Warner HR, Rutherford BD, Houtchens B. A sequential bayesean approach to history taking and diagnosis. *Comput Biomed Res* 1972:5:256-262.

A Sequential Bayesean Approach to History Taking and Diagnosis

HOMER R. WARNER, * BARRY D. RUTHERFORD,† AND BRUCE HOUTCHENS‡

Received January 3, 1972

Although a variety of programs have been developed to limit the number of questions which must be asked a patient to elicit the information needed to properly recognize and classify his problems, they have largely used binary decisions based on each response given by the patient (1). With such a branching scheme an atypical patient response at any branch point may result in failure to ask pertinent questions vital to the recognition of the patient's problem. Gorry and Barnett (2) described a program for diagnosis using conditional probability by which probabilities for each possible diagnosis may be estimated at each step in the process of data collection. We have developed a program using a somewhat similar strategy for automated history-taking and diagnosis which is described in this paper. This has been implemented for automated history acquisition in the admission screening clinic of a large general hospital. In this mode of operation a data base for continued improvement of the statistical matrix from which the program operates is being generated.

A table or matrix of estimated probabilities has been generated by the authors to characterize each disease or diagnosis in the set. Each row in the table represents a diagnosis and each column represents a question. Each position in the table is occupied by a number representing the likelihood that a patient having the disease represented on that row would answer "yes" to the question represented by that column. Each row and each column of the matrix is stored on magnetic disc so that only the row or column required for a particular calculation need be read into core memory at any one time, thus minimizing the amount of core memory required by the multiprocessing-time sharing system to run the program (3). Each probability is represented by an 8-bit character and all calculations are done using integer arithmetic.

For each question answered "yes", the following equation is solved:

$$P_{D_i} = P_{D_i} P_{Q_j/D_i} \Big/ \sum_{k=1}^{n} P_{D_k} P_{Q_j/D_k},$$

where P_{Di} is the probability of the patient having the i-th disease after (left hand side of equation) and before (right hand side of equation) answering "yes" to the j-th question, and $P_{Qj/Di}$ is the probability that a patient with the i-th disease will answer "yes" to the j-th question. The denominator is the numerator summed over all diseases and assumes the patient has only one disease in the set. This constraint, however, is relaxed by another tactic in the program which is described below. It can be seen that for this calculation all operations are performed on the column representing the j-th question.

There are some lists that must be used by the program beyond the row and column probability list already described (Table 1). Some of these lists are unaltered by the program as a particular case is processed The first is a list of 50 Key questions which must be asked each patient irrespective of his response to prior questions. These key questions are organized into ten sets, each representing a particular anatomical or functional area. Each time a new set of key questions is asked, the probability of all diagnoses are reset to a common value, thus

* Department of Biophysics and Bioengineering, University of Utah, Salt Lake City, Utah 84103.

Supported in part by N I H Grant RR-00012 and Research Career Award HE 18344.

† Cardiovascular Disease Mayo Clinic, Rochester, Minnesota.

‡ University of Utah Medical Center, Salt Lake City, Utah 84112.

TABLE 1

LISTS

A. Common to all cases

1. Key Questions
2. Linked Questions
   a. Modifiers
   b. Not appropriate
3. Average probability of each question

B. Unique to each case

4. Questions asked
5. Questions asked since last key question
6. Questions answered “yes”
7. Questions inappropriate in this case
8. Diagnostic suggestions
9. Linked questions yet to be asked in this case

making it possible to diagnose multiple abnormalities in a given patient.

The second list points to modifiers for each question which must be asked to further qualify an answer already given. For example, the question, "Do you have or have you recently suffered from pain in the chest causes sets of questions to be asked regarding the nature, time-relationship and location of the pain as well as what brings it on and relieves it. When a question with modifiers is asked, the new disease probabilities are not estimated until the answer to the modifiers is obtained. The probability for the original question is ignored in the calculation since it is not independent, of course, from its modifier. Another type of modification may result in a flag being set which prevents some other question (or set of questions) from being asked. An example of this type of question is “Are you a male?’ If “yes”, don’t ask questions only appropriate for a female.

Another permanent table is the probability of a “yes” answer to each question from any patient selected from the population under consideration and is obtained by averaging the probability for each question across all diseases after weighing each disease by its a priori probability. This table is used as part of the algorithm for selecting questions to ask as the history proceeds.

All items in lists 4-9 shown in Table 1 are set to zero at the start of each history. Items in lists 4-7 are represented by individual bits whose positions correspond to question numbers. A question is never asked more than once. However, when a new set of key questions is referred to, bits in list 5 are set to zero making all these questions eligible for selection. If a question already asked is selected, it is not asked again but the answer is accessed from list 6 and used to influence the calculation of new disease probabilities. List 6 is also used in printing out the final history. Bits in list 7 are set when an answer (yes or no) to a question makes another question, or set of questions, inappropriate as described already.

List 8 contains the current probabilities for each disease and also flags each disease whose probability has qualified it to be listed in the history as part of the differential diagnosis. The current probabilities for all diseases are reset to a common value when a disease has satisfied the diagnostic criteria or when, for some other reason, a new set of key questions is presented. List 9 is used to keep track of modifying questions which are appropriate as a result of answers already given but have not yet been asked. These questions are presented in sets of 3 to 8 at a time. Not only may several sets of modifiers be triggered by one answer, hut modifying questions may themselves have modifiers.

Each new set of questions is selected for presentation to the patient based on the following criteria: First, are there any modifying questions which have not yet been asked? If so, these will be asked before another set of independent questions will be selected. If no modifying questions are pending, new disease probabilities are calculated based on the latest answers given and the two most likely diagnoses are selected. One question is selected for each of these diseases from among the questions not already asked in this sequence by finding the question which maximizes the ratio of the probability of getting a "yes" answer in a patient having this disease to the probability of a "yes" answer in someone randomly selected from the general population (average probability for that question).

The other three questions are selected to distinguish between these two most likely diseases. To achieve this, the ratio of the probability of a "yes" answer to each question in one disease to the corresponding probability in the other is examined to find the three questions with the largest values. If a ratio is less than one, its reciprocal is used. If the probability of a "yes" answer to a question in either disease (hut not both) is zero, this question is immediately chosen since a "yes" answer would rule out that diagnosis.

These questions are presented on a display terminal to the patient with a one-digit number preceding each question. The patient uses a numerical keyboard to enter the numbers corresponding to the questions to which he would like to answer "yes". When he has completed his entry, he asks for more questions by pressing the return key. At this point the sequence of question selection just described is repeated.

This sequence continues until one of two criteria is met. First, if the probability of the patient having a particular diagnosis exceeds 0.90, further questioning is considered unnecessary and this diagnosis will he suggested on the history printout. All the disease probabilities will now be set to a common value so that a new line of questioning can begin without bias based on questions asked to this point. The next set of five key questions will be selected for presentation. If one or more of these have already been asked in an earlier sequence (i.e., appeared as the best questions for a particular circumstance), it will not be presented to the patient again, but the answer already given will be used in calculating probabilities.

The second criteria for returning to the key question list is based on the lack of sufficiently sensitive questions among those not already asked in this sequence to distinguish between the two most likely diagnoses. If the ratio of the probability of a "yes" answer to each question in a patient having the most likely disease (D_{max1}) to the corresponding probability in the next most likely disease (D_{max2}) (or the reciprocal of this ratio) is less than I .2 for all questions, this criterion is satisfied and a new set of key questions is sought. However, a decision must now he made as to whether either or both of the currently most likely diagnoses should be presented as part of the differential diagnosis on the printout. If the sum of these two probabilities exceeds 0.5, the most likely disease will be printed, and if the probability of the next most likely disease exceeds 0.2, it will be included also in the differential diagnosis. When a new set of key questions is sought and all have been asked, the history is complete.

In the history printout a statement corresponding to each of the questions answered "yes" by the patient is printed in the format shown in Fig. 1. Each disease in the differential diagnoses is followed by those statements which contributed to making that diagnosis a likely choice. This is accomplished by examining the ratio of the probability of a "yes" answer in a patient with the first diagnosis to each of the questions answered "yes" by this patient to the corresponding average probability for this question and printing those that have a ratio greater than 1.3. If a statement is printed under one disease, it is not repeated under another even though it may have been important in establishing the likelihood of the second disease as well. Those statements corresponding to "yes" answers which are not related to one of the diagnostic suggestions are printed under the appropriate part of the system review as determined from a list which classifies each question.

The goal of this project is to: (1) quickly and accurately collect the information from a patient which will be most pertinent to the decision-making process essential to his care; (2) proceed as far as possible with the intellectual task of organizing and classifying the information in order to save the M.D. time and effort and

improve his performance; (3) accomplish this at a cost less than that incurred by existing methods. Experience to date with the program used on patients being admitted to the hospital shows that it takes the average patient less than 10 minutes to complete the questionnaire. There are at present 320 questions of which at least 50 must always be answered. Tests are now underway to compare the relative accuracies of the answers obtained in this

Patient History

Moss George K 106651 1/5/72

History Suggests

Acute pericarditis
--Because the patient
Experiences pain aggravated by assuming supine position
Gets chest pain lasting longer than 30 minutes
Recently had a cold or sore throat
Gets chest pain which is aggravated by taking a deep breath
Is awakened at night by pain
Has recently had a fever
Is often short of breath
At times has palpitation

Prostatic hypertrophy
--Because, in addition, the patient
Notices urine stream is less forceful
Has difficulty starting or stopping urine flow
Gets up several times at night to pass urine
Is a male
Is over 40 years old

System Review

The patient also
1 Constitutional
2 Skin
3 EENT
Is aware of losing hearing
4 Neck and nodes
5 Breasts
6 Heart
Has had several attacks of the same chest pain
7 Lung
8 G-I
Experiences pain which is sharp and stabbing
9 G-U
10 M-S
Has varicose veins
11 Endocrine
12 Neuro
13 Psych

FIG. 1

project to those obtained by a nonbranching questionnaire and by a resident physician and/or the attending physician on the case.

The ability to make the correct diagnostic decisions from the data will depend primarily on the validity of the statistical matrix upon which the calculations are performed. Although the performance of the system is very encouraging even with the approximations which now constitute this table, two parallel efforts are underway to improve these values. Ten qualified consultants representing each of the key areas have agreed to participate in establishing a better approximation to the probability table. In addition, a system is in operation by which each patient taking the history not only has his answers stored in a computer-based medical record, but at the time of discharge is diagnosed by his attending physician and that diagnosis is coded using an expanded version of Systematized Nomenclature of Pathology (SNOP) by the computer and stored in the same record. Programs are operational to review this file of patients and determine just how many with a given disease did answer "yes" to each question. These patients are entering the system at the rate of 30 per day.

The costs of operating this history program on the MEDLAB time-sharing system is less than $1 .00 per history. However, this estimate depends on optimal utilization of the remaining system capability over a 16-hour period each day. In fact, the system is used (but not to capacity) 24 hours a day, 7 days a week for patient monitoring, screening, clinical laboratory automation, and other clinical chores. The history program is written in a reentrant form so that several patients may be using the same version of the program in core memory.

Two other versions of the program are under development. One uses the same statistical matrix in a consultation mode in which the M.D. or his assistant interacts with the program. The matrix will be expanded to include data from physical examination and laboratory for this mode of operation (5). The third mode of operation of the program is for teaching and testing medical students, house officers and other personnel. In this case a random number generator is used to produce a set of answers representing a simulated patient having one of the diseases chosen at random. Once the disease is chosen, this row of the matrix is scanned and each probability is compared to a new random number from 0 to 100. If the probability times 100 is greater than the random number, that symptom is present in the simulated patient. Thus, each simulated patient even with the same disease may have a different set of symptoms. These answers are stored for use as the simulation proceeds by presenting a chief complaint to the student and asking him for his differential diagnosis at each step and for his selection of questions to present to the patient.

In conclusion, a history program based on sequential estimates of conditional probability has been described. Although it is too early to be certain of its role in clinical medicine, it is apparent even with 134 diseases and 320 questions that much valuable information can be collected efficiently and structured around a differential diagnosis for the physician.

REFERENCES

1. Slack, W.V., Hicks, G.P., Reed, C.E., and VanCura, L.J. A computer-based medical history system. *New England J. Med*. **247**, 194 (1966).
2. Gorry, G.H. and Barnett, G.O. Experience with a model of sequential diagnosis. *Comp. Biomed. Res.* **1**, 490 (1968).
3. Pryor, T.A. and Warner, H.R. Time-sharing in biomedical research. *Datamation* 54, April (1966).
4. Warner, H.R. Experiences with computer-based patient monitoring. *Anes. And Analgesia Current Researchers* **47**, 453-461 (1968).
5. Bishop, C.R. and Warner, H.R. A mathematical approach to medical diagnosis. *Comp. Biomed. Res.* **2**, 486-493 (1969).