JACC Vol. 21, No. 5
April 1993:1064–74

Factors Affecting Uniformity in Interpretation of Planar Thallium-201 Imaging in a Multicenter Trial

FRANS J. TH. WACKERS, MD, FACC, MONTY BODENHEIMER, MD, FACC,*
JOSEPH L. FLEISS, PHD,† MARY BROWN, RN, MS‡ AND THE MULTICENTER STUDY ON SILENT MYOCARDIAL ISCHEMIA (MSSMI) THALLIUM-201 INVESTIGATORS§
New Haven, Connecticut and New Hyde Park and Rochester, New York

Objectives. This study was designed to assess factors affecting interobserver agreement in interpretation of planar thallium-201 stress imaging in the Multicenter Study on Silent Myocardial Ischemia (MSSMI).

Background. Five hundred fifty-six planar thallium-201 images were interpreted in 24 clinical centers and in a Radionuclide Core Laboratory. The trial's Coordinating and Data Center observed that the participating clinical centers interpreted a significantly greater number of thallium-201 stress studies as abnormal (i.e., myocardial ischemia or scar) than the Core Laboratory, and overall agreement was poor (kappa 0.27).

Methods. Agreement in image interpretation between clinical centers and the Radionuclide Core Laboratory was analyzed by kappa statistics. The reproducibility of the Core Laboratory results on 41 randomly selected test studies was excellent (kappa 0.77). In contrast, the reproducibility of interpretation in the clinical centers on their own studies was at best fair (kappa 0.45). It was hypothesized that the poor agreement and reproducibility in the clinical centers were caused by lack of standardization of image display and lack of objective criteria for image interpretation. To test the effect of standardization, 13 clinical investigators interpreted the same 41 test studies using 1) uniform image display, and 2) uniform quantification of images.

Results. The agreement in interpretation between clinical investigators and the Radionuclide Core Laboratory improved modestly with uniformity of image display (kappa 0.57) but improved markedly (kappa 0.66) with quantitative circumferential profile analysis.

Conclusions. Lack of standardization in image display and lack of objective criteria for interpretation of thallium-201 images are responsible for suboptimal reproducibility and poor interlaboratory agreement in the interpretation of thallium-201 stress imaging. The adoption of a uniformly accepted method for computer quantification of myocardial perfusion images is crucial to improve agreement in interpretation.

(J Am Coll Cardiol 1993;21:1064–74)

The clinical usefulness of an imaging modality depends on its sensitivity, specificity and predictive accuracy. However, of equal importance is the *reproducibility* of interpretation, not only for the same observer or within the same laboratory, but also in multiple laboratories. For instance, the prognostic value of thallium-201 stress testing has been well documented in reported studies (1–10), but it is not known to what extent these results can be reproduced in various laboratories.

The Multicenter Study on Silent Myocardial Ischemia (MSSMI) is a National Institutes of Health–sponsored multicenter study that aims to evaluate prospectively the prognostic significance of asymptomatic myocardial ischemia in patients who had a recent acute myocardial infarction or unstable angina. This study started in 1988 and enrollment of 1,084 patients was completed in May 1991. In this study the presence of myocardial ischemia was assessed in three ways: 1) ST segment depression on 24-h Holter ambulatory electrocardiographic (ECG) monitoring; 2) ST segment depression on stress testing, and 3) reversible exercise-induced myocardial perfusion defects on planar thallium-201 stress imaging.

The present report examines the agreement in interpretation of thallium-201 stress studies within the participating clinical centers and between these centers and the Radionuclide Core Laboratory. In addition, factors that are important for uniform interpretation of thallium-201 stress images among multiple observers are evaluated. The results indicate a need for standardization of image display and objective criteria for image interpretation.

From the Cardiovascular Nuclear Imaging Laboratory, Department of Diagnostic Radiology and Department of Internal Medicine, Section of Cardiovascular Medicine, Yale University School of Medicine, New Haven, Connecticut; *Division of Adult Cardiology, Long Island Jewish Medical Center, New Hyde Park, New York; †Division of Biostatistics, School of Public Health, Columbia University, New York, New York, and ‡Heart Research Follow-up Program, University of Rochester, Rochester, New York. §A complete list of the study investigators appears in the Appendix. This study was supported in part by grants from the National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland; Mallinckrodt Medical, Inc., Saint Louis, Missouri; Ciba-Geigy Corporation, Summit, New Jersey and Tanabe Seiyaku Comp, Ltd, Osaka, Japan.

Manuscript received March 5, 1992; revised manuscript received October 12, 1992, accepted October 22, 1992.

Address for correspondence: Frans J. Th. Wackers, MD, Yale University School of Medicine, 333 Cedar Street, TE-2, P.O. Box 3333, New Haven, Connecticut 06510.

0735-1097/93/$6.00

Methods

Clinical centers, Radionuclide Core Laboratory and Coordinating and Data Center. Twenty-four clinical centers participated in the MSSMI trial. Each center acquired thallium-201 studies in its own nuclear imaging laboratory using a variety of gamma camera/computer systems. The Cardiovascular Nuclear Imaging Laboratory at Yale University School of Medicine, New Haven, Connecticut served as the central Radionuclide Core Laboratory for uniform processing and analysis of thallium-201 images. A central data archive was established at the MSSMI Coordinating and Data Center at the University of Rochester, Rochester, New York.

Certification procedure preceding MSSMI study. Before centers could enroll patients in the MSSMI study, each of the participating nuclear laboratories was "certified" by the Radionuclide Core Laboratory. The purpose of this certification procedure was to ensure uniformity in image acquisition and acceptable quality of studies and to test the logistics of submitting digital data to the Radionuclide Core Laboratory. Each participating laboratory submitted for certification three thallium-201 stress studies, acquired according to the standardized imaging protocol described next (see below).

Standardization of exercise/imaging protocol. A standardized thallium-201 stress imaging protocol (11,12) was discussed in detail with all participating investigators before the beginning of the MSSMI study.

Exercise. One thousand thirty-two of 1,084 patients performed exercise on a motor-driven treadmill using the standard Bruce protocol, until one of the following end points was reached: target heart rate (220 beats/min minus age), severe fatigue, severe angina, hypotension or ventricular tachycardia. At peak exercise, 2.5 mCi of thallium-201 was injected intravenously and the patient was encouraged to continue to exercise for at least another 1 min at same level, then another minute at a slower speed.

Imaging. Planar myocardial imaging was started within 5 min after injection of thallium-201. Imaging was performed in three projections. A left anterior oblique image was obtained with the patient supine and camera head angulated in the position that provided the best separation between right and left ventricles. An anterior image was obtained with the patient supine and the camera head angulated 45° to the right of the left anterior oblique angulation. The left lateral image was obtained with the patient in a right decubitus position (lying on his or her right side) with the gamma camera angulated in the same way as for the anterior view. The gamma camera was peaked on the 176 keV thallium-201 gamma peak (25% window) and over the 68 keV X-ray peak (20% window). The gamma camera was equipped with a general all purpose parallel hole collimator. All images were acquired for preset time: 10 min per view, acquiring at least 600,000 counts in the field of view. Delayed imaging was performed 2.5 to 3 h later, in the same projections and for the same time. All data were acquired on commercially available computer systems in 128 × 128 matrix (word mode). Unprocessed image data were stored on floppy disk and mailed to the Radionuclide Core Laboratory.

Table 1. Characteristics of Laboratories and Physicians Involved in the MSSMI Study

	All	Test Group
Laboratories (no.)	24	13
MDs (no.)	37	13
Nuclear medicine physicians	23	8
Nuclear medicine cardiologists	14	5
MD experience (yr)	10 ± 5	12 ± 2
Range	1–20	8–15
Median	10	11
Thallium studies/day (no.)	5 ± 3	6 ± 3

Values are expressed as number of the variable or mean value ± SD.

Image interpretation at clinical sites. Thallium-201 stress images were interpreted by the clinical investigators for the purpose of clinical management of the patients using their routine analysis methodology. The experience with thallium-201 imaging in the laboratories and of the interpreting physicians is shown in Table 1. As shown in Table 2 the method of image display and interpretation varied in different laboratories. The investigators were allowed to use clinical and exercise information in the interpretation of the thallium-201 stress tests. The results of interpretation were recorded on MSSMI thallium-201 report forms. On these forms the left ventricle on each view was divided into three segments. Each segment was scored qualitatively as normal, partially reversible, completely reversible, fixed or uninterpretable. Finally, the entire study was scored as *normal*, *ischemia*, *scar* or *ischemia and scar*. The completed report forms were sent to the MSSMI Coordinating and Data Center.

Image processing and interpretation at Radionuclide Core Laboratory. *Transcription and display.* Table 3 shows the various computers and storage media used by the MSSMI centers. In the Radionuclide Core Laboratory the unprocessed digital data from the clinical centers were transcribed to PICKER PCS-512 format. Hard copies of unprocessed

Table 2. Modes of Image Display and Interpretation Employed in 24 Nuclear Laboratories

	Centers (no.)
Display	
Computer screen	10
X-ray film	11
Polaroid film	3
Color	
Black and white	11
Multicolor	3
Combined: color and black and white	10
Interpretation	
Visual analysis only	10
Combined quantification and visual analysis	14

Table 3. Computer Systems and Storage Media Employed in the MSSMI Thallium-201 Imaging Study

Computer	Media
ADAC	5.25" disk
Baird	Polaroid film
CDA	5.25" disk
DEC	8" disk
Elscint	5.25" and 8" disks
G E Star, Starcam	8" disk
MDS	8" disk
Picker	5.25" disk
Sophy	5.25" and 8" disks
Technicare	8" disk
Toshiba	8" disk and magnetic tape
VAX	Magnetic tape

images were printed on black and white photographic paper with use of a video imager. On these images radioactivity is white against a black background. A linear gray scale was employed. The images were normalized to the pixel with the greatest amount of radioactivity within the heart. Exercise and delayed images were displayed side by side on high quality glossy prints (Fig. 1).

Assessment of image quality. The quality of thallium-201 images was evaluated subjectively in the Core Laboratory according to the following criteria: 1) adequate count density in the total field of view (at least 600,000 counts) and adequate count density within the heart (at least 30,000 counts after background correction); 2) acceptable "appearance" of the images (adequate heart to background ratio and adequate image resolution); 3) adequate positioning and reproducible repositioning of the heart within the field of view; 4) complete visualization of all myocardial segments (i.e., true left anterior oblique, anterior and left lateral images of the heart); 5) adequate size of the heart within the field of view when zooming was employed. The heart should occupy approximately one third to one fourth of the diameter of the field of view.

Taking the preceding listed features into account, the quality of thallium-201 study was categorized as either excellent, good, fair or unacceptable.

Quantification. Quantification of thallium-201 images was performed by using software developed in the Radionuclide Core Laboratory. This computer program has been described previously (13). In brief, after modified interpolated background subtraction left ventricular regions of interest were determined on the stress and delayed thallium-201 images. Circumferential count distribution profiles were generated from these regions of interest. The left ventricle was divided into 36 sectors, each subtending a 10° arch. The mean count density in each of the 36 sectors was determined, and the sector with maximal mean count density was designated the value of 100%. The values for the remaining sectors were expressed as a percent of this maximum.

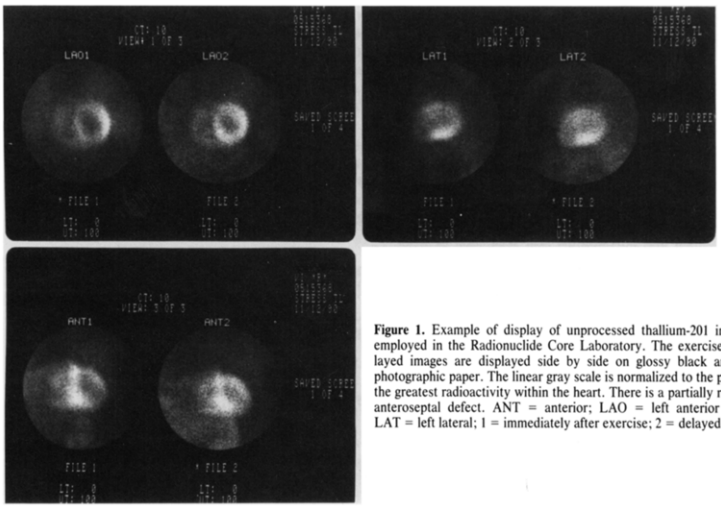

Figure 1. Example of display of unprocessed thallium-201 images as employed in the Radionuclide Core Laboratory. The exercise and delayed images are displayed side by side on glossy black and white photographic paper. The linear gray scale is normalized to the pixel with the greatest radioactivity within the heart. There is a partially reversible anteroseptal defect. ANT = anterior; LAO = left anterior oblique; LAT = left lateral; 1 = immediately after exercise; 2 = delayed imaging.

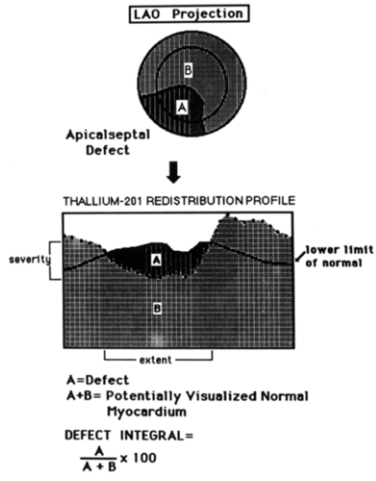

Figure 2. Method of quantifying perfusion defect size. **Top**, Diagram of the left ventricle in the left anterior oblique (LAO) projection: an apical-septal defect (A) is indicated. **Bottom**, Thallium-201 distribution profile. The apical-septal myocardial perfusion defect is graphically displayed as the portion of the circumferential profile below the lower limit of normal (A). The defect can be described in terms of extent (the number of data points below the lower limit of normal) as well as severity (the nadir of the curve below the lower limit of normal). Defect size is quantified as an integral of the defect area (A) and the potentially visualized normal myocardium (A + B). (Reproduced from Wackers et al. J Am Coll Cardiol 1989;14:861–73).

Myocardial defect size was quantitated by integrating the hypoperfused area below the lower limit of normal curve (mean minus 2 SD). This area was expressed as a proportion (×100) of the total, potentially visualized, normal myocardium. This integral is a value without units and reflects both the extent and severity of a perfusion defect (Fig. 2). This method has been validated previously in patients (14). The inter- and intraobserver variability of this methodology for quantifying thallium-201 defects has been reported (15).

Interpretation. All thallium-201 studies were interpreted in the Radionuclide Core Laboratory by two of us (F.J.Th.W. and M.B.) without knowledge of clinical information, the interpretation at the clinical centers or the results of exercise and of 24-h Holter ECG recordings. Two sets of data were available: 1) prints on glossy photographic paper of unprocessed digital exercise/delayed thallium-201 images (Fig. 1), and 2) quantification of these images as circumferential count distribution profiles with a reference normal data base (Fig. 3).

The interpretation of images was recorded on Radionuclide Core Laboratory MSSMI report forms. On these forms the left ventricle on each view was divided into five segments. The distribution of thallium-201 in each segment was scored qualitatively and measured quantitatively. The qualitative score method was identical to that used in the clinical centers. The quantitative defect size on the exercise image and on the delayed image, and the quantitative difference between the two, were measured as outlined earlier.

The final interpretation of a thallium-201 study was based on analysis of *quantitative circumferential profiles* with visual overread.

Apparent discordances between computer quantification and analog images were resolved by consensus. Discordances were rare and occurred predominantly in small defects (<5) or on studies with obvious attenuation artifacts (e.g., by overlying breast tissue).

The studies were categorized as either *normal*, *ischemia*, *scar* or *ischemia and scar*. The completed Radionuclide Core Laboratory report forms were sent to the MSSMI Coordinating and Data Center.

Assessment of reproducibility of interpretation in the Radionuclide Core Laboratory. Forty-one studies (from 10 clinical centers) were selected by stratified random selection from the data base by the Coordinating and Data Center for reinterpretation by the Radionuclide Core Laboratory. Studies previously graded by the Core Laboratory as being of "unacceptable" quality were excluded. These studies were originally interpreted by the Core Laboratory as *normal* in 14, *scar* in 12, *ischemia* in 8 and *scar and ischemia* in 7. The selected studies (test studies) were coded by assigning numbers 1 to 41. The unprocessed analog images and their quantitative circumferential profiles were reinterpreted by F.J.Th.W. and M.B. without knowledge of the prior interpretation. New MSSMI report forms were completed and submitted to the Data Coordinating Center.

Assessment of reproducibility of interpretation in the clinical centers. Eight clinical centers that had entered more than 20 thallium-201 stress studies in the MSSMI study were asked by the Coordinating and Data Center to reread 17 to 20 of their own, randomly selected studies. A total of 156 studies were reinterpreted by the eight investigators without knowledge of the original interpretation, first *without* clinical information (as performed at the Core Laboratory), and then *with* clinical information available (as routinely performed at the clinical centers). These two interpretations were made in the same reading session, one after the other. New MSSMI report forms were filled out by the clinical centers and submitted to the Coordinating and Data Center.

Evaluation of factors affecting agreement in interpretation. Because the clinical centers used different methods for image display and image interpretation (Table 2), it was hypothesized that agreement in interpretation of thallium-201 images could be improved by 1) identical display of

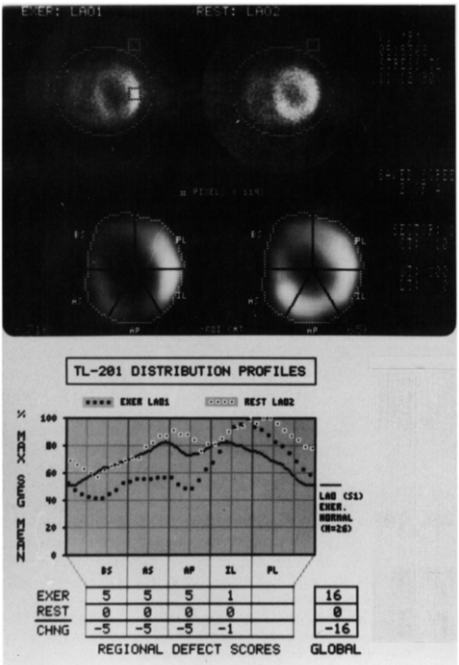

Figure 3. Quantification of the left anterior oblique (LAO) images in Figure 1. **Top,** Analog images with reference ellipse employed for generating interpolative background correction. The **small squares** are regions of interest to compute lung/heart ratio (normal [0.49] in this patient). **Bottom,** Circumferential count distribution profiles. The **solid black line** indicates the lower limit of normal thallium (TL)-201 distribution. The exercise (EXER) profile (**black squares**) is below the lower limit of normal in the basal-septal (BS), apical-septal (AS) and apical (AP) area. The delayed profile (**white squares with dot**) is largely within the normal range. Quantification of the defect size after exercise, at delayed imaging (REST) and change (CHNG) is shown in the table. IL = inferolateral; % MAX SEG MEAN = percent of maximal segment (mean counts). PL = posterolateral.

images; 2) identical quantitative data; and 3) comparison with a normal reference data base.

Identical image display visual analysis. Thirteen MSSMI thallium-201 interpreters from 13 centers (including the eight centers that assessed the reproducibility of their own studies) were asked to interpret the 41 thallium-201 *test studies.* Numerically coded high quality black and white photographic prints, showing unprocessed analog exercise and delayed images side by side (Fig. 1), were sent to the investigators. The investigators interpreted these images qualitatively by visual analysis. No clinical or quantitative data were made available. The results were recorded on new MSSMI report forms. Each study was to be categorized again as *normal, ischemic, scar* or *ischemia and scar.*

Identical image quantification and normal reference data. Within 6 months the same 13 investigators met for a reading session. During this meeting the investigators were familiarized with the quantitative circumferential profile analysis program employed in the radionuclide Core Laboratory. Examples of quantitative thallium-201 studies were shown, and guidelines and rules for interpretation were discussed. Subsequently, the 13 investigators interpreted the 41 *test studies* from 1) black and white photographic prints of unprocessed images, and 2) from circumferential count distribution profiles with a normal reference data base. The investigators had neither clinical information nor the results of their visual analysis available. The interpretation derived from quantitative analysis was recorded on new MSSMI forms. The final interpretation of each study was again recorded as *normal, ischemic, scar* or *ischemia and scar.*

Statistical analysis. Analysis of agreement was performed on 4 × 4 tables displaying the final categorization of the three-view studies as *normal, ischemia, scar* or *ischemia and scar.* No analysis was performed on the segmental image interpretation.

Agreement between the Core Laboratory and the clinical

Table 4. Interpretation of Kappa Values

+1	Complete agreement
>0.75	Excellent agreement beyond chance
0.40 to 0.75	Fair agreement beyond chance
<0.40	Poor agreement beyond chance
0	Chance agreement
<0 but >−1	Disagreement beyond chance
−1	Complete disagreement

centers and agreement between repeat analyses were evaluated by kappa statistics (16). The interpretations of different kappa values are indicated in Table 4 (17). Kappa values were determined for three comparisons: 1) each of four categories: *normal, ischemia, scar* or *ischemia and scar*; 2) normal versus abnormal; and 3) ischemia (including *ischemia and scar*) versus no ischemia. Differences between the Core Laboratory and the clinical centers in the proportions of studies classified as normal and in proportions classified as ischemic (alone or with scar) were tested by using McNemar's chi-square test with 1 degree of freedom (18).

Results

Quality of thallium-201 images before certification and during the MSSMI study. During the certification process (before the start of the MSSMI study) 24 clinical centers submitted 75 thallium-201 studies. The Radionuclide Core Laboratory judged that 19 (25%) of these studies from seven laboratories were of unacceptable quality (Table 5). The most common problems were 1) poor count statistics, 2) inadequate image resolution due to suboptimal energy window placement or too large distance between patient and camera, 3) failure to carefully reproduce the positioning angles between the exercise and delayed images, 4) incorrect angulation resulting in incomplete visualization of all myocardial segments, and 5) too large a zoom factor. In all instances these problems could be readily corrected by correspondence and telephone discussions between clinical centers and the Radionuclide Core Laboratory.

During the MSSMI study the clinical centers adhered well to the standardized imaging protocol. At the completion of the study (May 1991), 95% of all thallium-201 studies were considered to be of fair or better quality (Table 5); only 5% of studies were categorized as unacceptable.

Table 5. Quality of Thallium Images at Certification and During the MSSMI Study

Image Quality	March 1988 Certification (75 studies)	May 1991 MSSMI Study (1,032 studies)
Excellent	23	11
Good	29	60
Fair	23	24
Unacceptable	25	5

Data are presented as percent of studies.

Table 6. Agreement in Interpretation Between Radionuclide Core Laboratory and Clinical Centers (n = 556 patient studies)

	Clinical Centers				
	Normal	Scar	Ischemia	Scar and Ischemia	Total
Core Laboratory					
Normal	72	56	37	32	197
Scar	6	69	17	38	130
Ischemia	7	11	36	37	91
Scar and ischemia	0	37	28	73	138
Total	85	173	118	180	556
Agreement	Kappa				
All four categories	0.27				
Normal or abnormal	0.38				
Ischemia or no ischemia	0.36				

Comparison of interpretation by Radionuclide Core Laboratory and by clinical centers. After enrollment of approximately 50% of the MSSMI patients, the MSSMI Data Coordinating Center performed an interim analysis of the agreement on thallium study interpretation between clinical centers and the Radionuclide Core Laboratory (Table 6). Of the 556 studies, 197 studies (35%) were interpreted as normal by the Core Laboratory, compared with only 85 studies (15%) by the enrolling centers ($p < 0.001$). Furthermore, the Core Laboratory read 229 (41%) of the studies as showing *ischemia* or *scar and ischemia*, compared with 298 (54%) by the clinical centers ($p < 0.001$). Kappa (k) values indicated poor agreement whether studies were classified into one of four categories ($k = 0.27$) or dichotomized as normal versus abnormal ($k = 0.38$), or as ischemia versus no ischemia ($k = 0.36$). Table 7 shows that the agreement on interpretation of 156 patient studies by the eight clinical centers selected to reread their own thallium-201 studies and by the Core Laboratory was equally poor and no apparent bias existed. The relatively poor agreement for this sample of 156 thallium-201 studies was similar to that observed for the total group of 556 patients. (Kappa values were calculated for

Table 7. Agreement in Interpretation Between Radionuclide Core Laboratory and Eight Clinical Centers (n = 156 patient studies)

	Clinical Centers				
	Normal	Scar	Ischemia	Scar and Ischemia	Total
Core Laboratory					
Normal	23	12	6	9	50
Scar	2	22	4	20	48
Ischemia	1	1	5	13	20
Scar and ischemia	0	13	6	19	38
Total	26	48	21	61	156
Agreement	Mean Kappa ± SD		Median (range)		
All four categories	0.24 ± 0.15		0.21 (0.01 to 0.49)		
Normal or abnormal	0.52 ± 0.26		0.57 (0.00 to 0.78)		
Ischemia or no ischemia	0.28 ± 0.17		0.32 (0.01 to 0.47)		

JACC Vol. 21, No. 5
April 1993:1064–74

Table 8. Reproducibility of Interpretation in Radionuclide Core Laboratory (41 test studies)

	Reading 2				
	Normal	Scar	Ischemia	Scar and Ischemia	Total
Reading 1					
Normal	13	1	0	0	14
Scar	2	8	0	2	12
Ischemia	0	0	7	1	8
Scar and ischemia	1	0	0	6	7
Total	16	9	7	9	41
Agreement	Kappa				
All four categories	0.77				
Normal or abnormal	0.79				
Ischemia or no ischemia	0.84				

each of the eight clinical centers and summary statistics determined for the set of eight kappa values).

The agreement (between the Core Laboratory and clinical centers) in the original interpretation of the selected *41 test studies* was better than that for all studies (unacceptable quality studies were excluded) but still suboptimal. Kappa values for agreement on all four categories was 0.47, for *normal versus abnormal* 0.52, and for *ischemia versus no ischemia* 0.53.

Reproducibility of interpretation in Radionuclide Core Laboratory. Repeat interpretation of the 41 *test studies* by the Core Laboratory showed excellent reproducibility (Table 8). The kappa values all exceeded the threshold for excellent agreement (17).

Reproducibility of interpretation in clinical centers. Eight centers reread the 156 patient studies. The results of intracenter agreement are summarized as kappa values in Table 9. When the original interpretation in the centers was compared with the rereading without access to clinical and exercise information, the kappa value for categorization as *normal or abnormal* was good (0.70 ± 0.13). However, the kappa values for categorizations as either *ischemia or no ischemia* or into one of the four categories were only fair (0.50 ± 0.28 and 0.45 ± 0.21, respectively). When clinical information, such as the history and exercise ECG data, were made available, reproducibility did not improve discernibly (Table 9). Nevertheless, there was excellent agreement between reinterpretations first without and then with clinical information.

Table 9. Reproducibility of Interpretation in Clinical Centers (eight clinical centers)

	Mean Kappa Value (±SD)		
	Without Clinical Information*	With Clinical Information*	Without Versus With Clinical Information
All four categories	0.45 ± 0.21	0.44 ± 0.22	0.75 ± 0.16
Normal or abnormal	0.70 ± 0.13	0.71 ± 0.23	0.89 ± 0.14
Ischemia or no ischemia	0.50 ± 0.28	0.54 ± 0.24	0.78 ± 0.16

*Compared with original interpretation with clinical information.

Table 10. Agreement in Interpretation Between Radionuclide Core Laboratory and 13 Investigators, Using Uniform Image Display (41 test studies)

	Clinical Investigators				
	Normal	Scar	Ischemia	Scar and Ischemia	Total
Core Laboratory					
Normal	139	21	17	5	182
Scar	11	116	3	26	156
Ischemia	13	11	50	30	104
Scar and ischemia	1	22	10	58	91
Total	164	170	80	119	533
Agreement	Mean Kappa ± SD		Median (range)		
All four categories	0.57 ± 0.17		0.59 (0.29 to 0.90)		
Normal or abnormal	0.70 ± 0.18		0.71 (0.29 to 0.94)		
Ischemia or no ischemia	0.61 ± 0.14		0.60 (0.39 to 0.89)		

Effect of standardization on interpretation. *Standardization of display.* Table 10 and Figure 4 show the effect of uniform display of 41 test images on the agreement between 13 clinical investigators and the Radionuclide Core Laboratory. For categorization as *normal or abnormal*, kappa value improved from 0.53 (original agreement test studies, see earlier) to 0.70. However, there was a wide range in the kappa values (0.29 to 0.94). For categorization as *ischemia or no ischemia*, kappa value improved from 0.52 to 0.61, and for categorization in one of four categories it improved from 0.47 to 0.57.

Standardization of display and quantification. Table 11 and Figure 4 show the further improvement in agreement between the clinical investigators and the Radionuclide Core Laboratory obtained by adding quantification (with normal reference data base) to standardized image display. For categorization as *normal or abnormal*, kappa value improved to 0.80. For categorization as *ischemia or no ischemia*, kappa value improved to 0.71. The kappa value for the comparison of all four possible categories further improved to 0.66. This improved performance appears to be associated with smaller standard deviations and narrower range of kappa values.

The effect of quantification can be appreciated by comparing Tables 10 and 11. The addition of quantification achieved better agreement with the Core Laboratory because the investigators interpreted more studies as *normal* (31% vs. 36%) and fewer defects as *scar* (32% vs. 26%). The frequency of interpretation of *ischemia* and *ischemia and scar* remained essentially the same.

Discussion

This study demonstrates that among centers with well established radionuclide laboratories, the methodology of interpreting thallium-201 stress varies considerably.

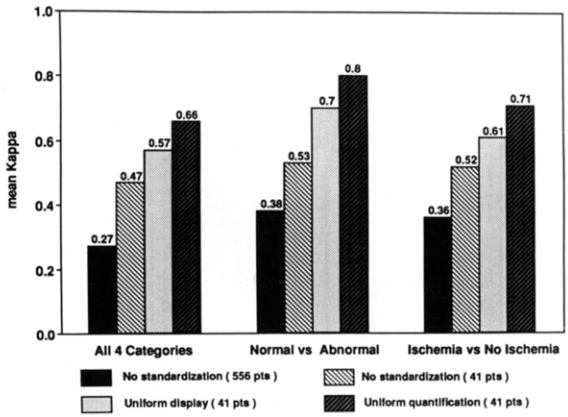

Figure 4. Mean kappa values for categorization as "all four categories" (normal, ischemia, scar, ischemia and scar), "normal versus abnormal" and "ischemia versus no ischemia". Kappa values indicate agreement between the Radionuclide Core Laboratory and clinical centers without standardization (all 556 patients [pts] and 41 test patients) and uniform image display and uniform image quantification (41 test patients). Kappa values improve considerably with uniformity of display and quantification.

Variation in image acquisition and image quality. The initial certification procedure for the MSSMI study aimed to optimize image quality by standardizing image acquisition (11,12). Our system for grading image quality was by nature subjective and reflects our esthetic preference. The majority of participating laboratories submitted studies that were of fair or better quality (Table 5). However, 25% of the initially submitted images were of unacceptable quality. It was clear that these suboptimal images were not caused by inadequate equipment but rather by deficiencies in imaging technique and, most frequently, by lack of attention to technical details of imaging. The most common deficiency was that of insufficient count density within the heart, resulting in poor signal to noise ratio. These and other deficiencies in general could readily be corrected by communication between the Radionuclide Core Laboratory and the clinical centers. As a result, the overall quality of thallium-201 studies at the completion of the MSSMI study was good to excellent. Only 5% of studies were of poor quality (Table 5).

Table 11. Agreement in Interpretation Between Radionuclide Core Laboratory and 13 Investigators Using Uniform Image Quantification (41 test studies)

	Clinical Investigators				
	Normal	Scar	Ischemia	Scar and Ischemia	Total
Core Laboratory					
Normal	162	10	6	3	181
Scar	12	113	7	24	156
Ischemia	17	7	58	22	104
Scar and ischemia	1	8	15	67	91
Total	192	138	86	116	532*
Agreement	Mean Kappa ± SD		Median (range)		
All four categories	0.66 ± 0.10		0.70 (0.46 to 0.76)		
Normal or abnormal	0.80 ± 0.07		0.77 (0.70 to 0.95)		
Ischemia or no ischemia	0.71 ± 0.11		0.70 (0.53 to 0.89)		

*One study was not interpreted by one investigator.

Variation in interpretation. Our study demonstrates further that good quality of thallium-201 images does not ensure uniform interpretation of images in multiple laboratories. The Coordinating and Data Center noticed that a significantly smaller proportion of thallium-201 studies were interpreted as showing ischemia by the Core Laboratory than by the clinical centers (Tables 6 and 7).

Reproducibility. Because no reference standard for ischemia on thallium-201 imaging exists, it was not clear whether the Core Laboratory or the clinical centers were correct. To examine this issue, we first examined the reproducibility of interpretation. The reproducibility by the Radionuclide Core Laboratory was found to be excellent (Table 8). In contrast, the reproducibility at eight representative clinical centers was no better than fair (Table 9). It made no apparent difference whether the interpreters were nuclear medicine physicians or nuclear medicine cardiologists, nor did the experience (in years or patient volume) of the laboratories with thallium-201 imaging appear to be a factor. The observed lack of agreement conceivably could be explained in several ways. Whereas image display and quantitative analysis at the Core Laboratory was rigorously standardized, at various clinical centers thallium-201 images were interpreted in many different ways: from Polaroid film, X-ray film, computer screen; in color or in black and white; by visual

inspection or by using quantitative analysis software. Furthermore, in several centers, the studies were not consistently interpreted by the same physicians in part because interpreting physicians employed a rotating reading schedule or moved to other institutions. Such changes can be expected to occur in any multicenter study that recruits patients over the course of several years.

Previous studies on observer agreement. Data on intraobserver and interobserver agreement in interpreting thallium-201 studies have been reported before. Okada et al. (19) and Atwood et al. (20) reported good observer agreement when studies were interpreted in a simple dichotomous fashion (normal or abnormal) with kappa values ranging from 0.56 to 0.74. However, percent agreement ranged from 11% to 79% when interpreters were asked to read the anatomic location of defects. Numerous other investigators (21–25) have reported observer reproducibility as part of studies on the diagnostic and clinical usefulness of thallium-201 stress testing. In all instances the investigators interpreted studies from their own laboratory employing their usual display format. Watson et al. (26) reported recently on quantitative computer-assisted analysis of planar technetium-99m sestamibi studies acquired in a different institution. They found high (90.2%) agreement between interpreters and between computer operators. Only one study compared agreement between observers of different institutions. Trobaugh et al. (27) evaluated agreement in interpretation among four interpreters from two different laboratories on thallium-201 images acquired in their respective laboratories. The images were interpreted visually as either normal, borderline or abnormal. Agreement in qualitative interpretation of studies occurred in 79% of studies.

The agreement between the Core Laboratory and clinical centers as reported in our present study appears to be substantially less than that reported before. However in most previous studies no kappa statistics were applied. When agreement is expressed as a percentage, results may be deceiving. For example, in Table 6, agreement on *normal or abnormal* was 75%, and the agreement on all four categories was 45%; both values are substantially higher than the corresponding kappa values. Moreover, our present study is unique because of the more detailed image interpretation and the comparison of relatively large number of laboratories participating in this multicenter trial.

Clinical information. It appeared also conceivable that the availability of clinical information might have biased image interpretation in the clinical centers. Surprisingly, clinical information appeared to have little effect on interpretation at the clinical centers (Table 9). With or without clinical information the agreement with the Core Laboratory was poor. However, at rereading, good intraobserver agreement existed at the clinical centers with and without utilization of clinical information. These observations suggest that different criteria for normality and abnormality were employed.

Effect of uniform image display. We hypothesized that the poor agreement between the clinical centers and Radionuclide Core Laboratory and the lack of intrainstitutional reproducibility could at least in part be due to a lack of standardization of data display. When the same investigators who demonstrated poor reproducibility on their own studies were asked to interpret thallium-201 test images using *uniform image display*, agreement with the Core Laboratory interpretations improved (Table 10).

Effect of uniform image display and quantification. Agreement in interpretation of thallium-201 images improved further by the addition of uniform quantification of defect size and comparison with a normal reference data base. Such quantification provides objective criteria for normality and abnormality, as well as for reversibility of defects. Kappa values indicated excellent agreement with the Core Laboratory and, as a corollary, excellent agreement among the clinical investigators. The 13 clinical investigators did not achieve the same high kappa values for agreement that the Core Laboratory demonstrated for reproducibility. This finding is explained by the extensive experience of the Core Laboratory with this quantitative method. Nevertheless, the clinical investigators achieved their considerable improvement in agreement after one teaching session of approximately 1 h.

The observations in this multicenter study are not unique. We have experienced similar variability in study quality and image interpretation in other multicenter clinical trials (Thrombolysis in Myocardial Infarction [TIMI] and Survival and Ventricular Enlargement Study [SAVE]), whether myocardial perfusion imaging or equilibrium radionuclide angiocardiography was involved. It appears however that thallium-201 imaging in particular lacks standardization of image acquisition and objective criteria for image interpretation.

The aim of the present report was not to demonstrate that the quantitative software used at the Radionuclide Core Laboratory is the superior methodology. After all, it is not possible to know whether the interpretation of the Core Laboratory is the correct one. However, the results do suggest that *standardization of image display* and the use of a *quantitative software with reference to a normal data base*, are instrumental in achieving uniform and reproducible interpretation of thallium-201 images in different laboratories. The varying methodology of image display and variable criteria for image interpretation reflect current reality in many laboratories.

Quantitative image analysis. Quantitative analysis of myocardial perfusion imaging has been employed in our laboratory *routinely* for >10 years. It should be emphasized that the results of computer quantification are not accepted slavishly. Computer quantification is an invaluable aid to visual analysis. In most cases the graphic display *confirms* the impression from the analog images and is employed as objective feedback to the interpreter. Quantification with reference to a *normal data base* enhances *consistency and reproducibility* of image interpretation. Quantification is

particularly helpful in questionable studies with minor abnormalities or minor defect reversibility. Quantification enhances the confidence of interpretation by objective feedback through graphic display. However, the computer does not recognize artifacts. Potential artifacts are recognized by inspection of the analog images, and computer quantification is "overread."

Investigative implications. Our observations have important investigative implications. In recent years various multicenter trials have used Radionuclide Core Laboratories. These laboratories are important for designing a uniform imaging protocol, for ascertaining acceptable quality of radionuclide studies before the start of the trial and for quality control during the trial. The ultimate goal is to collect consistently good quality data and to perform uniform data processing and analysis.

It has been suggested that such Core Laboratories are too costly and that radionuclide studies could equally well be analyzed at each of the participating clinical centers. The results of analysis at the clinical centers could then be submitted to a central data coordinating center. Our study indicates that, because of a wide variety in computer and acquisition equipment and variety in criteria for image interpretation, this may well be an illusion and that such a study design carries a significant risk of collecting nonuniform data. This could potentially seriously endanger the validity of conclusions drawn at the completion of the entire trial.

Clinical implications. The demonstrated lack of uniform standards and mediocre reproducibility of the interpretation of thallium-201 stress studies also may have direct practical relevance with regard to clinical usefulness of thallium-201 imaging and its impact on patient management. If the interpretation of thallium-201 stress studies lacks reproducibility, the test may not provide consistently useful information for clinical practice. In recent years numerous investigators have demonstrated the prognostic value of certain characteristics of thallium-201 stress imaging, such as the number and extent of reversible defects and the presence of increased lung uptake (10). To apply these observations to the general population of patients with coronary artery disease, it is crucial that these aspects of imaging be identified accurately and reproducibly in *all* nuclear laboratories. Our results indicate that this may not always be the case.

Imaging agents. A major limitation of myocardial imaging with thallium-201 is the relatively poor count density obtained within the heart because of the relatively low dose (2 to 3 mCi) that can be administered. Adequate count density is important for optimal planar imaging (11,12), but even more important for optimal single-photon emission computed tomographic (SPECT) imaging. Suboptimal quality perfusion images may be difficult to interpret, particularly when abnormalities are of moderate degree. Therefore, the confidence in interpretation and consequently reproducibility and agreement in interpretation among observers is adversely affected. The new ^{99m}Tc-labeled myocardial perfusion imaging agents allow administration of 25 to 30 mCi/study (28–30). The resulting high count density may aid substantially in improving overall images quality and thus improve consistency of interpretation of myocardial perfusion imaging.

Tomographic imaging. The findings in the present study apply to planar thallium-201 imaging. The majority of nuclear cardiology laboratories perform at the present time in addition to planar imaging, SPECT myocardial perfusion imaging. Tomographic imaging is more demanding than planar imaging (31,32). Acquisition, display and processing methods for SPECT also vary markedly in different laboratories and for different camera and computer systems. There is no reason to believe that tomographic imaging would be less affected than is planar imaging by the factors described in the present study. Recently, considerable effort has been directed toward optimizing and standardizing of SPECT imaging with ^{99m}Tc-sestamibi (33). For uniformity in interpretation of myocardial perfusion images, the need for standardization applies to both planar and tomographic imaging.

Conclusions. Our study indicates a need for a uniformly accepted protocol for all aspects of myocardial perfusion imaging. In our experience, acquisition protocols can be standardized in multiple laboratories with relatively little effort (11,12). Reproducible interpretation of myocardial perfusion images in multiple laboratories can be achieved by using a uniformly accepted method of quantitative image processing, reference to a normal data base and objective interpretative criteria.

The secretarial assistance of Irene Courtmanche is greatly appreciated. The tireless efforts of Pat Severski, MS in retrieving and tabulating data from the computer files at the MSSMI Coordinating and Data Center, University of Rochester, Rochester, New York were crucial for the preparation of this report, and her involvement is gratefully acknowledged. We thank Barry L. Zaret, MD for reviewing the manuscript and his invaluable comments and suggestions.

Appendix

Multicenter Study on Silent Myocardial Ischemia (MSSMI) Participants

MSSMI Thallium Investigators

Jesaia Benhorin, MD, Bikur Cholim Hospital, Jerusalem, Israel; David Blood, MD, Presbyterian Hospital, New York, NY; Robert Case, MD, St. Lukes/Roosevelt Hospital, New York, NY; Thomas Challis, MD, Kingston General Hospital, Kingston, Ontario, Canada; Paul Chandysson, MD, Washington Hospital Center, Washington, DC; Gordon DePuey, MD, St. Lukes/Roosevelt Hospital, New York, NY; Keith Fischer, MD, The Jewish Hospital, Saint Louis, MO; John Gillespie, MD, Highland Hospital, Rochester, NY; Robert E. Goldstein, MD, Uniformed Services University of the Health Sciences, Bethesda, MD; Henry M. Greenberg, MD, St. Lukes/Roosevelt Hospital, New York, NY; Alvin Greengart, MD, Maimonides Hospital, Brooklyn, NY; Eugenio Inglese, MD, Ospedale Maggiore-della Carita, Novara, Italy; James Korsten, MD, Over-

look Hospital, Summit, NJ; Yasuyuki Nakamura, MD, Kyoto University, Kyoto, Japan; Bob Shah, MD, Walter Reed Army Medical Center, Washington, DC; James F Walroth, MD, Andrews Air Force Base, MD; Walter Williams, MD, University of Arizona, Tucson, AZ;

Nuclear Coordinators

Monty Bodenheimer, MD, Long Island Jewish Medical Center, Hyde Park, NY; Mary Brown, RN, MS, University of Rochester, Rochester, NY; Joseph Fleiss, PhD, Columbia University, New York, NY; W. Jackson Hall, PhD, University of Rochester, Rochester, NY; Arthur J. Moss, MD, University of Rochester, Rochester, NY; Ronald Schwartz, MD, University of Rochester, Rochester, NY; Frans J. Th. Wackers, MD, Yale University School of Medicine, New Haven, CT.

References

1. Brown KA, Boucher CA, Okada RD, et al. Prognostic value of exercise thallium-201 imaging in patients presenting for evaluation of chest pain. J Am Coll Cardiol 1983;4:994–1001.
2. Gibson RS, Watson DD, Craddock GB, et al. Prediction of cardiac events after uncomplicated myocardial infarction: a prospective study comparing predischarge exercise thallium-201 scintigraphy and coronary angiography. Circulation 1983;68:321–36.
3. Wackers FJTh, Russo DJ, Russo D, Clements JP. Prognostic significance of normal quantitative planar thallium-201 stress scintigraphy in patients with chest pain. J Am Coll Cardiol 1985;6:27–30.
4. Pamelia FX, Gibson RS, Watson DD, Craddock GB, Sirowatka J, Beller GA. Prognosis with chest pain and normal thallium-201 exercise scintigrams. Am J Cardiol 1985;55:920–6.
5. Wahl M, Hakki HA, Iskandrian AS. Prognostic implications of normal exercise thallium 201 images. Arch Intern Med 1985;145:253–6.
6. Koss JH, Kobren SM, Grunwald AM, Bodenheimer MM. Role of exercise thallium-201 myocardial perfusion scintigraphy in predicting prognosis in suspected coronary artery disease. Am J Cardiol 1987;59: 531–4.
7. Ladenheim ML, Pollock BH, Rozanski A, et al. Extent and severity of myocardial hypoperfusion as predictors of prognosis in patients with suspected coronary artery disease. J Am Coll Cardiol 1986;7:464–71.
8. Brown KA, Weiss RM, Clements JP, Wackers FJTh. Usefulness of residual ischemic myocardium within prior infarct zone for identifying patients at high risk late after acute myocardial infarction. Am J Cardiol 1987;60:15–9.
9. Gill JB, Ruddy TD, Newell JB, Finkelstein DM, Strauss HW, Boucher CA. Prognostic importance of thallium uptake by the lungs during exercise in coronary artery disease. N Engl J Med 1987;317:1485–9.
10. Kaul S, Finkelstein DM, Homma S, Leavitt M, Okada RD, Boucher CA. Superiority of quantitative exercise thallium-201 variables in determining long-term prognosis in ambulatory patients with chest pain: a comparison with cardiac catheterization. J Am Coll Cardiol 1988;12:25–34.
11. Wackers FJTh. Myocardial perfusion imaging. In: Gottschalk A, Hoffer PB, Potchen EJ, eds. Golden's Diagnostic Radiology. 2nd ed., Diagnostic Nuclear Medicine. Baltimore: Williams & Wilkins, 1988; Vol 1:291–354.
12. Zaret BL, Wackers FJTh, Soufer R. Nuclear cardiology. In: Braunwald E, ed. Heart Disease, A Textbook of Cardiovascular Medicine. 4th ed. Philadelphia: WB Saunders, 1991: 276–311.
13. Wackers FJTh, Fetterman RC, Mattera JA, Clements JP. Quantitative planar thallium-201 stress scintigraphy: a critical evaluation of the method. Semin Nucl Med 1985;15:46–66.
14. Wackers FJTh, Gibbons RJ, Verani MS, et al. Serial quantitative planar technetium-99m-isonitrile imaging in acute myocardial infarction: efficacy for noninvasive assessment of thrombolytic therapy. J Am Coll Cardiol 1989;14:861–73.
15. Sigal SL, Soufer R, Fetterman RC, Mattera JA, Wackers FJTh. Reproducibility of planar thallium-201 scintigraphy: quantitative criteria for reversibility of myocardial perfusion defects. J Nucl Med 1991;32:759–65.
16. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measmt 1960;20:37–46.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
18. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153–7.
19. Okada RD, Boucher CA, Kirshenbaum HK, et al. Improved diagnostic accuracy of thallium-201 stress test using multiple observers and criteria derived from interobserver analysis of variance. Am J Cardiol 1980;46: 619–24.
20. Atwood JE, Jensen D, Froelicher V, et al. Agreement in human interpretation of analog thallium myocardial perfusion images. Circulation 1981; 64:601–9.
21. Bailey IK, Griffith LSC, Rouleau J, Strauss HW, Pitt B. Thallium-201 myocardial perfusion imaging at rest and during exercise. Circulation 1977;55:79–87.
22. Verani MS, Marcus ML, Razzak MA, Ehrhardt JC. Sensitivity and specificity of thallium-201 perfusion scintigrams under exercise in the diagnosis of coronary artery disease. J Nucl Med 1978;19:773–82.
23. Iskandrian A, Wasserman LA, Anderson GS, Hakki H, Segal BL, Kane S. Merits of stress thallium-201 myocardial perfusion imaging in patients with inconclusive exercise electrocardiograms: correlation with coronary arteriograms. Am J Cardiol 1980;46:553–8.
24. Ritchie JL, Trobaugh GB, Hamilton GW, et al. Myocardial imaging with thallium-201 at rest and during exercise. Comparison with coronary arteriography and resting and stress electrocardiography. Circulation 1977;56:66–71.
25. Lenaers A, Block P, van Thiel E, et al. Segmental analysis of Tl-201 stress myocardial scintigraphy. J Nucl Med 1977;18:509–16.
26. Watson DD, Smith WH, Beller GA, Vinson EL, Taillefer R. Blinded evaluation of planar technetium-99m-sestamibi myocardial perfusion studies. J Nucl Med 1992;33:668–75.
27. Trobaugh GB, Wackers FJTh, Busemann E, DeRouen TA, Ritchie JL, Hamilton GW. Thallium-201 myocardial imaging: an interinstitutional study of observer variability. J Nucl Med 1978;19:359–63.
28. Wackers FJTh, Berman DS, Maddahi J, et al. Technetium-99m hexakis 2-methoxyisobutyl isonitrile: human biodistribution, dosimetry, safety, and preliminary comparison to thallium-201 for myocardial perfusion imaging. J Nucl Med 1989;30:301–11.
29. Kiat H, Maddahi J, Roy LT. Comparison of technetium 99m-methoxy isobutyl isonitrile and thallium-201 for evaluation of coronary artery disease by planar and tomographic methods. Am Heart J 1989;117:1–11.
30. Hendel RC, McSherry B, Karimeddini M, Leppo JA. Diagnostic value of a new myocardial perfusion agent, teboroxime (SQ 30,217), utilizing a rapid planar imaging protocol: preliminary results. J Am Coll Cardiol 1990;16:855–61.
31. Maddahi J, Van Train K, Prigent F, et al. Quantitative single photon emission computed thallium-201 tomography for detection and localization of coronary artery disease: optimization and prospective validation of a new technique. J Am Coll Cardiol 1989;14:1689–99.
32. DePuey EG, Garcia EV. Optimal specificity of thallium-201 SPECT through recognition of imaging artifacts. J Nucl Med 1989;30:441–9.
33. Garcia EV, Cooke CD, Van Train KF, et al. Technical aspects of myocardial SPECT imaging with technetium-99m-sestamibi. Am J Cardiol 1990;66:23E–31E.