Use of the Help Clinical Database to Build and Test Medical Knowledge

Omar Bouhaddou, Peter J. Haug, Homer R. Warner

Department of Medical Informatics, University of Utah, School of Medicine

ABSTRACT

When building knowledge for expert systems, the analysis of patient observations and interpretations of those observations can provide a valuable source of information in addition to literature review and expert consultations. In the HELP system, a complete medical information environment, patient data is collected on a routine basis from all parts of the hospital and is made available to support knowledge base development projects. In return, the system interpretations enrich the patient database and further provide for data accuracy and validation. This work illustrates the integration of real patient data to the medical expertise and the medical literature for acquiring and validating medical knowledge.

KEY WORDS :

medical information system, patient clinical database, information retrieval, expert systems, knowledge base, knowledge engineering, H.E.L.P

INTRODUCTION

Building a decision model involves restructuring knowledge in a particular medical speciality into an explicit form "understandable" to a computer. This is generally followed by long periods of testing against real case material with many revisions before an accurate representation evolves. The sources of information available to assist the experts in defining decision logic are often limited to their expertise and the literature. The building of knowledge can also be effectively supported by a database of patient observations and interpretations of those observations [1, 2].

The HELP system, a complete information system[3, 4], has been built around an extensive patient database. This clinical database has served as a powerful alternative source of information in the development of the HELP knowledge base. In order to efficiently access this database a set of high level tools were created whose main purpose is to steamline the design phase of query generation while protecting the user from the complexities of an active patient database.

This paper describes these tools and illustrates the importance of clinical database in medical knowledge acquisition. It demonstrates the potential of the tools to access patient data and describes a knowledge engineering environment where real patient data have been used dynamically to supply alternative information to the knowledge engineers.

A COMPLETE MEDICAL INFORMATION ENVIRONMENT

HELP is a hospital-based computer system used for acquiring data and implementing medical logic[3, 4] (Figure 1). The time-oriented database consists of two elements: a long term abstract of demographic and clinical information and a short-term collection of all data gathered during the current hospital admission .

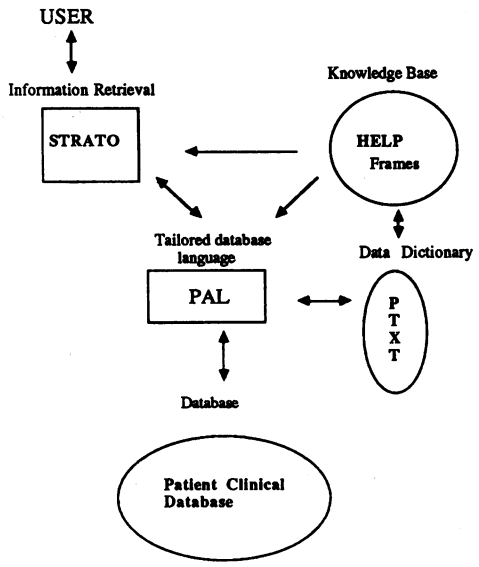

Figure 1: The HELP system environment: a comprehensive patient database, a hierarchical data dictionary (PTXT), a knowledge base and a tailored database language (PAL) provide for a powerful environment for a database analysis program (STRATO).

The data available in the clinical database include drugs prescribed, laboratory data, X-ray and ECGs interpretations, respiratory therapy data, biopsies, microbiology data and other clinical information. Efforts are currently being made to integrate into the database physician and patient derived data including history and physical examination.

All data items are stored in coded form (as opposed to free text). That is, any information that applies to a patient is

assigned a code sequence that gets stored in the patient file instead of the text. For instance, if one wants to specify a 30 mg tablet of Codeine, one will simply store 8 1 2 3 2 in the patient file (Figure 2). This coded format allows for data to be manipulated in decision logic. A computer based data dictionary called PTXT, containing some 70,000 terms, is evoked either for storing or for reviewing patient data. In addition to the text, key words are defined offering an easier access to the dictionary. The coding scheme of the dictionary follows a tree structure. In other words, the dictionary terms are arranged in a hierarchy that reflects knowledge about a domain. The PTXT dictionary also contains other medical nomenclatures such as ICD9 and SNOMED codes.

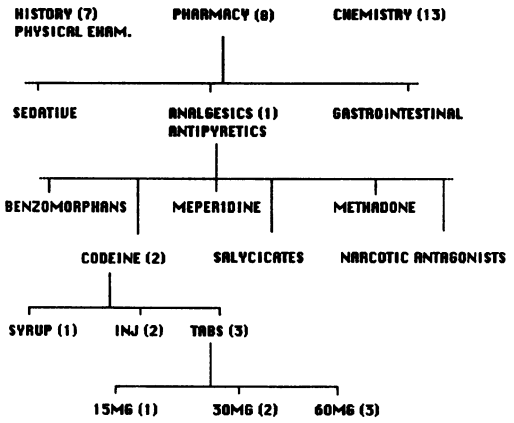

Figure 2 : The HELP dictionary terms are arranged in a hierarchy that reflects knowledge about a domain.

HELP contains a separate knowledge base consisting of the logic for a group of medical decisions that are triggered by the system when a prespecified set of clinical circumstances occur. These medical algorithms or sets of rules are called Help frames and are stored in a modular knowledge base integrated into the medical information system (over 2,000 different HELP frames are clinically operational). Multiple decision models are supported in the HELP system to allow flexible modeling of medical knowledge[6].

HELP's comprehensive clinical database constitutes a rich source of information. For the medical decision author, the availability of this resource offers assistance not only in defining the logic for the frames but also provides a setting to test the knowledge frames produced. In other words, patient data may be used as often as needed to refine, confirm or deny any clinical decision frames under development.

KNOWLEDGE BASE DEVELOPMENT: SOME TOOLS

Specific tools for examining the clinical database are provided in the HELP system. The HELP information retrieval program, STRATO, has been designed to enhance the ease of access to the clinical database while preserving powerful means to describe the information searched. An important part of STRATO is to provide an interface to powerful PC-based programs for database analysis.

The STRATO program retrieves and stratifies HELP patient database information for subsequent analysis.

STRATO prompts the user to define his selection criteria, processes the query against the patient database and structures the results in the form of populations (list of patient numbers), variables (a value for each patient) and time series (for each patient, at each time index, a set of values).

The query definition

The following phrase describes the generic format of a STRATO query.

BUILD, for each patient, a value corresponding to:
{ selected term (s) from the PTXT data dictionary }
WHERE
- 'existence conditions' are satisfied
- time of data is between the 'from' and the 'to' values
- when there is more than one candidate who satisfies the two first conditions consider the 'modifier'

Note that the link to the PTXT data dictionary, as the first step in defining the information searched, makes all fields pertaining to the description of a given patient attribute available for retrieval. The use of key words to search through the dictionary relieves the user from having to know the exact terminology HELP uses to represent medical information in order to retrieve it. Moreover the tree structure of the data dictionary permits the use of a wild card at any node of the tree.

Once dictionary terms have been selected STRATO offers the user the flexibility to further constrain the information to be retrieved . The methods of constraining the information retrieved are based on both the temporal and contextual components of the information. The chronological component is addressed by restrictions based on time range or on time in relation to occurrence of other data. The contextual component is specified as a value range or as a value in relation to value of other data or as coexistence condition with other data. Any combination of the above constraining methods is allowable.

Knowledge frames as search algorithms

The scenario described to extract data from the patient file is quite satisfactory for most of the user needs. However when the selection algorithm becomes sophisticated and leads to multiple classification groups, the original query may have to be converted into a series of simpler queries. Another approach is to use the HELP knowledge base. A HELP frame can serve to represent the selection algorithm of a query to the patient database. The use of HELP frames to define selection criteria supports very sophisticated queries involving medical data from different sources (e.g. history and physical exam data, laboratory results, radiology findings, ...) as well as complex data manipulations (e.g. Bayesian calculations).

Front Ends

Several front ends are available to initiate queries for the clinical database. These are accessible from either the HELP host computer or a remote desktop computer. For instance, for Macintosh users, a short communication program, written in C, has been developed to establish the link between the HELP system and the Macintosh. In fact, the Macintosh based program makes the link invisible. It takes all the commands in the nearly natural environment of the Macintosh then translates and sends them to the STRATO program.

Back Ends

STRATO also provides an interface to a number of 'back ends' or descriptive tools for looking at the extracted results. Indeed, to further manage and extract the pertinent information from the collected data, the user can choose to persue the investigations using his favorite data manipulation package on

the personal workstation . STRATO has built-in facilities to download data from the HELP host computer to microprocessors in a format readily accessible to several powerful commercial software packages. In addition, the possibility of carrying out the research activities on the remote workstation reduces the load on the clinical system.
The coupling of a rich database with convenient graphic routines and statistical analysis programs makes patient observations a direct and effective pathway to medical knowledge. The link to the personal work station stimulates the back-and-forth approach from data extraction to data description. Therefore the design of the system contributes to shortening the time between idea conception and testing. Note that no attempt is made to automate the process of medical knowledge acquisition such as in references 1 and 2 and medical involvement is necessary to make pertinent use of the clinical database.
In summary, a larger number of users in the research community are given access to the powerful HELP patient database without requiring from them any specific experience either with the host computer or with the HELP system syntax. In return, the use of the database improves accuracy of medical decision logic put back into the HELP system.

KNOWLEDGE BASE DEVELOPMENT: SOME EXAMPLES.

The development of "expert systems" for the HELP system is a major focus at the department of Medical Informatics at the University of Utah. For that goal three sources of information are used conjointly: expert experience, literature review and the patient data. All sources contribute to suggesting or confirming or rejecting relationships between medical variables as well as estimating specificities and sensitivities of findings within a disease context.
Several knowledge development projects are currently active in the fields of orthopedics, hematology, pharmacy, pain management and primary care medicine.
In each one of these knowledge production projects, the STRATO program is being used to provide the knowledge engineering team with pertinent information from patient data as well as prospective evaluation of the performance of the knowledge frames developed.

Hematology knowledge base

As an example, in the hematology group, a team of experts from different fields (hematology, medical informatics, library sciences, statistics) meets once a week to expand the HELP hematology knowledge base.The scope of hematology knowledge under development by this group covers a large range of blood disorders. An intermediate decision, the recognition of hemolysis, is important in the diagnosis of several specific blood diseases (e.g. acquired hemolytic anemias, thalassemias, etc).
The knowledge representation model used (Figure 3) for hemolysis involves sequential Bayesian calculations. Therefore the information needed to build our decision model includes :

-the best set of variables to use,
-evaluation of correlations (non-independence) among variables,
-estimations of conditional probabilities values (sensitivities, specificities) and
-aprioris.

A decision model for hemolysis

To define the set of criteria that best describes a hemolytic process, we first asked the two hematologists in the group to suggest appropriate data elements. These included the direct Coombs test, the absolute reticulocyte count, the hemoglobin or hematocrit value, the haptoglobin value, the total and the indirect bilirubin, the red cell morphology (target cells, spherocytes, ovalocytes, anisocytes, etc) and the urobilinogen. Considering each variable separately, STRATO was used to extract one value for each patient in the disease and the no-disease populations. The criteria specified the earliest value in the patient record to avoid using information altered by treatment. The discrimination power was then evaluated by means of statistical comparisons between the data distribution of the two populations. The associated sensitivity and specificity served also to describe the utility of a criteria in helping discriminate between the two populations.
To illustrate this process we will describe the investigation of the variable "indirect bilirubin". The hematologists suggested that indirect bilirubin is "elevated" in hemolytic anemia patients (below is given the experts definition of "elevated indirect bilirubin"). Patient data was collected to test for a statistical difference in the distribution of indirect bilirubin in the two groups. For example, given this PTXT terminology :

A) (SMAC) total bilirubin =.== (mg/dl)
B) (SMAC) direct bilirubin =.== (mg/dl)

the condition for the existence of an elevated indirect bilirubin was chosen as:

(A > 1.1 and A < 5.5 and B <= 0.5) OR (A-B >= 0.8 and A <= 1.1)

This represents what the experts meant by "elevated indirect bilirubin" in this context.
Using the "transfer data" feature the samples of indirect bilirubin values for each of the two populations were downloaded to a Macintosh workstation . Then switching to a statistical package, the user accesses data transferred from the HELP system and the two data distributions are plotted on the same graph (Figure 4) showing two distinct curves. The Kolmogorov- Smirnov test is used to further validate the statistical significance of the difference of distribution. Also the associated sensitivity (0.44) and specificity (.81) served to describe the utility of "indirect bilirubin" in helping to recognize a patient having hemolysis. Note that the whole process from data extraction to data analysis takes about 5 minutes using the STRATO program.
The Red blood cell Distribution Width (RDW), available with every CBC order, was considered as an eventual indicator for anisocytosis. Our patient data was not significant in supporting this hypothesis. However high values of RDW (RDW > 17.0) might be more reliable indicator for anisocytosis. Also our data showed that anisocytosis is seen in hemolytic anemia patients just as much as in the rest of the population (sensitivity = .79 and specificity = .18) and therefore was not considered in the frame. Our hematologist confirmed these results and stated that anisocytosis and RDW should be created as separate tests because anisocytosis is a measure of variability in cell diameter while RDW is a measure of variability in cell volume and that this may account for some of the observed anisocytosis in patients with low RDW.
The other information to include in the knowledge frame as suggested by the hematologists was extracted from the database and analyzed as described above. A version of the HELP knowledge frame is given in figure 3.

frame 75.1 16 HEMOLYTIC PROCESS

PURPOSE: To calculate probability of a hemolytic process in a patient, and print out probability, gain and reasons for decision.

LOGIC

A Apriori probability = .0116
B Search for direct coombs test
C Skip next search (positive direct coombs) if direct coombs test not done
D Search for positive direct coombs test
E Calculate probability given positive direct coombs test Sens. = .37, spec. = .94
F Search for value of retic count %
G Skip next two searches (rbc, transfusion) if retic count % not done
H Search for minimum value of rbc within 2 days of retic count %
I Calculate absolute retic count (retic count % * rbc)
J Search for existence of red cell transfusion
K Check for red cell transfusion prior to retic count % or rbc (= "false")
L Calculate probability given absolute retic count > 200,000, unless there was a prior transfusion Sens. = .41, spec. = .91
M Search for existence of anemia (frame "Anemia"#75.1 18) Hgb <12, Hct < 37
N Calculate probability, given anemia Sens. = .85, spec. =.83
O Search for value of haptoglobin
P Calculate probability given haptoglobin < 27 Sens = .57 spec = .83
Q Search for value of total bilirubin
R Skip next search (direct bilirubin) if total bilirubin not done
S Search for value of direct bilirubin
T Check for either condition of elevated indirect bili (= "true") 5.5 > total bili > 1.1 AND dir bili < .6 OR Indirect bili > .7 AND total bili < 1.2
U Calculate probability given elevated indirect bili Sens. = .44, spec. = .81
V Search for existence of differential blood count
W Skip next two searches (target cells,etc and polychromatophilia) if differential not done
X Search for existence of target cells, spherocytes, ovalocytes, schistocytes, sickle cells, basophilic stippling, howell-jolly bodies
Y Search for existence of polychromatophilia, store "1" only if retic ct data not already used
Z Calculate probability, given conditions in item X and Y Sens. = .76, spec. = .89
AA check if posterior probability is now less than apriori probability, store posterior prob and gain, and exit frame without asking for missing data.
BB Ask for missing data.

Figure 3: A HELP knowledge frame for hemolysis

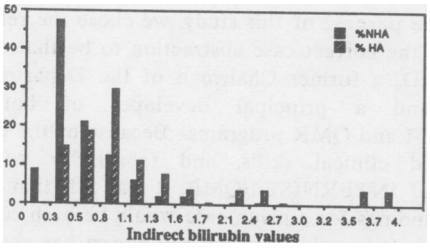

Figure 4: Distribution of indirect bilirubin values in hemolytic anemia (HA) and non hemolytic anemia patients (NHA). The P-value associated to the Kolmogorov-Smirnov statistic is 0.012.

Testing and refining knowledge

As a second example, in the hemolysis model the apriori was first estimated as :

$$\frac{\text{\# of Hemolytic Anemias(ICD9 Discharge Code)}}{\text{Total Population size}} = 0.00054$$

The processing of the hemolysis frame against a population of anemic patients (i.e. HGB<12 or HCT<37) has permitted refinement of this value. Indeed, assuming that a ratio posterior/aprior of 10 or more is a good indication of hemolysis, the proportion of anemic patients is 13.7%. Given the proportion of anemics in the total hospital population (8.5%) the proportion of hemolytic anemia patients in the hospital population is then derived as .137 x .085 = .0116. This value is compatible with the experts estimation of the apriori probability .

In the HELP environment the refinement process of a decision model is an iterative process where values and criteria are constantly adjusted until a given level of accuracy is obtained on real patient data.

CONCLUSION

Computer-based decision models or expert systems are now recognized as an exciting new way to provide the best of medical knowledge to the decision maker[1]. This paper stresses the importance of clinical database in the building of knowledge for these systems.

The environment of the HELP system has provided for an appropriate setting to use patient data to develop the knowledge base. HELP is designed specifically for efficient and flexible storage of clinical information. The information retrieval program, STRATO, has coupled the rich database produced with convenient statistical, sorting and graphics routines. The personal workstation interface has further enhanced the usefulness of the system . Moreover, because of the relatively fast response time, analyses performed on-line allow interaction between the user and the data so that results of the analysis may suggest further investigations.

To the knowledge engineer, the system has revealed the clinical database as a valuable alternative for acquiring and refining medical knowledge. In addition, patient database allows for testing and refining a decision model in a real world setting which gives the knowledge base a source for empirical validation.

In the near future, the availability of large inexpensive storage media (laser disk) , the development of computerized medical records and the genesis of a unified medical language system should stimulate the sharing of clinical databases as valuable and powerful alternatives for acquiring and validating medical knowledge.

This work was supported in part by grant #5 G08 LM 04403-02 to the University of Utah from the National Library of Medicine

REFERENCES

1. BLUM RL Discovery, confirmation, and Incorporation of Causal Relationships from a large time-oriented clinical database: The RX project. Comput. Biomed. Res. 15:164-187 (1982)

2. ADLASSNIG KP, GERNOT K Representation and semiautomatic acquisition of medical knowledge in CADIAG-1 and CADIAG-2 Comput. Biomed. Res. 19:63-79 (1986)

3. MILLER RA, MASARIE FE, MYERS JD Quick medical reference (QMR) for diagnostic assistance MD Computing vol 3 no 5 (1986)

4. WARNER HR Computer-assisted medical decision making Academic Press (1978)

5. PRYOR TA, GARDNER RM, CLAYTON PD, WARNER HR The HELP system. J. Med. Syst. 7(2): 87-102 (1983)

6. NGUYEN LT Transferability of medical knowledge base: a case study between Internist-1 and Help Ph.D. thesis,Department of Medical Biophysics and Computing,University of Utah (1986)