8

INTERNIST-1, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine

Randolph A. Miller, Harry E. Pople, Jr., and Jack D. Myers

One of the best-known AIM systems is the large diagnostic program constructed by researchers at the University of Pittsburgh during the 1970s. The work developed out of a collaboration between Harry Pople (a computer scientist with an interest in AI, logic programming, and medical applications) and Jack Myers, university professor (medicine) and prominent clinician, who was eager to try to encode some of his diagnostic expertise in a high-performance computer program. Rather than selecting a small subtopic in medicine for the work, Pople and Myers decided to consider the entire field of internal medicine. This necessarily required approaches that quickly narrowed the search space of possible diseases and also permitted case analyses in which two or more diseases could coexist and interact. The resulting program, now known as INTERNIST-1 (or INTERNIST, for short), is capable of making multiple and complex diagnoses in internal medicine. It differs from other programs for computer-assisted diagnosis in the generality of its approach and in the size and diversity of its knowledge base.

The knowledge base was developed over several years by Myers and medical student assistants. One of these students, Dr. Randolph Miller, became involved in the programming as well and, as a clinical faculty

member at the University of Pittsburgh, continues as a principal collaborator on the project. Those building the knowledge base would study the major diseases in medicine one by one, identifying both their major and minor clinical manifestations and developing weights that link each finding with the diseases in which it can occur. The resulting ad hoc scoring scheme proved to be capable of guiding excellent diagnostic reasoning. To test the program during its development, Myers and his students would select especially difficult cases for consideration, often ones drawn from published clinical pathological conferences in medical journals.

After several years of testing and refinement of the knowledge base, the study outlined in the following chapter was performed. To document the strengths and weaknesses of the program, the group performed a systematic evaluation of the program's capabilities. Its performance on a series of 19 clinicopathological exercises ("Case Records of the Massachusetts General Hospital"), published in the New England Journal of Medicine, *appeared qualitatively similar to that of the hospital clinicians but inferior to that of the case discussants. As a result, Miller, Pople, and Myers believe that the evaluation demonstrated that the present form of the program is not sufficiently reliable for clinical applications. They cite specific deficiencies that must be overcome before the program is ready for clinical use: an ability to construct differential diagnoses spanning multiple problem areas, new methods to avoid occasional attribution of findings to improper causes, and human-engineering enhancements to allow the program to explain its "thinking." A more detailed discussion of the serious limitations in the underlying representation and control methods used in INTERNIST-1 has recently been presented by Pople (1982). In that article Pople explains the contemplated enhancements that will be the basis for the next version of INTERNIST, to be known as CADUCEUS.*

8.1 Introduction

INTERNIST-1, an experimental program for computer-assisted diagnosis in general internal medicine, differs considerably in scope from other medical diagnostic computer programs. In the past, techniques including mathematical modeling, use of Bayesian statistics, pattern recognition, and other approaches (Wardle and Wardle, 1978; Wagner et al., 1978) (see also Chapter 3), have been shown to be useful in circumscribed areas such as the differential diagnosis of abdominal pain (deDombal et al., 1972) and the diagnosis and treatment of meningitis (Yu et al., 1979a). However, no program developed for use in a limited domain has been successfully adapted for more generalized use. From its inception, INTERNIST-1 has addressed the problem of diagnosis within the broad context of general internal medicine (Pople et al., 1975; Myers et al., 1982; Pople, 1982). Given a patient's

initial history, results of a physical examination, or laboratory findings, INTERNIST-1 was designed to aid the physician with the patient's work-up in order to make multiple and complex diagnoses. The capabilities of the system derive from its extensive knowledge base and from heuristic computer programs that can construct and resolve differential diagnoses.

The INTERNIST-1 program represents an example of applied symbolic reasoning (artificial intelligence). A variety of such techniques have been developed by computer scientists in an attempt to model the thought processes and problem-solving methods employed by human beings (Winston, 1977; Nilsson, 1980). An important aspect of the INTERNIST-1 approach to computer-assisted diagnosis is that the program attempts to form an appropriate differential diagnosis in individual problem areas. A *problem area* is defined as a selected group of observed findings, the differential diagnosis of which forms what is assumed to be a mutually exclusive, closed (i.e., exhaustive) set of diagnoses. Physicians routinely construct such closed differential diagnoses on the basis of causal considerations (e.g., bacterial pneumonias) or pathoanatomic considerations (e.g., causes of obstructive jaundice). By constructing specific differential diagnoses to address identified problem areas, a physician or computer program can narrow the set of possible diagnoses from all known diseases to well-defined collections of competing diagnoses in a small number of categories. Heuristic principles, such as diagnosis by exclusion, can then be employed to resolve each differential diagnosis. The use of such strategies in INTERNIST-1 represents an attempt to model the behavior of physicians.

Reported below is the first systematic evaluation of INTERNIST-1. The purpose of the study was to illustrate the strengths and weaknesses of the program and to provide a rough estimate of its clinical acumen. The trial was conducted with clinicopathological conferences (CPC's) that had been published in the *New England Journal of Medicine* (NEJM) but had not previously been analyzed by the system. The CPC's fulfill the criteria of being diagnostically challenging cases and of containing sufficiently detailed information to allow computer analysis. The evaluation was not intended to validate INTERNIST-1 for clinical use. CPC's should not be used for such a purpose, and as the trial demonstrated, the program does not yet possess sufficient reliability for clinical application. Nevertheless, INTERNIST-1 performed remarkably well, considering the simple, *ad hoc* nature of its algorithms.

8.2 The INTERNIST-1 Knowledge Base

A medical knowledge base must meet the needs of any associated diagnostic programs. In particular, the INTERNIST-1 knowledge base was designed to permit the consultant program to construct and resolve differential diagnoses. The knowledge base incorporates individual disease

profiles, which list findings that can occur in patients with each illness. By inverting the disease profiles with use of a computer program, an exhaustive differential diagnosis for each finding is obtained; these manifestation-based differential-diagnosis lists are retained as part of the knowledge base. The diagnostic program can use these lists to construct differential diagnoses in clinical cases.

How to group potential diagnoses into relevant problem areas is a separate consideration. The individual diseases in the INTERNIST-1 knowledge base are part of a *disease hierarchy* that is organized from the general to the specific. For example, acute viral hepatitis is classified as an hepatocellular infection, hepatocellular infection is a subclass of diffuse hepatic parenchymal disease, and diffuse hepatic parenchymal disease falls into the category of hepatic parenchymal disease, which is a major subclass of diseases of the hepatobiliary system. Initially, it was thought that access to the disease hierarchy would allow INTERNIST-1 to construct appropriate differential diagnoses (i.e., problem areas) based on higher-level concepts such as hepatocellular infection. If several diagnoses representing types of hepatocellular infection were under consideration, it would be simple to create a problem area for hepatocellular infection. However, early experience with the system showed that a rigid hierarchical classification scheme was inadequate, since a single disease often merits simultaneous categorization under more than one heading. Infectious mononucleosis is both a hepatocellular infection and a type of infectious lymphadenopathy. Hierarchical classification would require that it be listed as one or the other, but not both. An additional concern is that diseases may present differently in different patients. For example, alcoholic hepatitis may occur with predominance of intrahepatic cholestasis in one patient and with massive hepatocellular necrosis in another. Solution of the classification problem entailed development of algorithms (discussed below) that permit INTERNIST-1 to construct problem areas in an *ad hoc* manner.

The building block for the INTERNIST-1 data base is the individual disease. For each diagnosis entered into the system, a disease profile is constructed. The disease profile consists of findings (historical items, symptoms, physical signs, and laboratory abnormalities) that have been reported to occur in association with the disease, including demographic data and predisposing factors. Two clinical variables are associated with each manifestation in an INTERNIST-1 disease profile: an evoking strength and a frequency. The evoking strength answers the question "Given a patient with this finding, how strongly should I consider this diagnosis to be its explanation?" The frequency is an estimate of how often patients with the disease have the finding. In addition, each manifestation is assigned a disease-independent import. The import is the global importance of the manifestation—that is, the extent to which one is compelled to explain its presence in any patient. Although the evoking strengths, frequencies, and imports are expressed as numbers (on a scale of 0 to 5 or 1 to 5) in the INTERNIST-1 knowledge base, it is important to remember that they rep-

resent a shorthand for judgmental information, as their suggested interpretations in Tables 8-1 through 8-3 indicate. True quantitative information does not exist in the medical literature in most cases; the numbers used by INTERNIST-1 are judgmental in that they are compiled after a review of the available knowledge.

The current INTERNIST-1 knowledge base, which represents 15 person-years of work, encompasses over 500 individual disease profiles (an example appears in Figure 8-1) and approximately 3550 manifestations of disease. The disease profiles have been generated by review of the literature and by consultation with expert clinicians. In addition to the disease profiles, the knowledge base details relations among diagnoses and among manifestations. Within INTERNIST-1, important high-level pathophysiologic states (such as acute left ventricular failure, chronic congestive left heart failure, prerenal azotemia, and chronic uremia) are profiled as if they were diseases. The knowledge base contains links between such "diseases" and other diseases. The links are used to express causality or a predisposition of patients with one disease to have another. Because INTERNIST-1 formulates and resolves problem areas serially, it can piece together interdependent components of a multisystem illness one by one, using the links in the data base to promote consideration of diseases related to previously concluded diagnoses. The total number of links among the 500 diagnoses in the data base is about 2600. The 3550 manifestations in the INTERNIST-1 knowledge base are not independent. Men do not have oligomenorrhea, and a patient with oligomenorrhea must be presumed to be female. The knowledge base includes the properties of each manifestation that specify how its presence or absence may influence the presence or absence of other manifestations. There are roughly 6500 such interrelationships detailed in the knowledge base.

8.3 The Diagnostic Algorithms

The problem-solving algorithms represent the intellectual core of the INTERNIST-1 system. Although the scoring mechanism described below manipulates probabilistic data (evoking strengths, frequencies, and imports), it must be emphasized that the behavior of INTERNIST-1 results primarily from application of two heuristic principles: formation of problem areas via a partitioning algorithm, and conclusion of diagnoses within problem areas using strategies such as diagnosis by exclusion.

The steps on pages 197–200 are taken during an INTERNIST-1 diagnostic consultation. [Please refer to Section 8.6 for an annotated sample case analysis taken from a CPC published in the *New England Journal of Medicine* (Castleman, 1969).]

TABLE 8-1 Interpretation of Evoking Strengths

Evoking strength	*Interpretation*
0	Nonspecific—manifestation occurs too commonly to be used to construct a differential diagnosis.
1	Diagnosis is a rare or unusual cause of listed manifestation.
2	Diagnosis causes a substantial minority of instances of listed manifestation.
3	Diagnosis is the most common but not the overwhelming cause of listed manifestation.
4	Diagnosis is the overwhelming cause of listed manifestation.
5	Listed manifestation is pathognomonic for the diagnosis.

TABLE 8-2 Interpretation of Frequency Values

Frequency	*Interpretation*
1	Listed manifestation occurs rarely in the disease.
2	Listed manifestation occurs in a substantial minority of cases of the disease.
3	Listed manifestation occurs in roughly half the cases.
4	Listed manifestation occurs in the substantial majority of cases.
5	Listed manifestation occurs in essentially all cases—i.e., it is a prerequisite for the diagnosis.

TABLE 8-3 Interpretation of Import Values

Import	*Interpretation*
1	Manifestation is usually unimportant, occurs commonly in normal persons, and is easily disregarded.
2	Manifestation may be of importance, but can often be ignored; context is important.
3	Manifestation is of medium importance, but may be an unreliable indicator of any specific disease.
4	Manifestation is of high importance and can only rarely be disregarded, as, for example, a false-positive result.
5	Manifestation absolutely must be explained by one of the final diagnoses.

```
DISPLAY WHICH MANIFESTATION LIST?
ALCOHOLIC HEPATITIS

AGE 16 TO 25 ...0 1
AGE 26 TO 55 ...0 3
AGE GTR THAN 55 ...0 2
ALCOHOL INGESTION RECENT HX ...2 4
ALCOHOLISM CHRONIC HX ...2 4
SEX FEMALE ...0 2
SEX MALE ...0 4
URINE DARK HX ...1 3
WEIGHT LOSS GTR THAN 10 PERCENT ...0 3
ABDOMEN PAIN ACUTE ...1 2
ABDOMEN PAIN COLICKY ...1 1
ABDOMEN PAIN EPIGASTRIUM ...1 2
ABDOMEN PAIN NON-COLICKY ...1 2
ABDOMEN PAIN RIGHT UPPER QUADRANT ...1 3
ANOREXIA ...0 4
DIARRHEA ACUTE ...1 2
MYALGIA ...0 3
VOMITING RECENT ...0 4
ABDOMEN BRUIT CONTINUOUS RIGHT UPPER QUANDRANT ...1 2
ABDOMEN TENDERNESS RIGHT UPPER QUADRANT ...2 4
CONJUNCTIVA AND/OR MOUTH PALLOR ...1 2
FECES LIGHT COLORED ...1 2
FEVER ...0 4
HAND(S) DUPUYTRENS CONTRACTURE(S) ...1 2
JAUNDICE ...1 3
LEG(S) EDEMA BILATERAL SLIGHT OR MODERATE ...1 2
LIVER ENLARGED MASSIVE ...1 2
LIVER ENLARGED MODERATE ...1 3
LIVER ENLARGED SLIGHT ...1 2
PAROTID GLAND(S) ENLARGED ...1 2
SKIN PALLOR GENERALIZED ...0 2
SKIN PALMAR ERYTHEMA ...1 3
SKIN SPIDER ANGIOMATA ...2 3
SKIN TELANGIECTASIA ...1 1
ALKALINE PHOSPHATASE BLOOD GTR THAN 2 TIMES NORMAL ...1 2
ALKALINE PHOSPHATASE BLOOD INCREASED NOT OVER 2 TIME NORMAL ...1 4
BILIRUBIN BLOOD DECREASED ...2 2
BILIRUBIN URINE PRESENT ...2 4
CHOLESTEROL BLOOD DECREASED ...2 2
CHOLESTEROL BLOOD INCREASED ...1 2
HEMATOCRIT BLOOD LESS THAN 35 ...1 3
HEMOGLOBIN BLOOD LESS THAN 12 ...1 3
KETONURIA ...1 2
PROTEINURIA ...1 2
SGOT 120 TO 400 ...2 3
SGOT 40 TO 119 ...2 3
SGOT GTR THAN 400 ...1 2
UREA NITROGEN BLOOD LESS THAN 8 ...2 2
UROBILINOGEN URINE ABSENT ...1 1
UROBILINOGEN URINE INCREASED ...2 4
```

FIGURE 8-1 A sample manifestations list. The first number after each manifestation is its evoking strength for the diagnosis; the second is the frequency of the manifestation in the disease.

WBC 14000 TO 30000 ...0 3
WBC 4000 TO 139000 PERCENT NEUTROPHIL(S) INCREASED ...0 3
WBC LESS THAN 4000 ...1 1
ACTIVATED PARTIAL THROMBOPLASTIN TIME INCREASED ...1 3
ANTIBODY MITOCHONDRIAL ...1 1
ANTIBODY SMOOTH MUSCLE ...2 3
BSP RETENTION INCREASED ...1 5
ELECTROPHORESIS SERUM ALBUMIN DECREASED ...2 4
ELECTROPHORESIS SERUM GAMMA GLOBULIN INCREASED ...2 4
FACTOR VII PROCONVERTIN DECREASED ...1 2
LDH BLOOD INCREASED ...1 3
MAGNESIUM BLOOD DECREASED ...2 2
PROTHROMBIN TIME INCREASED ...2 3
SGPT 200 TO 600 ...1 2
SGPT 40 TO 199 ...2 3
SGPT GTR THAN 600 ...1 1
LIVER BIOPSY BILE PLUGGING ...1 2
LIVER BIOPSY FATTY METAMORPHOSIS ...2 4
LIVER BIOPSY FOCAL NECROSIS AND INFLAMMATION ...2 5
LIVER BIOPSY HEPATOCELLULAR NECROSIS MARKED ...2 3
LIVER BIOPSY MALLORY BODIES ...3 3
LIVER BIOPSY PERIPORTAL FIBROSIS MILD ...1 3
LIVER BIOPSY PERIPORTAL INFILTRATION NEUTROPHIL(S) ...3 5
LIVER BIOPSY PERIPORTAL INFILTRATION ROUND CELL(S) ...1 2
LIVER BIOPSY SMALL BILE DUCT(S) PROMINENT ...1 2

LINKS FOR ALCOHOLIC HEPATITIS:

Predisposes to	MALLORTY WEISS SYNDROME ...1 1
Causes	SINUSOIDAL OR POSTSINUSOIDAL PORTAL HYPERTENSION ...1 2
Causes	HEPATIC ENCEPHALOPATHY ...2 2
Causes	RENAL FAILURE SECONDARY TO LIVER DISEASE <HEPATORENAL SYNDROME> ...2 2
Coincident with	PANCREATITIS ACUTE ...2 2
Precedes	MICRONODAL CIRRHOSIS <LAENNECS> ...2 3

FIGURE 8-1 continued

1. Initial positive (present) and negative (absent) patient findings are entered by the user. As each new positive manifestation is encountered, the program retrieves its complete differential diagnosis from the inverted disease profiles in the knowledge base. A *disease hypothesis* is created for each item on the manifestation's differential-diagnosis list. A master list of all such disease hypotheses is maintained. Higher-level concepts from the classification hierarchy are retained on the differential-diagnosis list as long as the diagnoses that they subsume are indistinguishable in their ability to explain the observed data. The master differential list therefore comprises all possible diagnoses that can explain any of the observed findings (taken either individually or in groups).
2. For each disease hypothesis, four lists are maintained: all positive manifestations in the patient that are explained by the disease hypothesis (i.e., findings matching the disease profile stored in the data base); all manifestations that might occur in a patient with the disease but are

known to be absent in the patient being considered; all manifestations present in the patient but not explained by the disease hypothesis, that is, not found on the disease profile (these manifestations represent either "red herrings" or items that would have to be explained by a second disease present in the patient); and manifestations on the disease's profile about which nothing is known (this list is used in determining which questions to ask).

3. Each hypothesis on the master list of diagnoses is given a score. Scores are calculated as the sum of a positive and a negative component as follows. The positive component includes the weights of all manifestations explained by the hypothesis, based on the evoking strengths of the observed manifestations for the diagnosis. A nonlinear weighting scheme is used: an evoking strength of 0 counts as 1 point; a strength of 1 counts as 4 points; a 2 counts as 10 points; a 3 counts as 20; a 4 as 40; and a 5 as 80. Any disease hypothesis related to a previously concluded diagnosis (through links in the data base) is given a bonus score. The bonus awarded is 20 points times the frequency number listed for the hypothesized diagnosis in the disease profile of the concluded diagnosis. The negative component includes the weight of all manifestations that are expected to occur in patients with the disease but are absent in the patient under consideration. A nonlinear scale based on the expected frequency of the manifestation in the disease is used: a frequency of 1 counts as -1 point; a 2 as -4 points; a 3 as -7 points; a 4 as -15 points; and a 5 as -30 points. Also included are the weights of all manifestations present in the patient but not explained by the hypothesized diagnosis. The import (clinical significance) of each manifestation is used to assess this penalty: an import of 1 counts as -2 points; a 2 as -6 points; a 3 as -10 points; a 4 as -20 points; and a 5 as -40 points. The net score for any disease hypothesis is thus the sum of the above four component weights.
4. After all disease hypotheses have been scored, the master list of all hypotheses is sorted by descending score. Diagnoses whose scores fall a threshold number of points below the topmost diagnosis are temporarily discarded as unattractive. They may be reconsidered, however, if further evidence obtained during the case analysis raises their scores above the threshold (relative to the topmost diagnosis).
5. At this point, the sorted master differential-diagnosis list is a heterogeneous grouping of many disease hypotheses. A critical step in the diagnostic logic of INTERNIST-1 is to delineate a set of competitors for the topmost diagnosis (i.e., to create a problem area containing the topmost disease hypothesis). Only one of the set of diseases in a properly defined problem area is likely to be present in a patient. Problem area construction is carried out by the INTERNIST-1 partitioner, which employs a remarkably powerful yet simple heuristic rule. The rule states, "Two diseases are competitors if the items not explained

by one disease are a subset of the items not explained by the other; otherwise, they are alternatives (and may possibly coexist in the patient)." To paraphrase, if Disease A and Disease B taken together explain no more observed manifestations than does either one taken alone, then the diseases are classified as competitors. Competitors for the likeliest diagnosis are identified from the master differential list using the partitioning rule; including the topmost diagnosis, they constitute the *current problem area.* Because INTERNIST-1 defines problem areas in this *ad hoc* manner, its differential diagnoses will not always resemble those constructed by clinicians.

6. Once the problem area containing the most attractive diagnosis has been selected, criteria for establishing a definitive diagnosis can be applied. If the problem area contains only the topmost diagnosis, INTERNIST-1 will immediately decide on (conclude) that diagnosis. If there is more than one diagnosis in the problem area, INTERNIST-1 directly concludes the leading diagnosis when its score is 90 or more points higher than the nearest competitor. The value of 90 was chosen because it slightly exceeds the weight carried by a pathognomonic finding (80 points). This method of concluding a diagnosis is a hallmark of INTERNIST-1. The absolute score of the diagnosis does not matter. The only point of importance is whether the diagnosis is sufficiently higher in score than its reasonable competitors (other diagnoses that explain the same set of findings).
7. If it is not possible to conclude a diagnosis (which by default means that the current problem area contains more than one hypothesis), one of three questioning strategies is selected: pursuing, ruling out, or discriminating. The pursuing mode is selected if the second-best contender is 46 to 89 points behind the topmost diagnosis. In the pursuing mode, questions are asked to establish the topmost diagnosis, since it is close to fulfilling criteria for conclusion. The questions asked are those that are most specific for the leading diagnosis (i.e., those with high evoking strengths). If there are five or more diagnoses within 45 points of the topmost diagnosis, the ruling-out mode is used. Questions that have high frequency numbers under the contenders are asked, with the expectation that several negative responses will remove some of the diagnoses from contention. The discriminating mode is used when there are two to four diagnoses within 45 points of the leading diagnosis. The questions asked attempt to maximize the spread in scores.
8. In order to improve the efficiency of computations, questions are asked in small groups. The level of questioning is escalated (from history to physical-examination findings to gradations of laboratory results) only after the useful questions in a previous category have been exhausted. After the answers are processed, the disease hypotheses are again scored and partitioned. A new differential diagnosis

is formed on the basis of the (possibly) new topmost diagnosis. This *ad hoc* method for constructing a differential diagnosis gives INTERNIST-1 seemingly intelligent behavior, since the program will often change focus from one problem area to another when questioning in the first area has been counterproductive.

9. When a diagnosis is concluded, all observed manifestations explained by the diagnosis are removed from future consideration. The program then recycles using the remaining unexplained positive findings. Subsequent findings are marked as explained when a previously concluded diagnosis can account for them. However, it is not possible to undo a previous diagnostic conclusion when contradictory evidence becomes available.
10. When a problem area contains more than one disease hypothesis and all useful lines of questioning have been exhausted (without meeting criteria for concluding the topmost diagnosis), the program will defer making a diagnosis in that problem area. Diagnoses in the problem area are then displayed by descending score, along with an explanation that the differential diagnosis cannot be resolved.
11. When all remaining manifestations have an import of 2 or less, the program stops.

8.4 An Evaluation of INTERNIST-1

We have completed a preliminary evaluation of INTERNIST-1. The program was evaluated to compare its clinical acumen to that of human experts and to highlight its strengths and weaknesses. CPC's published in the *New England Journal of Medicine* (NEJM) as "Case Records of the Massachusetts General Hospital" were used for the computer analysis. During the trial, only the published findings available to the case discussant were presented to INTERNIST-1 (i.e., only findings mentioned before the presentation of the pathological findings). The knowledge base of INTERNIST-1 was not altered during the course of the evaluation.

During the development of INTERNIST-1, hundreds of miscellaneous individual cases, both simple and complex, have been presented to the system in order to evaluate and improve the data base and the diagnostic computer program. Since many of these test cases included NEJM CPC's, cases for the trial were selected from 1969, a year from which no previous NEJM cases had been presented to INTERNIST-1. Before entering any cases, project members serially reviewed the published final anatomic diagnoses. All cases in which one or more of the major diagnoses were not represented in INTERNIST-1's still incomplete knowledge base were rejected. The diagnostic program cannot conclude a diagnosis that is

missing from the knowledge base; such a case would not be a fair test for the system. The excluded diagnoses were neither more rare nor more complex than the diagnoses chosen for analysis. Cases 1-1969 through 42-1969 (inclusive) were reviewed, and 19 cases were obtained in which all major CPC diagnoses were included in the data base. That only 19 of the 42 cases reviewed qualified for the study is not unexpected. It is estimated that the current INTERNIST-1 knowledge base includes roughly 70–75% of the major diagnoses of internal medicine. If each case on the average contained three major diagnoses, the probability that all three diagnoses would be included in the knowledge base is (0.75) × (0.75) × (0.75) or 42%.

In establishing criteria for evaluating performance on the NEJM CPC's, one must classify final anatomic or clinical diagnoses as major or minor. Major diagnoses are defined as those central to the problem. Classified as minor diagnoses are diseases that were present in the patient but were clinically less relevant, including those diseases only partially described in the published case protocol, as well as conditions that were successfully managed and that subsequently resolved. Diagnostic decisions made by the clinicians at the Massachusetts General Hospital (MGH), by the case discussants, and by INTERNIST-1 were classified as correct when they were confirmed by the pathologists or when a clinical syndrome was universally agreed to be present. When either the physicians or INTERNIST-1 introduced an incorrect diagnosis, a separate notation was made because an incorrect diagnosis has a different meaning from that of a failure to make a correct diagnosis. We recognize two ways for a program or a clinician to make a correct diagnosis in the setting of a CPC: to state unequivocally that the patient has the disease *(definitive diagnosis)* or to offer an unresolved differential diagnosis that includes the correct diagnosis as its topmost element *(tentative diagnosis)*. INTERNIST-1 makes definitive diagnoses by conclusion and tentative diagnoses by deferral (see above). The hospital clinicians and the case discussants also made both types of diagnoses. A tentative diagnosis was counted as incorrect if its topmost element was not the correct diagnosis, even if the associated differential diagnosis included the correct diagnosis.

Table 8-4 summarizes the results for the 19 trial cases. There were 43 possible correct major diagnoses. INTERNIST-1, the clinicians at the MGH, and the case discussants made 17, 23, and 29 correct definitive diagnoses, respectively. A correct tentative diagnosis was offered 8, 5, and 6 times, respectively. Thus, of 43 anatomically verified diagnoses, INTERNIST-1 failed to make a total of 18, whereas the clinicians failed to make 15 such diagnoses, and the discussants missed only 8. Of the 18 situations in which INTERNIST-1 failed to make an anatomically correct diagnosis, the clinicians or the discussant or both failed to make the correct diagnosis 11 times. INTERNIST-1 made a correct diagnosis in 7 circumstances in which the clinicians or the case discussant failed to do so. INTERNIST-1 made 5 incorrect definitive diagnoses and 6 incorrect tentative

TABLE 8-4 Summary of results for major diagnoses in 19 cases used in the INTERNIST-1 evaluation

	No. of instances		
Category	*INTERNIST-1*	*Clinicians*	*Discussants*
Definitive, correct	17	23	29
Tentative, correct	8	5	6
Failed to make correct diagnosis	18	15	8
Definitive, incorrect	5	8	11
Tentative, incorrect	6	5	2
Total no. of incorrect diagnoses	11	13	13
Total no. of errors in diagnosis	29	28	21
Total possible diagnoses	43	43	43

diagnoses (naming diseases that were not present in the patients). The MGH clinicians made 8 incorrect definitive diagnoses and 5 incorrect tentative diagnoses. The case discussants made 11 incorrect definitive diagnoses and 2 incorrect tentative diagnoses. Of the 5 situations in which INTERNIST-1 made an incorrect definitive diagnosis, 4 were situations in which the discussants also made a wrong diagnosis.

The shortcomings of the program, which were highlighted by the evaluation, fall into two general categories. The first type are limitations due to the structure or content of the knowledge base. Examples include the absence of a manifestation required to describe an important finding; the use of overly simplistic manifestations for some circumstances; the inadvertent omission of a finding from a disease profile; the assignment of an incorrect evoking strength, frequency, or import; and the failure of a manifestation to convey adequate anatomic information. The second type of limitation resulted from deficiencies in the design or implementation (or both) of the computer program. Included in this category were failure to incorporate temporal reasoning capabilities; problems resulting from use of the scoring algorithm; the inability to take a broad overview in attacking a complex problem; and the improper attribution of findings to concluded diagnoses (i.e., invoking the wrong explanation for a finding). Specific reasons for INTERNIST-1's incorrect diagnoses (made both by omission and by commission) are listed in Table 8-5.

8.5 Discussion

Experience with INTERNIST-1 has reinforced our impression of medical diagnosis as a complex process. Diagnosis consists of two fundamental activities: the generation of one or more differential diagnoses (each for a separate problem area), and the resolution of individual differential di-

TABLE 8-5 Classification of errors made by INTERNIST-1 during the evaluation

Type of error	*No. of occurrences*
Knowledge-base errors	
Data base incomplete/omission	2
Data base incorrect	2
Lack of anatomic knowledge	1
Failure to represent degree of severity	2
Computer-program faults	
Lack of temporal reasoning	3
Failure of scoring algorithm	3
Failure to seek global overview	1
Improper attribution of finding to a concluded diagnosis	6

agnoses. The surprising ability of the program to make multiple and complex diagnoses in the broad field of internal medicine emphasizes the power of its underlying heuristic methods.

Several important shortcomings of the INTERNIST-1 approach to diagnosis merit further investigation. Feinstein (1977b) has emphasized the importance of explanation as part of diagnostic reasoning. INTERNIST-1's greatest failing during the evaluation (occurring in 6 instances) was its inability to attribute findings to their proper causes. Because of the *ad hoc,* serial nature of INTERNIST-1's formation of problem areas, the program cannot synthesize a general overview in complicated multisystem problems. The structure of the knowledge base, especially the form of the disease profiles, limits the program's ability to reason anatomically or temporally. The program cannot recognize subcomponents of an illness, such as specific organ-system involvements or the degree of severity of pathologic processes.

A diagnostic program must be able to recognize the appropriate cause or causes of observed findings in a patient. A justification for each diagnosis must be developed on a pathophysiologic or causal framework that is consistent with established medical knowledge. To its detriment, INTERNIST-1's handling of explanation is shallow. When the program concludes a diagnosis, that diagnosis is allowed to explain any observed manifestations that are listed on its disease profile. Once explained, a manifestation is no longer used to evoke new disease hypotheses or to participate in the scoring process. This situation is compounded by the inadequate representation of causality in the INTERNIST-1 knowledge base. Disease profiles contain, in an undifferentiated manner, factors predisposing to the

illness as well as findings that result from the disease process itself. An example of this problem occurred in analysis of Case 17-1969, when INTERNIST-1 allowed hepatic encephalopathy to explain the finding of hypokalemia. The program should have recognized hypokalemia as a predisposing factor for hepatic encephalopathy and initiated a search for an independent cause of the finding. At present, the limitations of the knowledge base prohibit such activity.

What is required is a restructuring of the knowledge base to include intermediate-level pathophysiologic states and the segregation of predisposing factors from findings actually caused by a disease. Diseases should be profiled in terms of their intermediate states, rather than as exhaustive lists of manifestations. If the program had such a feature, the presence or absence of each state would be independently determined, and a disease would be allowed to explain a finding only when the state causing the finding was confirmed.

A related problem not handled well by INTERNIST-1 is the interdependency of manifestations. For example, persons with elevated conjugated bilirubin levels in their blood usually have bilirubinuria. At present, the evoking strengths of each finding count redundantly toward any diagnosis that can explain them. This phenomenon causes INTERNIST-1 to favor disproportionately the most common explanation for a set of findings. A solution would be the creation of an intermediate-level state, "abnormal bilirubin metabolism and transport," which would explain both conjugated hyperbilirubinemia and bilirubinuria. Appropriate weight for the intermediate state (rather than for the interdependent manifestations) could be given to any diseases that cause it. Thus creation of a causal network of pathophysiologic states, interposed between observable manifestations and final diagnoses, would allow a diagnostic program to attribute findings to causes accurately and would help to diminish the influence of interdependent manifestations of disease.

INTERNIST-1 constructs differential diagnoses in an *ad hoc* manner, using a scoring algorithm to define the topmost (best) diagnosis and another program, the partitioner, to define reasonable competitors for the topmost diagnosis. By formulating and focusing attention on only one problem domain at any given time, the program is able to disregard "red herrings" and to set aside—temporarily—findings caused by disease processes falling outside the selected problem domain. By creating and processing problem domains serially, the program is able to make multiple diagnoses. But INTERNIST-1 cannot formulate a broad perspective in complicated multisystem patient problems. It is constrained to working with tunnel vision, discriminating among diagnoses within each problem area, unable to look at several problem areas simultaneously. Only after a specific diagnosis is concluded can INTERNIST-1 use the links in its data base to give bonus weight to interrelated diagnoses in separate problem domains. New programming approaches to complex reasoning processes have been developed (Pople, 1982) to enable CADUCEUS, the successor

to INTERNIST-1, to synthesize a broad overview incorporating causal relationships into an approach to a patient's problems.

INTERNIST-1 is unable to reason anatomically or temporally. The program could not differentiate gastric compression due to pancreatic mass effect from that due to hepatic mass effect in Case 23-1969, and as a result it erroneously concluded that the patient had a hepatoma rather than pancreatitis. Nor can INTERNIST-1 recognize the degree of severity of a finding or process in all instances. Two of INTERNIST-1's failures during the evaluation resulted from its inadequate recognition of the degree of severity of an individual manifestation (a decreased blood potassium level) and of an organ-system involvement by a pathologic process (disseminated vasculitis). Reorganization of the data base to allow representation of these concepts is also being undertaken.

INTERNIST-1 is only one of many computer-based tools with the purpose of extending the capabilities of the physician. Such programs can broaden the clinician's scope and awareness of data for the diagnosis and treatment of illness. For the present, INTERNIST-1 remains a research tool. After refinement of the knowledge base and diagnostic programs, a prospective clinical trial will be required to compare the program's behavior with that of clinicians in terms of diagnostic accuracy, cost effectiveness, and danger to the patient.

8.6 A Sample Case Analysis

The transcript of an INTERNIST-1 case analysis given in Figure 8-2 illustrates the operation of the diagnostic programs. The case was taken from a CPC published in the *New England Journal of Medicine* in 1969 (Castleman, 1969). The laboratory values are reported as measured in 1969. The bracketed paragraphs labeled "Comment" have been interpolated for clarification; they are not part of the actual consultation. Places where the transcript has been abridged are indicated by ellipses.

INTERNIST-1 consultation SUMEX-AIM Version
15-May-81 07:31:39
ENTER CLASS NAME: NEJM-CASE-30-1969-ADMISSION-1

[Comment: Here the user enters the initial positive findings (present in the patient) and negative findings (absent). The specialized INTERNIST-1 vocabulary of some 3550 manifestations must be used in describing the case. The plus (+) prompt precedes each positive finding entered by the user. Because INTERNIST-1 has no mechanism for the representation of time, all findings have been collapsed into a single list, independently of their order of appearance in the patient.]

FIGURE 8-2 Transcript of an INTERNIST-1 case analysis.

INITIAL POSITIVE MANIFESTATIONS:
+AGE GTR THAN 55
+ARTHRITIS HX
+DEPRESSION HX
+SEX FEMALE
+THYROIDECTOMY HX
+ULCER PEPTIC HX
+URINE DARK HX
+WEIGHT INCREASE RECENT HX
+ANOREXIA
+CHEST PAIN LATERAL EXACERBATION WITH BREATHING
+CHEST PAIN LATERAL SHARP
+DYSPNEA ABRUPT ONSET
+ABDOMEN DISTENTION
+ABDOMEN FLUID WAVE
+ASTERIXIS
+FECES LIGHT COLORED
+JAUNDICE
+JOINT(S) PERIARTICULAR THICKENING
+JOINT(S) RANGE OF MOTION DECREASED
+LIVER ENLARGED MODERATE
+PLEURAL FRICTION RUB
+PULSE PRESSURE INCREASED
+SKIN PALMAR ERYTHEMA
+SKIN SPIDER ANGIOMATA
+SPLENOMEGALY SLIGHT
+TACHYCARDIA
+TACHYPNEA
+THYROID ENLARGED ASYMMETRICAL
+ALKALINE PHOSPHATASE BLOOD GTR THAN 2 TIMES NORMAL
+BILIRUBIN BLOOD CONJUGATED INCREASED
+BILIRUBIN URINE PRESENT
+FECES GUAIAC TEST POSITIVE
+GLUCOSE BLOOD 130 TO 300
+SGOT 120 TO 400
+AMMONIA BLOOD INCREASED
+ASCITIC FLUID PROTEIN 3 GRAM(S) PERCENT OR LESS
+ASCITIC FLUID WBC 100 TO 500
+ELECTROPHORESIS SERUM ALBUMIN DECREASED
+ELECTROPHORESIS SERUM BETA GLOBULIN INCREASED
+ELECTROPHORESIS SERUM GAMMA GLOBULIN INCREASED
+GLUCOSE TOLERANCE DECREASED
+IMMUNOELECTROPHORESIS SERUM IGA INCREASED
+IMMUNOELECTROPHORESIS SERUM IGM INCREASED
+IMMUNOELECTROPHORESIS SERUM IGG INCREASED
+LDH BLOOD INCREASED
+LE TEST POSITIVE
+LIVER RADIOISOTOPE SCAN IRREGULAR UPTAKE
+PROTHROMBIN TIME INCREASED
+RHEUMATOID FACTOR POSITIVE
+STOMACH BARIUM MEAL HIATAL HERNIA
+*DENY*

[Comment: The keyword *DENY* indicates to INTERNIST-1 that the user now wants to enter negative findings. The minus (-) prompt precedes each negative finding entered.]

FIGURE 8-2 continued

INITIAL NEGATIVE FINDINGS:
-ALCOHOLISM CHRONIC HX
-DIARRHEA CHRONIC
-FECES BLACK TARRY
-FEVER
-PRESSURE VENOUS INCREASED ON INSPECTION
-HEMATOCRIT BLOOD LESS THAN 35
-UREA NITROGEN BLOOD 30 TO 59
-URIC ACID BLOOD INCREASED
-ASCITIC FLUID AMYLASE INCREASED
-ASCITIC FLUID CYTOLOGY POSITIVE
-ASCITIC FLUID LDH GTR THAN 500
-ESOPHAGUS BARIUM MEAL VARICES
-STOMACH BARIUM MEAL ULCER CRATER
-T3 RESIN UPTAKE INCREASED
-T4 TOTAL BLOOD INCREASED
-GO

[Comment: The user enters *GO* to indicate that all relevant positive and negative findings have been entered. The INTERNIST-1 consultant programs guide the user through the rest of the diagnostic work-up by asking questions. For clarity, all reponses typed by the user from this point will be in italics.]

DISREGARDING: DEPRESSION HX, WEIGHT INCREASE RECENT HX, CHEST PAIN LATERAL EXACERBATION WITH BREATHING, CHEST PAIN LATERAL SHARP, DYSPNEA ABRUPT ONSET, ABDOMEN FLUID WAVE, ASTERIXIS, JOINT(S) PERIARTICULAR THICKENING, PLEURAL FRICTION RUB, THYROID ENLARGED ASYMMETRICAL, FECES GUAIAC TEST POSITIVE, GLUCOSE BLOOD 130 TO 300, AMMONIA BLOOD INCREASED, ASCITIC FLUID WBC 100 TO 500, GLUCOSE TOLERANCE DECREASED

CONSIDERING: AGE GTR THAN 55, SEX FEMALE, URINE DARK, ANOREXIA, FECES LIGHT COLORED, JAUNDICE, LIVER ENLARGED MODERATE, SKIN PALMAR ERYTHEMA, SKIN SPIDER ANGIOMATA, SPLENOMEGALY SLIGHT, ALKALINE PHOSPHATASE BLOOD GTR THAN 2 TIMES NORMAL, BILIRUBIN BLOOD CONJUGATED INCREASED, BILIRUBIN URINE PRESENT, SGOT 120 TO 400, ELECTROPHORESIS SERUM ALBUMIN DECREASED, ELECTROPHORESIS SERUM GAMMA GLOBULIN INCREASED, IMMUNOELECTROPHORESIS SERUM IGA INCREASED, IMMUNOELECTROPHORESIS SERUM IGG INCREASED, IMMUNOELECTROPHORESIS SERUM IGM INCREASED, LDH BLOOD INCREASED, LE TEST POSITIVE, LIVER RADIOISOTOPE SCAN IRREGULAR UPTAKE, PROTHROMBIN TIME INCREASED, RHEUMATOID FACTOR POSITIVE.

DISCRIMINATE: HEPATITIS CHRONIC ACTIVE, BILIARY CIRRHOSIS PRIMARY

[Comment: At this point, INTERNIST-1 has constructed a master differential-diagnosis list, ordered its members with a scoring algorithm, and then focused attention on the most promising problem area, which appropriately contains two liver diseases. The DISREGARDING list consists of all findings that are inconsistent with the topmost (first-listed) diagnosis; the CONSIDERING list includes all findings explained by the topmost diagnosis (i.e., chronic active hepatitis). The number of plausible contenders in the problem area determines the strategy for questioning. Questions are asked to discriminate between the two diagnoses.]

Please Enter Findings of LIPID(S) BLOOD
GO
CHOLESTEROL BLOOD DECREASED?
N/A

[Comment: INTERNIST-1 asks questions in two forms. A question of the style "Please Enter Findings of . . . " asks about a generic class of findings, such as serum lipids or chest x-ray films. The user enters specific (positive or negative) findings and when finished types *GO*. If there are no positive or negative findings to enter, typing *GO* will cause the programs to ask the original single item that prompted the generic question.

FIGURE 8-2 continued

For example, an elevated cholesterol level would favor primary biliary cirrhosis over chronic active hepatitis, whereas the converse situation would apply if the cholesterol were decreased. Thus, INTERNIST-1 asked a generic question about serum liquids. The second form of questioning is direct. INTERNIST-1 simply prints a manifestation name followed by a question mark. The user can then respond by typing *YES* (present), *NO* (absent), or *N/A* (not available—either unknown or not performed). The published CPC did not mention the serum cholesterol value, so the user entered *N/A*.]

Please Enter Findings of INSPECTION AND PALPITATION SKIN
GO
SKIN XANTHOMATA?
NO
Please Enter Findings of BLOOD COUNT AND SMEAR
GO
RBC TARGET(S)?
NO

[Comment: INTERNIST-1 uses information about the type of each manifestation—history, symptoms, signs, and laboratory tests, both routine and two gradations of more expensive or invasive laboratory procedures. The types of manifestations determine which questions are asked. Although the greatest discriminating power for resolving a differential diagnosis often resides in an invasive test, such as a liver biopsy, INTERNIST-1 has been programmed to exhaust the useful history and physical-examination findings before escalating its questioning to more expensive or dangerous laboratory tests.]

Please Enter Findings of KIDNEY FUNCTION TEST(S)
GO
UREA NITROGEN BLOOD 60 TO 100?
NO
UREA NITROGEN BLOOD GTR THAN 100?
NO
Please Enter Findings of LIVER FUNCTION TEST(S)
GO
UROBILINOGEN URINE ABSENT?
N/A

[Comment: INTERNIST-1 did not seek to investigate renal function itself in the passage above. It was programmed to recognize that the urinary urobilinogen value may be unreliable in the presence of marked azotemia, so it first asked about the blood urea nitrogen.]

CONCLUDE: HEPATITIS CHRONIC ACTIVE

[Comment: It is likely that INTERNIST-1 had nearly made the diagnosis of chronic active hepatitis before it asked any questions. The negative answers favored chronic active hepatitis, and therefore INTERNIST-1 concluded the diagnosis.]

DISREGARDING: DEPRESSION HX, CHEST PAIN LATERAL EXACERBATION WITH BREATHING, CHEST PAIN LATERAL SHARP, DYSPNEA ABRUPT ONSET, ASTERIXIS, JOINT(S) PERIARTICULAR THICKENING, PLEURAL FRICTION RUB, THYROID ENLARGED ASYMMETRICAL, FECES GUAIAC TEST POSITIVE, GLUCOSE BLOOD 130 TO 300, AMMONIA BLOOD INCREASED, GLUCOSE TOLERANCE DECREASED

CONSIDERING: WEIGHT INCREASE RECENT HX, ABDOMEN DISTENTION, ABDOMEN FLUID WAVE, ASCITIC FLUID OBTAINED BY PARACENTESIS, ASCITIC FLUID WBC 100 TO 500, ASCITIC FLUID PROTEIN 3 GRAM(S) PERCENT OR LESS ...
CONCLUDE: TRANSUDATIVE ASCITES
DISREGARDING: ...
CONSIDERING: ASTERIXIS, PULSE PRESSURE INCREASED, TACHYCARDIA, AMMONIA BLOOD INCREASED
PURSUING: HEPATIC ENCEPHALOPATHY

FIGURE 8-2 continued

[Comment: The links in the INTERNIST-1 knowledge base between chronic active hepatitis and hepatic encephalopathy have resulted in a bonus weight's being given to hepatic encephalopathy here; previously, links had promoted the consideration of transudative ascites, since it can also be caused by chronic active hepatitis.]

CSF FLUID OBTAINED?
N/A

[Comment: Here INTERNIST-1 was about to ask about the glutamine level in the cerebrospinal fluid. Since no lumbar puncture was performed, the result is not available.]

CONCLUDE: HEPATIC ENCEPHALOPATHY

[Comment: In the above situation, there were no diagnostically helpful tests remaining for INTERNIST-1 to ask. INTERNIST-1 has been programmed to relax its criteria for concluding a diagnosis when all useful lines of questioning have been blocked. Since INTERNIST-1 had been close to making the diagnosis of hepatic encephalopathy, the program now concludes the diagnosis. The case analysis was intentionally stopped at this point, because all relevant major diagnoses had been covered. Without such intervention, INTERNIST-1 would try to explain any remaining important findings, such as the arthritis and pleurisy.]

FIGURE 8-2 continued

ACKNOWLEDGMENTS

We are indebted to Craig Dean, Charles Oleson, and Kenneth Quayle for their contributions in writing the INTERNIST-1 computer programs; to Zachary Moraitis for his assistance in the conceptual design of the project and in the development of the knowledge base; to a large number of medical students and several fellows in computer medicine for their assistance in the development of the INTERNIST-1 knowledge base; and to the staff of the SUMEX-AIM computing facility of the National Institutes of Health for providing expert assistance and a friendly environment for programming.

The INTERNIST-1/CADUCEUS project is supported by grants from the Division of Research Resources (R24 RR 01101) and the National Library of Medicine (R01 LM 03710 and R23 LM 035789), National Institutes of Health. The SUMEX computing project is supported by a grant (RR 00785) from the Biotechnology Resources Program, National Institutes of Health.