Radiology Image Interpretation System: Modified Observer Performance Study of an Image Interpretation Expert System

David Piraino, Bradford Richmond, Mark Schluchter, Daniel Rockey, and Jean Schils

Application of computer-based expert systems to diagnostic medical problems has been described in many areas including clinical diagnosis and radiology. Expert systems are computer programs that contain encoded expert knowledge to provide expert advice. A modified observer-performance study was done comparing the efficacy of the Radiology Image Interpretation System (RIIS), an expert system that diagnoses focal bone abnormalities, and radiology residents on a known set of 44 abnormal and 10 normal cases. Modified receiver operating characteristic curves for four inexperienced residents, five experienced residents, and RIIS were generated using the set of known radiographs. The true-positive rates of RIIS and the residents at false-positive rates of 0.05, 0.15, and 0.20 were estimated using the modified receiver operating characteristics curve and were compared using a paired *t* test. On the average, the RIIS system was less accurate when compared with experienced and inexperienced residents but the difference was only significant for experienced residents at a false-positive rate of 0.05. RIIS performed better than inexperienced residents when RIIS was used by experienced residents but this difference was not significant.

KEY WORDS: ROC, observer-performer, expert system, diagnoses, image.

EXPERT SYSTEMS have been applied to a wide variety of medical problems including clinical diagnosis, radiology, laboratory test interpretation, and chemotherapy protocols.[1] Expert systems are computer programs that contain encoded knowledge obtained from an expert in a specific field. These programs use simple inference techniques and the encoded knowledge to produce expert advice for the user in the specific field. Several investigators have been studying the application of expert systems to radiographic image interpretation and consultation.[2-4] These systems have been used as a cognitive aid in their areas of expertise.

Lodwick et al[4] produced a program for computer-aided diagnosis of primary bone tumors using an approach based on Bayes' formula of inverse probability. This type of program can estimate conditional probabilities, but may not be as useful as a cognitive aid as expert systems. A probability matrix was estimated using a large number of known cases. This program correctly predicted the pathological diagnosis in 77% of selected cases taken from the group on which the conditional probability matrix was based.[4]

RIIS was developed to study the application of expert systems to diagnosing bone tumors and other focal bone abnormalities. Knowledge is encoded in Radiology Image Interpretation System (RIIS) in two ways: (1) "if-then" rules and (2) a knowledge base of diseases and findings associated with those diseases. The "if-then" rules are used to make conclusions from specific information. When the "IF" portion of the rule is true then RIIS will conclude that the "THEN" portion of the rule is also true. Multiple "if-then" rules can be concluded sequentially allowing complex analysis and decisions.

The disease knowledge base contains 45 diseases or disease variants with 60 findings associated with each disease. The set of 60 findings were findings considered important in diagnosing focal bone abnormalities by radiologists with extensive experience with primary and secondary bone tumors. All findings were considered independent. Each finding for each disease is given a relative frequency and relative predictive value. The relative frequency is used to estimate how often that finding occurs with that disease. The relative predictive value is used to estimate how much that finding predicts that disease is present. The relative frequencies and relative predictive values were determined by review of standard radiographic textbooks. This approach is modeled after the Internist program.[5] Table 1 shows the findings, relative frequencies and relative predictive values for chondrosarcoma that are contained in RIIS.

RIIS interactively questions the physician

From the Department of Radiology and the Department of Biostatistics, Cleveland Clinic Foundation, Cleveland, OH.

Address reprint requests to David Piraino, Department of Radiology, A-21, 1 Clinic Center, 9500 Euclid Ave, Cleveland, OH 44195-5021.

0897-1889/91/0402-0005$03.00/0

Table 1. Findings with Relative Frequencies and Predictive Values for Chondrosarcoma

Findings	Relative Frequency	Relative Predictive Value
Sclerotic	low	low
Lytic	medium	low
Geographic	medium	low
Motheaten	medium	low
Permeative	low	low
Narrow margin	medium	low
Intermediate margin	medium	medium
Wide margin	medium	medium
Bone matrix	low	low
Chondroid matrix	medium	high
Calcified matrix	medium	medium
Groundglass matrix	low	low
No matrix	low	low
Nidus	low	low
Expansion	medium	medium
Endosteal scalloping	high	low
Cortical penetration	medium	low
Sclerotic margin	medium	low
Central	high	medium
Eccentric	medium	low
Cortical	low	low
Juxacortical	low	low
Cortical thickening	low	low
Soft tissue mass	medium	medium
Periosteal reaction	medium	medium
Interrupted periosteal reaction	medium	medium
Uninterrupted periosteal reaction	medium	low
Epiphyseal	low	low
Metaphyseal	medium	medium
Diaphyseal	medium	medium
Diffuse	low	low
Age		
0-10	low	low
11-20	low	low
21-30	medium	low
31-40	medium	medium
41-50	medium	medium
51-60	medium	low
61-70	low	low
70-. . .	low	low
Sacrum/coccyx	low	medium
Humerus-radius-ulna	medium	low
Femur-tibia-fibula	medium	low
Hands-feet	low	low
Skull	low	low
Flat-bone	medium	medium
Fallen fragment	low	low

NOTE. Relative frequencies: low—never or rarely occurs in disease; medium—finding commonly occurs in disease; high—almost always or always occurs in disease. Relative predictive value: low—rarely caused by diagnosis; medium—disease is commonly associated with finding; high—pathogonomonic or almost pathognomonic of disease.

about the findings on a radiograph. The questions are used to produce a list of the findings present on the radiograph. The sequence of questions and which questions are asked depends upon answers to previous questions. For example if the user states that a lesion in the tibia has a thin sclerotic margin, there is no need to ask whether the lesion has a permeative or motheaten pattern and RIIS does not ask for this.

RIIS uses the list of findings to make several intermediate conclusions using "if-then" rules. For example the following rule is used by RIIS to decide if a bone lesion is aggressive: IF there is cortical destruction or wide zone of transition or permeative pattern or motheaten pattern. THEN lesion is aggressive.

The intermediate conclusions are used to classify a broad category of possible diagnoses such as bone forming nonaggressive lesions, aggressive lesions with chondroid matrix, and others. Diagnoses in the most likely category or categories are then considered for the final differential diagnosis.

A matching approach using the disease knowledge base and the list of radiographic findings is used to rate the relative likelihood of each disease. The relative likelihood of each disease is calculated according to the following rules: 1) a finding present in the radiograph and with a high predictive value for that disease increases the likelihood of that disease; 2) a finding present in the radiograph but not or rarely occurring in a disease decreases the likelihood of that disease if the predictive value is low; 3) a finding that is not present in the radiograph but occurs frequently in a disease decreases the likelihood of that disease. The sum of the above possibilities for each finding for one disease is the relative likelihood of that disease. Calculation of the relative likelihood of the osteosarcoma shown in Fig 1A is illustrated in Table 2.

Diseases are then put in order according to their relative likelihoods. At present RIIS is limited to interpretation of focal bone abnormalities. RIIS contains 45 diseases or disease variants in its disease knowledge base.

The program runs on an IBM (White Plains, NY) compatible personal computer. It was constructed using an expert system shell, Person-

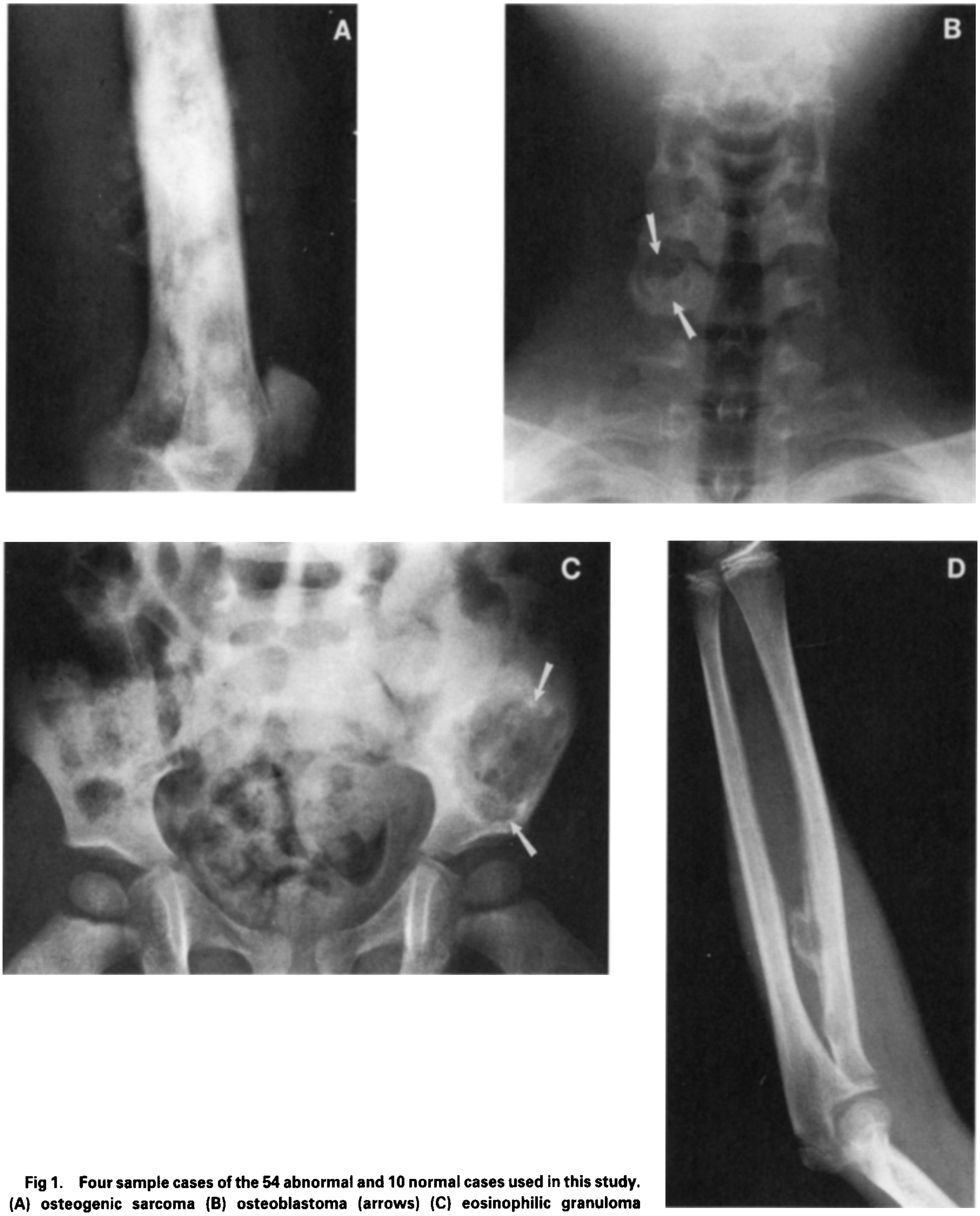

Fig 1. Four sample cases of the 54 abnormal and 10 normal cases used in this study. (A) osteogenic sarcoma (B) osteoblastoma (arrows) (C) eosinophilic granuloma (arrows) (D) subperiosteal aneurysmal bone cyst.

nel Consultant Plus (Texas Instruments, Austin, TX), and PC-Scheme (Texas Instruments, Austin, TX).

The advantages of observer performance studies using receiver-operating characteristics (ROC) analysis for medical imaging have been clearly established.[6,7] The application of observer-performance studies to expert systems has been limited.[8] Standard ROC analysis is used to measure the performance in detection of the presence or absence of one signal in an image. This type of analysis does not model the

Table 2. Calculation of the Relative Likelihood of Osteosarcoma for the Radiograph in Fig 1A for Selected Radiographic Findings

Findings	Relative Frequency in Osteosarcoma	Relative Predictive Value in Osteosarcoma	Finding Present (+) or Absent (−) in Figure 1A	Relative Likelihood
Sclerotic	medium	low	+	+1
Lytic	medium	low	—	+0
Geographic	medium	low	—	+1
Motheaten	medium	medium	—	+0
Permeative	medium	medium	—	+0
Narrow margin	low	low	—	+0
Intermediate margin	low	medium	—	+0
Wide margin	high	medium	+	+10
Chondroid matrix	low	medium	—	+0
Bone matrix	medium	high	+	+10
No matrix	low	low	—	+0
Nidus	low	low	—	+0
Sclerotic margin	low	low	—	+0
Total				+22

clinical situation of diagnosing bone tumors when a major problem is differentiating multiple diagnostic possibilities. In this clinical situation, which involves detection, localization, and classification, joint ROC curves and a modified ROC analysis comparing the performance at specific points on the joint ROC curve is suggested by Swets and Pickett[9] as the areas under the ROC curves are not directly comparable. Several types of errors can occur in this clinical situation: One may perceive an abnormality on a normal film, may incorrectly identify an abnormality, may fail to identify an abnormality, or may incorrectly localize an abnormality. This study compared RIIS, experienced radiology residents, and inexperienced radiology residents in the diagnosis of focal bone abnormalities using a modified ROC analysis.

METHODS AND MATERIALS

Forty four cases of known focal bone lesions and 10 normal cases were reproduced as black and white prints. The type of abnormalities included a wide variety of skeletal tumors. Table 3 shows the numbers of each tumor type. Several examples of the test radiographs are shown in Fig 1. The reproductions were used as test cases for RIIS and the radiology residents. The abnormal cases were selected from the Cleveland Clinic Teaching File and each had a known pathological diagnosis or were considered diagnostic of the disease by an experienced radiologist. The normals were selected from routine clinical exams considered normal by an experienced radiologist.

Five residents or fellows with at least 3 years of radiology training were the experienced residents. Four residents with less than 1 year of training were the inexperienced residents. The residents first select their own differential diagnosis of a case from a list of 34 diagnostic possibilities including normal. Each diagnosis was selected with a confidence level from 0 to 100 in increments of 10 (0, definitely not the diagnosis; 100, definitely the diagnosis).

RIIS was then used by the same resident and RIIS produced a differential diagnosis for the same case. RIIS interactively questioned the resident concerning the findings on the unknown radiograph and produced a list of the most likely diagnoses. The relative likelihood scale used by RIIS was not limited to the 0 to 100 range. The diagnostic likelihood estimate by RIIS was independent of the diagnostic ratings made by the residents. The resident's diagnoses as well as RIIS' diagnoses were automatically saved. Second-

Table 3. Number of Cases by Disease Type

Aneurysmal bone cyst	3
Brodies abscess	1
Brown tumor	1
Chondroblastoma	2
Chondrosarcoma	3
Chordoma	1
Enchondroma	1
Eosinophilic granuloma	3
Ewing's sarcoma	2
Fibrous dysplasia	2
Giant cell tumor	1
Hemangioma	1
Lymphoma	3
Malignant fibrous histiocytoma	1
Metastatic disease	1
Myeloma	1
Osteoblastoma	2
Osteochondroma	1
Osteoid osteoma	4
Osteosarcoma	6
Parosteal osteosarcoma	1
Simple bone cyst	3
Normal	10
Total	54

ary to time limitations not all residents read all the films. Residents were encouraged to work through the films in the same sequence.

Decisions for each possible diagnosis were clarified in four alternatives according to the relative likelihood and confidence estimates: (1) true-positive decision, decision is the true-positive or pathological diagnosis; (2) false-positive decision, decision is not the true-positive or pathological diagnosis; (3) true-negative decision, decision that is not the true-positive or pathological diagnosis is correctly avoided; and (4) false-negative decision, the true-positive or pathological diagnosis is not made.[8]

Observed ROC curves were constructed by calculating the true-positive rates (total true-positive diagnoses/[total true-positive diagnoses + total false-negative diagnoses] and false-positive rates [total false-positive diagnoses/total false-positive diagnoses + total true-negative diagnoses]) for RIIS and the residents using the confidence ratings and relative likelihoods. Fitted joint ROC curves were obtained separately for each reader and method from a least squares regression of the normal deviate of the true-positive rate on the normal deviate of the false-positive rate. The maximum likelihood method of Dorfman and Alf[10] was not used because the criteria for this method was not strictly satisfied. The true-positive rates at false-positive rates of 0.05, 0.15, and 0.2 were calculated from the fitted ROC curves. These true-positive rates were then compared using either a paired *t* test or a two-sample *t* test, depending on whether the test was comparing methods within the same set of residents or different groups of residents, respectively.

RESULTS

The observed ROC curves are shown in Figs 2 and 3. In general RIIS did not perform as well as the residents but this was not always the case. RIIS performed better than resident number 15, an inexperienced resident, as can be see in Fig 2.

Table 4 shows the average true-positive rates for the residents and RIIS at false-positive rates of 0.05, 0.15, and 0.2 along with differences and standard deviations. On the average, RIIS did not perform as well as experienced or inexperienced residents. This difference between the residents and RIIS was only significant for experienced residents at a 0.05 false-positive rate using a paired *t* test. The performance of RIIS with image descriptions by experienced residents was better than the performance of inexperienced residents but this difference was

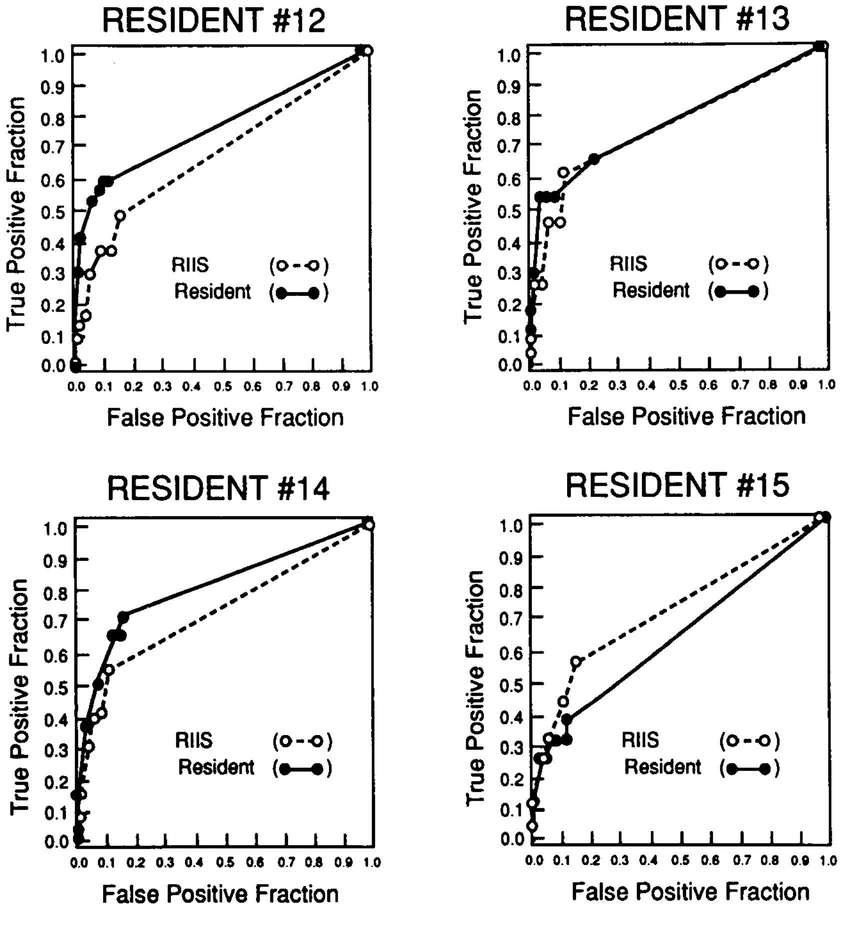

Fig 2. Observed ROC curves for inexperienced residents and RIIS with image descriptions by the same resident.

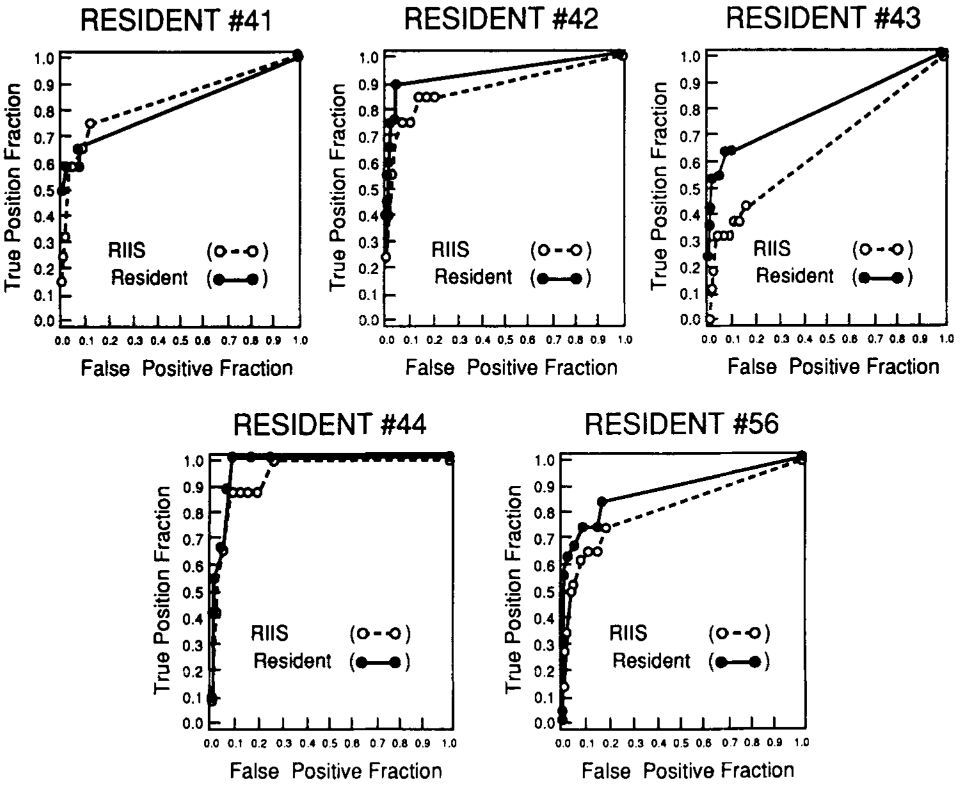

Fig 3. Observed ROC curves for experienced residents and RIIS with image descriptions by the same resident.

not significant (Table 5). Experienced residents also outperformed inexperienced residents and this difference was significant at a false-positive rate of 0.05 (Table 6). RIIS when using image descriptions from experienced residents outperformed RIIS when using image descriptions from inexperienced residents, and this difference was significant at a false-positive rate of 0.05 (Table 7).

DISCUSSION

Computer-based medical expert systems have been described in many areas but only a few of these systems have been evaluated for accu-

Table 4. True Positive Rates for RIIS, Inexperienced Residents and Experienced Residents

False-positive Rate	Average True-positive Rate, Standard Deviation and *P*-value of Inexperienced Residents or RIIS (N = 4)	Average True-positive Rate, Standard Deviation and *P*-value of Experienced Residents or RIIS (N = 5)
0.05		
Resident	0.42 SD 0.1	0.70 SD 0.12
RIIS	0.32 SD 0.06	0.55 SD 0.17
Difference (resident-RIIS)	0.10 SD 0.11	0.15 SD 0.10
P-value of difference	0.18	0.03*
0.15		
Resident	0.59 SD 0.16	0.81 SD 0.15
RIIS	0.55 SD 0.10	0.73 SD 0.18
Difference (resident-RIIS)	0.04 SD 0.16	0.08 SD 0.15
P-value of difference	0.65	0.28
0.20		
Resident	0.64 SD 0.18	0.84 SD 0.14
RIIS	0.62 SD 0.11	0.78 SD 0.17
Difference (resident-RIIS)	0.02 SD 0.17	0.06 SD 0.16
P-value of difference	0.84	0.44

*Statistically significant *P*-value ($P < 0.05$)

Table 5. Average True-Positive Rates for Inexperienced Residents and RIIS With Image Descriptions by Experienced Residents

False-positive Rate	Average True-positive Rate for Inexperienced Residents	Average True-positive Rate for RIIS with Image Descriptions by Experienced Residents	*P*-value from 2-sample *t* Test
0.05	0.42	0.55	0.20
0.15	0.59	0.73	0.26
0.20	0.64	0.78	0.29

racy.[11] As these expert systems become more numerous and available for use by physicians it is important that these systems be evaluated for efficacy. This study demonstrated the successful use of a modified ROC analysis for testing a radiologic expert system.

This modified or joint ROC analysis required the observer to classify the abnormality into a limited number of predefined categories or diagnoses. The use of a modified ROC analysis allowed the comparison of residents and RIIS, the expert system, over a range of false-positive values in an environment closely modeling the clinical environment. This is important as performance may be significantly different at different false-positive rates. This type of analysis also has the advantage of standardizing the evaluation of radiologic expert systems in a manner similar to other observer-performance imaging studies.

This modified ROC analysis of radiologic expert systems has the same disadvantages of all ROC studies including large number of observers needed, selection of number and type of cases, and several others.[7] This study does not take into account sample variability in the type of bone tumors used. However, an attempt was made to include many different types of tumors and different appearances of the same tumors. We feel that this type of modified ROC analysis of medical expert systems is a valuable tool in evaluating these systems.

This study demonstrated that RIIS, a prototype expert image interpretation system, on the average performed similar to inexperienced residents and only slightly worse than experienced residents given the variability among residents. We consider these results very encouraging for the future prospects of expert systems in radiology as there are several enhancements that could be made to RIIS that should improve its performance. These enhancements could include extension of the disease knowledge base to account for disease variation, more accurate estimates of relative frequencies and predictive values, inclusion of new rules to make better decisions concerning the categories of disease, and the inclusion of a case data base to allow the program to learn from previous known cases.

Studies such as this one that evaluate expert system's performance must be done before an expert system can be used in clinical practice. After performance of the expert system alone has been evaluated, as was done in this study, a further important study would be to test whether a radiologist's performance improved with the use of an expert system. This type of evaluation was not done in this study. Such a study could potentially answer the important question of whether or not expert systems could improve the quality of care provided by radiologists.

Comparison of this study with a study done on the program made by Lodwick[4] is difficult. An analysis of Lodwick's program demonstrated that a general radiologist with the help of Lodwick's program performed better than a different general radiologist without the program's help.[12] The present study and the previous study of Lodwick's program did not test

Table 6. Average True-Positive Rates for Experienced and Inexperienced Residents

False-positive Rate	Average True-positive Rate for Inexperienced Residents	Average True-positive Rate for Experienced Resident	*P*-value from 2-sample *t* Test
0.05	0.42	0.70	0.01*
0.15	0.59	0.81	0.06
0.20	0.64	0.84	0.11

*Statistically significant *P*-value ($P < 0.05$)

Table 7. Average True-Positive Rates for RIIS With Image Descriptions by Experienced and Inexperienced Residents

False-positive Rate	Average True-positive Rate for RIIS with Descriptions by Inexperienced Residents (N = 4)	Average True-Positive Rate for RIIS with Descriptions by Experienced Residents (N = 5)	*P*-value from 2-Sample *t* Test
0.05	0.32	0.55	0.03*
0.15	0.55	0.73	0.12
0.20	0.62	0.78	0.17

*Statistically significant *P*-value ($P < 0.05$)

whether a individual radiologist's performance improved when using the expert system. Direct comparison of Lodwick's program performance to this study would not be valid as different sets of cases were used and the evaluation methods are different.

It is interesting to note that the experienced residents' performance (true-positive rate = 0.84 at false-positive rate = 0.20) and the expert system's performance using experienced residents' image descriptions (true-positive rate = 0.78 at false-positive rate = 0.20) in the present study is similar to the performance of the radiologist with experience with bone tumors (true-positive rate = 0.86 at false-positive rate = 0.18), and the performance of the general radiologist with help from Lodwick's program (true-positive rate = 0.82 at false-positive rate = 0.17) in an analysis of Lodwick's program.[12] It is also interesting that the expert system's true-positive rate, 78%, with experienced residents' image description at a false-positive rate = 0.20 was similar to Lodwick's program performance of diagnosing 77.9% of bone tumors in a separate study.[4]

The fact that RIIS was capable of better performance when used by experienced residents than when used by inexperienced residents shows that the performance of RIIS with inexperienced users could be improved if an "accurate" description could be obtained from them. Future expert image interpretation systems need to account for variations in the users ability to "accurately" describe the important findings. This could be accomplished by direct input of the image, modification of the expert system by experience level, and user-specific knowledge bases.

ACKNOWLEDGMENT

Lata Paranandi, MSHP, for her help with data analysis.

REFERENCES

1. de Vries PH, de Vries Robbe PF: An overview of medical expert systems. Methods Inf Med 24(2):57-64, 1985

2. Swett HA, Fisher PR, Cohn AI, et al: Expert system—Controlled image display. Radiology 172:487-493, 1989

3. Piraino D, Richmond B, Uetani M, et al: Problems in applying expert system technology to radiographic image interpretation. J Digit Imag 1(2):21-26, 1989

4. Lodwick GS, Haun CL, Smith WE, et al: Computer diagnosis of primary bone tumors: A preliminary report. Radiology 80:273-275, 1963

5. Miller RA, Pople HE, Myers JD: Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 307:468-476, 1982

6. Swets JA, Pickett RM, Whitehead SF, et al: Assessment of diagnostic technologies. Science 205:753-759, 1979

7. Metz CE: Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 24:234-245, 1989

8. Adlasing KP, Scheithauer W: Performance evaluation of medical expert systems using ROC curves. Comput Biomed Res 22:297-313, 1989

9. Swets JA, Pickett RM: Evaluation of diagnostic systems: Methods from signal detection theory. New York, NY, Academic, 1982, 51-58

10. Dorfman DD, Alf E: Maximum likelihood estimation of parameters of signal detection theory and determination of confidence interval-rating data. J Math Psych 6:487-496, 1969

11. Lundsgaarde HP: Evaluating medical expert systems. Soc Sci Med 24(10):805-819, 1987

12. Virtama P, Katevuo K, Makela P, et al: Computer-aided diagnosis of bone tumors. Acta Radiol Diag 20:70-74, 1979