A Multicenter Trial on Interobserver and Intraobserver Reproducibility of Segmental Scoring of Thallium-201 Planar Myocardial Imaging Before and After Reinjection

Marco Brambilla, Eugenio Inglese, Giorgio Cannizzaro, Maurizio Dondi, Roberto Sara, Francesco Arrigo and Gian Luigi Tarolo on behalf of the Italian Group of Nuclear Cardiology

Department of Nuclear Medicine, Ospedale Maggiore, Novara; Department of Nuclear Medicine, Montescano Hospital; Department of Nuclear Medicine, Faenza Hospital, Faenza; Department of Nuclear Medicine, Messina University, Messina; Department of Nuclear Medicine, Milan University; and Department of Nuclear Medicine, Niguarda Hospital, Milan, Italy

Inter- and intraobserver reproducibility (R) of segmental ^{201}Tl scores after stress (ST), redistribution (RD) and reinjection (RI) planar imaging were evaluated. **Methods:** Images were examined from 396 patients with suspected coronary artery disease, demonstrated by means of post-ST imaging of at least one perfusion defect. To eliminate external sources of variability, the same gamma camera, acquisition protocol and computer software were used in this multicenter study. Thallium-201 images of the anterior, left anterior oblique and left lateral projections were obtained immediately, 4 hr after exercise and 30 min after the injection of additional ^{201}Tl either on the same day or on a different day. The left ventricle was divided into 15 segments and evaluated by three independent observers, blinded to clinical data, according to a five-point scale. **Results:** The R score for ST, RD and RI images, expressed as an intraclass correlation coefficient, was 0.76, 0.74 and 0.72, respectively. After averaging multiple observer scores, R increased to 0.91, 0.90 and 0.89, respectively. Individual observer measurement of the R score was 0.48, 0.51 and 0.32 for ST-RD, ST-RI and RD-RI image pairs, respectively, and multiple observer scores showed R increases to 0.74, 0.76 and 0.58. **Conclusion:** This qualitative scale reliably assesses the severity of ^{201}Tl perfusion defects, particularly when multiple-observer scores are averaged. Individual observer change scores should be taken with great caution, especially in studies involving the visual evaluation of RD-RI image changes.

Key Words: multicenter trial; thallium-201; planar myocardial imaging

J Nucl Med 1994; 35:601–608

Despite the introduction of quantitative techniques in the analysis of both planar (*1,2*) and tomographic (*3*) myocardial perfusion scintigraphies, the visual interpretation of scintigraphic examinations is still widely in use. To provide a scientific basis for such subjective judgement, a number of different nominal and ordinal scales have been developed (*4–6*). The adoption of a measurement scale based on subjective judgement should lead the researcher to gather evidence that the scale is designed to measure in a reproducible fashion, i.e., to demonstrate that measurements of individuals on different occasions (intraobserver R), or by different observers (interobserver R), produce the same or similar results. The expression of R is an alternative to reporting the measurement error when the observation is categorical.

The aim of this study is to assess R scoring on the five-point scale adopted in this Italian multicenter study on thallium reinjection (SIRT)* in order to rate the severity of myocardial perfusion defects. Both intra- and interobserver R were separately assessed for ST, RD and RI images in order to test differences between the R of the scoring when using images which are potentially different in terms of their signal-to-noise ratio, count statistics and other parameters that may affect the quality of the visual display. Furthermore, the R of ST-RD, ST-RI, RD-RI change scores (i.e., the scores obtained by simply subtracting post-test from pre-test scores) were also assessed in order to provide objective guidelines for reversibility after RD and after RI.

METHODS

Twelve Italian medical centers experienced in nuclear cardiology and equipped with Elscint gamma cameras (Apex series) participated in the study.

Received Jun. 16, 1993; revision accepted Oct. 26, 1993.

For correspondence or reprints contact: Marco Brambilla, Dept. of Nuclear Medicine, Ospedale Maggiore della Carità, corso Mazzini 18, 28100 Novara, Italy.

*A list of the SIRT investigators and associates appears in the Appendix.

Patient Selection

We enrolled 396 consecutive patients (351 males and 45 females, mean age 58.1 ± 9.3 yr) with at least one segment showing a ^{201}Tl defect on stress planar myocardial perfusion scintigraphy for this study. All patients had ischemic heart disease of various degrees of severity as revealed by their medical history, physical examination and/or instrumental signs of transient or stable damage of the ventricular function (EKG and/or echocardiography). Eighty-two patients were asymptomatic and 242 had previous attacks of angina; 300 (76%) had a history of previous myocardial infarction.

Exercise Thallium Imaging

All patients underwent a symptom-limited treadmill stress test in a fasting state. At peak exercise, 2 mCi of ^{201}Tl were injected intravenously and the patient continued to exercise for at least one additional minute. Immediately after exercise, sequential 8-min (or 750 kcts in the total field of view, whichever was reached first) planar images were recorded in the left anterior oblique "best septal" anterior and left lateral views (Fig. 1). The images were acquired using a general-purpose, parallel-hole collimator. The images were acquired using a 25% window on the 80-keV pack and a 20% window on the 167-keV peak in a 128 × 128 byte matrix, with a standardized zoom factor. A second set of RD images was acquired in the same views for the same duration of the ST images approximately 3–4 hr after exercise. The patients were asked to continue fasting until the delayed images were recorded.

Reinjection Thallium Imaging

All patients were also evaluated by ^{201}Tl reinjection under baseline conditions. Seven medical centers followed a same-day approach in which 226 patients (Group A) received a second injection of 1 mCi of ^{201}Tl immediately after the RD study. Five medical centers followed a different-day approach in which 170 patients (Group B) received a RI of 2 mCi of ^{201}Tl at rest 48–72 hr after the ST-RD study. In both groups, acquisitions started at 30 min after RI in the same views and followed the same criteria as those of the ST-RD study.

Image Analysis

Serial thallium images were visually analyzed by the observer of each center on an Elscint Apex black and white video terminal. Operators were not allowed to modify display brightness or contrast. The video display was automatically programmed to get the maximal R in visual analysis among different medical centers and patients. Dedicated programs performed normalization to the maximal myocardial activity in the ST images and accurate pairing of each triplet of planar views for a single-sight simultaneous display, before and after background subtraction following the Goris method modified by Watson (*1*). The left ventricle in each view was divided into five segments (Fig. 2), and each segment was visually graded by the peripheral readers according to a five-point scale (0 = normal, 1 = equivocal, 2 = mild, 3 = severe and 4 = absent uptake).

The studies were subsequently recorded onto floppy disks and mailed to the core center. Three experienced observers from three different institutions, whose good interobserver R had been previously assessed using a randomized subgroup of 40 cases, reread the studies separately and independently on the same type of black and white video terminal as those used by the peripheral readers, without any knowledge of the patients' clinical data.

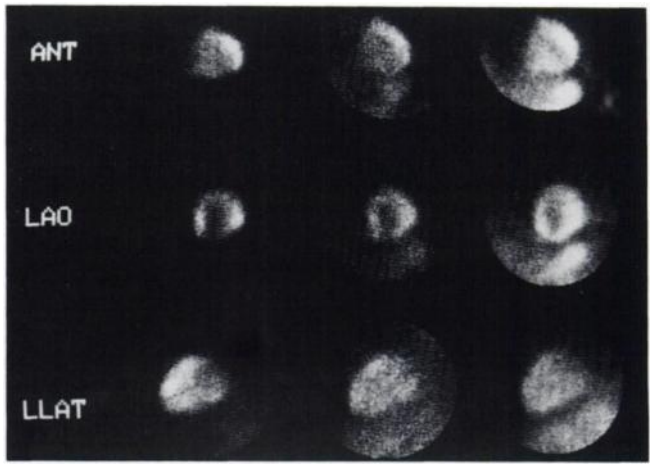

FIGURE 1. Representative case of redistribution and reinjection ^{201}Tl images. The left column images were acquired immediately after exercise (stress study), the middle column images were acquired 3–4 hr later (redistribution study) and the right column images were acquired 30 min after reinjection (reinjection study). This patient shows a perfusion defect on the inferior wall (anterior and LAO postexercise images) that remained fixed on the redistribution images and improved on the reinjection images.

Statistical Analysis

R was determined by using repeated measurement ANOVA methods (*7,8*) as the variance between subjects (σ^2_{sub}) divided by the sum of error variance (σ^2_{err}), observer variance (σ^2_{obs}) and the variance between subjects:

$$R = \frac{\sigma^2_{sub}}{\sigma^2_{sub} + \sigma^2_{obs} + \sigma^2_{err}} \cdot \qquad \text{Eq. 1}$$

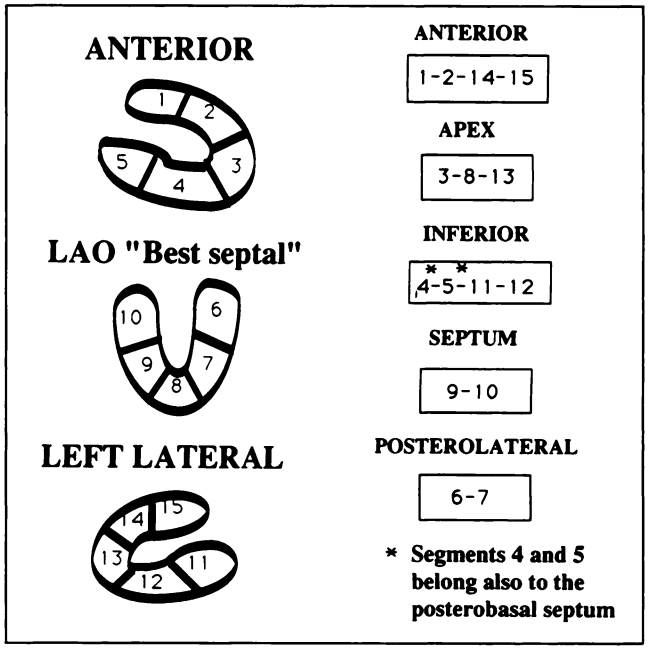

FIGURE 2. Diagram of the standard segmentation scheme used for scoring all ^{201}Tl images and assignment of individual segments to myocardial areas.

Thus, R is expressed as a number between 0 and 1, with 0 indicating no R and 1 indicating perfect R. A value of 0.75 is a fairly minimum requirement for a useful instrument. For instance, an R of 0.8 and 0.95 will result in a 20% and 2.2% chance of reversal of the order of two scores separated by an interquartile distance in repeated testing (*9*). A way of increasing R is to take multiple observers and average their scores since this allows the error variance plus the observer variance to be divided by the number of observers.

To facilitate the interpretation of R in terms of measurement errors on individual scores, it is useful to report the precision of subjective scoring. This requires the use of standard error of measurements (s.e.m.), defined in terms of the standard deviation (s.d.) and R as:

$$\text{s.e.m.} = \text{s.d.} \times (1 - R)^{1/2}. \qquad \text{Eq. 2}$$

This means that if the segment's true score is n, we can expect that its observed score will fall between n − 2 s.e.m. and n + 2 s.e.m. 95% of the time.

The relationship between sample size (N) and the confidence interval (CI) for the R coefficient is:

$$N = (Z_{\alpha/2}/CI)^2 + 3, \qquad \text{Eq. 3}$$

where $Z_{\alpha/2}$ = 1.96 for a 95% CI and 2.54 for a 99% CI.

The hypothesis of equality of intraclass correlation coefficients in independent samples was tested (*10*) ($p < 0.05$ was considered significant).

RESULTS

Count statistics in a rectangular ROI encompassing the left ventricle in the anterior view was 237 (s.d. 71), 169 (s.d. 63), 271 (s.d. 88) kcounts for ST, RD and RI images, respectively.

Table 1 shows the scoring R for ST, RD and RI images of individual observer scores (R_{single}) and the averaged scores of multiple observers (R_{all}). R_{single} was 0.76, 0.74 and 0.72 for ST, RD and RI, respectively. Averaging multiple-observer scores increased R_{all} to 0.91, 0.90 and 0.89. R_{single} for ST images was significantly higher than that for RD ($z = 2.49$, $p \leq 0.05$) and RI images ($z = 4.53$, $p \leq 0.01$). R_{single} for RD was significantly higher than that for RI images ($z = 2.05$, $p \leq 0.05$). A similar trend was found when comparing R_{all} between ST and RD ($z = 2.87$, $p \leq 0.01$), ST and RI ($z = 5.2$, $p \leq 0.01$), and RD and RI images ($z = 2.32$, $p \leq 0.05$). Significant differences ($z = 2.29$, $p \leq 0.05$; $z = 2.63$, $p \leq 0.01$) were also found when comparing R_{single} and R_{all} for RI images between Group A (same-day RI) and Group B (different-day RI) studies.

TABLE 1
R Scores for Stress, Redistribution and Reinjection of Individual Observer and Averaged Multiple Observer Scores

	Stress (n = 5874)	Redistribution (n = 5833)	Reinjection (n = 5854)
R_{single}	0.76	0.74	0.72
R_{all}	0.91	0.90	0.89

R_{single} = reproducibility of individual observer scores; R_{all} = reproducibility of averaged multiple observer scores; n = number of segments evaluated.
99% confidence interval = 0.03.

TABLE 2
Scoring Reproducibility for Stress, Redistribution and Reinjection of Individual Observer and Averaged Multiple Observer Scores for Individual Segments

Segment no.	Stress		Redistribution		Reinjection	
	R_{all}	R_{single}	R_{all}	R_{single}	R_{all}	R_{single}
1	0.82	0.60	0.81	0.58	0.79	0.56
2	0.90	0.75	0.90	0.75	0.90	0.74
3	0.85	0.65	0.84	0.64	0.83	0.62
4	0.88	0.70	0.89	0.72	0.87	0.70
5	0.89	0.72	0.89	0.68	0.86	0.67
6	0.83	0.61	0.81	0.59	0.82	0.60
7	0.91	0.77	0.92	0.80	0.90	0.76
8	0.88	0.71	0.89	0.73	0.89	0.73
9	0.92	0.79	0.90	0.75	0.89	0.73
10	0.92	0.80	0.90	0.75	0.89	0.73
11	0.91	0.77	0.91	0.76	0.90	0.74
12	0.90	0.76	0.89	0.74	0.87	0.70
13	0.86	0.68	0.87	0.68	0.85	0.66
14	0.92	0.78	0.90	0.75	0.89	0.74
15	0.80	0.57	0.79	0.55	0.78	0.54

R_{single} = reproducibility of individual observer scores; R_{all} = reproducibility of averaged multiple observer scores; the number of segments evaluated ranged from 381 to 393.
99% confidence interval = 0.13.

Table 2 shows R_{single} and R_{all} for ST, RD and RI images with respect to individual segments (Fig. 1). When considering ST images, R_{single} was lower ($R < 0.65$) for segments 15, 1 and 6; intermediate ($0.65 \leq R < 0.75$) for segments 3, 13, 4, 8 and 5, and higher ($R \geq 0.75$) for segments 2, 12, 11, 7, 14, 9 and 10.

The individual segments were subsequently assigned to six myocardial regions as follow: segments 1, 2, 14 and 15 to the anterior area; segments 3, 8 and 13 to the apex; segments 4, 5, 11 and 12 to the inferior area; segments 9 and 10 to the septum; and segments 6 and 7 to the posterolateral area. Since segments 4 and 5 not only represent the inferior wall, but also the posterobasal septum, we evaluated them as an additional region (Fig. 1). Table 3 shows R_{single} and R_{all} for the ST, RD and RI images of these myocardial regions. When considering ST images, maximal interobserver variability was observed in the apex ($R_{single} = 0.68$), and maximal R in the septal area in the LAO view ($R_{single} = 0.79$). Similar trends were observed for the RD and RI images. Significant differences were found when R_{single} of the septum was compared with R_{single} of the inferior ($z = 2.78$; $p < 0.01$), anterior ($z = 4.08$; $p < 0.01$), posterolateral ($z = 3.73$; $p < 0.01$), posterobasal septum ($z = 3.73$; $p < 0.01$) and apical areas ($z = 5.28$; $p < 0.01$). Significant differences were also found between the R_{single} of the inferior and apical areas ($z = 3.16$; $p < 0.01$).

TABLE 3
Scoring R for Stress, Redistribution and Reinjection of Both Individual Observer and Averaged Multiple Observer Scores for Myocardial Areas

Myocardial area	Stress			Redistribution			Reinjection		
	R_{all}	R_{single}	N	R_{all}	R_{single}	N	R_{all}	R_{single}	N
Anterior (*1,2,14,15*)	0.88	0.72	1564	0.87	0.70	1551	0.87	0.69	1557
Apex (*3,8,13*)	0.87	0.68	1175	0.85	0.65	1188	0.86	0.67	1171
Inferior (*4,5,11,12*)	0.90	0.74	1564	0.89	0.74	1545	0.88	0.71	1549
Septum (*9,10*)	0.92	0.79	785	0.90	0.75	785	0.89	0.73	788
Posterolateral (*6,7*)	0.88	0.71	785	0.88	0.71	785	0.87	0.69	788
Posterobasal Septum (*4–5*)	0.88	0.71	786	0.88	0.70	784	0.87	0.69	784

R_{single} = reproducibility of individual observer scores; R_{all} = reproducibility of averaged multiple observer scores; N = number of segments evaluated.

Table 4 reports the R_{single} and R_{all} for the change scores of the ST-RD, ST-RI and RD-RI sequences. R_{single} was suboptimal for all of the image pairs, extending from 0.31 for RD-RI to 0.51 for ST-RI. Averaging multiple observer scores considerably increased R (0.74 and 0.76 for ST-RD and ST-RI), but RD-RI R_{all} still remained weak (0.58).

Intraobserver R for the three core observers, evaluated over a randomized sample of 40 studies, is reported in Table 5.

The median scores of the three core observers were compared with the peripheral reader's scores to assess interobserver R between a representative sample of blinded and unblinded readers, and to see which scoring scale level contributed the most to observer variability. These results are summarized in Table 6 and Table 7.

The effect of reducing the number of the steps on the scoring scale is reported in Table 8. If we group together scores 0 and 1, and 2, 3 and 4, we have a two-level scale that simulates the dichotomous judgement normal-abnormal, while grouping together scores 0 and 1, 2, and 3 and 4 results in a three-level scale. For simplicity, we only reported R_{single} and R_{all} for the ST images.

TABLE 4
Change Score R for ST-RD, ST-RI, RD-RI of Individual Observer and Averaged Multiple Observer Scores

	ST-RD (n = 5821)	ST-RI (n = 5841)	RD-RI (n = 5802)
R_{single}	0.48	0.51	0.31
R_{all}	0.74	0.76	0.58

R_{single} = reproducibility of individual observer scores; R_{all} = reproducibility of averaged multiple observer scores; n = number of segments evaluated.
99% confidence interval = 0.03.

DISCUSSION

In clinical medicine it is common practice to examine observer agreement over the presence or absence of a particular sign or symptom, or a particular diagnostic pattern in a medical image.

The expression of R as a simple percentage agreement is not only theoretically incorrect (because it does not take chance agreement into account), but in many practical circumstances it can be very far from the reality.

Okada et al. (*11*) reported interobserver variance ($\sigma^2_{obs} + \sigma^2_{err}$) or its associated standard deviation as an expression of R. This is incorrect since it is equivalent to neglecting a major determinant of the R coefficient (σ^2_{sub}). Moreover, this makes it impossible to compare R coefficients, as referred by different study groups, since the equality of error variance does not, by itself, imply the equality of R, as long as the variance of true differences between subjects may differ.

The Pearson correlation coefficient is another index of R used by some authors (*12,13*). This is also a theoretically incorrect measure of R, since it does not take into account systematic differences between observers.

TABLE 5
Intraobserver R of the Three Core Observers

Observer no.	ST	RD	RI	ST-RD	ST-RI	RD-RI
1	0.77	0.79	0.80	0.47	0.56	0.37
	(578)	(580)	(573)	(574)	(567)	(569)
2	0.80	0.77	0.75	0.60	0.59	0.35
	(577)	(574)	(564)	(572)	(562)	(559)
3	0.81	0.80	0.78	0.52	0.60	0.38
	(578)	(580)	(578)	(574)	(572)	(574)

ST = stress, RD = redistribution, RI = reinjection.
Numbers in parentheses are the segments evaluated.

TABLE 6
Effects of Clinical Data Knowledge on Interobserver R Between Blinded Core Readers and Unblinded Peripheral Readers

	ST	RD	RI	ST-RD	ST-RI	RD-RI
N	5964	5923	5943	5843	5930	5891
R_{single}	0.78	0.76	0.74	0.49	0.60	0.36
R_{all}	0.88	0.87	0.85	0.66	0.75	0.53

R_{single} = reproducibility of individual observer scores; R_{all} = reproducibility of averaged multiple observer scores; N = number of segments evaluated. ST = stress, RD = redistribution, RI = reinjection.
99% confidence interval = 0.03.

TABLE 7
Percentage Agreement Between Averaged Core Observers and Peripheral Readers Five-Step Scoring Scale After ST Imaging

		Core observers					
		0	1	2	3	4	Total
Peripheral observers	0	88.8 *29.5*	46.9	16.4	7.9	3.9	3539
	1	8.3	35.1 *20.7*	25.3	7.2	1.6	864
	2	2.4	13.9	37.1 *25.0*	22.5	11.8	726
	3	0.3	3.6	18.8	46.0 *35.2*	35.4	587
	4	0.2	0.5	2.3	16.3	47.2 *43.1*	248
	Total	3300	826	1003	582	254	5964

Numbers in italics are percentage agreement beyond chance.

Another index of R is the kappa statistic (*14–17*) introduced by Cohen (*18*) to describe observer agreement in the classification of dichotomous nominal scales. This index scales the percentage of perfect agreement according to the percentage due to chance agreement. An extension of this approach is the weighted-kappa statistic (*19*), which considers partial agreement by weighting the degree of the discrepancy. The limitation of this index is that it only allows the pair-wise comparison of two observers and cannot be applied in situations when each subject is rated by the same group of more than two raters (*20*).

On the other hand, all forms of intraclass correlation coefficients based on ANOVA methods take into account chance agreement, partial disagreement and systematic differences between observers. Moreover, they also allow situations in which three or more observers are involved to be evaluated. Thus, these indexes should be considered as the most appropriate for assessing R.

The level of R so obtained is not to be understood as an absolute measure of a property referred to as a particular instrument; rather, this instrument will have a certain degree of R when applied to certain populations under certain conditions.

Our results show interobserver R in the scoring of segmental myocardial ^{201}Tl activity on ST, RD and RI images using a five-point scale. As can be seen from Table 1, the R is good but not excellent when a single observer is involved in the scoring. The situation is considerably improved when three observer scores are averaged, which leads to an interobserver R only slightly lower than those reported by Sigal et al. (*13*) in a quantitative analysis of myocardial perfusion abnormalities: the precision of an averaged subjective grading of ST, RD and RI images on a 0–4 scoring scale is respectively ±0.73, ±0.73, ±0.72. Although statistically significant (due to the high statistical power of the test), the differences in R between ST, RD and RI images are hardly interpretable as clinically significant, notwithstanding the presence of significant differences in count statistics among the three sets of images. The same is true for differences in R between the two RI protocols.

When examining R scores for individual segments, we found a broad range of values (from 0.57 to 0.80 for R_{single} in ST images). The segments 1, 6 and 15, showing lower R, belong to regions where scoring may be difficult due to the proximity of diaphragm and valve planes. Intermediate R values were found for apical segments 3, 8 and 13 (where the well-known phenomenon of apical thinning may affect the scoring), and for segments 4 and 5 reflecting both the inferior wall and the posterobasal septum in the anterior projection.

If the segments are grouped together into myocardial regions, a slightly different perspective emerges: the apical region shows minimal R since all of its segments are in the intermediate range of R, while the simultaneous presence of low and high R segments leads to an intermediate R value for the anterior, posterolateral and inferior regions. Maximal R is found in the septal area in the LAO view. The superimposition of different anatomical structures typical of planar imaging leads to a difficult interpretation of the R values when they refer to the inferior region in the anterior projection. If we consider segments 4 and 5 as representative of the posterobasal septum, we obtain an intermediate R value, significantly lower than the R of the septum in the LAO view. This might reflect the different anatomical location of the septum in the anterior and LAO view with respect to the gamma camera. The posterobasal septum in the anterior projection is located deep in the thorax and well-known problems of attenuation and self-attenuation

TABLE 8
Effects of Reducing the Number of Steps in the Scoring Scale for Interobserver Reproducibility After Stress Imaging

	Five levels	Three levels	Two levels
	0-1-2-3-4	0,1-2-3,4	0,1-2-3-4
R_{single}	0.76	0.74	0.69
R_{all}	0.91	0.89	0.87

might contribute to lower R in this area. On the other hand, if we consider only segments 11 and 12 as representative of the inferior region, we obtain a high R (R_{all} = 0.91, R_{single} = 0.77) not significantly different from the R of the septum in the LAO view. Similar findings have been previously reported by Atwood et al. (*21*). The discrepancy between our results and those reported by Okada et al. (*11*) might be explained by considering that, in our experiment, the scoring was based on both raw and background subtracted images. An appropriate background subtraction should reduce the problem of superimposition of activities belonging to different anatomical structures. Moreover, this procedure makes the target-to-background ratio uniform for all of the myocardial regions.

Exercise thallium myocardial imaging is widely used to evaluate patients with suspected coronary artery disease; defects on the initial image suggest coronary artery disease, and fill-in of these defects on the delayed images suggests hypoperfused viable myocardium (*22–24*). However, in many regions of viable myocardium, the defects that are detected during thallium exercise testing persist, and appear to be irreversible on RD images taken at 3–4 hr later in baseline conditions (*25,26*). Recent studies have shown that the RI of thallium at rest after RD imaging may lead to increased thallium uptake in apparently irreversible thallium defects, which is compatible with viable myocardium (*4,5,27–29*). The evaluation of change in ^{201}Tl myocardial activity between ST, RD and RI images is thus of crucial importance in the assessment of both coronary artery disease and viable myocardium.

The calculation of a change score is based upon the difference between pre-test and post-test scores. As can be seen from Table 4, the change score R is systematically lower than pre- and post-test R. Two reasons can be suggested to explain this: (1) both pre- and post-test scores have a certain degree of error (σ^2_{err}) that propagates when combining the two measurements by means of a difference; or (2) the true variance between subjects (σ^2_{sub}) is lower examining change scores, since pre- and post-test scores show a certain degree of correlation.

This last explanation also accounts for the lower value of the RD-RI change score R (in comparison with ST-RD and ST-RI), since there is less variation in scores when passing from RD to RI, than when passing from ST to RD or to RI. The precision of the averaged subjective scoring of change was ±0.52, ±0.60, ±0.46 for ST-RD, ST-RI and RD-RI, respectively. Notwithstanding the lower R, the precision of the change score assessment is higher than that of pre- and post-test scores. This confirms the general argument that a low R does not necessarily imply a lack of precision. Although individual differences in change are necessary for a high R, the absence of such differences does not preclude meaningful assessment of individual change (*30*).

Intraobserver R for all of the three observers is systematically higher than the interobserver R_{single} for ST, RD and RI images, but the difference is slight (Table 5). Thus, the evaluation of perfusion defects severity might be performed in the same patient by different observers without losing much in terms of R.

An important goal of our study was to attempt to identify and measure the main sources of variability in ^{201}Tl image interpretation. The standard measures of intra- and interobserver R do not exhaust the possible sources of variance.

A potential reason for discordant readings of ^{201}Tl images is the availability of clinical data at the moment of the scoring. The figures of R between blinded and unblinded readers shown in Table 6 are very close to the ones obtained when assessing interobserver R_{single} among the three blinded observers (Tables 1 and 4). Thus, the knowledge of clinical data does not seem to play a major role in interobserver variability, at least when a strictly controlled methodology of image acquisition and image display is adopted.

It could be asked if examinations judged as optimal by the core observers have a better R than examinations of poorer quality. In the study protocol, the three core observers were asked to rate the overall quality of the scintigraphic study as: suboptimal, good or excellent. Eighteen studies were rated by at least two observers as suboptimal. In these studies, we found R_{single} = 0.73 and R_{all} = 0.89 for ST images. These values are lower than those of the whole sample but the difference is not striking. Thus, the subjective judgement on the quality of examinations does not seem to discriminate between high and low R studies.

Table 7 shows that, at a first glance, the higher percentage of observed agreement for ST images is for the zero level, but if we take chance agreement into account, the higher percentage of agreement beyond chance is reached for levels 3 and 4, while levels 0, 1 and 2 have a similar percentage of agreement considerably lower than levels 3 and 4.

The choice of the number of the steps on a scoring scale is not primarily an esthetic issue. There are a number of theoretical reasons (*7*) and experimental evidences (*31*) suggesting that, if the number of the steps on a scale is less than the rater's ability to discriminate, the result will be a loss of information. This is indeed the case in our study; as fewer categories are used, the R drops (Table 8).

CONCLUSION

The interobserver R of the visual five-point scale used to rate myocardial perfusion defects severity after ST, RD and RI imaging, is acceptable when a single observer is involved in the scoring, but is far from ideal. Averaging multiple observer scores leads to a high R similar to those reported using quantitative analysis of myocardial scintigraphies (*2*). Change scores are sufficiently reproducible only when multiple observers are involved in the scoring and their score is averaged. In this case, the precision of the assessment of change is within one point at the 95% confidence level, and thus they may be used to assess the reversibility of defects. Inter- and intraobserver reproducibility do not differ significantly. The availability of clinical

data at the moment of the scoring does not seem to play a major role in interobserver variability. Reducing the number of steps on a visual scoring scale lowers reproducibility; thus two-level scales (normal-abnormal) should be avoided.

APPENDIX

Italian Multicenter Study on Thallium Reinjection (SIRT) Investigators

Study Coordinator

Eugenio Inglese, MD, Medicina Nucleare, Ospedale Maggiore di Novara.

Study Chairmen

Gain Luigi Tarolo, MD, Istituto di Medicina Nucleare, Università di Milano.

Francesco Arrigo, MD, Istituto di Cardiologia, Università di Messina.

Clinical Centers

Divisione di Cardiologia, Divisione di Medicina Nucleare, Ospedale Malpighi, Bologna, Italy. C. Corbelli, MD, M. Dondi, MD, S. Fanti, MD, N. Monetti, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Ospedale Bufalini, Cesena, Italy. F. Busi, MD, D. Della Vittoria, MD, S. Gherardi, MD, G. Moscatelli, MD, PL. Pieri, MD, A. Tisselli, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Università Firenze, Firenze, Italy. G. Bisi, MD, G.M. Santoro, MD, R. Sciagrà, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Ospedale Civile, La Spezia, Italy. D. Bertoli, MD, M. Cappagli, MD, T. Duce, MD, S. Gramenzi, MD, R. Leoncini, MD, A. Montepagani, MD, P. Poggi, MD, R. Russo, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Policlinico Università, Messina, Italy. F. Arrigo, MD, S. Baldari, MD, S. Carerj, MD, A. Marvelli, MD, A. Migliorato, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Ospedale Niguarda, Milano, Italy. A. Ascione, MD, G. Boni, MD, G. Cannizzaro, MD, D. Massa, MD, G. Piccalò, MD, S. Pirelli, MD, R. Sara, MD, F. Spinelli, MD, F. Sarullo, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Ospedale Cervello, Palermo, Italy. U. Ficola, MD, S. La Monica, MD, R. Lo Mauro, MD, P. Marozzi, MD, P. Sabella, MD.

Divisione di Cardiologia, Ospedale S. Maria delle Croci, Ravenna, Italy. G. Bellanti, MD, G. Berti, MD, S. Coccolini, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Università Cattolica, Roma, Italy. ML. Calcagni, MD, A. Giordano, MD, E. Rossi, MD, M. Salvatori, MD, G. Schiavoni, MD, C. Trani, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Centro Campano Ricerche, Salerno, Italy. V. Arienzo, MD, M. Bifulco, MD, V. Capuano, MD, N. La Maida, MD, M. Punzi, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Policlinico Università, Siena, Italy. A. Vattimo, MD, L. Baldi, MD, P. Bertelli, MD, C. Cataldi, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Ospedale Civile, Treviso, Italy. G. Favretto, MD, V. Cuzzato, MD, F. Palermo, MD, P. Zoli, MD.

Divisione di Cardiologia, Divisione di Medicina Nucleare, Ospedale Civile, Venezia, Italy. R. Anastasio, MD, A. Benzoni, MD, A. Bonazza, MD, S. Gravili, MD.

Italian Group of Nuclear Cardiology

G. Mazzotta, MD, Cardiologia, Ospedale Galliera, Genova.

ACKNOWLEDGMENTS

The authors thank Mallinckrodt Medical for the reinjection ^{201}Tl doses; Byk Gulden Italia for the excellent organization management and central data archive; and New Elscint Technologies for their technical assistance and support.

REFERENCES

1. Watson DD, Campbell NP, Read EK, et al. Spatial and temporal quantitation of plane thallium myocardial images. *J Nucl Med* 1981;22:577–584.
2. Garcia EV, Maddhai J, Berman D, Waxman A. Space/time quantitation of thallium-201 myocardial scintigraphy. *J Nucl Med* 1981;22:309–317.
3. Garcia EV, Van Train K, Maddhai J, et al. Quantification of rotational thallium-201 myocardial tomography. *J Nucl Med* 1985;26:17–26.
4. Dilsizian VD, Rocco TP, Nanette MT, et al. Enhanced detection of ischemic but viable myocardium by the reinjection of thallium after stress-redistribution imaging. *N Engl J Med* 1990;323:141–146.
5. Bonow RO, Dilsizian VD, Cuocolo A, et al. Identification of viable myocardium in patients with chronic coronary artery disease and left ventricular dysfunction. *Circulation* 1991;83:26–37.
6. Kiat H, Berman DS, Maddhai J, et al. Late reversibility of tomographic myocardial thallium-201 defects: an accurate marker of myocardial viability. *J Am Coll Cardiol* 1988;12:1456–1463.
7. Streiner DL, Norman GR. *Health measurement scales.* New York: Oxford University Press; 1991:79–95.
8. Wiener BJ. *Statistical principles in experimental design.* New York: McGraw Hill; 1971:283–285.
9. Thorndike RL, Hagen E. *Measurement and evaluation in education and psychology.* New York: Wiley; 1969.
10. Kleinbaum DG, Kupper LL, Muller KE. *Applied regression analysis and other multivariable methods.* Boston: PWS-Kent; 1988:91–93.
11. 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;46:619–624.
12. Van der Poel HG, Boon ME, van der Meulen EA, et al. The reproducibility of cytomorphometrical grading of bladder tumors. *Virchow Arch Patol Anat Histopathol* 1990;416:521–525.
13. Sigal SL, Soufer R, Fetterman RC, et al. Reproducibility of quantitative planar thallium-201 scintigraphy: quantitative criteria for reversibility of myocardial perfusion defects. *J Nucl Med* 1991;32:759–765.
14. Trobaugh GB, Wackers FJ, Sokole EB, et al. Thallium-201 myocardial imaging: an interinstitutional study of observer variability. *J Nucl Med* 1978;19:359–363.
15. Jarlov AE, Gjorup T, Hegedus L, et al. Observer variation in the diagnosis of solitary cold thyroid lesions. *Clin Endocrinol* 1990;33:1–11.
16. Vuillez JP, Peltier P, Mayer JC, et al. Reproducibility of image interpretation in immunoscintigraphy performed with indium-111 and iodine-131-labeled OC125 F(ab')$_2$ antibody injected into the same patients. *J Nucl Med* 1991;32:221–226.
17. Tiel-van Buul M, van Beek EJR, van Dongen A, et al. The reliability of the 3-phase bone scan in suspected scaphoid fracture: an inter and intraobserver variability analysis. *Eur J Nucl Med* 1992;19:848–852.
18. Cohen J. A coefficient of agreement for nominal scales. *Educational Psychological Meas* 1960;20:37–46.
19. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psych Bull* 1968;70:213–220.
20. Fleiss JL. Measuring nominal scale agreement among many raters. *Psych Bull* 1971;76:378–382.
21. Atwood JE, Jensen D, Froelicher V, et al. Agreement in human interpretation of analog thallium myocardial perfusion images. *Circulation* 1981;64:601–609.
22. Pohost GM, Zir LM, Moore RH, et al. Differentiation of transient ischemic from infarcted myocardium by serial imaging after a single dose of thallium-201. *Circulation* 1977;55:294–302.
23. Rozanski A, Berman DS, Gray R, et al. Use of thallium-201 redistribution

scintigraphy in preoperative differentiation of reversible and nonreversible myocardial asynergy. *Circulation* 1981;64:936–944.
24. Gibson RS, Watson DD, Taylor GL, et al. Prospective assessment of regional myocardial perfusion before and after coronary revascularization surgery by quantitative thallium-201 scintigraphy. *J Am Coll Cardiol* 1983; 3:804–815.
25. Liu P, Kiess MC, Okada RD, et al. The persistent defect on exercise thallium imaging and its fate after myocardial revascularization: does it represent scar or ischemia? *Am Heart J* 1985;110:996–1001.
26. Cloninger KG, DePuey EG, Garcia EV, et al. Incomplete redistribution in delayed thallium-201 single photon emission computed tomographic (SPECT) images: an overestimation of myocardial scarring. *J Am Coll Cardiol* 1988;12:955–963.
27. Rocco TP, Dilsizian V, McKusick KA, et al. Comparison of thallium redistribution with rest "reinjection" imaging for the detection of viable myocardium. *Am J Cardiol* 1990;66:158–163.
28. Tamaki N, Othani H, Yamashita K, et al. Metabolic activity in the areas of new fill-in after thallium-201 reinjection: comparison with positron emission tomography using fluorine-18-deoxyglucose. *J Nucl Med* 1991;32:673–678.
29. Dilsizian V, Freedman NMT, Bacharach SL, et al. Regional thallium uptake in irreversible defects. *Circulation* 1992;85:627–634.
30. Rogosa D, Brandt D, Zimowski M. A growth curve approach to the measurement of change. *Psych Bull* 1982;92:726–748.
31. Nishisato N, Torii Y. Effects of categorizing continuous normal distributions on the product-moment correlation. *Jpn Psych Res* 1970;13:45–49.

Condensed from
15 Years Ago:

Editorial: Teamwork in Cardiovascular Nuclear Medicine

Julia W. Buchanan and Henry N. Wagner, Jr.
The Johns Hopkins Medical Institutions, Baltimore, Maryland

Both cardiologists and nuclear medicine physicians are involved in the rapidly developing field of cardiovascular nuclear medicine. In July 1978 we sent a questionnaire to the heads of nuclear medicine residency training programs and directors of adult cardiology training programs to assess the type and degree of collaboration in performing these studies.

Sixty-four percent of the 351 questionnaires sent were returned within 2 mo. Only 10% of the returned questionnaires stated that these studies were not performed, and these institutions were excluded from further evaluation. In only six institutions are the studies performed exclusively in the department of cardiology. Most are done either exclusively in nuclear medicine or in both cardiology and nuclear medicine departments.

The most encouraging result of our survey is that cardiologists and nuclear medicine physicians work together as a team. Over 90% of both groups said there was collaboration between cardiology and nuclear medicine and that they believed the procedures should be performed as a joint effort.

Exercise testing was the area where collaboration was most often cited. The experience and expertise in selecting appropriate patients, performing the proper exercise protocol, and monitoring the patient's response were described as essential roles for the cardiologist. The expertise of the nuclear medicine physician was of greatest value in the technical aspects of the study. In some institutions a separate division of cardiovascular nuclear medicine has been created with the joint appointment of a cardiologist and a nuclear medicine physician, or of one physician trained in both disciplines. Some questionnaires stated that collaboration was primarily in joint research projects; others stated that there was joint interpretation of the studies. Some institutions have joint training programs for cardiology and nuclear medicine residents.

From the results of the questionnaire, we have concluded that cardiologists and nuclear medicine physicians should and do work together. Almost all have found that collaboration is the key to success.

J Nucl Med 1979; 20:377-378