DXplain: Experience with Knowledge Acquisition and Program Evaluation

G. Octo Barnett, M.D., James J. Cimino, M.D., Jon A. Hupp, M.D., and Edward P. Hoffer, M.D.

Laboratory of Computer Science
Massachusetts General Hospital
Boston, Massachusetts 02114

DXplain is a program which provides access to a medical diagnosis knowledge base via a nation-wide computer communications network. This report describes the methods used in the knowledge acquisition phase of development and the initial evaluations of the program.

Introduction

The prospect of using computers to assist physicians in the process of medical diagnosis has generated a great deal of interest among researchers and the public[1]. "Differential diagnosis" involves two competing operations: selecting a list of diseases that is broad enough to include all reasonable possibilities and narrowing the list to focus on the most likely candidates. Previous efforts have ranged from listing diseases related to an isolated laboratory abnormality[2] to attempting to provide definitive diagnoses[3]. The experiences of these previous works have identified a number of obstacles related to knowledge acquisition[4], program acceptance[5], and program evaluation[6,7,8,9] which have affected the practical use of these approaches.

DXplain is a program developed at the Massachusetts General Hospital Laboratory of Computer Science which takes a slightly different approach to the problem of computer assisted medical diagnosis. The program has been described previously[10,11]; a brief overview of the purpose and design of the system will be given here. This report describes our experience with compilation of the knowledge base and the means by which we are evaluating the system.

Overview of DXplain

DXplain provides clinicians access to a medical diagnosis knowledge base which is in the form of relationships between clinical terms (history, symptoms, physical findings, and laboratory data) and diagnoses. Access to this information is provided by a user-friendly interface which accepts a patient description from the user, translates the terminology into its controlled vocabulary of clinical features, generates a list of disorders having some of the features entered by the user, and provides tools for further exploration of the knowledge base. In this way, DXplain assists the clinician in two phases of the process of differential diagnosis. It reminds the user of disorders which may have been overlooked that might explain the patient's condition, thus helping to broaden the differential diagnosis. It then provides additional information helpful to the diagnostic process that aids the user in determining which diseases deserve further consideration, thereby allowing more precise focusing on appropriate possibilities.

In designing DXplain, we considered the barriers encountered by previous efforts to assist the diagnostic process. We have attempted to minimize these obstacles by appropriate selection of design objectives in the domain of the knowledge base, by choosing relatively narrow objectives for the program, and by providing a flexible user interface.

Knowledge Base Domain

Although a narrowly bounded problem domain (such as the set of diseases within a medical subspecialty) has appeal for comprehensive knowledge compilation and extraction of expert algorithms, prior efforts along these lines have met with limited enthusiasm due to their limited applications. Therefore, we have selected the realm of general internal medicine and included those disorders which might be encountered in primary medical care. We have concentrated on information available to physicians in outpatient settings. Many advanced tests (e.g., CT scans, cardiac catheterization, specialized laboratory analyses, etc.) are not represented in the knowledge base.

Purpose of DXplain

Past efforts in computer-based medical diagnosis, such as Internist-I[3], have demonstrated the difficulty of reliably generating the appropriate combination of diagnoses which can uniquely explain the patient's clinical findings. Our design objectives for DXplain did not focus on finding the "correct" diagnosis. We felt that such an ambitious goal would be impossible to achieve reliably in a large problem domain. We reasoned that the clinician will always a more complete picture of the patient than could be related to a computer program and that a more realistic goal would be to furnish a list of diagnostic possibilities. Indeed, the focus on a limited subset of diagnostic data (available at the initial work-up) will often preclude the possibility of the program having sufficient knowledge to make a definitive diagnosis. Thus, the design goal of DXplain is to provide a list of diseases which could explain some (or all) of the features of a clinical case. The user may then consider which diseases are appropriate to the case by applying common sense, clinical experience, full knowledge about the patient, and by obtaining further information from the program. DXplain can provide disease descriptions from the knowledge base, display its reasoning used in formulating a plausible diagnosis list, and textual descriptions from an on-line version of *Current Medical Information and Terminology*[12] (CMIT). The textual descriptions provide information not covered by DXplain's knowledge base, such as temporal aspects of disease processes, etiology, and pathology.

0195-4210/0000/0150$01.00 © 1987 SCAMC, Inc.

Style of User Interaction

One of the most important design objectives in a diagnostic decision-support system for physicians should be that access to the knowledge base be easy and rapid. We have given high priority to optimizing the pragmatic issues of user interface, response time, and program availability[11].

From the user's viewpoint, finding the appropriate terminology is one of the most difficult tasks involved in entering pertinent information into a computer-based reference source. Therefore, a great deal of our work has concentrated on the methods used to assist in translating clinical findings into the appropriate terms in the program's controlled vocabulary. We do not encourage the user to undertake to enter an exhaustive list of patient findings (particularly the usual large number of "pertinent negatives"). Once the user has entered a partial description of the patient, the "Question" mode may be entered whereby the program requests information about the patient. This provides a rapid means for extending the case description, and minimizes the time required by the user casting about for terms that the program will recognize.

Several techniques are used to optimize response time. For example, rather than attempting to consider all diseases in a case, heuristics enable consideration to be given only to the most appropriate subset of diseases. Because the knowledge base is accessed for many different purposes (disease selection, disease scoring, explanation of reasoning, etc.), no single internal data structure was found to be ideal for rapid information retrieval. Through the use of multiple representations, we have chosen to sacrifice storage economy in order to optimize response time. Thus far, we have not given serious consideration to natural language processing, temporal reasoning or pathophysiologic modeling of diseases.

In order to maximize the availability of DXplain and to allow us to make rapid modifications of the knowledge base, we have chosen to support the program on a central computer and provide access over two national computer networks: the MGH continuing education network (which is available to medical schools and teaching hospitals) and the American Medical Association's AMA/NET. Since to access the program by telephone adds little more time than if DXplain were executing on a personal computer, we do not believe this to be an impediment to use. In fact, we believe that access will be enhanced, since distribution is simplified and the users' hardware requirements are minimized. We are able to make updates instantly available to all users as both the program and the knowledge base undergo improvement and expansion. The use of the system over the past year supports our choice of this mode of distribution: thousands of clinical cases have been entered by physicians from across the continental United States, Alaska, Canada, and Japan.

Knowledge Acquisition

DXplain's knowledge base consists of descriptions of over 2000 diseases, over 4700 patient descriptors (signs, symptoms, etc.), and some 70 000 relationships among them. Each association between a descriptor and a disease includes the frequency with which one sees that feature in the disease (term frequency) and also the degree to which that feature suggests the disease (evoking strength). Compiling this information involved the work of thirteen board-certified internists. At present, five internists are participating in ongoing testing and enhancement of the knowledge base.

Diseases and Clinical Terms

Initially, medical diagnoses were selected for inclusion in DXplain from among the 3262 disorders in CMIT. Where appropriate, some conditions were combined to form more inclusive definitions, while others were subdivided into more specific forms of the disease (to reflect differences in stages of disease and etiologic agents). Additional diseases were included from other sources when deficiencies in CMIT were noted.

Clinical features are represented in a controlled vocabulary contained in the Term Directory. A list of some 6000 words and phrases was initially compiled; we then extensively reviewed and edited this list to give coherence to the vocabulary. Synonymous terms were merged, and similar terms were linked together in a hierarchical manner to express their relationships (usually in terms of such qualifiers as severity, duration, or anatomic location). The outcome was a Term Directory consisting of 4000 descriptors. As disease descriptions were compiled, 700 additional concepts were added to the Directory.

Initial Collection of Disease-Term Links

The generation of the knowledge base began with a three-step process to link appropriate terms to each disorder, thereby producing the disease descriptions. First, a list was compiled of all the terms that had some role in either supporting or ruling out a particular disease. Next, medical texts were reviewed to determine the frequency of each term in the disease. When standard references proved inadequate for recent discoveries or obscure conditions, literature searches were carried out. Finally, an estimate of the potential for the presence or absence of each term to evoke or refute the diagnosis was made.

Extensive discussions were required to reach a consensus on content and assignment of evoking potentials by the internists recruited to author the disease descriptions. Once guidelines were established, authors were trained in filling out worksheets for each disease. These forms included all terms which had been previously assigned to each disease, a standard set of demographic terms (age, sex, duration of symptoms), and a generous number of blank lines for writing in new terms. After the worksheets were completed, one of us (JJC) reviewed them to ensure consistency of content, terminology, and assignment of evoking potentials.

Problems in the Initial Knowledge Acquisition

It became evident that as the knowledge base grew the assignment of the evoking strengths became more difficult. Inconsistent use of evoking strengths occurred, not only from author to author, but among diseases described by individual authors. The problem was one of perspective, whereby the focus on a particular disease caused the author to lose sight of the general implications of a clinical finding - a "forest for the trees" problem. After spending some time reading about a particular disease, an author tended to assign term evoking strengths that favored the disease, without considering how strongly the term might evoke other diseases. The author's disease-oriented perspective, useful for establishing term frequency, had to be balanced with a term-oriented perspective which is more appropriate for assigning the evoking strengths.

An additional problem which quickly became apparent was the under-utilization of individual terms. This occurred because no author could hope to be facile with the entire Term Directory. For example, in describing a disease which has arthritis as a component, an author might include terms such as "arthritis", "joint effusion", "joint swelling", "joint tenderness", "joint stiffness" and "joint heat", and might easily overlook "joint erythema".

Evaluation of DXplain

The evaluation of DXplain has focused on assessing whether we have achieved a system that is easy to use, provides information helpful in making differential diagnoses, contains accurate and comprehensive medical knowledge, and is both acceptable and and of clinical usefulness to the intended users. DXplain's performance was initially evaluated using a variety of clinical cases. Random medical records were abstracted and entered into the system. We also recruited dozens of physicians from around the country for a ß-test site evaluation and provided them with unlimited access to the program to enter their own cases. The users were automatically queried after each case to ascertain their reactions and the resulting sessions were reviewed. Additional user responses were obtained through the use of a written questionnaire.

Evaluation of Selected Medical Cases

We have undertaken to use DXplain in several series of clinical cases chosen because they were considered to be diagnostic problems. We used this case method since we could conceive of no protocol which would provide an exhaustive evaluation of every possible combination of terms and relations of the terms to the different diseases. Also, a static evaluation of the diagnostic performance of the knowledge base has been impossible since we are continually adding terms and diseases and modifying their relationship.

The initial results which we have obtained using abstracts from clinical records are encouraging, both in terms of the ease and completeness with which it is possible to enter clinical manifestations, and in terms of the plausibility of the disease lists generated by DXplain. Our impression is that the knowledge base contained in DXplain and the algorithms which are used in the diagnostic selection process are sufficiently promising to justify the initial distribution of the system as an educational and decision-support resource.

ß-Test Evaluation

During a two month period early in 1987, forty-three physicians from all regions of the continental United States, Alaska, Canada and Japan entered 303 patients into the system. The interactions were automatically recorded in session logs and carefully reviewed. The users attempted to enter 2244 clinical findings, of which 1837 were accepted by the system (82%). Of the remaining 407 terms, an additional 142 findings (6%) were accepted after the user reworded the term. On average, 6.5 features were entered per patient (not including the age, sex and duration of symptoms, which are automatically requested on each patient).

Of the remaining 265 findings (12%) that were not understood by the system, ninety-two of these (4%) were absent from the Term Directory because they dealt with laboratory tests not presently included in the DXplain vocabulary. Most of the other 173 clinical findings (8%) were synonymous with DXplain terms, but weren't identified by the program. The users also entered findings by using the DXplain "Question" mode in about half the sessions, entering 1111 additional clinical features, or an average of 7.4 terms per case in those cases where this mode was used. Further review of session logs showed that the educational features provided by DXplain were frequently used. Users requested a display of the program's reasoning concerning 235 diagnoses (0.8 per case). They accessed the on-line textbook CMIT a total of 118 times (0.4 per case).

We asked users to enter the diagnosis or diagnoses they were considering prior to entering the cases, in order to determine the effect of DXplain on the user's differential diagnosis. The users entered a total of 268 diagnoses. The disease lists subsequently generated by DXplain included 120 (45%) of the original diagnoses. While the precise reason for the absence of the remaining 148 diseases was difficult to determine, it appears that they were not included because of one of three conditions, each of which accounted for about one third of the diagnoses: the user did not enter terms about the patient which would have led to considering the diagnosis, the knowledge base had some deficiency which prevented the disease from being considered, or the user entered a term which effectively (and appropriately) ruled out the diagnosis. The first cause could theoretically be rectified by requiring the users to enter more complete descriptions. The second can be addressed by continuing our efforts to improve the knowledge base. The third is not a deficiency; excluding diagnoses, where appropriate, is part of the function of the program.

Further information about the DXplain's effect on differential diagnosis was obtained by automatically questioning of the users immediately after the entry of each case. The users stated that following the use of DXplain, 49 diagnoses were added to their differential diagnoses by using DXplain. The users felt that DXplain *should* have mentioned an additional 39 diseases. We asked the users "Did using DXplain change your differential diagnosis?" after each case. They responded to this question 61 times: 41 with "No" and 20 with "Yes".

ß-Test Users' Questionnaire

A written questionnaire was compiled to assess user attitudes towards six aspects of DXplain: program accessibility, ease of use, knowledge of diseases, diagnostic reasoning, the ability to justify that reasoning and overall usefulness. A list was compiled of all ß-test site users during the first 8 months of the test period. Questionnaires were sent to the ninety-five physicians who had used the system for at least ten minutes. Thirty four of them responded. Forty-five of those who did not respond were casual users in medical centers where access to the program was made available to a large number of house officers.

Seven (21%) of the respondents were house officers, two (6%) were fellows and twenty-five (73%) had completed their formal medical training. Twelve (35%) of the respondents stated that they had extensive previous computer experience, seventeen (50%) described themselves as moderately experienced, four (12%) had little experience and one (3%) had no previous experience. Twenty-seven (80%) of the respondents stated that they found it very easy to access DXplain, despite the fact that only nine (26%) had easy access to the program from a patient care area. The amount of DXplain usage by each respondent varied: three (9%) had only used it once, ten (29%) had used it a few times, and (62%) twenty-two had used it enough to feel comfortable with it. The results of the questions were independent of medical background, computer experience, access to computer terminals and DXplain usage.

Two thirds of respondents stated that entry of patient findings was easy, with the remainder saying it was "somewhat difficult". None felt it was "very difficult". All respondents felt that the system understood most or all of their terms and that they were able to obtain results by entering a few or a moderate number of terms, although some felt that term entry could become tedious.

The "Question" mode was the most popular of the other program features, with fifteen respondents stating that they liked it and only two expressing dislike. Of the fourteen responses regarding program speed, ten liked the response time while four felt it was too slow.

When evaluating DXplain's disease list, twenty-two of the users felt that the program occasionally failed to include some important diseases. The disease lists were generally described as complete, with many extraneous diseases. Only three respondents felt that the number of irrelevant diseases made it difficult to discern which diseases should be considered plausible candidates for inclusion in the differential diagnosis.

Dxplain's general knowledge of disease was felt by six users to be complete, "pretty good" (few errors and omissions) by twenty-one, and missing much information or wrong by six. Most users felt that the program was knowledgeable about symptoms and physical findings, but less so about history and laboratory information. Most users felt that the questions asked in the "Question" mode were appropriate. The most commonly listed "strong" points in DXplain's knowledge were that it knew about many diseases and included appropriate diseases on its lists. The most commonly listed weak point was that the disease lists included many diagnoses which did not belong on the list.

DXplain was useful in expanding the differential diagnoses for twenty-nine of the users. Only one user stated that DXplain did not add diseases to the differential diagnosis. On the other hand, twenty respondents felt that DXplain did not help in excluding diagnoses, while nine found that it did. In general, thirteen users believed that the program helped inform them about diseases, while eight did not. Of the twelve who used the on-line text book (CMIT), all but one found it informative. Overall, DXplain was felt to be useful in many or most cases by four respondents and useful in some cases by twenty-three; five found it of minimal usefulness, but none stated that it was of no usefulness.

DXplain's reasoning was judged as good when adding diagnoses, but poor when excluding diagnoses. Eighteen felt that the reasoning was explained clearly, while two felt that it was not. Overall, DXplain's reasoning was judged as poor by one respondent, fair by five, good by fifteen, very good by twelve, and excellent by none.

Discussion

We have encountered three types of problems in the generation of the knowledge base. First, it was difficult to maintain consistency in assigning evoking strengths for particular terms. Another problem was the under-utilization of some terms, resulting in the absence of some term-disease links which should have been included. For example, diseases which included jaundice in their descriptions should also include elevation of bilirubin. Lastly, for those terms which are mutually exclusive, the term frequencies should sum up to unity. For example, a disease cannot occur usually in males *and* usually in females. These patterns of knowledge base deficiencies lend themselves to systematic efforts for discovering and eliminating them. To a large measure, the failure of DXplain to identify diseases that deserve consideration has been due to a simple lack of information in its knowledge base. While we can never hope to represent every manifestation of every disease, we believe that there is still much we can reasonably expect to accomplish by continuing our efforts at knowledge acquisition.

The initial evaluation by outside users indicates that providing the program via a nationwide computer network appears to be a reasonable choice. Both the extensive use by ß-testers and their opinions on the written questionnaire indicate that this access method is practical and does not hinder program acceptance.

The user interface has been generally successful in allowing the untrained user to enter clinical information. With only a three page user's guide and without personalized instruction, users were able to enter the majority of their patient descriptions and make use of the other features of the program. The impression that term entry was easily accomplished was initially derived from the session logs and was borne out by the written questionnaire. In addition, the "Question" mode appears to be an efficient means for adding clinical features to patient descriptions.

The patterns of knowledge base flaws we identified during the knowledge acquisition phase were confirmed during testing. The major problems were failure to include some disease features and incomplete history and laboratory terms. The evaluations showed that DXplain most frequently erred on the side of including too many diseases when generating disease lists. Since a design goal of DXplain was to suggest diseases which might be only partially supported by the evidence, we were not surprised by the tendency to include diseases on the list which users might find inappropriate.

We are most encouraged by the users' impressions that DXplain was useful for adding diseases to their differential diagnosis lists and giving them additional information about diseases they were considering.

Finally, in the ß-test site use, there were numerous occasions where DXplain was unable to focus on (or even include) what the users considered to be the leading diagnostic possibility. To a certain extent, this was due to the program's lack of representation of pathophysiologic processes. Since DXplain cannot adequately determine diagnostically important characteristics such as temporal and causal relationships between between terms and diseases, it sometimes did not include a disease which the user felt was a likely diagnostic possibility. This was similar to the experience with the Internist-I program[3]. We expect that physicians will continue to rely on their own abilities and use DXplain to supplement their clinical judgment with DXplain's broad knowledge base.

DXplain has recently been released on the American Medical Association's AMA/NET, a nationwide computer network that will provide access via a local phone call to any subscriber with a terminal and modem. This is the first time that a program offering assistance with medical diagnosis has been commercially available without expensive hardware and software purchases. Since the release of DXplain, the use of the program and the amount of feedback have begun to increase. This expanding use provides useful information about the system's strengths and weaknesses which will further our evaluation efforts.

Conclusion

The evaluation of the level of performance of any computer-based medical decision-support system is very difficult. The difficulty is compounded when the knowledge base is as large and complex as that used by DXplain. We have established a system for collection and synthesis of medical knowledge which is continually being extended to improve program performance. Both the program and the knowledge base are centralized and are readily updated for all users simultaneously. Our initial evaluations of DXplain have identified some deficiencies in the program and the knowledge base; however, we believe the problems to be largely remedial. The evaluations we have performed have given us only a crude estimate of the actual educational value of the DXplain disease list. At present, we have no way of measuring how critically our users examined the results of DXplain's diagnoses lists nor of the impact of using the program in their clinical practices. Nevertheless, our preliminary evaluations have been encouraging with regard to both user acceptance and program performance and have served to establish where we should concentrate our next efforts. The challenge will now be to respond to issues of method and content as they arise, and to continue to improve system performance.

Acknowledgements

This project has been developed with the support of the American Medical Association. Dr. Cimino is supported by National Library of Medicine training grant LM07037-03. We are indebted to Dr. David Cooks for his assistance in the preliminary evaluation effort, to Ian Kirkwood for assisting with the knowledge base audits, to all of the internists who authored the diseases descriptions, and to Kathy Famiglietti for invaluable help with the telecommunications and ß-test accounts.

References

[1] Stengel R. Calling Dr. SUMEX - The diagnostician's new colleague needs no coffee breaks. Time. May 17, 1982; 119:71.

[2] McDonald CJ: Profile: Charles Meader and his Differential Diagnosis Machine. Medcomp. 1983; 1(1):30-33,72.

[3] Miller RA, Pople HE, Myers JD. INTERNIST-I, An Experimental Computer-Based Consultant for General Internal Medicine. New England Journal of Medicine. 1982; 307(8):468-476.

[4] Davis R. Consultation, Knowledge Acquisition, and Instruction: A Case Study. In: Szolovits P, ed. Artificial Intelligence in Medicine. AAAS Symposium Series, Boulder, Colo.: Westview Press. 1982; Chapter 3:57-78.

[5] Teach RL, Shortliffe EH. An Analysis of Physicians' Attitudes. In: Buchanan BG, Shortliffe EH, ed. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, Mass.: Addison-Wesley, 1984; Chapter 34:635-652.

[6] Reggia JA. Evaluation of Medical Expert Systems: A Case Study in Performance Assessment. Symposium of Computer Applications in Medical Care. Baltimore, Maryland; 1985; 287-291.

[7] Kingsland LC. The Evaluation of Medical Expert Systems: Experience with the AI/RHEUM Knowledge-Based Consultant System in Rheumatology. Symposium of Computer Applications in Medical Care. Baltimore, Maryland; 1985; 292-295.

[8] Miller RA, McNeil MA, Challinor SM, Masarie FE, Meyers JD. The INTERNIST-I/Quick Medical Reference Project - Status Report. Western Journal of Medicine. 1986;145(6):816-822.

[9] Yu VL, Fagan LM, Wraith SM, Clancey WJ, Scott AC, Hannigan J, Blum RL, Buchanan BG, Cohen SN: Antimicrobial selection by a computer - A blinded evaluation by infectious diseases experts. Journal of the American Medical Association. 1979; 242(12):1279-1282.

[10] Hupp JA, Cimino JJ, Hoffer EP, Lowe HJ, Barnett GO. DXplain - A computer-based Diagnostic Knowledge Base. In: Salamon R, Blum B, Jorgensen M, ed.: MEDINFO 86. Amsterdam: Elsevier Science (North Holland), 1986; 117-121.

[11] Barnett GO, Cimino JJ, Hupp JA, Hoffer EP: DXplain- An Evolving Diagnostic decision-Support System. Journal of the American Medical Association. 1987; 258(1):67-74.

[12] Gordon B, ed. Current Medical Information and Terminology, Fourth Edition. Chicago: American Medical Association, 1971.