Problems in Applying Expert System Technology to Radiographic Image Interpretation

David W. Piraino, Bradford J. Richmond, Masataka Uetani, Thomas Luetkehaus, Daniel Rockey, George Belhobek, Joe Armistead, and Fred Jones

A prototype expert system was developed to study the problems applying expert system technology to radiographic image interpretation. The Radiographic Image Interpretation System (RIIS) was developed on a microcomputer using Turbo Prolog, a low cost implementation of the prolog programming language. The present implementation of RIIS was developed to highlight potential problems in applying expert system technology in the evaluation of radiographic images. It was believed that the evaluation of this prototype expert system should include a large number of users unfamiliar with the program's use as this would probably be the case in clinical use of an image interpretation expert system. At present, the expert system deals with a limited domain of focal bony lesions. Twenty cases of pathologically proven bony lesions of varying difficulty were used to evaluate potential problems in the use of this expert system technology. RIIS, with the 20 sample cases, was presented as an exhibit at the 1987 Radiological Society of North America (RSNA) meeting to evaluate the potential problems with inexperienced users. These results were compared with those of experienced users. When a musculoskeletal radiologist, familiar with the programs use, provided the "proper description," the program averaged the correct diagnosis in the top five 80% of the time. During the program's use at the RSNA meeting, the program selected the correct diagnosis in the top five 22% of the time.

© 1989 by W.B. Saunders Company

KEY WORDS: Diagnosis, computer, expert system, bone tumor, image.

EXPERT system technology, an outgrowth of artificial intelligence research, has been applied in a wide variety of medical situations including radiographic differential diagnosis, hematologic disorders, evaluation of anemia, chemotherapy protocols, and the diagnosis of medical disorders.[1] Difficulties related to expert system application to radiographic image interpretation must address all the standard problems in applying expert system technology in any environment. In addition, radiographic differential diagnosis or image interpretation systems must address the problem of an appropriate method to input image information into the expert system.

Evaluation of an expert system is an important part of its development. Evaluation of expert systems, however, remain difficult and a standard evaluation process has not been developed. In general, a system may be evaluated on its accuracy, method of construction, and user impact.

A system's accuracy can be judged in relationship to a standard such as pathology. Since an expert system uses information and knowledge used by experts to arrive at a conclusion, its relative accuracy should be judged against and by other experts in its area of expertise.[2]

The techniques used to develop the expert system influence its user interface, methods of deduction, potential conclusions, and explanation of its conclusions. If a specific knowledge structure or user interface is chosen that is not appropriate for a expert system's environment, it will not be useful.

The goal of expert system technology is to produce a system that can be used by non-experts to help them arrive at conclusions at a level comparable to experts. It is therefore important, when evaluating expert systems, to test the user interface for its ease of use and user acceptance.[3] It is also necessary to evaluate whether the expert system helps the non-expert to make decisions at an accuracy level comparable with that of an expert. Finally, the non-expert must be confident that the decision he or she reaches with the help of an expert system is appropriate.

RIIS was developed on a personal computer to explore problems in implementing a prototype expert system for radiographic image interpretation. The study of the problems in implementing this prototype were directed primarily toward expert system construction, user interface, and

From the Cleveland Clinic Foundation, Tripler Army Medical Center, Honolulu and Henry Ford Hospital, Detroit.

Address reprint requests to David Piraino, MD, Department of Radiology, The Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, Ohio 44195-5021.

© 1989 by W.B. Saunders Company.

0897-1889/89/0201-0004$03.00/0

accuracy. The investigation evaluated users familiar and unfamiliar with the program.

RADIOGRAPHIC IMAGE INTERPRETATION SYSTEM

The Radiographic Image Interpretation System (RIIS) was developed to produce a differential diagnosis for focal musculoskeletal lesions. Focal bone lesions were chosen because this domain can be defined relatively easily and because several of the authors are musculoskeletal radiologists. Turbo Prolog, a fourth generation language for microcomputers, was used to construct the program. The program uses relative likelihoods and relative predictive values to produce a list of differential diagnostic possibilities.

Several psychological models have been developed to explain the process of image interpretation. Differences in psychological models deal primarily with whether there is a preliminary expectation of the observer that affects the processing of the visual information before the actual inputting of the image information (Table 1).[4]

The RIIS prototype expert system only deals with the last transformation of the aforementioned model from a conceptional understanding of the image to a decision about the most appropriate diagnosis related to that image. This decision to start with a language description of the image abnormalities places constraints on the expert system and imposes potential problems concerning image information input.

The user of RIIS selects the positive radiographic findings from a list of possibilities presented to the user by the program. Typical findings included in the initial implementation include bony matrix, chondroid matrix, bony expansion, geographic lesion, permeative pattern, and many other findings. The program considers any findings that are not selective as not being present in the image. An information base relating each diagnostic entity to the radiographic findings is contained within the program. All findings associated with any specific diagnosis such as osteogenic sarcoma are given a relative frequency from 1 to 5 and a relative predictive value from 1 to 5. A relative frequency of 1 indicates a very unusual finding associated with that disease while a 5 represents a finding that is always associated with that disease. A predictive value of 1 represents a completely nonspecific finding while a value of 5 represents a pathognomonic finding.[6] Relative frequencies and relative predictive values were assigned by the consensus of two musculoskeletal radiologists after reviewing standard radiology textbooks.

Table 1. Example of an Image Interpretation Psychological Model

I. Input: Various light levels projected on retina
II. Preprocess: Retina preprocesses information
III. Segmentation: Grouping of input information
IV. Understanding: Image groups recognized
V. Decision: Decision on image diagnosis*

*Data adopted from Kundel et al.[4]

Each diagnosis in the information base is compared with the findings input by the user. In the first level of evaluation, each diagnosis is compared with the positive image findings. Any positive finding that a specific diagnosis can explain, raises the relative likelihood of that diagnosis. After all diseases have been evaluated using this technique, a second evaluation takes place. In the second evaluation process, the findings that are commonly seen in a specific disease entity but were not described as being present on the image, decrease the overall likelihood of that specific diagnosis. Finally, a correction is made for findings described as being present in the image, but do not occur in the specific diagnosis being considered. In this case, since the diagnosis does not explain all the findings on the image, the overall likelihood of that specific diagnosis is decreased.

RIIS uses a simple inference engine to accomplish the above results. A simple consecutive search is performed on the information base comparing the radiographic description of the specific image with the radiographic descriptions of each disease. The relative likelihoods of each diagnosis is then calculated.

RIIS arranges the diagnoses in order of most likely first, least likely last, according to their calculated relative likelihoods. The relative likelihoods are compared and different numeric selection criteria are applied to the relative likelihood for appropriate diagnosis selection. Different selection criteria include: (1) considering only those whose relative value is >50% of the relative likelihood associated with the most likely

diagnosis, and (2) considering only those diagnoses above any point in which there is a 50% difference in the relative likelihood of any two adjacent diagnoses.

PROTOTYPE EVALUATION

RIIS was evaluated with experienced and nonexperienced users. Twenty cases of focal bony abnormalities in print form were used to evaluate the program. The cases were selected to represent a range of diagnoses and a range of difficulties from classic cases to atypical cases. The cases were selected from the Cleveland Clinic teaching file and from clinical practice.

The program along with the 20 cases in print form were presented as an exhibit at the 1987 RSNA meeting in Chicago. People viewing the exhibit were instructed how to use the RIIS system in print form and were encouraged to use the 20 sample cases as examples. A monitoring program was produced to retain statistics throughout the meeting to determine how well the users and the program performed for each case. The users entered their experience level, selected what they considered to be the appropriate differential diagnosis from a list of diagnostic possibilities, and entered the positive findings they identified on the radiograph. The program then produced its differential diagnosis, with associated relative likelihoods, which the user could compare with their differential diagnosis. Statistics were maintained for the program's use throughout the week as well as for each experience level.

Subsequently, a musculoskeletal radiologist involved in the development of the program and familiar with its syntax and use developed a "proper description" of the abnormalities on each of the 20 cases. While the musculoskeletal radiologist knew the pathologic diagnosis, the proper descriptions were developed independently of the program. All the descriptions were reviewed before input into the program and any findings that were considered questionable were removed to avoid bias. The proper descriptions were entered into the program and its performance was then evaluated.

The differential diagnosis included only those diagnoses above a cutoff point where the first 50% difference between any two adjacent diagnoses was found in the differential diagnostic list. This same selection was applied both at the 1987 RSNA meeting and when the proper description was provided.

RESULTS

The results of the program's use at the RSNA meeting, the program's results with a proper description, and the results for RSNA users are presented graphically in Fig 1. At the RSNA meeting, the program was used 268 times by people with varying experience levels. The correct diagnosis was selected in the program's differential diagnostic list in 22% of attempts. The correct diagnosis was selected first in 33 cases (12%); second in 16 cases (6%); third in six cases (2%); fourth in two cases (0.7%); and fifth in two cases (0.7%). The correct diagnosis was not listed in 209 attempts (78%).

When the proper description was entered by our musculoskeletal radiologist, the program

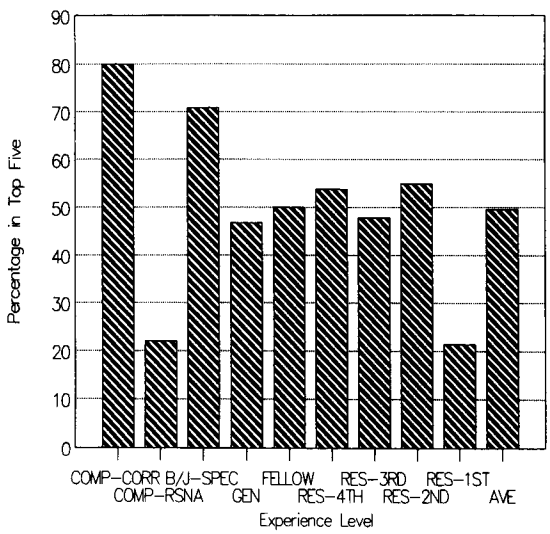

Fig 1. This figure graphically demonstrates the percentage by experience level in which the pathologic diagnosis was selected in the top five diagnoses. RIIS with a proper description (COMP-CORR) selected the pathologic diagnosis in the top five 80% of the time and RIIS at the 1987 RSNA (COMP-RSNA) selected the pathologic diagnosis 22% of the time. This compares with musculoskeletal specialists (B/J-SPEC) at RSNA who selected the correct diagnosis 71% of the time. The remaining bars demonstrate the percentage of attempts in which the diagnosis was listed in the top five for the remaining experience levels including general radiologists (GEN), fellows (FELLOW), fourth-year residents (RES-4TH), third-year residents (RES-3RD), second-year residents (RES-2ND), first-year residents (RES-1ST), and the average (AVE) for all experience levels.

selected the correct diagnosis in 16 of 20 cases (80%). In 11 of 20 (55%) of these cases, the correct diagnosis was listed as the most likely diagnosis and in five cases (25%), the correct diagnosis was listed as the second most likely.

By comparison, musculoskeletal radiologists at the 1987 RSNA meeting included the proper diagnosis in their differential diagnostic list in 34 of 48 attempts (71%). General radiologists included the correct diagnosis in their differential diagnostic list in 58 of 109 attempts (47%).

Statistics for individual cases at RSNA varied widely for the accuracy of both the program and the users. The best performance for the program at RSNA occurred in case no. 3, a chondrosarcoma, where the program listed the proper diagnosis as its most likely possibility in over 70% of attempts. This compares with musculoskeletal radiologists who listed chondrosarcoma 100% of the time as the most likely diagnosis but general radiologists listed chondrosarcoma only 30% of the time. The worse performance for the program at RSNA was case no 6 (Fig 2), a lytic osteogenic sarcoma. The correct diagnosis was never listed in the program's differential diagnostic list in 15 attempts. In the 15 attempts at RSNA, only one participant listed lytic osteogenic sarcoma in their differential diagnosis.

When the proper descriptions were entered, the program did not include the proper diagnosis in four cases. The first case was again the lytic osteogenic sarcoma. The program did not list a lytic osteogenic sarcoma as a possibility although it included osteogenic sarcoma, chondrosarcoma, metastatic disease, and Ewing's sarcoma in its differential diagnostic list. Another case in which the correct diagnosis was not included by the program when a proper description was input was a case of a parosteal osteogenic sarcoma involving a finger. The program listed a single differential diagnostic choice, a juxtacortical chondroma.

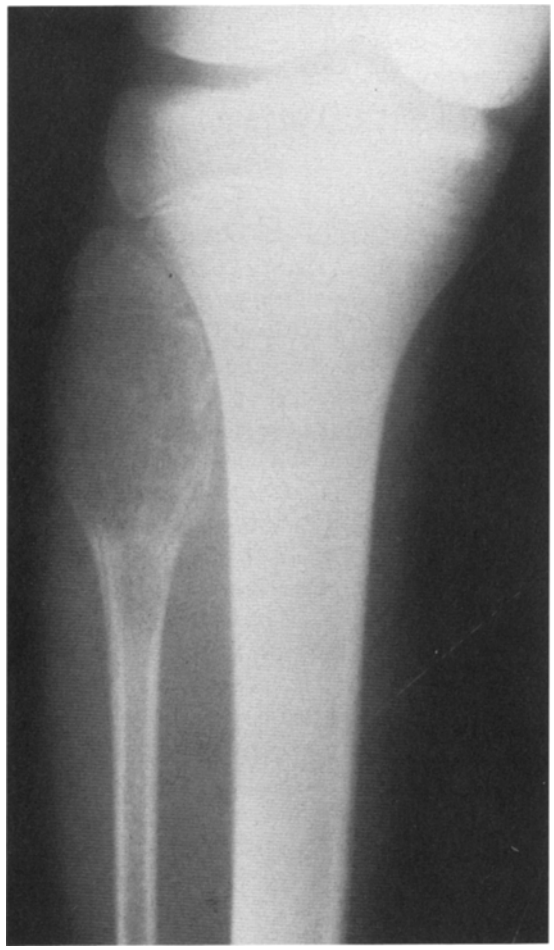

Fig 2. Representation of case no 6, a lytic osteogenic sarcoma in the proximal fibula. The correct diagnosis was listed only once out of 15 attempts during the RSNA meeting. RIIS never listed the diagnosis during the RSNA meeting. When the proper description was provided, RIIS's differential diagnosis was osteogenic sarcoma, chondrosarcoma, metastatic disease, and Ewing's sarcoma. However, the correct diagnosis of lytic osteogenic sarcoma that is considered separately by the RIIS system was not included.

DISCUSSION

The RIIS program was able to correctly diagnose the lesions in the 20 sample cases in 22% of attempts with inexperienced users. This accuracy level is significantly less than that of musculoskeletal experts. However, the system's accuracy improves remarkably with a user familiar with the program and its associated syntax and language. In fact, with an experienced user, RIIS performed at a level comparable with musculoskeletal radiologists. Also, the results from the RSNA meeting should not be strictly interpreted because (1) several of the users may have input descriptions slightly different from the sample cases to observe how the program performed in these instances and (2) 20 cases are a limited sample for evaluating the performance of an expert system.

The RIIS did perform worse on the average than the general radiology user at our 1987

RSNA meeting and much worse than a musculoskeletal specialist. There was marked improvement when the proper descriptions were provided by an experienced user. While standard descriptions for focal bony lesions were used by the RIIS program, it appears that these descriptions were interpreted differently by different radiologists. For example, in a case of chondrosarcoma with chondroid matrix, several users described the matrix as groundglass or bone. In these instances, this description of the matrix is considered incorrect and this incorrect description would adversely affect the program's diagnostic list. The proper description in such a case should include chondroid matrix. While all incorrect descriptions may not be as obvious as the above example, such descriptions lead the program further away from the correct diagnosis.

The possibility of providing an incorrect description to the system raises the question of whether providing a description of x-ray abnormalities is a legitimate method for inputting abnormal findings on an image or radiograph. Other possibilities for inputting image information include direct input of the image, using more descriptive terminology to describe the findings, or using graphic drawings of possible findings. Presently, direct input of the radiographic image is a relatively simple technical task with a state of the art video digitizer. However, the process of abstracting anatomic structures and anatomic abnormalities is an extremely difficult task that is only beginning to be investigated.[5] Therefore, the direct input of radiographic images would not be practical on a microcomputer.

Possible interim solutions would include diagramatic graphic representations of x-ray abnormalities and a more descriptive input environment. Simplified graphic representations of radiographic abnormalities might help eliminate potential discrepancies between use of standard radiographic descriptions. The user could simply select the graphic representation of specific abnormalities that could then be assembled by the system to produce a schematic drawing of the abnormality to provide user feedback. The graphic descriptions could then be assembled by the program for symbolic processing of either matched descriptive information or direct processing of the schematic diagrams.

A second and perhaps simpler solution relates to a more simplified descriptive input environment. In this type of environment, the program itself makes interim conclusions about the presence or absence of a finding from simplified descriptions. For example, instead of inquiring whether a bony matrix is present within a lesion, the system questions the user as to whether the abnormality demonstrates increased, decreased, or mixed density abnormalities. Subsequently, if the user selects increased density, the program questions more specifically about the location of the area of increased density and asks for a description of the characteristics of the increased density. The program then uses the more descriptive terminology to arrive at a conclusion about the presence or absence of a bony matrix.

Another major problem area involves the inference methodology or program logic. The inference engine in this system is simplistic in nature. The complete sequential match and score technique works relatively well in this small limited domain system as shown by the very good results observed with the proper descriptions. However, in larger more complex domains, this technique becomes computationally intensive and does not lend itself well to explanation of the conclusions made by the system.

A second important problem area in the program's logic is its selection criteria. A cutoff at the first 50% relative likelihood difference may not be appropriate, especially when the relative likelihood of all selected lesions are relatively low. This selection criteria is certainly simplistic in nature and may be inadequate for this type of decision. This problem relates to the expert system prototype construction and highlights how a specific expert system technique or programming technique can affect the accuracy of performance of an expert system.

An important evaluation criteria of an expert system is to determine that its information or knowledge base is factually and conceptually accurate. This task is quite difficult in the medical domain, especially with relative predictive values and frequency measures. Inaccuracies in determining relative frequencies or relative predictive values would be expected to significantly change the accuracy of the program. It is difficult to determine the relative frequency and predictive value units from standard radiographic textbooks. The inability to confirm the

accuracy of the knowledge base is a significant drawback to this type of knowledge or information structure.

User acceptibility and user confidence in the RIIS system was not evaluated. User "believability" in the system is an extremely important evaluation criteria especially if such expert systems are to be used in a clinical setting. It is therefore important to structure further evaluations of expert systems to evaluate user acceptance and to determine if the expert system helps non-experts to perform as experts. Furthermore, it is probably more important to provide the user with a good explanation as to why certain diseases were selected rather than simply presenting the user with a differential diagnostic list. There are several systems at present that provide such an explanation as a critiquing facility in their use.[7]

Finally, in order for expert systems to be used in a clinical setting, they must be widely available, relatively inexpensive, and user friendly. While there are many problems remaining to be solved before expert systems are clinically accepted, there is much potential in specific limited domain areas.

REFERENCES

1. Bar A, Feigenbaum EA: The Handbook of Artificial Intelligence (vol 3). Reading, Addison-Wesley, 1982

2. Quaglin S, Stefanelli M, Barosi G, et al: A performance evaluation of the expert system ANEMIA. Comput Biomed Res 21:307-323, 1988

3. Hudson DL, Cohen ME: The role of user-interface in a medical expert system, in Ackerman MJ, (ed): Proceedings Symposium on Computer Applications in Medical Care (SCAMC), Washington, DC, Institute of Electrical and Electronics Engineers, (IEEE) Computer Society, 1985, pp 232-236

4. Kundel HL, Nodine CF, Doi K: Human interpretation of displayed images, in Hendee WR, Wells PN (eds): Engineering Research in Visual Perception. Chicago, American College of Radiology, 1986

5. Vries JK, Banks G, McLinden S, et al: Three-dimensional neuro-imaging using octree encoding, in Ackerman MJ (ed): Proceedings SCAMC 1955, 697

6. Miller RA, Pople HE, Myer JD: Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 307:466-476, 1982

7. Swett HA, Miller PL: ICON: A computer-based approach to differential diagnosis in radiology. Radiology 163:555-558, 1987