Computer Methods and Programs in Biomedicine, 35 (1991) 261–266

COMMET 01208

An approach to evaluating the accuracy of DXplain

Mitchell J. Feldman and G. Octo Barnett

Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA, U.S.A.

DXplain is a computer-based decision support system which generates a differential diagnosis (ddx) from a given list of clinical manifestations (Barnett et al., J. Am. Med. Assoc. 258 (1987) 67-74). An approach was developed to evaluate the accuracy of the ddx's produced by DXplain. The first step involves the collection of 65 benchmark cases drawn from a variety of sources and authors. Despite their diverse origins, the cases share in common that they are all clinical cases upon which a consulting physician might be asked to produce a differential. This helps to ensure that the evaluation of the system will be done in an environment similar to that in which the system is actually used. In the second step, all cases are reviewed by five board-certified physicians (experts) as well as DXplain. For each case, the evaluators (experts and DXplain) produce a rank-ordered ddx list along with an indication of how strongly each disease was felt to be supported by the case findings. A scoring technique was devised which rewards concordance with the gold standard: a consensus of the evaluators' ddx lists. Each evaluator receives a score which is proportional to the degree of agreement achieved with the consensus on the ddx submitted. Preliminary results on a trial evaluation of 46 cases indicate that DXplain, on average, did well in agreeing with the consensus. Agreement was achieved both in regard to the specific diagnoses listed in the ddx and the degree to which the diseases were felt to be supported by the case findings. A discussion of some important issues in the evaluation of knowledge-based systems is undertaken.

Evaluation of Knowledge-Based Systems

1. Background

DXplain's knowledge base includes 2100 disease profiles. Each profile consists of from 10 to 200 terms (average 35). A term is a descriptor which gives information about the patient; the knowledge base contains 4500 terms, including demographic, historical, physical and laboratory findings. Each term has two disease-dependent attributes, Term Frequency (TF) and Evoking Strength (ES), and a disease-independent Term Importance (TI). TF is a measure of how frequently the finding occurs in the particular disease. ES indicates how strongly one should think of the particular disease given the presence of the finding. A high TI is given to findings that can be identified with high reliability or are rarely found in healthy people and, therefore, should be explained by some disease within the differential diagnosis [1]. The disease profiles and their associated TF, ES and TI values comprise the bulk of the DXplain knowledge base.

DXplain determines its list of plausible diseases, once a set of terms has been entered, by picking those diagnoses which best explain the case terms [2]. A score is given to a disease based on the ES and TI values of those case terms found in the disease profile. Diseases are displayed to the user in rank order of descending scores. For each disease in the rank-ordered list

Correspondence: M.S. Feldman, Laboratory of Computer Science, Massachusetts General Hospital, 50 Staniford Street, 5th Floor, Boston, MA 02114, U.S.A.

DXplain also provides an indication of how well that disease is supported by the case findings. Two plus signs (+ +) are listed next to a diagnosis when it is felt to be strongly supported and one plus sign (+) when it is partially supported. This degree of support for a disease is based on the disease score.

2. Methods

The first component of this evaluation process involves the assembling of a set of benchmark cases to be used as input to both the expert physician evaluators and to DXplain. Three sources for cases were chosen. The first source, from which we are collecting 30% of the cases, is the current DXplain user community. AMANET is a nationwide computer communications network sponsored by the American Medical Association through which the majority of users access DXplain. Interesting cases where actual patient descriptions were entered via AMANET will be used. The second source from which we are selecting 60% of the cases is physicians practising in a variety of settings from across the country. Each physician who participates submits four cases: two classic "textbook" cases and two describing clinical situations where the differential was not straightforward. The third source for the remaining 10% of the cases comes from abstractions of clinicopathologic conferences (CPCs) from the New England Journal of Medicine. These are taken from issues at least several years old so physician evaluators will not likely be familiar with them.

All cases consist of from eight to 20 individual terms. This range was chosen in part based on the experience that the average number of terms entered per case during previous Beta-testing of DXplain was ten [3]. The second component of the evaluation is having each case reviewed by five board-certified physicians, each of whom produces a rank-ordered ddx along with an indication of how well each disease is supported by the case findings. When formulating their differential lists, physician evaluators are free to use any textbooks, journals or other references that they might ordinarily use in a hospital or office-based consultation. They are, however, asked not to use DXplain or any other decision support computer program. The identical case terms are entered into DXplain and the resulting ddx's are compared. Each of the five physicians and DXplain is compared to the gold standard, which is a consensus list of diagnoses culled from the five physician evaluators and DXplain.

Since the physician evaluators and DXplain may use different terminologies when submitting the ddx lists, it is occasionally necessary to decide whether different disease names on two separate lists are in fact the same entity, e.g., Crohn disease, terminal ileitis and regional enteritis. Synonyms are always reconciled to be the same entity. The following guidelines are also used to help decide when to merge two disease names:

(1) Two diseases can be merged together if they describe the same or very similar basic pathophysiologic processes and neither disease is supported more than the other by the case terms.

(2) Two diseases can be merged together if they add no more valuable information to the differential when apart than when merged.

(3) If an evaluator puts down a specific dx, then any parent dx is implicit, e.g., 'diabetic neuropathy' implies also 'diabetes mellitus' with the same number of plus signs and rank order.

The gold standard consensus list contains a maximum of 15 diseases and is arrived at as follows:

(1) A disease must be listed by at least two evaluators to appear on the list.

(2) Diseases listed by multiple evaluators with + + are added, followed by diseases listed by multiple evaluators with + followed by diseases listed by multiple evaluators without plus signs.

(3) Diseases are added to the list according to criteria #1 and #2 until a total of 15 are reached or no diseases remain on any evaluator's list.

2.1. Scoring algorithm

Each physician and DXplain is given a score for each case. The score is based on both the number of diseases common to the list that is being scored and the consensus list, and the

		x # of other evaluators who put ++	x # of other evaluators who put +	x # of other evaluators who put 0+
If individual being scored places a	++	6	4	2
	+	4	5	3
	0+	2	3	4

Fig. 1. DXplain scoring system.

agreement about the number of plus signs. The score is computed as follows. For each diagnosis listed by the individual (DXplain or physician) which was also on the consensus list, the individual receives points as illustrated by Fig. 1. An individual receives no points for a disease listed which is not on the consensus list. The sum of the points received for all the diagnoses common to the individual's list and the consensus list is the individual's score for the case. The scoring mechanism rewards an individual's concordance with the group of evaluators. An individual is given more points for listing the number of plus signs with a disease that is the same as or closest to the number of plus signs listed by the other evaluators. As an example, if, of the six evaluators (five physicians and DXplain), five list the same disease, beriberi, with 0 + and the individual being scored (a physician or DXplain) lists this disease with + +, then this individual will receive 10 points (from Fig. 1, $2 \times 5 = 10$); however, if the individual had listed the disease with a 0 +, the points received would have been $4 \times 5 = 20$. The highest score will be obtained by the individual who agrees the most with the consensus list in terms of both the actual diagnoses on the list and the degree of support assigned.

3. Preliminary results

A trial evaluation using the main ideas of the outlined approach was completed prior to embarking on the present evaluation currently in progress. 46 cases, including classic cases, NEJM CPC abstractions and cases submitted by MDs were used. Each evaluator's ddx's were compared to the consensus ddx's culled from the physicians and DXplain. The scores achieved by DXplain were, on average, comparable to the scores achieved by the individual physicians. DXplain scored within one standard deviation of the mean of the experts' scores in 27 of 46 cases, above one standard deviation in 14 cases and below one standard deviation in five cases.

4. Discussion

Previous evaluations of knowledge-based systems in medical diagnosis have focused on the ability of the system to diagnose accurately *the* illness or illnesses which a patient has [3–8]. Since the goal of DXplain is to provide a list of plausible diagnostic hypotheses, it seems reasonable that the gold standard upon which to judge this output also be a list of diseases. We feel that using a consensus of evaluators, in this instance board-certified physicians and DXplain, will help to maximize the authenticity and validity of the diagnoses comprising the consensus and minimize the extraneous 'zebras.' Another issue which a consensus approach addresses is that of reliability. There is no guarantee that a ddx list from a particular physician based on a set of case terms will be the same if repeated over time. It often occurs that the diagnoses of which a physician thinks can be influenced by recent journal articles

read and patients encountered. Using a gold standard based on the consensus of evaluators would help to minimize these potential biases. The reliability of consensus ddx lists over time might be a useful issue for future study. By comparing the results of the ddx lists generated by DXplain to the consensus lists upon cases from various sources ranging from CPCs to classic cases, we hope to observe potential differences in the performance of DXplain in diverse clinical environments.

We recognize that it is artificially constraining to ask a physician for a list of possible diagnoses based on a set of 8–20 case terms. In clinical practice, a physician will always be able to obtain more information through history, physical examination and laboratory data. For the purposes of our evaluation however, it is imperative that the physician experts and DXplain see the same input.

One may argue that an individual who lists the 'correct' diagnosis (in a case where this is reported to be known) may not be assigned as many points as an individual who lists a more popular diagnosis listed by more of the evaluators, even though the latter diagnosis is not the 'correct' answer. We recognize there can be legitimate criticism of this protocol; however, we believe there is more justification to reward agreement with consensus than identification of the 'correct diagnosis.' There are a number of factors which led us to choose this reasoning.

The primary purpose of DXplain is not to make a single diagnosis but rather to suggest a list of plausible alternatives. When analyzing a set of discrete case terms, it is difficult to obtain as complete a picture as is achievable by talking with and examining a patient. In addition, the art of diagnosis in a broad domain such as internal medicine or pediatrics often necessitates the integration of information obtained from the sensory modalities of sight, hearing, touch and smell. Hence a list of diseases which explain the case findings or some subset thereof is a more realistic and attainable 'deliverable' for a knowledge-based system than a single diagnosis purported to be the 'absolute truth'.

At the time DXplain is likely to be used in the clinical setting, the 'correct diagnosis' may not be known, since further information may need to be accumulated. In such circumstances a differential diagnosis would be more useful than one disease name. Of particular importance is that the suggested disease list may stimulate consideration of additional diagnoses which the patient may manifest.

The importance of suggesting a list of diagnoses to be considered has been demonstrated in two other studies. In the AI/Rheum experience described by Kingsland [17] a list proposed by the system occasionally included a later confirmed secondary diagnosis which the clinician had failed to pursue after having assigned the primary diagnosis. In an evaluation of QMR in the setting of assisting the ward team at an academic medical center [8], a QMR-suggested diagnosis was added by the ward team to its initial ddx which had been generated prior to the computer-assisted consult in 14 of 31 cases.

Gasching et al. describe a checklist of pitfalls which they advise avoiding when evaluating expert systems [9]. It is instructive to review and address the applicable pitfalls:

- ☐ Evaluators may fail to clarify what is being evaluated.
  We are evaluating the list of plausible diagnostic hypotheses produced by DXplain.
- ☐ Evaluators may fail to clarify for whom the evaluation is intended.
  The evaluation is intended for the DXplain user community.
- ☐ Preselected cases may bias the results or their interpretation by narrowing the scope of problems with which the system is asked to deal.
  We propose a collection of cases spanning a broad spectrum of clinical medicine, from classic cases to less straightforward cases and CPCs. By obtaining cases from a diverse geographic distribution we hope to further broaden the range of material.
- ☐ Evaluators may fail to select an appropriate standard against which to compare the performance of the expert system.
  Our reasoning for using the gold standard mentioned is justified in the above discussion.
- ☐ Evaluators tend to overgeneralize from results

obtained in a highly constrained environment. We will attempt to ensure a challenging environment by selecting an assortment of case types from a variety of sources. We will also select evaluators from a wide geographic area.

- ☐ The evaluation may be inappropriate for the stage of development (because premature or entwined with conflicting goals of the funding agency and the researcher).
  DXplain has been in use since 1986, though it has evolved considerably since that time. A formal evaluation such as the one described has not yet been undertaken. Though a few previous efforts have looked at DXplain's output [2,10], none has compared its list of diagnoses with a consensus list incorporating the differential diagnoses of board-certified experts on a variety of case sources.
- ☐ There are often inherent difficulties in designing elegant tests: these may not have been appreciated and dealt with adequately in the study design.
  We have carried out considerable discussion, benchmarking and preliminary trials with different evaluation strategies, and feel that the present model of evaluation including the scoring protocol is appropriate.
- ☐ A formal evaluation may be a misallocation of scarce resources needed for other project activities (such as funds, computer time and work time of research staff, evaluators and those supplying input data).
  While this is a potential danger, we are persuaded that it is important the users of DXplain be provided some level of formal evaluation to enable them to decide the degree of confidence to be given to the results.

6. Summary

DXplain is a computer-based decision support system which provides a list of diagnostic hypotheses when presented with a set of case terms describing a patient. We propose that in the case of a diagnostic dilemma, when a knowledge-based system may be most useful, the ultimate diagnosis may not be knowable until further data become available. A list of plausible diseases which explain at least some of the case findings may therefore be more helpful. An approach is outlined to evaluate the accuracy of DXplain. A collection of benchmark cases from a variety of sources provides a diverse clinical environment for testing. A method for comparing the ddx of DXplain and expert physicians by rewarding concordance with the gold standard consensus of evaluators is described. Preliminary results indicate that the accuracy achieved by DXplain is, on average, comparable to that achieved by the physician evaluators.

Acknowledgements

The authors wish to express their thanks to the members of the DXplain editorial board for serving as physician evaluators during the preliminary in-house evaluation: Drs. Chris Cimino, Ed Hoffer, Marvin Packer, Peter Elkin, Bruce Forman, Diane Oliver, and Swati Bhave, Thanks also to Kathy Famiglietti and Richard Kim. This work was supported, in part, by an educational grant from the Hewlett Packard Corporation. M.J.F. Feldman is supported by NLM training grant [2-T15-LM07037-04]. We gratefully acknowledge the support of an educational grant from Paul Mongerson.

References

[1] G.O. Barnett, J.J. Cimino, J.A. Hupp, E.P. Hoffer, DXplain – An Evolving Diagnostic Decision-Support System. J. Am. Med. Assoc. 258 (1987) 67–74.

[2] J.A. Hupp, J.J. Cimino, E.P. Hoffer, N.J. Lowe, G.O. Barnett, DXplain- a computer based diagnostic knowledge base. Proceedings of Medinfo 86. pp. 117–21. Amsterdam: Elsevier (1986).

[3] G.O. Barnett, J.J. Cimino, J.A. Hupp, E.P. Hoffer, DXplain: Experience with Knowledge Acquisition and Program Evaluation. Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care. New York: pp. 150–154, (IEEE Computer Society Press, 1987).

[4] S.J. Nelson, M.S. Blois, M.S. Tuttle et al, Evaluating RECONSIDER. A computer program for diagnostic prompting. J. Med. Syst. 9 (1985) 379–88.

[5] J.A. Reggia, D.R. Tabb, T.R. Price, M. Banko, R. Hebel, Computer-aided assessment of transient ischemic attacks. A clinical evaluation. Arch. Neurol. 41 (1984) 1248–1254.

[6] R.A. Miller, H.E. Pople, J.D. Myers, INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. N. Engl. J. Med. 307 (1982) 468–476.

[7] L. Kingsland, The evaluation of medical expert systems: Experience with the AI/Rheum knowledge-based consultant system in rheumatology. Proceedings of the Ninth Annual Symposium on Computer Applications in Medical Care. pp. 292–295, (IEEE Computer Society Press; New York 1985).

[8] R.A. Bankowitz, M.A. McNeil, S.M. Challinor, R.C. Parker, W.N. Kapoor, R.A. Miller, A computer-assisted medical diagnostic consultation service. Ann. Intern. Med. 110 (1989) 824–832.

[9] J. Gaschnig, P. Klahr, H. Pople, E. Shortliffe, A. Terry, Evaluation of expert systems: issues and case studies, in Building Expert Systems, eds. F Hayes-Roth, DA Waterman, DB Lenat, pp. 241–280 (Addison-Wesley, Reading, 1983).

[10] J.R. Hammersley, K. Cooney, Evaluating the utility of available differential diagnosis systems. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. pp. 229–231, IEEE Computer Society Press New York, 1988).