A Comparative Study of Computer-Aided Clinical Diagnosis of Birth Defects

by

Howard Bruce Sherman

Submitted to the Department of Electrical Engineering and Computer Science on January 21, 1981 in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science

ABSTRACT

In recent years many computer systems have been developed to assist in medical decision making. Two of these systems in particular, INTERNIST and the Present Illness Program (PIP), have been proposed as suitable for performing general medical diagnosis. However, there has been no way of comparing the performance of these two programs since the medical data used by the programs differs extensively.

In order to make such a comparison versions of both systems have been implemented, and the medical data used by each has been abstracted from a single data base in the domain of birth defects. Although both systems use a common paradigm of constructing diagnostic hypotheses and then testing those hypotheses by suggesting further tests, variations in their implementation of this paradigm result in significant differences in performance. A detailed analysis of the strengths and weaknesses of these two approaches to computer-aided medical diagnosis, in the diagnosis of thirty-five clinical cases drawn from the congenital defects domain, is presented. The results of this analysis are used to generate suggestions for the improvement of such programs.

Thesis Supervisor: Dr. Peter Szolovits

Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank all those who made this thesis possible, and in particular:

Peter Szolovits, my advisor, for his generous support and guidance throughout the course of this work;

Ramesh Patil for his thought-provoking conversations and his invaluable suggestions;

Dr. Marylou Buyse for her advice on medical aspects of the thesis;

Seymour Small for his time and effort in explaining the BDDS system and for helping in the transfer of the BDDS database;

Harold Goldberger for careful reading and substantive suggestions on later drafts of this thesis;

Dr. Stephen Pauker for helping me get started on this project;

Bill Long and Ken Church for their useful comments on early drafts of this thesis;

Glenn Burke for programming assistance;

All of my colleagues in the clinical decision-making group and knowledge-based systems group for fine discussions and suggestions;

and, finally, Lynn Cococcia and my parents for moral support.

CONTENTS

1. INTRODUCTION 8

1.1 Hypothesis Driven Diagnosis 9
1.2 The Comparative Study 10

2. INTERNIST 12

2.1 Representation of Knowledge 12
2.2 INTERNIST's Algorithm 15
2.2.1 Hypothesis Generation 16
2.2.2 Scoring Algorithm 18
2.2.3 Strategy and Question Selection 20
2.2.4 Concluding Diseases 22
2.3 Summary 23

3. Present Illness Program 24

3.1 Representation of Knowledge 24
3.2 PIP's Algorithm 26
3.2.1 Hypothesis Generation 26
3.2.2 Scoring Algorithm 29
3.2.3 Strategy & Question Selection 30
3.2.4 Concluding Hypotheses 31
3.2.5 Summary of PIP and INTERNIST 33

4. The Birth Defects Database 35

4.1 The Domain 35
4.2 The Database 36

5. The Comparison of PIP and INTERNIST 38

5.1 Hypothesis Generation 38
5.1.1 The Overall Performance 39
5.1.2 Deactivation of Hypotheses 45
5.1.3 The Disease Hierarchy and Hypothesis Generation 50
5.1.4 The Separation of Scoring and Triggering Information 51
5.1.5 Summary 59
5.2 Scoring 59
5.3 Concluding Hypotheses 63
5.4 Diagnostic Strategy and Question Selection 65
5.4.1 The Discrimination Strategy 72
5.5 Summary 75

6. Suggestions for Improvements 76

6.1 Representation of Knowledge 77
6.1.1 Decomposition of Medical Knowledge 79
6.1.2 Representing Clinical Situations 81
6.1.3 Representation of Non-Medical Knowledge 81
6.1.4 Representation of Time 82
6.2 The Algorithm 83
6.2.1 Hypothesis Generation 83
6.2.2 Scoring 86
6.2.3 Diagnostic Strategy and Question Selection 87
6.2.4 Concluding Hypotheses 89
6.2.5 Multiple Syndromes 90
6.2.6 Exception Handling 92
6.3 Summary 92

7. Conclusions and Further Research 94

7.1 Summary 94
7.2 Further Research 95

Appendix I. List of Syndromes 97

Appendix II. List of Clinical Cases 99

Appendix III. The Design for an Improved System 100

III.3 Representation of Knowledge 101
III.3.1 The Frame Representation Language 102
III.3.2 Syndromes, States and Findings 103
III.3.3 Statements 107
III.3.4 Representation of Findings 108
III.3.5 Representation of Items 110
III.3.6 Representation of Time 111
III.4 The CDDS Algorithm 113
III.4.1 Hypothesis Generation 114
III.4.2 Scoring 116
III.4.3 Diagnostic Strategy and Question Selection 119
III.4.4 Multiple Syndromes 121
III.4.5 Concluding Hypotheses 123
III.4.6 Exception Handling 124
III.5 Implementation 125
III.6 Summary of CDDS 125

Appendix IV. The Performance of the Congenital Defects Diagnostic System .. 126

IV.7 Hypothesis Generation 126
IV.8 The Scoring Algorithm 128
IV.9 Concluding Syndromes 130
IV.10 Diagnostic Strategy and Question Selection 130
IV.11 Specialized Features 133
IV.12 Summary 134

References 135

FIGURES

Fig. 1. INTERNIST's Disease Hierarchy 13
Fig. 2. INTERNIST's Scoring Algorithm 19
Fig. 3. Typical Disease Frame in PIP 25
Fig. 4. Number of Hypotheses Generated after Entry of Initial Findings 43
Fig. 5. Number of Inappropriate Hypotheses After Entry of Initial Findings 44
Fig. 6. Number of Hypotheses at End of Session With Deactivation 48
Fig. 7. Number of Inappropriate Hypotheses at End of Session With Deactivation 49
Fig. 8. Nonterminal Hypotheses Generation by INTERNIST after Initial Findings 53
Fig. 9. Number of Hypotheses Generated 57
Fig. 10. Number of Inappropriate Hypotheses Generated 58
Fig. 11. Number of Questions Required to Conclude the Correct Hypothesis 67
Fig. 12. Number of Questions Required to Pursue the Correct Hypothesis 70
Fig. 13. Number of Questions Required to Pursue the Correct Hypothesis 71
Fig. 14. INTERNIST With and Without the Discriminate Strategy 74
Fig. 15. A Typical Frame 102
Fig. 16. Typical Syndrome or Clinical State Frame 104
Fig. 17. Typical Statement 108
Fig. 18. Typical Finding Frame 109
Fig. 19. Typical CDDS Item Frame 111

TABLES

Table I. Number of Hypotheses Generated 42
Table II. Number of Hypotheses Generated by PIP With and Without Deactivation 47
Table III. Nonterminal Hypotheses Generation by INTERNIST 52
Table IV. Hypotheses Generated by INTERNIST After Entering Initial Findings .. 55
Table V. Number of Hypotheses Generated 56
Table VI. The Rank of the Correct Hypothesis after Initial Findings Entered 61
Table VII. Number of Questions Required to Conclude the Correct Hypothesis .. 66
Table VIII. Number of Questions Required to Pursue the Correct Hypothesis 69
Table IX. INTERNIST With and Without the Discriminate Strategy 73

Table X. Number of Hypotheses Generated 127
Table XI. Rank of Correct Hypothesis after Initial Findings Entered 129
Table XII. Number of Questions Required to Conclude the Correct Hypothesis 131
Table XIII. Questions Required to Pursue Correct Hypotheses 133

1. INTRODUCTION

In recent years there has been much interest in the development of computer aids for medical diagnosis and management. Interest in these systems has been in response to the emergent needs of medicine as it becomes an increasingly broad field of knowledge. These systems hold the promise of employment within the ranks of practicing physicians, specialists and general practitioners alike. In addition these systems also present possibilities in myriad facets of medicine [48].

Currently, for example, the training of physicians stresses experiential learning in their development as competent practitioners. Computer based systems could provide students an opportunity to engage in a dialogue with the program to learn varied diagnostic and therapeutic skills not easily acquired from textbooks.

The development and employment of such systems could affect the quality of health care available to those living within inner cities and rural areas which have difficulty attracting and keeping adequate numbers of physicians. These examples of possible employment are but two of the many potential uses for these systems.

Toward these ends, and others which have been delineated in many papers [39,40,48,52], many systems have been created. These systems have been addressed to several areas of medical concern. Among these systems are the Present Illness Program [27,28,44], INTERNIST [29,30,31,32], CASNET/GLAUCOMA [51,52], MYCIN [6,7,38,39], and others. But, so far, none of the systems that have been developed have been perfected and tested to the point of being a releasable product.

1.1 Hypothesis Driven Diagnosis

Among the systems that have been built, several, which bear certain similarities, are often described as hypothesis driven [24]. This would include some of the Bayesian diagnostic systems [15] as well as systems using the artificial intelligence approach such as the Present Illness Program and INTERNIST. Although there is no precise accepted definition of hypothesis driven systems, for this thesis hypothesis driven diagnostic systems are those which, given a set of findings, select a small set of hypothesized diagnoses which are used to guide the search for the correct diagnosis by controlling the search for further evidence. This is often a useful approach when searching a large state space with many possible transitions between states.

Like Bayesian diagnostic systems, hypothesis driven systems use the association between findings and diseases to identify the likelihood of a disease being present. To determine this likelihood a matching (scoring) algorithm is required. However, unlike strict Bayesian systems, the hypothesis driven diagnosis systems apply the scoring algorithm to only the currently active hypotheses. The algorithms used for this matching are often Bayesian in nature but do not strictly adhere to Bayes' rule. So, given a set of findings, these systems estimate the likelihood of diseases being present.

The systems then use these likelihoods to select a strategy from which to proceed. This strategy is used to select an unknown finding(s) to ask the user. The matching algorithm can then be used again on the enlarged set of findings. This process continues until the system has found a satisfactory match between findings and diseases.

Although there is nothing that would, in theory, prohibit a hypothesis driven diagnostic system from reasoning in great depth about the physiological and anatomical mechanisms behind the evidence in the case, in most systems this is not done. Instead,

pattern matching between the findings and the diseases via the scoring algorithm is the major method for diagnosis. The systems that do not engage in physiological and anatomical reasoning require much less knowledge than those systems that do and hence are smaller and easier to construct. But if the domain is such that the interactions between findings affect the associations between findings and diseases, then physiological and anatomical knowledge about the findings may be necessary for diagnosis.

This thesis examines the above type systems; hypothesis driven medical diagnostic systems using a pseudo-Bayesian matching algorithm between the diseases and the findings without extensive reasoning about the physiological and anatomical mechanisms involved.

1.2 The Comparative Study

Many hypothesis driven medical diagnostic systems have been built. They bear certain similarities and certain distinctive features. It would be useful to be able to compare the implementation of the common ideas and the value of the unique ones, but it is difficult to make such comparisons between systems. This is due to the fact that each system or program has been developed for a specific domain, hence the difficulties in separating the domain sensitive ideas from the more general ideas preclude ready comparison.

Two such systems, INTERNIST and the Present Illness Program (PIP), have similar approaches to diagnosis with some interesting differences. This thesis will examine and compare these two systems using a common problem domain, congenital defects. Versions of both systems have been constructed for this study and a congenital defects database developed for each. Clinical cases have been entered into each system. This thesis will analyze the behavior of these versions of PIP and INTERNIST

during the diagnosis of these cases with respect to the creation of hypotheses, the matching of findings to diseases (syndromes), and the selection of the next action.[1]

Before the comparative study of PIP and INTERNIST is detailed, a review of the PIP and INTERNIST systems and the congenital defects problem domain is outlined.

---

1. All references to the comparison of PIP and INTERNIST refer to the versions of PIP and INTERNIST constructed for use with the congenital defects problem domain, not the original versions.

2. INTERNIST

INTERNIST is a diagnostic program which has been designed and constructed by H. E. Pople, J.D Myers and others at the University of Pittsburgh [29,30,31,32]. This system was designed to encompass all of internal medicine as its problem domain. Hence, it has quite a large database which, at present, contains over 500 diseases.

The following is a description of this system. This discussion of INTERNIST does not take into account modifications planned in INTERNIST II. Since INTERNIST II is still in the developmental stage, none of its features were available for use in this study.

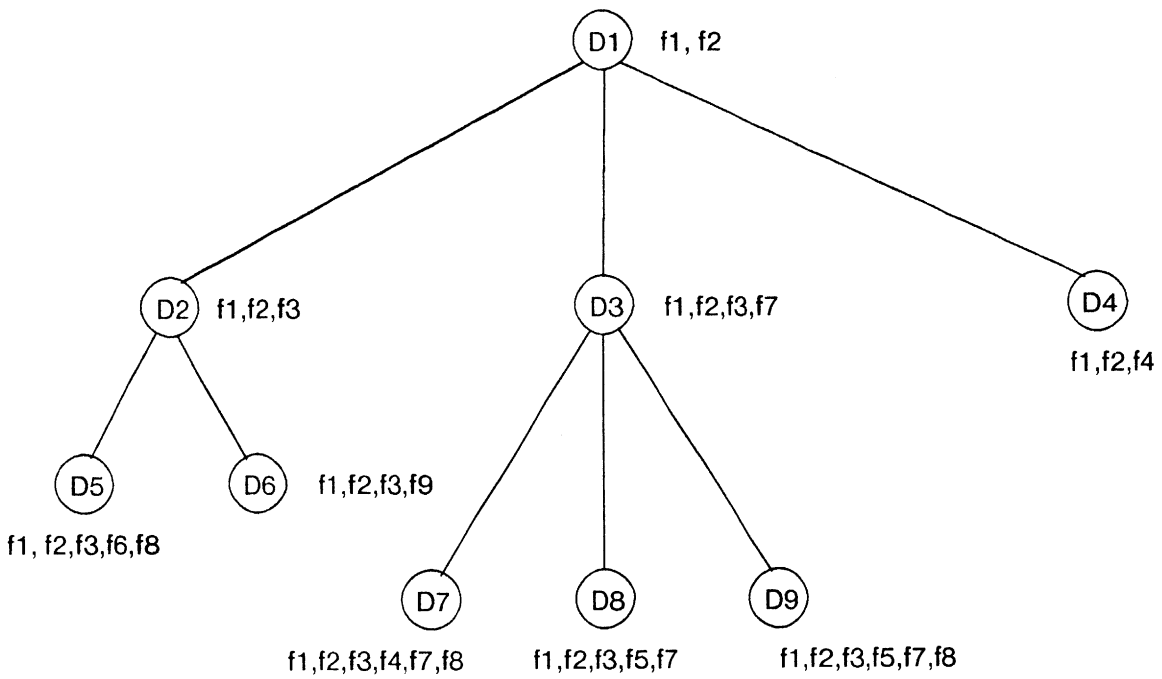
2.1 Representation of Knowledge

INTERNIST's database is organized into a tree hierarchy of diseases, the superior of any disease is a more general disease category represented in the same manner, e.g., the superior of hepatocellular disease might be liver disease. Each disease has a list of manifestations (or findings) associated with it. This list of associated manifestations for a non-leaf disease node is the intersection of the findings of all of its inferior disease nodes (see figure 1).

In turn, each associated finding in a disease has two associated weights, one of which is referred to as the *evoking strength* and the other as the *frequency of occurrence*. The *evoking strength* is a subjective estimate, on a scale from *zero* to *five*, of the likelihood of a disease given a finding. A *zero* indicates that the finding is too non-specific to lend support to that disease. A *five* indicates that the finding is pathognomonic for that disease.

The *frequency of occurrence* is a subjective estimate, using a scale from *one* to *five*, of a finding being present given that the disease is present. A *one* indicates that the finding is only rarely seen with the disease, while a *five* indicates that the finding is

Fig. 1. INTERNIST's Disease Hierarchy



Dn = the nth disease

fn = the nth finding

almost always present with the disease.

For non-leaf disease nodes these associated numbers are computed from the subnodes of the disease. The *evoking strength* of a manifestation is the maximum of the *evoking strengths* associated with that manifestation in its subnodes. The *frequency of occurrence* of a manifestation is the minimum of the *frequency of occurrence* in its subnodes.

Since diseases can present with one or more clinical pattern, INTERNIST allows one to split diseases into several nodes placing each in the appropriate spot in the disease tree. There are *equivalent* links between these nodes denoting that they belong to the same disease. There are also causal links between disease nodes. These are unidirectional links indicating that one disease can cause the other. These links have an associated weight which is a measure of the degree of association between the linked nodes.

There is other information associated with an instances of a disease that is being considered during the diagnosis (i.e., it is a hypothesis). This information takes the form of four associated lists of findings for each hypothesized disease. These include a list of the manifestations that have been observed but which are not explained by the disease, i.e., are not one of the findings associated with the disease. This is called the *shelf* of the hypothesized disease. There is also a list of all the observed manifestations that the disease explains. There is a list of findings that would be expected to occur with the disease and were found absent and, last, there is a list of findings that would be expected to occur with the disease but about which nothing is known yet.

Manifestations are objects, the presence or absence of which can be determined. They could be thought of as LISP atoms since they are not built from other entities. In addition, the manifestations have associated properties. Each manifestation

has a *type* property which can be one of the following: history, symptoms, signs, one of three types of laboratory data, or a syndrome. This enables the system to estimate the expense and danger of the medical procedures. This information is used in question selection (discussed later).

Each of the manifestations also has an associated number, called the *import*, which is an estimate, on a scale from *one* to *five*, of the importance of explaining that manifestation with the diagnosis. In addition, the manifestations also contain links to other manifestations, which allows the system to avoid asking redundant or nonsensical questions. For example, a finding's *prereqs* are links to manifestations which must be present before asking the finding, *unreqs* are links to manifestations which must be absent, etc.

2.2 INTERNIST's Algorithm

Internist's algorithm has two parts. The first phase of the diagnosis begins by having the user enter, one at a time, the initial findings of the case. After each finding has been entered, the system activates the disease nodes which it deems appropriate, i.e. it generates disease hypotheses. The second phase is an iterative loop in which INTERNIST attempts to determine which diseases are present through an interactive dialog with the user. The first step in the loop is the scoring of each of the active disease nodes (hypotheses). INTERNIST then ranks the hypotheses according to their score and chooses a strategy for diagnosis determined by this ranking.[2] Using the selected strategy, INTERNIST chooses several questions to ask the user, which in turn can generate new hypotheses. The second phase then repeats with the rescoring of the hypotheses. This process continues until all of the findings deemed important are

---

2. At this point, instead of selecting a diagnostic strategy, INTERNIST could conclude that a disease is present. One might consider confirming a disease is be one of the possible strategies.

explained by concluded diseases. Each of these processes is described in detail in the following sections.

2.2.1 Hypothesis Generation

INTERNIST's hypothesis generation strategy attempts to limit the number of hypotheses by two methods. First, it activates a disease node only when it contains a finding of *evoking strength* greater than zero. This prevents diseases which no finding strongly suggests from being considered. Second, INTERNIST activates a superior node in the disease hierarchy if findings which would differentiate between the inferior nodes are not known. This allows INTERNIST to generate one general disease hypothesis until more specific information is known and hence save the work of prematurely evaluating several very similar hypotheses.[3]

The system activates all of the disease nodes which contain the finding with an *evoking strength* greater than zero and whose superior does not contain that finding. This insures that the most superior node which contains the finding is the one activated. There are two exceptions to the above statement. If there is a superior node, node A, that is already evoked when an inferior node, node B, qualifies to be evoked (by the above criteria), then all of the inferiors of node A are evoked and node A is deactivated. After this is done the system tries to evoke node B again. If a newly evoked inferior of node A is still a superior of node B, that node is deactivated and its inferiors activated. This continues until node B can be evoked without one of its superiors also being evoked.

---

3. This also allows the strategy selection to consider a group of syndromes as one hypothesis.

The second exception occurs when a superior node is to be evoked and an inferior node, node C, has already been evoked. Then all of the superior node's inferiors are recursively evoked, in the same manner as in the first exception, so that the node C remains evoked and none of its superiors are evoked.

To illustrate INTERNIST's evoking algorithm, consider the disease hierarchy in Figure 1. Suppose that the first finding that was entered was f1, then D1 would be evoked since it is the only node which contains the finding and has no superior which does.[4] If f5 was then entered, D8 and D9 would qualify to be evoked. This would cause D1 to be deactivated and D2 and D4 to be evoked. D3 would not be evoked but all of its inferiors would be. So after the dust settles D2, D4, D7, D8, and D9 would be evoked..

If the order of entry was reversed, first f5 then f1, the other exception to the general rule would occur. After f5 was entered D8 and D9 would be evoked. When f1 was then entered D1, which qualifies to be evoked by the general rule, would not be evoked due to the fact that some of its inferiors (D8 and D9) were already active. Instead, the system would try to evoke D1's immediate inferiors. D2 and D4 would be evoked, but D3 would not be evoked for the same reason D1 was not. Instead D3's immediate inferior D7 would be evoked (only D7 needs to be evoked since D8 and D9 already were). The end result would be the evoking of D2, D4, D7, D8, and D9, the same as before.

This method evokes the most superior set of nodes that explain all of the important findings possible by that branch of the disease tree. This method for generating disease hypotheses is used throughout the diagnosis.

---

4. I am assuming for this example that all findings have evoking strengths greater than zero.

By activating the most superior node possible, the system is able to consider a general group of diseases until more discriminating findings are entered and hence decrease the number of active hypotheses. With the scheme, however, the number of hypotheses can only increase as the diagnosis proceeds.[5]

2.2.2 Scoring Algorithm

The scoring algorithm can be broken into four parts dealing with the four possibilities given the state of a finding and a disease hypothesis. Those possibilities include: that a finding is present and is expected to be present (in a given disease hypothesis), that a finding is present and is expected to be absent, that a finding is absent and is expected to be present, and that a finding is absent and is expected to be absent. The first possibility increases the hypotheses score, the next two possibilities decrease its score, and the last one does not alter the score. There is another part of the algorithm that increases a hypothesis' score for links to confirmed diseases.

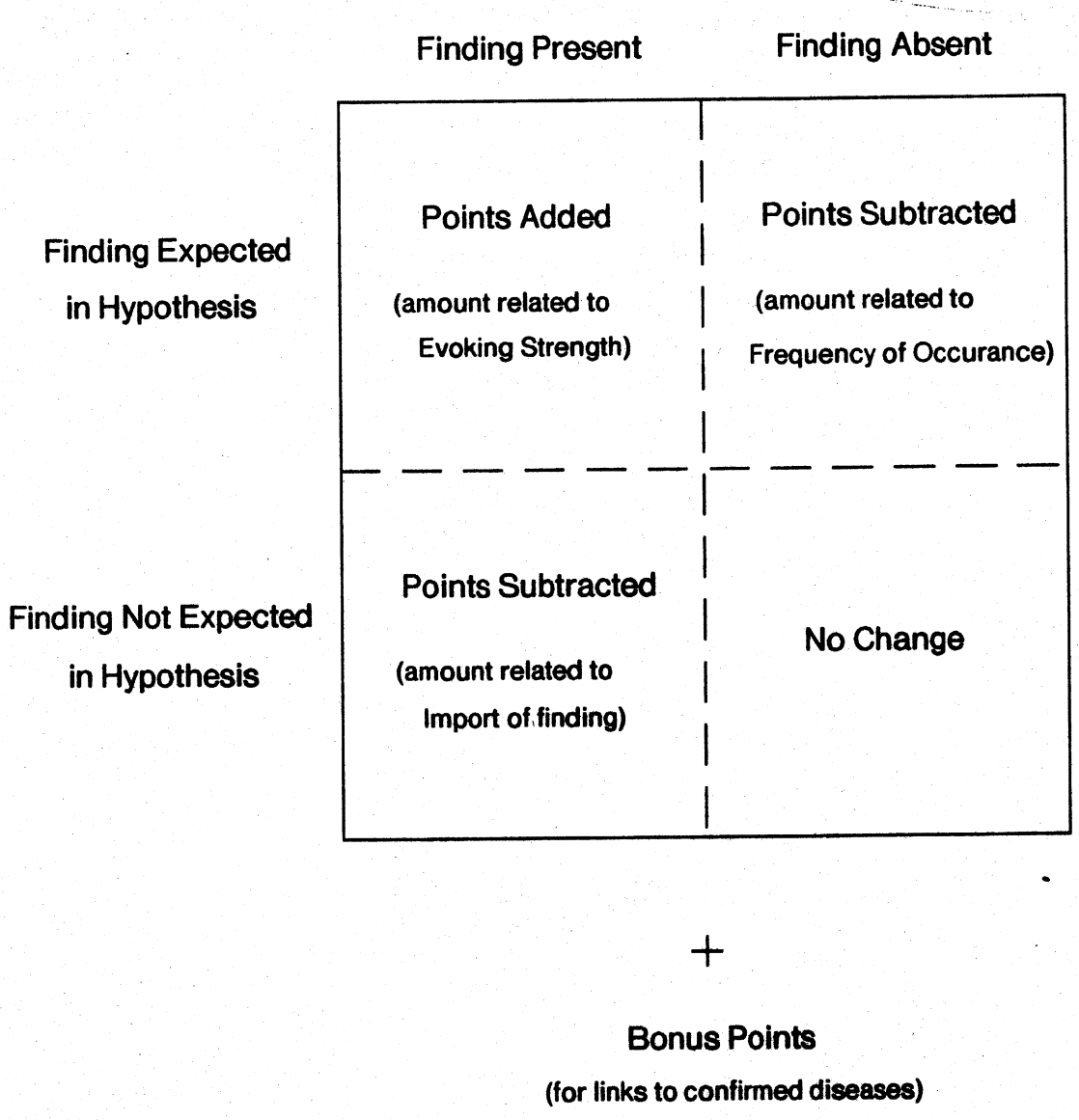
A more detailed description of each part is in order. First, for each manifestation that is known to be present and is associated with the hypothesis, the algorithm adds to that hypothesis' score an amount related to the manifestation's associated *evoking strength* (see figure 2). The scales used are exponential not linear, hence an *evoking strength* of four adds considerably more than would twice an *evoking strength* of two. This same nonlinear scale is used throughout the scoring algorithm.

The next part of the scoring algorithm subtracts from the score of the hypothesis an amount for each of the manifestations found to be present but not explained by that hypothesis (i.e. not associated with the hypothesis) related to the

---

5. There is one exception, confirmed hypotheses are removed from the list of active hypotheses and put on the list of concluded diseases. This, of course, decreases the number of active hypotheses by one, but it is not a method for controlling the number generated.

Fig. 2. INTERNIST's Scoring Algorithm



import of the manifestation.

The third part subtracts from the hypothesis' score an amount for each manifestation found to be absent but which would be expected to be present by the hypotheses, (i.e. the manifestation is associated with the hypothesis). The amount subtracted is related to the *frequency of occurrence* associated with the manifestation in that hypothesis.

The last part of the algorithm involves giving *bonus* points for links, either causal or equivalent, from the hypothesis to previously confirmed diseases. The hypothesis' score is increased in proportion to the weight of the link. These four parts are repeated for each hypothesis.

2.2.3 Strategy and Question Selection

INTERNIST first chooses a strategy for diagnosis and then chooses several questions to ask the user according to the strategy selected.

To choose the diagnostic strategy the system first divides all of the hypotheses into two groups: those hypotheses whose scores are within a given range of the leading hypothesis' score and those whose scores are not. It uses only the former group for the rest of the strategy selection.

The next step in the strategy selection is a partitioning of the remaining hypotheses. The system creates the subset of all the hypotheses that explain either a subset or a superset of the findings explained by the leading hypothesis. This partitioned subset is a list of diseases that are competing to explain the findings explained by the lead hypothesis.

Depending on the number of hypotheses in the partitioned subset and their scores, either a hypothesis is concluded or one of several different strategies is selected. The criteria for concluding a hypotheses is discussed later. The different strategies for diagnosis are: *confirm, discriminate, ruleout*, and *narrow*. The *confirm* strategy is used when there is only one hypothesis in the partitioned subset. It chooses questions[6] to ask that could cause an increase in that hypothesis' score. These questions are about the associated manifestations with high *evoking strengths*.

The *discriminate* strategy is used when there are two to four hypotheses in the partitioned set. The system tries to distinguish between the top two hypotheses by choosing questions that could raise the score of one hypothesis while lowering that of the other. Those findings associated with one hypothesis having a high *evoking strength*, which is not associated with the other hypothesis, as well as having a high *import* value are good choices for questions.[7]

The *ruleout* strategy is invoked when there are more than four hypotheses in the partitioned subset. This strategy chooses questions that could lower one of the hypothesis' score, thus ruling it out. These questions are about the associated manifestations with a high *frequency of occurrence* since these would decrease the hypothesis' score if it were answered negatively.

In addition to the above criteria for question selection, the system searches first for findings of low cost and risk before checking those of higher cost or risk manifestation. The cost and risk value is determined by the *type* property of the manifestation. The type order used for question selection is as follows: historical facts

---

6. Questions are just inquiries as to whether manifestations which have not been previously entered are present, absent, or unknown.
7. Also if the negation of a finding is associated with the other hypothesis with a high *evoking strength* the finding would be a good choice.

are checked first, then symptoms, signs, and finally the different levels of laboratory tests.

If the system is in *ruleout* mode it will not ask about laboratory results. If no historical findings, signs, or symptoms are found then INTERNIST changes to *narrow* mode. In this mode it narrows the field to the top two hypotheses and uses the *discriminate* mode's strategy to select questions. This is intended to prevent the system from asking costly questions when the probability of the disease being present is low.

Several questions are selected using the given strategy and asked before evaluating the hypotheses. The answers to these questions are processed in the same manner as the initial findings, evoking new disease nodes where necessary.

2.2.4 Concluding Diseases

A hypothesis can only be concluded when it is the only disease in the partitioned subset, i.e. the system is in *confirm* mode. When the leading hypothesis has been the only member in the partitioned set long enough for the difference between its score and that of the next closest hypothesis to be twice that of the initial difference (upon entering confirm mode) between the two, then the leading hypothesis is concluded to be present by the system. Viewed another way, to conclude the leading hypotheses, the difference between the top two hypotheses must be twice the difference required to cause the system to be in *confirm* mode.

After a disease has been concluded the manifestations that are explained by that disease are removed from further consideration. If there are *important* findings not explained by the concluded diseases then the second phase continues with bonus points being given to hypotheses linked to the concluded diseases. The importance of a finding is determined by its *import* value.

By removing the finding explained by concluded diseases, INTERNIST is requiring that co-occurring diseases have a sufficient number of distinct findings to identify both of them as being present or that there be a causal link between the diseases to offset the effect, by awarding bonus points, to the remaining disease's score of removing the findings.

2.3 Summary

INTERNIST is a hypothesis driven diagnostic system which uses a pseudo-Bayesian scoring algorithm to score those diseases which it has hypothesize are likely. Based on the scores INTERNIST selects a diagnostic strategy and using this strategy selects findings to question the user about. Concluding a syndrome's presence is done by exceeding a relative (rather than absolute) threshold between the top two hypotheses. All findings explained by a concluded hypothesis are removed from further consideration.

3. Present Illness Program

The Present Illness Program is a frame-based diagnostic program which has been constructed by S.G. Pauker, G.A. Gorry, J.P. Kassirer, W. B. Schwartz, and P. Szolovits [27,28,44] at M.I.T. and the Tufts University School of Medicine. It has been implemented using the renal disease problem domain, although it was intended to be a more general medical diagnostic system. Like INTERNIST, PIP uses a pseudo-Bayesian hypothesis driven approach to diagnosis, but its design differs from INTERNIST in many aspects. This chapter contains a description of PIP as well as a comparison with INTERNIST.

3.1 Representation of Knowledge

In PIP each disease is represented as a frame [22] with many entries or slots which contain the information about the disease. Each slot has an associated value or values. A typical diseases frame is given in figure 3.

The slots used by PIP in the disease frames include:*type*, *scoring-function*, *must-not-have*, *must-have*, *is-sufficient*, *triggers*, *findings*, *differential-diagnosis*, and some causal and associative link slots (e.g., *major-cause-of*). The use of each of these slots will be discussed when the algorithm is described. It suffices to say that not all of these slots have equivalent structures in INTERNIST's representation. In addition to the disease frames there are also clinical state and physiological state frames which are identical to disease frames (disease will be used to mean disease, clinical state, or physiological state in this chapter). The clinical state and physiological state frames are used to collect groups of findings that, although they are not diseases, have been identified by physicians as clinically useful intermediates between findings and diseases. It was the intention of PIP's designers that these frames be richly interconnected with causal and associative links.

Fig. 3. Typical Disease Frame in PIP

```
FRAME

     NAME:              ACHONDROPLASIA
     TYPE:              SYNDROME
     FINDING:           STATURE WITH LEVEL SHORT, TYPE DISPROPORTIONATE,
                              AND ONSET AT-BIRTH
     FINDING:           FRONTAL-BOSSING WITH STATUS PRESENT
       .
       .
       .
     MUST-NOT-HAVE:     STATURE WITH LEVEL NORMAL OR LEVEL TALL
                                   OR
                        RHIZOMELIA WITH STATUS ABSENT

     TRIGGERS:          FRONTAL-BOSSING WITH STATUS PRESENT
                        STATURE WITH LEVEL SHORT, TYPE DISPROPORTIONATE,
                                   AND ONSET AT-BIRTH
                         .
                         .
                         .
     IS-SUFFICIENT:     RHIZOMELIA WITH STATUS PRESENT
                                   AND
                        TRIDENT-LIKE-HAND WITH STATUS PRESENT

     DIFFERENTIAL-DIAGNOSIS:
                        IF: SHORT-STATURE WITH STATUS PRESENT
                                                AND TYPE PROPORTIONATE
                        THEN: SILVER-SYNDROME

     MAJOR-CAUSE-OF:    SCOLIOSIS

     SCORING-FUNCTION:
                        FRONTAL-BOSSING:  STATUS PRESENT: .9
                                          STATUS ABSENT: -.6

                        RHIZOMELIA:  STATUS PRESENT: 1.0
                                     STATUS PRESENT: -1.0
                         .
                         .
                         .
```

Findings are also represented as frames. The slots contain the possible items that the finding can have. Unlike INTERNIST's findings which can only be *present absent or unknown*, PIP's findings can have many values. Each item can take on one of two or more possibilities. So the finding *stature* might have level short, onset at birth and degree mild as associated items and values. This is quite different from INTERNIST's representation of findings as atomic objects. Of course, by enumerating all of the possible values one could convert one of PIP's findings into INTERNIST compatible findings. The frame representation merely allows the system to capture many different aspects of the finding in one place, hence reducing the number of findings and the number of links between findings. The links between findings can be placed in the frames when needed, but this is not well developed in PIP.

In PIP, the item frames contain the possible values of the item. These values can be mutually exclusive or multi-valued. PIP will create multiple instances of a finding if more than one value is given to a multi-valued item, each instance has a different value for the multi-valued item, but it will not allow instances to occur with two values of any single-valued item and the same value(s) of the multi-valued item(s). For example the in the finding *edema*, *time* item can have multiple values but *status* can take only a single value, so "edema with time now and status present" and "edema with time past and status absent" is allowed but "edema with time now and status present" and "edema with time now and status absent" is not allowed.

3.2 PIP's Algorithm

3.2.1 Hypothesis Generation

The diagnostic session begins, as in INTERNIST, by entering a list presenting complaints. Each of these findings is compared against the values of the findings in the trigger slot in each disease frame. When a match occurs between any of the input

findings and any of the findings in the trigger slot, that disease is activated (i.e., a frame representing an instance of the disease is created and its name is put on the list of active hypotheses).

After the trigger slots have been checked and the hypotheses generated, the values of the *must-have* and the *must-not-have* slots of the active hypotheses are checked. The value of the *must-not-have* slot (or *must-have* slot) is a finding (with its associated items and values for those items) or several findings connected via logical operators (see figure 3). These slots are evaluated, i.e. the values of the findings listed in their slots are compared with the values entered for those findings and any logical operator present is applied to the results of the comparison. If the slot evaluates to *true*, then the frame is removed from the list of active hypotheses (deactivated).

There is another slot which is checked for all active hypotheses at this time, the *differential-diagnosis* slot. This slot can have several values, each value has two parts. The first part is a condition identical to the *must-not-have* and *must-have* slots which can be evaluated in the above manner. The second part is a disease, clinical state or physiological state. If the first part evaluates to true then an instance of the frame referred to in the second part is created and marked as *semi-active*. If any other of the diseases findings are present or entered later then the disease is activated, (i.e., a semi-active frame is treated as if all its findings were triggers). Semi-activation is also caused by the causal or associative links between frames.

The causes of hypothesis activation in PIP differ in several ways from those in INTERNIST. First, in PIP, the findings that trigger a disease from an inactive state are independently chosen. In INTERNIST, the findings used to trigger a disease are chosen according to information used by the scoring algorithm. Hence, it is not possible, in INTERNIST for a finding to cause a hypothesis to be generated and have it contribute less than a given amount (i.e., they must have an evoking strength greater than zero) to

the hypotheses score. Similar to INTERNIST, only one trigger finding is needed to activate a disease. One might postulate that separating the information used to trigger hypotheses from the scoring information should allow one to increase the specificity of the triggering mechanism, hence decreasing the number of hypotheses generated and increasing the probability of activating the correct hypothesis. But it might be that INTERNIST mechanism is wholly adequate for generating hypotheses and scoring them. This will be looked at in the comparison.

Another difference between PIP's and INTERNIST's methods of controlling the number and appropriateness of hypotheses is PIP's ability to deactivate hypotheses via the *must-have* and the *must-not-have* slots.[8] INTERNIST does not deactivate any disease after it is activated, but it can use the disease tree to generate one abstract hypothesis until enough information is known to generate several more specific ones.[9]

PIP's other method of activation, using a *semi-active* state, does not represent different degrees of activation, but rather the dynamic nature of the causes of activation in PIP. So, depending on the state of the system, there can be different triggers used for hypothesis generation. This seems to be a reasonable idea in theory, but constraining a single finding to either trigger or not trigger a disease independent of the situation may be too simplistic a model of hypothesis generation in medical diagnosis. If this is the case PIP's *semi-active* state may be able to capture some of the missing knowledge, although the use of semi-activation still highly constrains any dynamic nature of the triggers. INTERNIST does not allow the causes of activation to change dynamically during the course of the diagnostic session.

---

8. PIP also deactivates hypotheses when their score drops below a given threshold.
9. The disease tree is also used in the strategy selection process discussed later.

3.2.2 Scoring Algorithm

The scoring algorithm in PIP is, like INTERNIST, pseudo-probabilistic. There is a *scoring-function* slot which contains a list of the findings of the disease with associated items and the possible values for those items. Associated with each value of each finding is a weight (see figure 3). The weight is a floating point number which varies from negative one to positive one, it could be thought of as a shifted, scaled, conditional probability of a finding given the hypothesis. For each active hypotheses the *scoring-function* slot is examined. When a value of a finding in the *scoring-function* slot of a disease matches the value of a known finding, the weight associated with that value is added to the score of that disease. This total score is divided by the maximum possible score to normalize the value. In addition, the fraction of known findings explained by each hypothesis is calculated. These two scores are averaged for each hypothesis to get a final "averaged" score.

Before PIP does the above calculations for a disease it checks for causal and associative links between that disease hypothesis and other active hypotheses. For scoring purposes the findings associated with linked frames that are activated are treated as if they were part of the original frame. The links can be either major or minor links. The associated weights of findings in the *scoring-function* slot are multiplied by 1.0 for major links and 0.3 for minor links before being used. These links can form arbitrarily long chains as long as the links are going in one causal direction. If no information about a linked frame is known then the frame is ignored in the scoring calculation.

PIP includes some information in the scoring algorithm that INTERNIST does not and vice-versa. The normalization of the hypotheses score over the maximum possible score for the disease is unique to PIP, there is no equivalent calculation in INTERNIST. This prevents a disease from being concluded due to a match with the first

few findings. But it would seem to penalize diseases with many findings and links. [10]

The calculation of the fraction of the known findings explained by a disease is analogous to INTERNIST's use of the *import* of a finding in the scoring algorithm, except that using the *import* gives a weighed measure to use in determining the fit. So in this calculation INTERNIST's algorithm makes use of more detailed information. These differences will be examined in the comparison.

3.2.3 Strategy & Question Selection

In PIP, the diagnostic strategy and question selection given the strategy are more simplistic than in INTERNIST. PIP always focuses on the leading hypothesis (disease, clinical state, or physiological state) and the question selector picks the first unanswered finding on the ordered list of findings (the finding slots are ordered) of that hypothesis. This list was ordered by the designer of the database.[11] If there are no active hypotheses the system searches for an unknown finding in the causally-related semi-active hypotheses.

This is contrasted to INTERNIST's more complex algorithm which partitions the hypotheses into competing sets, chooses a mode and questions dependent on the state of the diagnosis. Question selection in both systems uses information attached to each finding. As in hypothesis generation, INTERNIST uses some of the same information for many different tasks (in this case the scoring and hypothesis generation information is also used for question selection).

---

10. PIP's ignoring of links that have no known findings is an attempt to prevent this.
11. The findings in the birth defects database, for PIP, are ordered so that PIP first asks several questions which are the equivalent of INTERNIST's *ruleout* questions (i.e., findings that are usually found to be present with the syndrome). After these question, the ordering then causes PIP to ask the equivalent of INTERNIST's *confirm* questions.

One would tend to believe that the more sophisticated strategy used by INTERNIST would result in the selection of more appropriate questions. This will depend on whether the algorithm and information INTERNIST uses is sufficient to determine appropriate questions. On the other hand, it may be that the information embedded in PIP's ordering of the findings, with PIP's simpler algorithm, will prove to select better questions than the algorithm used by INTERNIST. This remains to be seen.

Although the designers' intent behind the *differential-diagnosis* feature of PIP appears to be one of controlling the strategy and question selection, the implementation, using *semi-activation*, does not allow for direct control over strategy selection or question selection. Hence the *differential-diagnosis* feature can be said to influence the strategy only indirectly via hypothesis generation.

3.2.4 Concluding Hypotheses

PIP has two methods of concluding hypotheses. The program concludes that a hypothesis is present if the normalized score computed using the scoring function is greater than 0.8 and the averaged score is greater then 0.5. This does not require that a hypothesis explain any minimum fraction of the known findings since a score of 1.0 will have an average score of greater than 0.5 (it must explain at least one finding so the fraction of findings explained must be greater than 0.0), but it does require that most of the important findings for the disease be present before concluding the disease.

Although both PIP and INTERNIST use thresholds to determine when to conclude that a disease is present, there is a difference between the methods. In PIP, it is the magnitude of the score that is being used to determine when to conclude a disease, whereas in INTERNIST it is the magnitude of the spread between the top two hypotheses that is used to conclude a disease's presence.

The second method for concluding diseases in PIP is through categorical reasoning via the *is-sufficient* slot. If, in an active hypothesis, the clause in this slot evaluates to *true* then that hypothesis is concluded. This is particularly useful for clinical and physiological states where minimal diagnostic criteria are well established.

After a hypothesis has been confirmed, it is removed from the list of active hypotheses but the findings that it explains are not removed. The diagnosis proceeds until all of the remaining active hypotheses are either confirmed or deactivated.

PIP's method for handling the findings explained by confirmed hypotheses is the opposite extreme from INTERNIST's. INTERNIST removes all explained findings and PIP leaves them. Both of these methods are simplistic solutions to a difficult problem. If multiple syndromes are present in a case it is not reasonable to assume that none of the findings will overlap which is INTERNIST's assumption.[12] One would predict that PIP's leaving all the explained findings to contribute to the other hypotheses' scores is not an acceptable solution either since this will require ruling out hypotheses whose findings have all been explained by the concluded diseases.

The solution to this problem may require a large amount of knowledge of the interactions between findings. Such problems are addressed in other on going work, e.g. [25]. The problem of handling multiple syndromes will not be treated in this thesis.

---

12. This is compensated to some degree if the two syndromes are known to interact, since a link can be placed between the two syndromes so concluding one will increase the score of the other. But this cannot be done for syndromes not known to occur together.

3.2.5 Summary of PIP and INTERNIST

To summarize these two systems a brief comparison is in order. Both systems essentially match a set of manifestations against a set of known diseases in an attempt to find the "best fit" of the manifestations to one or more of the known diseases. The match actually involves only a subset of the diseases known to the system; the active hypotheses. Both programs alternate between question asking and hypothesis generation and evaluation. During the hypothesis evaluation each system scores the current hypotheses with pseudo-probabilistic algorithms and ranks each hypothesis according to that score. Both systems attempt to expedite the diagnostic process by constraining the number of active hypothesis at any one time.

There are also important differences, most notably in the implementation of the above common ideas. The organization of the databases is quite different. PIP uses a interconnected frame representation in which the diseases have both competing and complementary links. INTERNIST has a complementary link which is used only after a disease is concluded.

PIP uses only one strategy, confirm, for diagnosis, whereas INTERNIST uses several; confirm, differentiate, narrow, ruleout. The scoring algorithms deviate from a pure Bayesian schema in different ways.

Each system uses different methods to control the number of active hypotheses. PIP uses categorical reasoning (i.e., triggers, must-have and must-not-have slots, etc.) and a semi-active state to limit the active hypotheses. INTERNIST uses an evoking threshold and a tree hierarchy of diseases which enables it to hypothesize one general disease type instead of many similar specific diseases when detailed information, needed to differentiate between the specific diseases, is not known.

Each of these differences represent design choices that were made. To determine which of these choices are advantageous requires a comparative study with an implementation of both systems using the same problem domain. The congenital birth defects field was chosen as the common domain to perform such a study.

4. The Birth Defects Database

4.1 The Domain

The birth defects field was chosen for several reasons. From a medical viewpoint this domain is an interesting one. Each year, in this country alone, more than a quarter of a million infants are born with birth defects, greatly affecting their lives as well as the emotional and financial lives of their families. Thus, the impact of birth defects is felt daily by more than twenty-five million Americans.

The rapid pace of developments in genetics and pediatrics has made it difficult for physicians to keep abreast of changes in the field. This, along with the sheer size and complexity of the birth defects domain makes it an area in which a computer aid might be useful to the physician who is not a specialist in the field. From a computer science view point, birth defects is not the original domain for either PIP or INTERNIST, hence neither program will have been designed specifically for this field. The domain is also broad enough to encompass many types of medical diagnostic problems which must be dealt with by these computer systems.[13]

There are several properties of this problem domain that make it well suited for the approach taken by PIP and INTERNIST. The diagnosis of birth defects is often syndromic in nature (i.e. a group of symptoms without precise cause [18]), which leads one to believe that the amount of physiological and causal reasoning needed can be minimized. Also birth defects usually occur as isolated events, thus interactions between syndromes are avoided. Theses two properties of the domain simplify, to some extent, the diagnostic process for PIP and INTERNIST since unanticipated interactions between

---

13. Some adjustments will be made to both systems to enable them to achieve better performance in this domain.

syndromes requiring physiological reasoning are difficult to handle in these systems. Hence by avoiding these interactions the performance of the two systems should improve. Of course, a comparison of these systems different methods of handling multiple diseases will not be possible.

4.2 The Database

Through collaboration with the Center for Birth Defects Information Services (CBDIS), a division of Tufts-New England Medical Center, the database used for both PIP and INTERNIST was adapted from an already existing database. This database, developed by CBDIS for use by their own diagnostic algorithm, contains information concerning over one thousand birth defects. It was hoped that the modifications required for PIP and INTERNIST to use this database would be minimal, and, possibly, could be computer generated. Unfortunately, this was not the case. Much of the information required was not derivable from their database and had to be provided by the physicians at the CBDIS.

To try to implement all birth defects would have been too large an undertaking, so a subset of the field is used. This subset contains skeletal and endocrine defects. These two areas have considerable overlap since hormonal irregularities often affect skeletal development. These areas also provide a range of known causality, from totally unknown causation to knowledge of precise biochemical and genetic defects.

There are a total of fifty diseases in the database for each system, these are listed in appendix one. Of these fifty diseases (or syndromes), thirty-three would be classified as endocrine defects and seventeen as skeletal defects, although as stated above there is some overlap between the two groups and a precise dividing line is difficult to define. In addition, for PIP, there are ten clinical states and forty-five links between the states and the diseases. INTERNIST has a total of sixty-nine syndromes of

which nineteen are superior nodes for other syndromes. The average syndrome frame in PIP has approximately thirty findings associated with it, as compared to about thirty-five in INTERNIST.

The number of findings in PIP's database is 847 and in INTERNIST's database 980. The reason for the difference is that many of INTERNIST's findings are combined together to form one of PIP's findings. For example, the serum sodium level had several findings associated with it in INTERNIST: increased, normal, and decreased. In PIP, this was combined into a single finding with a level item that could take on these values.

Due to the small size and similarity of the syndromes in these databases extrapolation of results to larger databases with larger numbers of syndromes may be difficult. But these databases should be sufficient for the comparison of PIP and INTERNIST, described in the next chapter.

5. The Comparison of PIP and INTERNIST

The following comparison examines the behavior of versions of PIP and INTERNIST when presented with clinical cases from the congenital defects problem domain. Although these two systems employ the same basic diagnostic model (hypotheses driven diagnosis), they differ in their implementions in many respects. These differences, presented in chapter 3 (and later in this chapter), reflect design choices that were made. This comparison attempts to determine the relative benefits of the choices made for each system.

The comparison between these two systems has been divided into four orthogonal areas in order to better contrast their functioning. These areas are: 1) hypothesis generation, 2) scoring, 3) strategy and question selection, and 4) hypothesis confirmation. Each section will outline the comparisons performed, report the results obtained, and give an explanation of these results.

The clinical cases used in this comparison were obtained from the records of the Tufts-New England Medical Center and other medical centers around the country. In addition, cases have also been taken from the literature. A description of these cases is provided in Appendix 2.

5.1 Hypothesis Generation

The goal of hypothesis generation is to consider (hypothesize) all reasonable explanations[14] (syndromes) while not taking into consideration those which are unreasonable. This is to say that one wishes to minimize the number of hypotheses

---

14. By "reasonable explanations" I mean all explanations for which there is medical justification for considering, given the known findings. The medical justification may be due to many reasons (e.g., a disease's likelihood of being present, or treatability if detected early, etc.).

generated without overlooking the correct one. This allows the system to avoid spending time on unlikely candidates.[15]

The optimal number of hypotheses is difficult to determine. A system would certainly be functioning non-optimally if it failed to hypothesize the correct syndrome. Indeed, one would like the correct syndrome to be hypothesized by the point in the diagnosis where it might begin to be investigated, if it was hypothesized. The other extreme would be if a system generated so many unreasonable hypotheses that, given the strategy and question selection mechanism, many unreasonable hypotheses were investigated, lengthening the diagnostic session and decreasing the physician's confidence in the system.

To determine how well PIP and INTERNIST generate reasonable hypotheses, several tests were performed, using the 35 test cases. Several different areas of hypothesis generation were investigated. These include overall performance, the usefulness of the disease hierarchy in hypothesis generation, deactivation to control hypotheses generation, and separation of scoring and evoking information. The following sections describe this investigation.

5.1.1 The Overall Performance

In both INTERNIST and PIP the presence of a single findings is used to generate a hypothesis. The difference between the two methods of evoking hypotheses is that in PIP only a few selected findings of a syndrome will cause it to be hypothesized (activated)[16] , whereas in INTERNIST any of the findings associated with a syndrome

---

15. Minimally each hypothesis must be scored and a decision made not to use it further on this iteration.

16. In PIP, if a syndrome is semi-activated, all of its findings act as triggers and hence any one of its findings, if present, will cause the syndrome to be hypothesized.

with an *evoking strength* greater than zero (i.e. most of the findings) will cause it to be hypothesized. PIP also deactivates hypotheses whereas INTERNIST uses a disease hierarchy to control hypotheses generation (but never specifically deactivates hypotheses).

To attain an overall measure of hypothesis generation in these two systems, a comparison of INTERNIST's evoking of disease nodes and PIP's triggering of syndromes was made. The number of hypotheses generated was determined at two points in the diagnostic session: 1) after the initial findings were entered and 2) at the end of the session.[17] In addition, a physician at the Center for Birth Defects Information Services was asked to judge the appropriateness of the hypotheses. The appropriateness of a hypothesis is of course a value judgment. To minimize the effect of personal preference, inherent in the one physician sample, only two categories were used, appropriate and inappropriate, and any hypothesis that could at all be included in a differential diagnosis given the findings was considered appropriate, (i.e only blatantly unsuitable hypotheses were considered inappropriate). The number of inappropriate syndromes was also determined both after the initial findings were entered and at the end of the session for all thirty-five cases. The following tables and figures give the results of these comparisons.

The graphs in this chapter show the item being compared (e.g., number of hypotheses activated, number of questions asked, etc.) on the vertical axis and the cases on the horizontal axis. The cases have been sorted according to the results for one of the systems so that the resulting data is in increasing order going from left to right. The corresponding data for the other system is paired with it, of course. For example, in figure 4 the cases are sorted so that the number of hypotheses generated by

---

17. The end of the session was defined as either when the system concluded the correct syndrome or when it was apparent that it could not conclude it.

PIP increases from left to right. This is done solely to make the graphs more readable.

The *initial findings* were defined as either the findings initially entered by a physician into the Center for Birth Defects Information Services algorithm or the findings obtained in the initial examination of a patient, including routine laboratory results for cases extracted from the literature. The initial findings entered were the same for each system with the proviso that PIP's findings sometimes combined two or more of INTERNIST's findings. The finding count used for comparison purposes is the number of findings entered into INTERNIST.

In this comparison there is no taking into account the fact that PIP uses clinical and physiological states that can break a syndrome into several hypotheses. This could either increase or decrease the number of hypotheses generated since a state could be activated in addition to a syndrome or a state could be activated instead of several syndromes. In this domain it appears that this is a minor effect and often mimics INTERNIST's disease hierarchy,(i.e., when INTERNIST has a superior node evoked only the state is hypothesized in PIP and when terminal nodes are evoked in INTERNIST the state and associated syndromes are activated in PIP). Concluded hypotheses were not considered active for either PIP or INTERNIST.

The results show that PIP evokes fewer hypotheses after the initial findings have been entered; a mean of 23.4 (standard deviation 4.9) for INTERNIST as compared to 5.4 (standard deviation 2.4) for PIP. This difference is statistically significant ($p < 0.001$ using the student's t). The difference in the number of inappropriate hypotheses, an average of 15.5, was also statistically significant at the 0.001 level. These results carried through to the end of the case. INTERNIST generates an average of 25.1 (standard deviation 5.2) hypotheses at the end of the case compared with 5.5 (standard deviation 2.5) by PIP. The mean number of inappropriate hypotheses generated by the programs at the end of a case was 18.7 (standard deviation of 5.4) and 1.7 (standard deviation 1.7) for

Table I. Number of Hypotheses Generated

Case	Number of Initial Findings	Number of Hypotheses Generated ** PIP Initial	PIP Total	INTERNIST Initial	INTERNIST Total
1	26	10/5	11/4	31/18	31/18
2	17	5/0	5/0	21/13	21/13
3	16	5/0	5/0	24/14	24/14
4	24	9/3	9/3	27/19	32/24
5	6	6/1	6/1	21/11	23/13
6	5	3/1	5/3	18/10	24/16
7	11	5/1	5/1	21/14	21/14
8	21	3/2	3/2	18/17	20/19
9	8	4/1	4/1	26/23	26/23
10	14	3/1	5/2	26/20	29/23
11	19	9/5	9/5	25/21	25/21
12	9	4/1	6/3	15/9	16/10
13	15	6/2	1/0	34/28	39/33
14	8	8/2	7/2	18/14	21/17
15	15	8/1	8/1	31/20	32/21
16	11	5/0	5/0	21/13	21/13
17	13	9/3	9/3	29/19	29/19
18	25	3/0	5/0	26/17	26/17
19	20	6/1	7/1	26/17	26/17
20	13	3/3	4/2	26/22	26/22
21	27	7/3	7/3	29/19	29/19
22	22	2/0	2/0	23/16	23/16
23	10	7/3	4/2	25/21	28/24
24	17	3/0	4/0	22/17	22/17
25	22	5/0	5/0	23/16	23/16
26	11	3/0	3/0	15/8	15/8
27	8	3/1	3/1	20/15	20/15
28	11	9/1	10/3	27/18	34/25
29	8	6/0	2/0	18/11	22/15
30	14	3/0	3/0	17/13	17/14
31	15	9/6	8/7	33/31	33/31
32	8	6/2	9/5	20/15	26/21
33	10	4/1	4/1	26/24	26/24
34	9	4/2	6/2	21/16	23/18
35	13	3/1	2/2	19/15	27/23
Mean	14.3	5.4/1.5	5.5/1.7	23.4/17.0	25.1/18.7
Median	13	5/1	5/1	23/17	25/18
S.D.	6.0	2.3/1.5	2.5/1.7	4.9/5.0	5.2/5.4

** Numerator is the total number of hypotheses generated;
Denominator is the number of inappropriate hypotheses generated.

Fig. 4. Number of Hypotheses Generated after Entry of Initial Findings

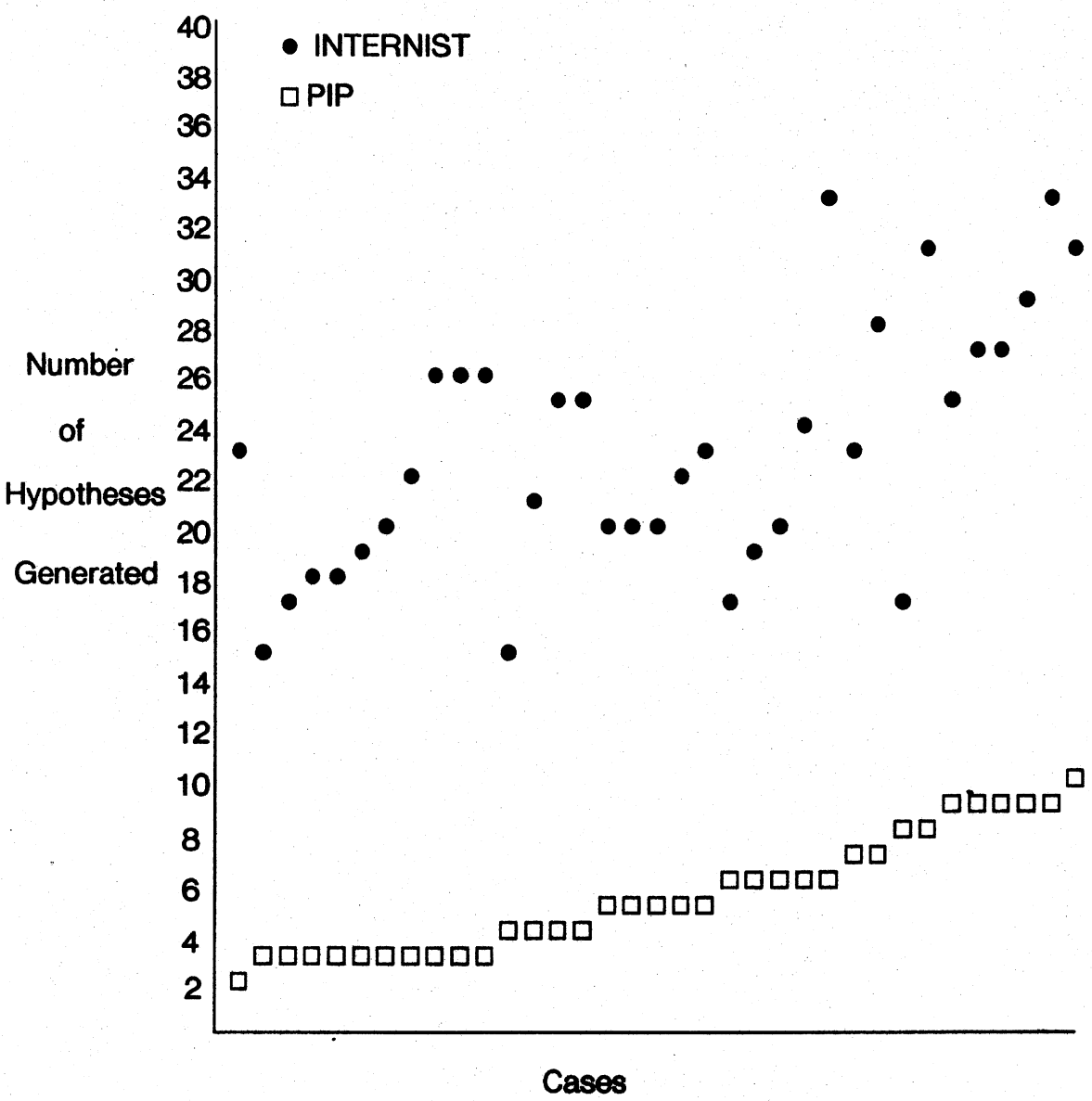
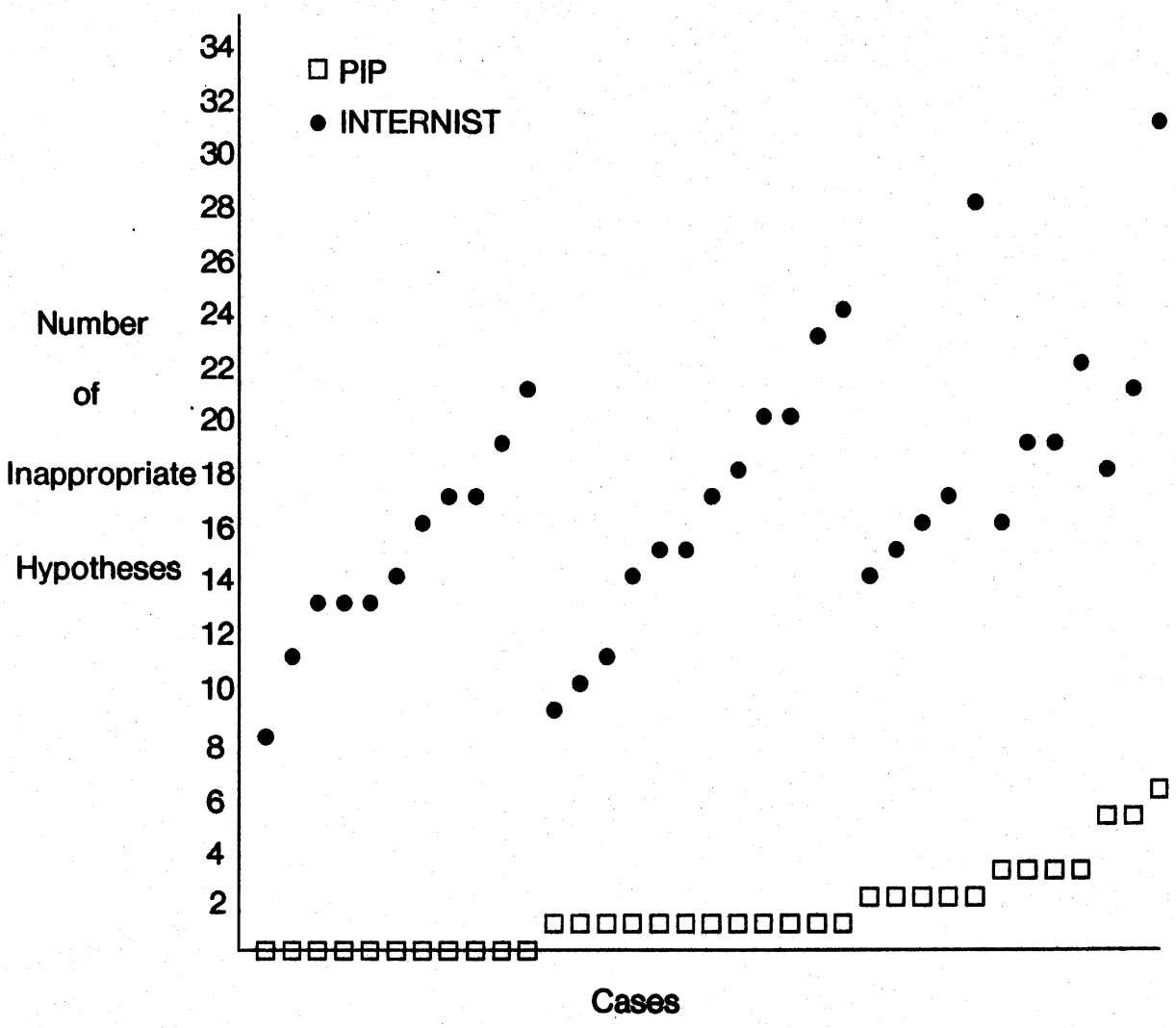


Fig. 5. Number of Inappropriate Hypotheses After Entry of Initial Findings



INTERNIST and PIP respectively.

After the initial findings were entered, in all thirty-five cases INTERNIST had evoked the correct syndrome, whereas PIP had triggered the correct syndrome after the initial findings were entered in all but one of the cases (case 12). In case 12 PIP triggered the correct syndrome after the second question was asked. This was before INTERNIST had even begun to inquire about the correct hypothesis.

So these results show that PIP's hypothesis generation algorithm generates fewer hypotheses as well as fewer inappropriate ones. This was not done at the expense of hypothesizing the correct syndrome. Whether this reduction improves the performance of the system is less certain. Due to the use of different scoring algorithms, strategies, and question selection mechanisms it is difficult to compare the effect of the hypothesis generation throughout the program. Both systems suggested to the user hypotheses that were deemed inappropriate and both then inquired about them. This could cause physicians to have less faith in the decisions of a system, because it proposes and pursues unreasonable hypotheses. However, there were cases where inappropriate hypotheses generated by INTERNIST, but not by PIP's algorithm, were suggested to the user as possibilities and inquired about. The inverse was not observed. This would indicate that PIP's more constrained method for hypothesis generation may reduce the number of inappropriate hypotheses pursued and hence increase user confidence in the systems.

5.1.2 Deactivation of Hypotheses

The acquisition of new findings may, of course, generate new hypotheses. In addition, one would expect that this new information might eliminate certain hypotheses. The analysis of protocols showed this to be the case with clinicians [20,23]. Although a total of over twenty hypotheses were usually considered at some time during the

diagnosis, the maximum number of hypotheses considered at one time was only five or six for specialists in the field and nine or ten for non-specialists [20,23].

INTERNIST does not permit any evoked nodes to be deactivated unless their inferior nodes are evoked or the node is concluded. Hence the number of active hypotheses cannot decrease unless a syndrome is concluded.

PIP, on the other hand, does allow for deactivation of hypotheses via the *must-not-have* feature.[18] So the use of deactivation for limiting the generation of hypotheses by PIP was investigated. The results are shown in the following two tables and associated figures. These tables show that using deactivation does decrease both the number of hypotheses generated and the number of inappropriate ones. The average decrease in the number generated was 1.8 after the initial findings were entered and 2.9 at the end of the diagnostic session. These differences were statistically significant ($p < 0.001$). The number of inappropriate hypotheses decreased an average of 0.9 after the initial findings were entered and 1.3 at the end of the session. These differences were also significant at the .001 level.

In addition to the above results it was also noticed that the leading hypothesis was deactivated on seven occasions during the 35 diagnostic sessions. Since PIP uses the leading hypotheses for question selection, deleting this hypothesis decreased the number of questions asked in some of these cases.[19]

---

18. PIP also deactivates hypotheses when their score drops below a threshold, although this deactivation is reversible since if another trigger finding is entered the hypothesis is reactivated.

19. Deactivating the leading hypotheses does not always decrease the number of question asked since the finding that caused its deactivation could have lowered its score to the point where it would no longer be the leading hypotheses, and hence no longer pursued.

Table II. Number of Hypotheses Generated by PIP With and Without Deactivation

Case	Number of Initial Findings	Number of Hypotheses ** After Initial Findings Deactivation No	Yes	At End of Session Deactivation No	Yes
1	26	17/7	10/5	17/7	11/4
2	17	10/2	5/0	10/2	5/0
3	16	9/1	5/0	9/1	5/0
4	24	10/4	9/3	10/4	9/3
5	6	6/1	6/1	6/1	6/1
6	5	3/1	3/1	5/3	5/3
7	11	10/3	5/1	11/3	5/1
8	21	4/3	3/2	4/3	3/2
9	8	4/1	4/1	4/1	4/1
10	14	3/1	3/1	5/2	5/2
11	19	10/6	9/5	11/7	9/5
12	9	4/1	4/1	10/5	6/3
13	15	6/2	6/2	10/2	1/0
14	8	9/4	8/2	9/4	7/2
15	15	8/1	8/1	8/1	8/1
16	11	10/2	5/0	10/2	5/0
17	13	15/5	9/3	16/5	9/3
18	25	6/1	3/0	10/1	5/0
19	20	11/3	6/1	16/8	7/1
20	13	3/1	3/3	5/2	4/2
21	27	12/6	7/3	14/8	7/3
22	22	2/0	2/0	2/0	2/0
23	10	7/3	7/3	7/3	4/2
24	17	3/0	3/0	4/0	4/0
25	22	10/3	5/0	10/3	5/0
26	11	3/0	3/0	3/0	3/0
27	8	3/1	3/1	3/1	3/1
28	11	10/2	9/1	12/4	10/3
29	8	8/2	6/0	8/2	2/0
30	14	3/0	3/0	3/0	3/0
31	15	10/8	9/6	11/9	8/7
32	8	11/5	6/2	14/9	9/5
33	10	4/1	4/1	4/1	4/1
34	9	5/2	4/2	8/2	6/2
35	13	3/1	3/1	6/2	2/2
Mean	14.7	7.2/2.4	5.4/1.5	8.4/3.1	5.5/1.7
Median	14	7/2	5/1	9/2	5/2
S.D.	6.0	3.8/2.1	2.3/1.6	4.0/2.6	2.5/1.7

** Numerator is the total number of hypotheses generated;
Denominator is the number of inappropriate hypotheses generated.

Fig. 6. Number of Hypotheses at End of Session With Deactivation

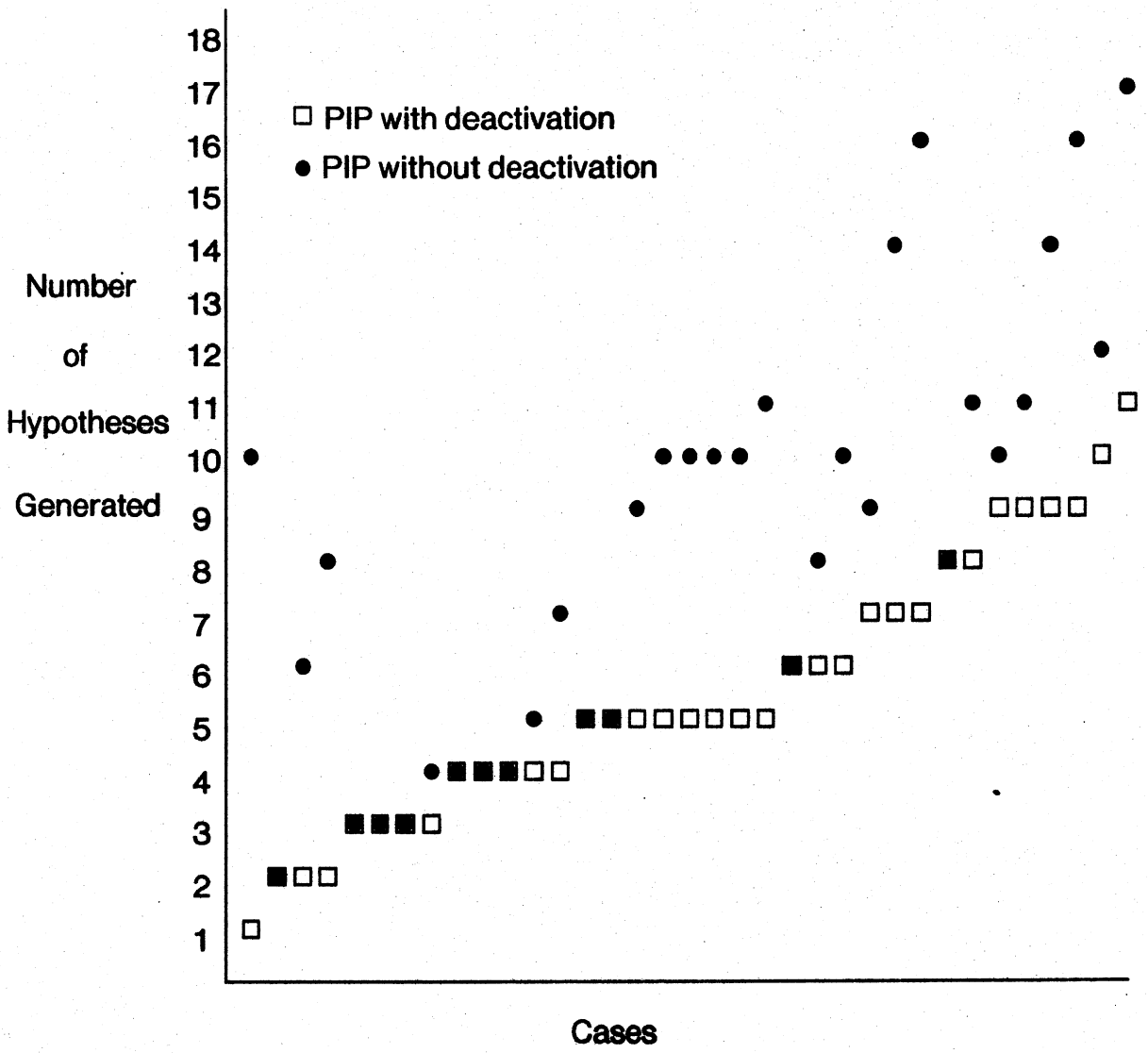
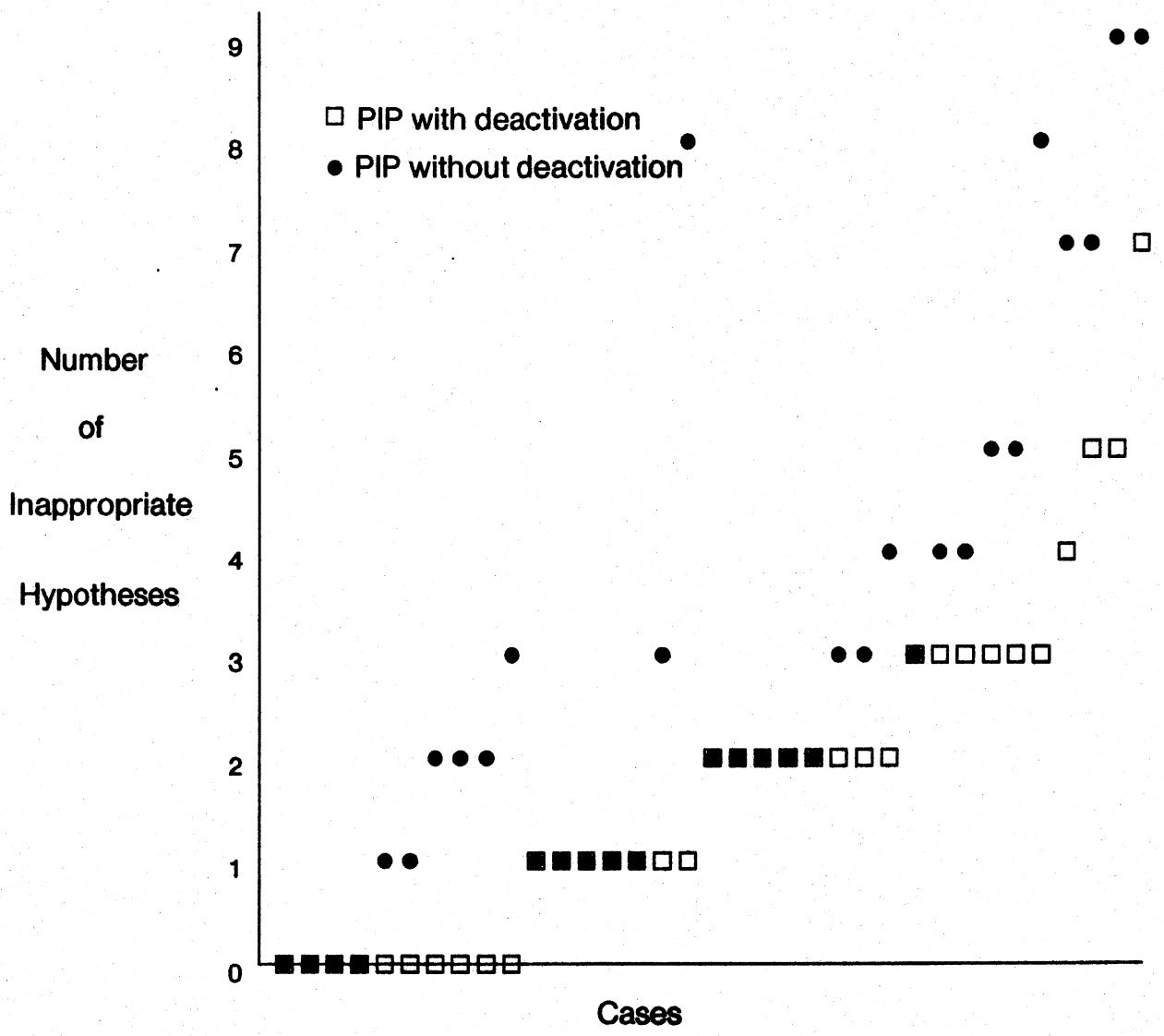


Fig. 7. Number of Inappropriate Hypotheses at End of Session With Deactivation



In PIP, it is highly desirable not to deactivate the correct hypotheses since deactivation via the *must-not-have* feature is an uncorrectable mistake for PIP, (i.e., once deactivated PIP can never reactivate the hypotheses). This did occur once, due to a stray finding, in case 35 causing the session to fail to conclude any hypothesis.

In conclusion, it appears that deactivation does decrease the number of hypotheses generated as well as the number of inappropriate ones, although these decreases are small compared to the number of hypotheses never generated (as compared to INTERNIST) via PIP's use of triggers. The birth defects domain appears to use the *must-have* and the *must-not-have* slots to a greater extent than other medical domains, possibly due to its syndromic nature. But since the birth defects domain also has more stray findings than most other medical domains, it may be desirable to have some method to reconsider a deactivated hypotheses if necessary, (i.e., if no other hypothesis proves to be correct).

5.1.3 The Disease Hierarchy and Hypothesis Generation

INTERNIST's use of a disease hierarchy is another method for controlling the number of hypotheses generated. The disease hierarchy constrains hypothesis generation by activating a superior syndrome node rather than several inferior nodes, hence there are fewer hypotheses to evaluate. In addition, the hierarchy could be useful in strategy selection by allowing the system to consider groups of similar syndromes as a single entity and possibly to choose strategies which support or eliminate entire groups rather than dealing with each syndrome individually. This aspect of the disease hierarchy will be investigated later. The desire presently is to determine to what degree this hierarchy decreases the total number of hypotheses and the number of inappropriate ones.

The data shows (see table III) that using the disease hierarchy does indeed reduce the number of hypotheses generated to a small extent (in these cases an average of 0.5 after the initial findings were entered and 0.6 at the end of the case). These reductions, as with those found with the use of deactivation, are small compared to those gained via the use of triggers.

From examining the non-leaf hypotheses that were generated by INTERNIST one sees that, with few exceptions, these hypotheses were those that explained very few of the findings, i.e. hypotheses with very low scores. The hierarchy basically prevented a single finding (or a few non-specific ones) from activating several similar syndromes each of which explained little else. This seems to be due to the fact that just a single finding associated with one inferior node and not with another will cause the inferiors to be evoked. This limited the usefulness of the hierarchy since by the time several findings were known about a group of syndromes, the leaf nodes were usually evoked.

5.1.4 The Separation of Scoring and Triggering Information

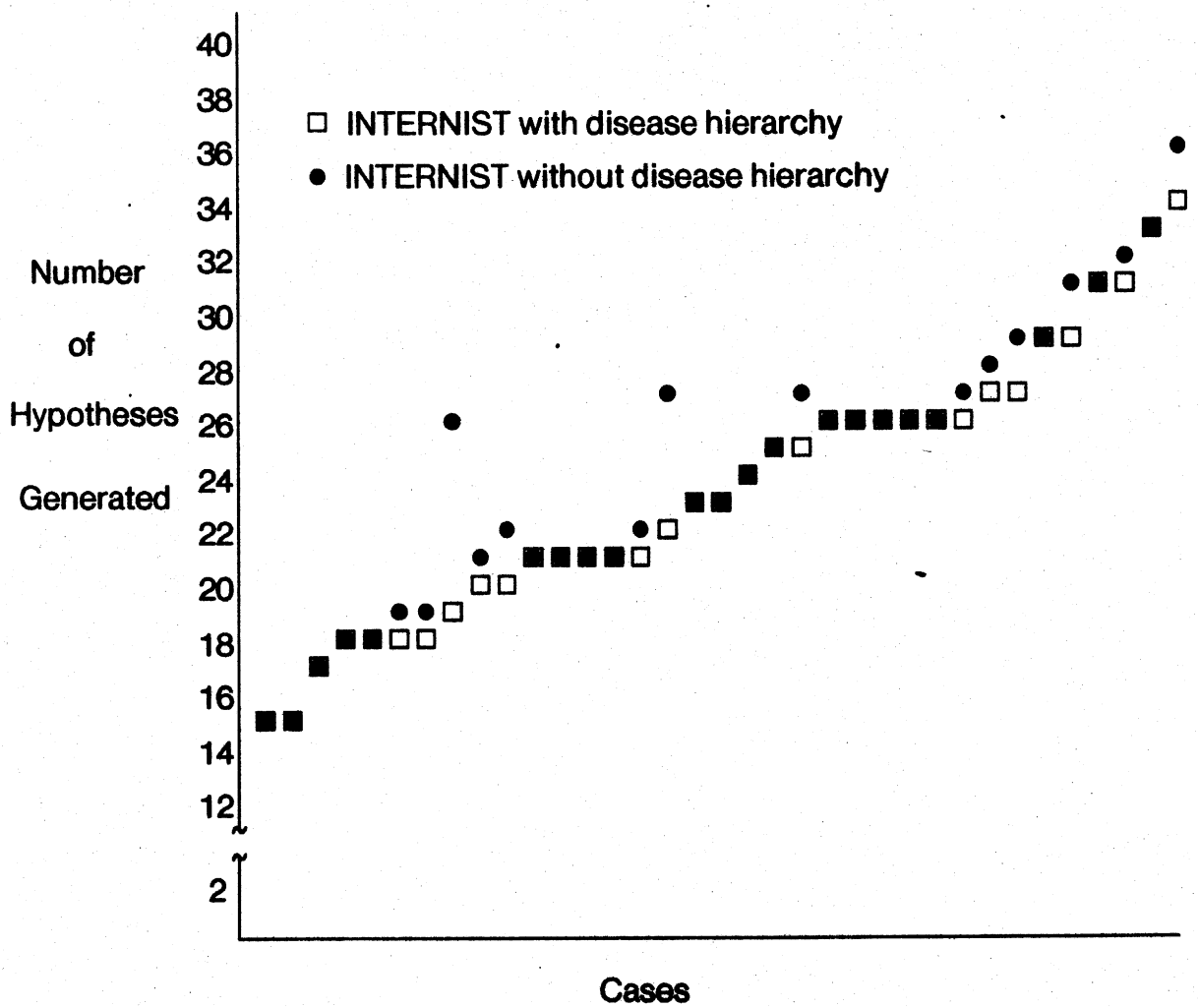
In both INTERNIST and PIP the presence of single findings is used to trigger the generation of hypotheses. In both cases only certain findings of a syndrome will cause it to be hypothesized. The only difference between the two methods of evoking hypotheses, if one ignores deactivation and the disease hierarchy, is that PIP has separated the evoking information from the scoring information and INTERNIST has not. Curiosity beckoned us to see if, by raising the threshold that the *evoking strength* was required to surpass in INTERNIST, one could duplicate the performance of PIP's triggers.

Before this comparison could be performed, the degree to which INTERNIST's evoking threshold for hypothesizing syndromes could be elevated without its performance deteriorating needed to be determined. The optimal evoking threshold for

Table III. Nonterminal Hypotheses Generation by INTERNIST

Case	Number of Initial Findings	Number of Hypotheses Generated After Initial Findings		At End of Session	
		Hierarchy		Hierarchy	
		Yes	No	Yes	No
1	26	31	31	31	31
2	17	21	21	21	21
3	16	24	24	24	24
4	24	27	29	32	34
5	6	21	22	23	24
6	5	18	19	24	25
7	11	21	21	21	21
8	21	18	18	20	20
9	8	26	26	26	26
10	14	26	26	29	29
11	19	25	27	25	27
12	9	15	15	16	16
13	15	34	36	39	39
14	8	18	19	21	22
15	15	31	32	32	33
16	11	21	21	21	21
17	13	29	29	29	29
18	25	26	26	26	26
19	20	26	27	26	27
20	13	26	26	26	26
21	27	29	27	29	31
22	22	23	23	23	23
23	10	25	25	28	28
24	17	22	27	22	27
25	22	23	23	23	23
26	11	15	15	15	15
27	8	20	21	20	21
28	11	27	19	34	36
29	8	18	18	22	22
30	14	17	17	17	17
31	15	33	33	33	33
32	8	20	22	26	26
33	10	26	26	26	26
34	9	21	21	23	23
35	13	19	26	27	27
Mean	14.3	23.4	23.9	25.1	25.7
Median	13	23	24	25	26
S.D.	6.0	4.9	5.0	5.2	5.4

Fig. 8. Nonterminal Hypotheses Generation by INTERNIST after Initial Findings



INTERNIST was found by entering the initial findings of thirty-five cases into INTERNIST using different values for the threshold (see table IV).

The results showed that when the threshold was set at zero, (i.e. any manifestation associated with a syndrome with an evoking strength of one or greater would cause a syndrome to be activated), all of the correct syndromes were hypothesized by the time all of the initial findings were entered. The same was true if the threshold was set at one or two. But if the threshold was set at three, (i.e. only manifestations with evoking strengths of four or five could cause a syndrome to be hypothesized), in ten out of the thirty-five cases the correct syndrome was not evoked. In addition, in three of those cases when the threshold was set lower the correct syndrome was not only evoked but was the leading hypothesis in either confirm or differentiate mode. Due to these results the evoking threshold was set at two for this comparison.

A comparison of INTERNIST using an evoking threshold of 2 and PIP not using deactivation shows that PIP still evokes fewer hypotheses after the initial findings have been entered; an average of 11.2 (standard deviation 4.1) for INTERNIST as compared to 7.2 (standard deviation 3.8) for PIP. This corresponds to a level of significance of less than 0.001 using the student's t. The difference in the number of inappropriate hypotheses, an average of 3.2, was also statistically significant difference ($p < 0.001$). These findings carried through to the end of the case. INTERNIST generating an average of 12.3 (standard deviation 4.6) hypotheses at the end of the case compared with 8.4 (standard deviation 4.1) by PIP. The average number of inappropriate hypotheses generated by the programs at the end of a case was 6.7 (standard deviation of 3.1) and 3.1 (standard deviation 2.6) for INTERNIST and PIP respectively.

Table IV. Hypotheses Generated by INTERNIST After Entering Initial Findings

Case	Number of Initial Findings	Number of Hypotheses Evoking Threshold 0	2	3	
1	26	31	23	4	*
2	17	21	12	4	
3	16	24	12	3	
4	24	27	12	4	*
5	6	21	13	2	*
6	5	18	6	3	
7	11	21	12	4	
8	21	18	9	1	
9	8	26	7	3	
10	14	26	8	2	
11	19	25	12	2	*
12	9	15	8	2	*
13	15	34	14	4	
14	8	18	10	5	
15	15	31	19	4	
16	11	21	12	4	
17	13	29	17	4	
18	25	26	12	6	
19	20	26	14	4	
20	13	26	5	1	
21	27	29	15	8	
22	22	23	13	3	
23	10	25	10	4	
24	17	22	7	1	
25	22	23	12	5	
26	11	15	5	1	
27	8	20	8	2	
28	11	27	18	1	*
29	8	18	11	2	*
30	14	17	6	2	
31	15	33	15	7	
32	8	20	12	3	*
33	10	26	7	2	
34	9	21	8	5	*
35	13	19	8	1	*
Mean	14.3	23.4	11.2	3.2	
Median	13	23	12	3	
S.D.	6.0	4.9	4.1	1.7	

* - Correct syndrome not evoked when evoking threshold = 3

Table V. Number of Hypotheses Generated

Case	Number of Initial Findings	Number of Hypotheses Generated ** INTERNIST * Initial	INTERNIST * Total	PIP Initial	PIP Total
1	26	23/11	23/11	17/7	17/7
2	17	12/5	12/5	10/2	10/2
3	16	12/4	13/5	9/1	9/1
4	24	12/4	17/9	10/4	10/4
5	6	13/5	15/7	6/1	6/1
6	5	6/2	6/2	3/1	5/3
7	11	12/5	12/5	10/3	11/3
8	21	9/8	12/11	4/3	4/3
9	8	7/5	7/5	4/1	4/1
10	14	8/4	12/7	3/1	5/2
11	19	12/8	14/10	10/6	11/7
12	9	8/2	9/4	4/1	10/5
13	15	14/9	20/15	6/2	10/2
14	8	10/6	11/7	9/4	9/4
15	15	19/9	22/10	8/1	8/1
16	11	12/4	12/4	10/2	10/2
17	13	17/9	17/9	15/5	16/5
18	25	12/4	14/5	6/1	10/1
19	20	14/7	14/7	11/3	16/8
20	13	5/2	5/2	3/1	5/2
21	27	15/7	15/7	12/6	14/8
22	22	13/9	13/9	2/0	2/0
23	10	10/6	14/10	7/3	7/3
24	17	7/3	7/3	3/0	4/0
25	22	12/5	12/5	10/3	10/3
26	11	5/1	5/1	3/0	3/0
27	8	8/3	8/4	3/1	3/1
28	11	18/7	19/11	10/2	12/4
29	8	11/5	12/5	8/2	8/2
30	14	6/2	6/2	3/0	3/0
31	15	15/13	15/13	10/8	11/9
32	8	12/7	12/7	11/5	14/9
33	10	7/5	7/5	4/1	4/1
34	9	8/4	9/5	5/2	8/2
35	13	8/5	11/7	3/1	6/2
Mean	14.3	11.2/5.6	12.3/6.7	7.2/2.4	8.4/3.1
Median	13	12/5	12/7	7/2	9/2
S.D.	6.0	4.1/2.7	4.6/3.3	3.8/2.1	4.1/2.6

* INTERNIST with the evoking threshold set at 2.

** Numerator is the total number of hypotheses generated;
Denominator is the number of inappropriate hypotheses generated.

Fig. 9. Number of Hypotheses Generated

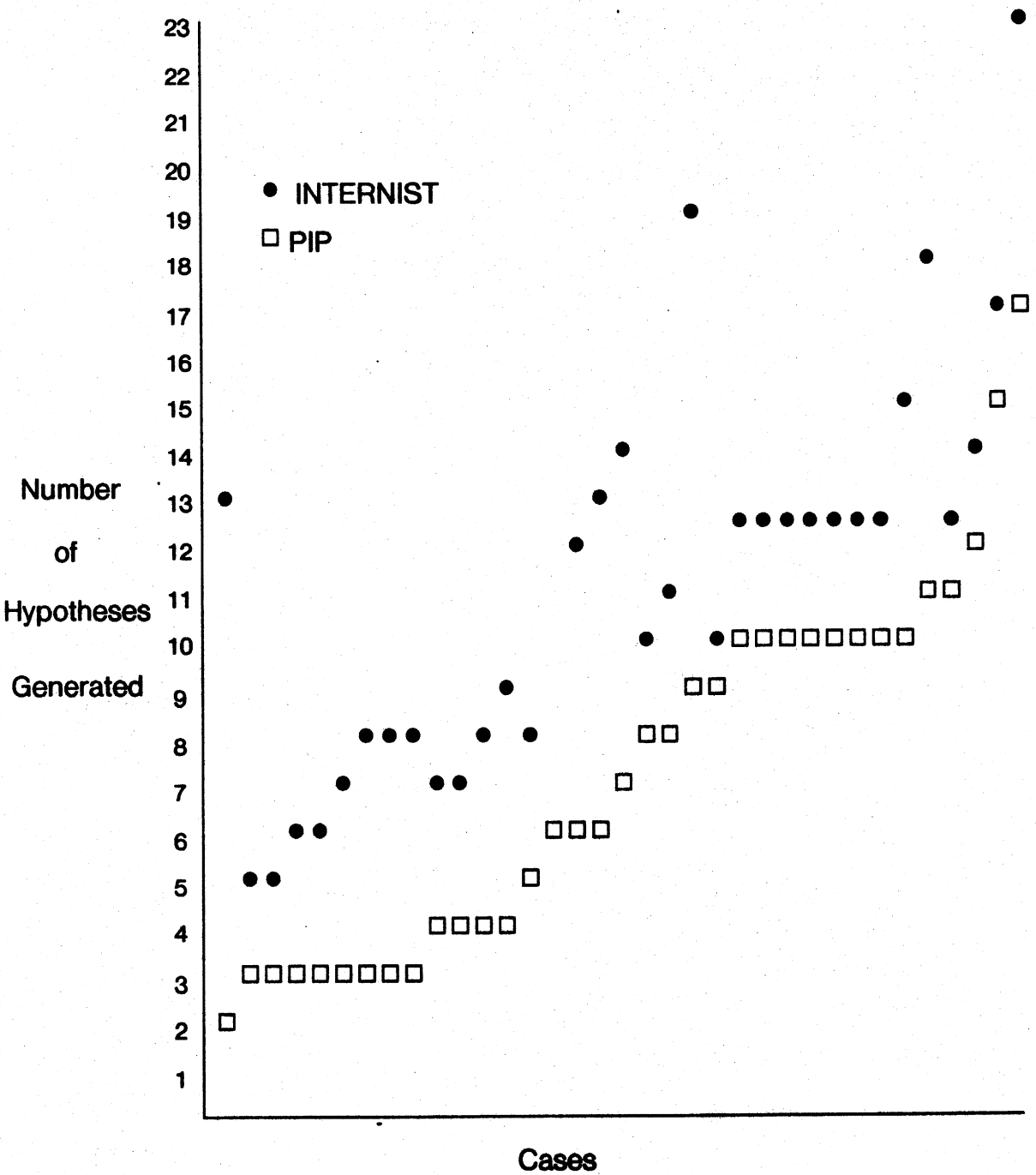
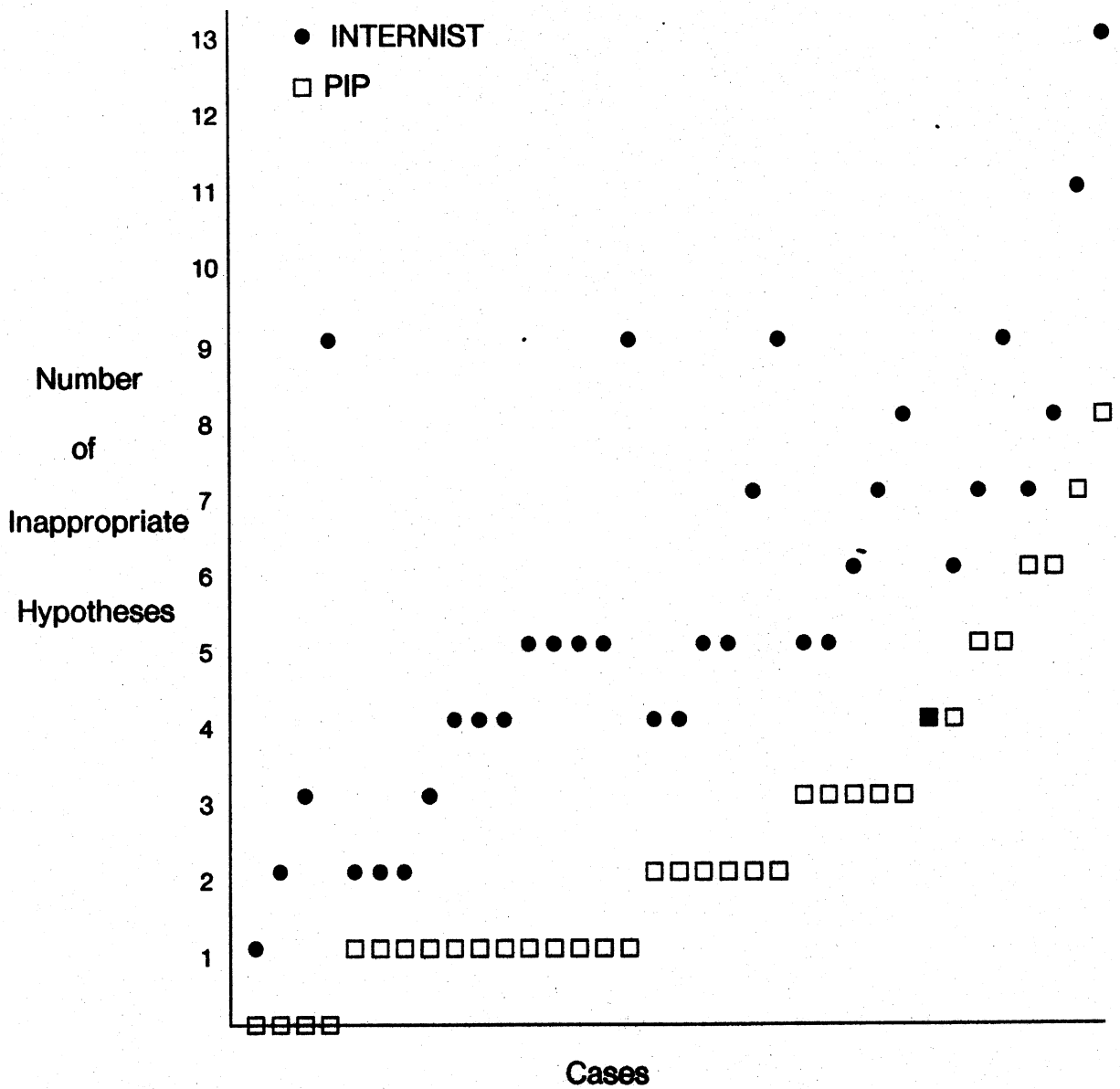


Fig. 10. Number of Inappropriate Hypotheses Generated



One can conclude from this data that separating information for scoring from the information for hypothesis generation does cause a reduction in the total number of hypotheses generated as well as a reduction in the number of inappropriate hypotheses generated.

5.1.5 Summary

It appears that PIP's triggering mechanism is an effective method for controlling the number of hypotheses generated while still hypothesizing the correct one early on in the diagnosis. Both PIP's deactivation of hypotheses and INTERNIST's disease hierarchy appear to have only minor effects on hypothesis generation. Separation of hypothesis generation information from scoring information in hypothesis generation appears to be beneficial.

5.2 Scoring

The scoring algorithm is the major method by which these programs determine the likelihood of a hypothesis being present. Hence a hypothesis' score should reflect the likelihood of that hypothesis' presence given the information known to the system. Since there is no reference likelihood for a syndrome being present given a set of findings, it is difficult to get absolute measures of the scoring algorithms' performance. But, since the correct syndrome is known, one test of the scoring algorithms is to compare the rankings of the syndromes after a set of findings has been entered. If one of the algorithms consistently ranks the correct syndrome higher than the other, this is a good indication that the former algorithm is better.

This test was done using the initial findings of each case as the set of findings. INTERNIST's hypothesis generation was used for both systems in order to remove any

effect due to generation of different hypotheses.[20] The results of this test are shown in table VI. The rank denotes the position of the correct syndrome in the list of the active hypotheses ordered according to score (highest score ranked one).

The results of this comparison do not show any statistically significant difference between the scoring algorithms. In both systems the correct syndrome's score ranked first in most cases. In those cases where in one system the correct hypothesis' score was not highest ranked, it was likely that the same was true in the other system.

The scoring algorithms were also tested using different sets of findings, for example the initial findings with all of the high *import* findings removed or with random findings removed or all laboratory data removed. The results from these tests also showed no statistical differences between the scoring algorithms.

This result was rather unexpected since PIP does not use any information equivalent to INTERNIST's *import* value and PIP has a normalizing scheme that would appear to penalize those hypotheses with greater numbers of associated findings. One reason that the normalization might not have been as detrimental as one would expect is due to the fact that these syndromes have a fair number of findings associated with them, the maximum possible score usually varied only 30 or 40 percent for either the skeletal syndromes or the endocrine syndromes.[21]

---

20. Although this was done, it had little effect on the results since in both systems the few leading hypotheses were usually the same.

21. The endocrine syndromes in general had ten to fifteen fewer associated findings, but it was seldom the case that both types of defects were among the leading hypotheses.

Table VI. The Rank of the Correct Hypothesis after Initial Findings Entered

Case	Number of Initial Findings	INTERNIST Rank	PIP Rank
1	26	6	3
2	17	1	1
3	16	1	1
4	24	1	1
5	6	3	5
6	5	1	1
7	11	1	1
8	21	1	1
9	8	1	1
10	14	1	1
11	19	1	2
12	9	4	5
13	15	4	6
14	8	2	1
15	15	1	1
16	11	1	1
17	13	1	1
18	25	1	2
19	20	1	1
20	13	1	1
21	27	1	1
22	22	1	1
23	10	2	2
24	17	1	1
25	22	1	1
26	11	1	1
27	8	1	1
28	11	6	5
29	8	5	2
30	14	1	1
31	15	2	1
32	8	4	5
33	10	1	1
34	9	1	1
35	13	6	3
Mean	14.7	1.9	1.8
Median	14	1	1
S.D.	6.0	1.6	1.5

To determine the reason that PIP's algorithm performed as well as INTERNIST's without the use of *import* information a test was done to ascertain the importance of the *import* to INTERNIST's algorithm. The cases were entered into an INTERNIST which was identical except that for all of the findings the *import* was set to the same value. The value that was used was equivalent to an *import* of 2.5 for scoring purposes. The ranks of the correct hypotheses after the initial findings were entered were again determined. The results of this test show that, with the *import* set at a constant value rather than its original value, in seven cases the correct syndrome was ranked lower, in four cases it was ranked higher, and in the rest of the cases there was no change. All of the changes in rank that occurred were only a change of one position (i.e. from fourth to third or first to second). These results indicate that the use of the *import* (or lack thereof) does not have a large effect on the hypotheses' relative scores in INTERNIST.[22] Hence one would not expect PIP's scoring algorithm to be unduly impaired for not using the information contained in INTERNIST's *import* values.

In summary, it appears that although there are several differences between PIP and INTERNIST's scoring algorithm there is no major difference in their performance in this domain. Any minor differences were not detected (probably due to the number of cases). In addition, it appears that in the scoring algorithm the use of a weighted value for each finding not normally associated with a hypothesis (i.e. the *import*) is not critical, at least not in the birth defects domain. It should be noted that due to the domain the effects of INTERNIST's partitioning algorithm in ranking of the hypotheses did not come

---

22. It is possible that the relatively large number of unexplained findings in these syndromes (as compared to diseases in internal medicine) causes an averaging effect which would reduce the error caused by not using the *import*. This error is that caused by assuming that all findings not explicitly mentioned in the disease description have equal importance in the diagnosis of a syndrome.

into play here.[23]

5.3 Concluding Hypotheses

Deciding when to conclude that a syndrome is present is a difficult problem. One must justify the cost of obtaining additional information against the possible benefits to the patient of a more precise and certain diagnosis. This decision may indeed require additional knowledge (e.g. about therapy, resource availability, etc.) [25].

PIP and INTERNIST take a simple solution to the problem of when to conclude that a syndrome is present. They use the scores of the hypotheses and a predetermined threshold value. As stated before, PIP requires that a hypothesis' score exceeds the threshold while INTERNIST requires that the difference between the scores of the top two hypotheses exceeds the threshold. To evaluate these two methods, the hypotheses concluded by each system were examined.

The results showed that INTERNIST concluded the correct syndrome in 34 of the 35 cases and in one case no syndrome was concluded. PIP, on the other hand, concluded the correct syndrome in thirty cases, failed to conclude any syndrome in four cases, and concluded an incorrect syndrome in two cases.[24]

The differences in the number of incorrect syndromes concluded and the number of cases where no syndromes were concluded, although not statistically significant, can be explained by the differences in the algorithms. In both of the cases where PIP concluded an incorrect syndrome there was another hypothesis which had a

---

23. If multiple syndromes were present INTERNIST's partitioning algorithm could increase the rank of one of the correct hypothesis and hence possibly cause INTERNIST to conclude that hypothesis sooner.
24. In one case PIP concluded both a correct and an incorrect syndrome.

score that was very close. INTERNIST would not have concluded the syndrome at that point since it uses the difference between the scores of the leading hypotheses and not the actual magnitude of the leading hypothesis' score.

In the three cases where PIP failed to conclude any syndrome but INTERNIST did conclude the correct syndrome, PIP was pursuing the correct hypothesis but due to stray and/or absent findings (quite common in birth defects) the correct hypothesis' score was not greater than the threshold for confirmation. In INTERNIST, although the magnitude of the correct hypothesis was not very large, the difference between its score and those of the other hypotheses was great enough to conclude the correct hypotheses.

One of the four cases (case 5) that PIP failed to conclude a syndrome was the same case for which INTERNIST failed to conclude any syndrome. The reason for the failure of these systems to conclude the correct syndrome in cases when a physician was able to do so appears to be due to the systems' lack of knowledge about therapy. The child in this case was diagnosed as having thyroid dysgenesis. Both INTERNIST and PIP were pursuing causes of congenital hypothyroidism (of which thyroid dysgenesis is the most common type) but due to lack of laboratory results they were not able to distinguish between the different possible causes. The physician realized that the therapy for all of these syndromes is the same, hence found no need to request costly lab tests to differentiate between the different possibilities. This is not really a deficiency in the algorithms for concluding hypotheses, since one can argue that the physician did not rule out the other rare causes of congenital hypothyroidism, but rather a deficiency in the algorithm that determines when to terminate the diagnostic session.

In summary, it appears that considering the score of the leading hypothesis alone is not as effective in determining when to confirm that the hypothesis is present as considering its score relative to the score of the other hypotheses.

5.4 Diagnostic Strategy and Question Selection

The purpose of the diagnostic strategy is to direct the acquisition of information via question selection so that the system will focus on (and conclude)[25] the correct diagnosis quickly and in a coherent manner. Determining what is a *coherent manner* is a difficult problem. Among physicians there is a large variability in diagnostic style. Since these matters are subject to personal preferences, no comparison of this aspect of the diagnostic strategy and question selection will be done.

The diagnostic strategy and question selection of the two systems differ in several ways. PIP uses the ordered list of findings of the leading hypothesis to select the next question (asking the first unknown finding in the list). This permits PIP to encode the equivalent of INTERNIST's *ruleout* and *confirm* strategies, but not the *discriminate* strategy.[26] The ordering of the finding list in PIP may encode more information than the simple selection rules of INTERNIST. Also PIP's diagnostic strategy calls for reevaluation after each question is asked whereas INTERNIST inquires about several findings before reevaluating the situation. In addition, the algorithm to conclude hypotheses, a part of the diagnostic strategy, differ in the manner described in the last subsection.

To determine the ability of the diagnostic strategies for each system to focus on and conclude the correct syndrome a comparison of the number of questions each system asked before concluding the correct hypothesis was done.[27] The results of the comparison are shown in table VII and figure 11.

---

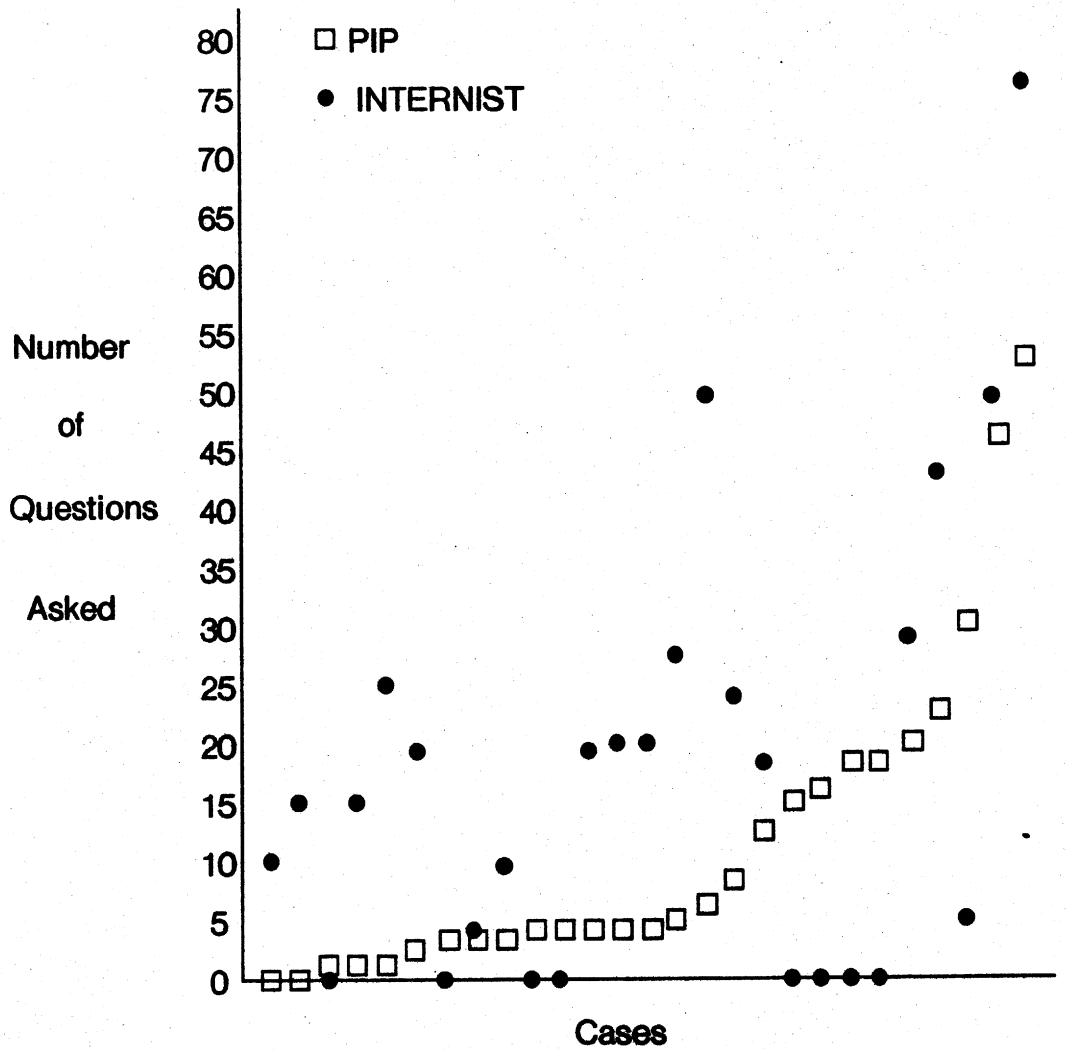
25. One possible option at any point is to conclude that a hypothesized syndrome is actually present.
26. Of course, with only one list PIP cannot dynamically switch between *ruleout* and *confirm* modes, but it can first attempt to ruleout a syndrome before attempting to confirm it.
27. Although the last section dealt with the algorithms for concluding hypotheses, the first comparison includes these algorithms since they are an integral part of the strategy.

Table VII. Number of Questions Required to Conclude the Correct Hypothesis

Case	Number of Initial Findings	Number of Questions Asked INTERNIST	PIP
1	26	49	46
2	17	0	14
3	16	18	12
4	24	10	0
5	6	*	*
6	5	9	3
7	11	0	18
8	21	15	0
9	8	0	1
10	14	20	4
11	19	19	4
12	9	31	*
13	15	58	*
14	8	43	22
15	15	15	1
16	11	0	18
17	13	0	16
18	25	24	8
19	20	5	29
20	13	4	3
21	27	19	2
22	22	0	0
23	10	78	*
24	17	0	0
25	22	0	4
26	11	0	0
27	8	0	4
28	11	28	5
29	8	49	6
30	14	0	0
31	15	20	4
32	8	76	53
33	10	0	3
34	9	25	1
35	13	29	20
Mean **	14.3	15.4	9.7
Median **	13	10	4
S.D. **	6.0	18.6	13.2

* -- Correct hypothesis was not concluded.
** -- Only cases in which data is present for both systems used in calculation.

Fig. 11. Number of Questions Required to Conclude the Correct Hypothesis



The results showed a large variation between cases in the number of questions asked before the correct hypotheses was concluded. This might be explained by the observation that (for both systems) when the system was focused on the correct hypothesis the number of question asked before concluding the syndrome was rarely very large. But when the systems were not pursuing the correct path it often required the investigation of several incorrect hypotheses before discovering the correct one. The investigation of these incorrect hypotheses usually required the asking of many questions.

The analysis of the data shows that PIP requires fewer findings to conclude the correct syndrome, an average of 5.6 fewer. Part of this difference may be due to the criteria for concluding hypotheses, INTERNIST requiring that no other hypotheses be near by whereas PIP does not.

To attempt to remove this factor, the number of questions required to make the correct hypotheses the leading hypotheses was determined.[28] This data is shown in table VIII and figure 12.

The results of this are similar to the results of the previous comparison. This indicates that more than the difference in methods for concluding syndromes is required to explain the differences in speed of focusing on the correct syndrome. There are two basic differences in the algorithms that could account for the discrepancy. One, the question selection algorithm is different in each system. Two, PIP reevaluates the situation (i.e. triggers new hypotheses, deactivates others, rescores and ranks the hypotheses) after each question is asked whereas INTERNIST asks a group of questions

---

28. The point at which the correct hypotheses was considered the *leading hypotheses* was defined as the point at which it became the number one ranked hypothesis and remained so until concluded.

before reevaluating.

To determine the effect this second difference has on the diagnosis, the cases were run on an identical INTERNIST except that only one question was asked before reevaluating the situation.[29]

The results of this are also shown in table VIII and figure 12.

Table VIII. Number of Questions Required to Pursue the Correct Hypothesis

	INTERNIST		
Case *	Normal	Single **	PIP
1	40	26	27
5	4	1	0
11	0	0	3
12	17	0	19
13	36	46	19
14	14	11	0
18	0	0	1
23	4	1	24
28	9	12	3
29	15	9	4
31	5	1	0
32	44	22	11
35	24	24	9
Mean	16.3	11.9	9.2
Median	14	9	4
S.D.	15.3	14.1	9.8

* Cases where neither system required any questions to pursue the correct hypothesis were removed.

** Only one question asked before re-evaluation

29. If the situation is the same, i.e. the mode and the ranking of the hypotheses (the relevant part) is the same, then the question selection would proceed as if the reevaluation did not occur.

Fig. 12. Number of Questions Required to Pursue the Correct Hypothesis

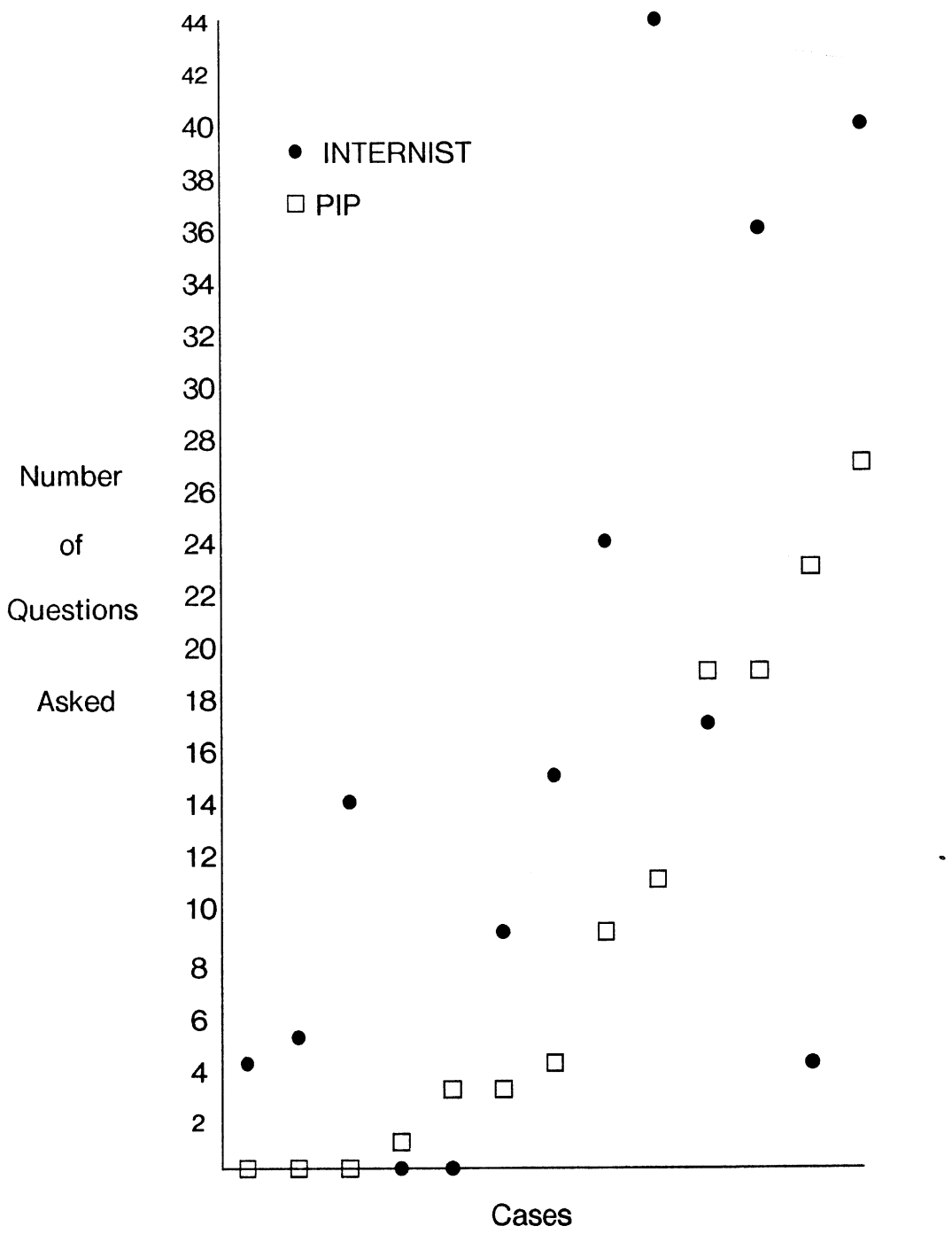
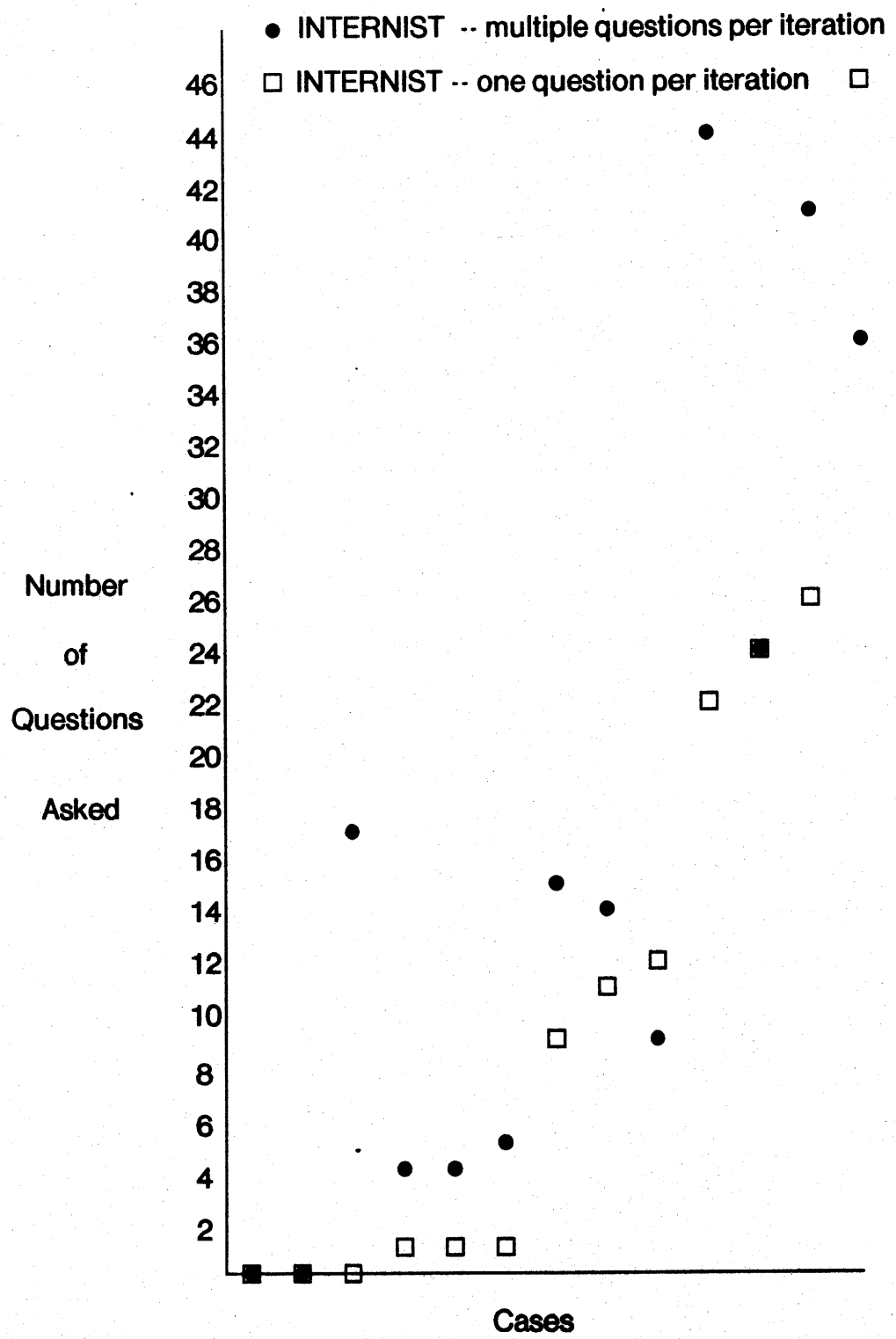


Fig. 13. Number of Questions Required to Pursue the Correct Hypothesis



By asking only one question before reevaluation the total number of questions asked in order to pursue the correct syndrome was reduced to a mean of 11.9 (standard deviations of 14.1). This was a statistically significant difference ($p < 0.05$) from the previous value. In addition, the difference between this result and the number of questions asked by PIP was not statistically different ($p > 0.05$). Hence the focusing power of each algorithm seems about equal if one evaluates the situation with the same frequency. This might indicate that the added knowledge embedded in PIP's ordered finding lists are offset by INTERNIST's *discriminate* strategy.

5.4.1 The Discrimination Strategy

One of the differences between INTERNIST's diagnostic strategies and PIP's is, as stated above, INTERNIST's use of the *discriminate* strategy. This strategy attempts to widen the difference between the top two hypotheses by asking about findings which will increase one hypothesis' score while decreasing that of the other.[30] This strategy requires more computation time (or memory) than do the *ruleout* and *confirm* strategies and it does not allow the physicians to easily alter the order in which the findings are asked.[31] Hence, it is of interest to determine the value of this strategy. To do this, the cases were entered into two identical INTERNIST systems except that in one system the *discriminate* mode was replaced by *ruleout* mode and in the other system the *discriminate* mode was replaced by *confirm* mode. The results of this comparison are shown in table IX and figure 14.

---

30. One could argue that, since PIP does not require a separation between the two top hypotheses for confirmation of the leading hypotheses, the discriminate strategy is not useful to PIP. But the results from the last section indicates that PIP should consider separation between the two top hypotheses and hence it is of value to determine the usefulness of this strategy.

31. With the *ruleout* or *confirm* mode the physician can construct an ordered list of findings for each mode that will (hopefully) optimize question selection for those diseases given the intent of the mode. This can not easily be done in discriminate mode since a different list would be needed for each pair of diseases.

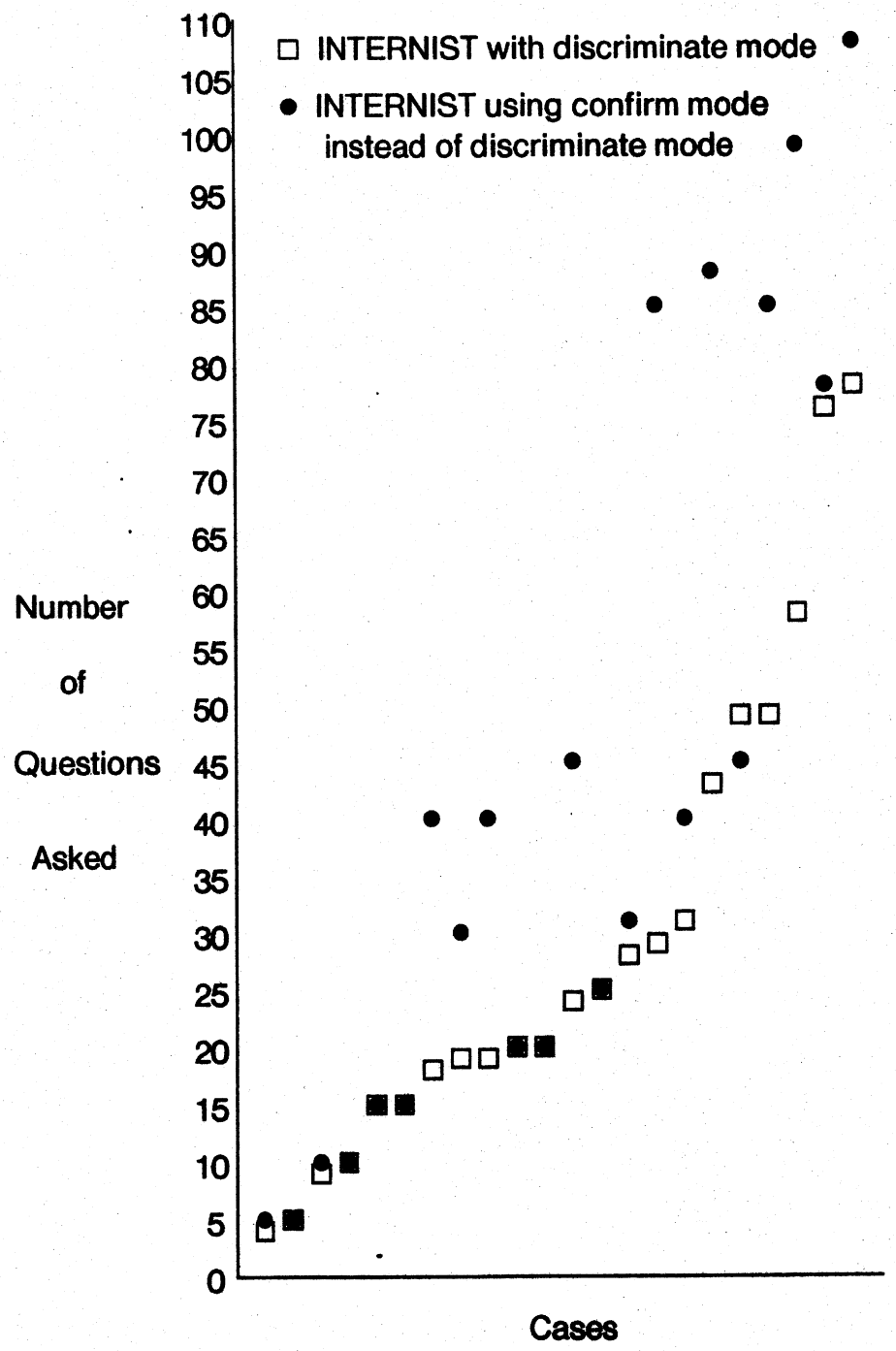
Table IX. INTERNIST With and Without the Discriminate Strategy

Case **	Number of Questions Asked When in Discriminate Mode		
	Strategy Used		
	Discriminate	Ruleout	Confirm
1	49	57	85
3	18	31	40
4	10	10	10 *
6	9	12	10
8	15	15	15 *
10	20	20	20 *
11	19	33	30
12	31	46	40
13	58	73	99
14	43	51	88
15	15	15	15 *
18	24	29	45
19	5	5	5 *
20	4	7	5
21	19	53	40
23	78	45	108
28	28	31	31
29	49	52	45
31	20	20	20 *
32	76	74	78
34	25	25	25 *
35	29	90	85
Mean	29.3	36.1	42.7
Median	22	31	35.5
S.D.	21.1	23.7	32.7

* Discrimination strategy (or replacement strategy) was not used.
** Cases where no questions were asked were removed.

The results show that INTERNIST asks an average of 6.8 more questions when the *ruleout* strategy is used in place of the *discriminate* strategy. An average of 13.4 additional questions are asked when the *confirm* strategy is used in place of the *discriminate* strategy. These are statistically significant differences ($p < 0.05$ for the *ruleout* strategy and $p < 0.005$ for the *confirm* strategy). This indicates that this strategy

Fig. 14. INTERNIST With and Without the Discriminate Strategy



is useful if the system requires a differential between the top two hypotheses' scores before concluding that the leading hypothesis is present.

5.5 Summary

These tests have shown that PIP's triggering mechanism generates fewer hypotheses and fewer unreasonable ones than does INTERNIST's evoking strategy. Both the INTERNIST's disease hierarchy and PIP's deactivation of hypotheses cause only small reductions in the number of hypotheses. Also separation of triggering information from scoring information appears to have some beneficial effects in hypothesis generation. Although there are several differences between scoring algorithms of PIP and INTERNIST, there were no significant differences found between them in their performance. INTERNIST's mechanism for concluding a hypotheses, using the difference in the scores between the top two hypotheses, proved to be superior to PIP's threshold mechanism. PIP's question selection algorithm was able to focus on the correct hypothesis quicker than was INTERNIST's. This appeared to be due, at least in part, to INTERNIST asking groups of questions before reevaluating the situation. Also INTERNIST's *discriminate* strategy appears to speed its diagnosis.

6. Suggestions for Improvements

The comparison of PIP and INTERNIST presented in the last chapter identified several of the strengths and weaknesses of the respective systems. This chapter will suggest improvements to such pseudo-Bayesian hypothesis driven systems. These suggestions were inspired by the experience gained in constructing the databases for the two systems as well as the results of the comparison. Although it is assumed in this chapter that the birth defects domain is the intended problem domain for the diagnostic system, most of the suggestions presented here are expected to be valid in other medical problem domains.

The ideas presented in this chapter are not intended as a general solution to medical diagnosis, or a wholly new approach to the problem, this is an attempt to improve upon the techniques of PIP and INTERNIST (beyond what can be done by fine tuning them for use in this particular domain) while keeping to their pseudo-Bayesian approach. Of course, this means that some basic flaws in the PIP and INTERNIST approaches will not be addressed here.

As shown in the comparison, both PIP and INTERNIST do a fairly good job of diagnosing the birth defects. Most of the changes suggested will be aimed at diagnosing that fraction of the situations with which these systems did poorly, while still diagnosing the rest of the situations as well as PIP and INTERNIST did. Attempting to squeeze the last drops of performance out of any technique usually requires a great deal of effort for modest improvements.[32]

---

32. In the AI community (and elsewhere) this is referred to as the 80-20 phenomenon, i.e. it takes 20 percent of the effect to get 80 percent of the performance. The other 20 percent will take 80 percent of the effort.

The philosophy guiding the ideas presented here has two basic components, completeness and flexibility. Any diagnostic system should be able to easily represent knowledge about the problem domain and about solving problems in that domain. Also, since no simple algorithm will be able to cover all of the possible situations, the system should be flexible enough to allow changes in the control flow caused by either information in the database or from the user. The following sections describe suggestions for implementing this philosophy for the representation of knowledge and the various parts of a diagnostic algorithm.

6.1 Representation of Knowledge

One of the problems found in both PIP and INTERNIST was the lack of ability to represent the needed knowledge in an efficient manner within the constructs provided. It was often noted, while constructing the database for these systems, that knowledge the physicians had was very difficult or impossible to encode in the system. Even simple relationships were hard to define.

In INTERNIST, for example, each finding is represented as a LISP atom with associated properties. Any time a variation in a finding is desired it must be represented as a new finding and its associations explicitly stated. So for the finding "short stature" where the time of onset is important, several finding must be created and their relationships asserted. The association between findings must be stated for each finding separately. To state that "long" is the opposite of "short" in all findings referring to length, one must state for each finding pair that they are opposites.

In PIP, the situation is somewhat better in that a finding can have many items associated with it, each with several mutually exclusive values. So some of these relationships only need be specified once, i.e. the item "length" can be created with values: long, short and normal, or "short stature" can have associated item "onset"

which can have numerous values. But it is still not possible to conveniently encode all of the knowledge that is known by the physician. For example, it is difficult to encode the fact that a normal level of a substance (e.g. normal serum sodium level) should not decrease the score of those hypotheses that do not explicitly mention the finding (one does not want to have to list all of the normal findings that are expected with a disease or syndrome). It is also difficult to represent the relationships between the different items in a finding. This can cause needless questions to be asked. For example, one can state that the finding of edema is not recurrent and PIP will still inquire about the frequency of recurrence.

PIP's frame representation could have captured more of the clinical knowledge but the implementation did not include many of the features described by Minsky [22] which give the frame representation much of its flexibility. The algorithm also restricts the types of values the slots may take. This further restricts the type of knowledge that can be represented. Because the implementation was not as general as one might like, information that could have been put in the database was dispersed into the algorithm and hence created difficulties when trying to alter the code. For example, the default values of slots were placed in the code rather than in the database, hence changing a default value of a slot (because the default assumptions are different in the new domain) requires searching through the programs for the appropriate line of code.

The control structures of both of the systems are even more rigid and difficult to alter. In INTERNIST there is no way to affect the flow of control of the algorithm. In PIP, only slight control is possible via *semi-activation* with the *differential diagnosis* slot and the causal links and by allowing the user to enter a different finding from the one PIP chooses to inquire about. So, if a user or the database designer wishes to change the focus of the program due to some particular circumstance, there is no direct way of doing so. This is certainly undesirable especially since many of the problems are specific to one syndrome or a few special cases.

Ideally, the desired representation should be able to capture all of the knowledge that the physician has that could be used in diagnosis by this method. The representation should be able to capture the causal, associative inter-relationships between objects (syndromes, states, findings, etc.) in a way that facilitates reasoning about these relationships. It should be able to represent the temporal relationships as well as the static ones. Abstraction to a level convenient for making inferences should be supported.

The following sections outline some suggestions for accomplishing these goals. First a comment on implementing these suggestions. To accomplish these goals the use of a general knowledge representation language (e.g., KRL, FRL, OWL, or one designed by the implementor etc.) would be advisable. PIP's and INTERNIST's representations fail to include many of the features found in knowledge representation languages (e.g., inheritance, defaults, procedural attachments, etc.). When the need for these features arose they were implemented as special cases and embedded in the program. This leads to the difficulties in expanding and modifying the system, as noted above.

6.1.1 Decomposition of Medical Knowledge

Since these systems are pseudo-Bayesian in principle, it would be reasonable to believe that one wants to decompose the medical knowledge into events (findings) whose presence or absence is known and events (syndromes) whose presence is to be determined, all of the syndromes being mutually exclusive. This was done in some of the early Bayesian schemes [15, 21]. But as the number of syndromes and findings increases, the importance of the inter-relationships between findings and between syndromes increases. Both PIP and INTERNIST introduced causal links to help capture some of these inter-relationships. PIP introduced clinical and physiological states to enable the representation to group commonly occurring clusters of findings.

The decomposition of medical knowledge into findings, states, and syndromes as done by PIP (or INTERNIST) has two basic problems. First, PIP uses each entity in different ways, but there is no consistency in the way in which these different entities are used by physicians, as noted by Feinstein[12]:

> The main contemporary impediment to detailed specificity in physical examination is the failure of many clinicians to distinguish the three different intellectual disciplines — description, designation, and diagnosis... Designatory or diagnostic gestalts are used instead of descriptions when the findings of physical examination are reported in such terms as *opening snap of mitral valve, pericardial friction rub*,...

Feinstein's descriptions, designations, and diagnoses are analogous to the findings, clinical states, and syndromes in PIP.

The ability to treat clinical states and syndromes in the same manner as findings is needed. This dual perspective of the states and syndromes is useful in several ways. For the constructor of the knowledge base it eases the task of incorporating knowledge about these entities into the knowledge base. From the physician's point of view, it facilitates the entry of data since it allows known states and syndromes to be directly entered as if they were findings without requiring an exhaustive listing of their findings.

Another problem that occurs with clinical and physiological states is known as the "X" phenomenon (described by Rubin [35]). The "X" phenomenon occurs when the presentation of a clinical state varies depending on the cause. There is no easy way to represent this knowledge in PIP without removing the clinical state from the database and placing the appropriate findings in the appropriate syndrome (or creating two clinical states). A better way to handle this problem is to have specializers associated with the clinical and physiological states just as PIP does for findings (the specializers for PIP's findings are called items). The specializers for states would be determined by a logical combinations of findings rather than by having the user enter their value. For

example, the presence of the findings mental retardation and low serum T3 might cause the severity of hypothyroidism to be set to moderate. The strengths of the links between states and syndromes (or other states) would depend on which of the specializers were present. This is analogous to the way, in PIP, that the weight associated with a finding is dependent on the values of the items.

6.1.2 Representing Clinical Situations

One problem that was noted in the development of PIP's knowledge base was that many of the restrictions placed on the possible values of the slots in the frames interfered with the ability to represent the medical knowledge. For example triggers in PIP can only be single findings but the value of the *must-not-have* slot can be logical combinations of findings. Neither of these slots can have the presence or absence of a clinical state as a value. There are times when one would like to use logical combinations of findings or clinical states as triggers.

To overcome this problem I propose removing the restrictions on the values of slots and using a uniform representation for the values, called a *situation*. A *situation* should be able to contain combinations of findings, states, syndromes, and other information which could be relevant to the diagnosis (e.g., stage of the diagnosis). This allows the system developer to represent the clinical knowledge more accurately by not forcing it into a rigid predetermined form.

6.1.3 Representation of Non-Medical Knowledge

Non-medical knowledge, (e.g., knowledge about time, fluids, pressures, etc.) should be represented using the same knowledge representation as the medical knowledge. This, along with procedural attachments, permits one to place much of the information encoded in the program in PIP and INTERNIST in the database, making it much easier to modify the database and the program.

6.1.4 Representation of Time

One of the deficiencies of both PIP and INTERNIST is their lack of ability to adequately capture temporal knowledge. Temporal knowledge is an important source of information in the diagnostic process. Often knowing the time of onset of a finding can greatly reduce the number of hypotheses generated to explain it. INTERNIST's only way to represent time is to create several findings one for each possible time of onset, duration, or temporal relation to other findings, i.e. *short-stature-with-onset-at-birth*, *short-stature-with-onset-during-childhood*, *short-stature-with-onset-during-adolescence*, etc. The inter-relationships between all of these finding would have to be specified also. This would, of course, lead to great proliferation in the number of findings if this were done for a large proportion of the original finding set.[33]

In PIP, time was represented as an item in a finding. Originally, *now*, *near past*, *distant past*, *near future*, and *distant future* were the possible values of the time item. This was changed to an absolute scale since in congenital defects (and pediatrics in general) the age of the patient at onset of the findings is frequently more important than the order of occurrence of the findings. The possible values included: prenatal, birth, neonatal, infancy, early childhood, late childhood, adolescence, adulthood. This proved adequate for most situations, but not all. A finer grain was sometimes needed, but increasing the number of possible values was impractical since this would cause lengthy OR statements for most of the findings. In addition, the relative order between findings is sometimes important and not easily represented. Lastly, PIP does not provide a way to enter uncertainty about the temporal knowledge. This is often required especially when entering historical data.

---

33. This was done in INTERNIST only when it was deemed important enough for the syndrome to justify creating more findings.

What is needed is a representation of time which attempts to solve the problems listed above without increasing complexity too greatly. A scheme similar to Kenneth Kahn's system [19] (although possibly simplified) might be adequate. Each finding (syndromes and states too) could have an associated fuzzy interval (it could be an item in the PIP formalism). This fuzzy interval would represent the onset and duration of the finding. In addition before-after chains could be associated with the finding. The before-after chains would be used to specify that a finding occurs during, before, or after the onset or cessation of another finding, syndrome or state. They can specify the amount of time between the findings, that time being a fuzzy interval. This allows the order of occurrence of the findings to be explicitly denoted. Associated with this representation of time would be a "time expert". This is a pattern matcher which can compare fuzzy intervals (and before-after chains) to see whether they are equivalent.

6.2 The Algorithm

The following sections outline improvements in different parts of the algorithms discussed in the comparison.

6.2.1 Hypothesis Generation

The above comparison has shown that the methods used to restrict the number of hypotheses generated in both PIP and INTERNIST fail to achieve optimal performance. They both generate hypotheses that are not reasonable given the situation.

PIP's use of triggers to activate certain syndromes, even though others could explain some of the findings, has great appeal, but its use of only single findings as triggers is not sufficient. The problem is that in certain situations one may want a finding to trigger a syndrome and in different situations one may not. Analysis of hypothesis generation by physicians has shown that pairs and combinations of findings caused

triggering of hypotheses [20, 23]. PIP does not allow one to embed that sort of knowledge in the triggers. So one is often forced to use commonly occurring symptoms as triggers, knowing that they may trigger the frame at inappropriate times, in order not to overlook a syndrome.

One way to overcome this is to allow the designer to give a more complete description of the situations (via the *situation* feature described in the previous section) which trigger a syndrome and those that deactivate hypotheses. The description of the situation may include logical combinations of findings, clinical states, and syndromes. The phase of the diagnostic process (e.g., the beginning of the diagnosis or the end) may also be important in describing the situation required to trigger a syndrome.[34]

INTERNIST's disease hierarchy is also a useful idea in principle. In addition to reducing the number of hypotheses, this allows one to choose strategies based on a more general view of the problem, but one often finds, when using INTERNIST, that after entering ten or so manifestations, almost all of the relevant activated diseases are the terminal entries of the hierarchy (non-terminal entries are mostly the unreasonable hypotheses). In some sense INTERNIST is using the hierarchy to limit the number of unreasonable hypotheses rather than using the hierarchy to help in strategy selection and problem formation.

The problem is similar to PIP's trigger problem. If one wishes to use the hierarchy for problem formulation a single manifestation cannot always be allowed to change the level of the hypothesis. The more specific triggers described above might alleviate this problem. The leaf nodes in the hierarchy could have very specific triggers whereas the more superior nodes would allow more general situations to trigger them.

---

34. Studies seem to indicate that physicians are more willing to generate new hypotheses early in the diagnosis than towards the end [23,20].

This would allow the more general disease nodes to be triggered before the more specific inferior nodes. It might be that the more detailed information provided in the triggers could obviate the need to use the hierarchy for controlling hypothesis generation. In that case, the hierarchy might be used only to guide the strategy selection. Only leaf nodes would be hypothesized, superior nodes would be used to determine whether many of these hypotheses are similar, (i.e., have the same superior). If so, a different strategy might be used. This is similar to the Patil's "group and differentiate" strategy [25], but strategies other than "differentiate" can be used.

Another method for controlling the proliferation of hypotheses is PIP's *must-have* and *must-not-have* features. The evidence from the comparison indicates that "always present" or "always absent" findings are rare, but "almost always present" or "almost always absent" findings are common. In an attempt to enhance the usefulness of these features, a recourse for deactivated hypotheses might be added. By allowing these deactivated hypotheses to be reconsidered later, if necessary, mistakes caused by atypical presentations can be detected. This should allow the usage of the *must-have* and *must-not-have* features and hence greater pruning of unlikely hypotheses without incurring the errors found in PIP. As with the triggers, the values of these slots should allow a more complete description of the situations to which they apply.

PIP also uses causal, associational links and the *differential-diagnosis* slot to first semi-activate and then activate frames. The experience gained from running test cases and developing the syndrome frames indicated that semi-activation was not the correct action in many cases. It was often the case that there was a more precise action that could have been taken if there was a way to encode that information. A method for encoding these actions should be developed.

6.2.2 Scoring

Although the comparisons of the scoring algorithms did not find any significant difference in their performance, the ad hoc nature of these scoring strategies leads one to wonder if a strategy founded in Bayesian analysis would have superior performance. It was noted that each algorithm did have some peculiarities. PIP, as noted previously, penalizes syndromes for having a large number of associated findings. It also fails to use the importance of findings (weights) in calculating the fraction of the findings explained by the hypothesis.[35] In INTERNIST's scoring algorithm the magnitude of the score does not seem to relate to the probability of the hypothesis being present since the magnitude of the hypothesis' score does not correspond to the likelihood of the hypothesis being concluded (I am assuming that one wants a hypothesis to be concluded only when the probability of it being present is very high). It would be interesting to attempt to develop a scoring algorithm which is as close to Bayes' rule as possible, justifying any deviations or relaxation of assumptions that are required. This has two advantages over the ad hoc algorithms of PIP and INTERNIST: 1) by doing this one has a mathematical justification for decisions made using these scores and 2) where available, probabilities can be objective measures rather than subjective measures.

An attempt to do this was outlined in Szolovits' paper: *Remarks on Scoring* [49]. In this paper Szolovits outlines a scoring algorithm based on Bayes' rule, but containing useful additions. First, Szolovits' algorithm allows for uncertainty of the observations. This is useful since the physician does not always know for certain that a finding is present, especially historical findings and laboratory data. Second, the algorithm allows one to incrementally take into account the interdependencies between findings. Last, this scoring algorithm allows for the interdependencies between hypotheses. Although

---

35. Although it is debatable to what extent these problems effect the performance of PIP's scoring algorithm, they are certainly not desirable properties of a scoring strategy.

some modifications would be needed in Szolovits' algorithm,[36] it would be interesting to compare the results for this algorithm to those of the ad hoc algorithms of PIP and INTERNIST.

6.2.3 Diagnostic Strategy and Question Selection

From observing the systems' choice of questions it appears that both PIP's and INTERNIST's algorithms were not able to capture all of the clinical knowledge used by physicians in question selection. INTERNIST sometimes asked questions which seemed inappropriate at that point in the session, (e.g., it would ask a very specific question before asking the basic findings of the syndrome). PIP's method asked fewer unreasonable questions, even though the questions were sometimes not optimal.[37] The reason for the unexpected questions asked by INTERNIST, I feel, was that its dynamic selection algorithm is too simplistic. It fails to capture many subtleties of the situations. For example, often a laboratory test may be routinely run and hence be no more costly to ask than a symptom, etc. The reason that PIP asked fewer inappropriate questions appears to be that some of the knowledge about clinical style was embedded in the ordered list of findings. Unfortunately, PIP fails to consider the overall situation in selecting question to ask the user, so it sometimes asks questions which will not help to differentiate between the top hypotheses.

---

36. Modifications are required because Szolovits assumes that all syndromes are active (i.e. hypothesized) throughout the session and this is not the case for PIP and INTERNIST. Also clinical and physiological states (if present) need to be scored and Szolovits' algorithm may require modifications to do score these states.

37. PIP's method worked best when there was one clear leading hypothesis. If several were grouped at the top it would occasionally ask a question that was present in more than one of the leading hypotheses and hence did not help to differentiate between these hypotheses.

To overcome some of these problems, one might construct a system which would attempt to analyze the situation, in a more sophisticated way than INTERNIST, and determine which question to ask. The problem with this approach is that the analysis may be quite complex. Instead, since PIP's ordered list of findings did work well when the confirm or ruleout strategy was required, several ordered lists for each disease for different strategies could be used. The question selection algorithm could use lists from more than one hypothesis (as well as other available information) for strategies where an order for the findings cannot be predetermined, (e.g., differentiate). This would allow the system to use the embedded knowledge of the ordered lists, hence simplifying the question selection algorithm, while still allowing dynamic question selection where required.

Another weakness in the strategies of PIP and INTERNIST is their lack of knowledge about the diagnostic process. This causes the systems to behave in unusual ways which can cause the physician-user to lose confidence in the system. For example, PIP will activate a new hypothesis late in the diagnostic session, immediately before concluding that a syndrome is present. Indeed, the finding that triggers the creation of the new hypothesis may be explained by the hypothesis which is about to be concluded. INTERNIST, one might claim, has some plan for the diagnostic session in that the session can go from *ruleout* mode to *differentiate* mode to *confirm* mode. But it can change from *confirm* to *ruleout* mode at any time during the diagnosis. In addition, neither system has any method for detecting when it is not making headway or when the problem might not be in its field of expertise.

One way to capture some of the knowledge about the diagnostic process would be to break the session into several parts: beginning, middle and end. Each of these

parts could have different expectations[38] and effects on the strategy algorithm. For example, the ease of triggering a new hypothesis might differ for different parts of the session. Also a systems might use the the expectations to determine when the diagnosis is proceeding normally. If not, a system could alter its strategy, (e.g., it might try to see if a deactivated hypothesis should be reconsidered). Having a model of the diagnostic process seems necessary to improve the diagnostic style of the system.

Another problem with PIP's question selection strategy is that there is no ability to determine when to pursue a clinical state which is linked to a hypothesized syndrome versus pursuing the hypothesized syndrome. This sometimes causes PIP to stop pursuing a clinical state and begin pursuing a linked syndrome before the clinical state has been concluded to be present. This led to very specific findings (associated with the syndrome) being asked before more general findings (associated with the clinical state) have been inquired about. This problem can be (at least partially) rectified by allowing clinical states to be placed in the ordered lists of findings. When a clinical state is encountered by the question selection algorithm it would use the equivalent list in the clinical state. This would permit the system to know whether a clinical state should be investigated early on in the diagnosis of a syndrome or at a later point in the diagnosis.

6.2.4 Concluding Hypotheses

The comparison of PIP and INTERNIST showed that the conclusion algorithm needs to examine the scores of at least the top two hypotheses. This works well but it appears that INTERNIST occasionally concluded a syndrome sooner than the physician did (i.e., before all of the major findings noted by the physician were entered into INTERNIST). Although this caused no errors with the 35 cases, if one were using PIP's

---

38. An expectation might be that after a certain number of questions were asked there would be one hypothesis clearly leading the others.

triggering algorithm it would be possible to conclude the wrong syndrome. To prevent this one could have a list of findings that must be inquired about before the syndrome can be concluded.

6.2.5 Multiple Syndromes

In the birth defects area and pediatrics in general multiple diseases and syndromes are much less common than in internal medicine (only about 5% of the birth defects are diagnosed as multiple syndromes, whereas 70% of birth defects are undiagnosed). In fact there were no multiple birth defects in the clinical cases used in this comparison. So it is difficult to judge the effectiveness of PIP and INTERNIST's methods of handling multiple birth defects and propose corrections. It was possible to determine when multiple hypotheses where incorrectly hypothesized (since any time they were hypothesized it was incorrect). Of course, this information is not sufficient to propose an algorithm for handling multiple syndromes.

It is abundantly clear that PIP's method of handling multiple syndromes is inadequate. In essence it merely pursues all active hypotheses until their scores become so low that they are deactivated, quitting only when there are no hypotheses left. This proves to be a very frustrating way to terminate the algorithm. PIP does not remove findings explained by a syndrome after it has been concluded so they remain to be used by the other active hypotheses.

INTERNIST's approach is the exact opposite of PIP's approach. INTERNIST deletes all findings explained by concluded syndromes and then proceeds if one or more findings with high import score remains unexplained. This algorithm will work correctly if the findings of the co-occurring syndromes do not overlap. But syndromes may have findings in common which would not be taken into account if all of the explained findings were removed.

It is doubtful whether an algorithm can be developed to correctly diagnosis multiple syndromes using only the information available to PIP and INTERNIST. This is due to the fact that the findings could interact and overlap to the extent that neither syndrome is discernible without knowledge of these interactions and the knowledge bases for PIP and INTERNIST do not know about these interactions. Adding information about the interactions of different syndromes and findings is a large undertaking that would substantially increase the complexity of these systems. [39]

Although diagnosis of multiple syndromes with interactions between the findings is difficult, it may be possible with only a small amount of additional information, for systems like PIP and INTERNIST to detect when these interactions are likely. Multiple syndromes without interactions can be diagnosed by an algorithm similar to INTERNIST's. If interactions are expected, then the system can either notify the user that this may be a case that the system cannot handle, or ship the case off to another system which can handle these interactions. The interactions between syndromes could be detected by checking the intersection of the findings of the syndromes in question. In addition *interaction* links between findings could be added to the database and these links could also be checked (to a given depth) to detect interactions between two syndromes.

In the congenital defects domain diagnosing the single syndromes and multiple syndromes whose findings do not interact should account for most of the cases. I believe that cases in which the findings interact and overlap to the extent that neither syndrome is discernible without knowledge of these interactions are usually classified as unknown (or given a unique name if they occur often enough).

---

39. There are research projects presently under way to implement such systems in other domains [25, 32].

6.2.6 Exception Handling

A diagnostic system should be able to alter the flow of control of the system. This is desirable for two reasons. First, it was found, when running PIP and INTERNIST on cases, that certain situations arose where the action taken by the system was inappropriate. This is bound to occur when using a few simple methods to handle all possible situations. A better action was often known but encoding this knowledge into these systems was not possible.

The second reason stems from the observation that experts in medicine appear to acquire information which, while they are pursuing one hypothesis, allows them to recognize a pattern of answers indicating that a different hypothesis may be worth investigating. This allows these experts to avoid backtracking and arrive at the solution in a more direct manner. This is the intent of PIP's semi-activated state with causal and associative links and with the *differential-diagnosis* feature. Unfortunately semi-activation alone does not capture the experts' ability to redirect their focus. This sort of knowledge is not easily put into simplistic algorithms, the correct action is usually dependent on the situation.

To overcome these problems a feature should be added to the system so that it would be able to detect that a *special* situation has occurred and be able to take the appropriate action. This can be accomplished via the use of daemons. These are often already available in knowledge representation languages.

6.3 Summary

The representations of knowledge in both PIP and INTERNIST are not robust enough to capture all of the knowledge used by physicians. To overcome this problem, the use of a general knowledge representation language is recommended. In addition, viewing syndromes and clinical and physiological states as either states or findings,

since physicians do, is proposed to increase flexibility and ease user interaction. Also providing a general uniform representation for medical situations should allow the designer of the knowledge base to more accurately represent the clinical situation. A more detailed representation of temporal and non-medical knowledge is also advocated.

The comparison of PIP and INTERNIST presented in the last chapter showed that both systems' algorithms took inappropriate or incorrect actions during the diagnoses. The reasons for the inappropriate actions were that the systems used a few simple algorithms to handle all possible cases and that the information used by the algorithms was highly restricted and hence the entire medical situation was not taken into account. This chapter presented several suggestions for preventing the inappropriate or incorrect actions by allowing the algorithms to use all of the information available to the system and allowing the system to recognize *special* situations and take the appropriate actions. The specific suggestions included: a more detailed description of the clinical situation for hypothesizing a syndrome or state in order to reduce the number of inappropriate hypotheses; a scoring algorithm which is developed from Bayes' rule instead of the ad hoc algorithms of PIP and INTERNIST; a strategy selection algorithm which uses a model of the diagnostic process and reconsiders discarded hypotheses when progress is not forthcoming.

These suggested improvements should create a more flexible system which can represent more of the medical knowledge and use this knowledge in a more appropriate manner.

7. Conclusions and Further Research

7.1 Summary

This thesis examined the performance of the Present Illness Program and the INTERNIST system using the Congenital Defects problem domain. These systems are both hypothesis driven pseudo-Bayesian diagnostic systems. Both systems use simple representations and rather general algorithms to determine the diagnoses (although PIP has somewhat more exception handling capability). The examination of PIP and INTERNIST found that these general algorithms worked well much of the time, but in a small percentage of the cases the general approach did not work. In addition, the algorithms employed by the different systems proved to perform at different levels of proficiency. PIP's hypothesis generating algorithm (using triggers) generated fewer hypotheses than did INTERNIST while still generating the correct ones. INTERNIST's confirmation algorithm concluded more correct hypotheses and caused fewer errors than did PIP's. The scoring and question selection algorithms performed at similar levels for each system.

The last part of the thesis discussed several improvements for hypothesis driven medical diagnostic systems, of the PIP and INTERNIST variety. The suggested improvements attempt to increase the flexibility and completeness of the diagnostic system. Additional constructs have been proposed to allow the representation to capture more of the medical knowledge. Additional methods were suggested to allow the system to be flexible enough to correctly handle more of the situations that occur during the diagnosis.

7.2 Further Research

There are areas of clinical diagnosis which can be further investigated with respect to the comparison of PIP and INTERNIST. The comparisons in this thesis were all done using thirty-five clinical cases and a database which contained fifty syndromes. This was adequate to determine many of the properties of these systems. With more clinical cases some of the more subtle properties may also be detected. The performance of these systems in clinical cases with more than one syndrome present was not investigated. Although multiple syndromes are not as common as in internal medicine, they do occur and this aspect of PIP's and INTERNIST's diagnostic algorithms should be investigated. Also, since any complete system is likely to have hundreds or thousands of syndromes, it would be interesting to determine the behavior of the systems as the number of syndromes known to each system increases.

The development of a new system using the suggested improvements discussed in chapter 6 is another area for further research. If a such a system was constructed, it would be of interest to determine the extent to which the performance of this system can be improved without including a mechanism for physiological reasoning.

One issue not addressed in this thesis is the handling of unanticipated interactions between findings when two or more syndromes are co-occurring. Handling these interactions will require reasoning about the physiology and pathophysiology involved. Incorporating this knowledge into diagnostic systems is a difficult problem but an important one if these systems are to be able to handle all of the clinical situation presented to physicians. This problem is presently being researched at M.I.T. by Patil[25].

The issues involved in physician interaction with these systems have not been investigated in this thesis. The ease of interaction, length of the diagnostic session, and coherence of the diagnostic style will all effect the acceptance of these systems by physicians. These issues must be addressed before any diagnostic system is ready for use by the medical community.

Appendix I - List of Syndromes

ACHONDROPLASIA
ACROFACIAL DYSOSTOSIS
ACROOSTEOLYSIS DOMINATE TYPE
ACROPECTOROVERTEBRAL DYSPLASIA
ADRENAL HYPOALDOSTERONISM OF INFANCY TYPE TRANSIENT ISOLATED
ADRENAL HYPOPLASIA TYPE CONGENITAL
ADRENOCORTICAL UNRESPONSIVENESS TO ACTH TYPE HEREDITARY
ADRENOCORTICOTROPIC HORMONE DEFICIENCY TYPE ISOLATED
ASPHYXIATING THORACIC DYSPLASIA
CHONDRODYSPLASIA PUNCTATA CONRADI
CHONDRODYSPLASIA PUNCTATA RHIZOMELIC
CLEIDOCRANIAL DYSPLASIA
CORTICOSTEROID BINDING GLOBULIN ABNORMALITIES
DIABETES INSIPIDUS TYPE VASOPRESSIN RESISTANT
DWARFISM TYPE PANHYPOPITUITARY
ENDOCRINE NEOPLASIA I TYPE MULTIPLE
ENDOCRINE NEOPLASIA II TYPE MULTIPLE
ENDOCRINE NEOPLASIA III TYPE MULTIPLE
GOITER TYPE GOITROGEN INDUCED
GONADOTROPIN DEFICIENCY TYPE ISOLATED
HYPERALDOSTERONISM TYPE FAMILIAL GLUCOCORTICOID SUPPRESSIBLE
HYPERPARATHYROIDISM TYPE NEONATAL FAMILIAL
HYPOCHONDROPLASIA
HYPOGLYCEMIA TYPE FAMILIAL NEONATAL
HYPOGLYCEMIA TYPE LEUCINE INDUCED
HYPOMAGNESEMIA TYPE PRIMARY
HYPOPARATHYROIDISM TYPE X LINKED INFANTILE
HYPOPHOSPHATASIA
HYPOPHOSPHATEMIA
JUVENILE DIABETES MELLITUS
JUVENILE DIABETES MELLITUS TYPE OPTIC ATROPHY AND DEAFNESS
LIDDLE SYNDROME
PSEUDOHYPOALDOSTERONISM
RICKETS TYPE VITAMIN D DEPENDENT
SILVER SYNDROME
SPONDYLOEPIPHYSEAL DYSPLASIA CONGENITA
SPONDYLOEPIPHYSEAL DYSPLASIA TARDA
SPONDYLOTHORACIC DYSPLASIA

STEROID 11 BETA HYDROXYLASE DEFICIENCY
STEROID 17 ALPHA HYDROXYLASE DEFICIENCY
STEROID 17 TYPE 20 DESMOLASE DEFICIENCY
STEROID 18 HYDROXYLASE DEFICIENCY
STEROID 18 HYDROXYSTEROID DEHYDROGENASE DEFICIENCY
STEROID 20 22 DESMOLASE DEFICIENCY
STEROID 21 HYDROXYLASE DEFICIENCY
STEROID 3 BETA HYDROXYSTEROID DEHYDROGENASE DEFICIENCY
THYROGLOSSAL DUCT REMNANT
THYROID DYSGENESIS
THYROID PEROXIDASE DEFECT
THYROTROPIN DEFICIENCY TYPE ISOLATED
THYROTROPIN UNRESPONSIVENESS

Appendix II - List of Clinical Cases

Case Number	Case Name
1	SPONDYLOEPIPHYSEAL DYSPLASIA CONGENITA
2	ACHONDROPLASIA
3	SPONDYLOEPIPHYSEAL DYSPLASIA CONGENITA
4	SPONDYLOTHORACIC DYSPLASIA
5	THYROID DYSGENESIS
6	SILVER SYNDROME
7	ACHONDROPLASIA
8	ASPHYXIATING THORACIC DYSPLASIA
9	CLEIDOCRANIAL DYSPLASIA
10	HYPOPHOSPHATASIA
11	SPONDYLOTHORACIC DYSPLASIA
12	ASPHYXIATING THORACIC DYSPLASIA
13	HYPOPHOSPHATEMIA
14	HYPOCHONDROPLASIA
15	HYPOPHOSPHATASIA
16	ACHONDROPLASIA
17	ACHONDROPLASIA
18	SPONDYLOEPIPHYSEAL DYSPLASIA CONGENITA
19	ACHONDROPLASIA
20	CLEIDOCRANIAL DYSPLASIA
21	SPONDYLOEPIPHYSEAL DYSPLASIA CONGENITA
22	CLEIDOCRANIAL DYSPLASIA
23	CHONDRODYSPLASIA PUNCTATA CONRADI
24	CLEIDOCRANIAL DYSPLASIA
25	ACHONDROPLASIA
26	SPONDYLOTHORACIC DYSPLASIA
27	CLEIDOCRANIAL DYSPLASIA
28	THYROTROPIN UNRESPONSIVENESS
29	HYPOCHONDROPLASIA
30	SPONDYLOTHORACIC DYSPLASIA
31	SILVER SYNDROME
32	ACHONDROPLASIA
33	CLEIDOCRANIAL DYSPLASIA
34	CHONDRODYSPLASIA PUNCTATA CONRADI
35	ACROOSTEOLYSIS DOMINATE TYPE

Appendix III - The Design for an Improved System

In the previous comparison of the PIP and INTERNIST, several strengths and weaknesses of the respective systems were identified. This section will define a framework for a new diagnostic system, the Congenital Defects Diagnostic System (CDDS), which attempts to utilize the desirable features, eliminate the undesirable ones from each system and combine these features with new ideas in order to create a better system for the birth defects domain.[40]

Rather than proposing a general solution to medical diagnosis, or a wholly new approach to the problem, this is an attempt to improve upon the techniques of PIP and INTERNIST (beyond what can be done by fine tuning them for use in this particular domain) while keeping to their pseudo-Bayesian approach. Of course, this means that some known flaws in PIP and INTERNIST will not be corrected.

As shown in the comparison, both PIP and INTERNIST do a fairly good job of diagnosing the birth defects included in this data base. Most of the changes will be aimed at diagnosing those situations with which these systems did poorly. Attempting to squeeze the last drops of performance out of any technique usually requires a great deal of effort for modest improvements.[41] Hence, the added complexity of Congenital Defects Diagnostic System will effect only modest improvements in performance most of the time. But in the portion of the cases not handled well by PIP and INTERNIST, this new system should substantially improve performance. It is precisely this population of cases that CDDS is designed to handle in an improved manner, while still, naturally, dealing with the rest.

There is still a population of cases that this new system is not attempting to handle. This population is comprised of those cases involving multiple syndromes, whose interactions are not explicitly stated and where these interactions significantly alter the findings of each syndrome significantly. Since the system, like PIP and INTERNIST, does not try to do any reasoning regarding the effects of one syndrome on

---

40. This system was a preliminary design which was constructed before the last chapter was written. CDDS incorporates some of the suggestions made in the last chapter, but not all. In addition, some of the ideas implemented in this system are only first order solutions which the author no longer defends as adequate. They will be replaced in newer versions of CDDS.

41. In the AI community (and elsewhere) this is referred to as the 80-20 phenomenon, i.e. it takes 20 percent of the effect to get 80 percent of the performance. The other 20 percent will take 80 percent of the effort.

another, unless this information is explicitly entered in the database, there is no way for the system to determine these interactions.

The philosophy guiding the design of this system has two basic components, completeness and flexibility. The system should be able to easily represent knowledge about the problem domain and about solving problems in that domain. Also, since no simple global algorithm will be able to cover all of the possible situations, the system should be flexible enough to allow special situations to cause specific actions to occur (i.e., to allow changes in the control flow). These special situations could be caused by either information in the database or by the user. The following sections describe the implementation of this philosophy for the various parts of the diagnostic algorithm and the represention of knowledge.

III.3 Representation of Knowledge

One of the problems found in both PIP and INTERNIST was the lack of ability to represent the needed knowledge in an efficient manner using the constructs provided. It was often noted, while constructing the database for these systems, that the physician's knowledge was very difficult or impossible to encode in the system. Even simple relationships were hard to define.

In INTERNIST, for example, each finding is represented as a LISP atom with associated properties. Any time a variation in a finding is desired it must be represented as a new finding and its associations explicitly stated. So for the finding "short stature" where the time of onset is important, several finding must be created and their relationships asserted. The association between findings must be stated for each finding separately. To state that "tall" is the opposite of "short" in all findings referring to length, one must state for each finding pair that they are opposites.

In PIP, the situation is somewhat better in that a finding can have many items associated with it, each with several mutually exclusive values. So some of these relationships only need be specified once, i.e. the item "length" can be created with values: long, short and normal, or "short stature" can have associated item "onset" which can have numerous values. But it is still not possible to conveniently encode all of the knowledge that is known by the physician. For example, it is difficult to encode the fact that a normal level of a substance (e.g. normal serum sodium level) should not decrease the score of those hypotheses that do not explicitly mention the finding (one does not want to have to list all of the normal findings that are expected with a disease or syndrome). It is also difficult to represent the relationships between the different items in a finding. This can cause needless questions to be asked. For example, one can state the finding of edema is not recurrent and PIP will still inquire as to the frequency of the recurrence.

PIP's frame representation could have captured more of the clinical knowledge but the implementation did not include many of the features described by Minsky [22] which gives the frame representation much of its flexibility. The algorithm also restricts the types of values the slots may take. This further restricts the type of knowledge that can be represented. Because the implementation was not as general as one might like, information that could have been put in the database was dispersed into the algorithm and hence created difficulties when trying to alter the program.

III.3.1 The Frame Representation Language

The representation developed for CDDS attempts to overcome the problems encountered in developing a database for PIP and INTERNIST. The representation chosen, to attempt to accomplish these aims, is a frame representation [22]. Hence, the medical knowledge as well as the other world knowledge (e.g., knowledge about time, properties of objects, state of the system) are represented as frames. This representation is implemented using an autonomous frame representation language very similar to FRL [33,34].[42] A full description of FRL is presented in other papers [33,34]; the following is only a brief description of some of the useful features of FRL.

In FRL, objects are represented as frames with slots, facets, datum, comments, and messages containing the information about the frame (see figure 15). Frames are created and the information contained in them is modified and retrieved using the predefined functions of FRL. FRL supports defaults, inheritance, constraints, and procedural attachment and other techniques useful in knowledge representation.

Fig. 15. A Typical Frame

Frame	Slot	Facet	Datum	Label	Message
Minsky					
	Name	Value	|Marvin Minsky|		
	Address	Value	|Tech Sq.|		
	Interests	Value	Robotics Music	Source	RBR

42. The language is actually a subset of FRL with the options not deemed necessary for this application eliminated in order to increase speed and conserve memory.

FRL provides a mechanism for the inheritance of properties by allowing frames to be organized in an **AKO** (a-kind-of) hierarchy and by having functions to use the hierarchy to retrieve information. If no value is found when attempting to retrieve the data for a particular frame, slot and facet, the **AKO** links are searched to see if a superior frame has a value for the slot and facet in question. This is a convenient way for frames to share information. The FRL functions allow the programmer to specify whether or not to use inheritance. If the information retrieved has been inherited, the name of the frame in which it was found is attached as a comment to the retrieved data. FRL also allows default values for slots using the *default* facet. Defaults can be used in conjunction with inheritance.

FRL supports several types of attached procedures which are evaluated in specific circumstances. Attached procedures allow values to be computed rather than explicitly stated. In FRL one can indicate that the datum is to be evaluated and the result of evaluation returned rather than the datum itself. The data can be retrieved from another frame using an indirection pointer. Also available are the *if-needed*, *if-added*, and *if-removed* facets which evaluate the procedures (the data of the facets) when the value of the slot is needed and not present, when a value is added to the slot, and when a value is removed from the slot, respectively. The *require* facet is used to constrain the possible values that a slot can have. The values of the slot can be checked against the attached predicates in the *require* facet of the slot.

FRL distinguishes between generic frames and individual frames by the value of the *classification* slot. Instantiation of a generic frame causes an individual frame to be created with an *ako* link to the generic frame. This is needed in CDDS to distinguish between the generic findings and findings entered into the system and between the prototype syndromes and the hypothesized syndromes.

III.3.2 Syndromes, States and Findings

In CDDS, the medical knowledge is divided into syndromes, clinical and physiological states, and findings, in a manner similar to PIP. But the way in which these entities are used and the information associated with each is different from PIP. As in PIP, syndromes are represented as frames with associated findings. These associated findings are listed in the *findings* slot of the syndrome frames. The slot actually contains pointers to finding frames which contain the description of the findings associated with the syndrome. There are other slots in a syndrome frame which contain additional information about the syndrome. These include: *type*, *scoring-function*, *must-not-have*, *must-have*, *is-sufficient*, *triggers*, *confirm*, *items*, *inferiors*, *superiors*, *ruleout*, *links*, and *special-action*.

Fig. 16. Typical Syndrome or Clinical State Frame

Frame	Slot	Facet	Value
Hypothyroidism			
	Type	Value	Clinical-state
	Classification	Value	Generic
	AKO	Value	Thyroid-Disorder
	Inferiors	Value	Hypothyroidism-1, Hypothyroidism-2
	Prior-prob	Value	0.0003
	Findings	Value	(DRY-SKIN (STATUS (VALUE PRESENT)) (RESPIRATORY-DISTRESS (STATUS (VALUE PRESENT)) (ONSET (VALUE (FROM BIRTH))) (DURATION (VALUE (3 MONTHS)))) etc.
	Confirm	Value	((DELAYED-ERUPTION-OF-TOOTH (STATUS (VALUE PRESENT))...) (DISTENTION-OF-ABDOMEN (STATUS ...)) (SLOW-RADIAL-PULSE-RATE (STATUS ...)) etc.)
	Ruleout	Value	((SLOW-RADIAL-PULSE-RATE (STATUS...)) (DELAYED-ERUPTION-OF-TOOTH (STATUS ...)) etc.)
	Trigger	Value	(or (SLOW-RADIAL-PULSE-RATE (STATUS (VALUE PRESENT))) ...)
	Items	If-needed	(cond ((compare (MENTAL-RETARDATIONS (STATUS ...) (DEGREE SEVERE))) '(DEGREE (VALUE SEVERE))))
	Must-Not-Have	Value	(or (THYROXINE (LEVEL (VALUE INCREASED))) etc.)
	Is-Sufficient	Value	(and (THYROXINE (LEVEL (VALUE DECREASED)))...)
	Links	Value	(((THYROID-DYSGENESIS (TYPE (VALUE (CAUSED-BY))) (ONSET (VALUE (AT BIRTH)))) .76) etc.)
	Scoring-Function	Value	(((SLOW-RADIAL-PULSE-RATE (STATUS (VALUE PRESENT))) .85) ((SLOW-RADIAL-PULSE-RATE (STATUS (VALUE ABSENT))) .15) etc.)

The *type* slot indicates the frame's type (e.g., syndrome, clinical state, finding, etc.). This information is needed in order for the program to correctly use the information in the frame. This slot is also present in other frames.

The *scoring-function* slot contains lists of findings and their associated conditional probability given the syndrome (i.e., Prob(Finding | Syndrome)). This information is used by the scoring algorithm.

The *must-not-have*, *must-have* and *is-sufficient* slots are essentially the same as these slots in PIP. The exception to this is that CDDS does not restrict the value of the slots to just findings. CDDS allows, via a *statement* frame (discussed in section 8.1.3), the use of clinical-states, physiological-states, syndromes and system states, as well as findings in the datum of the *value* facet of these slots. This allows the designer of the database to specify clinical situations more precisely and hence decrease inappropriate actions by the program.

The values of the *confirm* and *ruleout* slots are ordered lists of pointers to finding frames. These lists are used to determine the order in which questions are asked. This will be discussed later. Finding frames are used rather than just the finding name since the additional information is used to deduce if a finding is worth asking about. For example, if the system is considering asking about a finding in reference to a particular syndrome, but its value in the *confirm* slot of the syndrome states that the finding onset occurs at an age greater than the patient's age, then the finding is no longer considered. There can also be syndromes and states in this list (their use is discussed in the diagnostic strategy section).

The *inferiors* and *ako* slots are used to create an INTERNIST like disease tree which will be used in strategy formulation and question selection. This also will be discussed later.

The *trigger* slot is evaluated in order to determine whether or not to activate the syndrome or state. This slot also contains pointers to a *statement* frame, so virtually any combination of findings, states or syndromes can activate a frame. This facility has been added to allow a more complete description of the activation process by allowing the designer of the database to require more complex patterns of findings to be present for activation and also separates the scoring information from the triggers.

The *special* slot is used to alter the flow of control of the system via condition-action pairs which are checked on each iteration. These are also discussed later.

This system attempts to allow clinical and physiological states to be used as if they were findings. In order to do this a state should be able to have modifiers (*items* in PIP's terminology) associated with it, such as severity, onset, etc. Without this feature one would need many similar states and syndromes, (e.g., mild hypothyroidism, moderate hypothyroidism, severe hypothyroidism, etc.) in order to specify the correct clinical situation. These modifiers are placed in the *items* slot of the state or syndrome frame. Since, unlike findings, the user is not usually entering the information about the modifiers (i.e., the states are hypothesized and concluded by the system), an *if-needed* facet containing a function that computes the value of the modifier is used. If the value of the slot is requested it will be computed. CDDS also allows physicians to enter states and syndromes as if they were primitive findings. This dual perspective of states (and syndromes) as either syndromes or findings is consistent with the observations of the use of these entities by physicians, as noted by Feinstein[12].

The causal and associative links between states are contained in the *links* slot. The slot contains pointers to *link* frames. The information contained in these frames includes the type of link, the strength of the link, and any special contexts associated with this link (i.e. a link might only be valid if the severity of the state is mild or if the duration is greater than three weeks). The strength of the link is of the identical form as entries in the *scoring-function*. It is more convenient to place the information here since the scoring algorithm takes links between states and syndromes into account at a different time than links between findings and syndromes. The items of the state or syndrome (and the values of the associated findings) are used in determining the strength of the link given different presentations of the state or syndrome. This allows CDDS to capture information that is not easily represented in PIP. For example this allows one to overcome the "X" phenomenon described by Rubin [35]. The "X" phenomenon occurs when the presentation of a clinical state varies depending on the causing syndrome or disease. A given presentation may support one disease more than another. For example, both ascites and facial edema is evidence for the clinical state of sodium retention, but sodium retention with ascites supports cirrhosis more than acute glomerulonephritis whereas sodium retention with facial edema supports acute glomerulonephritis more than cirrhosis. There is no easy way to represent this knowledge in PIP without removing the clinical state and placing the appropriate findings in the appropriate syndromes.

The Syndrome and state frames (as well as all other frames) are placed in a hierarchy, all having the **REAL-WORLD** as the most superior frame in the hierarchy. The more superior frames contain more general information which can be shared by the inferior frames. As with all hierarchical representations, this allows one to place common information in a common superior frame. The frames are also connected, via links, to other frames.

When syndromes or states are hypothesized by the system a new instantiation of the generic syndrome or state is made (using the FINSTANTIATE function of FRL). The new frame is placed in the hierarchy but marked as an individual frame rather than generic frame via the *classification* slot.

III.3.3 Statements

Before describing the *findings* frame, it is worth discussing the *statement* frames since they have been mentioned several times in the previous section. In order to describe clinical situations in greater detail a pointer to a *statement* frame is used as the value of many slots. This frame that can contain information about syndromes, states, findings, or the state of the system (i.e., the diagnostic strategy, the leading hypothesis, number of findings asked, etc.). Statements can be compared against each other to determine if they are equivalent. Usually this comparison is made against statements that represent the current state of knowledge about the patient. Since there may be uncertainty about the true state of the patient, this comparison determines the probability that the statements are equivalent (i.e., match).

A statement consists of a frame which contains an *answer* slot whose value is a list. The first item in the list is either a function name (usually a boolean operator) or a finding, clinical state, syndrome, system variable or item name. The rest of the items are pointers to other *statement* frames or details about the first item (e.g., the value of an *item* of a finding) (see figure 17).

When retrieving *statements* the system follows the pointers concatenating the values as it goes (see figure 17). The *statement* frames are inferior frames (in the ako hierarchy) to both the generic *statement* frame and to whatever the first item in the value *answer* slot is (e.g., a finding, clinical state, boolean operator, etc.). This allows *statements* about findings, for example, to inherit properties of the finding if needed.

Statements, by allowing syndromes, clinical states, and system states to be used in the same manner as findings, eliminate most of the restrictions in the possible values for the slots imposed by PIP. This allows the constructor of the knowledge base to represent the clinical knowledge more accurately by not forcing it into some rigid predetermined form the program expects. It also eases the task of incorporating knowledge about these entities into frames since one uniform representation can be used. From the physician-user's point of view, it facilitates the entry of data since each entry is converted into a statement and hence known clinical states and syndromes can

Fig. 17. Typical Statement

Frame	Slot	Facet	Value
Statement-54			
	Type	Value	Statement
	Classification	Value	Individual
	AKO	Value	And-Statement
	Answer	Value	(and Statement-7 Statement-32 Statement-55)

Retrieving this statement would yield an expression which might look like:

```
(And
      (Stature ;;; this is a finding
          (Height (Value Short))
          (Onset (Value (From 6 Months To 1 Year)))
          (Type (Value Disproportionate)))
     (Hypothyroidism ;;; this is a clinical state
          (Status (Value Present))
          (Onset (Value Birth)))
     (Diagnostic-Mode ;;; this is a system state
          (Status (Value Differentiate))))
```

be directly entered without requiring an exhaustive listing of their constituent findings.[43]

III.3.4 Representation of Findings

Findings in this system are meant to represent the signs, symptoms, historical facts, and laboratory tests which the physician uses in diagnosing birth defects. The representation of findings is similar to their representation in PIP in that the findings are represented as frames with the details about the findings found in the slots of the frames.

43. The system could still ask about these findings if that knowledge is required.

But the *finding* frames in CDDS contain additional information not found in PIP's *finding* frames (see figure 18). The slots used by generic *finding* frames include: *type*, *items*, *items-required*, *superiors*, *inferiors*, *prerequisites*, *triggered-by*, *natural-frequency*, *importance*, and *special*.

The *Type* slot merely identifies the frame as being a *finding* frame. The *item* slot lists the items that pertain to the finding. The *items-required* slot contains links between the items of the finding. This information indicates whether an item is worth inquiring about, thus preventing the asking of unnecessary items. For example, there is no reason for inquiring about the degree of mental retardation if its status in absent. This slot can prevent items from being asked at all.

Fig. 18. Typical Finding Frame

Frame	Slot	Facet	Value
Stature			
	Type	Value	Finding
	Classification	Value	Generic
	AKO	Value	Growth-related-findings
	Inferiors	Value	Stature-1, Stature-2, etc.
	Items	Value	Height, Severity, Type, Distribution, Onset
	Answer	If-needed	(determine-stature-from-age-and-height)
	Also-ask	Value	Growth-Rate
	Implies	Value	...
	Items-required	Value	...
	Prerequisites	Value	...
	Natural-Freq	Value	((stature (height short)) .23) etc.

The *implied-by* and *implies* slots are links between findings and are used in deducing findings from other findings (i.e., finding A implies finding B). The *implies* links are traversed as soon as the finding is known. The *implied-by* slot is used only when the finding is unknown but needed. The knowledge as to whether the finding was deduced or given by the user is kept (via the labels and messages of FRL) and the user can override deductions made by the system.

The *natural-frequency* and *importance* slots are both similar to INTERNIST *import* value. The *import*, in INTERNIST, is used in two ways. One, it is used in scoring in syndromes when the finding has been reported to be present but is not expected in that syndrome. Two, it is used to determine whether an unexplained finding warrants the consideration of multiple syndromes, both in the partitioning process and in deciding whether to continue the diagnosis after a syndrome has been concluded.

Two slots have been created here to attempt to separate this knowledge. The *natural-frequency* slot is used to give the scoring function of findings not mentioned in the individual syndromes and states. Since there are many different possible values for the finding's items, the value of this slot is not a single number as is the *import*. The value is a list of statement-value pairs as in the scoring function slot (i.e., a possible value of the finding and its associated probability).

The importance slot stores information as to the circumstances under which this finding should contribute to the continuation of the diagnosis or a consideration of multiple syndromes. The value of this slot is also a list of statement-value pairs. The statements are usually the different possible values of the finding but can be more situation specific conditions (i.e., the importance of explaining a finding may depend on the situation). The possible values of these statement-value pairs include: *need-not-be-explained*, *should-be-explained*, and *must-be-explained*. The use of these answers is discussed in section 8.2.4.

III.3.5 Representation of Items

Items contain the medical and non-medical information about the findings (and syndromes and states). For example, items contain information about the severity, time of onset, distribution, and other properties of the finding. Generic *item* frames contain the default list of possible values for the item which is contained in a *required* facet. This facet is inherited so that it need not be copied into every instance of the item. This slot could contain a function which merely checks a list of possibilities or a parser that can check the validity of more complex structures retrieved (for statements and other non-atomic values). There is also a slot for an optional evaluation function. This function is used in matching non-atomic values. The main use of this slot presently is for comparing time intervals, although it may be used for any item where an identity check

(EQ test) is not sufficient. The hierarchy is used to store information that is valid for all items in a common place.

III.3.6 Representation of Time

One of the deficiencies of both PIP and INTERNIST is their lack of ability to adequately capture temporal knowledge. Temporal knowledge is an important source of information in the diagnostic process. Often knowing the time of onset of a finding can greatly reduce the number of hypotheses generated to explain it. INTERNIST's only way to represent time is to create several findings, one for each possible time of onset, duration, or temporal relation to other findings, i.e. *short-stature-with-onset-at-birth*, *short-stature-with-onset-during-childhood*, *short-stature-with-onset-during-adolescence*, etc.. The inter-relationships between all of these finding would need to be specified also. This would, of course, lead to great proliferation in the number of findings if this were done for a large proportion of the original finding set.[44]

Fig. 19. Typical CDDS Item Frame

Frame	Slot	Facet	Value
Onset			
	Type	Value	Item
	Classification	Value	Generic
	AKO	Value	Item
	Inferiors	Value	Onset#1, onset#2, etc.
	Answer	Required	(parse-onset)
	Eval-function	Value	check-onset

44. For the congenital defects database for INTERNIST this was done only when it was deemed important enough for the syndrome to justify creating more findings.

In PIP, time was represented as an item in a finding. Originally, *now*, *near past*, *distant past*, *near future*, and *distant future* were the possible values of the time item. This was changed to an absolute scale since in congenital defects (and pediatrics in general) the age of the patient at onset of the findings is frequently more important than the order of occurrence of the findings. The possible values included: prenatal, birth, neonatal, infancy, early childhood, late childhood, adolescence, adulthood. This proved adequate for most situations, but not all. A finer grain was sometimes needed, but increasing the number of possible values was impractical since this would cause lengthy OR statements for most of the findings. In addition, the relative order between findings is sometimes important and not easily represented. Lastly, PIP does not provide a way to enter uncertainty about the temporal knowledge. This is often required especially when entering historical data.

In CDDS, the representation of time attempts to solve the problems listed above without increasing complexity too greatly. The scheme developed is somewhat similar to Kenneth Kahn's system [19], although greatly simplified. Each finding, or syndrome, or state can have an onset and a duration item associated with it. The value of these items is a pointer to an *onset* or *duration* frame. These frames can contain a *date* slot for the absolute time of the event and slots for before-after chains.

The date slot can contain a value which is a single number or two numbers which define a fuzzy interval. A fuzzy interval represent the maximum and minimum dates on a absolute scale (birth = 0.0) for the onset of the finding or state. For the *duration* frame the datum represents the maximum and minimum length of time possible for the duration of the finding. So two additions are all that are required to find the possible interval for the end of a finding. If the date slot contains a single number rather than two, it is assumed that the information is known with certainty.

The before-after chains are used to specify that a finding occurs during, before, or after the onset or cessation of another finding, syndrome or state. It can specify the amount of time between the findings, that time possibly being a fuzzy interval. There can be many chain slots the value of which is a list with two, three, or four elements. The first element is the type of chain, i.e. before, after, or during. The second is the finding or state being compared (the other is assumed known via the context in which the item occurs). The third and fourth define an optional interval of time required between the events, two numbers being a fuzzy interval.

There is a time expert which analyzes and compares the values of these items returning true or false or the percentage of overlap. It searches through before-after chains making simple deductions in an attempt to verify them. So if A occurs before B and C occurs after B it would conclude that A occurs before C. Any before-after chains which are deduced and required (i.e. not intermediate results) are stored so that they do

not need to be recomputed.

There are also special reference events which include: conception, birth, puberty, now (or age), and death. These are used as constraints by the time expert. So if a finding is supposed to occur at a time greater than the patient's age, the time expert can know that fulfilling that requirement is not possible, and it need not pursue it further.

There is a simple parser that lets one enter times in a more natural way. It converts the string into the appropriate time frame. For example, the input stream "between birth and 3 months" would be converted to a fuzzy interval and placed in the *date* slot of the appropriate *onset* frame.

III.4 The CDDS Algorithm

The algorithm used by this system is organized into three sections: the initial entering of findings, an iterative diagnostic loop, and a case summary.

The initial entering of data is similar to that of PIP and INTERNIST, with two exceptions. First, the user can enter clinical and physiological states, syndromes and control information as well as findings. Second, after the physician finishes entering facts, the system does a review of systems making sure systems not mentioned are normal. It asks about very general findings. If any abnormal findings are entered it looks down the hierarchy of findings with a menu selection format attempting to determine which inferiors of the abnormal finding are present. The user can terminate this review at any time. The review of systems is intended to help prevent long searches in incorrect areas due to incomplete initial data and help the user enter data if they are unfamiliar with the system.

The iterative diagnostic section is more similar to PIP's than to INTERNIST's. First, it asks about a finding. Then using this new piece of information it decides one of the following: whether to conclude any syndromes or states, whether to hypothesize any new syndromes or states, and/or whether to deactivate any old hypotheses. Then the system scores the active hypotheses, chooses a strategy, and selects the next question to ask given the strategy chosen. Each of these activities, as well as the reasoning behind the changes from PIP's and INTERNIST's algorithms, is described in detail in the following sections.

The case summary prints the reasons for the conclusions of the system and gives the system's explanation of the patient's findings. This should be useful to the physician as well as in the programmer debugging the system and the database. The physician can also request a summary at any point during the diagnosis. This summary lists all the concluded states and syndromes and the findings which each explains. In

addition, the summary lists the unexplained findings and the syndromes hypothesized to explain them.

III.4.1 Hypothesis Generation

The previous comparison of PIP and INTERNIST has shown that the methods used to restrict the number of hypotheses generated in both PIP and INTERNIST fail to achieve optimal performance, they generate hypotheses that are not reasonable given the situation.

PIP's use of triggers to activate certain syndromes, even though others could explain some of the findings, has great appeal, but its use of only single findings as triggers is not sufficient. The problem is that in certain situations one may want a finding to trigger a syndrome and in different situations one may not. Analysis of hypothesis generation by physicians has shown that pairs and combinations of findings caused triggering of hypotheses [20,23]. PIP does not allow one to embed that sort of knowledge in the triggers. So one is often forced to use commonly occurring symptoms as triggers in order to not overlook a syndrome, even though they may trigger the frame at inappropriate times.

To overcome this, CDDS uses statements as triggers, so any possible situation can be specified in the triggers. Actually the trigger is just one statement but since that statement can be a logical "OR" of other statements any number of patterns can trigger a frame. As shown in the next chapter, this system does decrease the number of unreasonable hypotheses.

INTERNIST's disease hierarchy is also a useful idea in principle. In addition to reducing the number of hypotheses, this allows one to choose strategies based on a more general view of the problem, but one often finds, when using INTERNIST, that after entering ten or so manifestations, almost all of the relevant activated diseases are the terminal entries of the hierarchy, usually the non-terminal hypotheses are unreasonable. In some sense INTERNIST is using the hierarchy to limit the number of unreasonable hypotheses rather than using the hierarchy to help in strategy selection and problem formation. The problem here is similar to PIP's trigger problem. If one wishes to use the disease hierarchy for problem formation a single manifestation cannot always be allowed to change the level of the hypothesis.

In this system a syndrome hierarchy is employed, but it is not used to generate hypotheses or reduce the number of hypotheses. Rather, it is used to allow the program to reason about groups of diseases (see section {8.2.4}). The non-terminal frames need not even have triggers, although they can. They can be exclusively used in strategy and question selection. If triggers are used and a non-terminal syndrome is concluded, the

inferiors are activated. If an inferior syndrome is triggered its superior one is deactivated, but it still can be used in strategy and question selection. The findings of a non-leaf syndrome are not restricted to be the intersection of the findings of its inferiors. So, even if very specific syndromes are triggered early, the question selection strategy can still decide to use their superior. This is similar to the "group and differentiate" strategy [25], but strategies other than "differentiate" can be used.

Another method for controlling the proliferation of hypotheses is PIP's *must-have* and *must-not-have* features. The evidence from the comparison indicates that "always present" or "always absent" findings are rare, but "almost always present" or "almost always absent" findings are common. So, in an attempt to enhance the usefulness of these features, the *reevaluate* strategy has been added as a recourse for deactivated frames (discussed later). By allowing these deactiveated frames to be reconsidered later mistakes caused by deactivating hypotheses because of atypical presentations can be detected. This should allow increased usage of the *must-have* and *must-not-have* features, and hence greater pruning of unlikely hypotheses without incurring the errors found in PIP. As with the triggers, the values of these slots are statements so syndromes and states can be used.

PIP also uses causal and associative links and the *differential-diagnosis* slot to first semi-activate and then possibly activate frames. The experience gained from running test cases and developing the syndrome frames indicated that semi-activation was not the correct action in many cases. It was often the case that there was a more precise action that could have been taken if it were possible to encode that information.

To try to capture some of this knowledge the *special* slot has been added. This slot has a list of condition-statements. These are used as pattern-action pairs: when a pattern is matched a specific action is taken. Since one of the possible actions is to activate a syndrome or state, this slot can be used to create hypotheses in the same manner as the *differential-diagnosis* slot in PIP but without first semi-activating the hypothesis. Semi-activation has not been implemented at all. Also, since statements are used as the patterns in the pattern-action pairs the situations can be specified more precisely than with PIP's *differential-diagnosis* slot. The details are discussed in subsection {8.2.6}.

Hypotheses in CDDS are more detailed than in PIP or INTERNIST. In addition to creating an instance of a syndrome or state, CDDS also hypothesizes the values of various properties of the syndrome or state, (e.g., the time of onset of the syndrome or state, the severity, etc.). These values are calculated from attached procedures in the syndrome or state frames and are updated as more information is known to the system.

III.4.2 Scoring

The major thrust in the development of the scoring algorithm in this system was to keep it as close to Bayes' rule as possible, justifying any deviations that are required. By doing this one has a mathematical justification for decisions made using these scores and, where available, probabilities can be objective rather than subjective measures.

An attempt to do this was outlined in Szolovits' paper: "Remarks on Scoring" [49]. I have used the ideas in that paper as a guide in developing a scoring algorithm, but some of the assumptions made there must be modified in this system.

Starting with the sequential version of Bayes' rule:

$$P_{i+1}(H_j) = \frac{P_i(S_k|H_j)}{P_i(S_k)} P_i(H_j) \qquad (1)$$

where

$$P_i(S_k) = \underset{j}{\text{SUM}}\, P_i(S_k|H_j)P_i(H_j) \qquad (2)$$

S_k is the k^{th} finding and H_j is the j^{th} hypothesis.

$P_i(H_j)$ is the Probability of H_j on the i^{th} iteration.

and

$P_i(S_k|H_j)$ is the conditional probability of S_k given H_j.

The first change in Bayes' rule that Szolovits proposes is to allow for uncertain observations. This will be useful since the physician does not always know for certain that a finding is present, especially historical findings and laboratory data. This modification results in:

$$P_{[i+1]}(H_j) = P_i(H_j)\left(Op(S_k)\frac{P_i(S_k|H_j)}{P_i(S_k)} + (1 - Op(S_k))\frac{1 - P_i(S_k|H_j)}{1 - P_i(S_k)}\right) \qquad (3)$$

where

Op(S) is the probability that a symptom has actually been observed.

These equations (eqs. 5 and 6 in [49]) have the correct limiting behavior for $Op(S_k)$. When $Op(S_k) = 1$ (i.e., when one is certain that S_k has been observed), this equation reduces to eq. (1). When $Op(S_k) = 0$ (i.e., when one is certain that $\sim S_k$ has

been observed) the equation reduces to (1) if one substitutes $\sim S_k$ for S_k as is appropriate. And when $Op(S_k) = P_i(S_k)$ (i.e., when the probability of S_k's having been observed is equal to the the probability of S_k occurring), the equation reduces to $P_{[i+1]}(H_j) = P_i(H_j)$, which is what one wants since no information was gained from the observation.

The above equation is the first part of the scoring algorithm used by the system. The second part of the algorithm is used for scoring hypotheses linked to other hypotheses. This algorithm is discussed below. Only the active syndromes are considered in either part of the algorithm. First the above equation is repeated for all of the known findings, then the next part of the algorithm is applied and then a normalization is performed.

Because of the representation of syndromes and findings, there are several problems that arise in trying to implement the first part of the algorithm. As in PIP a finding in CDDS can have many values, not just present or absent. To fit these findings into a usable form for equation 3, one must define a new finding whose presence or absence can be determined. This is easily done by making this new finding the concatenation of the original finding entered into the system with all its known items and their values. To find $P_i(S_k|H_j)$ one matches this pattern against those listed in the part of the scoring function of H_j which is associated with the finding. Each pattern in the scoring function has an associated probability which is $P_i(S_k|H_j)$. The patterns in the scoring function are statements and hence can encode interdependencies between findings by using logical operators. This allows the finding interdependencies to be incrementally extendable as described in [49].

A probability for each of the possible values of a finding need not be explicitly stated; the omission of an item implies that the same probability applies for all values of that item. Also all findings need not be listed in the scoring function of each syndrome or state. There is an a priori probability, a natural frequency of occurrence, for each value of each finding. This value is used if the finding is not matched to any of the patterns in the scoring function of a hypothesis and the finding cannot be explained by a linked clinical or physiological state. For example, one would not wish to list the finding *cleft palate* in all syndromes where that was not expected to occur since this would greatly increase the size of the database if it were done for all findings. So instead a default probability is given to the finding. This probability is defined as the probability of the finding, in this case cleft palate, occurring in a syndrome where it is not specifically expected to occur. Any syndrome for which this is an acceptable estimate of the probability of a cleft palate given the syndrome need not list the finding in its scoring function. All other syndromes will list the probability of cleft palate given that syndrome.

If the finding in a given hypothesis is explained by a linked state, the hypothesis is not used in calculations of the first part of the algorithm. Its previous score is used for the second part of the algorithm. Information about the effect of the finding on that hypothesis will be incorporated in the score by the second part of the scoring algorithm. By not using all of the hypotheses in this equation, the scores will not be normalized after this part of the algorithm.

Another problem involves the ability to use uncertainty of observations in the scoring algorithm. This becomes more complex when the findings have many items. The doctor may be unsure whether the patient has edema and whether that edema (if it exists) is cyclic, etc. Determining the certainty of the finding not being there is no longer a straightforward process. This system assumes that uncertainty of a finding is the product of the uncertainties of all its items.

After the first part of the algorithm has been completed for all of the findings and the active hypotheses, the second part is run. This part uses a similar equation except that syndromes and states are used in place of findings. It incorporates each of the links of a hypothesis into its score using equation 24 from [49]:

$$P'_{[h+1]}(H_j) = P'_h(H_j)\left(Op(H_k)\frac{P_h(H_k|H_j)}{P_h(H_k)} + (1 - Op(H_k))\frac{1 - P_h(H_k|H_j)}{1 - P_h(H_k)}\right) \quad (4)$$

where $P'_h(H_j) = P_m(H_j)$ which is the probability of H_j computed in the first part of the algorithm and

$$P_h(H_k) = \underset{j}{\text{SUM}}\, P_h(H_k|H_j)\, P'_h(H_j) \quad (5)$$

The summation is done over all $j \neq k$ if a link between H_j and H_k exists. If no link exists, the score remains unchanged. Also since $Op(H_k)$ can not be observed directly, $P'_h(H_j)$ is

used as an estimate of $Op(H_k)$.[45]

After this second part of the algorithm is completed the scores still do not necessarily sum to one, so a normalization (division of each score in a set by the sum of all the scores in that set) over all active hypotheses is done.

The scoring algorithm is not calculated incrementally as is the algorithm in [49] because scoring in this system is being done over only the active hypotheses, which is a dynamic set. So whenever a hypothesis is added or deleted the score is recalculated from scratch.[46]

In addition, this scoring algorithm (as well as the algorithm described in [49]) is somewhat dependent on the order in which the links are scored. This system attempts to order the scoring of hypotheses in the second part of the algorithm so that the evaluation is done from findings to states to syndromes. In other words, the system tries to score those states which are linked to other states, before those which are linked only to syndromes. Although this ordering could present a major problem, in practice one does not find a large interwoven net of links, rather one sees links that can easily be ordered in the way described above.

III.4.3 Diagnostic Strategy and Question Selection

There are several stages which the system goes through in determining an appropriate strategy. First, all of the hypotheses whose scores are not near the leading hypothesis are removed from further consideration. Next, the system decides whether multiple syndromes should be considered. This process is described later. If so, the system partitions these leading hypotheses into those competing with the leading hypothesis (called the partitioned hypotheses) and all others. These first steps are

---

45. Another scoring algorithm has been implemented. It avoids some of the difficulties involved in not having all of the hypotheses competing, (i.e., a clinical state and a syndrome can both be present). It uses the first part of the previous algorithm but for $P_i(S_k|H_j)$ when S_k occurs in linked state H_m it uses $P_i(S_k|H_m)\ P_i(H_m|H_j)$. After the the first part of the previous algorithm has been completed for all the findings, the system normalizes the scores of different sets of competing hypotheses depending on the situation. For example, if the system is attempting to conclude a clinical state it would exclude linked syndromes from the normalization, if it were concluding a syndrome it would exclude the linked states from the normalization. A comparison of the two scoring algorithms using the cases from the PIP and INTERNIST comparison showed little difference in their performance.

46. It could, in theory, be done incrementally but it would require more work than recalculating the score from scratch.

similar to INTERNIST's.

The system then determines, using only the partitioned hypotheses, whether certain hypotheses should be grouped together and treated as one. Grouping is done when more than two hypotheses have a common superior node in the syndrome hierarchy and their scores are close to other hypotheses in the partitioned hypotheses. The common superior syndrome is substituted for its inferiors in the set of partitioned hypotheses. This helps to prevent asking many questions of similar incorrect hypotheses, a problem encountered with INTERNIST and PIP.[47] After this is done the remaining hypotheses are used to select the strategy for question selection.

As in INTERNIST there are several different possible strategies that the system can use for question selection depending on the state of the program. The strategies of CDDS include: confirm, differentiate, ruleout, reevaluate, floundering.

These strategies use the pursue and ruleout slots, which each contain a list of findings, to choose the next question to be asked. The order of the findings in the slots is used to determine which finding should be inquired about next. This order can embed some of the knowledge about clinical style. This approach resulted from noting that INTERNIST's algorithm was not able to capture all of the clinical knowledge used in question selection, sometimes asking inappropriate questions, and that PIP's method rarely asked unreasonable questions even though they were sometimes not optimal.[48] The reason for the unexpected questions asked by INTERNIST, I feel, was that its dynamic selection algorithm is too simplistic. It fails to capture many subtleties of the situations. For example, often a laboratory test may be routinely run and hence be no more costly to ask than a symptom, etc.. By using these ordered lists of findings many of the subtleties of each syndrome can be captured. Yet by having more than one of these lists, by determining the possible effects of findings on other hypotheses' scores and by grouping similar hypotheses, I believe that many of PIP's flaws can be eliminated.

Clinical and physiological states and other syndromes can also be listed in the pursue and ruleout slots. When present they indicate to the program to use the same slot in that frame if the state or syndrome has not been concluded. If no question is found using the list in the other slot then the system returns to the list from the original slot continuing at the place of departure.

---

47. Asking only one question per iteration should also help prevent this.

48. PIP's method worked best when there was one clear leading hypothesis. If several were grouped at the top it would occasionally ask a question that was present in more than one of the leading hypotheses and hence did not help to differentiate between these hypotheses.

The confirm strategy is used when only one hypothesis is being considered. First, the system checks to see if the hypothesis can be concluded at this time (the conclusion of hypotheses is discussed later). If the hypothesis cannot be concluded yet, the system uses the first unasked question in the pursue slot of the finding as the next question.

The differentiate strategy is used if there are only two hypotheses left for consideration. The question selected is the first unasked finding in the pursue slot of the leading hypotheses which, if present, would cause the other hypothesis' score to decrease. This could be caused either by the finding having a low probability in the second hypothesis' scoring function or by the finding not being present in the second hypothesis (or active linked hypotheses) and having a low probability of natural occurrence.

The ruleout strategy uses the first unasked finding in the ruleout slot of the leading hypotheses. There are also backup methods in case the above question selection techniques fail to find an appropriate question.

The floundering strategy is used when the system has asked a large number of questions and none of the hypotheses have a very high score. The system first rescores all of the hypotheses that were ever active. Then, if no hypotheses appear very likely, it states that this might be an unknown birth defect (about 70% are) and inquires whether further investigation is desired. The user can also request that the deactivated hypotheses be reconsidered (the reevaluate strategy).

III.4.4 Multiple Syndromes

In the birth defects area and pediatrics in general multiple diseases and syndromes are much less common than in internal medicine (only about 5% of the birth defects are diagnosed as multiple syndromes, whereas 70% of birth defects are undiagnosed). In fact there were no multiple birth defects in the clinical cases used in this comparison. So it is difficult to judge the effectiveness of PIP and INTERNIST's methods of handling multiple birth defects and proposed corrections. It was possible to determine when multiple hypotheses where incorrectly hypothesized (since any time they were hypothesized it was incorrect). Of course, this information is not sufficient to develop an algorithm for handling multiple syndromes, so the method proposed here is very tentative and not tested.

It is abundantly clear that PIP's method of handling multiple syndromes is inadequate. In essence it merely pursues all active hypotheses until their scores become so low that they are deactivated, quitting only when there are no hypotheses left. This proves to be a very frustrating way to terminate the algorithm. PIP does not remove

findings explained by a syndrome after it has been concluded so they remain to be used by the other active hypotheses.

INTERNIST's approach of deleting all findings explained by concluded syndromes and then proceeding if an unexplained finding with a high import score remains works better than PIP's approach. But syndromes may have findings in common which would not be taken into account if all of the explained findings were removed.

This system employs a compromise solution: when a hypotheses is concluded all of the explained finding are temporarily removed. All of the hypotheses that were deactivated because of one of these findings are reactivated. The system continues to pursue the diagnosis if any of three conditions are met. First, if there are any findings remaining that have "must-be-explained" as the value of the *importance* slot (or several findings with values of "should-be-explained"). Second, if after rescoring with this smaller set of findings, any of the remaining hypotheses have a high score compared to the others and to the "unknown" syndrome. Third, if there are any links between the concluded syndrome and hypothesized syndromes (not clinical or physiological states) present or if there is an entry in the *special* slot of the concluded syndrome stating that if this hypothesis is concluded another hypothesis should be pursued.

If the system decides to continue with the diagnosis, it deactivates all the hypotheses that do not pass at least one of the above criteria, so only hypotheses that are deemed "reasonable" hypotheses without the manifestations explained by the concluded syndromes remain active. Then it continues the diagnosis using all of the known findings, although any new hypotheses must pass the above criteria and no hypotheses are deactivated because of a finding that is explained by a concluded syndrome. This process removes hypotheses that explain only already accounted for findings and no important other findings while allowing the remaining hypotheses to use the already explained findings.

The same three conditions mentioned above are also used to determine when consideration of multiple syndromes is necessary before concluding a syndrome. One or more of these conditions should be met when more than one of the leading hypotheses are actually present, hence the partitioning of the active hypotheses would be desirable to prevent the two correct hypotheses from competing with each other. If multiple syndromes are hypothesized to be present, CDDS does an INTERNIST like partition of the hypotheses but findings with low importance are ignored. It then removes from consideration the hypotheses which are not in the partitioned set with the leading hypothesis. It also removes any finding which is only explained by the hypotheses just removed from consideration. The removed hypotheses and findings are reinstated after the system chooses a diagnostic strategy and asks the next question of

the user (or concludes the leading hypothesis).

This system will not try to solve the problem of diagnosing multiple syndromes whose findings interact and overlap to the extent that neither is discernible without knowledge of these interactions. I believe that these cases are usually classified as unknown (or given a unique name if they occur often enough).

III.4.5 Concluding Hypotheses

Hypotheses are concluded by two methods in CDDS: by fulfilling the *is-sufficient* feature (if this is present in the frame) or by having the hypothesis' score exceed a threshold after a sufficient number of the hypothesis' findings have been asked. This second method is checked only when in the confirm strategy.

A "sufficient" number of findings for a hypothesis is defined as all of the findings in its *must-not-have* slot and any other findings that are specifically stated to be required to be asked before confirming the hypothesis (this can be done via use of the *special* slot). It is necessary to check to make sure that a sufficient number of questions have been asked before concluding the hypothesis since, with normalization over the active hypotheses, the scores can be very high without there being any confidence in the hypotheses. The best example of this is when one has entered only one finding and the leading hypothesis is the only one triggered (hence it would always have a score of 1.0). Of course, it usually would not be the case that one finding would be sufficient to justify concluding the hypothesis but its score is certainly above any possible threshold. Another measure to prevent this from happening is the introduction of an unknown syndrome which is always active. Its score denotes the probability of the findings randomly occurring (assuming all findings occur independently).

The score which is checked against the threshold is not precisely the same score calculated in the scoring algorithm. The score calculated in that algorithm is renormalized over a subset of the active hypotheses. This subset includes all the active hypotheses with all hypotheses linked to the leading hypothesis removed. This leaves only the hypotheses competing with the leading hypothesis and the leading hypothesis itself. If multiple syndromes are postulated then in addition to the above only the set of partitioned hypotheses are included (i.e., only hypotheses that are competing with the lead hypothesis are used).

If a syndrome is concluded and it is not a terminal node in the syndrome hierarchy, the system activates all of its inferiors (noting that the syndrome is concluded in case no further conclusions are possible). This is somewhat similar to Patil's *refine* strategy [25], although he may refine the hypothesis even though a superior is not concluded. If a clinical or physiological state is concluded its score is set to 1.0 and it is

put on the list of concluded syndromes and removed from the list of active hypotheses.

The categorical decisions of the IS-SUFFICIENT feature is used mostly by the clinical and physiological states in CDDS. The probabilistic scoring is more useful in concluding syndromes. This has been noted by others [25,45].

III.4.6 Exception Handling

The *special* slot is used to alter the flow of control of the system. This is desirable for two reasons. One, it was found, when running PIP and INTERNIST on cases, that certain situations arose where the action taken by the system was inappropriate. This is bound to occur when using a few simple methods to handle all possible situations. A better action was often known but encoding this knowledge into these systems was not possible.

The second reason stems from the observation that experts in medicine appear to acquire information which, while they are pursuing one hypothesis, allows them to recognize a pattern of answers indicating that a different hypothesis may be worth investigating. This allows these experts to avoid backtracking and arrive at the solution in a more direct manner. This is the intent of PIP's semi-activated state with causal and associative links and with the *differential-diagnosis* feature. Unfortunately, semi-activation alone does not capture the experts' ability to redirect their focus. This sort of knowledge is not easily put into simplistic algorithms; the correct action is usually dependent on the situation.

To overcome these problems the *special* slot has been introduced. This slot contains "conditional statements", i.e. condition-action pairs, which are checked for each active hypothesis after the evaluation of new findings and again after the scoring is done. These pairs or rules are written using a small set of built-in primitives. The list of possible action functions are: deactivate, activate, trigger, conclude, ask, set-strategy, keep-active and check-for-multiple-syndromes. The conditional part is a *statement* and hence can contain any information about the entered findings, hypotheses, or the global state of the system.

In addition the user can interrupt the diagnosis at any time and check the status of the system. The user can also change the state of the system by entering the actions listed above, i.e. he can activate or deactivate hypotheses, select the strategy, etc.. Assuming that the system is not perfect, this should avoid some user frustration by allowing him to guide the system when he feels its algorithm is non-optimal.

III.5 Implementation

The system described above has been implemented in Maclisp on a PDP-10.[49] A database has been constructed using the information from the PIP and INTERNIST congenital defects databases and some additional information. Unfortunately to construct a database that would take advantage of all the features of CDDS would require a large amount of additional effort, so only a small fraction of the syndromes and findings were totally modified. The rest were only modified to the extent that was required for CDDS to function, and to check its performance.

III.6 Summary of CDDS

CDDS is a frame based system which offers increased flexibility for the designer of the database and the physician using the system. The system has a representation of knowledge which allows more precise encoding of information about the syndromes as well as information about the diagnostic process pertaining to those syndromes. The algorithm has been altered to eliminate some of the problems found in PIP and INTERNIST. A more robust method for handling exceptional cases has also been added.

---

49. Some of the features described here but not used by the database or the test cases have not been implemented yet.

Appendix IV - The Performance of the Congenital Defects Diagnostic System

A database for CDDS was constructed using the information contained in the INTERNIST, PIP, and Center for Birth Defects Information Services databases. Where available, probabilities taken from the literature were used in the scoring algorithm. In addition, CDDS was capable of representing knowledge not contained in any of the other three databases. Since it is a major undertaking to extract this knowledge from the literature and/or physicians in order to represent in the CDDS database, this information was added only for one group of defects, those involving congenital hypothyroidism. After this database was constructed the performance of CDDS, compared to that of PIP and INTERNIST, was investigated using the same cases that were used in the comparison of PIP and INTERNIST. This chapter reports the results.

IV.7 Hypothesis Generation

The Congenital Defects Diagnostic System's triggering algorithm differs from PIP's and INTERNIST's in two aspects: 1) it allows activation or conclusion of clinical states and syndromes to directly trigger hypotheses and 2) it allows combinations of findings (and clinical states and syndromes) to cause a hypothesis to be created. The CDDS database converted the *major-cause-of* links of PIP's database to cause triggering of associated syndromes after the clinical state was concluded. CDDS also allows greater flexibility in the deactivation of hypothesis as described in the previous chapter.

The cases were entered and the number of hypotheses generated and the number of inappropriate hypotheses generated both after the initial findings were entered and at the end of the session were determined. These results are listed in table X.

CDDS's hypothesis generating algorithm generated slightly fewer inappropriate hypotheses both after the initial findings were entered and at the end of the case, an average of 1.0 fewer ($p < 0.005$) after the initial findings were entered and 1.2 fewer ($p < 0.005$) at the end of the case. In all of these cases CDDS's algorithm hypothesized the correct syndrome or the major clinical state of the correct syndrome after the initial

Table X. Number of Hypotheses Generated

Case	Number of Initial Findings	Number of Hypotheses Generated ** CDDS		PIP	
		Initial	Total	Initial	Total
1	26	5/0	4/0	10/5	11/4
2	17	5/0	5/0	5/0	5/0
3	16	5/0	4/0	5/0	5/0
4	24	8/2	6/1	9/3	9/3
5	6	4/0	2/0	6/1	6/1
6	5	1/0	1/0	3/1	5/3
7	11	4/1	4/1	5/1	5/1
8	21	3/2	3/2	3/2	3/2
9	8	2/0	2/0	4/1	4/1
10	14	2/0	3/0	3/1	5/2
11	19	5/1	5/1	9/5	9/5
12	9	5/0	5/0	4/1	6/3
13	15	6/0	5/0	6/2	1/0
14	8	3/1	3/1	8/2	7/2
15	15	4/0	4/0	8/1	8/1
16	11	4/0	4/0	5/0	5/0
17	13	5/1	5/1	9/3	9/3
18	25	6/1	5/0	3/0	5/0
19	20	5/0	5/0	6/1	7/1
20	13	2/0	2/0	3/3	4/2
21	27	5/0	4/0	7/3	7/3
22	22	1/0	1/0	2/0	2/0
23	10	5/1	4/1	7/3	4/2
24	17	1/0	1/0	3/0	4/0
25	22	1/0	1/0	5/0	5/0
26	11	1/0	1/0	3/0	3/0
27	8	1/0	1/0	3/1	3/1
28	11	3/0	5/0	9/1	10/3
29	8	5/0	3/0	6/0	2/0
30	14	2/0	2/0	3/0	3/0
31	15	8/6	8/6	9/6	8/7
32	8	3/0	3/0	6/2	9/5
33	10	3/2	3/2	4/1	4/1
34	9	3/0	3/0	4/2	6/2
35	13	3/1	3/0	3/1	2/2
Mean	14.3	3.7/0.5	3.4/0.5	5.4/1.5	5.5/1.7
Median	13	4/0	3/0	5/1	5/1
S.D.	6.0	1.9/1.1	1.7/1.1	2.3/1.5	2.5/1.7

** Numerator is the total number of hypotheses generated.
Denominator is the number of inappropriate hypotheses generated.

findings were entered.[50]

These results seem to indicate that CDDS's methods for generating hypotheses (the use of combinations of findings as triggers, the use of clinical states as triggers, etc.) is an improvement over PIP's method. It decreases the number of inappropriate hypotheses generated as compared to PIP or INTERNIST while still generating the correct hypothesis as early in the diagnosis as PIP or INTERNIST (earlier than PIP in one case).

IV.8 The Scoring Algorithm

CDDS's scoring algorithm was tested in the same manner as PIP's and INTERNIST's, by examining the rank of the correct hypothesis after the given set of findings were entered. As with the comparison with PIP and INTERNIST, INTERNIST's hypotheses generation algorithm was used so CDDS's decreased number of hypotheses did not artificially raise the rank of the correct hypothesis. The results of this comparison, using the initial findings as the set of findings entered, are shown in table XI.

The results of this comparison indicate that the scoring algorithm used by CDDS on the average ranked the correct hypotheses slightly better than did PIP's or INTERNIST's algorithm, a mean of 1.5 for CDDS as compared to 1.8 for PIP and 1.9 for INTERNIST. This difference is just statistically significant ($p < 0.05$). This result indicates that CDDS's scoring algorithm performs at least as well as PIP's and INTERNIST's scoring algorithms, although a larger sample size is needed to determine with certainty if there is any increased performance using CDDS's algorithm. This comparison did not take into account the effect of CDDS's partitioning algorithm on the ranking of the correct hypothesis since there were no cases with multiple syndromes (this was, of course, also true of INTERNIST's partitioning algorithm).

---

50. The reason that only clinical states were hypothesized by CDDS in some cases whereas in PIP both clinical states and the associated syndromes were hypothesized is that PIP often pursued specific reasons for a clinical state's presence before it had proven that the clinical state was present (i.e., pursuing vitamin-d-resistant-rickets before determining rickets is likely to be present. To avoid this CDDS uses the clinical state as a trigger so the more specific syndromes are not triggered until the clinical state is shown to be present. This is not done for all syndromes with linked clinical states, only those syndromes which are really explanations for the clinical state.

Table XI. Rank of Correct Hypothesis after Initial Findings Entered

Case	Number of Initial Findings	Rank PIP	Rank INTERNIST	Rank CDDS
1	26	6	3	2
2	17	1	1	1
3	16	1	1	1
4	24	1	1	1
5	6	3	5	2 *
6	5	1	1	1
7	11	1	1	1
8	21	1	1	1
9	8	1	1	1
10	14	1	1	1
11	19	1	2	1
12	9	4	5	5
13	15	4	6	5
14	8	2	1	1
15	15	1	1	1
16	11	1	1	1
17	13	1	1	1
18	25	1	2	2
19	20	1	1	1
20	13	1	1	1
21	27	1	1	1
22	22	1	1	1
23	10	2	2	1
24	17	1	1	1
25	22	1	1	1
26	11	1	1	1
27	8	1	1	1
28	11	6	5	3 *
29	8	5	2	2
30	14	1	1	1
31	15	2	1	1
32	8	4	5	3
33	10	1	1	1
34	9	1	1	1
35	13	6	3	3
Mean	14.7	1.9	1.8	1.5
Median	14	1	1	1
S.D.	6.0	1.6	1.5	1.1

* Rank of hypothesized clinical state of correct syndrome.

IV.9 Concluding Syndromes

The algorithm used to confirm syndromes in CDDS is similar to PIP's algorithm in that it requires a syndrome's score to surpass a predetermined threshold. But, since CDDS's scoring algorithm normalizes the score over all of the competing hypotheses,[51] its confirmation algorithm does require that the difference between the scores of the top two hypotheses exceed a certain amount before the top hypothesis can be concluded. The threshold used in CDDS in these comparisons is set at .98, hence the difference between the top two hypotheses must be greater than .96.[52]

The results from running the thirty-five cases showed that CDDS concluded the correct syndrome in thirty-four of the cases. In one case, case 5, no syndrome was concluded (both PIP and INTERNIST also failed to conclude the correct hypothesis in this case). The reason for CDDS's failure to conclude the correct syndrome was the same as PIP's and INTERNIST's, lack of laboratory data. CDDS did conclude (as did PIP) the major clinical state in this syndrome, hypothyroidism. CDDS went on to detect that the session was not getting anywhere and asked the user what he wished to do.

These results indicate that CDDS's algorithm avoids the pitfalls of PIP's confirmation algorithm. It performs as well as INTERNIST'S algorithm. This is to be expected since CDDS, like INTERNIST, takes into account the scores of the nonleading hypotheses in its confirmation algorithm.

IV.10 Diagnostic Strategy and Question Selection

To compare diagnostic strategy and question selection in CDDS to that in PIP and INTERNIST the number of questions required to conclude the correct syndrome by each system was determined. This is shown in table XII.

The results of this comparison show that CDDS asks fewer questions before concluding the correct hypothesis. The mean number of questions asked for CDDS was 5.6 as compared to 9.7 and 15.4 for PIP and INTERNIST respectively. These differences were statistically significant for both PIP and INTERNIST ($p < 0.05$ for PIP and $p < 0.001$ for INTERNIST). There are still wide case to case variations in the number of questions asked for all systems, probably due to the variation in the difficulty of the cases. The

---

51. As stated in the last chapter, there is a partitioning of hypotheses into competing subsets, if the system thinks that there might be more than one hypotheses present.
52. Actually, CDDS uses a stronger criterion: the difference between the leading hypothesis and the sum of all the scores of the other hypotheses must be greater than .96.

Table XII. Number of Questions Required to Conclude the Correct Hypothesis

Case	Number of Initial Findings	Number of Questions Asked		
		INTERNIST	PIP	CDDS
1	26	49	46	10
2	17	0	14	6
3	16	18	12	6
4	24	10	0	1
5	6	*	*	*
6	5	9	3	6
7	11	0	18	0
8	21	15	0	0
9	8	0	1	0
10	14	20	4	2
11	19	19	4	7
12	9	31	*	11
13	15	58	*	48
14	8	43	22	17
15	15	15	1	1
16	11	0	18	0
17	13	0	16	0
18	25	24	8	14
19	20	5	29	3
20	13	4	3	0
21	27	19	2	4
22	22	0	0	0
23	10	78	*	47
24	17	0	0	0
25	22	0	4	0
26	11	0	0	0
27	8	0	4	0
28	11	28	5	23
29	8	49	6	33
30	14	0	0	0
31	15	20	4	0
32	8	76	53	23
33	10	0	3	0
34	9	25	1	2
35	13	29	20	16
Mean **	14.3	15.4	9.7	5.6
Median **	13	10	4	1
S.D. **	6.0	18.6	13.2	8.6

* -- Correct hypothesis was not concluded.
** -- Calculations based on cases where data is present for all systems.

variations between the systems on the same case appear to be due to three factors: 1) the diagnostic strategy and question selection algorithm, 2) the confirmation algorithm[53], and 3) luck. The third factor occurs when the system asks a fortuitous question which leads it in the correct direction and decreases the total number of questions asked. An example of this is when one system is attempting to pursue the leading (but incorrect) hypothesis and asks about a finding which happens to be important to the correct hypothesis and this raises its rank to the top. Another system might ask about a different (but just as valid) finding that did not happen to be important to the correct hypothesis, thus causing more questions to be asked before the correct hypothesis becomes the leading hypothesis. Since the first system was not considering the correct hypothesis in its question selection algorithm it was just coincidental that a finding was inquired about that caused the system to pursue the correct hypothesis and hence shorten the diagnostic session.

The contribution of the confirmation algorithm to the length of the diagnostic session can be eliminated in the same manner as in chapter 5, by determining the number of questions asked in order to cause the correct hypothesis to be pursued.[54] The results of this comparison is shown in table XIII.

The results of this comparison show that the number of findings inquired about by CDDS in order to pursue the correct hypothesis is less that the number required by PIP or INTERNIST. This difference is statistically significant between CDDS and INTERNIST ($p < 0.01$) but not between CDDS and PIP ($p > 0.05$). An increased sample size is required to determine whether the quicker focusing on the correct hypothesis by CDDS is indeed the case.[55]

---

53. This is most noticeable between PIP's algorithm and those of the other two systems. Unfortunately, PIP's confirmation algorithm does not appear to work as well as INTERNIST's and CDDS. Some of the cases where PIP quickly concludes the correct hypothesis but INTERNIST and CDDS do not are probably premature conclusions. This is borne out by the fact that PIP concluded incorrect hypotheses several times when CDDS and INTERNIST did not.

54. As in chapter 4, the point at which the correct hypothesis is being pursued is defined as the point in the diagnostic session where the correct hypothesis is the leading hypothesis and remains so for the remainder of the session.

55. One reason that the difference between PIP and CDDS was not statistically significant might be that the ordering of the findings in CDDS's confirm and ruleout slots was essentially the same as the order generated by INTERNIST. The ordering in PIP's disease frame's was more detailed than CDDS's, since it also used information from the Center for Birth Defects Information Services' algorithm. With a more careful ordering of the findings the difference between CDDS and PIP in the number of findings required to focus on the correct hypothesis might increase to a statistically significant level.

Table XIII. Questions Required to Pursue Correct Hypotheses

Cases*	Number of Initial Findings	Number of Questions Asked INTERNIST	PIP	CDDS
1	26	40	27	6
5	6	4	0	1
11	19	0	3	0
12	9	17	19	6
13	15	36	19	36
14	8	14	0	5
18	25	0	1	1
23	10	4	24	0
28	11	9	3	6
29	8	15	4	12
31	15	5	0	0
32	8	44	11	9
35	13	24	9	7
Mean	14.3	16.3	9.2	7.4
Median	13	9	4	6
S.D.	6.0	14.1	9.8	9.7

* Cases initially pursuing the correct syndrome in all systems removed.

IV.11 Specialized Features

CDDS and the associated database contains features and knowledge that PIP and INTERNIST did not contain, among these are: the time expert, the floundering and careful modes, and the control of the diagnostic process by daemons associated with syndromes. These additions were motivated by the results of the comparison of PIP and INTERNIST. It would be unfair to evaluate these special case features using the same cases that motivated the changes,[56] of course if they failed to work on the these cases it would raise grave doubts on their effectiveness. It appears that floundering and careful modes worked on two cases they were designed for, case 5 and case 35. Careful mode reactivated the correct syndrome in case 35 and floundering mode was able to detect when the system was not making progress in case 5. The more detailed representation of time, which was used in the syndromes with hypothyroidism, was able to distinguish

56. In order to prevent the addition of information motivated by the clinical cases from effecting the previous comparisons, the syndrome specific information was kept in a separate database and tested at a different time.

between appropriate and non-appropriate instances of a finding as they related to hypothyroidism. Some of the other added information was not used in the diagnosis, but it did not hinder the system. More testing is required to determine the benefits of this information.

IV.12 Summary

The tests described in this chapter have shown that the more explicit control of the generation of hypotheses by CDDS does improve performance in that it decreases the number of inappropriate hypotheses (as compared to PIP and INTERNIST) while still hypothesizing the correct syndrome at least as early as PIP or INTERNIST. CDDS's use of a scoring algorithm based more closely on Bayes' rule has been shown to be at least as good as (and possible slightly better than) the more ad hoc approaches of the other two systems. The confirmation algorithm of CDDS, being very similar to INTERNIST's, performs as well as that of INTERNIST. The diagnostic strategy used by CDDS focuses on the correct hypothesis quicker than INTERNIST and possibly PIP.

The comparison of CDDS to PIP and INTERNIST did not examine the methods by which any of systems handle multiple syndromes due to the dearth of such cases (the same was true of the previous comparison of PIP and INTERNIST in chapter four). Because of this lack of multiple syndromes the value of the partitioning algorithms of both INTERNIST and CDDS could not be examined in this comparison.

These results are consistent with my expectations. CDDS seemed to perform approximately the same as or slightly better than PIP and INTERNIST in most cases. Any slight overall improvement was probably due to the selection of the superior algorithms from PIP and INTERNIST. In a few instances CDDS performed much better. In general, these instances were ones that the straightforward algorithms of PIP or INTERNIST did not cope with. In these instances CDDS was able to perform better because of either additions to its algorithm or the presence of instance-specific information. One would expect even better performance as more of the case-specific knowledge was incorporated into the database.

References

1. Betaque, N. E., and Gorry, G. A., Automating Judgmental Decision Making for a Serious Medical Problem, *Management Science* **17**:8: B-421 - B-434 (April 1971).

2. Bleich, H. L., Computer-Based Consultation: Electrolyte and Acid-Base Disorders, *American Journal of Medicine* **53**:285 (1972).

3. Buchanan, B. G., and Lederberg, J., "The Heuristic DENDRAL Program for Explaining Empirical Data, *Proceedings of the IFIP Congress 1971*, Ljubljana, Yugoslavia (1971). Also Artificial Intelligence Laboratory, Stanford University, Technical Report AIM-141 (1971).

4. Charniak, E., Toward a Model of Children's Story Comprehension, Ph.D. thesis, MIT Artificial Intelligence Laboratory Technical Report TR-266, December 1972.

5. Conklin, J. M., Munzenrider, J., Neurath, P. W., and Ross, W. M., Computer-Aided Medical Decision Making in Radiotherapy, *Radiology* **123**:411-466 (1977)

6. Davis, R., Applications of Meta Level Knowledge to the Construction, Maintenance and Use of Large Knowledge Bases, Ph.D. thesis, Stanford Artificial Intelligence Laboratory Memo 283 (1976).

7. Davis, R., Buchanan B., and Shortliffe, E., Production Rules as a Representation for a Knowledge-Based Consultation Program, *Artificial Intelligence* **8**:15-45 (1977).

8. Davis, R., Generalized Procedure Calling and Content-Directed Invocation, *SIGART Newsletter Proceedings of Symposium on Artificial Intelligence and Programming Languages*, August (1977).

9. Doyle, J., A Glimpse of Truth Maintenance, MIT Artificial Intelligence Laboratory Memo 461a (1978).

10. Duda, R. O., Hart, P. E., and Nilsson, N. J., Subjective Bayesean methods for rule-based inference systems, *Proceedings of the 1976 National Computer Conference*, AFIPS Press (June 1976).

11. Elstein, A. S., Shulman, L. A., and Sprafka S. A., *Medical Problem Solving*, Harvard University Press (1978).

12. Feinstein, A. R., *Clinical Judgment*, Krieger Publishing Co., Inc. (1967).

13. Feinstein, A. R., Clinical Biostatistics XXXIX. The haze of Bayes, the aerial palaces of decision analysis, and the computerized Ouija board, *Clinical Pharmacology and Therapeutics* **21**:4:482-496 (April 1977).

14. Gorry, G. A., and Barnett, G. O., Sequential Diagnosis by Computer, *JAMA* **205**:12:141-146 (1968).

15. Gorry, G. A., Kassirer, J. P., Essig, A., and Schwartz, W. B., Decision Analysis as the Basis for Computer-Aided Management of Acute Renal Failure, *The American Journal of Medicine* **55**:473-484 (1973).

16. Gorry, G. A., Silverman, H., and Pauker, S. G., Capturing Clinical Expertise: A Computer Program that Considers Clinical Responses to Digitalis, *American Journal of Medicine* **64**:452-460, (March 1978).

17. Gorry, G. A., Pauker, S. G., and Schwartz, W. B., The Diagnostic Significance of the Normal Finding, *N Engl J Med* **298**:486 (1978).

18. *Harrison's Principles of Internal Medicine*, 7th ed., McGraw-Hill (1978).

19. Kahn, K.M., Mechanization of Temporal Knowledge, Laboratory for Computer Science, Massachusetts Institute of Technology, MAC-TR 155 (1975).

20. Kassirer, J. P., and Gorry, G. A., Clinical Problem Solving: A Behavioral Analysis, *Ann. Int. Med.* **89**:245, (1978).

21. Knill-Jones, R. P., Stern, R. B., Grimes, D.H., Maxwell, J.D., Thompson, R.P.H., Williams, R., Use of a Sequential Bayesian Model in Diagnosis of Jaundice by Computer, *Brit. Med.* 1:530-533, (1973).

22. Minsky, M., A Framework for Representing Knowledge, *The Psychology of Computer Vision*, Edt. Winston, P. W., McGraw-Hill (1975).

23. Miller, P. B., Strategy Selection in Medical Diagnosis, Project MAC, Massachusetts Institute of Technology, Technical Report TR-153 (September 1975).

24. Newell, A., Simon, H.S., *Human Problem Solving*, Prentice-Hall (1972).

25. Patil, R. S., Design of a Program for Expert Diagnosis of Acid Base and Electrolyte Disturbances, MIT Laboratory for Computer Science TM-132 (May, 1979).

26. Pauker, S. G., and Kassirer, J. P., Therapeutic Decision Making: A Cost-Benefit Analysis, *N Engl J Med* **293**:229-234 (July, 1975).

27. Pauker, S. G., Gorry, G. A., and Kassirer, J. P., and Schwartz, W. B., Toward the Simulation of Clinical Cognition: Taking a Present Illness by Computer, *The American Journal of Medicine* **60**:981-995 (June 1976).

28. Pauker, S. G., and Szolovits, P., Analyzing and Simulating Taking the History of the Present Illness: Context Formation, in Schneider/Sagvall Hein, eds., *Computational Linguistics in Medicine*, North-Holland (1977).

29. Pople, H. E., Jr., Artificial-Intelligence Approaches to Computer-Based Medical Consultation, IEEE Intercon Conference (1975).

30. Pople, H. E., Jr., Myers, J. D., and Miller, R. A., DIALOG: A Model of Diagnostic Logic for Internal Medicine, *Advance Papers of the Fourth International Joint Conference on Artificial Intelligence*, available from the Artificial Intelligence Laboratory, Massachusetts Institute of Technology (1975).

31. Pople, H. E., Jr., Presentation of the INTERNIST System, Proceedings of the A.I.M Workshop, Rutgers Research Resource on Computers in Biomedicine, Rutgers University (June 1976).

32. Pople, H. E., Jr., The Formation of Composite Hypotheses in Diagnostic Problem Solving: an Exercise in Synthetic Reasoning, *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, available from the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213. (1977).

33. Roberts, R. B., and Goldstein I.P.,The FRL Primer, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Technical Report AI Memo 408 (1977).

34. Roberts, R. B., and Goldstein I.P.,The FRL Manual, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Technical Report AI Memo 409 (1977).

35. Rubin, A. D., Hypothesis Formation and Evaluation in Medical Diagnosis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Technical Report AI-TR-316 (January 1975).

36. Schwartz, W. B., Medicine and the Computer: The Promise and Problems of Change, *N Engl J Med* **283**:1257-1264 (1970).

37. Schwartz, W. B., Gorry, G. A., Kassirer, J. P, and Essig, A., Decision Analysis and Clinical Judgment, *The American Journal of Medicine* **55**:459-472 (October 1973).

38. Shortliffe, E. H., and Buchanan, B. G., A Model of inexact reasoning in medicine, *Mathematical Biosciences* **23**:351-379 (1975).

39. Shortliffe, E. H., *Computer Based Medical Consultations: MYCIN*, Elsevier North Holland Inc. (1976).

40. Shortliffe, E. H., Knowledge Engineering for Medical Decision Making: A Review of Computer-Based Clinical Decision Aids, *Proceeding of the IEEE*, forthcoming (1979).

41. Silverman, H., A Digitalis Therapy Advisor, Project MAC, Massachusetts Institute of Technology, Technical Report TR-143 (1975).

42. Smith, B. C., A Proposal for a Computational Model of Anatomical and Physiological Reasoning, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, AI-Memo 493 (1978).

43. Swartout, W. R., A Digitalis Therapy Advisor with Explanations, Laboratory for Computer Science, Massachusetts Institute of Technology, Technical Report TR-176 (February 1977).

44. Szolovits, P., and Pauker, S. G., Research on a Medical Consultation System for Taking the Present Illness, *Proceedings of the Third Illinois Conference on Medical Information Systems*, University of Illinois at Chicago Circle (November 1976).

45. Szolovits, P., Pauker, S. G., Categorical and Probabilistic Reasoning in Medical Diagnosis, *Artificial Intelligence* 11,:115-144 (1978).

46. Szolovits, P., Coordinating the Use of Categorical and Probabilistic Reasoning, *Proceeding of the 1978 Computer Laboratory Health Care Resources Workshop,* Texas Technical University School of Medicine, Lubbock, Texas (January 1978).

47. Szolovits, P., The Lure of Numbers--How to Live With and Without Them in Medical Diagnosis, *Proceeding of the Colloquium in Computer-Assisted Decision Making using Clinical and Paraclinical (Laboratory) Data,* 1978.

48. Szolovits, P., and Pauker, S. G., Computers and Clinical Decision Making: Whether, How, and For Whom? *Proceeding of the IEEE*, in press (1979).

49. Szolovits, P., Remarks on Scoring, Unpublished classnotes, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (1979).

50. Tversky, A., and Kahneman, D., Judgment under Uncertainty: Heuristics and Biases, *Science* 185:1124-1131 (September 1974).

51. Weiss, S. M., A System for Model-Based Computer-Aided Diagnosis and Therapy, Ph.D. Thesis, Computers in Biomedicine, Department of Computer Science, Rutgers University, CBM-TR-27-Thesis (June 1974).

52. Weiss, S. M., Kulikowski, C. A., Amarel, S., and Safir, A., A Model-Based Method for Computer-Aided Medical Decision-Making, *Artificial Intelligence* 11: 145-172 (1978).

This blank page was inserted to preserve pagination.

CS-TR Scanning Project
Document Control Form

Date : 8/3/95

Report # LCS-TR-283

Each of the following should be identified by a checkmark:
Originating Department:

☐ Artificial Intellegence Laboratory (AI)
☒ Laboratory for Computer Science (LCS)

Document Type:

☒ Technical Report (TR) ☐ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 139 (144-IMAGES)
Not to include DOD forms, printer instructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☐ Laser Print
☐ InkJet Printer ☒ Unknown ☐ Other:_____

Check each if included with document:

☐ DOD Form ☐ Funding Agent Form ☐ Cover Page
☒ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages (by page number): _____

Photographs/Tonal Material (by page number): _____

Other (note description/page number):

Description : Page Number:

IMAGE MAP: (1-2) UNNUMBERED TITLE AND ACKNOWLEDGEMENTS PAGES.
(3-139) PAGES # 3-139
(140-144) SCANCONTROL, SPINE, TRGT'S (?)

Scanning Agent Signoff:
Date Received: 8/3/95 Date Scanned: 8/21/95 Date Returned: 8/21/95

Scanning Agent Signature: Michael W. Cook

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**

Scanned

Date: 8/21/1995

M.I.T. Libraries
Document Services