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 (*13, 14*).

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.

7.2. Methodology of the Discovery Module

The function of the *discovery module* is to find candidate causal relationships. The discovery module exploits only the first two properties of causal relationships to do this: 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 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.

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 Fig. 3. Given a tentative cause A and effect B, the basic tool for uncovering a casual relationship is the Spearman correlation coefficient $r_s(A, B, \tau)$, where τ is the time delay used in computing the correlation.

7.2.1. 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

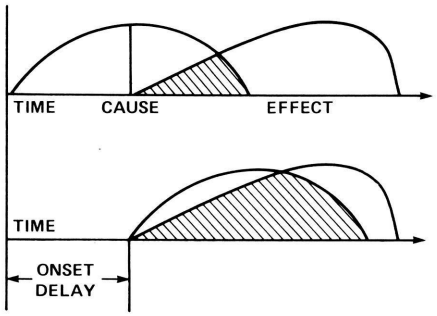

FIG. 3. The principle underlying lagged correlation.

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 in 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

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.

7.2.2. Combining Correlations across Patients

That various correlations within and across patient records are based upon 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

	τ_1	τ_2	$\cdots$	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 $p_{i,j}$ 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 (*15*) 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 χ^2. However, 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.

8. The Study Module

The study module is the core of the RX algorithm. It takes as input a causal hypothesis gotten 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 database study. We may think of the *data-base research team* as consisting of a doctor, a statistician, an

TABLE 2

STEPS PERFORMED BY THE STUDY MODULE

1. Parse the hypothesis.
2. Determine the feasibility of the study on the database.
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 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 significance.
10. Incorporate the results into the knowledge base.

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 in the study module appear in Table 2.

8.1. Determination of Feasibility of 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 shall assume that it is being run interactively. Throughout this section we shall 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 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.

Cholesterol

---

records: ((P78 32)(P118 25) · · · (P967 1))

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

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 in 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.

$$C = \text{intersection[antecedents}(A), \text{ antecedents}(B)]$$

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% activity} is treated by Prednisone {30 mg/day},
Glomerulonephritis can cause Nephrotic Syndrome {4 g proteinuria/24 hr}
which is treated by Prednisone {20 mg/day},
Glomerulonephritis can cause Nephrotic Syndrome {4 g proteinuria/24 hr}
which increases Cholesterol {65 mg/dl}.

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 in the present computer program by the following algorithm. Assume we are interested in determining whether A causes B. Let us designate by P the set of proximate causes of B; that is, P is simply the set of nodes on B's *affected-by* list. We first check to determine whether any nodes in P can reach A, i.e., may also causally influence A, however indirectly. Any of those nodes in P that can reach A are appended to the set C^* of causal dominators, which is initially empty. Call this set of nodes P_1. Then the nodes in P_1 are blocked by placing flags on them. This prevents flow through them on subsequent iterations. Next, consider the set of nodes $P_2 = P - P_1$, and generate the set of all proximate causes of the nodes in P_2. Call this set Q. If we now assign $P = Q$ and iterate the above sequence, the set of causal dominators C^* is generated. The algorithm is admittedly inefficient, but adequate for the size of networks with which we have dealt. In the current example, *glomerulonephritis* is deleted from the confounders since its confounding influence is entirely through *nephrotic syndrome*.

8.4. Controlling Other Variables

8.4.1. 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. These are the *siblings*(prednisone) = {dexamethasone ACTH}: nodes in the same class, *steroids*.

8.5. Determination of Methods for Controlling Confounding Variables

Three general methods are used by RX to control confounding variables: (1) eliminate entire patient records, (2) eliminate time intervals containing confounding events, and (3) control statistically for the presence of the confounder. Eliminating patient records is always the safest and most intellectually reassuring. With stastical control, doubt always remains as to

whether the confounder has been entirely eliminated. When eliminating time intervals, the possibility that the confounding influence extends beyond the interval is always possible. 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*.

8.6. Production Rules

Production rules have been widely used in artificial intelligence research to store domain knowledge (*16, 17*). 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 consist of a few lines of machine-readable code. In some systems (*17*), 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.

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 data base
ACTH	No control needed
Nephrotic Syndrome	Control statistically using *albumin* as a proxy
Hepatitis	Eliminate affected time intervals
Ketoacidosis	Eliminate affected time intervals

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 versus 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 Ref. (*18*) or (*19*).

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.

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

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

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. It does make available tables and plots of residuals, however, 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 in selecting a method that may be mechanically applied.

8.9. Composition 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

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

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

and

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

The time t_{pchol} denotes the time of measurement of the 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. Some of the time relationships that might be seen in one patient's record are illustrated in Fig. 4.

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 need for the log transform are retrieved from the schemata for nephrotic

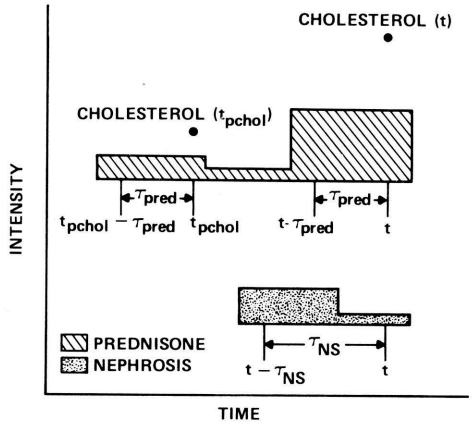

FIG. 4. Time relationships in prednisone: cholesterol study.

syndrome and prednisone. The estimated time delay for the effect of prednisone on cholesterol is obtained from the discovery module.

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 upon 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 (= 1 degree of freedom for the mean + 2 df for Δalbumin and for Δprednisone). Furthermore, patient records are excluded in which the coefficient of variation in log prednisone is below threshold.

8.11. Statistical Analysis: Fitting the Model

Until July 1980, all statistical analyses were performed using SPSS (*20*) 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 (*6*). 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, then calls the study module.

8.11.1 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 patient's 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, e.g., 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 Ref. (*1*). 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.

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 var(β), its variance. The ratio $\beta/[\text{var}(\beta)]^{.5}$ is approximately distributed as a t statistic 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 magnitudes of β and its corresponding p value. A significant p value does not necessarily mean the result is medically significant. Even an inconsequentially small change in the effect will become significant at a given p value, 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, it assigns a higher validity to the result than 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 3, which shows the expected distribution of cholesterol given prednisone at 30 mg/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 be obtained only from replicated studies, the highest scores requiring experimental manipulation and known mechanism of action. A score of 6 means that "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

TABLE 3

DISTRIBUTION OF THE PREDNISONE/CHOLESTEROL EFFECT ACROSS PATIENTS[a]

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 +

[a] Distribution across patients of cholesterol (mg/dl), given a baseline value of 230 mg/dl and given a change in prednisone from 0 to 30 mg/day.

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 controlled?" "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 (*21*); however, the automation of this kind of inquiry will require building much greater knowledge into the "robot statistician."

9. MEDICAL RESULTS

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

The effects that were confirmed by the study module for the steroid drug prednisone are shown in Table 4. To illustrate the interpretation of Table 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 of the studies into the knowledge base in the format discussed above.

TABLE 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–RBCs	−	Chronic	.17
Creatinine	−	Chronic	.19

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 been reported only 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 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 Ref. (*1*).

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 enable 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 National Institutes of Health 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 Grant HS-04389.

BIBLIOGRAPHY

1. BLUM, R. L. Discovery and representation of causal relationships from a large time-oriented clinical database: The RX project. Ph.D. thesis, Stanford Univ., 1982 in Computer Science and Biostatistics.
2. FRIES, J. F. Time-oriented patient records and a computer databank. *J. Amer. Med. Assoc.* **222,** 1536 (1972).
3. WIEDERHOLD, G., AND FRIES, J. F. Structured organization of clinical data bases. AFIPS Conference Proceedings, AFIPS, 1975, pp. 479–485.
4. BLUM, R. L. Displaying clinical data from a time-oriented database. *Comput. Biol. Med.* **11**(4) (1981).
5. TEITELMAN. "INTERLISP Reference Manual." Xerox Palo Alto Research Corp., Palo Alto, Calif., 1978.
6. KAPLAN, R. M., SHEIL, B. A., AND SMITH, E. R. "The Interactive Data-analysis Language Reference Manual." Xerox Palo Research Corp., Palo Alto, Calif., 1978.
7. BLUM, R. L., AND WIEDERHOLD, G. Inferring knowledge from clinical data banks utilizing techniques from artificial intelligence. Proceedings, 2nd Annual Symposium on Computer Applications in Medical Care, Washington, D.C., IEEE, November, 1978, pp. 303–307.
8. BYAR, D. P. Why data bases should not replace randomized clinical trials. *Biometrics* **36,** 337–342 (1980).
9. DAMBROSIA, J. M., AND ELLENBERG, J. H. Statistical considerations for a medical data base. *Biometrics* **36,** 323–332 (1980).
10. ISSELBACHER, K. J., *et al.*, "Harrison's Principles of Internal Medicine." McGraw–Hill, New York, 1980.

11. WIEDERHOLD, G. "Database Design." McGraw–Hill, New York, 1977.
12. STEFIK, M. J. An examination of a frame-structured representation system. Proceedings, Sixth International Joint Conference on Artificial Intelligence, IJCAI, 1979, pp. 845–852.
13. KENNY, D. "Correlation and Causality." Wiley, New York, 1979.
14. SUPPES, P. "A Probabilistic Theory of Causality." North-Holland, Amsterdam, 1970.
15. MOOD, A. M., GRAYBILL, F. A., AND BOES, D. C. "Introduction to the Theory of Statistics." McGraw–Hill, New York, 1974.
16. DAVIS, R. B., BUCHANAN, B. G., AND SHORTLIFFE, E. H. Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence* **8,** 15–45 (1977).
17. SHORTLIFFE, E. H., DAVIS, R., AXLINE, S., BUCHANAN, B., GREEN, C., AND COHEN, S. Computer-based consultation in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Comput. Biomed. Res.* **8,** 303–320 (1975).
18. ARMITAGE, P. "Statistical Methods in Medical Research." Blackwell, Oxford, 1971.
19. BROWN, B. W., AND HOLLANDER, M. "Statistics: A Biomedical Introduction." Wiley, New York, 1977.
20. NIE, N. H., HULL, C. H., JENKINS, J. G., STEINBRENNER, K., AND BENT, D. H. "SPSS: Statistical Package for the Social Sciences," McGraw–Hill, New York, 1975.
21. DRAPER, N. R., AND SMITH H. "Applied Regression Analysis," Wiley, New York, 1966.