, mitral regurgitation can have a systolic murmur that is usually holosystolic but can be shorter in duration. It is usually loudest at the apex, but sometimes in the third, fourth, or fifth left interspace. It is often also heard in the left axilla, or possibly in a number of other locations. It can be of any intensity, while most other murmurs have restrictions on their intensity. Mitral regurgitation can also have a diastolic rumble murmur in association with the systolic murmur when severe. Since the descriptions for the murmurs from different valvular lesions overlap considerably, we use a scheme for adjusting the probability that a murmur is caused by a particular lesion. The probability is multiplied by a factor for each of the characteristics reflecting how typical the value is for that disease. Since there may be multiple values for other locations, only the most typical is used.

This scheme has proven fairly effective for setting the probabilities to reflect how likely a disease is to cause a particular murmur, but there are still several problems. First, it is very difficult to fill in the table of allowed murmur characteristics. It is easy to describe the typical murmurs of particular lesions. It is hard to imagine all of the values that would still be consistent with a particular lesion. In the process of reviewing the cases, we made seven changes to the table of murmur characteristics. Second, the location and extent of murmurs means that a single murmur may be mistaken for two and two murmurs may be mistaken for one. The program independently considers each murmur and will often attribute multiple murmurs to the same lesion, more often than it should, since the intent on the part of the physician is that separately described murmurs (of the same type) do not seem to be coming from the same lesion. The problem of finding two causes for a singly described murmur is harder. Usually the evidence that this is happening is multiple alternate locations, some of which are rare or impossible for any single lesion. Separating the murmurs is complicated because there is no way to tell where the primary location of the second murmur might be. In cases 100 and 153 two murmurs are probably described as one, since the multiple locations specified are not consistent with any single murmur. Case 139 has a murmur description that is just inconsistent. In that case, the cardiologists discount the location and make the proper attribution using other findings. The program ignored the murmur entirely but still had the correct diagnosis as the second hypothesis on the final run and the first hypothesis on several other runs.

5.2 Severity and Chronicity

The lack of reasoning about severity and chronicity of diseases is a problem that caused many of the diagnostic failures. In addition, there are many unreasonable alternate hypotheses that should be eliminated for reasons of severity and chronicity. The program should not use an acute cause to account for a chronic effect or (in most situations) a mild cause to account for a severe effect. For example, an acute MI can cause pulmonary congestion and pulmonary congestion in general can produce findings such as nocturnal dyspnea (PND) over days, but the pulmonary congestion caused by an acute MI a few hours previous has not been present long enough to account for PND. Acute and chronic versions of the same pathophysiologic state can have different findings. As mentioned above, the acute pulmonary congestion in case 128 had only tachypnea and none of the X-ray or chest findings that come usually within a few hours. In cases 125 and 213, the opposite problem existed. When high LAP has existed for years, the lungs adapt and there are again few findings indicating pulmonary congestion, but the program took this lack of findings as evidence against high LAP.

The severity of a disease also affects how it presents, with milder forms having fewer findings. In cases 79, 133, and 214 there was known congestive cardiomyopathy and mild pedal edema with no other findings consistent with right heart failure. Since the causal pathway from the disease to pedal edema goes through several nodes, each of which may have negative evidence, the program estimates the probability to be higher if pedal edema is considered as having an external cause or was a therapy side effect than to include the additional nodes in the hypothesis. In reality, if the right heart failure is mild, the abnormal state of the intermediate nodes may be undetectable, so the lack of evidence should not count against the hypothesis. Whenever there is a physiological state that represents the cumulative effect of a dynamic process, there may be a difference between the process currently and the state. This makes the appropriate handling of chronicity and severity a necessity.

One approach we used to avoid the problem of severity is to make some of the causal probabilities between nodes conditional on other nodes being in the hypothesis. For example, only chronically very high pulmonary artery pressure can cause pulmonic regurgitation. That limits the ultimate causes to primary pulmonary hypertension or mitral stenosis, two states that produce very high pulmonary artery pressure over years. Making the immediate link conditional can interfere with the generation of hypotheses (as it did in cases 121 and 209) since the probability of a pathway may change when other nodes are added to the hypothesis, increasing the risk of missing good hypotheses. It only covers up the real problem of reasoning about severity and chronicity.

The KB does not currently differentiate between different kinds of renal diseases. This has proved to be a problem in several cases, because some renal disease can cause pedal edema. When pedal edema is the only manifestation of right heart failure and there is known renal disease, the program often hypothesizes that the pedal edema is due to the renal disease rather than hypothesizing the right heart failure mechanism, which may have considerable negative evidence along the causal chain. The problem is that only nephrotic syndrome or renal disease severe enough to cause oliguria produce pedal edema, and these constitute less than 10% of renal disease. In case 141 there was proteinuria, which is characteristic of nephrotic syndrome, but not knowing the severity of the proteinuria it is still more likely that the pedal edema is caused by right heart failure than by renal disease.

5.3 Independence of Diseases

In several cases the primary problem is the nature of the independence of diseases. Currently, each disease that is considered primary is assigned a prior probability, usually dependent on age or other attributes in the input. In a hypothesis with two or more primary diseases, these prior probabilities are multiplied together. This simple notion of independence overlooks the differences between diseases that are chronic and therefore have higher probability of coexisting with other diseases and those that are over by the end of the hospital stay. For example, in case 50, COPD with infection was considered less likely than pneumonia even though the COPD can be present in a patient for thirty or more years.

The independence assumption also overlooks the fact that a patient with chronic diseases is more likely to need hospitalization for a mild disease that would not require hospitalization in an otherwise healthy patient. As a result, the program can have probabilities that are too low on hypotheses involving multiple chronic diseases and probabilities that are too high on ones with multiple acute diseases. (This problem is addressed in a recent paper[5].)

Because of the low probability of primary diseases, the program has a tendency to invoke relatively unusual causal mechanisms rather than add a new primary disease. For example, in case 107 it explained an unusual prosthetic valve murmur as evidence for a ventricular septal defect caused by a known MI. In case 171, known aortic stenosis was used to account for unstable angina, rather than adding coronary artery disease. While aortic stenosis can cause exertional angina, it is rare that the angina would be unstable because of the fixed nature of the aortic stenosis, unless there were coexisting coronary artery disease. This problem is partially a question of independence and partially a more complex relationship among pathophysiologic states. In case 116, with known mitral stenosis HFP invoked the causal mechanism to pulmonic regurgitation to account for a murmur. Even though mitral stenosis can cause pulmonic regurgitation, this is sufficiently unusual that an additional primary disease, aortic regurgitation, is a better explanation. HFP missed that explanation because it underestimated the probability of the two chronic primary diseases coexisting. Because mitral regurgitation is often secondary to cardiac dilitation from low LV systolic function, the program sometimes will invoke this explanation to account for an underspecified murmur, as it did in case 183. Some of the problems with the degree of dependence among valvular lesions could also be addressed by including more of the disease processes in the KB that lead to multiple valvular lesions. These include rheumatic heart disease and endocarditis, especially that caused by intravenous drug abuse.

5.4 Reasoning with Causal Pathways

The process of generating hypotheses is heuristic in HFP. This has led to three kinds of errors in the cases: not finding the best causal path, inappropriate handling of conditional probabilities, and not accounting for unlikely combinations of nodes. In cases 180 and 231, there were better causal paths to account for the findings but taking the findings in the heuristic order and searching for causal paths, they are missed. In cases 121, 148, and 209, there are conditions on the causal probabilities that depend on multiple nodes (as mentioned in section 5.2). In the generation of the hypotheses, these nodes do not assume truth values until after the wrong paths have been chosen. Both of these problems can be addressed with a mixed strategy of developing the hypotheses from both the diseases and the findings, rather than exclusively using the findings to search for causal pathways.

In cases 102 and 202 there are unlikely combinations of nodes. In case 102 there are COPD and hypocapnia (low pCO_2). Since COPD often causes hypercapnia (high pCO_2) and always tends to increase the pCO_2 even when it stays in the normal range, this is unlikely. Similarly, case 202 has both low blood volume and high cardiac output, which are almost incompatible. Both of these problems require further modifications to the KB.

One final error illustrates the considerations that go into a diagnosis. In case 96, HFP misses pulmonary embolism in the hypothesis and leaves pleuritic chest pain unexplained. Muscloskeletal or other non-cardiopulmonary causes for the chest pain may indeed have a higher probability than the desired hypothesis with pulmonary embolism, but it is so important that the program catch the possibility of the treatable disease that we left the pulmonary embolism as a mandatory part of the diagnosis. This kind of problem must be handled outside of the current mechanism for hypothesis generation using a scheme for assessing the utility of the hypothesis as well as its probability.

6 Summary

Using the 242 cardiac cases collected from discharge summaries to conduct a formative evaluation of the Heart Failure Program has shown that with some adjustments to the knowledge base the mechanism for generating differential diagnoses using a probability network and a heuristic hypothesis generator is effective enough to produce appropriate diagnoses about 90% of the time in this training set. This seems quite respectable for several reasons: 1) The cases included the more complicated ones in a tertiary care hospital and excluded more simple ones. 2) The criteria for appropriate diagnosis was demanding, requiring that the first hypothesis exactly match one of the interpretations of the cardiologists' diagnosis. 3) In many of the failing cases the second hypothesis was satisfactory. 4) Almost all of the errors involved single aspects of complicated diagnoses.

However, the errors made by the program are ones that are obvious to the cardiologists using the same data. All the errors have been reviewed, eliminating any due to oversight. Thus, if the program can not produce an acceptable diagnosis, there is some weakness in the representation or use of the knowledge. It is probably possible to make further modifications to the KB until all of the top hypotheses are correct, but from the changes that have been made it is apparent that some of the diagnoses are becoming sensitive to small changes in causal probabilities. This indicates that the changes may not improve the performance of the program on future cases and that we are now placing too much reliance on a mechanism that is not using all of the information available about the diagnostic problem. Therefore, it is more appropriate to enhance the reasoning mechanisms to deal appropriately with the problems identified in the previous section. These include relationships of chronicity, severity, disease coexistence, more complicated dependencies among pathophysiologic states, and context sensitive interpretation of findings.

The second objective of the evaluation was to determine the applicability of the program to the types of cases that appear in a tertiary care hospital. The sample of cases in the study included all of those for whom discharge summaries were available and by DRG fit in the general categories of complicated cardiovascular disease. There were no cases that the program was unable to handle because the relevant diseases were not included in the KB. There were eight primary diseases covered by the KB that did not appear in the sample, because of their rarity. This is no guarantee that all future cases will be covered, because there are rare diseases that we know are not included, but it is a good indication that these situations will be rare.

References

[1] Long, W. J., Naimi, S., Criscitiello, M. G., and Larsen, G. Differential Diagnosis Generation from a Causal Network with Probabilities. *In* "Proceedings, Computers in Cardiology Conference," Washington, DC, 1988, pp 185-188.

[2] Long, W. Medical Diagnosis Using a Probabilistic Causal Network. *Applied Artificial Intelligence* **3,** 367 (1989).

[3] Long, W. J. Flexible Reasoning about Patient Management using Multiple Models. *Artificial Intelligence in Medicine* **3,** 3 (1991).

[4] Pearl, J. Fusion, Propagation, and Structuring in Bayesian Networks. *Artificial Intelligence* **29,** 241 (1986).

[5] Long, W. J. The Probability of Disease. *In* "Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care," Washington, DC, 1991, pp 619-623.