Reliability of Visual Assessment of Global and Segmental Left Ventricular Function: A Multicenter Study by the Israeli Echocardiography Research Group

David S. Blondheim, MD, Ronen Beeri, MD, Micha S. Feinberg, MD, Mordehay Vaturi, MD, Sarah Shimoni, MD, Wolfgang Fehske, MD, Alik Sagie, MD, David Rosenmann, MD, Peter Lysyansky, PhD, Lisa Deutsch, PhD, Marina Leitman, MD, Rafael Kuperstein, MD, Ilan Hay, MD, Dan Gilon, MD, Zvi Friedman, PhD, Yoram Agmon, MD, Yossi Tsadok, BSc, and Noah Liel-Cohen, MD, *Hadera, Jerusalem, Tel Aviv, Petah Tikva, Rehovot, Haifa, Zerifin, and Beer Sheva, Israel; Cologne, Germany*

Background: The purpose of this multicenter study was to determine the reliability of visual assessments of segmental wall motion (WM) abnormalities and global left ventricular function among highly experienced echocardiographers using contemporary echocardiographic technology in patients with a variety of cardiac conditions.

Methods: The reliability of visual determinations of left ventricular WM and global function was calculated from assessments made by 12 experienced echocardiographers on 105 echocardiograms recorded using contemporary echocardiographic equipment. Ten studies were reread independently to determine intraobserver reliability.

Results: Interobserver reliability for visual differentiation between normal, hypokinetic, and akinetic segments had an intraclass correlation coefficient of 0.70. The intraclass correlation coefficient for dichotomizing segments into normal versus other abnormal was 0.63, for hypokinetic versus other scores was 0.26, and for akinetic versus other scores was 0.58. Similar results were found for intraobserver reliability. Interobserver reliability for WM score index was 0.84 and for left ventricular ejection fraction was 0.78. Similar values were obtained for the intraobserver reliability of WM score index and ejection fraction. Compared to angiographic data, the accuracy of segmental WM assessments was 85%, and correct determination of the culprit artery was achieved in 59% of patients with myocardial infarctions.

Conclusion: Among experienced readers using contemporary echocardiographic equipment, interobserver and intraobserver reliability was reasonable for the visual quantification of normal and akinetic segments but poor for hypokinetic segments. Reliability was good for the visual assessment of global left ventricular function by WM score index and ejection fraction. (J Am Soc Echocardiogr 2010;23:258-64.)

Keywords: Echocardiography, Variability, Reliability, Wall motion, Ejection fraction

The accurate evaluation of segmental left ventricular (LV) wall motion (WM) abnormalities is of paramount importance for the interpretation of echocardiograms. The detection of abnormally contracting segments facilitates the diagnosis of myocardial damage, the distribution of the abnormal segments suggests the coronary artery involved,[1] the number of abnormal segments suggests the extent of myocardial damage, and the severity of the contraction abnormalities pertains to viability. The evaluation of WM abnormalities includes primarily the visual assessment of displacement and thickening of LV myocardial segments and the classification of each segment according to the severity of the detected abnormalities.[2] Visual estimation is also used for evaluating LV ejection fraction (LVEF) as a measure of systolic function and has important prognostic and therapeutic implications.[3] Recent advances in echocardiographic technology (eg, the incorporation of second harmonics, new transducers) have enhanced endocardial visualization and improved the detection of WM abnormalities.[4,5]

The purpose of this multicenter study was to determine the reliability and variability of the visual assessment of segmental WM abnormalities and global LV function among highly experienced

From Hillel-Yaffe Medical Center, Hadera, Israel (D.S.B.); Hadassah University Hospital, Jerusalem, Israel (R.B., D.G.); Sheba Medical Center, Tel Aviv, Israel (M.S.F., R.K., I.H.); Rabin Medical Center, Petah Tikva, Israel (M.V., A.S.); Kaplan Medical Center, Rehovot, Israel (S.S.); St Vinzenz Hospital, Cologne, Germany (W.F.); Shaare Zedek Medical Center, Jerusalem, Israel (D.R.); GE Healthcare, Haifa, Israel (P.L., L.D., Z.F.), Assaf Harofeh Medical Center, Zerifin, Israel (M.L.); Rambam Medical Center, Haifa, Israel (Y.A.); Ben-Gurion University, Beer Sheva, Israel (Y.T.); and Soroka University Medical Center, Beer Sheva, Israel (N.L.-C.).

Conflicts of interest: Drs Lysyansky and Friedman work for GE Healthcare and are involved in developing echocardiography software for the automatic detection and quantification of left ventricular wall motion abnormalities and function.

Reprint requests: David S. Blondheim, MD, Hillel Yaffe Medical Center, Hadera, Israel (E-mail: *dsblond@gmail.com*).

0894-7317/$36.00

doi:10.1016/j.echo.2009.12.020

echocardiographers using contemporary echocardiographic technology in patients with a variety of cardiac conditions.

METHODS

Patients

Echocardiograms of 105 patients (mean age, 59 ± 14 years; 68% men) were reviewed for this study. Ninety patients hospitalized for chest pain (mean age, 57 ± 14 years; 68% men) were catheterized and underwent standard echocardiographic studies within 48 hours of admission. Of these, 62 patients (mean age, 59.2 ± 13.1; 75% men) had electrocardiographic and enzymatic criteria indicating an acute myocardial infarction (MI; 58 with single-vessel disease), and 28 patients (mean age, 53.4 ± 15.7; 54% men) had ischemic disease excluded by history, nonspecific electrocardiographic changes, a lack of cardiac enzyme elevation, and negative findings on coronary angiography (performed in all but 3 patients). Echocardiograms of an additional 15 patients with known nonischemic dilated cardiomyopathy (mean age, 70.5 ± 10.9 years; 60% men) were also evaluated.

Analysis by Experienced Physicians

Twelve physicians with extensive experience in reading echocardiograms, who interpret echocardiograms as their main daily activity, were included in this study. Eight head echocardiography laboratories or noninvasive units, and all are from 9 major cardiology centers. Each reader received compact discs with studies of the 105 patients for blinded assessment. For each patient, 2-dimensional clips of a single cardiac cycle (to verify that all readers were relating to the same cycle) recorded from the 4-chamber, 2-chamber, and apical long-axis views were provided. All echocardiographic recordings were made using a Vivid 7 digital ultrasound scanner (GE Vingmed Ultrasound AS, Horten, Norway). The readers were instructed to assess WM abnormalities and LV function as they would in their routine clinical work, so that this study would reflect real-world practice. Segments considered inadequate for analysis ("unreadable") were not scored. The analysis was done on workstations installed for this purpose with EchoPAC software (GE Vingmed Ultrasound AS). The left ventricle was automatically divided into 18 segments (6 basal, 6 middle, and 6 apical) to facilitate comparison among readers, and the readers scored each of the segments (1 = normal, 2 = hypokinetic, 3 = akinetic, 4 = dyskinetic, 5 = aneurysmatic). There were only a few dyskinetic and aneurysmatic segments, so they were grouped together with the akinetic segments in the analysis. All scores were automatically stored in an Excel file (Microsoft Corporation, Redmond, WA) that was later sent to a central laboratory for statistical analysis.

Estimation of Systolic LV Function

For each patient, a WM score index (WMSI) was calculated by each reader as the average score of all readable segments.[6] The LVEF of each patient was visually estimated by each reader solely on the basis of the apical views. An average LVEF and standard deviation of the estimates made by the 12 readers was assigned to each patient. The WMSI and LVEF were repeated in 10 patients by 11 of the readers.

Relation of WM Abnormalities to the Culprit Artery

We assessed the ability of experienced readers to detect segmental WM abnormalities correctly by comparing their findings with angiographic data from the catheterized patients. Correct identification of WM abnormalities was considered when such segments were located in territories supplied by culprit arteries and incorrect when WM abnormalities were detected in territories supplied by normal arteries. Sensitivity, specificity, positive and negative predictive values, and accuracy were calculated for each reader (the means, standard deviations, and ranges of the readers are reported) and for each segment (segments were classified according to the scores assigned to them by the majority of readers).

The ability of each reader to predict a patient's angiographically determined culprit artery was assessed. The artery supplying the territory in which the majority of segments with WM abnormalities were detected was considered the culprit artery. Obviously, by this method, the culprit artery could not be determined in patients who had an identical number of abnormal segments in >1 coronary territory or who had no WM abnormalities. The averages, standard deviations, and ranges of successful determinations of the culprit artery by the individual readers are presented, as well as the rate of correct identification using the segmental majority score.

For the purpose of these analyses, we adopted the American Society of Echocardiography's classification of territories supplied by each coronary artery[2] with necessary modifications.

Statistical Analysis

Data were analyzed using MATLAB (The MathWorks Inc, Natick, MA) and SAS version 9.1 (SAS Institute, Inc, Cary, NC). Categorical data are presented as counts and percentages and continuous data as mean ± SD. The intraclass correlation coefficient (ICC) was used as an index of interobserver and intraobserver reliability.[7] Interobserver reliability was assessed by fitting several random-effects models with the SAS PROC MIXED procedure, whereby the score was modeled with physician and segment within subject entered as random effects. The score was modeled as an ordinal variable by classifying WM abnormalities into a trichotomous variable (normal = 1, hypokinetic = 2, and akinetic or worse = 3) and then as 3 separate binary variables (normal vs other, hypokinetic vs other, and akinetic or worse vs other). Segments that were unreadable were considered missing values. The ICCs were then calculated from each of the models' variance components as the ratio of the between-reader error variance and the total variance. For estimation of intraobserver reliability, similar models were used, but data entered into the models represented a sample of 10 randomly selected subjects' segments scored by 11 of the physicians a second time, enabling measurement of the ICC for repeated measurements (ie, read-reread reliability). The average intraobserver reliability was calculated from each reader's individual reliability.

ICCs were interpreted in a similar manner as correlation coefficients: ICC > 0.80, excellent; 0.60 ≤ ICC ≤ 0.80, good; 0.40 ≤ ICC ≤ 0.60, moderate; and ICC < 0.40, poor.

The interobserver and intraobserver ICCs were calculated for WMSI and LVEF measurements, using similar models as described above. Because these variables are continuous, we also estimated the interobserver and intraobserver variability (represented by the standard deviation) from the models using the variance components and the mean WMSI score of the study sample adjusted for physicians using the estimated intercept and its 95% confidence interval. Interobserver variability was calculated as the square root of the interobserver variance component, whereas intraobserver variability was calculated as the square root of the error variance component per reader (from 11 separate models) and then averaged.

The sensitivity, specificity, positive and negative predictive values, and overall accuracy of the visual determination of normal, hypokinetic, and akinetic segments by the "average" reader in the study were estimated using a bootstrapping method.[8] At each step, 5 physicians' segmental scores were randomly selected, and the majority's score was assumed to be the "true" score of a specific segment for that run. If no majority existed, the segment was considered "unreadable" and dropped from that run. The remaining 7 physicians' WM classifications were then evaluated versus the scores assigned to each segment by the other 5 physicians, and all 5 measures of accuracy were calculated. This process was repeated 10,000 times. Mean and 95% bootstrap confidence intervals for sensitivity, specificity, positive and negative predictive values, and overall accuracy were calculated.

RESULTS

Interobserver Reliability

A total of 1890 segments from 105 patients were visually read, and WM quality was scored by 12 readers (Table 1). An average of 17.4 ± 1.3 segments per patient were scored by the readers (range, 15.8-17.8). There were 78 "unreadable" segments; most were apical (69%) and least were in the basal area (22%). According to the average of 10,000 bootstrap iterations performed, 74.3% of segments were normal, 11.2% were hypokinetic, and 14.5% were akinetic.

Interobserver reliability for classifying segments into 3 categories had an ICC of 0.70. The ICC for dichotomizing segments into normal versus abnormal was 0.63 and for akinetic versus others was 0.58. The reliability for separating hypokinetic from normal or akinetic segments was poor (ICC, 0.26).

Intraobserver Reliability

Of 180 segments from 10 studies arbitrarily selected and reread by 11 of the readers for the assessment of intraobserver variability and reliability, the average number of segments scored per reader was 146.5 ± 6.3 (range, 138-159). There were a total of 182 changes in WM scoring between the duplicate readings (11.3% of segments), and the number of scores changed per reader ranged from 4 to 31 (average, 16.5 ± 8.9; median, 8.5). These alterations in scoring between the first and the second readings changed the assessment from normal to abnormal or vice versa in an average of 8.5 ± 7.4 segments per reader (5.8% of segments per reader).

Intraobserver reliability for classifying segments into 3 categories had a mean ICC coefficient of 0.79 (Table 1, Figure 1). Reliability for dichotomizing segments into normal versus abnormal was 0.71 on average. The lowest reliability was for differentiating hypokinetic from normal and akinetic segments (mean ICC, 0.37).

Effect of Segment Location on Interobserver Reliability

The variability for the assessment of segments in basal, middle, and apical sections of the left ventricle showed a trend for lower concordance and reliability in basal ventricular segments. However, regardless of segmental level, the reliability coefficient for interobserver assessments of hypokinetic segments (range, 0.23-0.29) was almost half of the corresponding coefficients for normal or akinetic segments (range, 0.49-0.73).

Diagnostic Accuracy of Visual Segmental Scores

Sensitivity, specificity, positive and negative predictive values, and accuracy for the visual determination of normal, hypokinetic and akinetic segments by individual readers compared with a segmental score determined by bootstrapping (see "Methods") are presented in Table 2. Whereas overall sensitivity, positive predictive value, and accuracy for the qualitative determination of segmental WM were high, specificity and negative predictive value were lower. The lowest specificity and negative predictive value were for hypokinetic segments (49% and 38%, respectively). A moderately low specificity for akinetic segments (74%) and for the negative predictive value of normally contracting segments (75%) was also noted.

Table 1 Interobserver and intraobserver reliability of visual assessment of WM assessed by ICCs

Variable	Interobserver ICC	Intraobserver Mean ICC ± SD	Intraobserver ICC range
Normal, hypokinetic, or akinetic	0.70	0.79 ± 0.06	0.68-0.88
Normal vs other (abnormal)	0.63	0.71 ± 0.12	0.50-0.91
Hypokinetic vs other	0.26	0.37 ± 0.19	0.03-0.62
Akinetic vs other	0.58	0.70 ± 0.10	0.56-0.83

Interobserver and intraobserver reliability for the determination of WM abnormalities between trichotomized ordinal segmental scores (normal = 1, hypokinetic = 2, and akinetic =3) and dichotomized segmental scores (normal vs abnormal, hypokinetic vs other, and akinetic vs other scores). Note the low values for hypokinetic segments.

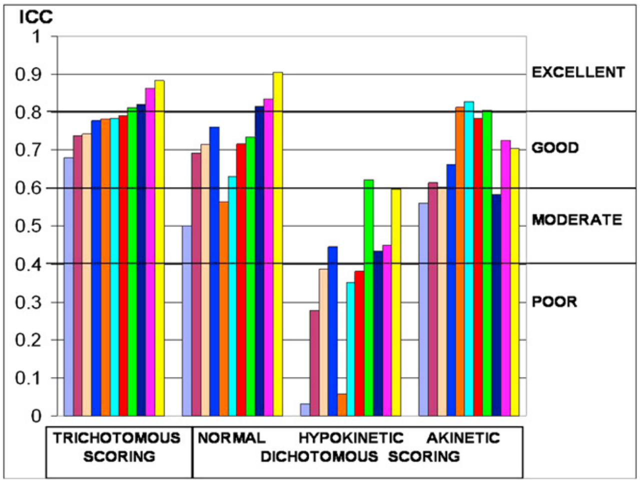

Figure 1 ICCs for intraobserver reliability for trichotomous ordinal scoring (normal, hypokinetic, and akinetic segments) and for dichotomous scoring (each score classification vs the other scores) for each reader. Ranges of the ICCs considered poor, moderate, good, and excellent are specified. Colors represent results of individual readers. Note worst results (ie, lowest reliability) for hypokinetic segments.

Assessment of WMSI

The average WMSI assigned by the readers to the 105 patients was 1.5 ± 0.2 (range of averages, 1.4-1.6; Figure 2). The mean WMSI for patients with normal left ventricles was 1.0 ± 0.0, for patients with ischemic heart disease was 1.3 ± 0.7, and for patients with dilated cardiomyopathy was 2.0 ± 0.9 ($P < .001$ for differences between all pairs). Among patients with MIs who had left anterior descending coronary artery (LAD) disease, the WMSI was 1.4 ± 0.8, compared with 1.1 ± 0.4 for circumflex coronary artery disease ($P < .02$) and 1.3 ± 0.6 for right coronary artery disease ($P =$ NS).

Table 2 Diagnostic accuracy of the visual assessment method for the determination of WM quality versus the mean scores determined for each segment by the bootstrap iterative procedure of 12 experienced readers

	Normal		Hypokinetic		Akinetic	
Variable	Average	95% CI	Average	95% CI	Average	95% CI
Sensitivity	90	87-93	90	88-92	95	93-97
Specificity	85	80-90	49	44-54	74	64-83
PPV	95	92-97	93	91-95	96	93-98
NPV	75	68-82	38	31-46	73	62-83
Accuracy	89	87-90	85	84-86	92	91-93

CI, Confidence interval; *NPV*, negative predictive value; *PPV*, positive predictive value.

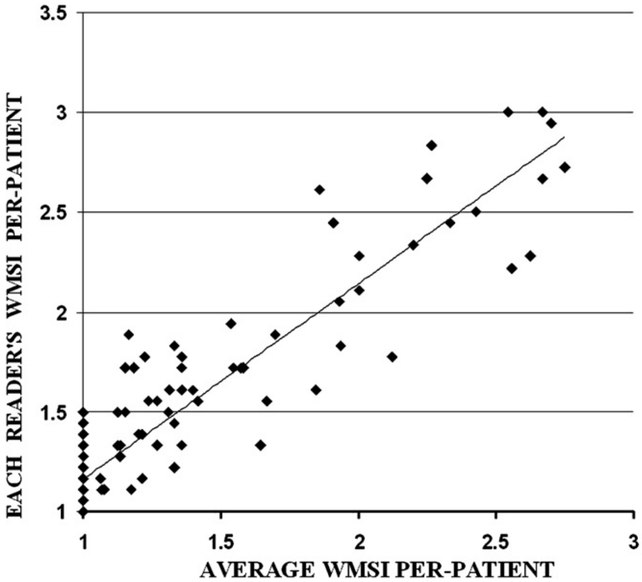

Figure 2 Visual assessment of WMSI by each individual reader per patient versus average of all visual scores per patient. Some points in the figure contain a number of superimposed determinations.

Interobserver variability (by standard deviation) was 0.19 (coefficient of variation, 13%), and reliability (by ICC) was 0.84. Similarly, the mean intraobserver variability was 0.13 ± 0.04 (range, 0.06-0.21), and mean reliability was 0.90 ± 0.06 (range, 0.77-0.98). The relation between individual readers' WMSIs assigned to each patient and their mean is presented in Figure 2.

Visual Assessment of LVEF

The average LVEF was 51.1 ± 5.8% (range of averages, 43.2%-59.5%; Figure 3). The interobserver variability was 5.8% (by standard deviation; coefficient of variation, 11%), and reliability was 0.78 (by ICC). Similarly, the mean intraobserver variability was 5.1 ± 3.4% (range, 2.0%-12.1% for individual readers), and mean reliability was 0.72 (range, 0.16-0.96). The difference of individual readers' estimations of LVEF (for each of the 105 patients) from the mean LVEF assigned to each patient by the 12 readers ranged from −8.1 to +4.4 percentage points (Figure 3), and the mean difference between duplicate LVEF estimates of the individual readers was 3.4 percentage points (range, 0.2-12.7 percentage points).

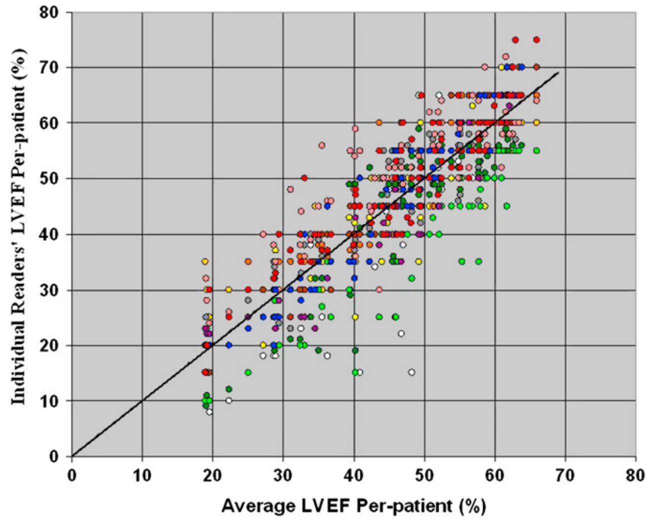

Figure 3 Visual estimation of LVEFs of all patients by each reader versus the average of the 12 readers' LVEF estimations for each patient. Estimations of each reader are represented by a different color.

Relation of Visual Segmental Scores to Angiographic Findings

Eighty-three patients were catheterized; 58 of them had acute MIs with single-vessel disease on angiography (LAD, n = 25; right coronary artery, n = 22; circumflex coronary artery, n = 11), and 25 did not have significant coronary artery disease (Table 3). Sensitivity for the detection of segmental WM abnormalities (segmental majority score) in territories supplied by culprit lesions was 52% (mean of individual readers, 57 ± 7%; range, 44%-68%), specificity was 94% (mean of readers, 86 ± 4%; range, 79%-92%), positive predictive value was 71% (mean of readers, 58 ± 7%; range, 46%-71%), negative predictive value was 88% (mean of readers, 86 ± 2%; range, 83%-88%), and accuracy was 85% (mean of readers, 79 ± 3%; range, 74%-84%). Sensitivity among the subgroup of patients with single-vessel disease was 71%, and specificity among subjects with normal coronary anatomy was 99%.

Correct identification of the culprit artery by location of the majority of segments with WM abnormalities was achieved in an average of 37 ± 5 patients (64% of 58 patients with single-vessel disease, 44% of 83 catheterized patients). When segments were assigned the majority score, 59% of the 58 patients with single-vessel disease had the culprit artery correctly identified, with the LAD being identified most frequently (Table 3). None of the 25 patients with normal coronary arteries were identified as having culprit arteries. Thus, when assigning each segment the majority's score and by the algorithm we used, 77% of the 83 catheterized patients had either correct determinations of their culprit arteries or were correctly identified as not having culprit arteries.

DISCUSSION

This is the most extensive study to date of the reliability, variability, and diagnostic accuracy of visual assessment of WM from echocardiographic recordings performed on a large number of patients with a wide range of pathologies commonly encountered in clinical practice. The largest and most highly qualified cohort of echocardiographers to be reported took part in this multicenter study: 12 experienced echocardiographers, most of them heads of noninvasive

Table 3 Echocardiographic detection of angiographically determined culprit artery by location of the majority of WM abnormalities

Variable	LAD	Circumflex	Right coronary artery	Total
Total number of patients by culprit location	25	11	22	58
No WM abnormality detected	4	5	4	13
No majority of WMA in any territory	0	0	6	6
Correct identification of culprit artery	21 (84%)	5 (45%)	8 (36%)	34 (59%)

Each segment was assigned the majority score of the 12 readers. The table presents the number of patients who had culprit arteries, the number of patients in whom no WM abnormalities were detected, the number of patients without a majority of WM abnormalities in any coronary territory, and the number and percentage (from total) of patients in whom the culprit arteries would have been correctly detected by the algorithm.

cardiology or echocardiography units; thus ours was the only study to have an appropriate number of observers to make meaningful estimates of interobserver and intraobserver variability and reliability.[9] The readers classified each segment by the current classical visual scoring method and also estimated global LVEF. Interobserver and intraobserver reliability of scoring (per segment [segmental score] and per patient [WMSI]) were calculated, agreement with coronary anatomy (defined angiographically) was determined, and interobserver and intraobserver variability and reliability of visual LVEF estimations were calculated.

We found that a significant variability among readers still exists in experienced hands despite improvements in modern echocardiographic hardware and software: interobserver reliability was good (ICC, 0.70) for classifying segments into 3 categories and for dichotomizing segments into normal versus abnormal (0.63); corresponding numbers for intraobserver reliability were somewhat higher (0.79 and 0.71). Not surprisingly, hypokinetic segments had much lower reliability than normal and akinetic segments. Reliability tended to decrease from apical to basal segments but was similar for all segments within each level (basal, middle, or apical level). WMSI had interobserver and intraobserver variability of 0.19 and 0.13 $\pm$ 0.04, and reliability was excellent (0.84 and 0.90 $\pm$ 0.06), respectively) for both. Interobserver and intraobserver variability for LVEF were 5.8 and 5.1 $\pm$ 3.4 percentage points, and their reliability was very good (0.78 and 0.72, respectively). Agreement between the WM abnormalities detected and the territories supplied by culprit arteries was 44% to 68% among the readers.

Significant improvements in software and hardware (eg, second harmonics, better transducers) have improved endocardial detection and myocardial visualization, yet surprisingly, no large-scale studies have been designed to examine the impact of such significant technological improvements on reliability of visual assessment of WM and function. The visual method for assessment of regional WM is highly observer dependent, because the human eye must visualize the endocardial border of the heart, follow it throughout the cardiac cycle, judge its excursion, assess wall thickening and longitudinal shortening during systole, and, as stressed in recent years, also notice minute alterations in timing of contraction.[10] As a result, long training periods and extensive experience are necessary to make echocardiographers proficient in this facet of echocardiography. Thus, current recommendations for level 2 training (ie, "independent performance and interpretation of echocardiograms") include a 6-month training period and the performance of 150 and interpretation of 300 examinations; for level 3 (ie, "a high level of expertise") a training period of 12 months and the performance of 300 and interpretation of 750 echocardiograms are required.[11] All 12 readers who participated in this study had been level 3 readers for many years. Despite their prominent professional status, all agreed to take part in this study without any coordination of definitions for WM abnormalities or any consensus reading sessions prior to the initiation of the study, because this study was intended to reflect the best scenario of "real-world," "eyeball" assessments of segmental WM.

Data in the literature about interobserver and intraobserver variability and agreement for the visual detection and assessment of WM are scarce. Peart et al[12] analyzed LV regional WM (the left ventricle divided into 5 segments, 30 subjects, 2 readers, 4 categories, and 2 occasions) and found "large variability" both between and within readers. Vermes et al[13] also showed large interobserver variability (the left ventricle divided into 7 segments, 15 subjects, 2 readers, and 2 categories). An intraobserver difference of 7% was reported by Badano et al[14] (the left ventricle divided into 16 segments, 105 subjects, only 2 readers, and 5 categories), similar to the 5.6 $\pm$ 2.9% of segments in our study that were assigned different scores by the same reader in duplicate readings. Hoffmann et al[15] addressed the issue of variability in the detection of WM as a byproduct of a multicenter study on dobutamine stress echocardiography (which requires the ability to identify new WM abnormalities) and reported a κ coefficient of only 0.37 for interobserver variability (150 subjects, 5 readers, and 4 categories), although with better imaging, a κ coefficient of 0.55 was achieved.[16] Recently Hoffmann et al[17] again addressed the issue of interobserver agreement in a multicenter comparison of methods and found a κ coefficient of 0.41 for detecting regional WM abnormalities on conventional echocardiography (the left ventricle divided into 16 segments, 100 subjects, 3 readers, and 4 categories). Whereas all these studies used the κ coefficient to describe interobserver and intraobserver variability (or, more precisely, interobserver and intraobserver agreement), we used the more statistically appropriate ICC to describe reliability,[7] although both are forms of correlation coefficients and are comparable. In general, the above studies showed moderate agreement between observers, whereas our data show moderate to good reliability for both interobserver and intraobserver reliability, but this difference is small and may be related in part to the interpretation of the different statistical indices used.

We found the lowest reliability, specificity, and negative predictive value in hypokinetic segments. This is not surprising, considering that, as Hoffmann et al[15] and others have pointed out, no clear criteria have been established for defining hypokinesis: some may consider mildly hypokinetic segments "normal" and severely hypokinetic areas "akinetic." Such segments may be classified differently even by the same observer on separate occasions. Another problem arises when contraction abnormalities do not overlap the arbitrary definition of segments, leading to nonhomogenous contraction within segments, thus forcing the reader to choose a score that in his or her opinion best represents the entire segment's contraction. Finally, it may be argued that there are readers who are even more experienced than ours, who may be able to categorize hypokinetic segments with better concordance, yet there is no doubt that the readers who participated in this study are more experienced than the average echocardiographic reader in real-world practice.

To avoid the obstacle of correctly separating hypokinetic from akinetic segments, we also looked at interobserver and intraobserver reliability for discriminating normal segments from those with abnormal contraction of any degree. Among our 12 readers, the interobserver and intraobserver reliability for dichotomizing segments as normal versus abnormal was only moderate (ICCs of 0.63 and 0.71 ± 0.12, respectively). This compares favorably with other modalities Hoffmann et al[17] tested, such as magnetic resonance imaging ($\kappa = 0.43$) and cine ventriculography ($\kappa = 0.56$), but is inferior to contrast-enhanced echocardiography ($\kappa = 0.77$).

The echocardiographic method for WM evaluation reflects real-life practice, as this qualitative way of scoring WM is accepted worldwide. Comparing the results of our readers with a reference standard formed by the bootstrap sample of echocardiographic data may not appear as strong as comparing them with a completely different modality, but such comparisons have traditionally served as a basis for clinical decision making, prognostication, and guideline formation.[18,19] Others have also used echocardiography as a reference[20] and as a gold standard.[21] In addition, we used this opportunity to compare our readers' results with angiographic data. Clearly, angiographic findings of epicardial coronary vessels are no reference for WM abnormalities, especially in the setting of an acute MI when echocardiography and angiography were not performed at the same time. However, it is safe to assume that when segmental WM abnormalities were detected in territories supplied by culprit lesions, they were correct. Although sensitivity for detection of WM abnormalities in territories supplied by culprit lesions was somewhat disappointing (52%, and even among patients with single-vessel disease, it was only 71%), it was reassuring to find a good accuracy rate for the assessment of WM abnormalities (85%) and to note 99% specificity for assessments of WM in subjects with chest pain but normal coronary anatomy. The relatively good concordance between the visual assessment of WM abnormalities and the actual coronary anatomy again demonstrates the robustness of the time-honored visual method. On the other hand, in keeping with clinical experience, the visual assessment of WM abnormalities could only predict the culprit artery (defined as the coronary artery supplying the territory in which the majority of WM abnormalities were found) in slightly more than half of our patients with acute MIs. The best prediction of the culprit artery was made among patients whose LADs were the culprit arteries (Table 3). It is not surprising that there was an incomplete match between the location of segments with WM abnormalities and territories supplied by culprit arteries. There are many factors that explain this, among them the well-known discrepancies between anatomic territories allegedly and actually supplied by coronary arteries, the presence of collateral blood supply, undetectable subendocardial damage, dependence on the proximity of obstructive lesions, and the fact that most patients underwent echocardiography within 48 hours of primary percutaneous coronary intervention. Our data for the detection of WM abnormalities in the setting of acute MI are similar to data on acute ischemia provoked by dobutamine in single-vessel disease, for which agreement of only 59% between observers was described by Hoffmann et al,[15] and a better result was reported by Takeuchi et al,[22] who analyzed a small number of patients by 2 readers.

Global LV function, estimated visually either by LVEF or by WMSI, is an important assessment performed routinely. For WMSI, the ICC was marginally better than for LVEF (interobserver, 0.84 vs 0.78; intraobserver mean, 0.90 vs 0.72), suggesting that the assessment of LV systolic function by WMSI may be slightly more reliable than by LVEF, although the latter had a slightly lower coefficient of variation (11% vs 13%). The overall interobserver variability we report of 5.8% is smaller than the 15% to 17% mentioned in a review by McGowan and Cleland[23] and even the 8% reported by Jensen-Urstad et al,[24] and the mean intraobserver variability of 5.1 ± 3.4% we found is smaller than the 11% to 13% they quoted. The 11% coefficient of variation for visual estimations of LVEF that we found is almost identical to the 10% reported by the latter.

Limitations

First, visual determination of WM scores is not expected to reflect the actual quality of the WM contraction but rather the way it is perceived by experienced echocardiographers. In this study, we were interested in comparing the reliability of agreement between them (interobserver reliability) and the consistency of each reader's assessments (intraobserver reliability), not necessarily with the objective, true WM score (for which there is no clinically useful gold standard) but with the method used in everyday life, on which major clinical decisions are based. For determination of the diagnostic accuracy parameters of the visual assessment method, it was compared with the score given to each readable segment by the majority of readers, using the bootstrap method. Previous clinical studies used "gold standards" based on magnetic resonance imaging tagging,[25] on independent expert panel decisions,[17] on comparison with catheterization data,[15] or even on a consensus of readers.[21]

Second, no parasternal views were available for the readers, though they are an integral part of echocardiographic examinations and have their contribution to the final assessment of WM. Thus, it must be emphasized that our data do not relate to the issue of the overall accuracy of visual assessment of echocardiograms, for which multiple views and observation of many cardiac cycles are needed. Our data are limited to determination of the reliability of visual WM abnormality assessments by experienced echocardiographers from only 3 apical views and from only 1 loop per view, thus ensuring that all observers were analyzing exactly the same segment in the same cardiac cycle in each view.

Third, in this study, only apical views were used, which are more difficult to assess because of endocardial dropout from the parallel ultrasound beam, which may account for part of the variability.

Fourth, the patients studied were not sequentially recruited but were selected for adequate quality echocardiograms and to represent patients with a range of WM abnormalities, such as MI and dilated cardiomyopathy, and normal subjects. This selection may have affected our results.

CONCLUSIONS

We have established that interobserver and intraobserver reliability for assessments of WM, WMSI, and LVEF using the classical, visual method for analysis of echocardiographic loops obtained with contemporary echocardiographic machinery when done by expert echocardiographers is reasonable. Reliability for the determination of hypokinetic segments is poor, even among experienced echocardiographers. A precise definition of hypokinesis and the development of objective methods for qualitative and quantitative WM analysis would undoubtedly improve the uniformity of assessments and may assist in the interpretation of echocardiograms by novice as well as experienced readers.

REFERENCES

1. Cerqueira MD, Weissman NJ, Dilsizian V, Jacobs AK, Kaul S, Laskey WK, et al. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. Circulation 2002;105:539-42.
2. Lang RM, Bierig M, Devereux RB, Flachskampf FA, Foster E, Pellikka PA, et al. Recommendations for chamber quantification: a report from the American Society of Echocardiography's Guidelines and Standards Committee and the Chamber Quantification Writing Group, developed in conjunction with the European Association of Echocardiography, a branch of the European Society of Cardiology. J Am Soc Echocardiogr 2005;18:1440-63.
3. Solomon SD, Anavekar N, Skali H, McMurray JJ, Swedberg K, Yusuf S, et al. Influence of ejection fraction on cardiovascular outcomes in a broad spectrum of heart failure patients. Circulation 2005;112:3738-44.
4. Nixdorff U, Matschke C, Winklmaier M, Flachskampf FA, Ludwig J, Pohle K, et al. Native tissue second harmonic imaging improves endocardial and epicardial border definition in dobutamine stress echocardiography. Eur J Echocardiogr 2001;2:52-61.
5. Sozzi FB, Poldermans D, Bax JJ, Boersma E, Vletter WB, Elhendy A, et al. Second harmonic imaging improves sensitivity of dobutamine stress echocardiography for the diagnosis of coronary artery disease. Am Heart J 2001;142:153-9.
6. Feigenbaum H, Armstrong WF, Ryan T. Feigenbaum's echocardiography. 6th ed. Philadelphia, PA: Lippincott, Williams & Wilkins; 2004.
7. Shrout PE, Fleiss JE. Intraclass Correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-6.
8. Efron B, Tibshirani R. An introduction to the bootstrap. Boca Raton, FL: Chapman & Hall/CRC; 1994.
9. Sadatsafavi M, Najafzadeh M, Lynd L, Marra C. Reliability studies of diagnostic tests are not using enough observers for robust estimation of interobserver agreement: a simulation study. J Clin Epidemiol 2008;61:722-7.
10. Kvitting JP, Wigstrom L, Strotmann JM, Sutherland GR. How accurate is visual assessment of synchronicity in myocardial motion? An in vitro study with computer-simulated regional delay in myocardial motion: clinical implications for rest and stress echocardiography studies. J Am Soc Echocardiogr 1999;12:698-705.
11. Quinones MA, Douglas PS, Foster E, Gorcsan J III, Lewis JF, Pearlman AS, et al. ACC/AHA clinical competence statement on echocardiography: a report of the American College of Cardiology/American Heart Association/American College of Physicians-American Society of Internal Medicine Task Force on Clinical Competence. J Am Soc Echocardiogr 2003;16:379-402.
12. Peart I, Austin A, Hall RJ. Subjective analysis of cross-sectional echocardiograms: reproducibility and sources of variability. Eur Heart J 1987;8:171-8.
13. Vermes E, Guyon P, Weingrod M, Otmani A, Soussana C, Halphen C, et al. Assessment of left ventricular regional wall motion by color kinesis technique: comparison with angiographic findings. Echocardiography 2000;17:521-7.
14. Badano L, Stoian J, Cervesato E, Bosimini E, Gentile F, Giannuzzi P, et al. Reproducibility of wall motion score and its correlation with left ventricular ejection fraction in patients with acute myocardial infarction. Am J Cardiol 1996;78:855-8.
15. Hoffmann R, Lethen H, Marwick T, Arnese M, Fioretti P, Pingitore A, et al. Analysis of interinstitutional observer agreement in interpretation of dobutamine stress echocardiograms. J Am Coll Cardiol 1996;27:330-6.
16. Hoffmann R, Marwick TH, Poldermans D, Lethen H, Ciani R, van der Meer P, et al. Refinements in stress echocardiographic techniques improve inter-institutional agreement in interpretation of dobutamine stress echocardiograms. Eur Heart J 2002;23:821-9.
17. Hoffmann R, von Bardeleben S, Kasprzak JD, Borges AC, ten Cate F, Firschke C, et al. Analysis of regional left ventricular function by cineventriculography, cardiac magnetic resonance imaging, and unenhanced and contrast-enhanced echocardiography: a multicenter comparison of methods. J Am Coll Cardiol 2006;47:121-8.
18. Kirkpatrick JN, Vannan MA, Narula J, Lang RM. Echocardiography in heart failure: applications, utility, and new horizons. J Am Coll Cardiol 2007;50:381-96.
19. Thygesen K, Alpert JS, White HD, Jaffe AS, Apple FS, Galvani M, et al. Universal definition of myocardial infarction. Circulation 2007;116:2634-53.
20. Dirksen MS, Jukema JW, Bax JJ, Lamb HJ, Boersma E, Tuinenburg JC, et al. Cardiac multidetector-row computed tomography in patients with unstable angina. Am J Cardiol 2005;95:457-61.
21. Lessick J, Mutlak D, Rispler S, Ghersin E, Dragu R, Litmanovich D, et al. Comparison of multidetector computed tomography versus echocardiography for assessing regional left ventricular function. Am J Cardiol 2005;96:1011-5.
22. Takeuchi M, Sonoda S, Miura Y, Kuroiwa A. Reproducibility of dobutamine digital stress echocardiography. J Am Soc Echocardiogr 1997;10:344-51.
23. McGowan JH, Cleland GF. Reliability of reporting left ventricular systolic function by echocardiography: a systematic review of 3 methods. Am Heart J 2003;146:388-97.
24. Jensen-Urstad K, Bouvier F, Höjer J, Ruiz H, Hulting J, Samad B, et al. Comparison of different echocardiographic methods with radionuclide imaging for measuring left ventricular ejection fraction during acute myocardial infarction treated by thrombolytic therapy. Am J Cardiol 1998;81:538-44.
25. Zhang Y, Chan AK, Yu CM, Yip GW, Fung JW, Lam WW, et al. Strain rate imaging differentiates transmural from non-transmural myocardial infarction: a validation study using delayed-enhancement magnetic resonance imaging. J Am Coll Cardiol 2005;46:864-71.