INDUCTION OF CAUSAL RELATIONSHIPS FROM A TIME-ORIENTED CLINICAL DATABASE: AN OVERVIEW OF THE RX PROJECT

Robert L. Blum

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

ABSTRACT

The RX computer program examines a time-oriented clinical database and attempts to derive a set of (possibly) causal relationships. First, a Discovery Module uses lagged, nonparametric correlations to generate an ordered list of tentative relationships. Second, a Study Module uses a small knowledge base (KB) of medicine and statistics to create a study design to control for known confounders. The study design is then executed by an on-line statistical package, and the results are automatically incorporated into the KB as a machine-readable record. In determining the confounders of a new hypothesis the Study Module uses previously "learned" causal relationships.

INTRODUCTION

One of the most important reasons for accumulating patient data on computers is the possibility of deriving medical knowledge from the stored observations. The long range objectives of the RX Project are 1) to increase the validity of medical knowledge derived from large time-oriented databases containing routine, non-randomized clinical data, 2) to provide knowledgeable assistance to a research investigator in studying medical hypotheses on large databases, and 3) to fully automate the process of hypothesis generation and exploratory confirmation. The objective of this paper is to provide an introduction to the RX Project. The project is described in full detail in [1] and in summary form in [2]. Methods for storing and displaying the clinical data are described in [3]. A discussion of the difficulties in drawing inferences from clinical databases appears in [4].

Designed to emulate standard methods of epidemiological research, the RX computer program is a prototype system for automating the discovery, confirmation, and incorporation of knowledge from large clinical databases. While the program is a research prototype, parts of it have been operational since 1979, and have been demonstrated at national conferences.

A medical researcher enters a causal hypothesis of interest into the RX Study Module, for example, "does aspirin decrease blood hemoglobin?" The Study Module uses a small on-line knowledge base (KB) of medicine and statistics to produce a study design of this hypothesis. In doing this, the Study Module also uses pre-computed information on the amount of data on each variable stored in the database. This study design is then executed by a statistical package using the appropriate data from the database. The results of this study are then automatically encoded into the on-line medical knowledge base in a machine-readable form. The KB also contains knowledge that was entered directly into it by clinicians. Both kinds of knowledge are accessible to the Study Module while it is designing a study.

Now, instead of obtaining the initial hypothesis from a medical researcher, it is easy to imagine deriving it empirically from the database. Following this concept, a prototype Discovery Module was added to the RX Project in 1980. The Discovery Module combs through a subset of the patient database to derive an ordered list of hypotheses for exploration. These hypotheses are studied by the Study Module as though they had been entered by a researcher.

RX consists of five major parts: the Database, the Discovery Module, the knowledge base, the Study Module, and a statistical analysis package. A brief description of each follows.

THE TIME-ORIENTED DATABASE

The database we use is the ARAMIS database, the American Rheumatism Association Medical Information System, developed at Stanford University and implemented on TOD, a Time-Oriented Database System [5] [6] [7]. A recent review of clinical databases appears in [8]. Our research, so far, has been done entirely on a subset of the ARAMIS Database collected at the Stanford University Division of Immunology Clinics and containing the records of fifty patients with systemic lupus erythematosus.

Each patient's record consists of a matrix of values for a set of attributes that may be recorded each time the patient is seen in the clinic. Values for several hundred attributes can be recorded in ARAMIS. The attributes include signs, symptoms, lab tests, therapies, and indices of patient functional status. In general, the time intervals between clinic visits are not uniform, and patients are not on treatment protocols.

TOD is implemented in PL/1; ARAMIS is stored on an IBM 370/3081 computer at the Stanford University Center for Information Technology. On the other hand, the RX Project is implemented at two other 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 DEC20/60 running TOPS-20. Data transfer from ARAMIS is done by magnetic tape.

All RX computer programs are written in INTERLISP, a dialect of LISP, a language highly suited for knowledge manipulation. The RX source code with knowledge base comprises approximately 200 disk pages of 512 words of 36 bits each.

THE DISCOVERY MODULE

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". The current Discovery Module uses lagged nonparametric correlations across variables but within individual patient records. The p-values of the correlations across patients are then combined to yield a score that is used to order the list of hypotheses. Knowledge from the medical KB is used to determine the range of time lags examined.

THE KNOWLEDGE BASE

The leitmotif of the RX Project is that derivation of new knowledge from databases can best be performed by integrating existing knowledge of relevant parts of medicine and statistics into the medical information system. In the RX computer program the medical KB 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 KB grows by automatically incorporating new knowledge into itself. Hence, it is designed in such a way that relationships derived from the database are translated into the same machine-readable form as knowledge entered from the medical literature by clinicians.

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 system schemata. The medical knowledge in the RX KB covers only a small portion of what is known about systemic lupus erythematosus and limited areas of general internal medicine. The present KB is merely a test vehicle; its size is 50 disk pages or 120,000 bytes.

The most important class of properties in the schema corresponding to each medical object is that specifying the causal relationships of an object to other objects. Causal relationships are stored between objects using an "effects" and an "affected-by" property list for each object. The resulting causal model is a directed cyclic graph; that is, the representation allows for the possibility that A causes B causes A with appropriate time lags.

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 >

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.

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 database.

In creating a study design the Study Module follows accepted principles of epidemiological research. It determines study feasibility and study design: cross-sectional versus longitudinal. It uses the KB to determine the confounders of a given hypothesis, and it selects methods for controlling their influence: elimination of patient records, elimination of confounding time intervals, or statistical control. The Study Module then determines an appropriate statistical method, using knowledge stored as production rules. Most studies have used a longitudinal design involving a multiple regression model applied to individual patient records. Results across patients are combined using weights based on the precision of the estimated regression coefficient for each patient. The steps in the Study Module appear below.

1) Parse the hypothesis and determine the classification of variables in it.
2) Determine the feasibility of the study on the database.
3) Select confounding variables and causal dominators using the KB.
4) Select methods for controlling the confounding variables.
5) Determine proxy variables.
6) Determine eligibility criteria.
7) Create a statistical model of the hypothesis using knowledge from the KB.
   a) Select an overall study design.
   b) Select statistical methods.
   c) Format the appropriate database 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 medical and statistical significance.
10) Incorporate the results into the knowledge base.

THE STATISTICAL PACKAGE: IDL

Until July 1980, all statistical analyses were performed using SPSS as a subroutine. Currently all statistical analysis is done using IDL [9]. Written in INTERLISP, IDL makes available fast numerical computation, matrix manipulation, and a variety of high-level primitives for statistical computation. To the basic IDL package we have added fifty disk pages of other statistical routines. The Study Module writes the study design to disk, then calls IDL. IDL reads the study design, executes it, writes the results to disk, then calls the Study Module.

The method of analysis we have used most often involves performing a separate multiple regression on each patient record, then combining results across patients. Our method of analysis accounts for autocorrelation and for differing quantities of data across patients.

RESULTS, AVAILABILITY, AND LIMITATIONS

The current RX system was applied to a sample database containing the longitudinal records of 50 patients with systemic lupus erythemotosus followed for an average of 50 clinic visits. Several well-known effects of the steroid drug prednisone were confirmed by the Study Module. The Study Module automatically incorporated these new links and details of the studies into the KB in the format previously discussed.

The RX computer program is currently only a research prototype. It is not available outside our lab except for program development.

We must emphasize that any methodology that draws causal inferences based on nonrandomized data is subject to an important limitation: unknown covariates cannot be controlled. The strength of a particular knowledge base lies in its comprehensiveness, but even so, it cannot guarantee nonspuriousness. Only through repeated studies, particularly through experimental manipulation of the causal variable, can a given result become more definitive.

ACKNOWLEDGEMENTS

I am grateful to Guy Kraines, Kent Bailey, and Byron William Brown for their assistance with the statistical models, to Gio Wiederhold for project management and conceptual guidance, to Ronald Kaplan and Beau Shiel for their assistance with IDL, and to James Fries, Dennis McShane, Alison Harlow, and James Standish for kindly providing access to the database.

Funding for this research was provided by the National Center for Health Services Research through grants HS-03650 and HS-4389, by the National Library of Medicine through grant LM-03370, and by the Pharmaceutical Manufactureres 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, supported by grants AM-21393 and HS-03802.

REFERENCES

[1] Blum, Robert L.; Discovery and Representation of Causal Relationships from a Large Time-Oriented Clincial Database: The RX Project, Ph.D. Thesis, Stanford University, January, 1982.

[2] Blum, Robert L.: Discovery, Confirmation, and Incorporation of Causal Relationships from a Large Time-Oriented Clinical Database: The RX Project. Computers and Biomedical Research, 15:2, 164-187, 1982.

[3] Blum, Robert L.: Displaying Clinical Data from a Time-Oriented Database. Computers in Biology and Medicine, 11:4, 197-210, 1981.

[4] Blum, Robert L. and Wiederhold, Gio: Inferring Knowledge from Clinical Data Banks Utilizing Techniques from Artificial Intelligence. "Proc. 2nd Annual Symp. on Comp. Applic. in Med. Care", 303-307, IEEE, Washington, D.C., November 4-9, 1978.

[5] Weyl, Stephen; Fries, J.; Wiederhold, G.; Germano, F.: A Modular Self-Describing Clinical Databank System. Computer and Biomedical Research 8:3, 279-293, June, 1975.

[6] Wiederhold, Gio; Fries, James F.: Structured Organization of Clinical Data Bases. "AFIPS Conference Proceedings" 44: 479-485, 1975.

[7] Wiederhold, Gio: Database Design, McGraw-Hill, 1977.

[8] Wiederhold, Gio: Databases for Health Care, Springer-Verlag, 1981.

[9] Kaplan, Ronald M., et al.: The Interactive Data-analysis Language Reference Manual. Xerox Palo Alto Research Corp., 1978.