Reprinted from *Deep Models for Medical Knowledge Engineering*, E. Keravnou, ed., Elsevier Sciences Publishers B.V., pp. 147-160, 1992

Validation of a causal probabilistic medical knowledge base for diagnostic reasoning*

William J. Long[a], Shapur Naimi[b], M. G. Criscitiello[b], and Ravi K. Adusumilli[b]

[a]MIT Lab for Computer Science, 545 Technology Sq. Rm 420A, Cambridge, MA, USA

[b]New England Medical Center Hospital, 750 Washington St., Boston, MA, USA

Abstract
This paper discusses the issues involved in validating a large medical knowledge base for diagnostic reasoning using a pseudo-Bayesian network of physiologic causal relations, based on our on-going experience in validating the Heart Failure Program. Validation is addressed at two levels, 1) methods for determining the extent, topology, and probabilities in the network and 2) methods for assessing, comparing, and learning from the diagnoses produced by the program. The local validation of the knowledge base requires the developer to face the issues of the boundaries of the domain, both in coverage and detail, versus the purposes of the program. The validation of the topology includes issues of imposing causal structure on everything from associations to feedback relations and determining what level of granularity captures clinical relevance. The probabilities require the use of clinical expertise checked against the literature and the available statistical data.

Validation at the level of the diagnoses produced by the program involves another set of issues. Program diagnoses have a very different character than expert diagnoses. Program diagnoses are very detailed and contain a small number of completely specified hypotheses. Expert diagnoses include intentional ambiguity and only specify the primary aspects of the problem. To compare such differing diagnoses, we have developed a program that assesses whether the program diagnosis is consistent with the expert diagnosis. It is still necessary to validate the form of presentation used by the program, since these are important justifications of the diagnosis that are not usually conveyed by the expert diagnosis.

1. INTRODUCTION

This paper addresses some of the issues of validation of a large knowledge based medical system operating in the context of complex medical cases using the same data as the physicians and providing information intended to be integrated into the physician's decision making process. The medical context is the diagnosis of patients with symptoms suggestive of heart failure. That is, patients with shortness of breath, fatigue, edema of the extremities, peripheral vasoconstriction or other findings that are consistent with hemodynamic dysfunction, or patients with known cardiac disease who are at risk for low cardiac output and the sequelae of compensatory mechanisms that produce the clinical presentation of heart failure. Since there are many diseases other than those involving the heart itself which have these consequences, the medical domain of disorders is quite broad. Since the underlying diseases that cause heart failure are usually chronic, many cases

*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.

involve repeated hospitalizations with progression of the disease, complications, effects of therapies, or additional unrelated diseases.

Over the past several years we have been developing the Heart Failure Program to assist physicians in reasoning about such patients. The program takes the same kind of case description a physician would record about the patient, including information about the history, symptoms, physical examination, and test results. It uses that information to generate a differential diagnosis consisting of hypotheses, each of which explains all of the findings, except those better explained by something outside of the domain. The program can also suggest additional measurements to refine the diagnosis and therapies to manage the problem. It can also predict the hemodynamic effects of the therapies, but this paper will be restricted to the validation issues that arise in the differential diagnosis process (see other papers about other aspects of the system[1, 2, 3]).

In the following sections we will outline the basic structure of the program, highlighting the aspects that need validation; the issues and methods involved in validating the knowledge base; and the issues and experience of assessing the correctness of the diagnoses.

2. DIAGNOSTIC REASONING MECHANISM

The Heart Failure Program (HFP) is a computer system which assists the physician by computing differential diagnoses for cases from the findings, represented as detailed graphical physiologic causal diagrams justifying each of the hypotheses. There are three major parts of the program involved in generating differential diagnoses: 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 heuristic hypothesis generator 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.

2.1. Input Interface

The input interface is a dynamically expanding menu with fields for specifying symptoms, known diseases, current therapies, detailed physical examination findings including vital signs, and the pertinent results of laboratory tests, both usual and unusual. The intent is to capture the information pertinent to the diagnosis and immediate management of cardiovascular disease without requiring the system to do reasoning outside of the problem 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. The nature of this interface means that decisions have to be made in the design of the program about what information is potentially pertinent to the diagnosis and at what level of detail. Obviously, anything the user cannot enter, the program cannot use in diagnosis. One might argue that the input should be as extensive as possible in hopes that someday the program could be expanded to use the information. However, that gives the users a false sense of security, believing that the information entered is all used in generating diagnoses.

2.2. Knowledge Base

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 qualitative physiologic *states*, *measurement categories* (the categories of patient information), and probabilistic causal *generic links*, constituting the general diagnostic knowledge about the domain and used as a general template from which the more specific knowledge about a case is generated. Using the information from a case, the states, measurement categories, and generic links are instantiated as *nodes*, *findings*, and specific *links* 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 laboratory results. The links model the causality as probability relations, both between states and from states to the values of 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. An important feature of the KB that influences the validation is the lengthy causal chains in the model. While a few diagnostic systems like MUNIN[4] have causal chains of four or five levels, those in the HFP can be a dozen or more nodes long. An important disease such as myocardial infarction (heart attack) includes the etiology of coronary artery disease, coronary obstruction, leading to the infarction, which causes ventricular dysfunction, and a whole chain of consequences leading to signs of pulmonary congestion or other effects of heart failure produced by the infarction.

There are two features of this knowledge representation that simplify the reasoning mechanisms 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 cannot 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. For example, high heart rate can cause low cardiac output, but only when the heart rate is very high, say above 120. This is captured by making the probability on that link a function of the actual heart rate. Probabilities that are functions of input values are resolved when the input is entered. Probabilities that are functions of nodes must be handled by the process of hypothesis generation.

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, there are nodes for both acute mitral regurgitation and chronic mitral regurgitation to distinguish the different presentations.

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, valvular disease, 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 probabilities on the links between nodes are combined using a "noisy-or" combination rule[5] 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:

$$(1 - \prod_{i \in P,W} (1 - p_i)) \prod_{i \in C} (1 - p_i)$$

*Levels of severity and time intervals for states have been recently added, but await further testing before they can be adequately discussed.

Similarly, each finding has a probability of being produced by nodes. The model is similar to those investigated by Pearl[5] 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). The forward loops imply that the product of the probabilities of the nodes computed locally from the states of immediate predecessor nodes is not a consistent interpretation of probabilities. That is, the probabilities of all combinations of node states computed in this manner, does not necessarily sum to one. However, a consistent interpretation is possible by eliminating loops in specific hypotheses. Because of these complications, heuristic methods are necessary for generating hypotheses or estimating the probability of a node and the program is pseudo-Bayesian.

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. The algorithm is described elsewhere[2], but may be thought of as a heuristic approximation to finding the maximum likelihood explanation for a set of findings.

3. KNOWLEDGE BASE VALIDATION

There are a number of aspects to the problem of validating the knowledge base for a large medical diagnostic system. These include defining the limits of the domain, specifying the allowed input vocabulary, specifying the causal structure, and validating the probabilities and other constraints on the linkages.

Medical domains are difficult to define cleanly. Since the human body is a highly integrated functioning unit, it is hard to define the limits of a particular class of diseases. Other diseases may complicate or mimic the diseases of the desired domain. Since the patient presents with a set of symptoms, the criteria for inclusion or exclusion must be decidable in terms of the presenting symptoms, or by some simple test. Furthermore, the criteria (and the program) cannot exclude the complexities of the difficult cases if it is going to be of value to the target audience. Because of these difficulties, the inclusion criteria for the HFP have remained loosely defined as: *adults with symptoms resembling heart failure, potential heart failure, or its complications.* That definition is not as circular as it sounds because the low output and congestive symptoms present a recognizable cluster of findings. On the other hand, the looseness of the definition means that it is important that the program handle sets of findings that could conceivably be mistaken for heart failure at least by identifying the correct medical domain.

Besides identifying the range of the medical domain, it is important to define the depth. For the HFP the focus is the short-term diagnosis and management of hemodynamic compromise. This implies that details of the target diseases beyond those that determine the hemodynamic compromise may be irrelevant to the diagnostic problem. For example, the etiology may not be important to the diagnostic problem once the presence of the disease has been determined. In the heart failure domain, the problems are:

1. To what extent should etiologies be included in the KB?

2. What diseases complicate the heart failure and how much detail is relevant in representing them?

3. What cardiac diseases should be excluded because they have negligible effect on the cardiac hemodynamics?

4. What cardiac diseases should be excluded because they are rarely encountered?

3.1. Disease Etiologies

The problem of deciding on the etiological detail to include in the HFP is an ongoing struggle. The problem is that while etiology may not affect the management of the patient, it does affect the likelihood of coexisting diseases and therefore the ability of the program to produce correct diagnoses. For example, once one has aortic stenosis, the hemodynamic effects and therapy are independent of the etiology. However, the probability that the patient also has mitral stenosis would be high if the etiology were rheumatic but low if the etiology were a degenerative process. One might include rheumatic heart disease in the model to account for this difference. However, the same argument can be made in the case of cardiomyopathy if the cause is alcohol consumption, because then the probability of liver disease is greatly increased which in turn can account for some of the same findings as cardiac diseases. From our experience with cases, it is clear that using etiology, when it is known, to adjust the probabilities of other diseases, is important. For the physician it may represent the difference between requiring ample evidence for consideration of a disease and including a disease as a strong possibility unless it is ruled out. On the other hand, it is not reasonable to pursue an etiology if the only evidence is the cardiac lesion it causes. In particular, if the program adds to aortic stenosis an etiology of rheumatic heart disease because, given the age, that is the most likely etiology, cardiologists find it unjustifiable. The only exception is mitral stenosis, for which almost all cases are due to rheumatic heart disease. For these reasons, we have included etiologies in the KB as factors that change the probabilities if known, but are not nodes themselves. For example, there is a node for *valvular heart disease*, which causes specific valvular lesions. The probability that it causes particular lesions is adjusted by input findings of history of rheumatic fever, Marfans, calcification, history of endocarditis, or drug abuse.

3.2. Complications from Other Diseases

There are several classes of diseases that have a significant impact on heart failure, including pulmonary, renal, and hepatic disease. Each of these is a domain in which a knowledge based system the size of HFP could be developed. Since the purpose of the program is to diagnose heart failure, we have restricted the knowledge about diseases of the other organ systems to those aspects that influence the heart failure. For example, there is a single node representing the various kinds of primary liver disease. The effect of this node in the model is to cause ascites, hepatomegaly, pedal edema, hypoalbuminemia and elevated liver function tests (which are only considered in total). This lumps very different liver diseases together, but to first approximation it is sufficient to account for the possibly different presentation of a patient who has both liver disease and heart failure. Similarly, renal disease is only differentiated into acute and chronic, although we found it necessary to adjust probabilities when there is evidence of nephrotic syndrome because that disease can account for pedal edema without the degree of elevation of BUN and creatinine associated with other forms of renal insufficiency severe enough to cause pedal edema. Without that differentiation and the common occurrence of renal insufficiency, the program would often use known mild renal insufficiency to account for pedal edema that was more likely the result of heart failure. Pulmonary disease is closely tied to cardiac disease, but have only a few patterns of hemodynamic effect. As a result, chronic bronchitis and chronic obstructive pulmonary disease are treated as a single node (with an unwieldy name), while primary pulmonary hypertension and pulmonary embolism are separate nodes because of their direct influence on cardiac hemodynamics. Thus, each of the complicating groups of diseases has been represented by a small number of nodes sufficient to represent their type of impact on the presentation of heart failure.

3.3. Hemodynamic Compromise of Cardiac Diseases

Not all cardiac diseases have significant effects on the hemodynamics. In particular, there are a number of different kinds of arrhythmias defined by the type and level of electrical disturbance in the heart. The degree of hemodynamic compromise is more dependent on severity than on type of disturbance. As a result, we have a node representing *high degree AV block* intended to cover complete block, Mobitz II block, and Wenckebach block greater than 2:1. Lesser degrees of block are only used as evidence for the other diseases that often cause them.

Because there are a large number of very rare cardiovascular diseases that can cause heart failure, it was clear from the outset that some would have to be left out of the model. Most of these are similar to diseases already in the model from a hemodynamic perspective, except for a few features. For example, left atrial myxoma, a tumor in the left atrium, has the same hemodynamic effects as mitral stenosis except that it has much more rapid onset and can cause intermittent symptoms. Another example is the combined pulmonic regurgitation and pulmonic stenosis resulting from the correction of tetralogy of fallot in childhood. As isolated lesions, both of these are unusual and together they are extremely rare. Because of the hemodynamic similarity of the rare diseases to more common ones, we have chosen to leave them out of the model. Instead, we have made sure that the KB includes at least one disease covering each type of hemodynamic compromise with the intent that missing diseases will instead generate diagnoses of the similar hemodynamic disturbance.

A related issue is what therapies to include in the KB. The purpose of including the therapies is to enable the HFP to account both for missing disease findings from successful therapy and for indications of toxicity. For example, furosemide (a diuretic) can cause hypokalemia and can account for the lack of pedal edema in a patient with right sided heart failure. It is sufficient to include one drug from each class to accomplish this, since to first approximation, the hemodynamic effects of other drugs in the same class are the same. (For example, propranolol covers the hemodynamic effects of all the beta adrenergic blocking agents.) The only exception we have encountered is the calcium channel blockers, which had to be individually represented because of the large differences in effects between individual agents.

4. VALIDATION OF PROBABILITIES

Since the KB consists of causal relations connecting pathophysiologic states by probabilities and conditions of causation, one problem is how to find the probabilities. Medical literature does not usually address questions such as how often a particular state causes a particular finding, beyond occasional qualitative statements such as *common* or *rare*. On the other hand, experienced cardiologists have little trouble giving approximate frequencies for these findings. From our experience with the program it is most important that these lists of findings be complete. If a finding does not have links to all of its possible causes, the program may have to add an unjustified disease to a hypothesis or even be directed in the wrong direction. For example, because pericarditis did not include high white blood count as a finding, the program added pneumonia to its diagnosis in a pericarditis case. In most cases there are a few findings that point strongly at the correct causes, so the exact probabilities are less important. That is, the association of diseases and findings presents such stark diagnostic alternatives that Occam's razor would be a sufficient guide. However, there are still a significant number of cases in which the same findings can be accounted for by different causes, with no certain way of distinguishing. Even in cases where the main part of the diagnosis is clear, some part of the mechanism or attribution of findings to multiple causes may be ambiguous.

For these cases, the probabilities do make a difference and need to be validated. The problem is determining how the probabilities elicited from experts might be biased. They may reflect memorable cases, recent cases, or may be a mixture of importance along with the actual frequency of the events. To assess and refine the probabilities in the KB, we have used several approaches:

1. Computation and assessment of evidential probabilities
2. Assessment of disease-to-finding probabilities
3. Comparison of frequencies with disease data bases
4. Analysis of cases in which diagnoses were faulty

4.1. Assessing Probabilities of Disease Given Evidence

The first method was to compute the distribution of causes for each effect. The information provided to the KB is the probability of an effect given a cause and the prior probabilities of the primary causes. This information along with Bayes' formula is sufficient, in principle, to compute the inverse probabilities. That is, for finding f and cause c, the relative frequency of c as a cause for f is:

$$p(c|f) = \frac{p(f|c)p(c)}{\sum_i p(f|c_i)p(c_i)}$$

where the c_i are all of the possible causes for f. The complexity of the actual KB structure means that these calculations are also estimations. Since the KB has multiple links in the causal chains from primary cause to ultimate findings, this inverse computation can be done for the immediate causes of the findings, the primary causes, or some important intermediate node in the network. For example, it is possible to compute the expected frequency of hypertrophy versus dilatation as the cause for an enlarged cardiac silhouette on X-ray or to compare congestive cardiomyopathy, chronic mitral regurgitation, chronic hypertension, aortic regurgitation, and aortic stenosis as potential disease causes of the finding. In the first case, since cardiac hypertrophy is an intermediate node, its prior probability is estimated by combining the probabilities along the pathways from its possible causes and the probability of an enlarged cardiac silhouette given hypertrophy is taken from the link. In the second case, the priors for the diseases are part of the KB and the probability of an enlarged cardiac silhouette given aortic regurgitation (for example) is computed from the causal pathways that connect them. By reviewing these lists it was possible to identify probabilities that were out of line. For example, if most of the findings of a disease seemed to have too high a fraction of their causes indicated as that disease, the probability of the disease was too high. If only one or two findings of the disease seemed out of line, the problem was in the link probability. This approach provides a different perspective on the probabilities, but one with which physicians are quite comfortable.

4.2. Assessing Spanning Probabilities

The second method was to examine the expected ultimate findings of diseases. Because there are many intermediate physiologic states, the path from a disease to some of the findings can be lengthy. By determining all expected effects of a disease with significant probability, the combinations of probabilities through the causal mechanisms were validated. The probabilities were estimated as:

$$\max_r \prod_{i=f}^{c} p(n_i|n_{i+1})$$

where the r are the possible paths from cause c to finding f. The maximum was used to simplify the calculations, simplify the explanations to the reviewers, and because particular hypotheses will normally have only one mechanism in effect.

4.3. Comparison of Probabilities to Data

If the probabilities in the KB are to reflect the actual frequencies of the corresponding events in patients, the logical way to validate the probabilities is with statistical data. We have used two sources of patient data to conduct this kind of testing. The easiest comparison is against the case base we have built over the years in the development of this program. At present we have about 400 cases that have been entered, run, and analyzed. The collection process has emphasized difficult cases and has tried to achieve as broad a coverage as possible. Thus, this is not a good source for estimating prior probabilities of disease or the frequency of various complications. However, it is an appropriate source for estimating the probabilities on links for frequently occurring nodes. For example, since all patients with aortic stenosis were included it will provide estimates of the frequencies of the effects of aortic stenosis.

We also have access to a data base of 5773 cases of patients with suspected acute cardiac ischemia[6]. These include all patients admitted in the emergency rooms of six hospitals over a two year period with chest pain or new onset of shortness of breath. About a hundred parameters were collected on these patients with an emphasis on the characteristics of the chest pain and shortness of breath. This has been a source for some information about the frequencies of different events, but it is somewhat limited. One of the problems with a data base like this is that the variables of concern in heart failure are not the most important to consider in the acute presentation of possible ischemia. As a result, there are a large number of missing values which may represent either absence of the finding or that the finding was not checked. A second problem with this or any data base collected in an environment different from that in which it is being used is the correspondence between the variables collected and the findings in the KB. When the variables are somewhat different, the applicability of the data is strained. For example, the HFP has an input finding of *syncope or near syncope* while patients in the acute cardiac ischemia study were asked about *dizziness.* These are closely related findings but not exactly the same.

4.4. Analysis of Failed Diagnoses

The most effective way of validating the probabilities is by analyzing faulty hypotheses produced by the HFP and determining the reason for the errors. This can be accomplished by constructing an appropriate hypothesis and comparing it to the generated hypothesis by computing the overall probability of each. If the computed probability of the correct hypothesis is greater than that of the generated hypothesis, the probabilities appropriately distinguish the hypotheses and the problem is that the heuristics used to generate the differential did not find the highest probability hypothesis. If the correct hypothesis had a lower computed probability than the generated hypothesis, something is wrong with the probabilities on the causal links that contributed to the computed probabilities. A problem with the probabilities can be analyzed by looking at the difference set between the two hypotheses. Even though the differences between the correct and faulty hypotheses are usually small, there are usually several probabilities that could correct the problem. The lists of possible corrections are collected from several faulty hypotheses and the multiple occurrence of the same possible corrections are used to guide the changes to the probabilities. When there are no other errors that help to disambiguate the possible corrections, questions can be formulated for the experts that will guide the changes, such as considering hypothetical patients distinguished by the findings of concern.

5. FORMATIVE EVALUATION

Over the past year we have conducted a formative evaluation of the program (reported in[7]). This process exposed a number of difficult issues not often discussed in the evaluation literature. The two aspects of the evaluation in which issues had to be resolved were 1) the collection of patient data and 2) the comparison of differentials.

5.1. Collection of Patient Data

The case material used in an evaluation is selected from the available patient population, some time point is selected for analysis, and the data is summarized for the program and for the reviewers. Each of these parts of the process can introduce biases in the evaluation that must be carefully considered in the design of the study.

For any diagnostic system covering a significant number of diseases, it will be difficult to find examples of all of the diseases to include in the evaluation because of the relative infrequent occurrence of many diseases. This problem is compounded in a domain, such as heart failure, where many of the diseases are chronic and already known when the patient is admitted. Thus, for a disease like aortic stenosis, there are several distinct diagnostic situations:

1. The disease is not known and needs to be diagnosed.
2. The disease is known and has become worse.
3. The disease is known but the patient is now suffering from a complication or independent disease.
4. The disease has been corrected but residual effects remain.
5. Although the disease has been corrected, it has recurred.

Each of these presents a distinct diagnostic challenge and greatly multiplies the size of the domain. For example, any of these situations can exist in the course of the natural history of aortic stenosis. Initially, the problem is to diagnose the valve disorder. Later, the patient may return with overt signs of failure from the increasing stenosis. The patient may also be admitted with pneumonia in the context of some degree of aortic stenosis. If the valve is replaced, the patient will still residual left ventricular hypertrophy (LVH). Finally, a prosthetic valve itself may become restricted, again producing symptoms of aortic stenosis.

In the HFP evaluation, we collected 242 cases from discharge summaries, taking all available cases with DRGs (disease classifications) that would indicate complicated disease. Even so, we were only able to collect examples of 21 of the 30 primary diseases in the KB in which the disease was not already known at the time of diagnosis and examples of 19 of the 30 in which the disease was already known but the patient had further complications or additional diseases. Eight of the primary diseases are not represented in the study, although we have since found examples of most to use for testing the program.

Since diagnosis is an evolving process, the time point or points selected for diagnosis influence the kind of diagnosis that is possible or appropriate. In the emergency room, one can focus on the presenting complaint, physical findings, and an electrocardiogram to decide questions such as the likelihood of acute ischemia. For the HFP, the appropriate time for diagnosis is after the initial laboratory findings are available. Thus, the program can make use of chest X-ray findings and blood test results. The program could be used later as well, but the information provided by the program would be less useful when the diagnostic task is almost complete, although it would be useful in identifying findings that might not be consistent with the current diagnostic hypothesis.

The final issue is what information the program and reviewers receive about the case. The program is restricted to the particular vocabulary of its input routines, which is never exactly the same as would appear on a chart or physician summary of the case. Thus, there is translation or summarization that must take place to enter the data into the program. The issue of variability in this summarization task has been addressed in other domains[8], but not the correspondence between that data and the information available to the reviewers. There are three levels of information one could give the reviewer, depending on the objectives of the evaluation. 1) The reviewers could be given the same information as the computer. This was our approach with the HFP. In fact, the reviewers (the cardiologists on the project) used the textual summary generated by the program from its input. In this way it is possible to evaluate whether the program is drawing the appropriate conclusions from the information it is given without the confounding possibility that the information may have been inappropriately abstracted. 2) The reviewers could be given a short, hand generated description of the case. This would be appropriate for comparing the program to the performance of an independent consultant not present to examine the patient. 3) In a study comparing the performance of the program to a physician, the physician needs to see the patient. The reviewer could see the patient, but a more practical approach is just to use the diagnosis of the primary care provider for the comparison. The problem with this approach is that there is no way to tell whether the differences between the program and the physician are the result of the program, findings that could have been entered but were not, findings for which the input vocabulary was inadequate, or impressions of the patient not easily verbalized. We considered using the diagnoses from the discharge summaries in the HFP study, but those diagnoses had the additional problems that they did not reflect the same time point as the information given to the program, were given in widely varying degrees of detail, and often were not adequately supported by the data available in the summary.

5.2. Comparison of Differentials

Once the patient data has been given to the program and reviewers and each has produced a differential diagnosis, the problem is to compare the differentials. The solution most often adopted is to require the reviewers to use the same diagnosis vocabulary as the program. Then the problem is only to assess the degree of match between a pair of ordered lists. This approach is fine if the diagnoses are adequately expressed by a choice from a fixed list. When the diagnosis is stated as a complex structure, the problem is more difficult. In the HFP the individual diagnostic hypotheses that make up the differential are causal graphs of pathophysiologic states and findings. It would be unreasonable for us to expect the reviewers to construct anything comparable. Therefore, we asked the reviewers to give their diagnoses in the terms that seemed natural. As a result, we had to confront a more general set of issues including the meaning of the terms used by the reviewers, how to match the diagnoses to the causal graphs, the interpretation of a differential, and the nature of the reviewers' diagnoses as a standard.

A typical expert diagnosis consists of terms of varying degrees of specificity, with alternatives, and some indications of priority or uncertainty. An example from the formative evaluation is:

> Either an acute myocardial infarction or unstable angina with previous evidence of coronary artery disease in the form of an old myocardial infarction; atrial flutter; mild left heart failure; and possible aortic stenosis.

Some of these terms have a direct correspondence with the names of states in the HFP hypothesis, such as aortic stenosis. Others do not correspond to a single node but cover a set of nodes. Indeed, even those with a direct correspondence actually cover the typically occurring consequences of that node. For example, the aortic stenosis diagnosis does not include LVH because that would be expected. Each of the terms in the diagnosis has a set of nodes it could acceptably cover. Some of the modifiers make this more explicit. For example, compensated heart failure (which appears in a number of diagnoses) includes the nodes normally associated with heart failure except for those that indicate significant pulmonary or systemic congestion. To accommodate this use of terms, we wrote a program for matching terms to the appropriate causal structure.

Since the diagnoses of the reviewers and the causal structures in the program's hypotheses each have several items in them, the next question is how close the match needs to be. The HFP designates 40 nodes as *diagnostic*, implying that they make important distinctions in the hypothesis. These also correspond to many of the terms used by the reviewers in stating the diagnoses. They were used as the essential characteristics of the diagnosis. To match, a hypothesis had to have all of the diagnostic nodes implied by the reviewer's diagnosis and no more. There were two exceptions to this rule. 1) If a node was directly implied by the input (*e.g.*, a known diagnosis), the reviewer did not have to state it. 2) If the reviewer did not state the etiology of heart failure or said it was unknown, we allowed the program to pick a common etiology to accommodate the hypothesis generation mechanism, which always generates a complete hypothesis.

The uncertainty in the reviewers' diagnoses represented their differential nature. That is, by stating that the diagnosis was myocardial infarction or unstable angina, the physician is including both in the differential. Stating that a disease is possible includes hypotheses with and without the disease. This is a very economical way of giving a differential because it focuses directly on the uncertainty without actually stating each complete possibility. To compare the diagnoses to the hypotheses in the computer generated differential, we expanded the implied complete diagnoses. The number of possibilities varied, but was four or less in 88% of the cases. Sometimes the reviewers rank ordered the alternatives, but only infrequently, so we treated the differentials as unordered. The reviewers could answer questions about which alternatives were most likely, but that was not an important objective of diagnosis since the differential was sufficient to direct further investigation by which the questions could be answered definitively.

Given the differential implied by the reviewers' diagnosis, the next question is to determine how many the program should match. This is a question the HFP was not ready to address. Since differences in hypotheses were defined by the different terms used in the diagnoses, many of the hypotheses included in the computer differential would not be considered different. Furthermore, the program used local optimization techniques to improve its hypotheses, which also had the effect of eliminating alternatives and hence some of the other hypotheses that should have been in the differential. Still, the question of what should be in a differential is important. From the use of the terms by the physicians, a working definition is: any diagnosis description that requires different terms to describe, has relatively high probability, has positive evidence, and is not ruled out. The terms determine what is different. While some of the terms are more specific than others, the more specific terms are only used when there is evidence to support a more specific diagnosis, not just a higher prior probability for one variant of the general diagnosis than others. The fact that diagnoses are ruled out by evidence and not just probabilities (within limits) is something that systems based on probability networks will have to address. (Similar observations have been made in other domains[9].)

For the HFP, experienced cardiologists are the best available standard for comparison of the diagnoses. However, with complicated differentials there is clearly room for adjustment of the diagnoses. Our procedure for evaluating the computer diagnoses was to review as many as possible of the leading hypotheses that did not match any of the diagnostic possibilities in the reviewers' diagnoses. In the process, 61 of the diagnoses were changed in some way, usually relatively minor. Most of these changes were to add a disease that was overlooked before. For example, adding possible chronic obstructive pulmonary disease (COPD) for which there was some evidence. From a medical standpoint, the changes were not significant, mainly because they involved chronic stable diseases that were not part of the acute problem. Acuteness and impact on management are clearly important parts of focusing the diagnosis and should be taken into account both in generating differentials and in evaluating the results.

6. CONCLUSION

The task of validation is fundamental to the development of diagnostic decision aids, but it is far from straightforward. In this paper we have addressed a number of issues which arose in the context of the Heart Failure Program, but are of concern in many similar systems.

7. REFERENCES

1 W. J. Long, S. Naimi, M. G. Criscitiello, and G. Larsen, Differential Diagnosis Generation from a Causal Network with Probabilities, in: *1988 Computers in Cardiology Conf.* (1988) 185-188.

2 W. Long, Medical Diagnosis Using a Probabilistic Causal Network, *Applied Artificial Intelligence* 3 (1989) 367-383.

3 W. J. Long, Flexible Reasoning about Patient Management using Multiple Models, *Artificial Intelligence in Medicine* 3 (1991) 3-20.

4 K. G. Olesen, U. Kjaerulff, et al., A MUNIN Network for the Median Nerve — A Case Study on Loops, in W. Horn, *Causal AI Models: Steps Toward Application* Hemisphere Publishing Company, New York (1990) 301-319.

5 J. Pearl, Fusion, Propagation, and Structuring in Bayesian Networks, *Artificial Intelligence* 29 (1986) 241-288.

6 M. W. Pozen, R. B. D'Agostino, et al., A Predictive Instrument to Improve Coronary-Care-Unit Admission Practices in Acute Ischemic Heart Disease, *New England Journal of Medicine* 310 (1984) 1273-1278.

7 W. J. Long, S. Naimi, and M. G. Criscitiello, Development of a Knowledge Base for Diagnostic Reasoning in Cardiology, to appear in *Computers in Biomedical Research.*

8 R. A. Bankowitz, B. H. Blumenfeld, N. B. Giuse, et al., User Variability in Abstracting and Entering Printed Case Histories with Quick Medical Reference, in: *Symposium on Computer Applications in Medical Care Conference,* (1987) 68-73.

9 S. Tuhrim, J. Reggia, and S. Goodall, An experimental study of criteria for hypothesis plausibility, *J. Expt. Theor. Artif. Intell.* 3 (1991) 129-144.