ORIGINAL ARTICLE

 CrossMark

Inter-reader variability of SPECT MPI readings in low- and middle-income countries: Results from the IAEA-MPI Audit Project (I-MAP)

Maurizio Dondi, MD,[a,g] Carlo Rodella, MSc,[b] Raffaele Giubbini, MD,[a]
Luca Camoni, BSc,[a] Ganesan Karthikeyan, MD, DM, MSc,[c] Joao V. Vitola, MD,[d]
Andrew J. Einstein, MD, PhD, FASNC,[e] Bertjan J. Arends, MSc,[f] Olga Morozova,
BA, MA,[g] Thomas N. Pascual, MD, MHPED,[g] and Diana Paez, MD, MsED[g] on
behalf of the I-MAP investigators

[a] Nuclear Medicine Department, University of Brescia, Brescia, Italy
[b] Department of Medical Physics, Spedali Civili di Brescia, Brescia, Italy
[c] Department of Cardiology, All India Institute of Medical Sciences, New Delhi, India
[d] Quanta Diagnóstico & Terapia, Curitiba, Brazil
[e] Department of Medicine, Division of Cardiology, and Department of Radiology, Columbia
   University Irving Medical Center, New York
[f] Department of Medical Physics, Catharina Hospital, Eindhoven, The Netherlands
[g] Nuclear Medicine and Diagnostic Imaging Section, Division of Human Health, International
   Atomic Energy Agency, Vienna, Austria

Background. Consistency of results between different readers is an important issue in medical imaging, as it affects portability of results between institutions and may affect patient care. The International Atomic Energy Agency (IAEA) in pursuing its mission of fostering peaceful applications of nuclear technologies has supported several training activities in the field of nuclear cardiology (NC) and SPECT myocardial perfusion imaging (MPI) in particular. The aim of this study was to verify the outcome of those activities through an international clinical audit on MPI where participants were requested to report on studies distributed from a core lab.

Methods. The study was run in two phases: in phase 1, SPECT MPI studies were distributed as raw data and full processing was requested as per local practice. In phase 2, images from studies pre-processed at the core lab were distributed. Data to be reported included summed stress score (SSS); summed rest score (SRS); summed difference score (SDS); left ventricular (LV) ejection fraction (EF) and end-diastolic volume (EDV). Qualitative appraisals included the assessment of perfusion and presence of ischemia, scar or mixed patterns, presence of transient ischemic dilation (TID), and risk for cardiac events (CE). Twenty-four previous trainees from low- and middle-income countries participated (core participants group) and their results were assessed for inter-observer variability in each of the two phases, and for

---

466   Dondi et al.
Inter-reader variability of SPECT MPI readings in low- and middle-income countries

Journal of Nuclear Cardiology®
March/April 2020

changes between phases. The same evaluations were performed for a group of eleven inter-national experts (experts group). Results were also compared between the groups.

Results. Expert readers showed an excellent level of agreement for all parameters in both phase 1 and 2. For core participants, the concordance of all parameters in phase 1 was rated as good to excellent. Two parameters which were re-evaluated in phase 2, namely SSS and SRS, showed an increased level of concordance, up to excellent in both cases. Reporting of categorical variables by expert readers remained almost unchanged between the two phases, while core participants showed an increase in phase 2. Finally, pooled LVEF values did not show a significant difference between core participants and experts. However, significant differences were found between LVEF values obtained using different software packages for cardiac analysis.

Conclusions. In this study, inter-observer agreement was moderate-to-good for core group readers and good-to-excellent for expert readers. The quality of reporting is affected by the quality of processing. These results confirm the important role of the IAEA training activities in improving imaging in low- and middle-income countries. (J Nucl Cardiol 2020;27:465–8.)

Key Words: CAD · Myocardial ischemia and infarction · Gated SPECT · Image interpretation

Abbreviations	
IAEA	International Atomic Energy Agency
MPI	Myocardial perfusion imaging
LVEF	Left ventricle ejection fraction
SSS	Summed stress score
SRS	Summed rest score
SDS	Summed difference score
EDV	End-diastolic volume
TID	Transient ischemic dilation
CE	Cardiac events
PHR	Patient high risk

INTRODUCTION

The International Atomic Energy Agency (IAEA) is an independent, intergovernmental science and technology-based organization which is part of the United Nations family of organizations.[1] The IAEA works with its 170 Member States (MS) and multiple partners worldwide to promote the safe, secure and peaceful use of nuclear technologies. The IAEA supports nuclear medicine through activities of the Nuclear Medicine and Diagnostic Imaging Section (NMDI) within a quality assurance framework.[2,3] The nuclear medicine programme contributes to achieving the sustainable development goals (SDGs) set by the United Nations, one of which is ''by 2030, reduce by one third premature mortality from non-communicable diseases through prevention and treatment and promote mental health and well-being''.[4]

Considering the burden of cardiovascular diseases (CVD) as a major threat to public health worldwide,[5,6] and the important role of nuclear techniques such as myocardial perfusion imaging (MPI) in the management of patients with ischemic heart disease (IHD),[7-9] the NMDI Section adopted a strategic decision of strengthening capacity building in nuclear cardiology (NC), providing training through national and regional projects,[10] supported by the Technical Cooperation Programme (TCP), which is the IAEA's main mechanism for transferring nuclear technology to Low- and Middle-Income Countries (LMICs).[11] Educational activities in NC include several Regional Training Courses (RTC) carried out over the past ten years.

This paper reports the results of an audit of NC practices (the I-MAP study), initiated in 2015 to assess whether and how training provided through RTCs impacted the quality of clinical practice. The primary goal was to assess homogeneity (i.e. intra- and inter-observer variability) within a group of core participants from LMICs. As secondary goals the study aims at a) evaluating the impact of IAEA activities in NC; b) comparing the readings of MPI studies in limited resource centres with those of international experts; c) evaluating the quality of reporting and d) assessing the impact of the reconstruction of MPI studies on the quality of reporting.

METHODS

Recorded contact data from all attendees to RTCs in NC was retrieved. In the preceding 10 years, 896 participants had attended a total of 41 RTCs. Their regional distribution is reported in Appendix (Table 5). To make sure that those trainees, prospective participants to this study, were still actively involved in NC, that list was cross-checked with data from an international database managed by the IAEA.[12] Of the 896 participants, 275 were identified as being currently active as nuclear cardiologists, and were approached for potential

Figure 1. Worldwide distribution of both participants and experts. Red dot identifies the Core Lab (Nuclear Medicine Dept, University of Brescia). The size of dots reflects more than one participant from the same country.

participation this study. Of these, 24/275 (8.7%), participated in the study. They formed the group referred to as ''core participants.'' Figure 1 reports their distribution around the world. The ''core participants'' group included physicians trained in nuclear medicine, with limited formal training in nuclear cardiology, in most cases acquired through short-term fellowships supported by the IAEA and/or trained ''on the job.'' Their yearly average volume of MPI studies was 880, with a minimum of 559 and a maximum of 1200.

The second group of ''expert readers'' consisted of eleven international experts identified by the Agency from a pool of its consultants and lecturers, and internationally recognized nuclear cardiologists. Overall, for the experts, the yearly volume of SPECT-MPI studies was on average double that of the core participants.

Both core participants and expert readers were requested to report anonymized case studies provided by a Core Lab, chosen on the basis of sound NC practice and significant record of research. The core lab identified 15 studies which, after anonymization, were uploaded onto a cloud-based collaborative platform (SharePoint™) and then downloaded from both core participants and experts.

All studies were carried out with the two-day protocol, using Tc99m labelled perfusion agents, and patients were

Table 1. Patients' data and clinical status

	Males	**Females**
#	9	6
Age (mean)	71.1	64.5
BMI (mean)	26.4	29.8
Mean HR at rest (bpm)	64.3	74.8
Mean systolic BP	141.1	150.8
Mean diastolic BP	87.8	91.7
Previous MI (yes/no)	5/2 + 2 CABG	3/3
Stressor (Ex/Pharm)	4/5	1/5
Symptoms at peak stress (yes/no)	5/4	4/2

BMI, Body Mass Index; *HR*, heart rate; *BP*, blood pressure; *MI*, myocardial infarction; *Ex*, exercise stress test; *Pharm*, pharmacologic stress test; *CABG*, Coronary Artery by-pass graft

imaged only in supine position. To provide readable studies for centres with limited technical resources, the core lab was asked to send studies processed with neither resolution recovery, nor scatter or attenuation correction, nor studies acquired with CZT cameras. Clinical data, including patients' history, rest
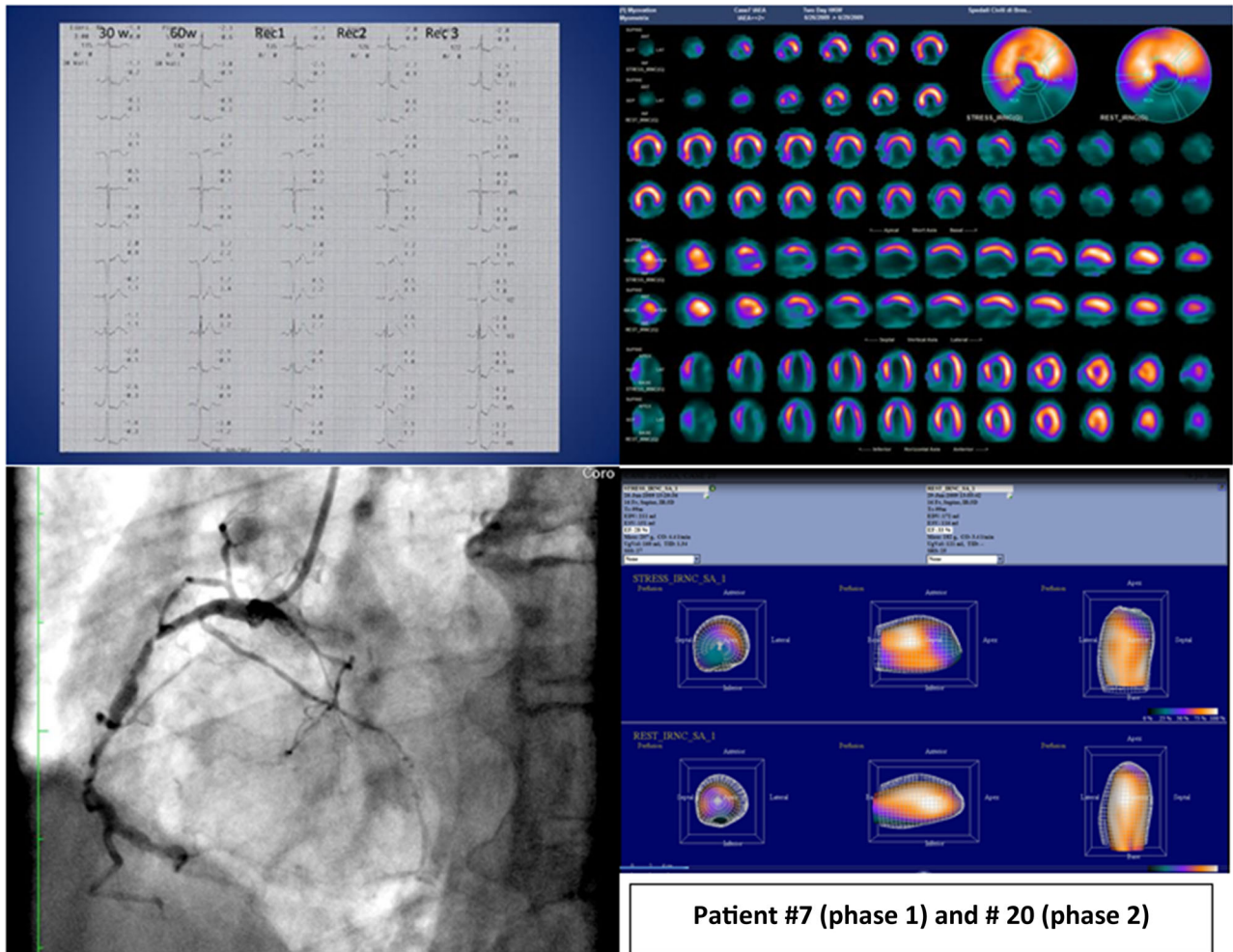
Patient #7 (phase 1) and # 20 (phase 2)

Figure 2. Example of a case study as distributed in phase 2: Male; 76 y-o; Family history of CAD; Hypertensive; Inferior AMI in 1970; relapse 1 year later; 1991 Coro: 75% stenosis RCA; distal occlusion RCA; 75% stenosis distal LAD; occlusion D2; patent LCX; OMT until 2009; Referred for MPI in 2009; Bicycle exercise; max workload 60 W for 3'; typical chest pain; ECG: inferior-lateral ST downslope and runs of NSVT. *CAD*, Coronary artery disease; *AMI*, acute myocardial infarction; *RCA*, right coronary artery; *LCX*, left circumflex; *OMT*, optimized medical therapy; *MPI*, myocardial perfusion imaging; *NSVT*, non-sustained ventricular tachycardia.

and stress ECG recordings and symptoms during stress were made available to participants. Relevant demographic and clinical data are summarized in Table 1.

We designed I-MAP to be run in two phases. In Phase 1, all 15 patient studies were provided as raw data. Both groups were requested to process them according to their own routine practice. For Phase 2, the same 15 cases were re-submitted in a different order, but pre-processed at the core lab using Myovation v3 software (GE Health Care; Haifa, Israel) with an iterative reconstruction ordered subset expectation maximization algorithm (2 iterations, 10 subsets) and motion correction. The ''cool'' GE colour scale was applied for tomographic slices representation. Both groups of participants were unaware that they were re-reading the same studies. This second phase was aimed at assessing whether reconstruction

could have any impact on the overall quality of the study and consistency of interpretation. An example of a pre-processed patient study, as distributed in phase 2, is illustrated in Figure 2.

We used standardized forms for data collection which were forwarded to the core lab for statistical analysis. After on-site processing for phase 1, and based on images provided by the core lab for phase 2, readers were requested to score tracer uptake in polar maps using a 17-segment model (Figure 3A). An important distinction is that while in phase 1 readers could accept any score given by the cardiac software, in phase 2 they had to digit their own interpretation. The severity of perfusion defects in each of the 17 myocardial segments, as defined by the American Heart Association[13] is scored on a 0-4 scale.
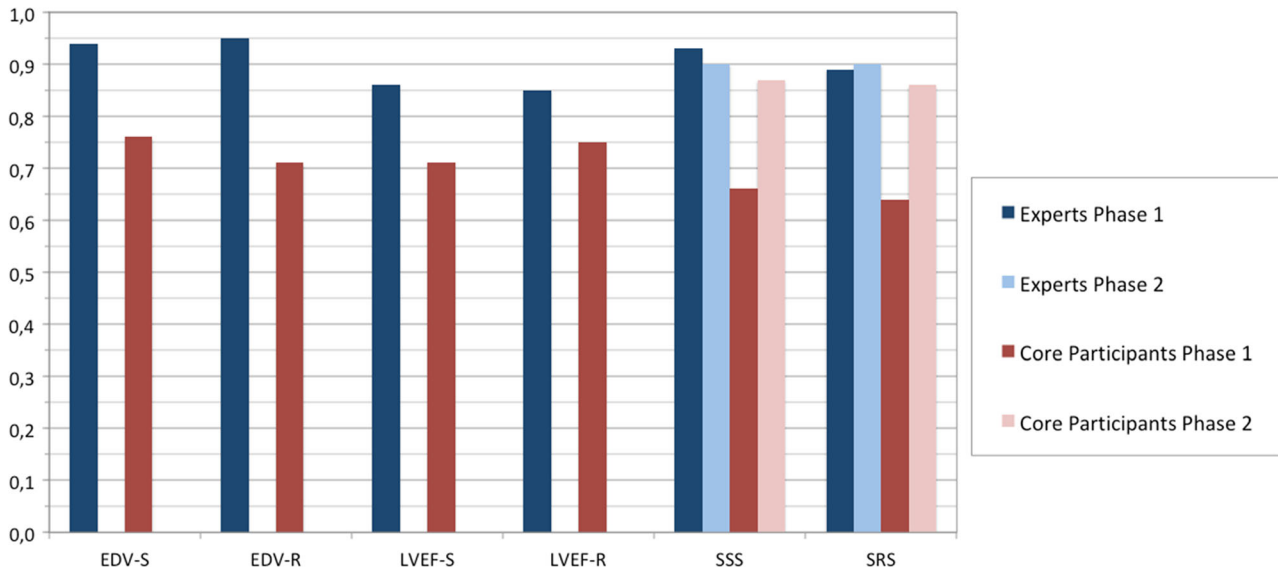
Figure 3. ICC values for continuous variables (EDV, LVEF, SSS, SRS). Calculation of EDV and LVEF was not requested for phase 2. *EDV-S/R*, end-diastolic volume post-stress/at rest; *LVEF-S/R*, left Ventricle Ejection Fraction post stress/at rest; *SSS*, summed stress score; *SRS*, summed rest score.

Table 2. ICC results (values and category of agreement)

Parameter	Experts				Participants			
	Phase 1		Phase 2		Phase 1		Phase 2	
EDV post stress	0.94	(excellent)	–	–	0.76	(excellent)	–	–
EDV rest	0.95	(excellent)	–	–	0.71	(good)	–	–
LVEF post stress	0.86	(excellent)	–	–	0.71	(good)	–	–
LVEF rest	0.85	(excellent)	–	–	0.75	(excellent)	–	–
SSS	0.93	(excellent)	0.90	(excellent)	0.66	(good)	0.87	(excellent)
SRS	0.89	(excellent)	0.90	(excellent)	0.64	(good)	0.86	(excellent)

Data to be reported included quantitative perfusion metrics such as Summed Stress Score (SSS); Summed Rest Score (SRS); and Summed difference Score (SDS). SDS results were pooled to generate three categories: (a) SDS ≤ 3; (b) $4 \leq SDS \leq 7$ and (c) SDS ≥ 8.[14].

For left ventricular function, quantitative data were reported on Left Ventricular Ejection Fraction (LVEF) and End Diastolic Volume (EDV), while regional wall motion was reported based on visual assessment. Other qualitative, or visual, appraisals included the assessment of perfusion, classified as normal or abnormal. In this latter case, readers had to report presence of ischemia, scar or mixed patterns. Another parameter visually analysed was presence or absence of Transient Ischemic Dilation (TID). Both groups were also requested to provide an overall judgment about patients being at high risk or not (PHR).

Furthermore, we aimed at assessing the relationship between the overall judgment of the status of perfusion, either normal or abnormal, and uptake scores (SSS; SRS) as the sum of scores assigned to each single segment. To this purpose and to avoid the possibility that high SSS values could just be the result of the sum of mild defects scattered throughout the myocardial wall, not representing significant perfusion defects, we defined ''hypoperfusion cluster'' as the presence of a real perfusion defect, when two adjacent segments scored ≥ 2. Then, we assessed the relationship between SSS values and the number of hypoperfusion clusters identified in the polar maps.

To evaluate the inter-reader concordance of hypoperfusion assessments, SDS values were stratified into three categories, a) SDS ≤ 3; b) $4 \leq SDS \leq 7$ and c) SDS ≥ 8.[14] For each study, each group of readers (both experts and core participants), and for both phase 1 and 2, the rate of responses
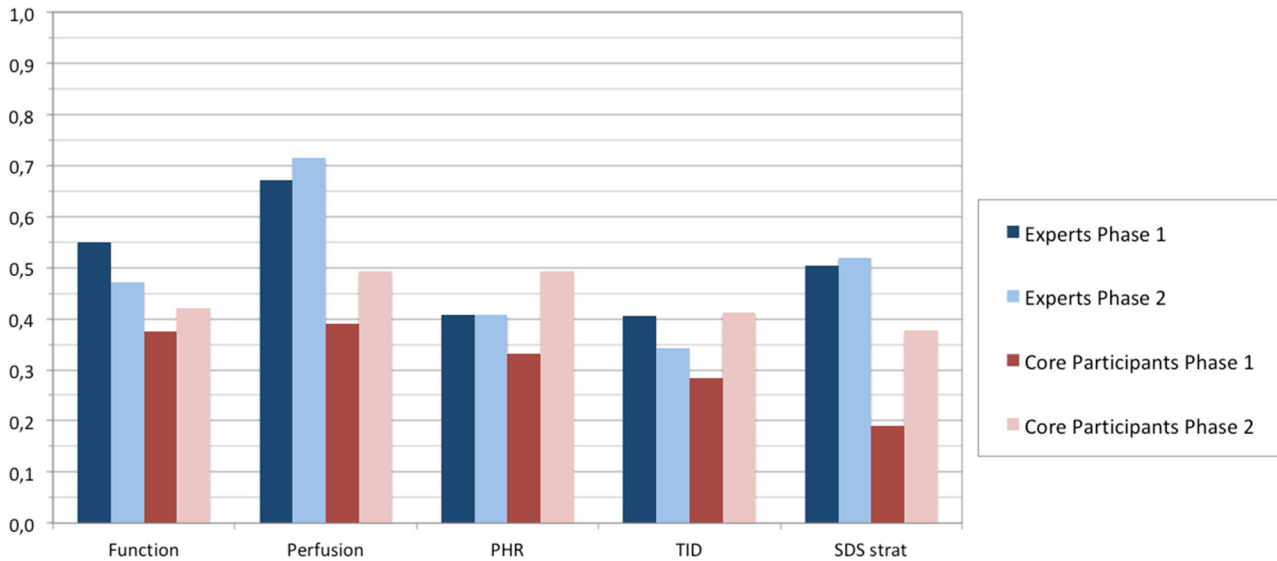
Figure 4. Calculated Fleiss' kappa values for categorical variables. *TID*, Transient Ischemic Dilation; *SDS*, Summed Differential Score stratified; *PHR*, patient high risk).

Table 3. Fleiss' kappa results (values and category of agreement)

Parameter	Experts				Participants			
	Phase 1		Phase 2		Phase 1		Phase 2	
Function	0.55	(moderate)	0.47	(moderate)	0.38	(fair)	0.42	(moderate)
Perfusion	0.67	(substantial)	0.71	(substantial)	0.39	(fair)	0.49	(moderate)
PHR	0.41	(moderate)	0.41	(moderate)	0.33	(fair)	0.49	(moderate)
TID	0.41	(moderate)	0.34	(fair)	0.28	(fair)	0.41	(moderate)
SDS strat	0.50	(moderate)	0.52	(moderate)	0.19	(slight)	0.38	(fair)

PHR, patient high risk; *TID*, transient ischemic dilation; *SDS strat*, stratification of perfusion defects

for each of the three different SDS categories, was evaluated. These three categories have been called ''SDS strat.''

For phase 1 we also tested the consistency of quantitative data, such as LVEF and EDV, since they were calculated using different software. This evaluation was run only for phase 1, since in phase 2 participants were provided pre-processed studies. Variables LVEF post stress and LVEF rest were analyzed using univariate analysis of variance (ANOVA).

Finally, we tested the repeatability of LVEF values when different processing software was used. To avoid the increased risk of Type I errors because of the multiple simultaneous hypotheses being tested, we adjusted *P* values using the Bonferroni method.[15]

Statistical Analysis

For statistical analysis, data were collected on Excel spread sheets and analysed using the Statistical Package for Social Sciences (SPSS; IBM® SPSS® Statistics Release 24); For hypothesis testing, Student's t-test, analysis of covariance (ANCOVA), ANOVA, and Chi-square test for proportions were used as appropriate, the latter for assessing difference in response rates between groups and phases. Intra-rater and inter-rater agreement were assessed:

- by means of the intra-class correlation coefficient (ICC), for continuous measurements (EDV, LVEF, SSS, SRS, SDS). ICC is a measure of agreement that combines information on both the correlation and the systematic differences between readings[16,17]; using ICC, the level of agreement is classified into four categories
- by means of the Fleiss' kappa, for categorical variables (Function, Perfusion, TID, SDS strat, patient high risk). Using Fleiss' kappa (κ) scores, the level of agreement is classified into seven categories.[18-20]

Values for either SSS and SDS reported from the two groups in phase 1, when MPI studies were supplied as raw data and each participant had to completely process and assess
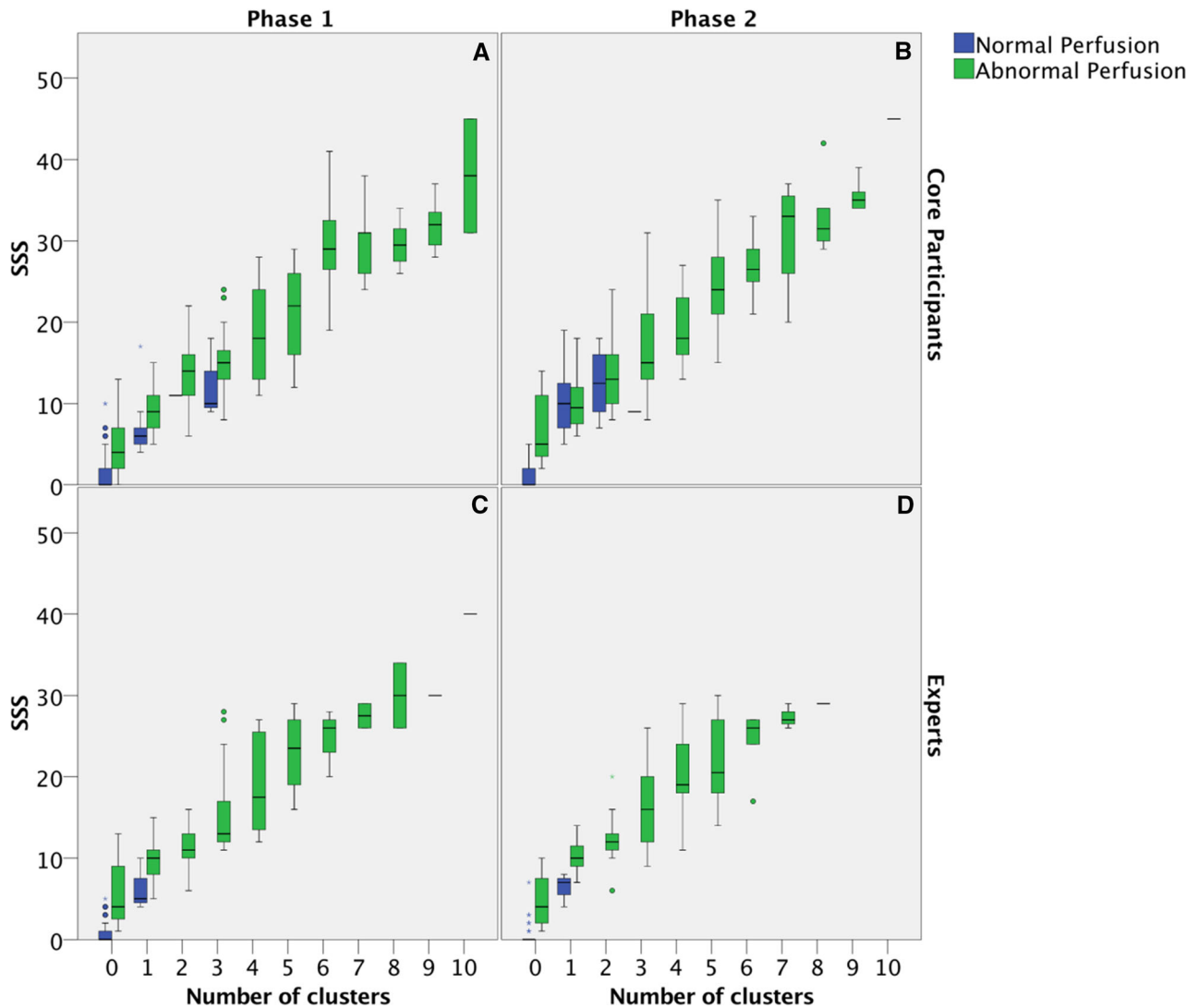
Figure 5. SSS value as function of the number of hypoperfusion clusters. Green boxes represent patients whose perfusion has been judged as abnormal. Blue boxes represent normal perfusion judgments. (**A**) Experts phase 1, (**B**) Experts phase 2, (**C**) Core participants phase 1, (**D**) Core participants phase 2. The figure shows box-and-whiskers plot, showing the median, quartiles, and outlier and extreme values for a scale variable. The interquartile range (IQR) is the difference between the 75th and 25th percentiles and corresponds to the length of the box. Circles outside the boxes represent OUTLIERS. Outliers are values between 1.5 IQR's and 3 IQR's from the end of a box. Stars represent EXTREME whose values are more than 3 IQR's from the end of a box.

using their own software, were compared with those reported from phase 2, where studies were supplied pre-processed at the core lab and participants had to visually score segmental perfusion.

RESULTS

For continuous variables (EDV; LVEF; SSS; SRS) ICC values and the corresponding concordance category are reported in Figure 3 and in Table 2.

Metrics for EDV and LVEF are assessed only for phase 1, as in phase 2 these data were already calculated at the Core lab. Expert readers showed an excellent level of agreement for all parameters in both phase 1 and 2, spanning from 0.85 for LVEF at rest to 0.94 for EDV post-stress. In phase 1, concordance levels for core participants were rated as good for all parameters (from 0.64 to 0.71), except for LVEF at rest and EDV post stress, which were rated as excellent (0.75 and 0.76, respectively). Interestingly, both parameters which were re-evaluated in phase

Table 4. Linear interpolation slopes for average SSS values vs cluster numbers

| Comparison | Items | ANCOVA | | |
| | | Linear regression slope coefficient comparison | | |
		Phase 1	Phase 2	Test results
Phase 1 vs Phase 2	Experts	Slope $= 3.940$	Slope $= 4.068$	$F = 0.90$
		$R^2 = 0.831$	$R^2 = 0.835$	$P = 0.533$
		$s_{slope} = 0.141$	$s_{slope} = 0.149$	
	Core participants	Slope $= 3.756$	Slope $= 4.019$	$F = 3.64$
		$R^2 = 0.836$	$R^2 = 0.830$	$P = 0.057$
		$s_{slope} = 0.087$	$s_{slope} = 0.110$	
Comparison	Items	Experts	Core participants	Test results
Experts vs core participants	Phase 1	Slope $= 3.940$	Slope $= 3.756$	$F = 1.22$
		$R^2 = 0.831$	$R^2 = 0.836$	$P = 0.270$
		$s_{slope} = 0.141$	$s_{slope} = 0.087$	
	Phase 2	Slope $= 4.068$	Slope $= 4.019$	$F = 0.06$
		$R^2 = 0.835$	$R^2 = 0.830$	$P = 0.801$
		$s_{slope} = 0.149$	$s_{slope} = 0.110$	

2, i.e. SSS and SRS, showed an increased level of concordance, up to 0.87 and 0.86 (excellent).

Fleiss' kappa values for categorical variables are summarized in Figure 4 and Table 3, along with the significance of concordance. In this case, reports from phase 1 and 2 are compared for all variables. For those variables, categories of agreement for expert readers between the two phases remained almost unchanged, with the exception of TID, while core participants showed an increase for all variables.

Relationship between SSS values as reported from both experts and core participants and the number of ''hypoperfusion clusters'', as derived from polar maps, is summarized in Figure 5. In more detail, Figures 5A and B represent results from experts in phases 1 and phase 2, respectively; while in Figures 5C and D the same analysis is reported for Core participants.

If we consider SSS mean values as a function of cluster number and then we determine a linear interpolation between the experimental data, we observe a tendency towards statistical significance (F=3.64 and p=0.057) for curve slopes only between phase 1 and phase 2 for core participants (Table 4).

As already described, based on SDS values, patients have been stratified (SDS-strat) as ''low risk'' (SDS ≤3); ''intermediate risk'' (4 ≤ SDS ≤ 7) and ''high risk'' (SDS ≥ 8), according to their SDS value. Analysing differences in risk stratification as described by SDS values between phases, we found that there is a significant difference for SDS strat between phases 1 and 2, for 3 studies out of 15 in the core participants group, and in 2 of 15 in experts.

As already described, participants were encouraged to analyze and report submitted studies according to their daily routine, including use of their cardiac software. Well aware of the possible impact on calculated values such as LVEF and EDV, information was also collected on type of cardiac software utilized. For both Core participants and Experts, the distribution of the different cardiac software available on the market is reported in Appendix (Table 6).

Evaluations on LVEF values included the following factors: Group (2 levels: Experts, Core Participants); cardiac software (5 levels: 4DMCardio; CedarsSinai; EmoryCardiacToolBox; InterView; Other); Case Study number (15 levels: patient studies 1-15). Changes in variables were assessed as a function of factors and interaction between factors themselves. Results are shown in Table 7 of the Appendix. There were significant differences in the LVEF values calculated both post-stress and at rest and for values calculated from the different types of software. The Bonferroni post-hoc analysis of multiple comparisons shows that one of the software packages (EmoryCardiacToolBox) systematically produces an LVEF value significantly lower than 4DMCardio, CedarsSinai, and Other software (range of differences: − 8.2% to − 10.8%); while no significant differences are found with the InterView software (see Table 8 in Appendix for details).

Overall, LVEF post-stress values are not significantly different between core participants and experts (Table 9 in Appendix). Average SD levels for the readings of core participants were about twice as high as the average SD levels for the experts group (10.4% vs
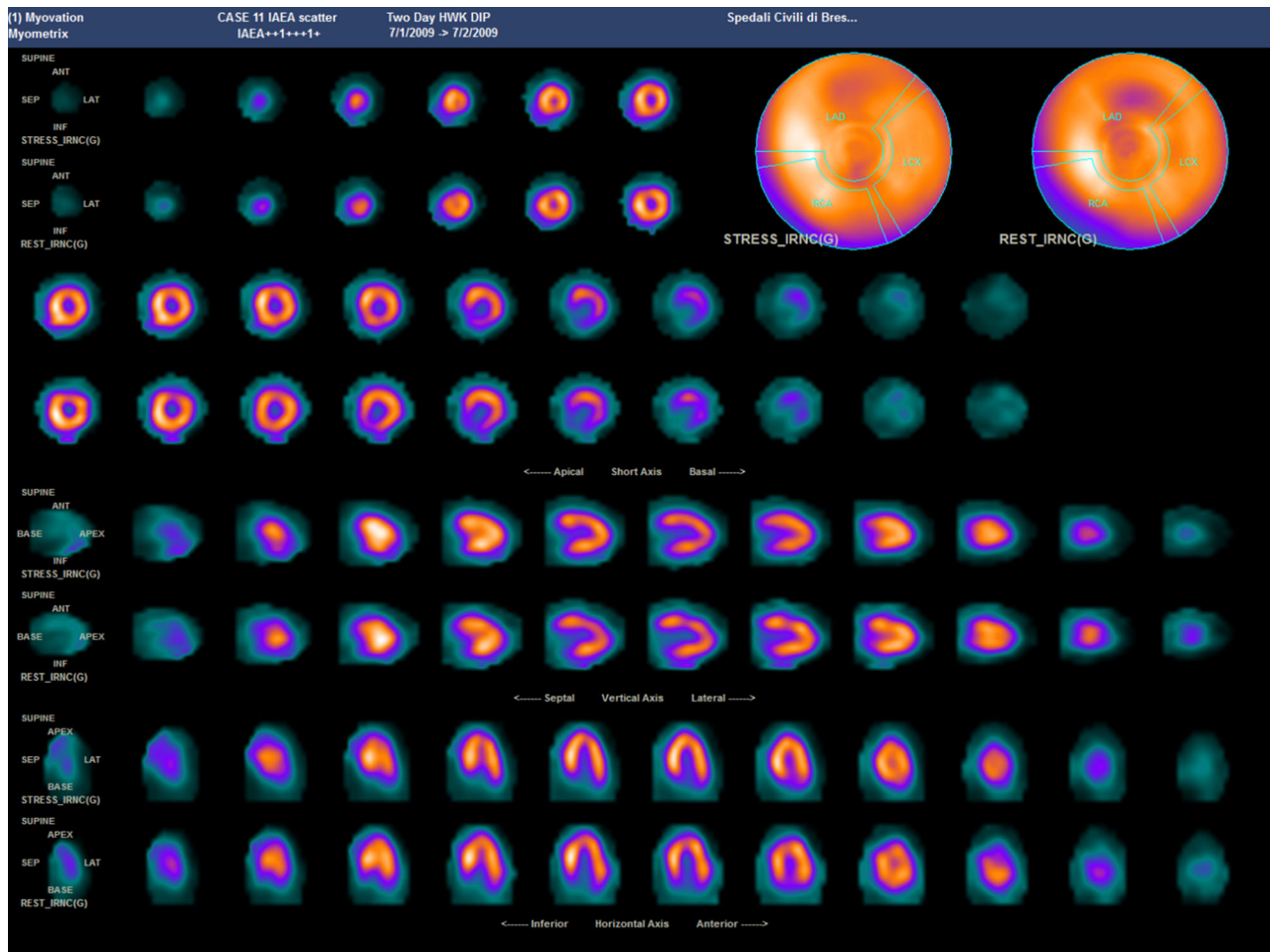
Figure 6. Case #11 of phase 1: Female, 71 y-o; Type 2 diabetes on medication with metformin; Hypertension treated by vasodilators (Enalapril); 5 yrs before MPI cardioversion for atrial fibrillation. On chronic therapy with warfarin and propafenone; H 162 cm W 79 kg; Referred for chest pain not related to efforts; Dipyridamol stress test (0.84 mg/kg/5 minutes); Rest BP 170/95, 67 BPM; At the end of Dipyridamol infusion BP = 150/100, 95 BPM.

5.8%), a finding which was also expressed in the higher ICC for the latter group (Figure 3). Case 11 that caused relatively larger SD values in both core participants and experts readers groups (18.5 and 19.4, respectively) is represented in Figure 6.

DISCUSSION

Often, in medical imaging, interpretation of results is subjective[21-30] and can be influenced by technical considerations. Quality plays a pivotal role when analysing and reporting an imaging study. Several factors can affect the results of the analysis and the value of the studies. This is true for all modalities and in the case of SPECT MPI,[31-35] which is the subject of this study, it is crucial to ensure that the acquisition and reconstruction parameters are consistent and optimized, thus allowing accurate and reproducible results.

Several factors, in different phases of the procedure, might influence the final results of MPI studies and require scrutiny. They include, but are not limited to, pre-examination checks, such as appropriateness of reference, QA/QC of equipment and radiopharmaceutical preparation, to steps to be taken during examination, such as QA/QC of acquisition parameters and of processing and reporting. We geared the I-MAP study towards assessing the quality of processing and reporting.

We examined the reliability of SPECT MPI studies using inter-observer variability within two groups of participants: one made of practitioners from LMICs, which are indeed the target of IAEA's educational activities, and a second group of expert readers. The first group of ''core participants'' was composed of nuclear cardiology professionals who attended training events managed by the IAEA, many of them working in

settings where financial resources might be limited, therefore with limited experience and limited resources for improving their expertise. As regards the study, it was run in two phases and in both of them participants had to report the same group of 15 cases, with the important difference that in phase 1 all participants were provided raw data and were requested to process them according to their routine practice and then report. In phase 2, all participants were given, in different order, the same 15 cases pre-reconstructed and were requested to provide their segmental uptake score, visually assessed, as well as other qualitative interpretations. Both groups were unaware that in phase 2 they were re-evaluating the same studies.

For quantitative data such as EDVs and LVEFs, an excellent level of concordance was found within both groups for both phase 1 and 2 (Table 2; Figure 3). Concordance was also excellent within the experts group for SSS and SRS values in both phases.

It's very Interesting that, for the latter two parameters (SSS and SRS), core readers showed an excellent intra-group agreement in phase 2 when they had to provide their own evaluation on pre-processed images (0.87 and 0.86; for SSS and SRS respectively), while in phase 1, when they had to process the studies and scores were automatically calculated by their software, concordance was only good, being 0.66 and 0.64; for SSS and SRS respectively).

It should be remembered that while in phase 1 readers could accept segmental scores from their own software, or override if needed, in phase 2 scores had to be visually assessed and manually entered into the forms, therefore reflecting a qualitative rather than a semi-quantitative evaluation. Therefore, we relate this improvement to the central role of processing: when less experienced readers are presented with well processed studies and are forced to score perfusion status, their readings are as good as experts' readings. This finding confirms that processing remains a crucial step for the overall SPECT MPI evaluation and that experience and training plays a major role for good quality processing. Furthermore, this finding tells us that, besides physicians who actually are those who read studies, IAEA training events should also involve technologists who often perform the processing.

Further confirmation of the importance of processing is found when we compare performances between the two groups for risk stratification. In this case, when we analysed differences between the experts panel and the core participants group, we have found that in phase 1 a significant difference could be seen in 2/15 cases, while no difference could be seen between the two groups for phase 2, when the core lab distributed pre-processed studies.

Fleiss' kappa value is a rather stringent index, very sensitive to even small deviations between readers

which may cause an important worsening of calculated values. In this study, it showed that experts, as expected, had a greater concordance in interpretation, in both phases of the study, while for core participants concordance improved significantly between phase 1 and 2. This finding holds true for both the analysis of continuous variables and for SSS and SRS indexes. Once more, this finding supports the notion that interpretation in itself is not the issue, but what is going to be interpreted is. When study processing is not properly carried out, then interpretation suffers.

A tendency of core participants to give an overall evaluation of "normal perfusion" even in presence of significant SSS values and hypoperfusion clusters was observed (Figure 5).

The greater variability in interpreting on-site processed images, as requested in phase 1, might well be affected by poor alignment of slices because of bad selection of left ventricular axes, valve planes and apex. So, while experts were able to minimize the impact of processing on the quality of images, this was not the case for core participants, who indeed markedly improved their performance when they were given studies which had been pre-processed at the core lab. Pre-processing included motion correction, careful slice realignment between stress and rest acquisitions, correct choice of slice thickness to avoid artefacts due to partial volume effect, and correct colour scale levelling in presence of extracardiac hot-spots such as sub-diaphragmatic activity.

Finally, we found, as reported by other groups[36,37] that important parameters such as LVEF, calculated through gated SPECT, may differ significantly when different processing software packages are used, as shown in Table 8. One software deviates substantially and significantly from almost all the other software packages, with a systematic bias in LVEF of -8.3% down to -10.8% which could be clinically significant when LVEF is used in clinical decision making, such as in longitudinal studies of cardio-oncological patients.

The univariate analysis of variance for LVEF post-stress and LVEF at rest was run considering the different factors involved and their interactions. Results of that analysis reported in Table 7 also show significant differences for LVEF values calculated both post-stress and at rest, and for values calculated from the different types of software.

Overall, LVEF values are not significantly different between the two groups, core participants and experts, as shown in Table 9. A relatively wide SD shown for case #11 could be attributed to factors such as patient movement during acquisition (which could have been corrected for by readers), small heart with partial volume effect, hypertrophic left ventricular walls due to hypertension, and attenuation due to obesity (Figure 6).

NEW KNOWLEDGE GAINED

This study has shown that the quality of processing remains a crucial step for SPECT MPI and that experience helps overcome possible artefacts that may hamper the quality of reporting. As concerns the IAEA, this study shows that the outcomes of training events in NC are satisfactory, as the performance of NC professionals from LMICs does not differ significantly from expert readers in many circumstances, and particularly when good quality processing was applied to clinical studies. This latter consideration supports the concept that training courses should necessarily cover basic issues such as study processing. In addition, this study shows that LVEF values may differ significantly depending on the cardiac package employed and this should be kept in mind particularly when patients are studied in different institutions or when an institution adopts a different software package.

LIMITATIONS OF THE STUDY

The small sample size of 24 participants from LMICs is a very low response rate for survey data, challenging the generalizability of findings. Furthermore, we don't know to what extent ''core participants'' are representative of the reading pattern in LMICs. This is, however, unavoidable when dealing with centres from developing world because of difficult communication as well as technical problems affecting data transfers and report transmission, which may affect active participation.

One more important limit of the study design is the choice of not requiring participants to provide images along with reporting forms. This choice was made to minimize image transmission problems, but prevented full quality checks from being performed for the processed studies.

CONCLUSIONS

The quality of reporting SPECT MPI could be rated as moderate-to-good for participants from emerging economies and good-to-excellent for expert readers. It is clearly affected by the quality of processing. Indeed, when readers with less experience are asked to report on studies pre-processed at an experienced core lab and by professionals well-trained to avoid sources of artefacts, inter-observer agreement between readers with less experience improves substantially. To our knowledge, this is the first study reporting these findings.

Significant differences were found between LVEF values obtained using different software packages for cardiac analysis. This should be kept in mind particularly when patients are studied in different institutions or when an institution adopts a different software.

This study calls for attention from scientific societies on the issue of the quality of study processing, suggesting the need for more stringent guidelines about this aspect of NC practice.

Finally, these results suggest that the outcomes of training events conducted by the IAEA in NC are satisfactory. However, in order to improve the quality of processing, future training courses should necessarily cover this issue, and should also involve technologists.

Table 5. Summary of training events managed by NMDI

Region	No. of training courses	No. of participants
Africa	8	155
Asia	10	208
Middle East	3	55
Eastern Europe	13	285
Latin America and Caribbean	7	193
Total	41	896

Table 6. Distribution of cardiac SW among both groups (core participants and experts)

Cardiac Software	Core participants (%)	Experts (%)
4DMCardio	28.4	9.4
CedarsSinai	43.0	90.6
EmoryCardiacToolBox	15.8	–
Interview	10.6	–
Other	2.3	–

Table 7. ANOVA univariate analysis of LVEF post stress and at rest for factors SW; group; case study number and their interactions

	LVEF post stress		LVEF rest	
Source	F	Sig.	F	Sig.
Software	12.041	**< 0.001**	11.772	**< 0.001**
Group	1.204	0.273	1.075	0.300
Case Study Number	28.384	**< 0.001**	23.657	**< 0.001**
Software × Group	0.048	0.826	0.500	0.480
Software × Case Study Number	1.293	0.096	0.719	0.922
Group × Case Study Number	0.316	0.992	0.334	0.989
Software × Group × Case Study Number	0.364	0.980	0.288	0.993

Highly significant P values are given in bold

Table 8. Bonferroni multiple comparisons post hoc test for the different cardiac SW utilized by study participants

Factor		LVEF post stress (%)			LVEF rest (%)		
Software (I)	Software (J)	Mean Difference (I–J)	SE	P value	Mean difference (I–J)	SE	P value
4DMCardio	CedarsSinai	− 0.017	1.059	1.000	0.350	0.972	1.000
	EmoryCardiacToolBox	− 8.266	1.560	**< 0.001**	− 8.244	1.555	**< 0.001**
	Interview	− 3.540	1.793	0.489	− 3.497	1.646	0.342
	Other	2.532	3.474	1.000	1.858	3.189	1.000
CedarsSinai	4DMCardio	0.017	1.059	1.000	− 0.350	0.972	1.000
	EmoryCardiacToolBox	− 8.248	1.399	**< 0.001**	− 8.594	1.420	**< 0.001**
	Interview	− 3.523	1.654	0.338	− 3.847	1.519	0.117
	Other	2.549	3.405	1.000	1.508	3.126	1.000
EmoryCardiac ToolBox	4DMCardio	8.266	1.560	**< 0.001**	8.244	1.555	**< 0.001**
	CedarsSinai	8.248	1.399	**< 0.001**	8.594	1.420	**< 0.001**
	Interview	4.725	2.013	0.193	4.747	1.944	0.150
	Other	10.798	3.593	**< 0.001**	10.102	3.353	**< 0.001**

Table 8 continued

Factor		LVEF post stress (%)			LVEF rest (%)		
Software (I)	Software (J)	Mean Difference (I–J)	SE	P value	Mean difference (I–J)	SE	P value
Interview	4DMCardio	3.540	1.793	0.489	3.497	1.646	0.342
	CedarsSinai	3.523	1.654	0.338	3.847	1.519	0.117
	EmoryCardiacToolBox	− 4.725	2.013	0.193	− 4.747	1.944	0.150
	Other	6.072	3.700	1.000	5.355	3.396	1.000
Other	4DMCardio	− 2.532	3.474	1.000	− 1.858	3.189	1.000
	CedarsSinai	− 2.549	3.405	1.000	− 1.508	3.126	1.000
	EmoryCardiacToolBox	− 10.798	3.593	**< 0.001**	− 10.102	3.353	**< 0.001**
	Interview	− 6.072	3.700	1.000	− 5.355	3.396	1.000

Highly significant *P* values are given in bold

Table 9. Comparison of results for LVEFs (phase 1 only) between experts and non-experts

	LVEF post-stress				
	Expert				
	No		Yes		
Case study #	Mean	SD	Mean	SD	*P* value
1	54.8	8.1	54.0	4.6	0.761
2	50.0	9.5	44.3	6.2	0.078
3	76.5	11.3	76.1	4	0.908
4	56.8	8.8	54.2	1.9	0.335
5	49.8	5.1	47.8	2.8	0.226
6	72.2	10.3	74.2	6.4	0.551
7	28.6	8.1	26.5	4.2	0.441
8	47.0	9.9	45.5	4.9	0.657
9	44.6	9.3	43.8	3.8	0.792
10	69.3	15.4	70.7	2.9	0.767
11	79.0	18.5	79.5	19.4	0.952
12	67.8	5.7	67.5	4.5	0.887
13	41.1	12.4	41.6	7.8	0.902
14	80.1	13.6	83.7	8.4	0.446
15	47.6	6.6	44.2	2.9	0.136

APPENDIX

See Tables 5, 6, 7, 8 and 9.

References

1. https://www.iaea.org/. Last accessed 22 May 2018.
2. https://www.iaea.org/about/organizational-structure/department-of-nuclear-sciences-and-applications/division-of-human-health. Last accessed 22 May 2018.
3. https://www.iaea.org/topics/nuclear-medicine-and-diagnostic-imaging-section. Last accessed 22 May 2018.
4. https://sustainabledevelopment.un.org/sdg3. Last accessed on 08 June 2018.
5. http://www.who.int/nmh/en/. Last accessed 08 June 2018.

6. Pradeepa R, Prabhakaran D, Mohan V. Emerging economies and diabetes and cardiovascular disease. Diabetes Technol Ther. 2012;14(Suppl 1):S59-67. https://doi.org/10.1089/dia.2012.00.

7. Jaarsma C, Leiner T, Bekkers SC, Crijns HJ, Wildberger JE, Nagel E, Nelemans PJ. Schalla S Diagnostic performance of noninvasive myocardial perfusion imaging using single-photon emission computed tomography, cardiac magnetic resonance, and positron emission tomography imaging for the detection of obstructive coronary artery disease: A meta-analysis. J Am Coll Cardiol. 2012;59(19):1719-28.

8. Metz LD, Beattie M, Hom R, Redberg RF, Grady D. Fleischmann KE The prognostic value of normal exercise myocardial perfusion imaging and exercise echocardiography: A meta-analysis. J Am Coll Cardiol. 2007;49(2):227-37.

9. Hachamovitch R, Hayes SW, Friedman JD, Cohen I, Berman DS. Comparison of the short-term survival benefit associated with revascularization compared with medical therapy in patients with no prior coronary artery disease undergoing stress myocardial perfusion single photon emission computed tomography. Circulation. 2003;107(23):2900-7.

10. Dondi M, Andreo P. Developing nuclear medicine in developing countries: IAEA's possible mission. Eur J Nucl Med Mol Imaging. 2006;33:514-5.

11. Casas-Zamora JA, Kashyap R. The IAEA technical cooperation programme and nuclear medicine in the developing world: Objectives, trends, and contributions. Semin Nucl Med. 2013;43(3):172-80. https://doi.org/10.1053/j.semnuclmed.2012.11.007.

12. https://humanhealth.iaea.org/HHW/NuclearMedicine/NUMDAB/index.html. Last accessed 22 May 2018.

13. 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. Circulation. 2002;105:539-42.

14. IAEA Human Health Series No. 23 (Rev. 1) Nuclear Cardiology: Guidance on the Implementation of SPECT Myocardial Perfusion Imaging International Atomic Energy Agency, Vienna; 2016.

15. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34(5):502-8. https://doi.org/10.1111/opo.12131.

16. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials. 1991;12(Suppl 4):142S-58S.

17. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutor Quant Methods Psychol. 2012;8(1):23-34.

18. Fleiss JL, Levin B, Paik MC. The measurement of interrater agreement. In: Fleiss JL, Levin B, Paik MC, editors. Statistical methods for rates and proportions. 3rd ed. New York: Wiley; 2003. p. 598-626.

19. Cicchetti DV. Guidelines, criteria and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284-90.

20. Hartling L, Hamm M, Milne A, Vandermeer B, Lina Santaguida P, Ansari M, et al. Validity and inter-rater reliability testing of quality assessment instruments. Rockville, MD: Agency for Healthcare Research and Quality (US); 2012. Table B, Interpretation of Fleiss' kappa (κ) (from Landis and Koch 1977).

21. Papadopoulou SL, Garcia-Garcia HM, Rossi A, Girasis C, Dharampal AS, et al. Reproducibility of computed tomography angiography data analysis using semiautomated plaque quantification software: Implications for the design of longitudinal studies. Int J Cardiovasc Imaging. 2013;29:1095-104. https://doi.org/10.1007/s10554-012-0167-5.

22. Herzog C, Kerl JM, De Rosa S, Tekin T, Boehme E, Liem S, et al. Influence of observer experience and training on proficiency in coronary CT angiography interpretation. Eur J Radiol. 2013;82:1240-7.

23. Taylor AJ, Patrick J, Abbara S, Berman DS, Halliburton SS, Hines JL, et al. Relationship between previous training and experience and results of the certification examination in cardiovascular computed tomography. JACC Cardiovasc Imaging. 2010;9:976-80.

24. Chauvela C, Abergel E, Renault L, Chatellier G, Cohen I, Attane C, et al. Improving stress echocardiography accuracy for detecting left circumflex artery stenosis: A new echocardiographic sign? Arch Cardiovasc Dis. 2012;105:196-202.

25. Kataoka A, Scherrer-Crosbie M, Senior R, Gosselin G, Phaneuf D, Guzman G, et al. The value of core lab stress echocardiography interpretations: Observations from the ISCHEMIA Trial. Cardiovasc Ultrasound. 2015;13:47.

26. Knight DS, Schwaiger JP, Krupickova S, Davar J, Muthurangu V, Coghlan JG, et al. Accuracy and test-retest reproducibility of two-dimensional knowledge-based volumetric reconstruction of the right ventricle in pulmonary hypertension. J Am Soc Echocardiogr. 2015;28:989-98.

27. Berman DS, Kang X, Gransar H, Gerlach J, Friedman JD, Hayes SW, et al. Quantitative assessment of myocardial perfusion abnormality on SPECT myocardial perfusion imaging is more reproducible than expert visual analysis. J Nucl Cardiol. 2009;16:45.

28. Nakajima K, Higuchi T, Taki J, Kawano M, Tonami N, et al. Accuracy of ventricular volume and ejection fraction measured by gated myocardial SPECT: Comparison of 4 software programs. J Nucl Med. 2001;42:1571-8.

29. Larghat AM, Maredia N, Biglands J, Greenwood JP, Ball SG, Jerosch-Herold M, et al. Reproducibility of first-pass cardiovascular magnetic resonance myocardial perfusion. J Magn Reson Imaging. 2013;37:865-74.

30. Meriki N, Izurieta A, Welsh A. Reproducibility of constituent time intervals of right and left fetal modified myocardial performance indices on pulsed Doppler echocardiography: A short report. Ultrasound Obstet Gynecol. 2012;39:654-8.

31. Wackers FJ. Artifacts in planar and SPECT myocardial perfusion imaging. Am J Card Imaging. 1992;6:42-57.

32. Germano G, Kavanagh PB, Waechter P, Areeda J, Van Kriekinge S, Sharir T, et al. A new algorithm for the quantitation of myocardial perfusion SPECT. I: Technical principles and reproducibility. J Nucl Med. 2000;41:712-9.

33. Ljungberg M, Pretorius PH. SPECT/CT: An update on technological developments and clinical applications. Br J Radiol. 2018;91(1081):20160402. https://doi.org/10.1259/bjr.20160402. Epub 2017 Jan 16.

34. Malek H, Yaghoobi N, Hedayati R. Artifacts in quantitative analysis of myocardial perfusion SPECT, using Cedars-Sinai QPS Software. J Nucl Cardiol. 2017;24(2):534-42. https://doi.org/10.1007/s12350-016-0726-6 **Epub 2016 Nov 10**.

35. Chrysanthou-Baustert I, Polycarpou I, Demetriadou O, Livieratos L, Lontos A, Antoniou A, et al. Characterization of attenuation and respiratory motion artifacts and their influence on SPECT MP image evaluation using a dynamic phantom assembly with variable cardiac defects. J Nucl Cardiol. 2017;24:698-707. https://doi.org/10.1007/s12350-015-0378-y **Epub 2016 Feb 4**.

36. Foley TA, Mankad SV, Anavekar NS, Bonnichsen CR, Morris MF, Miller TD, et al. Measuring left ventricular ejection fraction—Techniques and potential pitfalls. Eur Cardiol. 2012;8(2):108-14.

37. Steyn R, Boniaszczuk J, Geldenhuys T. Comparison of estimates of left ventricular ejection fraction obtained from gated blood pool imaging, different software packages and cameras. Cardiovasc J Afr. 2014;25:44-9.