Use of expert systems in medical research
data analysis: The POSCH AI project

by JOHN M. LONG, JAMES R. SLAGLE, MICHAEL WICK, ERACH IRANI,
JOHN MATTS and the POSCH GROUP*

University of Minnesota
Minneapolis, Minnesota

ABSTRACT

Artificial Intelligence (AI), especially expert systems technology, has very promising possibilities for more fully automating the analysis of medical research data. The Program on the Surgical Control of the Hyperlipidemias (POSCH), a national multi-centered clinical trial has been experimenting with using AI in its data analyses for several years. Three projects are described in this paper.

The first experiment, designed to automate the clinical judgment used to evaluate the data from serial graded electrocardiograms has been a success. Early efforts to automate one step of the evaluation of serial coronary arteriography data has also been successful, but major difficulties must be overcome to extend the work.

The initial motivation for the POSCH AI Project was to build an expert system that can search for interesting relationships among, and between, the POSCH variables. The problems encountered are different from those of other investigators who are attempting similar projects. Efforts to segment and/or collapse the problem into smaller units are explained. Issues related to the data structure and query languages have been identified. Statistical and other logically based reasoning methods, fundamental to the entire project, are discussed in conjunction with the heuristic methods used by expert systems.

* The work of J.R. Slagel, M.R. Wick, and Erach Irani is partially supported by NSF Grant DCR8512857 and by the Microelectronics and Information Sciences Center of the University of Minnesota.
INTRODUCTION

Artificial Intelligence (AI), primarily expert systems technology, is being widely explored in medicine. There are several large, multi-centered clinical studies that accumulate extensive files of data on a large number of patients. AI has promising possibilities for them. As an example, the Program on the Surgical Control of the Hyperlipidemias (POSCH) has been experimenting with using expert systems to automate more fully the analyses of its large database of clinical trial research data. In our first two projects, we are using expert systems to automate the element of clinical judgement required for some analyses. Ultimately we hope to use AI to automate the search for interesting relationships between, and among, the variables that are collected annually on POSCH patients.

The POSCH database comes from a randomized controlled clinical trial that uses a standard patient visit protocol and records all data on standard data forms. The timing and frequency of patient visits are also standard. POSCH has some missing data and a few missed visits, but the frequency is quite low. This allows us to focus our efforts on the interesting aspects of building a system to search for medical knowledge that may now be hidden in the database. In other words, the POSCH database is ideal in many ways for experimentation in the use of AI for data analysis.

POSCH AI PROJECT

The long range objectives of the POSCH AI Project are to use this new technology to:

1. Develop an automated system that will assist our staff of analysts in a comprehensive examination of all of the POSCH data.
2. Identify and test relationships that exist over time between and among the variables.
3. Identify an optimal subset of variables that can predict the outcome of the POSCH study endpoints (for example, a heart attack or atherosclerotic death).
4. Compare the AI based automated system's results with the results accomplished in a more conventional manner and thus provide an evaluation of the AI methods.

Basis For the Study

The POSCH study is a national, multi-clinic, clinical trial designed to test the lipid-atherosclerosis hypothesis in a population where all have had a heart attack. It is known that higher rates of heart attacks and atherosclerotic deaths occur in people with higher levels of cholesterol. The lipid-atherosclerosis hypotheses, yet to be fully proven, says that lowering cholesterol will also lower the rates of heart attacks and atherosclerotic deaths. In order to test this hypotheses, POSCH lowered the cholesterol in half of its patients, the intervention group, using the other half as a control group. POSCH is now collecting a large amount of data on each of the POSCH patients over a 7 to 14 year period.

POSCH's data collection, editing, storage, retrieval, and analyses problems are typical of a number of other large medical research projects. The focus of this paper is on the uses of AI in data analyses. A variety of manual and automated methods are used by POSCH, and other large clinical trials, to analyze the medical research data. The analysis of certain complex tests, especially for changes in the patient's performance over time, requires clinical judgement. This requirement is especially difficult because it has traditionally required manual processing by human experts, usually busy and expensive medical specialists.

POSCH has an additional problem. The designers of POSCH chose to collect about 1400 data items at each annual visit, 600 of which are placed in our computerized database. Even the 600 variables stored in our database are more than our staff can examine in a comprehensive way. At the present time, the POSCH analysis staff examines in detail about 100 of the 600 variables. These so called major variables are known to be related to the lipid-atherosclerosis hypothesis, the focus of our study, and/or the partial ileal bypass surgery, the intervention modality used to lower cholesterol. We are especially concerned with the potential effects of changes in these variables on the POSCH patients, either beneficial or not beneficial. The other 500 so called minor computerized variables are collected because there is some possibility that they could be related to our study focus. They are examined in much less detail. The remaining 800 items that are not entered into the computer are useful at the time of data collection but are judged to have little long term value. They consist primarily of series of questions asked by the physician in order to determine the patient's health status in a certain area. The relevant health status is entered into the automated portion of the patient's record.

POSCH is attempting to resolve the problem of analyzing so many variables by finding supplemental and automated methods for searching for relationships between, and among, all of the computerized variables; especially the ones that...
cannot receive the special attention our staff has devoted to the 100 or so major variables.

There is a potential trap in such a detailed examination of the POSCH data, such as finding spurious or irrelevant relationships. Statistically significant differences that are clinically unimportant are apt to be found when one examines such a large database in so much detail. The analyses must be performed in an intelligent way. We expect to avoid such pitfalls by designing the expert system rules to do the same thing that a human would do to avoid such errors. Direct human intervention will be used as a last resort. That is to say, we accept the possibility that we may not be able to fully automate the process. Even partial success, which we have already achieved, makes this work very practical.

The project has been under development for several years and this paper presents the current status. At this stage a basic plan has been formulated, the necessary resources have been assembled, and two prototype expert systems have been developed.

The Resources

1. The POSCH Database. The database is an hierarchically organized database that includes about 600 computerized variables collected on 838 patients upon entry into the study (that is, the baseline data) and at annual visits for a minimum of seven, and up to 14, subsequent annual visits. At present the database is about 75% of its ultimate size and includes a total of about 200 million characters broken down into about 160 million in MEDDB, the main data storage area, and 40 million in STATFILE/LOCATOR, an on-line administrative type database that includes the status and location of all patient visits, data forms and documents.

2. Catalog of Variables. The catalog classifies and organizes, into clinically logical groupings, all data items collected by POSCH. We may use the classification system as a basis for simplifying and directing the discovery phases of the project.

3. Human Resources. The POSCH Group has the necessary statistical and medical expertise. POSCH has clinical expertise related to the lipids, electrocardiographic stress testing, coronary arteriography, surgery, and other areas. The project is a joint effort with the AI group within the Department of Computer Sciences at the University of Minnesota. They have many years of experience in AI including the design, development, and implementation of many expert systems. The Project is using an expert system development tool built by members of our AI team called AGNESS (a generalized network-based expert system shell).8

4. Computer Hardware and Software. Existing mini-computer facilities in the POSCH data center have been supplemented with an IBM PC AT workstation and with SUN AI workstations in Computer Sciences. Virtually any computer type or size is available for our use at the University, including the largest supercomputers in existence and parallel computers. We anticipate the need for one or both of these in future phases.

AN EXPERT SYSTEM FOR ANALYSES OF SERIAL GRADED ECG TESTS

Our first expert system was designed to compare the data from a pair of graded exercise electrocardiogram (ECG) tests taken several years apart in order to determine whether the patient’s performance was better, unchanged, or worse over time. This system can approximate the decision reached by a cardiologist evaluating the same data. We will only briefly describe this system because it has been reported previously.5,8

The expert system rules for the ECG system were developed using an iterative process in which the knowledge engineer and expert met to discuss and analyze sample cases. The set of cases were carefully selected so as to present a variety of typical situations and to stimulate explanations by the clinician as to what he was doing to solve the case. The clinical expert explained the factual knowledge he used from scientific literature, often citing results of research performed by himself and others. He also used and explained the “rules of thumb” or heuristics that he found helpful. These are based on his experience, rather than on book knowledge, and their incorporation into the system is one of the unique reasons expert systems work. As these sessions progressed, the knowledge engineer formulated, modified, discarded, replaced, and expanded the rules used by the domain expert, either stated or implied. The computer version of most of the rules is of the “IF . . . THEN” type. The IF part, called the antecedent, or premise, contains the pattern or attributes that must be matched for the rules to be used. The THEN part, called the consequent, contains the action to be taken, or the assertion to be made when the antecedent is satisfied.

The resulting expert system was tested on 100 cases that were used to validate the system. Each case consisted of a pair of tests taken by a POSCH patient two years apart. The cases were selected to be representative of a variety of situations. The cases were also evaluated individually by two different members of a panel of five expert cardiologists. Each one evaluated 40 cases. The 100 pairs were evaluated in such a way that each reader’s cases were equally distributed among the other four readers for the other reading. Table I illustrates how this was done. We then compared the conclusions made by the expert system with the individual cardiologist’s evaluations.

<table>
<thead>
<tr>
<th>Reader A</th>
<th>P-P</th>
<th>P-N</th>
<th>N-P</th>
<th>N-N</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>B with A</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>C with A</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>D with A</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>E with A</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

Typical pattern of overlap of each reader with the other readers for the set of 100 pairs of exercise ECG tests evaluated by them.

P = Positive Test
N = Negative Test
Since a more conventional way to automate this situation would be to use a statistical approach, we also developed a set of multiple linear regression equations as a third method of evaluation. We briefly digress here to mention that, although the equations worked, several variables used in the multiple regression equations had obscure clinical meanings, a matter of concern to POSCH clinicians.

All three methods used a seven category scale to describe their conclusions. When a given pair of tests was evaluated by either of the two cardiologists, or the expert system, or the multiple regression equations, they concluded whether or not a patient’s result was better or worse from the first to the second test using the seven category scale shown below:

<table>
<thead>
<tr>
<th>much worse</th>
<th>worse</th>
<th>no slighty better</th>
<th>better</th>
<th>much better</th>
</tr>
</thead>
</table>

Because of the strong element of subjective clinical judgement in these evaluations, the three methods for evaluating the test data were not necessarily expected to draw the same conclusions. For this reason the evaluation methods were compared in two ways as to how well they agreed. “Exact” agreement means that the same category on the seven category scale was selected as the conclusion for both of the evaluations being compared. Agreement “within one category” means that the two evaluations selected the same or immediately adjacent categories of the seven category scale. The comparisons were made based on the percentage of agreement.

Table II summarizes the results. For “exact” agreement the expert system agreed with the cardiologists about as well as the cardiologists agreed among themselves and did better than the multiple regression equations. For agreement “within one category,” the expert system performed best and the multiple regression equations’ evaluations did better than the cardiologists among themselves.

After allowances for normal variation, it can be seen that even a very basic expert system can evaluate serial graded exercise ECG test data about as well as either an individual cardiologist or the multiple regression equations.

**AN EXPERT SYSTEM FOR ANALYSIS OF SERIAL CORONARY ARTERIOGRAMS**

The next system we are attempting to build is one to evaluate serial coronary arteriograms. The general concept is the same as for the ECG system but the medical aspects are far more complex. A brief background on the medical nature of the problem is needed in order to explain this. Narrowing of coronary vessels by lipid based deposits (stenosis) can cause decreased blood circulation to the heart muscles. In extreme cases the blood flow can be restricted enough to cause severe chest pain (angina pectoris) and can cause the death of those parts of heart muscle (myocardial infarction) whose blood supply depends on the blocked artery. Arteriography (a procedure that photographs the pattern of blood flow in the coronary arteries) yields useful information about the condition of the coronary vessels. The technique involves injecting a contrast medium sensitive to x-ray film into the heart vessels followed by a series of 35mm x-rays taken in rapid succession. A cine film strip is thus produced that shows how the blood fills the arteries of the heart. By repeating this procedure from several angles, a radiologist can examine the films and tell which vessels have stenosis as well as the nature and extent of stenosis. Narrowing of the blood vessels by either a lipid deposit stenosis or a blood clot (thrombus) shows up on the arteriogram as dark regions within the artery.

Cases from the POSCH study are used to build and test the experimental system. Coronary arteriograms of participants in POSCH are taken at the time they enter the study and at 3, 5, and either 7 or 10 years later. An arteriography review panel of cardiologists and radiologists has been formed by POSCH to clinically evaluate pairs of these serial coronary angiograms taken three to ten years apart. The methods they use to evaluate these clinical data are subtle and require a considerable amount of clinical judgement. The current method of assessment is for a subpanel of two members from among the eight doctors on the arteriography review panel to take turns meeting for two days, about once a month, to review about 30 to 40 pairs of arteriograms. This is an extremely tedious task for the doctors on the panel and logistically complex. The panel members live all across the United States.

Here are some of the details of the assessment process. The subpanel review is conducted in a double-blinded fashion. The members know neither the identities of the participants nor the temporal sequence of the arteriograms. The film pairs are identified simply as Film A and Film B. Film A is evaluated and all stenoses found are recorded. Film B is then evaluated for change from Film A. In the final step of their evaluations, the subpanel carefully reviews all of their findings and provides a global assessment of change using a scale similar to the one used for the ECG system. The total process requires about 20 minutes of the subpanel’s time to review one pair of films with the global assessment taking only a few minutes at the end. The findings of the subpanel are recorded on a standard form and the information is entered into the computerized database.

The coronary vessels in the arteriogram appear in a tree-like branching structure wrapped around the heart myocardium (muscle). What is visible in one frame of the cine may be obscured in another. Stenoses near the branching point of arteries are especially difficult to estimate. A factor affecting visualization is the presence of collateral arteries. When the normal blood flow in one branch of the system is blocked, collateral arteries (arteries at the ends of an adjacent branch
of the system) will sometimes open up and extend their perfusion field (provide a blood supply) to the affected muscle tissue. This amazing ability of the heart to adjust can complicate the task of determining stenoses. For example, if the blockage of the normal flow decreases, the collateral flow may also decrease and can disappear altogether. Another factor that must be assessed by experts is whether the blockage is caused by a thrombus or stenosis.

Assessing the change in stenoses is further complicated by the fact that vessels tend to develop stenosis more quickly after coronary bypass grafts have been placed on them. Medical procedures such as recanalizations (opening up the vessel by angioplasty) are also fairly common and complicate the evaluation.

These are examples of the many complex and interactive factors that make assessing the percentage change in stenosis difficult. For these reasons we chose to build only one part of the evaluation process in our first phase. As the first phase of this system, we chose to approximate just the global assessment process.

The initial knowledge base was built to perform this one task. Data elements from the consensus report (all previous steps) of the experts are used as the expert system's input and they form the leaf (entry) nodes of the network. The top-node represents the system's global assessment of the overall disease change. The interactions between the tree-like structure of arteries are not required at this point because the heart, as a pumping organ, is not to be evaluated. Thus, each artery can be treated independently. The change in each artery is assessed and the individual changes combined to obtain the overall change. The inference network therefore consists of a sub-network that is evaluated 22 times; once for each of the 22 arterial segments under study by POSCH; and a top-level network that merges the information passed up by the sub-networks.

Table III gives the results obtained by the global assessment expert system when applied to 56 test cases. Terminology and comparison methods are similar to those used for the ECG system.

The POSCH quality surveillance program has shown that two different panel evaluations will agree "exactly" on a seven category scale 55% of the time and agree "within one category" (select the same or adjacent categories on the scale) 91% of the time. The expert system's assessments of the set of 56 test cases agreed "exactly" with the panel 50% of the time, and agreed "within one category" of the panel assessments in 96% of the cases. The system's performance is roughly comparable to that of the panel for the global assessments.

The current, early version of the system does not consider many of the factors used by the subpanel. For this reason, the results obtained by this system are somewhat surprising and should be reviewed with caution. Further tests are needed. It is doubtful that the final version of the expert system for evaluating serial coronary arteriograms can perform this well, because the system does not include many of the facets of subjective clinical judgment that are used by the human experts for these evaluations. For example, the current version of the expert system does not translate many of the subtle things observed on the film by the arteriographers. Furthermore, we do not know how to interpret and record many of these data. Nonetheless, our success encourages us to hope that we can build a system that can closely approximate the results of these subpanels using a single simplified reading of each film by trained technicians.

### Table III—Average agreement rate of subpanel (SP) compared to expert system (ES)

<table>
<thead>
<tr>
<th></th>
<th>SP vs. SP</th>
<th>SP vs. ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>50%</td>
<td>55%</td>
</tr>
<tr>
<td>Within one Category</td>
<td>96%</td>
<td>91%</td>
</tr>
</tbody>
</table>

### Table IV—Comparison of POSCH study to RX project

<table>
<thead>
<tr>
<th></th>
<th>RX (RADIX)</th>
<th>POSCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>Large</td>
<td>838</td>
</tr>
<tr>
<td>Number of Variables</td>
<td>Non-Standard</td>
<td>1400</td>
</tr>
<tr>
<td>Number of Visits</td>
<td>Variable (up to 50)</td>
<td>8 to 15</td>
</tr>
<tr>
<td>Protocol</td>
<td>Nonrandomized</td>
<td>Randomized</td>
</tr>
<tr>
<td>Time of Visits</td>
<td>Variable</td>
<td>12 Months</td>
</tr>
<tr>
<td>Data Elements</td>
<td>Variable</td>
<td>Standard</td>
</tr>
<tr>
<td>Total Size</td>
<td>Very Large</td>
<td>Very Large</td>
</tr>
<tr>
<td>Setting</td>
<td>Clinical</td>
<td>Clinical Trial</td>
</tr>
</tbody>
</table>

An expert system to automate the search for new medical knowledge

The initial motivation for the POSCH AI Project was to build an expert system that can identify relationships that exists between, and among, the POSCH variables. This entails building an expert system that can automate what biostatisticians do when faced with the analysis of a large database like that in POSCH. The idea for this was inspired by the early work of Robert L. Blum in his RX discovery project, now renamed RADIX. His effort to build an expert system to discover and confirm causal relationships in a medical record database appears to be the first attempt to use this approach in medicine. The system obtained its initial hypothesis by selectively combing through a database using a discovery module. It combed through a selected subset of 50 patient records to produce an hypothesis such as A causes B. What it actually did was to determine that A precedes B and is correlated to B. A study module then designed a comprehensive study of the most promising hypotheses as determined by the human researcher. A statistical module was then used to test the hypothesis on the entire database. Newly discovered data were added to the knowledge base and then used in future phases of the study. In the more recent work of the RADIX project, he and his co-workers have a more advanced system using more sophisticated statistical methods.
Much of the work required in the development of the original RX system had to deal with the fact that the data were nonrandomized and included many missing data elements. The frequency and timing of patient data were also variable. POSCH does not have these problems. Instead, it must deal with a much larger number of variables and other issues. Table IV summarizes the differences.

As we attempt to put substance to our efforts, we have identified several things that must be done before we can begin to build the expert system designed to search for knowledge now hidden in the POSCH database.

Unifying Concepts

Some unifying concept must be used to pull together into fewer units the diverse variables in POSCH's database. We have been examining several classification schemes to accomplish this. Logically formed clinical groupings already exist within our Catalog of Variables. We can use either a statistical clustering method or clinical knowledge or some combination of them to produce a dozen or so groups into which all POSCH variables are placed. By using a single entity as a representative of each group, we will reduce the initial complexity of the problem.

Role of Statistics and Reasoning Systems

Standard statistical methods provide the basis for examining the database and for describing and explaining the relationships that exist between and among the variables. These methods are based in probability theory. However, the key to the success of the system is its ability to imitate what a biostatistician does in the discovery phases of his work.

It may seem to be a contradiction for us to attempt to use expert systems in data analysis because most successful expert systems use heuristics. These expert defined heuristics or “rules of thumb” are usually necessary in order for the system to work at the level of an expert. The heuristics often have no conventionally based scientific foundation and rely solely on the expert’s experience. They represent that “knowledge” of an expert that goes beyond book knowledge and this “knowledge” is what makes the expert an expert. He may not always realize that a part of his expertise lies in those intuitive things he relies on when selecting, using, and interpreting the “tools of his profession.” It appears that heuristics will be especially important in the search and discovery phases of our system and it may not be possible to accurately place them into any conventional system of logic because we are trying to automate the creative phases of data analysis. Once the search and discovery phases are done, confirmation can be provided using the more routine statistical processes.

Database Issues

Another issue to be resolved has to do with finding the most efficient database structure to use in order to facilitate the needed logical manipulations of the data. Once identified, we must rebuild the database into that structure. This seems to mean that we should convert to some form of a relational database. We also need to obtain or build a reasonably efficient query language that will allow our searching module to run efficiently.

The computing capacity requirements of our system are apt to be enormous. We may need a supercomputer, but preliminary investigations indicates that a parallel processing architecture is suitable for the expert systems processing.

SUMMARY AND CONCLUSIONS

We have briefly described the POSCH AI Project. In this project we are attempting to automate many aspects of clinical research data analysis previously requiring manual processing; usually because clinical and statistical judgment are necessary components of the analysis. We are meeting with success in automating the evaluation of serially administered tests that require such clinical judgment in their evaluations. Our efforts to automate the search for medical knowledge in the POSCH database are now focused on preliminary problems that need to be resolved. This involves some means of simplifying the problem by use of unifying concepts, identifying the role of statistical reasoning, and of other reasoning systems, and developing an efficient database structure and query language.

ACKNOWLEDGEMENTS

The Program on the Surgical Control of the Hyperlipidemias (POSCH) is funded by NHLBI Grant HL15265.

REFERENCES


The challenge was seemingly never-ending. Six months of hand calculation could be carried out in fifteen minutes on one of these stored program machines, and the resulting time and cost advantages rapidly made these machines a scarce resource. A few of us dedicated ourselves to devising methods of using this resource more efficiently, and one result was a set of programs called an operating system. This is the story of those efforts and the theme of the 1987 NCC Pioneer Day: "Early Operating Systems."

The first session of the Pioneer Day program describes a period of time in the 1950s when the ideas for operating systems were germinating. Pre-operating system days are described when a programmer entered a room full of computer equipment, card deck in hand, and bent every effort to get the most out of a scheduled fifteen-minute time period. The experiences of the people at MIT with Whirlwind and at North American Aviation with the IBM 701 computer are related. Out of this background several organizations worked on methods of improving efficient use of both programmers and machines. In the words of George Mealy "There was no single line of development...[so]... I have chosen to examine a number of threads in the tapestry, for it is far from clear how to describe the tapestry as a whole." Some groups were working toward increased efficiency by providing libraries of programs such as assemblers, decimal to binary conversion, and debugging aids. Others concentrated on reducing computer idle time through automatic sequencing of a number of jobs in a batch. Both are part of the modern operating system.

The second session explores a number of systems that were implemented in the mid to late 1950s. An operating system developed jointly by General Motors and North American Aviation is described. Both automatic job sequencing and libraries of programs were features of this system which began operation in 1956 on an IBM 704 computer. The Bell Telephone Laboratories followed with BESYS in late 1957, also running on an IBM 704. The story of BESYS takes us through a series of versions culminating in a powerful system with file handling capabilities, buffered I/O, and many other features found in modern systems. The program continues with a paper on the FORTRAN Monitor System developed by IBM in the late 1950s. A number of organizations had installed FORTRAN on their own systems and, through SHARE, were urging IBM to provide and maintain a system for FORTRAN. The author describes the IBM interactions with SHARE and the development of IBM's first operating system (FMS). IBM's follow-on work with IBSYS and IBJOB in the 1960s is also recounted. Concurrent with the FMS effort, SHARE was also working with IBM on the development of the SHARE Operating System (SOS). The final paper in the afternoon session describes this last major effort undertaken in the 1950s. Many of the features revealed in this paper, such as job management, data management, and run-time facilities, are direct ancestors of today's IBM operating system.

The 1950s were exciting times, and the papers in the Pioneer Day sessions reflect that excitement. Young and old conferees will share in the experiences of thirty years ago, not only to learn why and how operating systems came about, but also to learn how to build better systems in the future.