COMPUTERS AND BIOMEDICAL RESEARCH 15, 164–187 (1982)

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

ROBERT L. BLUM

Department of Computer Science, Margaret Jacks Hall, Stanford University, Stanford, California 94305

Received July 31, 1981

The objectives of the methods and computer implementation presented here are (1) to automate the process of hypothesis generation and exploratory analysis of data in large nonrandomized, time-oriented clinical data bases, (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. The RX computer program consists of a knowledge base (KB), a discovery module, a study module, and a clinical data base. Utilizing techniques from the field of artificial intelligence, the KB contains medical and statistical knowledge hierarchically organized, and is used to assist in the discovery and study of new hypotheses. Confirmed results from the data base are automatically encoded into the KB. The discovery module uses lagged, nonparametric correlations to generate hypotheses. 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 studies have used a longitudinal design involving a multiple regression model applied to individual patient records. Data for system development were obtained from the American Rheumatism Association Medical Information System.

1. INFORMATION

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 this data for clinical management of individual patients. Hospital administrators require it for billing and resource allocation. Government agencies require data for quality of health care assessments. Third party insurers require it 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 6

0010-4809/82/020164-24$02.00/0

Reprinted from Computers and Biomedical Research
Volume 15, Number 2, April 1982
Academic Press, Inc.

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. To illustrate this method consider the following hypothetical scenario.

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 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 Fig. 1. At

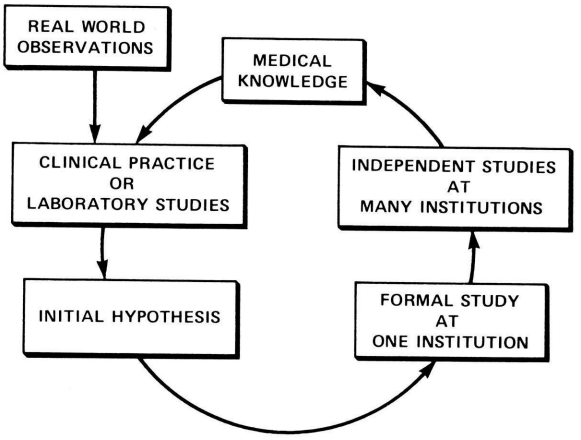

FIG. 1. The evolution of medical knowledge.

each stage of discovery and confirmation existing medical knowledge is used to design and to interpret the studies.

3. THE RX PROJECT

It is easy to imagine automating at least parts of the above discovery and confirmation cycle. 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 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 (Fig. 2).

- The *discovery module* produces hypotheses "*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

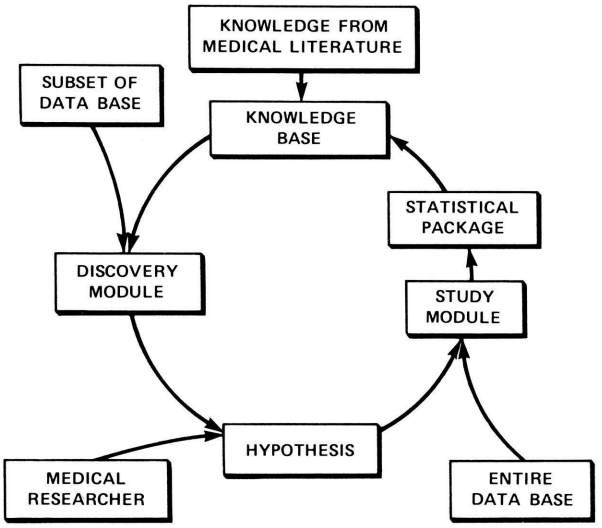

FIG. 2. Discovery and confirmation in RX.

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 paper 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 Ref. (*1*).

4. TIME-ORIENTED DATA BASES

The general format of a patient record is illustrated in Table 1. Each time a patient is seen in clinic a number of observations are made. These are recorded

TABLE 1

HYPOTHETICAL TIME-ORIENTED RECORD FOR ONE PATIENT

Visit number	1	2	3
Date	January 17, 79	June 23, 79	July 1, 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

with the date of observation in the data base. The recorded 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 Data Base) collected at Stanford University from 1969 to the present (*2, 3*). The subset contains the records of 50 patients with severe systemic lupus erythematosus (SLE). The average number of clinic visits for each patient was also 50, and the average length of follow-up was 5 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 Ref. (*4*).

5. COMPUTER FACILITIES AND LANGUAGES

Research was performed at two computer facilities at Stanford University: SUMEX/AIM and SCORE. SUMEX/AIM features a DEC dual processor KI-10 running the TENEX operating system. SCORE has a DEC 20/60 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 (*5*), a dialect of LISP, a language that is highly suitable for knowledge manipulation. Statistics are performed in IDL (Interactive Data-Analysis Language) (*6*), discussed later. The RX source code with knowledge base comprises approximately 200 disk pages of 512 words each.

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, non-protocol 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 (*7–9*). 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 RXs 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 system schemata. 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.

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, diseases* and into *drugs, surgery,* and *physical therapy.* The categories of diseases and other entities follow the conventional nosology based upon organ systems and/or pathology found in any standard textbook of medicine (*10*). I shall 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 are *derived variables,* that is, objects whose values must be derived from primary data. The properties in the schema of the object 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.*

6.1.1. 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 (*11*). 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: 0.1 {i.e., values are rounded to the nearest 0.1}
units: grams per deciliter

6.1.2. 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 below.

Inheritance mechanisms (*12*) 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.

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	

6.1.3. *Properties Pertaining to the Definition and Intrinsic Characteristics of an Object*

If an object is a primary data base attribute such as 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 definition of the object. The didactic example below shows a definition for pneumonia.

Pneumonia

definition:		Temperature >38°C
	and	WBC > 10,000 cells/mm^3
	and	Chest X-ray = 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, carryover, onset-delay,* and so on. These parameters are used by the time-dependent predicates when definitions for objects are evaluated.

6.1.4. *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 below.

⟨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 IDs, 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 the details of a particular causal relationship: its mathematical form, the evidence supporting it, or its distribution across patients. In the example below 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 mg/day {modal effects}

---

usually increases CHOLESTEROL by 50 to 130 mg/dl,
usually increases WEIGHT by 3 to 7 kg,
regularly attenuates NEPHROTIC SYNDROME by 1 to 2 g protein/24 hr,
regularly attenuates GLOMERULONEPHRITIS by 10 to 30% activity,
regularly decreases EOSINOPHILS by 2 to 3% of WBC,
commonly decreases ANTI-DNA by 50 to 90% activity,
occasionally increases GLUCOSE by 20 to 100 mg/dl.

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.

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.