Thallium-201 Myocardial Imaging: An Interinstitutional Study of Observer Variability

Gene B. Trobaugh, Frans J. Th. Wackers, Ellinor Busemann Sokole, Timothy A. DeRouen, James L. Ritchie, and Glen W. Hamilton

Veterans Administration Hospital and the University of Washington, Seattle, Washington, and the Cardiology and Nuclear Medicine Clinics, Wilhelmina Gasthuis, Amsterdam, The Netherlands

The sensitivity of myocardial imaging with thallium-201 for detection of infarction and ischemia depends on reproducible image interpretation. Agreement in interpretation among four experienced readers at two institutions was studied in 100 consecutive patients, 50 from each center. Studies were performed at rest following a 2-mCi injection of Tl-201 on similar scintillation cameras with high-resolution collimation. Each study (three views: ANT, 45° LAO, LLAT) was read independently from trilens Polaroid scintiphotos. All observers were unaware of other clinical information. Each study was interpreted as normal (N), borderline (B), or abnormal (AB) and quality graded as poor, adequate, or excellent.

Complete agreement among all possible observer pairs (six) was similar: 67% (range: 61–73%); complete disagreement (e.g. N/AB) was 4% (range: 2–8%). For all four readers combined, there was complete or essential (e.g. NNNB) agreement in 79, minor disagreement in 8, and major disagreement (e.g. at least 1 N and 1 AB) in 13. Poor-quality scans (as judged by two or more observers) were nearly equally divided between the two centers: nine in Seattle, and seven in Amsterdam.

The interobserver agreement found is similar to that reported for other images and for coronary arteriography. This study defines interobserver limitations on Tl-201 sensitivity in detecting infarction and ischemia.

J Nucl Med 19: 359–363, 1978

Myocardial imaging with various radionuclides is used increasingly to evaluate patients with known or suspected cardiac disease. Thallium-201 is currently widely used for this purpose (*1–6*), and provides clinically useful images with standard gamma-camera and collimator systems (*3,5*). The inherent quality of the Tl-201 myocardial image, however, is no more than marginal or adequate compared with other commonly used imaging procedures, such as bone or lung scans. Accurate interpretation is therefore often difficult, even for experienced observers (*2,3*). Intelligent clinical use of Tl-201 imaging is dependent on a realistic knowledge of the degree of observer variability in interpretation of the myocardial images.

This study compares, in detail, the observer variability between four exeprienced investigators from two separate institutions*.

MATERIALS AND METHODS

One hundred sequential Tl-201 myocardial images were retrospectively selected; 50 from each institution. Selection criteria were identical at the two institutions: simply the first 50 resting studies performed after a mutually agreed-upon date. The technical staff retrieved the studies, rephotographed the trilens Polaroid scintiphotos, and mounted the anterior, left anterior oblique, and left lateral views on a sheet with the sequential case number (1–50). Each observer then read the series of 100 studies without

Received July 29, 1977; revision accepted Dec. 6, 1977.

For reprints contact: Glen W. Hamilton, Chief, Nuclear Medicine Sec. (115), Veterans Administration Hospital, 4435 Beacon Ave. So., Seattle, WA 98108.

any knowledge of the patients' clinical history or the other observers' readings†, and noted whether the image was normal, abnormal, or borderline; and classified the image quality as adequate, excellent, or poor. Normal was defined as images showing homogeneous activity throughout the entire myocardium; abnormal images demonstrated discrete regions of diminished activity visually estimated to have 50% or less than normal activity and usually seen in more than one view; borderline images exhibited some degree of inhomogeneity but of insufficient magnitude to be called definitely abnormal. The combined results were collated by a fifth investigator not participating in the image reading.

Imaging was performed with similar techniques at the two institutions. Imaging was begun 20–30 min following a resting i.v. injection of 1.5–2.0 mCi of Tl-201‡ and images obtained in the anterior, left anterior oblique, and left lateral positions over a period of 34–45 minutes. Both centers used a camera§ with an energy window of 20% and a low-energy, high-resolution collimator¶ with 1800 parallel holes. Other details of the imaging procedure have been reported previously (*2,3,9*).

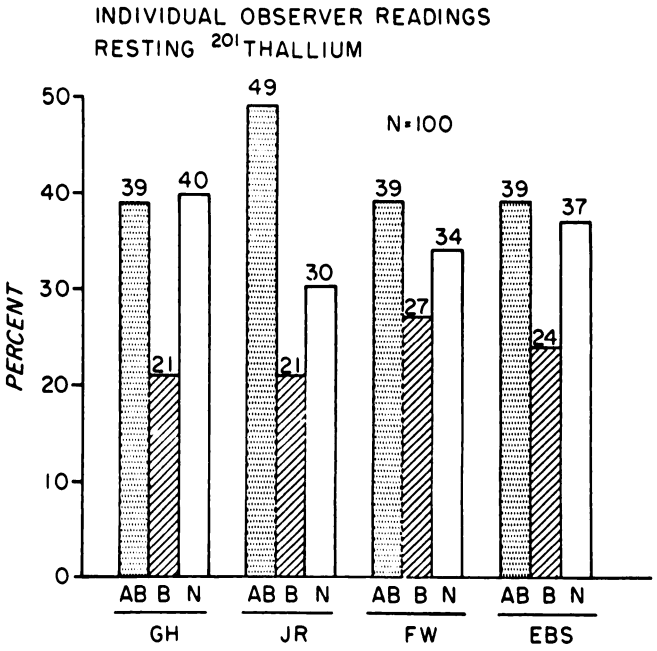

FIG. 1. Percentages of abnormal (AB), borderline (B), and normal (N) as read by each of the four observers. One observer (J.R.) tended to read more abnormals and fewer normals than the other three.

RESULTS

Based on the average reading of all four observers combined, 35 studies were normal, 23 borderline, and 42 obnormal. Because of the timing of the study, the majority of patients from Amsterdam were studied to detect acute myocardial infarction, whereas patients from Seattle were predominately those with chronic angina pectoris.

Figure 1 shows the individual readings of each observer. Three of the four appear virtually identical regarding the number of cases read as normal and abnormal. One observer (J.R.) tended to read more studies as abnormal and correspondingly fewer as normal. The number of cases read as borderline was similar among observers: mean 23, range 21–27.

Figure 2 illustrates the reading variability between all six possible observer pairs. For this analysis, complete agreement was defined as an identical reading—both normal (N), borderline (B), or abnormal (AB). Complete disagreement was a normal and an abnormal reading. Minor disagreement was defined as one borderline reading combined with either a normal or abnormal reading. Complete observer agreement was similar for all six observer pairs: 67% (range 61–73%). Complete disagreement was

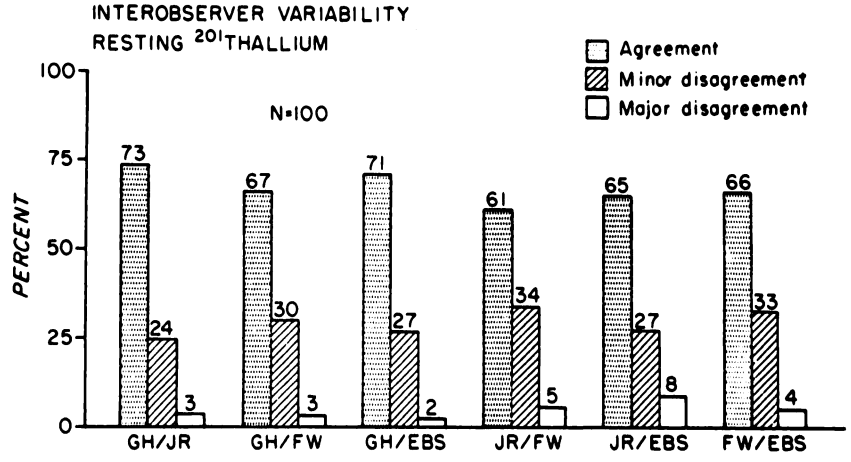

FIG. 2. Percentages of exact agreement, minor disagreement (AB/B, N/B), and major disagreement among the six possible observer pairs. Agreement was consistent among the pairs of readers.

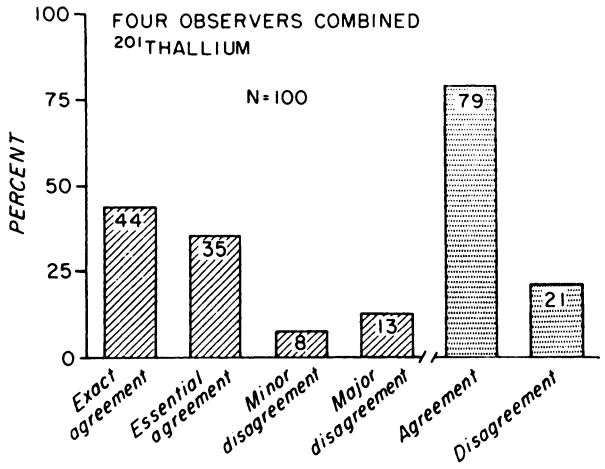

FIG. 3. Percentages of exact agreement (NNNN or AB AB AB AB), essential agreement (NNNB, AB AB AB B, BBBN, BBB AB), minor disagreement (NNBB, AB AB BB), and major disagreement (N/AB). *Agreement* is sum of exact and essential agreements. *Disagreement* is sum of minor and major disagreements.

also similar between all observer pairs (4%; range 2–8%). Neither Fig. 1 nor Fig. 2 suggests institutional bias in the number of cases read as abnormal (or normal) or reader variability.

The variability among all four observers combined is shown in Fig. 3. Exact agreement was defined as four identical readings and essential agreement as three identical readings with one minor disagreement (i.e., NNNB or AB, AB, AB, B). Minor disagreement was recorded when the four observers were equally divided between normal and borderline or borderline and abnormal (N, N, B, B, or B, B, AB, AB). Major disagreement was any combination containing a normal and abnormal reading. Exact agreement between all four observers was recorded in 44% of the 100 studies; essential agreement in 35%; and minor or major disagreement in 8% and 13% of the studies. Overall (Fig. 3), there was agreement in 79% and disagreement in 21%.

Of the 100 studies, two or more observers judged 16 of the studies to be of poor quality (9 Seattle, 7 Amsterdam). Exclusion of these studies did not significantly after the percentage of agreement or disagreement shown in Fig. 3.

Observers from one institution showed less variability when only studies from their institution were included. F.W. and E.B.S. had exact agreement on 74% of their own studies and only 58% of studies from Seattle. Likewise, J.R. and G.H. agreed exactly on 78% of the Seattle studies and 68% of the studies from Amsterdam.

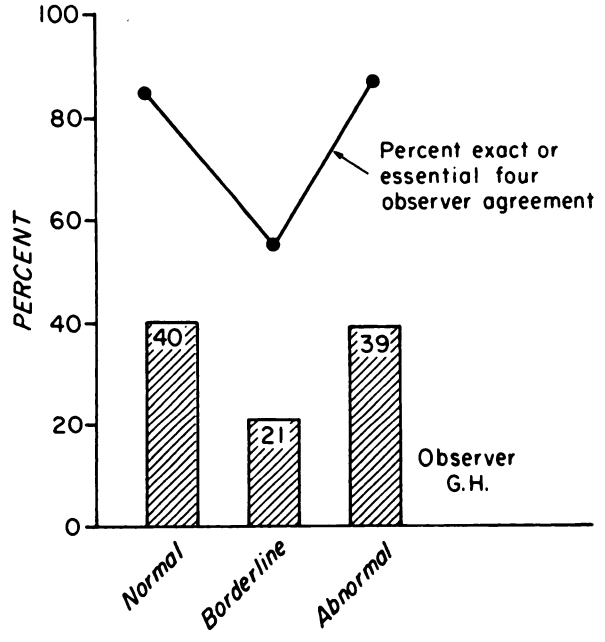

FIG. 4. Percentages of exact or essential four-observer agreement in the studies read as normal, borderline, or abnormal by a single observer (G.H.). The four-observer agreement is higher in group he read as normal or abnormal than group he read as borderline.

DISCUSSION

This study was designed to investigate the degree of observer variability in reading Tl-201 myocardial images. By design, we included patients with both acute and chronic stages of coronary disease in an effort to simulate the types of patient that would be seen in a general nuclear medicine laboratory. Likewise, the ratio of normal to abnormal studies should approximate that generally encountered. Whether the sample selected is truly representative in this regard is difficult to ascertain, but the fraction of normal-to-abnormal cases and the types of disease studied seem to be reasonably typical.

The imaging instrumentation used can affect image quality and could contribute to observer variability. The two gamma cameras were 37-photomultiplier types with good inherent energy and spatial resolution**. More observer variability might be encountered if imaging were done on systems with poorer resolution.

The major conclusion of the study is that observer agreement occurs in about 80% of studies when experienced readers interpret studies blindly. This degree of variability is similar to that previously found with coronary arteriography (7) and liver scans (8).

The overall figure of 79% for exact or essential agreement (Fig. 3) may not apply to any single individual study. In fact, greater observer agreement could be demonstrated for studies read as normal or abnormal than for those read as borderline. This effect is illustrated in Figs. 4 and 5. Note that four-observer agreement was greater with studies that a

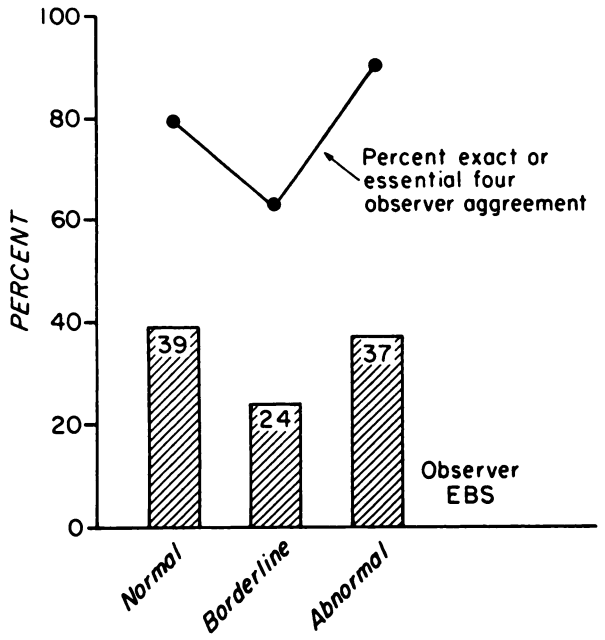

FIG. 5. Counterpart of Fig. 4 for observer E.B.S.: there is greater agreement in reading normal and abnormal scans than there is in reading borderline scans.

given observer read as normal or abnormal than those read as borderline. There was exact or essential four-observer agreement in 34 of the 40 scans read as normal by observer G.H. (Fig. 4); similar agreement was also observed in those scans read as abnormal (34/39; 87%). In those scans he read as borderline, however, agreement was noted in 52% (11/21). Similar (although less marked) findings were noted for observer E.B.S. This effect was found with all four observers and is illustrated in Fig. 6, which shows the average number of cases read as normal, borderline, or abnormal by any one observer and the observer agreement in these categories. Four-observer exact or essential agreement was found in 86% and 83% of studies read as normal or abnormal, and in only 63% of those read as borderline.

A single reader (J.R.) read significantly more abnormals and fewer normals, compared with any of the other three observers ($p = 0.06$; $p = 0.03$; 2 shows no difference in the agreement between $p = 0.0007$)††. Careful inspection of the data in Fig. 2 shows no difference in the agreement between observer pairs, including J.R., compared with those not including this observer. This suggests that this reader shifts normals to borderline, and borderline to abnormal compared with the other observers, resulting in little or no increase in total interobserver variability. In fact, the highest observer-pair agreement (G.H. against J.R., 73%) was found between the two readers with the greatest difference in number of normal readings (40 against 30).

The accuracy of any particular individual or combined study interpretation was not tested. This has been addressed by both participating laboratories (*2,3,9*) and others (*4,10*). Observer variability will be an important factor in efforts to improve the sensitivity and specificity of Tl-201 imaging. This study presents the variability using currently available cameras and unprocessed Polaroid scintiphotos. Presumably, image enhancement using background subtraction (*11*), resolution recovery (*12*), or improved display devices (*13*) would reduce observer variability, but this has not been demonstrated.

The improved observer agreement noted when observers read studies from their own institution suggests unintentional bias. This could also be explained by the different types of patients at the two centers (acute infarction and chronic angina pectoris). It is possible that Amsterdam observers had less variability with scans in suspected acute infarction than chronic angina, owing to greater experience with the former group. Likewise, the Seattle observers had greater experience reading studies in patients with chronic angina. Nonetheless, possible bias cannot be eliminated and would have tended to increase the level of agreement found between observers.

The overall agreement of 80% among four experi-

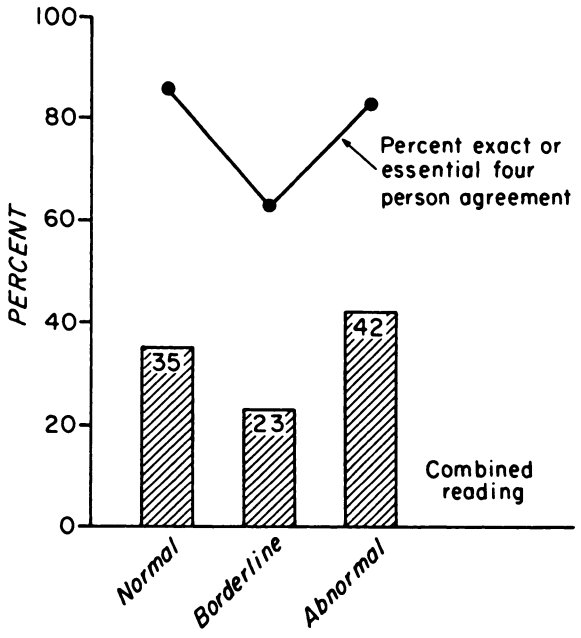

FIG. 6. Percentages of exact or essential four-person agreement for those studies read as normal, borderline, or abnormal by any one reader. Bar graph portrays average percent normal, borderline, and abnormal for the four observers. Above, average percentage of exact or essential agreement are indicated. The data are derived from the information in Figs. 4 and 5, and for the two other observers.

enced observers is lower than one might wish. Furthermore, 23% of the images were considered to contain insufficient diagnostic information and could not be read as either normal or abnormal. This finding may reflect in part the relatively poor quality of the resting Tl-201 image compared to the images following exercise. A study of 101 rest-exercise studies by two of the readers (J.R. and G.H.) however, had a virtually identical interobserver agreement (79%) (*3*). This suggests that Tl-201 studies read from standard unprocessed scintiphotos will exhibit significant observer variability, which will impair clinical application for the detection of coronary heart disease.

FOOTNOTES

* G. H. and J. R. from the Seattle VA Hospital, Seattle, Washington, USA; and F. W. and E. B. S. from the Cardiology and Nuclear Medicine Clinics, Wilhelmina Gasthuis, Amsterdam, The Netherlands.

† Due to differences in the numbering, the observer could tell the source institution.

‡ Tl-201 was obtained from Philips Duphar, Petten, The Netherlands, for studies in Amsterdam; and from New England Nuclear Corp., N. Billerica, Mass., for the Seattle studies.

§ Ohio-Nuclear series 100 gamma scintillation camera.

¶ Model #HRF-251.

** 22% for 80 keV x-rays; 11.8 mm FWHM.

†† Two way analysis of variance for single observations.

ACKNOWLEDGMENTS

This work was supported by the Medical Research Service of the Veterans Administration.

The contribution of Gerard Samson, Ph.D. and J. B. van der Schoot, M.D. is gratefully acknowledged.

REFERENCES

1. Ritchie JL, Hamilton GW, Trobaugh GB, et al: Myocardial imaging and radionuclide angiography in survivors of sudden cardiac death due to ventricular fibrillation: Preliminary report. *Am J Cardiol* 39: 852–857, 1977

2. Hamilton GW, Trobaugh GB, Ritchie JL, et al: Myocardial imaging with intravenously injected Thallium-201 in patients with suspected coronary artery disease: Analysis of technique and correlation with electrocardiographic, coronary anatomic and ventriculographic findings. *Am J Cardiol* 39: 347–354, 1977

3. Ritchie JL, Trobaugh GB, Hamilton GW, et al: Myocardial imaging with ^{201}Thallium at rest and during exercise: Comparison with coronary arteriography and resting and stress electrocardiography. *Circulation* 56: 66–71, 1977

4. Bailey IK, Griffith LSC, Rouleau J, et al: Thallium-201 myocardial perfusion imaging at rest and during exercise: Comparative sensitivity to electrocardiography in coronary artery disease. *Circulation* 55: 79–87, 1977

5. Wackers FJTh, van der Schoot JB, Sokole EB, et al: Noninvasive visualization of acute myocardial infarction in man with Thallium-201. *Br Heart J* 37: 741–744, 1975

6. Parkey RW, Bonte FJ, Stokely EM, et al: Acute myocardial infarction imaged with ^{99m}Tc-stannous pyrophosphate and ^{201}Tl: A clinical evaluation. *J Nucl Med* 17: 771–779, 1976

7. DeRouen TA, Murray JA, Owen W: Variability in the analysis of coronary arteriograms. *Circulation* 55: 324–328, 1977

8. Nishiyama H, Lewis JT, Ashare AB, et al: Interpretation of radionuclide liver images: Do training and experience make a difference? *J Nucl Med* 16: 11–16, 1975

9. Wackers FJTh, Sokole EB, Samson G, et al: Value and limitations of Tl-201 scintigraphy in the acute phase of myocardial infarction. *N Engl J Med* 295: 1–5, 1976

10. Berman DS, Salel AF, DeNardo GL, et al: Noninvasive detection of regional myocardial ischemia using rubidium-81 and the scintillation camera: Comparison with stress electrocardiography in patients with arteriographically documented coronary stenosis. *Circulation* 52: 619–626, 1975

11. Narahara KA, Hamilton GW, Williams DL, et al: Myocardial imaging with Thallium-201: An experimental model for analysis of the true myocardial and background image components. *J Nucl Med* 18: 781–786, 1977

12. Brown DW, Kirch DL, Ryerson TW, et al: Computer processing of scans using Fourier and other transformations. *J Nucl Med* 12: 287–291, 1971

13. Williams DL, Hamilton GW, Ritchie JL, et al: Implementation of an interactive image superimposition algorithm utilizing the display switching capabilities of the Medical Data Systems, Inc. display controller. Proceedings of Seventh Symposium on Sharing Computer Programs and Technology in Nuclear Medicine Computer Assisted Data Processing, January 16–17, 1977, pp 395–399