# 18

# A System for Empirical Experimentation with Expert Knowledge

**Peter Politakis and Sholom M. Weiss**

*When CASNET (Chapter 7) evolved into the general system-building tool known as EXPERT, one of the first applications was a rheumatology consultant program called AI/Rheum (Kingsland and Lindberg, 1983). Developed collaboratively by researchers at Rutgers University and the University of Missouri, AI/Rheum quickly became large and complex, thereby complicating the process of knowledge base maintenance. Peter Politakis, a Rutgers graduate student working with Sholom Weiss and Casimir Kulikowski, accordingly developed a program, named SEEK, that was designed to assist with both expansion and verification of the AI/Rheum knowledge base.*

*SEEK illustrates how a model of expert reasoning (in this case the rules of rheumatology diagnosis) can be refined with program assistance. The program suggests possible experiments involving generalization or specialization of the preexisting rules in the system. A library of stored patient cases with known conclusions is used as a basis for proposing the experiments. This approach has proven particularly valuable in assisting the expert in a domain like rheumatology where two diagnoses are often difficult to distinguish.*

*The research on SEEK also has its origins in the knowledge-acquisition tool TEIRESIAS, developed by Davis for MYCIN (Davis, 1979). However, SEEK is able to go a step further by using a somewhat more articulated representation than MYCIN's rules. In AI/Rheum evidence is classified according to major and minor findings, plus required and*

From *Proceedings of the Fifteenth Hawaii International Conference on Systems Science*, 2: 649–657 (1982).

*excluded findings. Specialization and generalization are accomplished by adding or deleting elements in these lists. The use of symbolic categories of belief (definite, probable, and possible) provides a specifiable means for manipulating the rules.*

*While based on a simple idea, the SEEK program convincingly demonstrates the value of a richly structured representation and of reasoning from cases as a way of constructing a model. That is, expert knowledge is inseparable from case experience (Schank, 1983), in so far as knowledge explains the cases. The use of a knowledge base to provide an explanatory model has characterized other recent AIM work as well (cf. the diagnostic approach used by Patil, Chapter 14). Another important strength of the SEEK approach is its exhaustive analysis of the entire library of cases, thereby revealing the overall effect of a modification. Experts building the system can accordingly avoid being swayed by one or two cases; they must explain their experiences as a whole.*

## 18.1 Introduction

Over the past decade, much of the research in the development of expert systems has been focused on the acquisition of knowledge in various medical areas: CASNET (Chapter 7)—ophthalmology; INTERNIST (Chapter 8), PIP (Chapter 6)—internal medicine; and MYCIN (Chapter 5)—infectious diseases. A relatively difficult task is to find effective methods for validating a system's knowledge base and evaluating its performance. A step in this direction has been taken in recent work to develop knowledge-engineering tools that would facilitate the building and testing of an expert system. Two examples of generalized knowledge-engineering tools are the EXPERT (Weiss and Kulikowski, 1979) and EMYCIN (van Melle, 1979) systems. These systems provide the builder of an expert system with a prespecified control strategy, a production rule formalism for encoding expert knowledge, explanatory tools for tracing the execution of rules during a consultation session, and a data base system in which cases can be stored for empirical testing. Other work on empirical testing of expert systems has been reported in the development of the PROSPECTOR consultation model for mineral exploration (Gaschnig, 1979). The PROSPECTOR scheme uses sensitivity analysis to determine the effect on the model's conclusions as a result of making changes to certainties in the input data. The empirical testing is based on matching the expert's conclusion to the overall result and also to the intermediate conclusions reached by the model.

As has been demonstrated in the TEIRESIAS system (Davis, 1977), the knowledge-engineering tools that explain a system's decisions are invaluable aids in expert knowledge acquisition and in improving performance.

During a consultation session on a patient case, TEIRESIAS assists the user in composing new rules to correct erroneous conclusions. TEIRESIAS generates its advice about the contents of a new rule by using a *rule model* that summarizes relationships within a subset of the rules in the knowledge base. It does not, however, directly determine the impact of changes to the knowledge base on other cases previously processed by the consultation program.

The approach described in this paper is to integrate performance information into the design of an expert model to automatically provide advice about rule refinement. A system called SEEK has been developed that generates advice in the form of suggestions for possible experiments in generalizing or specializing rules in an expert model. Case experience, in the form of stored cases with known conclusions, is used to interactively guide the expert in refining the rules of a model. In particular, SEEK looks for certain regularities about the performance of the rules in misdiagnosed cases as a basis for suggesting changes to the rules. An expanded description of methods and the uses of SEEK can be found in Politakis (1982).

## 18.2 The Model

A table of criteria, which is a specialized type of frame or prototype (Aikins, 1979), is prepared for each potential diagnosis. The table consists of two parts:

- major and minor observations that are significant for reaching the diagnosis
- a set of diagnostic rules for reaching the diagnosis

The following example shows observations, grouped under the headings *Major criteria* and *Minor criteria*, for mixed connective tissue disease:

*Major criteria*

1. Swollen hands
2. Sclerodactyly
3. Raynaud's phenomenon or esophageal hypomotility
4. Myositis, severe
5. CO diff. capacity (normally < 70)

*Minor criteria*

1. Myositis, mild
2. Anemia
3. Pericarditis
4. Arthritis ≤ 6 wks
5. Pleuritis
6. Alopecia

The second part of the table contains the diagnostic rules. In the following example, each column consists of a rule for a specific degree of certainty in the diagnosis:

| | *Definite* | *Probable* | *Possible* |
|---|---|---|---|
| | 4 majors | 2 majors, 2 minors | 3 majors |
| *Requirements* | Positive RNP antibody | Positive RNP antibody | No requirement |
| *Exclusions* | Positive SM antibody | No exclusion | No exclusion |

There are three levels of confidence: definite, probable, and possible. A diagnostic rule is a conjunction of three components, one taken from each row: specific numbers of major or minor observations, requirements, and exclusions. *Requirements* are those combinations of observations that are necessary beyond simple numbers of major and minor findings (although major and minor findings also may be requirements). *Exclusions* are those observations that rule out the diagnosis at the indicated confidence level. The three fixed confidence levels are an important attribute of the model. They substitute for complex scoring functions, which can be a major difficulty in analyzing and explaining model performance (see Chapter 9). It is understood that if a definite diagnosis for a particular disease is made, then even if the rules for the probable or possible diagnosis for the same disease are satisfied, the definite conclusion is appropriate.

As an example, the rule for concluding definite mixed connective tissue disease can be stated as follows: if the patient has 4 or more major observations for mixed connective tissue disease, and RNP antibody is positive, and SM antibody is not positive, then conclude definite mixed connective tissue disease. In most applications, multiple rules are described for each confidence level.

In terms of refinement of a model, the following sections will focus on tools that facilitate identifying two classes of changes that can be made to the rules—generalizations and specializations. *Generalizations* are changes to a rule R that result in a different rule Rg where Rg logically includes R. For example, this can be accomplished by dropping a requirement or decreasing the number of major and minor findings for a rule. *Specializations* are changes to a rule R that result in a different rule Rs where Rs is logically included by R. For example, this can be accomplished by increasing the number of major and minor findings in a rule.

Framelike schemes have been used to represent medical knowledge in the PIP (see Chapter 6) and CENTAUR (Aikins, 1979) systems, which were designed to provide diagnostic consultations in subspecialties of medicine. In addition to representing various clinical states, findings with typical values and frequencies, and related diseases in each disease frame, there were slots containing relatively complex scoring functions that could be specialized for the evaluation of the disease frame. The tabular model is a simple type of frame representation requiring for each diagnostic conclusion fixed types (e.g., majors, exclusions) of observations that are relatively easy to understand. Also, scoring follows directly from the three confidence levels of definite, probable, and possible.

## 18.3 The Rheumatology Application

In collaboration with rheumatologists at the University of Missouri, a consultation model for connective tissue diseases has been realized using the EXPERT system (Weiss and Kulikowski, 1979) for developing consultation models. This subpart of rheumatology is a particularly difficult area for the physician and includes seven diseases: rheumatoid arthritis, systemic lupus erythematosus (SLE), progressive systemic sclerosis, mixed connective tissue disease, polymyositis, primary Raynaud's syndrome, and Sjogren's disease. Some of the difficulties in the differential diagnosis of these diseases may be appreciated by noting that even the experts in this area disagree about some of the diagnoses, that the disease process evolves in atypical ways within patients, and that there is a general lack of pathognomonic criteria to confirm diagnoses objectively (Lindberg et al., 1980).

In terms of building the model in this area, a key aspect throughout its development has been testing the model against a data base of clinical cases that includes the correct diagnosis for each case; a correct diagnosis was decided by an agreement of at least two out of three rheumatologists. After an initial design consisting of 18 observations and 35 rules, the model has undergone many cycles of testing and revision. This incremental process resulted in the expansion of the model to include 150 observations, of which several observations were combined by rules to reach intermediate conclusions, and a total of 147 rules. The model has been critiqued by an external panel of expert rheumatologists, and a review of performance has shown the model to achieve diagnostic accuracy in 94% of 145 clinical cases (Lindberg et al., 1980). Current efforts include expanding the model to cover other rheumatic diseases and to provide advice about treatment management.

## 18.4 Stages of Model Development

The use of SEEK assumes the specification of a tabular model for each final diagnosis and the entry of cases, including the correct final diagnosis assigned to each case. The stages of model development that will be discussed are listed below.

*Stages in the Design of an Expert Model*

- Initial design of the model
- Data entry: cases with correct conclusions
- Performance summary of the model
- Analysis of the model
- Generation of model refinement experiments
- Refinement of the model
- Impact of model changes on the data

### 18.4.1 Initial Design of the Model

A text editor is used to specify an initial design of the model. Any one of three editing modes can be specified by the model designer: table input, table update, or table review and store. For each newly identified final diagnosis, table input mode allows the model designer to list major and minor observations and to specify components of the rules that would conclude the diagnosis. In table update mode, the table for a specified final diagnosis is retrieved, and the model designer can revise the rules or the lists of major and minor observations. When the additions and updates are completed, the table is stored and translated into a format used by SEEK. The translation of the table is to the EXPERT format (Weiss and Kulikowski, 1979) so that a consultation session (to be described in the next section) looks the same as one in EXPERT.

### 18.4.2 Entry of Data in a Consultation Session

A questionnaire is used to enter the observations, including the correct final diagnosis for a case. Editing facilities are available to review and to change the responses to questions. A case is stored in a data base that is maintained by the system. Figure 18-1 shows the entry of data for a particular case. After all questions have been asked, the system provides a summary of the data in the case. From this, the expert can correct any data entry errors, and, later, the case can be stored in a data base. Cases are usually entered in large groups during a single session. Typically, the tedious cycle that is repeated for each case consists of data entry, fixing errors, and saving the case. However, the expert can request the model's diagnosis for any case and at any time during this stage. An example (continuing with the case entered above) of the interpretative analysis output provided is shown in Figure 18-2. This includes the differential diagnosis (i.e., definite rheumatoid arthritis and possible SLE) followed by detailed lists of findings that provide a more complete picture of the case. These lists are

```
CASE TYPE:
                    (1) Case Entry     (2) Visit Entry     (3) Case Review
                    (4) Case Deletion  (5) Demo Entry      (6) Program Exit: 1

Enter Name or ID Number: test

Case Type: (1)Real (2)Hypothetical *2
Enter Date of Visit: 6/22/81

Enter Initial Findings (Press RETURN to begin questioning):
*

 1. Extremity Findings:
    1) Arthralgia
    2) Arthritis ≤6 wks. or non-polyarticular
    3) Chronic polyarthritis >6 wks.
    4) Erosive arthritis
    5) Deformity: subluxations or contractures
    6) Swollen hands, observed
    7) Raynaud's phenomenon
    8) Polymyalgia syndrome
    9) Synovial fluid inflammatory
    10) Subcutaneous nodules
 Checklist:
 *1,2,3,4,10
   .
   .
   .
31. Presumptive Diagnosis:
    1) Mixed Connective Tissue Disease
    2) Rheumatoid Arthritis
    3) Systemic Lupus Erythematosus
    4) Progressive Systemic Sclerosis
    5) Polymyositis
    6) Primary Raynaud's
    7) Sjogren's
 Checklist:
 *2
```

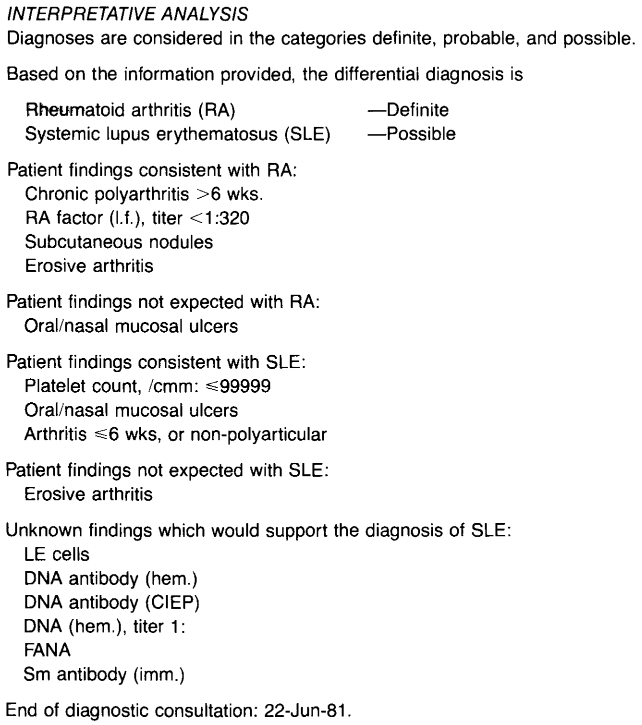
FIGURE 18-1 The entry of data for a case.

obtained by matching findings from the case data to prespecified lists that are associated with each final diagnosis in the model; the lists include those findings consistent, not expected, and unknown for the diagnosis.

### 18.4.3 Model Performance

A typical mode of interaction with SEEK involves iterating through these steps:

- obtain performance of rules on the stored cases
- analyze the rules
- revise the rules



*INTERPRETATIVE ANALYSIS*
Diagnoses are considered in the categories definite, probable, and possible.

Based on the information provided, the differential diagnosis is

| | |
|---|---|
| Rheumatoid arthritis (RA) | —Definite |
| Systemic lupus erythematosus (SLE) | —Possible |

Patient findings consistent with RA:
Chronic polyarthritis >6 wks.
RA factor (l.f.), titer <1:320
Subcutaneous nodules
Erosive arthritis

Patient findings not expected with RA:
Oral/nasal mucosal ulcers

Patient findings consistent with SLE:
Platelet count, /cmm: ≤99999
Oral/nasal mucosal ulcers
Arthritis ≤6 wks, or non-polyarticular

Patient findings not expected with SLE:
Erosive arthritis

Unknown findings which would support the diagnosis of SLE:
LE cells
DNA antibody (hem.)
DNA antibody (CIEP)
DNA (hem.), titer 1:
FANA
Sm antibody (imm.)

End of diagnostic consultation: 22-Jun-81.

**FIGURE 18-2 The interpretative analysis for the case in Figure 18-1.**

In reviewing the performance of a model, the expert's conclusions are matched to the model's conclusions. The expert's conclusion is stored with each case, while the model's conclusion is taken as that conclusion reached with the greatest certainty.

## Conditions for Performance Evaluation

The first step is to produce performance results on all stored cases. As mentioned earlier, evaluating performance involves matching the expert's conclusion to the model's conclusion in each case. A practical problem for scoring the results in a particular case occurs when ties in certainty between the expert's conclusion and the model's different conclusion are noted. Whether the model is scored as correct or incorrect for such a case affects the direction of subsequent rule refinements. A decision on how ties should be treated in performance evaluation rests with the problem domain. Whereas ties may be acceptable in particular medical areas for which it is

| *Current Performance* | | | *False positives* |
|---|---|---|---|
| Mixed connective tissue disease | 9/33 | (27%) | 0 |
| Rheumatoid arthritis | 42/42 | (100%) | 9 |
| Systemic lupus erythematosus | 12/18 | (67%) | 4 |
| Progressive systemic sclerosis | 22/23 | (96%) | 5 |
| Polymyositis | 4/5 | (80%) | 1 |
| Total | 89/121 | (74%) | |

**FIGURE 18-3 Summary of the model's performance.**

difficult to discriminate between competing diagnoses, they probably would not be acceptable in areas for which the diagnostic choices are well understood and mutually exclusive. Rheumatology is an area that exemplifies the former condition. For instance, particular rheumatic diseases do coexist during the progression of the respective disease processes, and therefore a final diagnosis is difficult to make. In such cases, a tentative diagnosis may be made that does not rule out other related diseases. An interpretation of a model's conclusions could reflect this situation by treating ties in certainty as correct (e.g., ties in certainty at the possible or probable confidence level). There may be exceptions. For example, ties at the definite level and at the null level (i.e., no conclusion was reached by the model) may be considered incorrect for diagnostically related diseases. The point of this discussion is to motivate the need for specifying a condition under which performance evaluation is to be performed. SEEK allows the model designer to specify how ties in confidence are to be treated.

Another condition is to allow the model designer to determine which rules and cases are to be ignored during the evaluation process. This has been found useful when either there are insufficient numbers of cases for a particular final diagnosis or the rules are not deemed to be in a satisfactory state by the model designer. If not turned off, these rules usually interfere in several case diagnoses, and their performance over all cases is therefore quite low. SEEK allows the model designer to specify rules to be turned off for performance evaluation.

## Performance Summary of the Model

The results are organized according to final conclusions and show the number of cases in which the model's conclusion matches the expert's conclusion. The column labeled *False positives* shows the number of cases in which the indicated conclusion was reached by the model, but did not match the stored expert's conclusion. In Figure 18-3, the summary of performance for mixed connective tissue disease indicates that 9 cases out of

| | |
|---|---|
| Rule 72: | 2 or more Majors for RA (MJRA)<br>2 or more Minors for RA (MNRA)<br>No Exclusion for RA (EX102)<br>→ Probable Rheumatoid arthritis (RA) |
| 43 Cases: | in which this rule was satisfied. |
| 13 Cases: | in which the greatest certainty in a conclusion was obtained by this rule and it matched the expert's conclusion. |
| 7 Cases: | in which the greatest certainty in a conclusion was obtained by this rule and it did not match the expert's conclusion. |

**FIGURE 18-4 Summary of a specific rule's performance.**

33 were correctly diagnosed. Furthermore, there are no cases that were misdiagnosed by the model as mixed connective tissue disease. The rules that conclude rheumatoid arthritis perform quite well for the stored rheumatoid arthritis cases, but they also appear to be candidates for specialization because of the 9 false positives.

In addition to the results shown in Figure 18-3, performance results about a specific rule can be obtained that show the number of cases in which the rule was satisfied. An example of this is shown in Figure 18-4, and includes the number of cases in which the rule was used successfully (i.e., matching the expert's conclusions stored with the cases) and the number of cases in which the rule was used incorrectly (i.e., not matching the expert's conclusions stored with the cases).

## 18.4.4 Analysis of the Model

Interactive assistance for rule refinement is provided during the analysis of the model. The model designer has the option of selecting either "single case" or "all cases" as a basis of analysis.

### Analysis of the Model in a Single Case

Analysis in a single case proceeds after a case has been chosen from the data base of stored cases. The objective of single case analysis is to provide the model designer with an explanation of the model's results in the case. This is done by first showing the model's confidence in both the expert's conclusion and the model's conclusion. Rules are cited that were used to reach these conclusions. Rules for the expert's conclusion are selected from those rules in the model with the same conclusion as the conclusion stored (by the expert) for a case. If the model's conclusion does not match the expert's conclusion in the case, the system attempts to locate a partially

```
CASE:           3

Expert conclusion: Progressive systemic sclerosis
Model conclusion: Probable Rheumatoid arthritis

This is the strongest satisfied rule for the expert's conclusion:

Rule 111: 1 or more Majors for PSS (MJPSS) (1 Majors Satisfied)
    1 or more Minors for PSS (MNPSS) (3 Minors Satisfied)
    → Possible Progressive systemic sclerosis (PSS)

This is the rule for the model's conclusion:

Rule 72: 2 or more Majors for RA (MJRA) (2 Majors Satisfied)
    2 or more Minors for RA (MNRA) (3 Minors Satisfied)
    No Exclusion for RA (EX102) (Satisfied)
    → Probable Rheumatoid arthritis (RA)

There exists 1 partially satisfied rule for PSS with weight
assignment ≥ that set by RA rule

Rule 112: Requirement 1 for probable PSS (RR105) (Not set)
    No Exclusion for probable PSS (ER105) (Satisfied)
    → Probable Progressive systemic sclerosis (PSS)
```

**FIGURE 18-5 Results of a case analysis.**

satisfied rule for the expert's conclusion that is the "closest" to being satisfied and would override the model's incorrect conclusion. A procedure for finding the "closest" rule is described later. An example of the results of single case analysis is shown in Figure 18-5. Case 3 is misdiagnosed by the model, which has assigned the certainty value of "possible" to progressive systemic sclerosis. The model's conclusion is rheumatoid arthritis with a certainty value of "probable." Rule 111 and Rule 72 are responsible for reaching these conclusions. Each line printed for a rule contains an internal label for reference purposes, such as MJPSS. In this example, Rule 72 was triggered because two majors and three minors for rheumatoid arthritis are present, and Case 3 did not have the (exclusion) findings that would deny Rule 72. Given this information the model designer can pursue either of two directions to refine the rules: to weaken Rule 72 so that it will not override Rule 111, or to find a stronger rule concluding progressive systemic sclerosis. In response to this latter possibility, SEEK cites Rule 112 as a likely candidate to generalize. A procedure that SEEK uses to identify rules such as Rule 112 is described in the next section.

Besides this information provided in single case analysis, SEEK allows the model designer to interrogate any conclusion in the model, both final and intermediate results. The rules for any conclusion can be cited by specifying a rule number or the internal label tagged to a conclusion (e.g., PSS). In the latter situation, all rules for a conclusion are cited, both totally satisfied and partially satisfied rules in the case. This aids the model designer in reviewing the performance of a subset of the rules on the case data.

## Analysis of the Model Based on Case Experience

The first step for the analysis of the model for all cases is to specify a final diagnosis for which rules are to be analyzed. In this manner, the model designer focuses the analysis on the subset of the rules in the model. The analysis is usually done after performance results have been obtained. SEEK assists the model designer in the analysis of a subset of the rules that are relevant to the misdiagnosed cases. An important design consideration for SEEK is to provide the model designer with a flexible means to perform experiments in refining the rules. In this section, advice will be described that helps in determining the specific experiments for rule refinement. Heuristic procedures are needed to select experiments from the many possibilities. For example, SEEK uses a heuristic procedure by tracing rules that conclude the stored expert's conclusion to determine which rules are "closest" to being satisfied. It looks for a partially satisfied rule for which the following conditions hold:

1. the rule concludes at a minimum confidence level that is greater than (or equal to, depending on the treatment of ties) the certainty value for the model's conclusion;
2. the rule contains the maximum number of satisfied components for all rules concluding at that confidence level.

A rule satisfying these conditions is marked for generalization, so that it may be invoked more frequently. The rule used to reach the model's conclusion is marked for specialization, so that it may be invoked less frequently.

In the following example, SEEK analyzes the rules for the specified diagnosis, mixed connective tissue disease, with regard to their use on the stored cases. After analysis, SEEK reports the results by numbering and listing rules that conclude mixed connective tissue disease, for which there exists information to indicate that the rule is a potential candidate for generalization or specialization. Figure 18-6 is a summary of this rule analysis and shows unsatisfied rules in the misdiagnosed cases for mixed connective tissue disease that are candidate rules for generalization. The column labeled *Generalization* contains the number of cases suggesting the generalization of a rule, and the column labeled *Specialization* contains the number of cases suggesting the specialization of a rule.

In Figure 18-6, rules at the possible level of certainty are strong candidates for generalization. Although Rule 56 is not satisfied in eight misdiagnosed cases, if Rule 56 had been satisfied, these eight cases would have been correctly diagnosed. In the eight cases cited for Rule 56, Rule 56 is "closer" to being satisfied than Rule 55 is. A more detailed analysis of each rule, summarizing the satisfied and unsatisfied components of the rule, is normally obtained at this point. Rule 55 can be stated as follows: if the

*Mixed Connective Tissue Disease*

| *Rule* | *Certainty* | *Generalization* | *Specialization* |
|---|---|---|---|
| 54. | Possible | 2 | 0 |
| 55. | Possible | 7 | 0 |
| 56. | Possible | 8 | 0 |
| 57. | Probable | 2 | 0 |
| 58. | Probable | 2 | 0 |

**FIGURE 18-6 Summary of rule analysis for the diagnosis of mixed connective tissue disease.**

patient has two or more major observations for mixed connective tissue disease and RNP antibody is positive, then conclude possible mixed connective tissue disease. Rule 56 can be stated as follows: if the patient has three or more major observations for mixed connective tissue disease, then conclude possible mixed connective tissue disease. A simple experiment for generalization of Rule 56, which might be tried first because it is the simpler rule, is to decrease the number of major observations required.

The scheme for analysis in all cases focuses on a subset of the rules by gathering empirical information suggesting the generalization and specialization of rules in the set. This can be viewed as a learning system. In Mitchell's version space approach (Mitchell, 1979), two sets of rules are maintained as bounds on the "maximally specialized" rules and the "maximally generalized" rules that are consistent with the training cases presented for a conclusion. A training case is prespecified as either *positive*—a rule must be found to cover the case—or *negative*—no rule should match the case. The scheme seeks to cover all positive cases while allowing no negative cases to match any of the rules. There are no certainty values assigned to the rules in the version space. Our scheme seeks to refine expert-derived rules that have been categorized by confidence levels in the model. Correct classification for all cases is not required. That is, a negative case is allowed to be covered so long as there is a rule for another conclusion that overrides the matched rule(s). A rule is marked for generalization or specialization based on the comparison of the certainty values assigned to the final conclusion expected to that reached by the model. Finally, our scheme is interactive in nature, requiring the involvement of the model designer. It is not intended to be an autonomous learning system.

### 18.4.5 Generation of Model Refinement Experiments

As was shown in Figure 18-6, SEEK indicated several mixed connective tissue disease rules that are candidates for generalization. In general, there are many possibilities that can be tried for refining the rules in a model. A difficult task is to select a rule or group of rules to work on and then to

24 cases in which the expert's conclusion MCTD does not match the model's conclusion:

1, 4, 11, 12, 14, 15, 42, 47, 49, 57, 60, 67, 71, 75, 78, 80, 84, 93, 99, 100, 104, 105, 107, 130

**Proposed Experiments for Mixed Connective Tissue Disease**

1. Decrease the number of majors in rule 56.
2. Delete the requirement component in rule 55.
3. Delete the requirement component in rule 54.
4. Decrease the number of minors in rule 57.
5. Delete the requirement component in rule 58.

**FIGURE 18-7 List of misdiagnosed cases of mixed connective tissue disease and proposed experiments for improving the rules.**

determine plausible refinements beyond classifying a rule as a candidate for generalization or specialization. In this section, an approach to suggest automatically plausible experiments for refining the rules in a model is described.

A heuristic rule-based scheme is used to suggest experiments. The heuristic rules are called EX-rules so as not to confuse them with the expert-modeled rules. The IF part of an EX-rule contains a conjunction of predicate clauses that essentially looks for certain features about the performance of rules in the model, while the THEN part of an EX-rule contains a specific rule refinement experiment. An example of an EX-rule is shown below and is used to suggest the specific generalization experiment to decrease the number of major findings in a rule. Currently, there are eleven EX-rules, which are divided almost equally with respect to the types of experiments (i.e., generalizations or specializations) that may be suggested.

```
IF:  the number of cases suggesting generalization of the rule is greater
     than the number of cases suggesting specialization of the rule and the
     most frequent missing component in the rule is the major component,
THEN:  decrease the number of major findings in the rule.
```

Evaluation of an EX-rule begins by instantiating the clauses with the required empirical information about a specific rule in the model. Function calls are used to gather the information. After instantiation, the clauses are evaluated in order beginning with the first clause in the EX-rule. If all clauses are satisfied, then the specific experiment is posted. All EX-rules are evaluated in this manner for a specific rule in the model. The experiments suggested by the EX-rules are narrowed by the expert to those changes consistent with his or her medical knowledge. In Figure 18-7, the experiments for improving the rules used in reaching the diagnosis of mixed connective tissue disease are presented after listing the misdiagnosed cases of mixed connective tissue disease.

The experiments are ordered based on maximum potential performance gain on the cases. Other criteria for ordering can be used such as

```
:why(1)

    If rule 56 had been satisfied, 8 currently misdiagnosed MCTD cases would have been diagnosed correctly.
    Currently, rule 56 is not used incorrectly in any of the cases. In rule 56 the component missing with the
    greatest frequency is Major.

    Therefore, we suggest decreasing the number of majors in rule 56.
    This would generalize the rule so that it will be easier to satisfy.
```

**FIGURE 18-8 Explanation of a proposed experiment.**

ease of change (e.g., an experiment that suggests changing the minors in a rule may be preferred over an experiment that suggests changing the majors). An explanation of a particular experiment is provided by a translation of the specific EX-rule used to suggest the experiment into a narrative statement containing the empirical information about the rule. As an example, the support for the first experiment is shown in Figure 18-8. It should be emphasized that a decision as to which experiments, if any, are to be tried is left to the model designer. Even though a particular experiment is supported empirically, the ultimate decision should include justifying an experiment in terms of other knowledge about the domain. For example, is a rule resulting from the first experiment for Rule 56 "medically sound" to make the diagnosis? This can lead to reconsidering the lists of major and minor findings for a particular final diagnosis and to potentially refining these findings.

It should be noted that one is not absolutely certain of a net gain in performance before an experiment is tried. In the case of a generalization experiment, there may be more than one unsatisfied component in a rule marked for generalization; the marking procedure picks the first unsatisfied component in the rule. Facilities for performing experiments and for determining the impact of changes on the cases are described later.

## 18.4.6 Refinement of the Model

After an experiment to revise the rules has been determined, the model designer can test his or her proposed revision on the cases. This is facilitated by editing capabilities that permit the model designer to interrogate and to modify the rules in the model. The changes are logged separately from the rules in the model so that the original rules can be restored. The editing functions include changing:

- the number of major or minor observations,
- the requirement component,
- the exclusion component, and
- any rule reaching an intermediate result that is used by other rules.

```
Candidate for Change is MJMCT in rule 56
Rule 56 is:
    3 or more Majors for MCTD (MJMCT)
    → Possible Mixed connective tissue disease (MCTD)
Generalization of Rule 56 is:
    2 or more Majors for MCTD (MJMCT)
    → Possible Mixed connective tissue disease (MCTD)
```

**FIGURE 18-9 SEEK's description of the proposed rule change.**

Continuing with our example, Figure 18-9 shows the response by SEEK for the model designer's suggested change to Rule 56: to change the number of majors required by Rule 56 to be 2 or more majors. The commands that allow the model designer to interrogate and to modify the rules require rule numbers or symbolic labels to reference parts of the model.

### 18.4.7 Impact of Model Changes on the Data

The results of a specific experiment are obtained by conditionally incorporating the revised rule(s) into the model. The updated model is then executed on the data base of cases. The results are summarized in Figure 18-10 for making the change to Rule 56. In this example, such a modification significantly improves performance. Several misdiagnosed cases of mixed connective tissue disease are now correctly diagnosed by the model.

| | *Before* | *False positives* | *After* | *False positives* |
|---|---|---|---|---|
| MCTD | 9/33 (27%) | 0 | 17/33 (52%) | 0 |
| Others | 80/88 (91%) | (see below) | 80/88 (91%) | (see below) |
| Total | 89/121 (74%) | | 97/121 (80%) | |
| | *Details of Effect on Other Diseases* | | | |
| RA | 42/42 (100%) | 9 | 42/42 (100%) | 8 |
| SLE | 12/18 (67%) | 4 | 12/18 (67%) | 3 |
| PSS | 22/23 (96%) | 5 | 22/23 (96%) | 3 |
| PM | 4/5 (80%) | 1 | 4/5 (80%) | 1 |

**FIGURE 18-10 Results of executing updated model on the data base of cases.**

Moreover, there was no adverse side effect of this change on other cases with different stored conclusions. The model designer has the option either to accept or to reject the experiment. If a simple modification does not lead to desirable results, more complicated changes may be tried, such as multiple modifications or dropping a condition in a requirement.

## 18.5 Discussion

The tabular model appears to be a reasonable framework for encoding expert knowledge in a real and complex application. Excellent performance was achieved for the diagnosis of mixed connective tissue disease (Lindberg et al., 1980). This approach has proven particularly valuable in assisting the expert in domains where two diagnoses are difficult to distinguish. For example, there is a general lack of deterministic clinical criteria to confirm the diagnoses in the connective tissue disease area. The experts obtain by means of empirical testing a measure of the usefulness of the observations expressed in the tabular model. There are limitations to this approach—for some applications it may be difficult to express rules using major and minor observations or using only three levels of confidence. Although this model may not be the most expressive model for capturing expert knowledge, it is a model that is suitable for an empirical analysis leading to experimentation with rule refinement. Samples of cases are not completely representative and cannot begin to match the scope of the expert's knowledge. But as others have found (Gaschnig, 1979), even with small samples of cases, empirical evidence can be of great value in designing and verifying an expert model.

Ideally, a tabular model abstracts the expert's reasoning in diagnostic criteria, while cases cite evidence that is accurately diagnosed by the model. The use of SEEK attempts to achieve this harmony by pointing out potential problems with these dual sources of knowledge. Given the performance of the cases, potential problems with the rules can be identified with the tools described earlier. The summarized performance results are a means for the expert to rethink a tabular model that is performing poorly for a specific diagnosis. The analysis of the tabular rules based on case experience sharply focuses the expert's attention on modifications that potentially result in improved performance and that are medically sound. This can lead to reviewing individual cases for inaccuracies in the data and to reconsidering the importance of specific criteria in the model. It should be emphasized that this process is not intended to "custom-craft" rules solely to the cases, but rather to provide the expert an interactive environment with explicit performance information that needs to be accurately explained. From an artificial intelligence perspective, this may be viewed as a learning process based on experience in developing the model. From the empirical testing and successive improvements in the performance of the model, the human expert will obtain not only a better formulation of the model but also a better understanding of the explicit diagnostic criteria used in his or her reasoning.

## ACKNOWLEDGMENTS

This research was supported in part by grant RR-643 of the National Institutes of Health Biotechnology Resources Program. We gratefully acknowledge the contributions of our collaborators in developing a rheumatology consultation system: G. Sharp, M.D., and D. Lindberg, M.D., and their fellow researchers at the University of Missouri at Columbia. We are also grateful to Casimir Kulikowski for his review of this work.