17

Discovery, Confirmation, and Incorporation of Causal Relationships from a Large Time-Oriented Clinical Data Base: The RX Project

Robert L. Blum

In the mid-1970s Robert Blum, a physician with an interest in medical AI, went to Stanford as a fellow in clinical pharmacology and a doctoral student in computer science. He soon learned about the well-known TOD data base work of James Fries and Gio Wiederhold (the time-oriented data bank that is used as the basis for an international rheumatology network known as ARAMIS—the American Rheumatism Association Medical Information System). Working with Wiederhold, he developed the concept of a computer program to derive new clinical knowledge from such data. His doctoral dissertation, known as RX, used a subset of the ARAMIS data base for this kind of investigation. RX differs from the other systems described in this book because the emphasis is not on consultation but on the use of AI techniques to guide the analysis of collected data. RX is knowledge-based in the sense that it requires not only the observations from a data base, but also underlying knowledge of pathophysiology, causality, and statistics.

As Blum describes in this chapter, the objectives of the RX research are threefold: (1) to automate the processes of hypothesis generation and ex-

From *Computers and Biomedical Research*, 15: 164–187 (1982).

ploratory analysis of data in a large nonrandomized, time-oriented clinical data base, (2) to provide knowledgeable assistance in performing studies on large data bases, and (3) to increase the validity of medical knowledge derived from nonprotocol data (i.e., data that are collected without formal guidelines or an experimental question in mind). In addition to the ARAMIS data and knowledge of pathophysiology and statistics, RX is composed of a discovery module and a study module. The knowledge in the system is organized hierarchically and is used to assist in the discovery and study of new hypotheses. Confirmed results from the data are automatically entered into the knowledge base for future use. Thus the work is related to research in learning, where the goal is to develop programs that can assimilate new knowledge by observing and analyzing past experience.

When RX starts running, it begins the "discovery" process by scanning the ARAMIS data. The discovery module uses lagged, nonparametric correlations to generate hypotheses of clinical interest. These are then studied in detail by the study module, which automatically determines confounding variables and methods for controlling their influence. In determining the confounders of a new hypothesis, the study module uses previously "learned" causal relationships. The study module selects a study design and statistical method based on knowledge of confounders and their distribution in the data base. Most of the RX experiments have used a longitudinal design involving a multiple regression model applied to individual patient records.

The importance of this work lies in its merging of AI, data bases, and statistics and in the thoughtful characterization of causality that Blum has devised. In characterizing the difference between data and knowledge (see Chapter 3), authors have often noted that knowledge is derived from data that are analyzed and validated. In RX we see that this process of data analysis is itself a knowledge-based task. Note, also, that new knowledge, once derived and added to the knowledge base, can then be used to guide further data analyses in the future. The analogy to intellectual growth and learning is clear, but equally evident is the importance of validation before new correlations are accepted as fact. RX continues to be an active area of research for Blum and his colleagues.

17.1 Introduction

Every year, as computers become more powerful and less expensive, increasing amounts of health care data are recorded on them. Motivation for collecting data routinely into ambulatory and hospital medical record systems comes from all quarters. Health practitioners require sets of data for clinical management of individual patients. Hospital administrators require them for billing and resource allocation. Government agencies re-

quire data for assessments of the quality of health care. Third-party insurers require them for reimbursement. Data bases may also be used for performing clinical research, for assessing the efficacy of new diagnostic and therapeutic modalities, and for the performance of postmarketing drug surveillance.

The various uses for data bases may be grouped into two fundamentally distinct categories. The first category pertains to uses that merely require *retrieval of a set of data.* For example, we may wish to know the names of all patients who had a diastolic blood pressure greater than 100 for more than six months and who received no treatment. Uses of medical record systems for patient management, billing, and quality assurance usually fall into this category. The second use of data bases is for *deriving or inferring facts* about the world in general. For example, we might request data from a health insurance data base on occupation and hospital diagnoses to determine whether certain occupations are associated with an increased prevalence of heart disease. Here the predominant interest is in generalizing from the data base and only secondarily in the particular values in the data base. The use of data bases for determining causal effects of drugs, for establishing the usefulness of new tests and therapies, or for determining the natural history of diseases falls into this latter category.

The possibility of deriving medical knowledge from data bases is an important reason for establishing them. Given a collection of large, geographically dispersed medical data bases, it is easy to imagine using them for discovering new causal relationships or for confirming hypotheses of interest.

The RX Project, as this research project is called, is a prototype system for automating the discovery and confirmation of hypotheses from large clinical data bases. The project was designed to emulate the usual method of discovery and confirmation of medical knowledge that characterizes epidemiological and clinical research. The following hypothetical scenario serves to illustrate this method.

17.2 Evolution of Empirical Knowledge

Suppose a medical researcher has noticed an interesting effect in a small group of patients, say unusual longevity. He carefully examines those patients' records looking for possible explanatory factors. He discovers that heavy physical exertion associated with occupation and sports is a possible factor in promoting longevity.

Interested in pursuing the hypothesis that "heavy physical exertion predisposes to long life," the medical researcher consults with a statistician, and together they design a comprehensive study of this hypothesis. First, they analyze the results of the study on their local data base, controlling

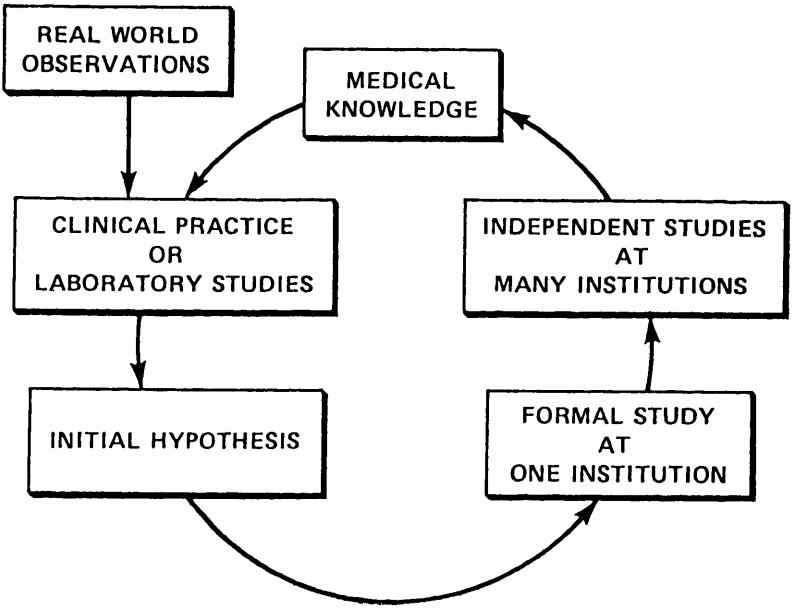

FIGURE 17-1 The evolution of medical knowledge.

for factors known to be associated with longevity. Having confirmed the hypothesis on one data base, they proceed to test the hypothesis on many other data bases, modifying the study design to allow for differences in the type and quantity of data. Having confirmed the hypothesis, they publish the result, and other researchers proceed with further confirmatory studies, attempting to elucidate the mechanism of the "exercise effect." When future researchers study other factors that influence longevity, they control for physical activity.

This cycle in which knowledge gradually evolves from data through a succession of increasingly comprehensive studies is illustrated in Figure 17-1. At each stage of discovery and confirmation existing medical knowledge is used to design and to interpret the studies.

17.3 The RX Project

It is easy to imagine automating at least parts of the above cycle of discovery and confirmation. We obtain our initial hypotheses by selectively combing through a large data base, examining a few patient records guided by prior knowledge. These clues are then studied more comprehensively on the

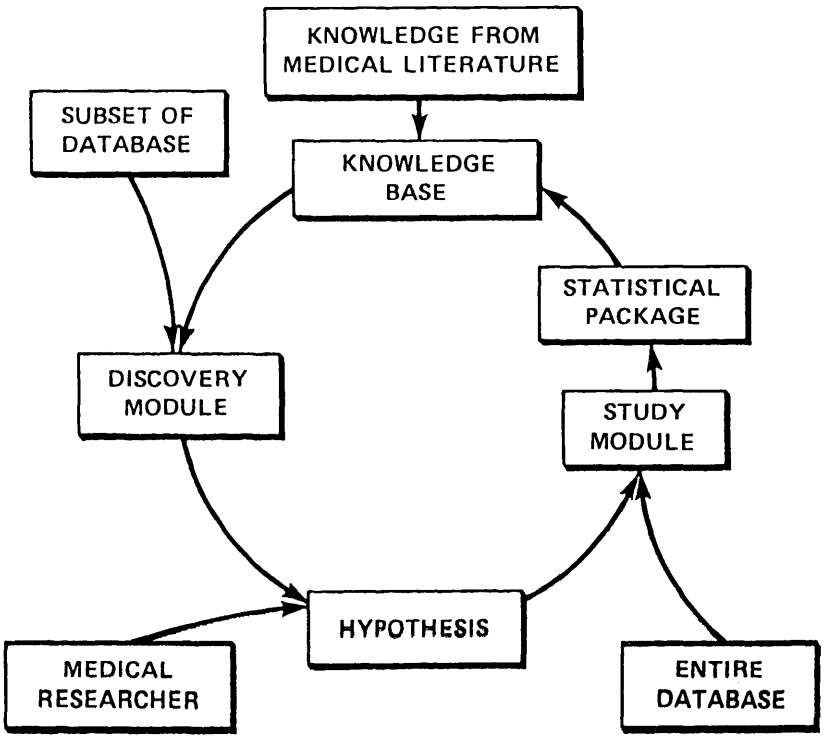

FIGURE 17-2 Discovery and confirmation in RX.

data base as a whole. To design and interpret these studies, medical and statistical knowledge from a computerized knowledge base is used. The final results are incrementally incorporated into the knowledge base, where they can be used in the automated design of future studies.

This describes the RX computer program, a prototype implementation of these ideas. Besides a data base, the RX program consists of four major parts: the discovery module, the study module, a statistical analysis package, and a knowledge base (Figure 17-2).

- The *discovery module* produces hypotheses of the form "*A* causes *B*." The hypotheses denote that in a number of individual patient records "*A* precedes and is correlated with *B*." Information from the knowledge base is used to guide the formation of initial hypotheses.
- The *study module* then designs a comprehensive study of the most promising hypotheses. It takes into account information in the knowledge base in order to control for known factors that may have produced a spurious association between the tentative cause and effect. The study module

uses statistical knowledge in the knowledge base to design an adequate statistical model of the hypothesis.

- The *statistical analysis package* is invoked by the study module to test the statistical model. The analysis package accesses relevant data from patient records, and then applies the statistical model to the data. The results are returned to the study module for interpretation.
- The *knowledge base* is used in all phases of hypothesis generation and testing. If the results of a study are medically and statistically significant, they are tentatively incorporated into the knowledge base where they are used to design further studies. Newly incorporated knowledge is appropriately labeled as to source, validity, evidential basis, and so on. As the knowledge base grows, old information is updated.

Currently, the RX program uses only one data base: a subset of the ARAMIS data base. Also, the extent of medical and statistical knowledge is limited, since the purpose of the research was primarily the development of methodology.

While the program is a prototype, it has been operational since 1979 and has been widely demonstrated. Several interesting medical hypotheses (in varying states of confirmation) have been discovered by the program, including some with little prior supporting evidence.

The objective of this chapter is to present an overview of the RX Project. Details on statistical methods, modeling of causal relationships, and methods of knowledge representation may be found in Blum (1982).

17.4 Time-Oriented Data Bases

The general format of a patient record is illustrated in Table 17-1. Each time a patient is seen in a clinic a number of observations are made. These are recorded with the date of observation in the data base. The recorded

TABLE 17-1 Hypothetical Time-Oriented Record for One Patient

VISIT NUMBER:	1	2	3
DATE:	17 Jan 79	23 Jun 79	1 Jul 79
KNEE PAIN:	severe	mild	mild
FATIGUE:	moderate		moderate
TEMPERATURE:	38.5	37.5	36.9
DIAGNOSIS:	systemic lupus		
WHITE BLOOD COUNT:	3500	4700	4300
CREATININE CLEARANCE:	45		65
BLOOD UREA NITROGEN:	36	33	
PREDNISONE:	30	25	20

characteristics of a patient are known as *primary attributes*, or simply *attributes*. Attributes may be real-valued, rank, categorical, or binary. The term *attribute* includes all recorded signs and symptoms, lab values, diagnoses, therapy, and functional states.

The defining characteristic of a time-oriented data base is that *sequential values for each attribute may be recorded.* Note that different attributes may be recorded on different patients and that the time intervals between values will usually differ. Some attributes may have values that are only sporadically recorded or not at all. In general, the quantity and character of data across patients may vary greatly.

All of the research reported here was done using a subset of the ARAMIS/TOD data base of rheumatology (American Rheumatism Association Medical Information System/Time-Oriented Databank) collected at Stanford University from 1969 to the present (Fries, 1972; Wiederhold et al., 1975). The subset contains the records of 50 patients with severe systemic lupus erythematosus (SLE). The average number of clinic visits for each patient was 50, and the average length of follow-up was five years. Patient records contained 52 attributes.

The size of the data base used in this project, a small sample of the ARAMIS data base, is approximately a half-million characters—much greater than available core storage on our computers after programs have been loaded. Patient records are kept in hash files on disk, where they are stored in compressed and transposed format. Indices for each attribute are maintained specifying numbers of values for each patient. Details of data storage and display methods may be found in Blum (1981).

17.5 Computer Facilities and Languages

Research was performed at two computer facilities at Stanford University: SUMEX-AIM and SCORE. At the time SUMEX-AIM featured a DEC dual processor KI-10 running the TENEX operating system. Currently both SUMEX and SCORE have DEC 20/60's running TOPS-20. The ARAMIS data base *per se* is stored at the Stanford Center for Information Technology on an IBM 370/3033. Data transfer was accomplished by magnetic tape.

All computer programs are written in Interlisp (Teitelman, 1978), a dialect of LISP, a language that is highly suitable for knowledge manipulation. Statistics are performed in IDL (Interactive Data-Analysis Language) (Kaplan et al., 1978), discussed later. The RX source code with knowledge base comprises approximately 200 disk pages of 512 words each.

17.6 The Knowledge Base

While the prospect of using clinical data bases to discover or to confirm medical hypotheses is tantalizing, there are formidable problems in making inferences from nonrandomized, nonprotocol data. These include numerous forms of treatment and surveillance bias, poor adjustment for covariates, inadequate specification of patient subsets, and improper use of statistical analysis (Blum and Wiederhold, 1978; Byar, 1980; Dambrosia and Ellenberg, 1980). The use of nonrandomized data for clinical inference demands more stringent data analysis, study designs of greater sophistication, and more thoughtful interpretation than does the use of data gathered in a randomized trial.

The leitmotif of the RX Project is that derivation of new knowledge from data bases can best be performed by integrating existing knowledge of relevant parts of medicine and statistics into the medical information system. During the evolution of a medical hypothesis, as was illustrated, existing medical knowledge comes into play at every stage.

In the RX computer program the medical knowledge base determines the operation of the discovery module, plays a pivotal role in the creation of subsequent studies in the study module, and, finally, serves as a repository for newly created knowledge. The medical knowledge base grows by automatically incorporating new knowledge into itself. Hence it must be designed in such a way that relationships derived from the data base can be translated into the same machine-readable form as knowledge entered from the medical literature by a researcher. In any case knowledge relevant to a study must be automatically accessible.

The main data structure of RX's knowledge base (KB) is a tree representing a taxonomy of relevant aspects of medicine and statistics. Each object in the tree is represented as a schema containing an arbitrary number of property:value pairs. The RX KB contains approximately 250 schemata pertaining to medicine, 50 pertaining to statistics, and 50 pertaining to the overall system. The medical knowledge in the RX KB covers only a small portion of what is known about systemic lupus erythematosus and some areas of general medicine. The present KB is merely a test vehicle; its size is 50 disk pages or 120,000 bytes.

17.6.1 Medical Knowledge

The medical knowledge base is a subtree of the KB distinct from the statistical knowledge base. Its first-order subtrees are *states* and *actions*, which in turn are broken down into *signs, symptoms, lab findings,* and *diseases* and into *drugs, surgery,* and *physical therapy*. The categories of diseases and other entities follow the conventional nosology based on organ systems and pa-

thology found in any standard textbook of medicine (Isselbacher et al., 1980). I will occasionally refer to each of the objects in the medical KB as a *node* and to the information stored at each node as its *schema.*

The schema for each object is represented as a collection of property:value pairs called a *property list.* In general, the objects in the KB are either primary attributes in the data base or *derived variables,* that is, objects whose values must be derived from primary data. The properties in an object's schema may be grouped into the following categories: *data base schema properties, hierarchical relationship properties, properties describing the definition of an object and its intrinsic properties,* and *properties describing cause/effect relationships to other objects.*

Data Base Schema Properties

Each of the attributes in the clinical data base is represented by a schema in the KB describing its units of measurement, how its values are stored, and so on. This kind of schema is typical of most data bases today (Wiederhold, 1977). As an example, part of the schema for the attribute hemoglobin appears below:

Hemoglobin
---
attribute-type: point-event
value-type: real {i.e., a real-valued number}
range: 0 < value < 25
significance: .1 {i.e., values are rounded to the nearest .1}
units: grams per deciliter

Hierarchical Relationship Properties

Two properties are used to store the position of an object in the medical hierarchy: *specialization* and *generalization,* abbreviated *spec* and *genl* as shown below.

Respiratory Diseases
genl: All Categories of Disease
spec: Pneumonia, Asthma, Emphysema

Pneumonia	Asthma	Emphysema
genl: Respiratory Dis.	*genl*: Respiratory Dis.	*genl*: Respiratory Dis.
spec: Pneumoncoccal Pn.	*spec*: Allergic Asthma	*spec*: CO_2 retention
Klebsiella Pn.	Intrinsic Asthma	

Inheritance mechanisms (Stefik, 1979) are used by the study module as a means for exploiting the knowledge implicit in the hierarchy. For example, in the course of a study, if the expected duration of klebsiella pneumonia was required to construct a statistical model, then a default value might be inherited from the schema for pneumonia.

Properties Pertaining to the Definition and Intrinsic Characteristics of an Object

If an object is a primary data base attribute like hemoglobin, then no definition is required, at least not from a standpoint of deriving values for it. Values for hemoglobin are simply those in the data base.

On the other hand, if the values for an object are derived from primary attributes, the specification of the means for derivation must be recorded in the KB. That is the object's definition. The didactic example below shows a definition for pneumonia.

Pneumonia

definition:		Temperature > 38 degrees C.
	and	WBC > 10,000 cells per mm^3
	and	Chest-XRAY = Lobar Infiltrate

In the RX KB the specification and use of definitions are far more complicated than is suggested by this example. Recall that data base attributes are time-oriented with nonuniform time intervals and frequently missing values. Hence definitions of derived objects must contain time-dependent predicates and mechanisms for handling sporadic values. Definitions can also refer to other derived objects. The temporal characteristics of an object may be specified using other properties in the schema: *expected duration, carry-over, onset delay,* and so on. These parameters are used by the time-dependent predicates when definitions for objects are evaluated.

Properties Specifying Causal Relationships to Other Objects

The final class of properties are those specifying the causal relationships of an object to other objects. In RX all causal relationships are stored using two properties: *effects* and *affected-by*. The *effects* property records a list of those objects directly affected by the object. The *affected-by* property contains a list of objects that directly affect it. Additionally, the detailed characteristics of the causal relationship between a pair of objects is stored on the *affected-by* property. The resulting causal model is a directed cyclic graph; that is, the representation allows for the possibility that A causes B causes A.

Besides the simple fact that A may affect B, each causal relationship is represented by a set of features as follows:

<intensity, frequency, direction, setting, functional form, validity, evidence>

Briefly, these take the following form when both the cause and effect are real-valued:

- *intensity*: the expected change in the effect given a change in the cause, expressed as an unstandardized regression coefficient
- *frequency*: the distribution of the effect across patients, expressed as deciles of the expected effect given a "strong" change in the causal variable
- *direction*: increase or decrease
- *setting*: the clinical circumstances specifically included or excluded from the study, expressed as a Boolean with time-dependent predicates
- *functional form*: the complete statistical model used to study the relationship, expressed in machine-readable form
- *validity*: a 1-to-10 scale distinguishing tentative associations from widely confirmed causal relationships
- *evidence*: a summary of the study performed by the study module, including patient ID's, methods, and intermediate results

The entire causal relationship is machine-readable. This enables it to be used automatically by the study module during subsequent studies. The causal relationships in the KB can also be interactively displayed in a variety of forms. All paths connecting two nodes may be displayed, or only the details of a particular causal relationship: its mathematical form, the evidence supporting it, or its distribution across patients. In the following example the effects of prednisone have been displayed. The verbs and adverbs in the phrases are supplied by a lexicon during machine translation.

```
PREDNISONE, at a level of 30 mgms/day,     {modal effects}
-------------------------------------------
usually increases CHOLESTEROL by 50 to 130 mgms/dl,
usually increases WEIGHT by 3 to 7 kgms,
regularly attenuates NEPHROTIC-SYNDROME by 1. to 2. gms protein/24 hrs,
regularly attenuates GLOMERULONEPHRITIS by 10. to 30. percent activity,
regularly decreases EOSINOPHILS by 2 to 3 percent of WBC,
commonly decreases ANTI-DNA by 50 to 90 percent activity,
occasionally increases GLUCOSE by 20 to 100 mgms/dl.
```

17.7 The Discovery Module

The general methodology used by RX to discover and then to study causal relationships is known as a generate-and-test algorithm. Briefly, the discovery module proposes causal links based on a test for strength of association and time precedence. After a number of tentative links have been added, the study module performs an exhaustive study of them in the same order in which they were added. In the course of this study many tentative links will be removed, and the remaining ones will be labeled with

detailed information on the respective relationships. After a link has been incorporated into the model, it may be used to refine the study of further links.

17.7.1 An Operational Definition of Causality

Underlying the discovery module and the study module is the following operational definition of causality: A is said to cause B if over repeated observations (1) A generally precedes B, (2) the intensity of A is correlated with the intensity of B, and (3) there is no known third variable, C, responsible for the correlation.

These properties are the foundation of the RX algorithm. I will refer to these properties as (1) time precedence, (2) covariation or association, and (3) nonspuriousness (Kenny, 1970; Suppes, 1970).

Causality can never be proven using observational data. The persuasiveness of a given demonstration simply depends on the extent to which the three properties have been shown.

17.7.2 Methodology of the Discovery Module

The function of the discovery module is to find candidate causal relationships. To do this, the discovery module exploits only the first two properties of causal relationships: time precedence and covariation.

The discovery module considers all pairs of variables, $\{A, B\}$, where A and B are either primary attributes in the data base or are derivable from primary attributes. It attempts to determine whether the data suggest that A causes B, B causes A, both, or neither. The output of the discovery module is an ordered list of hypotheses. A researcher may designate which potential causes and effects are of interest. For example, certain drugs and diseases might be tagged as being of interest in exploration. The algorithm is intrinsically slow, $O(n^2)$ where O is Order and n is the number of variables; however, it makes up for this inefficiency by its sensitivity and the speed with which simple correlations can be performed.

A pairwise algorithm was chosen for the discovery module after months of experimentation with multivariate methods. The latter cannot be applied to data of the type recorded in the ARAMIS data base without extensive loss of information. The reason is that values are only sporadically recorded and patients differ widely on covariates. The general philosophy in all RX procedures in either the discovery module or the study module is to analyze data only within *individual patient records*. That is, data in two patient records are never combined before statistical analysis. The computational expense incurred by analyzing individual patient records will decrease markedly when multi-CPU machines become standard.

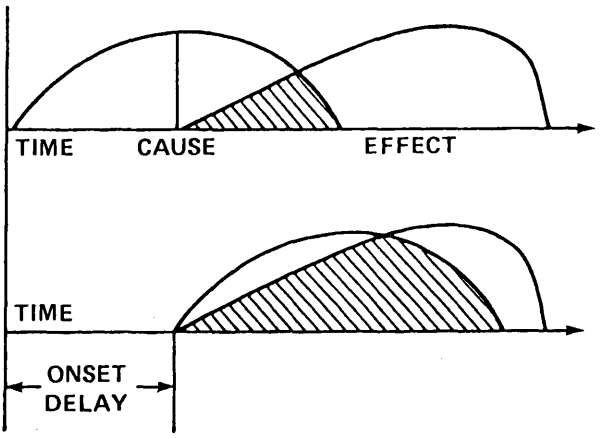

FIGURE 17-3 The principle underlying lagged correlation.

The basic algorithm uses a sliding nonparametric correlation performed on data from an individual patient's record. The principle underlying a lagged correlation is illustrated in Figure 17-3. Given a tentative cause, A, and an effect, B, the basic tool for uncovering a causal relationship is the Spearman correlation coefficient, $r_s(A, B, \tau)$, where τ is the time delay used in computing the correlation.

Selection of Patients for Correlation

In the discovery module only a sample of the patient records are analyzed. The sampling procedure uses a precomputed index called a *records list* associated with every variable in the data base. The records list is a sorted list of the form $((\text{patient}_1, n_1), (\text{patient}_2, n_2), \ldots (\text{patient}_m, n_m))$. The list identifies patients in descending order by their number of recorded values for the variable. That is, patient_1 has n_1 measurements of the variable, and so on.

The sample of records that are analyzed for a given pair of variables, $\{A, B\}$, is the sample $P^*_{\{A, B\}}$, where this is the set with the largest number of pairs of measurements of A and B. Let K denote the number of pairs in the set $P^*_{\{A, B\}}$. In experimental trials of the discovery module, K was set to 10.

The advantage of choosing the sample to be those patients with the most data on A and B is that "one might as well look where the looking is best." If a relationship exists between A and B, then it will be easiest to detect in patients with lots of data on A and B. This heuristic is particularly

valid for medical data when variables are more apt to be recorded when they are abnormal. Therefore, the frequency of observation tends to be correlated with the variance of the variable.

Correlations for the records in $P^*_{\{A, B\}}$ are computed as follows:

for each record in $P^*_{\{A,B\}}$ collect

[for each τ in T^* collect $r_s(A, B, \tau)$]

The *collect operator* denotes assembling a set composed of the value of each iterand. The time delays in T^* over which the correlations are performed are based on information from the knowledge base. That is, the algorithm makes use of prior information on the expected time delays of broad classes of causes and effects.

Combining Correlations Across Patients

That various correlations within and across patient records are based on different numbers of measurements poses a difficulty in combining them. Given equal correlations, we would like to assign more weight to records with more data. Using the p-value of the correlation achieves this and also facilitates combining correlations.

The p-values from the above procedure may be diagrammed as follows:

	τ_1	τ_2	. . .	τ_q
patient_1	$p_{1,1}$	$p_{1,2}$		$p_{1,q}$
patient_2	$p_{2,1}$	$p_{2,2}$		$p_{2,q}$
.	.	.		.
.	.	.		.
.	.	.		.
patient_K	$p_{K,1}$	$p_{K,2}$	. . .	$p_{K,q}$

Here p_{ij} denotes the p-value on the ith patient at the jth time delay. By the method of Fisher, the p-values may be combined to form an overall score s for each time delay τ_j:

$$s(A, B, \tau_j, P^*_{\{A, B\}}) = -2\Sigma\log(p_{i,\tau_j})$$

where the sum is over all patient records in $P^*_{\{A, B\}}$. It can be shown (Mood et al., 1974) that the scores s are distributed as χ^2 on $2p$ degrees of freedom. Since the distribution of the scores is known, their statistical significance may be calculated. Because of autocorrelation, the differences between scores determined at different time lags may not be distributed as χ^2. How-

ever, the significances are not taken literally by the discovery module, but are merely used to rank the hypotheses in terms of promise.

If the difference between the forward and backward sets of scores is large, a strong time precedence of association is implied. Since time precedence is not a sufficient condition for causality, spurious associations may also be reported as significant.

The output of the discovery module is a list of dyadic relations ranked in descending order by strength of unidirectionality of association. The algorithm has proven to be a sensitive, if nonspecific, detector of causal relationships, and is usually capable of accurately discriminating time precedence and determining approximate onset delay.

In the discovery module, only the properties of time precedence and covariation are used in a blind search for clues to causal relationships. Included in its output are many spurious relationships. The objective of the study module is to eliminate those relationships and to carefully examine those that remain in order to detail their characteristics and to store them in the KB.

17.8 The Study Module

The study module is the core of the RX algorithm. It takes as input a causal hypothesis obtained either from the discovery module or interactively from a researcher. It then generates a medically and statistically plausible model of the hypothesis, which it analyzes on appropriate data from the data base.

The study module is patterned after a sequence of steps usually undertaken by designers of large clinical studies. Its design may be considered an exercise in artificial intelligence insofar as it emulates human expertise in this area. There are at least six persons whose knowledge is brought to bear in designing, executing, reporting, and disseminating a large data base study. We may think of the *data base research team* as consisting of a doctor, a statistician, an archivist, a data analyst, a technical writer, and a medical librarian. The study module, in conjunction with the knowledge base (KB), emulates part of their expertise. The steps performed by the study module appear in Table 17-2.

17.8.1 Determination of the Feasibility of a Study

The study module may be operated automatically in batch mode, or it may be run interactively, enabling a researcher to modify the evolving study design. In this presentation we will assume that it is being run interactively.

TABLE 17-2 Steps Performed by the Study Module

1. Parse the hypothesis.
2. Determine the feasibility of the study on the data base.
3. Select confounding variables and causal dominators.
4. Select methods for controlling the causal dominators.
5. Determine proxy variables.
6. Determine eligibility criteria.
7. Create a statistical model.
   a. Select an overall study design.
   b. Select statistical methods.
   c. Format the appropriate data base access functions.
8. Run the study.
   a. Fetch the appropriate data from eligible patient records.
   b. Perform a statistical analysis of each patient's record.
   c. Combine the results across patients.
9. Interpret the results to determine significance.
10. Incorporate the results into the knowledge base.

Throughout this section we will use as an example the hypothesis that the steroid drug prednisone elevates serum cholesterol.

The first general task of the study module, or of the "data base research team," is to determine whether a particular study is feasible given the knowledge and the data available. The first step is the recognition by the program of the terms used in the hypothesis.

Suppose a researcher enters the hypothesis "prednisone elevates cholesterol." A top-down parser is applied to this input string. The pattern that matches is <variable relationship variable> where a variable may be any primary attribute or derived variable in the medical KB. As the parser matches the tokens in the input, it determines their classification in the KB.

```
Prednisone is a known concept.
It is classified as a Steroid which is a Drug which is an Action.

Elevates is a known concept.
It is classified as a Relationship.

Cholesterol is a known concept.
It is classified as a Chemistry which is a Lab-Value which is a State.
```

The classifications are simply determined by following the *generalization* pointers in the knowledge tree. The classification of each variable is not only of interest to the user but facilitates the inheritance mechanisms discussed above. For example, properties of the class steroids may be inherited by the drug prednisone, if they are needed in the course of the study.

To study the relationship between prednisone and cholesterol both variables must have been recorded in some patient records. Hence, the program next examines the intersection of their records lists.

The following list denotes that patient 78 had 32 recorded values for cholesterol, patient 118 had 25 values, and so on.

```
Cholesterol
-----------
records: ((P78  32)  (P118  25)  . . .  (P967  1))
```

17.8.2 Confounding Variables and Causal Dominators

The principal objective of the study module is the demonstration of *nonspuriousness*. In any observational drug study, as in the current one, the possibility must always be addressed that the effect of interest was caused by the disease for which the drug was given rather than by the drug itself. The first step in demonstrating nonspuriousness is identifying the set of possible confounding variables.

A confounding variable is any node, C, that may cause a clinically significant effect on both the causal node, A, and the effect node, B, in our hypothesis. The "clinical significance" of a given change in a variable is determined by a prior partitioning of that variable's range. Every real-valued object in the knowledge base has stored in its schema a *partition list* that divides its range into clinically significant regions.

Let C be the set of known confounders. The determination of C involves tracing the directed graph in the KB starting from A and B.

$$C = \text{Intersection}[\text{Antecedents}(A), \text{Antecedents}(B)]$$

where the list Antecedents(A) is the set of nodes that may produce a clinically significant effect on A. The *antecedents set* of a node is calculated by traversing the causal network in the KB. In the current example, the set C is determined to be {ketoacidosis, hepatitis, glomerulonephritis, nephrotic syndrome}.

Having determined the variables in C, the program displays the causal paths connecting them to A and B. The paths for glomerulonephritis appear below. The intensities of intermediate nodes are calculated using the regression coefficients stored in sequential causal relationships.

Glomerulonephritis {50 percent activity} is treated by Prednisone {30 mgms/day},

Glomerulonephritis can cause Nephrotic Syndrome {4 gms proteinuria/24 hrs} which is treated by Prednisone {20 mgms/day},

Glomerulonephritis can cause Nephrotic Syndrome {4 gms proteinuria/24 hrs} which increases Cholesterol {65 mgms/dl}.

17.8.3 Causal Dominators

To increase statistical power and stability of estimation it is usually desirable to control for as few confounding variables as possible. Since the set C in any real study is apt to be quite large, it is desirable to control for only the essentials. The set of *causal dominators*, C^*, is the smallest subset of C through which all known causal influences on both A and B flow.

The set of causal dominators, C^*, is determined by the following algorithm. The nodes in C are sorted into descending order according to their expected magnitude of effect on the relationship between A and B. More potent confounders appear earlier in the list. To determine C^*, the nodes in the ordered list are checked to determine whether paths to A and B still exist after earlier (more proximal) nodes have been blocked. In the current example, glomerulonephritis is deleted from the confounders since its confounding influence is entirely through nephrotic syndrome.

17.8.4 Controlling Variables Related to the Cause

Suppose prednisone affects cholesterol in some fashion; it is possible that related drugs may also affect cholesterol. We may also want to remove their influence by controlling them. Generally, we would like the program to suggest to us variables related to the cause, since they may also be confounders. These variables may not be in the set C, since causal paths between them and the effect may be unknown.

To select this set of variables related to the causal variable, the program uses the hierarchical structure of the KB. For example, since prednisone is one of the steroids, RX controls for the other steroids [i.e., *siblings* (prednisone) = {dexamethasone ACTH}, the other nodes in the same class, steroids].

17.8.5 Determination of Methods for Controlling Confounding Variables

Three general methods are used by RX to control confounding variables: (1) eliminating entire patient records, (2) eliminating time intervals containing confounding events, and (3) controlling statistically for the presence of the confounder. Eliminating patient records is always the safest and most intellectually reassuring. With statistical control, doubt always remains as to whether the confounder has been entirely eliminated. When eliminating time intervals, there is always the possibility that the confounding influence extends beyond the interval. On the other hand, eliminating patient records is the strategy most wasteful of data. There may be too few records left to analyze, or the generalizability of the result may be diminished.

To determine which method to use for each confounder, some decision criteria must be used. In making this decision and others discussed later, the study module uses decision criteria stored in the KB in the form of *production rules.*

17.8.6 Production Rules

Production rules have been widely used in artificial intelligence research to store domain knowledge (Shortliffe et al., 1975) (see also Chapter 5). A production rule is an IF/THEN rule consisting of a premise and conclusion.

The rule below is stored with other similar rules in the schema for control methods. To choose a control strategy, the rules are exhaustively invoked. Some rules may be used to resolve conflicts, if more than one control method is suggested.

IF the number of patients affected by a variable
is a small percentage of the number of
patients in the study,

AND the variable is present throughout those records,

THEN eliminate those records from the study.

The premise and conclusion of each production rule consists of a few lines of machine-readable code. In some systems (Shortliffe et al., 1975), the code may be mechanically translated into English upon request. To avoid the attendant complexity and to improve the quality of translation, the RX KB simply stores an English translation of each production rule.

In writing programs that use much domain knowledge, it is advantageous to separate the specific knowledge from the general algorithms that use it. Production rules are one method for achieving this modularity. The advantages are that (1) knowledge is more easily examined and updated, (2) dependencies among the knowledge are more easily discovered, and (3) the homogeneous format lends itself to machine translation.

17.8.7 Controlling Confounders

To determine how a particular confounder is to be controlled, the following information is first determined: N, the number of patient records in the study; *%records,* the fraction of records affected by the confounder; and *%visits,* the average fraction of visits affected. Each of these parameters is calculated using the information in the records list for each confounding variable.

If %records or %visits are low, then either records or time intervals may be eliminated. The rules tend to favor the elimination of records if N

is high. Only if N is low and %records or %visits is high is statistical control of the confounder considered.

While the program is running, the user may request a display of the rules that determined the choice of strategy. The user, as always, may override the decision made by the program.

In the prednisone/cholesterol study the program makes the following selections:

Dexamethasone	No control needed, since no values were recorded in the database
ACTH	No control needed
Nephrotic Syndrome	Control statistically using albumin as a proxy
Hepatitis	Eliminate affected time intervals
Ketoacidosis	Eliminate affected time intervals

17.8.8 Choice of Study Design and Statistical Method

Both the study design and the statistical method are selected using decision criteria stored in production rules in the KB. The choice of study design in the present system is simply a choice between a cross-sectional and a longitudinal design. In a cross-sectional design each variable is sampled once in a patient's record; in a longitudinal design variables are repeatedly sampled over time. The longitudinal study design has the advantage of making use of temporal information and multiple observations of variables within individual patient records. A cross-sectional design is only chosen when a longitudinal design is not feasible.

The selection of a particular statistical method uses knowledge encoded in a hierarchically organized, statistical knowledge base. The organization follows the conventional classification as in Armitage (1971) or Brown and Hollander (1977).

On the property list of each node in the tree is an *objectives*, a *prerequisites*, and an *assumptions* property. The objectives property describes the goals of the method. The prerequisites property describes the conditions that must hold for the method to be mechanically applied. The assumptions property describes the assumptions that must hold for the result to be valid.

An example of the schema for multiple regression appears below. The schema stores not only the English text but the equivalent machine-executable code.

```
Multiple-Regression
-------------------
objectives: linear-model
prerequisites:
    one dependent variable
    two or more independent variables
    measurement-level of dependent variable = real valued
```

measurement-level of independent variables = real valued
number of observations > 1 + number of independent variables
assumptions:
independent and identically distributed errors
normally distributed errors
linear and additive effects

To select a statistical method the objectives and prerequisites properties must satisfy the constraints of the study. The tree structure of the KB is used to prune limbs that are not applicable. When there is more than one applicable method, production rules at intermediate nodes arbitrate among methods. The present program does not determine whether the assumptions of a method have been fulfilled; they are merely displayed. However, it does make available tables and plots of residuals, so that the assumptions can be manually checked.

The present version of this *robot statistician* is rudimentary. Each of the nodes in the statistical KB contains about as much knowledge as is shown for multiple regression. No knowledge or methods are present for critically analyzing a fitted model or for revising the model. The current emphasis is simply on selecting a method that may be mechanically applied.

17.8.9 Formatting of Data Base Access Functions

In order to apply the selected analytical methods to the appropriate data, the data must be sampled from patient records at times that reflect the time delays inherent in the underlying processes. These time parameters are obtained by the study module from information in the KB.

For the longitudinal design in the present example the following model is created:

$$\Delta\text{cholesterol} = \beta_0 + \beta_1 \Delta\text{albumin} + \beta_2 \Delta\log(\text{prednisone})$$

where

$$\Delta\text{cholesterol} = \text{cholesterol}(t) - \text{cholesterol}(t_{\text{pchol}})$$

$$\Delta\text{albumin} = \text{albumin}(t - \tau_{\text{NS}}) - \text{albumin}(t_{\text{pchol}} - \tau_{\text{NS}})$$

$$\Delta\log(\text{prednisone}) = \log[\text{prednisone}(t - \tau_{\text{pred}})] - \log[\text{prednisone}(t_{\text{pchol}} - \tau_{\text{pred}})]$$

The time t_{pchol} denotes the time of the preceding measurement of cholesterol (previous to the present one), and τ_{NS} denotes the estimated delay from the start of nephrotic syndrome to the establishment of a steady state for cholesterol. The symbol τ_{pred} is the analogous onset delay for prednisone. No values are sampled during episodes of hepatitis or ketoacidosis. Figure 17-4 illustrates some of the time relationships that might be seen in one patient's record.

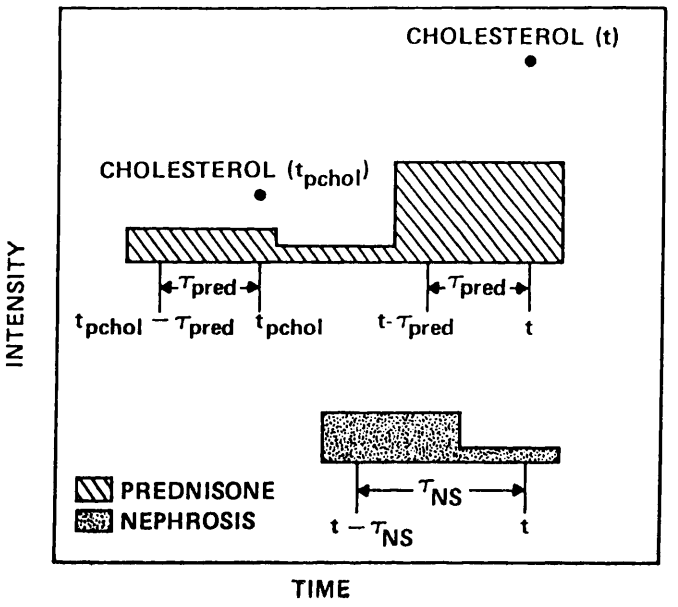

FIGURE 17-4 Time relationships in prednisone/cholesterol study.

Next, the mathematical model must be translated into the appropriate data base access functions. The function *create-access-functions* uses information in the schemata for the variables in the model to format the appropriate access functions. For example, the values for the onset delays and the indicator that there is a need for the log transform are retrieved from the schemata for nephrotic syndrome and prednisone. The estimated time delay for the effect of prednisone on cholesterol is obtained from the discovery module.

17.8.10 Determination of Eligibility Criteria

All patients in a data base may not be eligible for a particular study. Eligibility criteria in the current example are automatically formatted based on the number of relevant observations in a patient's record and the within-patient variance in the causal variable.

The study design cannot be executed on patient records in which there are less than four sets of observations (note that there is 1 degree of freedom for the mean plus 2 degrees of freedom for Δalbumin and for Δprednisone). Furthermore, patient records are excluded in which the coefficient of variation in log(prednisone) is below threshold.

17.8.11 Statistical Analysis: Fitting the Model

Until July 1980, all statistical analyses were performed using SPSS (Nie et al., 1975) as a subroutine; however, this incurred the inefficiency of having to write and read files in formats intended for human usage. Currently, all statistical analysis is performed using IDL (Kaplan et al., 1978). Written in Interlisp, IDL makes available fast numerical computation, matrix manipulation, and a variety of primitive operators for statistical computation.

Most of our studies are sufficiently large that statistical analysis requires use of a separate core image (separate job). The study module writes the study design to disk, then calls IDL. IDL reads the study design, executes it, writes the results to disk, and then calls the study module.

Longitudinal Design Using Weighted Multiple Regressions

The method of analysis that we have most extensively developed combines the results of separate multiple regression analyses performed on individual patients. Recall that individual patient records differ in quantity of data and greatly vary on covariates. By analyzing each patient's record separately, we can determine the distribution of an effect across patients and obtain information as to why some patients exhibit an effect and others do not.

Naturally, we are interested in knowing whether a given causal relationship is statistically significant in the study sample as a whole. The analysis of significance is complicated by the fact that patients have widely varying amounts of data. Intuitively, one would like to weight most heavily those patients in whom a relationship has been most precisely determined, i.e., the patients with the most data; however, these patients may be unrepresentative.

The approach we use is a mixed model. The regression coefficient for each patient is weighted by the inverse of its variance. The mathematical justification for this procedure lies beyond the scope of this paper but may be found in Blum (1982). When there is a large variation in the effect across patients, perfect precision on any one patient is of little advantage, and all patients are weighted nearly equally. When across-patient variation is small, weighting by precision is more appropriate, and the weights diverge.

17.8.12 Interpretation of Results

The final result of the longitudinal design is an estimate of β, the unstandardized regression coefficient of the effect on the cause, and $\text{var}(\beta)$, its variance. The ratio $\beta/[\text{var}(\beta)]^{5}$ is approximately distributed as a t statistic

TABLE 17-3 Distribution of the Prednisone/ Cholesterol Effect Across Patients (given a baseline value of 230 mg/dl and a change in prednisone from 0 to 30 mg/day)

Range of cholesterol		*Percentage of patients*	*Magnitude of change*
100	150	0	extreme −
150	195	0	strong −
195	210	0	moderate −
210	225	0	weak −
225	230	0	equivocal −
230	235	0	equivocal +
235	250	0	weak +
250	280	10	moderate +
280	360	82	strong +
360	700	8	extreme +

on $n - 1$ degrees of freedom, where n is the number of patients in the study. A two-sided p-value is calculated using the t statistic.

Presently, the interpretation of the results of a study depend only on the magnitude of β and its corresponding p-value. A significant p-value does not necessarily mean the result is medically significant; a p-value can always be made significant if the number of patients is large enough. The program for interpretation uses the following heuristic: if β is large, then for a given p-value, the program assigns a higher validity to the result than it does if β is small.

The clinical significance of β is determined by the magnitude of its expected influence on the effect variable in the study. This is illustrated in Table 17-3, which shows the expected distribution of cholesterol given prednisone at 30 mgms per day.

Recall that the *validity score* is a component of every causal relationship stored in the KB. The validity score is measured on a scale from 1 to 10 summarizing the state of proof of a relationship. The highest score that a study based on a single nonrandomized data base can achieve is 6. Higher scores can only be obtained from replicated studies, the highest scores requiring experimental manipulation and a known mechanism of action. A score of 6 means "strong correlation and time relationship have been demonstrated after known covariates have been controlled in a single data base study."

The discovery module populates the KB with causal links of validity between 1 and 3. The study module overwrites the links that it explores, assigning to those that it confirms scores between 4 and 6.

A statistician or researcher might choose to pursue a given study further, asking "Have the confounding variables in C^* been adequately con-

TABLE 17-4 Effects of Prednisone

	Direction	*Onset delay*	*p-value*
Weight	+	chronic	< .0001
Cholesterol	+	acute	.0001
WBC	+	acute	.0004
% Neutrophils	+	acute	.003
% Lymphs	−	acute	.003
BP-diastolic	+	acute	.004
Glucose	+	acute	.007
Hemoglobin	+	chronic	.009
Wintrobe ESR	−	chronic	.01
Platelets	+	acute	.02
Temperature	−	chronic	.05
anti-DNA	−	chronic	.08
% Eosinophils	−	acute	.15
Urine-RBC's	−	chronic	.17
Creatinine	−	chronic	.19

trolled?" "Are the residuals in each of the regressions independent and identically distributed?" "What accounts for the differences among patients?" A researcher can pursue these questions interactively in RX, incrementally improving the mathematical model (Draper, 1966); however, the automation of this kind of inquiry will require building much greater knowledge into the robot statistician.

17.9 Medical Results

The medical results reported here were generated by running the discovery module and then the study module on a sample data base containing the records of 50 patients with systemic lupus erythematosus (SLE). Many patients had multisystem involvement including glomerulonephritis and nephrotic syndrome.

Table 17-4 shows the effects that were confirmed by the study module for the steroid drug prednisone. The study module automatically incorporated these new links and details of the studies into the knowledge base in the format discussed above.

The effects that were confirmed by the study module for the steroid drug prednisone are shown in Table 17-4. To illustrate the interpretation of Table 17-4, the second row of the table means that prednisone is thought to cause an increase (+) in cholesterol, that the time delay is "acute" (less than one average intervisit interval), and that the effect is highly statistically

significant ($p = .0001$). The study module automatically incorporated these new links and details into the knowledge base in the format discussed above.

Almost all of the acute effects appearing in the table have been extensively confirmed in the medical literature. The effect of prednisone on cholesterol, strongly supported by this study, has only been reported a few times previously. No previous study has recorded the reproducibility of the effect over time or the interpatient variability, as was done here.

The chronic effects of prednisone shown in Table 17-4 are those appearing in a setting of severe SLE. Literature confirmation of these effects has been scant. Because of small numbers of patients, the chronic effects shown here must be further studied. Tables of other empirical results and a discussion of the statistical models used in these studies may be found in Blum (1982).

17.10 Summary

The methods described here emanate from a small set of operational properties of causal relationships. The discovery module uses a nonparametric method for producing a ranked list of causal hypotheses based on strength of time precedence and association. The study module uses a consensual causal model stored in a knowledge base to determine all known confounding variables and to determine appropriate methods of adjusting for them. The statistical model of the tentative causal relationship is then applied to a set of data. If the results indicate that a relationship is significant after controlling for confounding influences, then a new relationship is incorporated into the KB. Subsequent studies may make use of this new link.

All components of the study module can be used in an interactive mode to give a researcher more control in determining the course of the study. For example, the causal model stored in the KB can be queried interactively or changed in the course of a study as new information becomes available. All phases of the statistical analysis can also be interactively modified.

Any methodology that draws causal inferences based on nonrandomized data is subject to an important limitation: *unknown covariates cannot be controlled.* The strength of the knowledge base lies in its comprehensiveness, but even so, it cannot guarantee nonspuriousness. Any single study, particularly one using nonrandomized data, must be viewed skeptically. For this reason, the most conclusive causal relationships that RX discovers are always assigned a modest validity. Only through repeated studies, particularly through experimental manipulation of the causal variable, can a given result become more definitive.

ACKNOWLEDGMENTS

I am grateful to Guy Kraines, Kent Bailey, and Byron William Brown for their assistance with the statistical models; to Gio Wiederhold for project administration and guidance; to Beau Shiel and Ronald Kaplan for their assistance with IDL; and to James Fries, Alison Harlow, and James Standish for assistance in obtaining clinical data.

Funding for this research was provided by the National Center for Health Services Research through grant HS-03650, by the National Library of Medicine through grant LM-03370, and by the Pharmaceutical Manufacturers Association Foundation. Computation facilities were provided by SUMEX-AIM through NIH grant RR-00785 from the Biotechnology Resources Program. Clinical data were obtained from the American Rheumatism Association Medical Information System. The project is continuing under the sponsorship of NCHSR through grant HS-04389.