INTERPRETATION OF RADIONUCLIDE LIVER IMAGES: DO TRAINING AND EXPERIENCE MAKE A DIFFERENCE?

H. Nishiyama, J. T. Lewis, A. B. Ashare, and E. L. Saenger

Nuclear Medicine Laboratory, Bureau of Radiological Health and Radioisotope Laboratory, University of Cincinnati, Cincinnati General Hospital, Cincinnati, Ohio

Individual observers interpreted images in order to assess their accuracy and sources of error. Seventy-six liver images were presented to nine readers for interpretation. Readers of differing experience participated in the study: four radiology residents, three fellows in nuclear medicine, and two full-time nuclear medicine physicians. A higher incidence of false-positive reading was more common in inexperienced observers (11–50%) and the most correct readings were obtained by staff physicians (88% overall percentage accuracy). Heterogeneity in radionuclide uptake was the most frequently mentioned abnormality, and heterogeneity of an ill-defined nature was the most common false–positive finding in normal cases. Early stages of cirrhosis, mild hepatitis, and rare diseases such as hepatic sarcoidosis were difficult to detect. Skill of interpretation improved with experience, especially in judging heterogeneity of an ill-defined nature, and the rate of accurate readings was proportional to the level of training of the observer.

One of the commonly performed procedures in nuclear medicine is liver imaging; however, these images impose a variety of problems for interpretation because of difficulty in reliably differentiating normal and abnormal. Various studies have appeared in the literature describing the anatomical variation in size (*1,2*) and shape of the liver in normal individuals (*1,3*). Further, limitations in instrument capability to detect focal areas of decreased or absent uptake and many variations of disease pattern have been reported (*3–5*). Computer applications to detect minimal heterogeneity (*6*) and dynamic liver scintigraphy (*7*) present further difficulty in application and interpretation.

Training and experience are thought to be important factors in accurate image interpretation. The purpose of this report is to assess how individual observers interpret images and their sources of error. Image-reading ability was tested by comparing three groups of observers of differing experience.

MATERIALS AND METHODS

Liver images of 40 patients with a variety of liver diseases and 36 of normal patients were interpreted by separate observers without knowledge of the other clinical findings. The number and type of cases chosen reflected the average weekly case load in the

Received May 17, 1974; revision accepted Aug. 5, 1974.

For reprints contact: H. Nishiyama, Nuclear Medicine Laboratory, Cincinnati General Hospital, Cincinnati, Ohio 45267.

hospitals at which the authors practice (Cincinnati General Hospital, Jewish Hospital, Children's Hospital, and Veterans Administration Hospital combined services); that is, 70–80 liver images per week with slightly over 50% of these cases reported as abnormal.

Table 1 lists the different clinical indications for performing the liver imaging procedure. There is close correlation with the average weekly census of images requested by referring physicians. All cases with liver disease included in the study were documented by either needle biopsy, open surgery, or autopsy: 36 of these within 2 months and 4 within 4 months after imaging. Final histologic diagnoses of the 40 abnormal cases are given in Table 2. There is a variety of liver diseases including not only "space-occupying lesions" but also benign, diffuse liver parenchymal diseases. Thirty cases were determined to be normal after long-term followup without evidence of liver diseases ranging from at least 6 months (1 case) to 24 months. Six cases were proven normal by surgical procedures within 2 months after imaging.

The nine observers who constituted the study group included four radiology residents, who participated in the study within 1 month after completion of a 3-month nuclear medicine training period, three nuclear medicine fellows with length of experience ranging from 6 to 18 months, and two full-time nuclear medicine staff physicians with 4 and 20 years of experience. None of the observers knew the proportion of normal and abnormal cases. There were 684 interpretations (76 images read by nine observers) to be considered. The observers were asked to read 15–16 cases at a time so as not to become unduly fatigued. The observers completed an evaluation sheet listing the factors on which they based their decision (Table 3).

The upper limits of normal liver size were assumed to be approximately 20 cm in vertical height (*1,2*). All images were obtained by scintillation cameras following intravenous administration of 3 mCi of ^{99m}Tc-sulfur colloid. Imaging was usually performed 10 min postdose. Anterior, posterior, and lateral views were obtained. An additional view with a lead marker (10 cm long) was obtained in the anterior view for the assessment of organ size.

Results of interpretation of images by each observer were categorized as true-positive, false-negative, true-negative, and false-positive, as described by Lusted (*8*). Overall percentage accuracy (OPA) was computed for each individual and for each group of observers. A receiver-operating-characteristic (ROC) graph was established with each observer

TABLE 1. CLINICAL INDICATIONS FOR REQUESTING LIVER IMAGING IN 76 STUDY CASES

Indications	Proven abnormal (cases)	Proven normal (cases)
Chronic liver diseases	14	3
Search for primary and 2-deg tumors	9	18
Search for abscess	8	5
Acute liver diseases	7	—
Hepatomegaly	2	5
Abnormal enzymes	—	3
Chronic granulomatous disorders	—	2
Total	40	36

TABLE 2. HISTOLOGIC DIAGNOSES OF 40 ABNORMAL CASES

Diagnosis	Cases (No.)
Metastasis	12
Cirrhosis	9
Hepatitis	8
Traumas	3
Abscess	2
Polycystic liver disease	2
Hepatic sarcoidosis	1
Miliary TBC involving liver	1
Chronic passive congestion	1
Fatty degeneration (diabetes)	1
Total	40

TABLE 3. EVALUATION SHEET OF LIVER IMAGE

A. Normal ________ or abnormal ________
B. If Abnormal, why?
  1. Size: large ________ small ________
  2. Shape (describe, e.g., enlarged left lobe):
  ____
  3. Heterogenity* (more than one factor can be marked:
    a. ill-defined ________
    b. discrete ________
    c. single ________
    d. multiple ________
  4. Bone marrow uptake:
    a. Anterior view ________
    b. Posterior view ________
  5. Abnormal liver uptake:
    a. Increased ________
    b. Decreased ________
C. If 5 above is marked, what is your rationale:
  1. Compared with other view of liver ________
  2. Compared with other view of spleen ________
D. Other comments: ________

* Heterogeneous radionuclide uptake includes both discrete (focal) and ill-defined defects.

representing a single point. The data were also analyzed using the average information content per image (AIC), a new measure of observer performance derived from information theory and suggested by Metz, et al (*9*) (Table 4). By using the AIC, the amount of information obtained by different observers operating at different points on an ROC graph can be compared quantitatively.

TABLE 4. A NEW MEASURE OF OBSERVER PERFORMANCE DERIVED FROM INFORMATION THEORY (*9*)

Average information content per image =

$$TP \times F \times \log_2\left(\frac{TP}{TP \times F + FP \times (1 - F)}\right)$$

$$+ FP \times (1 - F) \times \log_2\left(\frac{FP}{TP \times F + FP \times (1 - F)}\right)$$

$$+ (1 - TP) \times F \times \log_2\left(\frac{1 - TP}{1 - TP \times F - FP \times (1 - F)}\right)$$

$$+ (1 - FP) \times (1 - F) \times \log_2\left(\frac{1 - FP}{1 - TP \times F - FP \times (1 - F)}\right)$$

Where TP = Conditional true-positive frequency or rate

$$= \frac{\text{No. of positive responses to abnormal cases}}{\text{No. of abnormal cases}}$$

FP = Conditional false-positive frequency or rate

$$= \frac{\text{No. of positive responses to normal cases}}{\text{No. of normal cases}}$$

F = Frequency of abnormal cases

$$= \frac{\text{No. of abnormal cases}}{\text{No. of abnormal cases + No. of normal cases}}$$

In this problem, $F = \frac{40}{40 + 36} = 0.526$

The above equation for AIC is derived from information theory for the case where only a binary decision is considered, i.e., positive or negative.

RESULTS

Composite data from the nine observers are presented in Table 5 and the ROC graph is shown in Fig. 1. The highest percentage of true-positive readings was obtained by residents and fellows; however, their percentage of false-positive readings was also high. A reciprocal relationship between the percentages of false-negative and false-positive diagnoses was noted in all of the participants as predicted by Lusted (*8*) (Table 5). The most accurate interpretation associated with both low false-negative and false-positive readings was performed by the staff physicians. Less experienced observers had a tendency to overread (high false-positive), a trend noted in both the residents and the fellows.

Table 5 also shows that the highest AIC values were scored by the staff members (0.4730 and 0.4879). This information content value obtained by the staff was almost twice that scored by the fellows (0.2057–0.2913). The values obtained by the residents were not significantly different from those of the fellows (0.1196–0.3039). Both staff members obtained OPA values of 88% but the staff member with the lower false-positive rate received a higher AIC value even though his true-positive rate was lower.

As shown in Table 6, there were 360 possible truly abnormal readings of which 312 were correctly read. Although there were 324 readings that should have been read as normal, 97 were falsely read as abnormal. The complete description of heterogeneity is somewhat different between true-positive and false-positive groups as seen in Table 7.

Among the true-positive cases, 50 observations were correctly made on the basis of either single or multiple ill-defined heterogeneity alone. The true-positive cases with this finding were found in patients with metastasis (19/50), hepatitis (7/50), cirrhosis (7/50), trauma (6/50), miliary tuberculosis involving liver (3/50), chronic passive congestion (with fibrotic changes histologically) (2/50), hepatic sarcoidosis (2/50), and four miscellaneous diagnoses. Difficulty in defining the heterogeneity of an ill-defined nature was well substantiated by 30 false-positive observations (Table 7). Thirteen false-positive observations were based on single discrete lesions. Some but not all of these were misreadings of normal anatomical variations such as gallbladder fossa or impressions of right kidney, inferior vena cava, and/or vertebra (*3*). Heterogeneity is not thought to be due to instrumentation malfunction which was minimized by daily quality-control tests including a field flood image to document uniformity.

The matter of which disease entities were most commonly missed (false-negative) is also important. A total of 48 false-negative observations (Table 5) was made. These occurred in patients with hepatitis (16/48), cirrhosis (10/48), hepatic sarcoidosis (7/48), trauma (6/48), miliary tuberculosis (3/48), chronic passive congestion (2/48), fatty degeneration of liver (2/48), and metastasis (2/48). As shown in Table 2, three cases with trauma were included in this study. One patient sustained laceration of the capsule in an auto accident which was interpretable as ill-defined heterogeneity. Two cases with gunshot wounds were observable as abnormal in shape. Two observers failed to recognize this ill-

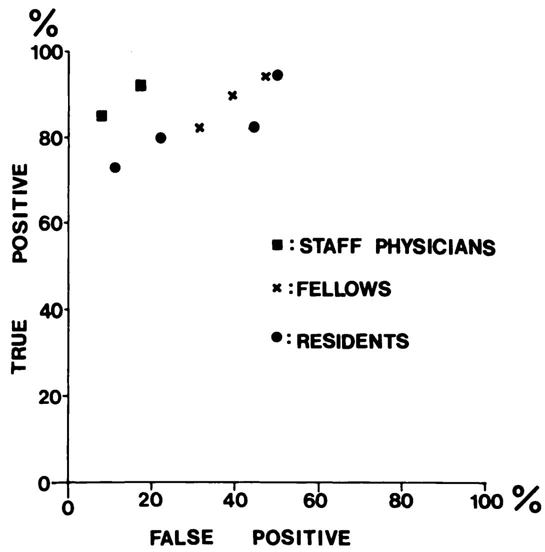

FIG. 1. ROC graph obtained from nine observers. Note difference between staff physicians and others with differing experience.

defined heterogeneity in the first case and four observers judged that the unusual shape alone was insufficient evidence to call images abnormal in the latter two (false-negative).

DISCUSSION

Our study population consisted of images of patients with a variety of liver diseases and normal livers, a close reflection of daily practice. By performing the reading of images without knowledge of clinical history, however, the results obtained may differ from what actually takes place in a practice situation.

The objective of a physician interpreting liver images is to decrease the number of false-positive readings and, simultaneously, increase the number of true-positives. From the data presented here it is apparently better to suffer a loss of true-positives (92.5–85%) and thereby decrease the false-positives (17–8%) for an improvement in the AIC values. In practice, however, the high AIC value at the cost of a lower true-positive rate may not be beneficial to the patient.

Observer variability in the interpretation of liver images has been documented (*10,11*). In our study, there was little difference in OPA between groups of fellows and residents, which might have been due in part to individual ability to recognize the pattern rather than experience alone. Skills of pattern recognition in inexperienced physicians can usually be improved by learning the sources of error but this is not necessarily so in experienced individuals (*10*).

One major source of false-positive reading can be attributed to difficulty in defining hepatomegaly. The study done by Drum, et al (*5*) set criteria of 17 cm in vertical dimension as maximum size (*1*). This will divide about 50% of the population as abnormal since the value was obtained from 55 "selected normal" scans on adult patients as a mean figure (16.7 cm ± 2.1 cm) (*1*). We have set 20 cm of vertical height as an upper limit of normal size and consequently enlarged size alone was not a common source of error. Emphasis on abnormal shape or

TABLE 5. COMPOSITE DATA FROM NINE OBSERVERS

	Reading of images*						
	True-positive	False-negative	True-negative	False-positive			
	(40 Abnormal cases)		(36 Normal cases)		OPA†	Average percentage	
Observer	(%)	(%)	(%)	(%)	(%)	accuracy	AIC‡
Resident 1	33 (82.5)	7 (17.5)	20 (56.0)	16 (44.0)	70		0.1196
2	38 (95.0)	2 (5.0)	18 (50.0)	18 (50.0)	74	76	0.2076
3	32 (80.0)	8 (20.0)	28 (78.0)	8 (22.0)	79		0.2581
4	29 (72.5)	11 (27.5)	32 (89.0)	4 (11.0)	80		0.3039
Fellow 1	38 (95.0)	2 (5.0)	19 (53.0)	17 (47.0)	75		0.2286
2	33 (82.5)	7 (17.5)	25 (69.0)	11 (31.0)	76	77	0.2057
3	38 (95.0)	2 (5.0)	22 (61.0)	14 (39.0)	79		0.2913
Staff 1	37 (92.5)	3 (7.5)	30 (83.0)	6 (17.0)	88		0.4730
						88	
2	34 (85.0)	6 (15.0)	33 (92.0)	3 (8.0)	88		0.4879
Total & average	312 (87.0)	48 (13.0)	227 (70.0)	97 (30.0)		78	

* Numbers in parentheses indicate percentages of true and false readings.

† Overall percentage accuracy $= \frac{\text{No. of correct interpretations}}{\text{Total number of cases}} \times 100.$

‡ Average information content per image, see Table 4.

TABLE 6. REASONS QUOTED TO MAKE FACTORS INFLUENCING TRUE- AND FALSE-POSITIVE READINGS

Observation	Frequency of observation: True-positive (%)	False-positive (%)
Heterogeneity	114 (36.5)	44 (45.4)
Heterogeneity + size	57 (18.3)	3(3.1)
Heterogeneity + marrow uptake	19 (6.1)	4 (4.1)
Heterogeneity + size + shape	12 (3.8)	0 —
Diffusely decreased liver uptake + marrow uptake	12 (3.8)	0 —
Heterogeneity + marrow uptake + size	10 (3.2)	0 —
Heterogeneity + diffusely decreased liver uptake + marrow uptake	9 (2.9)	1 (1.0)
Marrow uptake	5 (1.6)	14*(14.4)
Shape	1 (0.3)	10 (9.3)
Heterogeneity + shape	8 (2.6)	9 (9.3)
Other combinations	65 (20.8)	12 (12.4)
Total	312 (100)	97 (100)

* Fourteen observations in six cases were made. Four cases were judged to have marrow uptake by one observer and marrow uptake was at best equivocal. Two cases were judged by six and four observers, respectively, as showing marrow uptake which was unequivocal in retrospect. Diagnoses were pneumonia and congestive heart failure at the time of imaging. No enzyme abnormalities were noted in either case.

TABLE 7. HETEROGENEITY OBSERVED IN TRUE- AND FALSE-POSITIVE READINGS

Heterogeneity	No. of observations: True-postive (%)	False-positive (%)
Multiple ill-defined	40 (35.1)	19 (43.2)
Single ill-defined	10 (8.8)	11 (25.0)
Single discrete	24 (21.0)	13 (29.5)
Multiple discrete	37 (32.5)	1 (2.3)
Ill-defined and discrete	3 (2.5)	0 —
Total observations	114 (100)	44 (100)

position can increase the false-positive rate unless the range of normal variations is fully appreciated (*1,3*). Nine percent of our false-positive readings was based on changes in shape alone. A higher percentage (27%) has been noted in another study (*12*).

Early stages of metastatic carcinoma or small metastatic foci and early stages of diffuse liver diseases such as hepatitis and cirrhosis will continue to pose problems for interpreters (*4,12*) so long as the only detectable abnormality is heterogeneity of an ill-defined nature. Among the 48 false-negative observations in our study (Table 5), only two of these were made in patients with metastatic carcinoma. In a previous study quantitative analysis aided by a computer indicated that the cirrhotic liver was frequently indistinguishable from the normal (*6*). Our study indicated that not only cirrhosis but also hepatitis is frequently inseparable from the normal. Diffuse liver parenchymal diseases will continue to be difficult to detect.

We believe that the results reported demonstrate the value of training and experience to improve interpretation of radionuclide liver images. The 3-month resident training program is valuable for orientation; however, further experience is required to reduce a high false-positive rate. There is also a need for further training of fellows since, as a group, they also show high false-positive rates. The ability to differentiate between the normal and abnormal liver with minimal heterogeneity may be possible since the OPA was proportional to the observers' experience. Heterogeneity of an ill-defined nature is the most difficult factor to detect correctly; however, with experience this is minimized by a careful and categorical examination of each liver image. One should be extremely cautious in calling this abnormality if other factors, e.g., size, marrow uptake, shape, are all normal. Daily reading of various clinical cases and subsequent followup serve best to develop and maintain skill in liver image interpretation.

ACKNOWLEDGMENT

We would like to thank radiology residents and fellows in nuclear medicine who participated in this study.

REFERENCES

1. McAfee JG, Ause RG, Wagner HN: Diagnostic value of scintillation scanning of the liver. *Arch Intern Med* 116: 95–110, 1965

2. Lin S, Mansfield CM, Kramer S, et al: Liver scanning in patients with suspected or proven cancer. *Am J Roentgenol Radium Ther Nucl Med* 108: 98–101, 1970

3. Covington EE: Pitfalls in liver photoscans. *Am J Roentgenol Radium Ther Nucl Med* 109: 745–748, 1970

4. Covington EE: The accuracy of liver photoscans. *Am J Roentgenol Radium Ther Nucl Med* 109: 742–744, 1970

5. Drum DE, Christacopoulos JS: Hepatic scintigraphy in clinical decision making. *J Nucl Med* 13: 908–915, 1972

6. Grassman ED, Horgan JD, Meade RC: Quantitative analysis of liver scan data as an aid to diagnosis. *Radiology* 95: 517–523, 1970

7. Saba TM, Kaplan E, Graham L, et al: Hepatic phagocytic, metabolic, and blood-flow evaluation by dynamic scintigraphy. *J Nucl Med* 13: 300–306, 1972

8. LUSTED LB: Decision-making studies in patient management. *N Engl J Med* 284: 416–424, 1971

9. METZ CE, GOODENOUGH DJ, ROSSMANN K: Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. *Radiology* 109: 297–303, 1973

10. LUDBROOK J, SLAVOTINEK AH, RONAI PM: Observer error in reporting on liver scans for space-occupying lesions. *Gastroenterology* 62: 1013–1019, 1972

11. CONN HO, SPENCER RP: Observer error in liver scan. *Gastroenterology* 62: 1085–1090, 1972

12. HAYNIE TP, JHINGRAN SG, JAHNS MF: Sources of error in diagnostic scintigrams of the thyroid, liver and brain. *Tex Med* 68: 132–136, 1972

THE SOCIETY OF NUCLEAR MEDICINE
22nd ANNUAL MEETING

June 17–20, 1975 **Philadelphia Civic Center** **Philadelphia, Pa.**

FOURTH CALL FOR ABSTRACTS FOR SCIENTIFIC PROGRAM

The Scientific Program Committee welcomes the submission of abstracts of original contributions in nuclear medicine from members and nonmembers of the Society of Nuclear Medicine for the 22nd Annual Meeting. Abstracts for both the regular scientific program and for works-in-progress papers will be published in the June issue of the Journal of Nuclear Medicine.

This year the Committee plans to divide the program into five major categories: Basic Science, Clinical Practice, Clinical Research, Pediatric Nuclear Medicine, and Special Topics. Papers on the following subjects will be considered in these sessions:

- Bone/joint
- Cardiovascular
- Competitive binding assays
- Computer/data analysis
- Computerized axial tomography
- Dosimetry
- Gastroenterology
- Hematology
- Instrumentation (and ultrasound)
- Metabolism
- Neurology
- Oncology
- Pediatrics
- Pulmonary
- Radiopharmaceuticals
- Renal/electrolytes

GUIDELINES FOR SUBMITTING ABSTRACTS

This year abstracts will be printed from camera-ready copy provided by the authors. Therefore only abstracts prepared on the official abstract form will be considered. These abstracts forms are available from the Society of Nuclear Medicine, 475 Park Ave. South, New York, N.Y. 10016. The forms will not be sent to the entire membership as in former years, but must be requested from the Society office in New York. Be sure to request enough forms since only original forms can be used for each submission. The original and six copies must be submitted.

The deadline for submitting abstracts for the regular scientific program and for works-in-progress papers is:

January 15th, 1975

Send the original abstract form, supporting data, and six copies to:

James H. Christie, M.D.
Dept. of Radiology
University Hospital
Iowa City, Iowa 52240