Developing an Integrated Natural Language Database for Gastrointestinal Disease

Peter J. Feczko
Laurens V. Ackerman
Robert D. Halpert
Stuart M. Simms

Follow this and additional works at: https://scholarlycommons.henryford.com/hfhmedjournal

Part of the Life Sciences Commons, Medical Specialties Commons, and the Public Health Commons

Recommended Citation
Available at: https://scholarlycommons.henryford.com/hfhmedjournal/vol32/iss3/5

This Article is brought to you for free and open access by Henry Ford Health System Scholarly Commons. It has been accepted for inclusion in Henry Ford Hospital Medical Journal by an authorized editor of Henry Ford Health System Scholarly Commons.
Developing an Integrated Natural Language Database for Gastrointestinal Disease

Peter J. Feczko, MD,* Laurens V. Ackerman, MD, PhD,* Robert D. Halpert, MD,* and Stuart M. Simms, MD*

Using a mainframe computer connected to the Henry Ford Hospital computer network, we developed a database for gastrointestinal disease which includes data from radiologic and endoscopic gastrointestinal examinations, along with corresponding pathologic diagnosis. Because of the large volume of procedures in our practice, we developed several unique features for our system. The user enters data by responding to a series of question-and-answer sets constructed by the clinical staff, who do not have experience in computer programming. Data is stored in a hierarchical format using natural language. Boolean logic is used to retrieve data so that different procedures can be correlated with each other. In addition, several on-line functions permit us to retrieve data on a given patient immediately and provide computer-generated reports. Because the computer is connected to the hospital network, the database can be accessed from various terminals; data also can be transmitted through the network. Research, educational, and quality assurance functions are other applications of the system.

The management of information in medical specialties is becoming increasingly difficult, particularly in the field of gastrointestinal disease (1). In the last decade, there has been a revolution in diagnostic techniques: fiberoptic endoscopy, double-contrast radiographic techniques, and technologic breakthroughs in computed tomography and ultrasound. Typically, the patient with gastrointestinal disease undergoes many diagnostic procedures which often involve these modalities. In most large institutions, gastrointestinal procedures are performed by diagnostic radiologists, gastroenterologists, colon and rectal surgeons, as well as others to a lesser degree. The ability to manage these data within a specialty is difficult, but to integrate data between specialties is even more challenging. Moreover, as diagnostic studies become more complex, communication problems are compounded. As a result, it is increasingly apparent that there is a need to integrate data across departments in order to compare results, study the efficacy of procedures, provide quality assurance, and record information that reflects the incidence and demographic characteristics of gastrointestinal disease (1-3).

The task of managing these data at our institution is formidable since, during an average year, more than 10,000 radiologic procedures, 2,000 endoscopic studies, and 2,500 pathologic reports are obtained for patients with gastrointestinal disease. In an attempt to meet these needs and take advantage of the latest computer technology, we implemented a database system which integrates these studies so that a clinical group, without programming knowledge, can format and enter as well as retrieve and analyze data. The system was developed on an existing computer within the Department of Radiology.

How the System Works

Component Parts

The system is built around a mainframe computer, a Digital Equipment Corporation VAX 11/750. Although the storage capacity is several hundred megabytes, the database occupies approximately 10 megabytes for its files and programs. The computer is accessed from a terminal located next to the film alternators where gastrointestinal radiologic studies are interpreted; it can be accessed also from other terminals connected to the hospital's computer network.

The operating system is the Berkeley 4.1 UNIX (4), and the database programs are written in "C" language (5). With them, a hierarchical, or branching, file structure was created for data storage. In this system, multiple users can access the database simultaneously, and several functions (ie, data input, data retrieval, statistical analysis) can occur interactively at any time without interference.

Entering data

One of our main concerns when we developed the system was to design a method for entering data...
quickly. Not only do we have a large volume of data, but it must also be managed by clinical physicians without ancillary support. The system employs natural language so that it can be understood by anyone who uses it without the need to learn codes (6). A by-product of this natural language system is that the computer can generate a report when data entry is complete, and the information also can be transmitted over the network.

An interrogative language was created so that the user enters data by responding to a series of questions (7-8). A unique feature of this system is that the question-and-answer sets were written and altered by the clinical staff who had no programming knowledge or support. Accordingly, the staff can tailor data entry to their own clinical needs. Indeed, through customized question-and-answer sets, the database can be adapted to much wider applications; virtually any clinical group can design a database to fit their particular needs. Because the system responds to a series of questions with selected answers, data input is fast, reproducible, and easily learned. Typing and computing skills are not prerequisites for using the system.

After the user signs on to the system, the first series of questions concerns the patient's name, sex, and age. If the patient is already registered in the database, the physician can review all previous studies, if necessary, before entering new data (Fig. 1). The user then chooses the appropriate study from a menu of different procedures (Fig. 2). Each procedure contains a set of questions and answers appropriate for that study, including clinical history, the name of the physician performing the study, the quality or extent of disease, abnormalities, and diagnoses. All questions and answers appear in menu format, and the user is required to choose only the appropriate number, then press the RETURN key to enter data. These question sets usually define type, location, number, and size of abnormalities.

To begin data entry, the user types in the patient's seven digit medical record number. Patients with previous studies in the database are listed along with their names, as above. If there are no previous studies, the computer asks for the patient's name, sex, and age, and then establishes a file.

Fig. 1

PATIENT ID (7 digits)
7654821

Patient's Name: F——, P——-

Previous Exams:
UGI
12/3/83
LGI
12/5/83
EGD
1/4/84
PATH.UGI
1/4/84
1/14/84

Fig. 2

At present, 10 procedures are listed in the gastrointestinal database. Each represents a different question-and-answer set for data entry. All of the procedures may be entered in any patient file. More procedures may be added to the list in the future.

Although this method of data entry seems constrained, it has several advantages: it is rapid, reproducible, and easily controlled by the physician. "Free text" can be entered, if desired, at several points in the program. Once data entry is complete, the data are then organized by the computer and can either be printed (Fig. 3) or transmitted over the network.

Fig. 3

Example of a report generated by the computer, given the data entry from the question-and-answer set. The time needed to enter such data is less than one minute. The computer then formats the data in approximately one second and transmits it to an adjacent printer. GIMS: gastrointestinal information system.
Retrieving and analyzing data

The user searches for data by using well-defined search statements (Fig. 4). A search pattern begins when the studies and dates desired are chosen, and terms or phrases to be matched are then selected. The terms can be either simple or complex and can be diagnoses, a history, features, or even the name of the operator who performed the studies. Boolean logic* allows correlation and matching between procedures (9-10).

<table>
<thead>
<tr>
<th>SEARCH PATTERN</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) (sex f) &amp; (age 8.) &amp; (./83 .)</td>
<td>Women in their 80s with exams in 1983</td>
</tr>
<tr>
<td>b) (UGI ./83 HIST Dysphagia) (EGD ./83 HIST Dysphagia)</td>
<td>Exams done in 1983 for history of dysphagia</td>
</tr>
<tr>
<td>c) (PATH.LGI . DIAG Carcinoma)</td>
<td>Pathologically proven cases of carcinoma of colon</td>
</tr>
<tr>
<td>d) (UGI . DIAG Ulcer, stom) &amp; (EGD . DIAG Ulcer, stom)</td>
<td>UGI exams with endoscopically verified gastric ulcers (True positive)</td>
</tr>
<tr>
<td>e) (UGI . Rep PJF) &amp; (EGD)</td>
<td>Personal quality assurance of UGI exams with endoscopic correlation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OUTPUT PATTERN</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) who, DIAGNOSIS</td>
<td>Only diagnoses are listed</td>
</tr>
<tr>
<td>b) sex age UGI EGD PATH.UGI</td>
<td>All upper gastrointestinal data</td>
</tr>
<tr>
<td>c) sex age LGI COLONOSCOPY PATH.LGI</td>
<td>All lower gastrointestinal data</td>
</tr>
<tr>
<td>d) sex age LGI PATH.LGI</td>
<td>LGI exams with disease</td>
</tr>
<tr>
<td>e) who UGI EGD PATH.UGI</td>
<td>All upper gastrointestinal data</td>
</tr>
</tbody>
</table>

Fig. 4
Examples of data retrieval search patterns. Search output is not completely dependent on the search pattern itself.

Once retrieved, data are placed in a sequential file for analysis by a series of programs. Like the other portions of the system, this function can be done without programmer support for the program's match words, phrases, or numbers, then tabulated or analyzed statistically (11).

Discussion

With over 7,000 patients and 12,000 procedures in its files, this database presently is one of the world's largest active databases of gastrointestinal disease (12). This large volume of data, which occupies less than 10 megabytes of disk space, has been entered entirely by physicians involved in the project, without ancillary help. We have been able to accomplish this task because of the interrogative approach to data entry.

Although the system was designed for research purposes, it has several on-line functions which aid in clinical operations. First, all data on a given patient can be retrieved in approximately one second. As a result, we are less reliant on obtaining records from x-ray folders since we use the computer to retrieve data. Second, because the computer system is able to generate radiology reports, there are several advantages. In addition to financial savings, the workload of the secretarial services is reduced. Since fewer steps are involved in handling the reports, they can be distributed faster with less chance for error.

The ultimate goal is to integrate this database into a hospital-wide network with terminals in clinics and nursing stations. Under this arrangement, radiology reports can be transmitted via the computer network to the involved groups at the time of film interpretation, providing the fastest possible delivery of reports.

The computer database system has helped expand our approach to gastrointestinal disease research. Previously, we were able to evaluate relatively few specific cases or procedures. Now we can create a file of up to several thousand cases in approximately 15 minutes. For example, we are currently evaluating all patients who presented with the clinical symptom of dysphagia, 450 in all. The computer retrieved these patients and matched their endoscopic studies with their pathologic reports. This computation was performed with approximately one minute of physician time and 15 minutes of computer time. In addition, we recently completed an analysis of colonic polyps detected at Henry Ford Hospital over the last eight months and matched these patients with their pathologic findings (13) in a matter of minutes.

Because the system can integrate data from multiple procedures, it readily provides reliable information to assess quality assurance. With its ability to match or exclude data sets, the system is able to isolate true positives, false positives, true negatives, and false negatives in separate files for further evaluation (14). Figure 5 shows an example of the raw analysis of radiologic accuracy in the diagnosis of gastric ulcers as it was retrieved from the computer. Such an analysis requires less than an hour to perform. We have already completed a detailed analysis of the accuracy of detecting colon carcinoma by contrast colon x-ray; retrieval and analysis of an entire year's data required only a few hours. By contrast, a similar quality assurance study performed in 1978 consumed three weeks of time and involved three members of the department. Not only

*Ed. note: Boolean logic is a mathematical system of deductive logic based on the work of the English mathematician Boole.
can quality assurance assessment be performed according to the diagnosis, but retrieved data can also be based upon identity of the radiologist who performed or reported the study, date, sex, age, quality or type of examination, and other variables.

<table>
<thead>
<tr>
<th>COMPUTER SEARCH COMMAND</th>
<th>RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(UGl.DIAG!Ulcer, stomach) &amp; (EGD.DIAG!Ulcer, stomach)</td>
<td>True positives</td>
</tr>
<tr>
<td>(UGl.DIAG!Ulcer, stomach) &amp;! (EGD.DIAG!Ulcer, stomach)</td>
<td>False positives</td>
</tr>
<tr>
<td>(EDG.DIAG!Ulcer, stomach) &amp;! (UGl.DIAG!Ulcer, stomach)</td>
<td>False negatives</td>
</tr>
<tr>
<td>(UGl) &amp; (EDG) &amp; (. . DIAG! Ulcer, stomach)</td>
<td>True negatives</td>
</tr>
</tbody>
</table>

Fig. 5
Example of a raw analysis of accuracy of radiologic exams in detecting gastric ulcers compared to endoscopic exams. By using different search patterns, the computer will automatically isolate the data into different categories of diagnostic accuracy. Sensitivity and specificity can then be calculated ($\& = and; \&! = and not)$.

The system is also designed to include many educational functions. Any patient or procedure can be designated for subsequent follow-up, inclusion in the teaching file, case conference on diagnostic error, or research. These cases can be retrieved at any time for teaching or conferences (15). Finally, house officers are able to interact with a sophisticated computer and operating system, and thus obtain computer experience with practical hands-on exposure (16). The ability to be “computer literate” is increasingly important for physicians in training.

This database system has the potential to perform more sophisticated functions. Because of the hierarchical nature of the data storage, the system has the capability of artificial intelligence. We already have developed several programs which will predict a diagnosis based on the patient’s history. Unlike other systems, the computer prediction is based on the results of the procedures performed, evaluated in terms of data derived from our own patient population. Diagnostic predictions are not dependent on an artificial weighing system obtained from other studies. It is a dynamic program that constantly changes as more data are entered into the files. Although computer-assisted diagnosis based on radiographic and clinical criteria has not yet been implemented in our system, clinical gastroenterologists elsewhere have done so (1).

Acknowledgments
We thank Ms. Patricia L. Lind, Medical Secretary, for the preparation of this manuscript.

References