Diagnostic performance of an artificial intelligence-driven cardiac-structured reporting system for myocardial perfusion SPECT imaging

Ernest V. Garcia, PhD,[a] J. Larry Klein, MD,[b] Valeria Moncayo, MD,[a] C. David Cooke, MSEE,[a,c] Christian Del'Aune, BSEE,[c] Russell Folks, CNMT,[a] Liudmila Verdes Moreiras, MSN,[a] and Fabio Esteves, MD[a]

[a] Department of Radiology and Imaging Sciences, Emory University, Atlanta, GA
[b] Division of Cardiology, Department of Medicine, UNC School of Medicine, University of North Carolina, Chapel Hill, NC
[c] Syntermed, Inc., Atlanta, GA

Received Jun 5, 2018; accepted Aug 27, 2018
doi:10.1007/s12350-018-1432-3

Objectives. **To describe and validate an artificial intelligence (AI)-driven structured reporting system by direct comparison of automatically generated reports to results from actual clinical reports generated by nuclear cardiology experts.**

Background. **Quantitative parameters extracted from myocardial perfusion imaging (MPI) studies are used by our AI reporting system to generate automatically a guideline-compliant structured report (sR).**

Method. **A new nonparametric approach generates distribution functions of rest and stress, perfusion, and thickening, for each of 17 left ventricle segments that are then transformed to certainty factors (CFs) that a segment is hypoperfused, ischemic. These CFs are then input to our set of heuristic rules used to reach diagnostic findings and impressions propagated into a sR referred as an AI-driven structured report (AIsR).**

The diagnostic accuracy of the AIsR for detecting coronary artery disease (CAD) and ischemia was tested in 1,000 patients who had undergone rest/stress SPECT MPI.

Results. **At the high-specificity (SP) level, in a subset of 100 patients, there were no statistical differences in the agreements between the AIsr, and nine experts' impressions of CAD ($P = .33$) or ischemia ($P = .37$). This high-SP level also yielded the highest accuracy across global and regional results in the 1,000 patients. These accuracies were statistically significantly better than the other two levels [sensitivity (SN)/SP tradeoff, high SN] across all comparisons.**

Conclusions. **This AI reporting system automatically generates a structured natural language report with a diagnostic performance comparable to those of experts. (J Nucl Cardiol 2020;27:1652–64.)**

Key Words: Expert systems • artificial intelligence • myocardial perfusion SPECT • quantitative analysis, structured reporting

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12350-018-1432-3) contains supplementary material, which is available to authorized users.

The authors of this article have provided a PowerPoint file, available for download at SpringerLink, which summarises the contents of the paper and is free for re-use at meetings and presentations. Search for the article DOI on SpringerLink.com.

Reprint requests: Ernest V. Garcia, PhD, Department of Radiology and Imaging Sciences, Emory University, 101 Woodruff Circle, Room 1203, Atlanta, GA 30322; *ernest.garcia@emory.edu*

1071-3581/$34.00

Abbreviations

AI	Artificial intelligence
AIsR	AI-driven structured report
CAD	Coronary artery disease
CF	Certainty factor
CI	Confidence interval
ECTb	Emory Cardiac Toolbox
LAD	Left anterior descending coronary artery
LCX	Left circumflex coronary artery
LLK	Low likelihood
LV	Left ventricle
MPI	Myocardial perfusion imaging
NC	Nuclear cardiology
RCA	Right coronary artery
TID	Trans-ischemic dilatation
SN	Sensitivity
SP	Specificity
sRs	Structured report
SSS	Sum stress score

INTRODUCTION

Artificial intelligence (AI) methods to aid diagnosticians in making clinical image interpretation of SPECT myocardial perfusion studies have been reported. Examples include neural networks,[1–4] case-based reasoning,[5] support vector machines,[6] machine-learning,[7] and knowledge-based expert systems.[8,9] In expert systems, a knowledge base of heuristic rules is obtained from human experts capturing how they make their interpretations. Yet, to date, no one has developed automatically generated and/or validated natural language structured reports (sRs) that follow society guidelines. The convergence of the high prevalence of heart disease, increased complexity of cardiac imaging techniques, the increasing amount of patient-specific clinical information, and the reduced time the diagnostician has to dedicate to each patient inevitably lead to misdiagnosis and potential patient mismanagement. Hence, AI tools could assist physicians in interpreting and reporting studies at a faster rate and at the highest level of up-to-date expertise.

Here we report on the development and validation of an expert system in 1,000 patients, which applies its knowledge to extracted patients' left ventricle (LV) perfusion and function information from myocardial perfusion imaging (MPI) imagery to propagate this AI-driven structured[9] report (AIsR) following society guidelines.[10] Although physicians can easily modify any aspect of the AIsR, here we only evaluate the automatically generated results.

METHODS

Study Design

This is a single-center retrospective study designed to compare the diagnostic agreement between an automatically generated AIsR and the clinical rest/stress MPI report dictated by human experts. One of the nine nuclear cardiology (NC) experts dictated these clinical reports. The primary hypothesis was to demonstrate that the per-patient and per-vessel diagnostic performance of the AIsR in reporting hypoperfusion [coronary artery disease (CAD)] and reversibility (ischemia) is comparable (i.e., not inferior) to that of human experts' clinical reports. Agreement between the AIsR and the clinical report was compared in a 100-patient cohort to the agreement between the same MPI studies interpreted and reported a second time by another independent—10th human expert (VM) who started at Emory after the last MPI study in the trial was acquired (2010) and thus was never privy to their clinical reports. The second goal was to apply the same methodology to the entire 1,000 study group to determine agreement rates between AIsR and experts.

Study Population

One thousand consecutive MPI conventional studies used for this evaluation were obtained from our cardiac database of patients (589 men) referred to Emory University Hospital for clinically indicated attenuation-corrected (AC) rest/stress myocardial perfusion SPECT imaging between May 2008 and March 2010. Note that none of these 1,000 patients was used for the development of the method. Patients imaged with a CZT SPECT camera and/or lower doses during this period were excluded due to differences in technology and changing protocols. Emory's Institutional Review Board approved this research.

Clinical Data

Age, gender, body mass index, and risk factors data were extracted from the patients' medical records in Emory's data warehouse (Table 1). Risk factors mined were hypertension, hyperlipidemia, diabetes mellitus, smoking history, prior myocardial infarction, and prior revascularization. Representative quantitative MPI parameters were also extracted (Table 1) to characterize the population.

Standard Dual-Detector SPECT

All patients underwent eight-frame ECG-gated 1-day AC low-dose rest, high-dose stress Tc-99m tetrofosmin myocardial perfusion dual-detector SPECT according to the ASNC guidelines.[11] Rest-stress doses were determined based on patient's

body weight starting at < 200 lbs [370 MBq rest (10 mCi), 1,110 MBq stress (30 mCi)]. Acquisition times were 14 minutes for rest imaging and 12 minutes for stress imaging. Conventional SPECT projections were obtained utilizing the simultaneous emission/transmission acquisition method that uses a scanning gadolinium-153 line source as the transmission source. The emission transaxial images were reconstructed with an OSEM algorithm with 4 subsets and 10 iterations and a uniform initial estimate. The scatter distribution obtained from the scatter window was used to correct both the scatter from the patient onto the photopeak window and the scatter from the patient onto the transmission energy window. Attenuation maps were reconstructed by means of a Bayesian algorithm with Butterworth filter preprocessing at 0.43 critical frequency and an order of 5.0. The attenuation map reconstruction used 30 iterations with a uniform initial estimate.

MPI Reporting as Reference Standard

In each patient, the detection of hypoperfusion at stress and the presence of reversibility at rest for each major vascular territory reported by AIsR were compared to those from clinical reports generated by one of nine possible NC experts, each with at least 5 years of experience. The clinical interpretations reported were used as the reference standard. The image interpretations for the clinical reports were performed in the routine conventional way. The diagnosticians had full use of Emory Cardiac Toolbox (ECTb) V3.0 images and quantitative results[12] as well as all the usual clinical information requested by the interpreter. Neither the nine interpreters had access to the AIsR results from ECTb V4 developed after 2010, nor did any of these nine participate in developing any of the heuristic rules in the program's knowledge base.

Thus, because of the differences in the approaches, the sum stress score (SSS), and the SDS global and regional values between V3 and V4 could be quite different. Disease was assigned to one or more vascular territory combinations: left anterior descending artery (LAD), left circumflex artery (LCX), and the right coronary artery (RCA).

Interobserver Variability Subgroup

A subgroup of the last 100 consecutive patients was extracted from the 1,000-patients to determine the interobserver variability between experts. A tenth NC expert (VM) recruited to our institution, after the last patient in the study was acquired, performed as an independent reader to determine how the diagnostic variability between human experts reports compared to the variability between experts and the AIsR.

Image Analysis and AIsR Interpretation and Reporting

All MPI studies were reconstructed and reoriented into oblique-axis tomograms using conventional techniques according to ASNC guidelines.[11] The studies were then submitted by a technologist to a well-established automatic method of extracting 3D rest, stress distributions of myocardial perfusion, and function.[12] The technologist reviewed the processing and manually modified the automatically determined parameters if deemed incorrect, which was done less than 10% of the times and usually at the LV base.

Table 1. Characteristics of the study population.

Sample size	1000
Age (years)	61 ± 13
Male gender	59% (586)
Body mass index (kg·m^2)	29.2 ± 6.0
Hypertension	74% (741)
Hyperlipidemia	87% (867)
Diabetes mellitus	42% (415)
Smoking history	8.7% (87)
Prior myocardial infarction	11% (105)
Prior revascularization	30% (304)
Prevalence of CAD*	34.7%
Prevalence of ischemia*	12.0%
SSS^	2.24 ± 4.57
SDS^	1.11 ± 2.64
TID^	1.01 ± .13
Stress LVEF^	64 ± 13%
Rest LVEF^	63 ± 13%

*From clinical MPI reports
^From ECTb4

These 3D distributions were then submitted to our iterative method of database quantification implemented in ECTb V4.0. This iterative approach determines the 0 to 4 score for each of the conventional 17 segments using three iterations through the rest and stress AC, and non-AC perfusion, and non-AC function distributions. The iterative steps were as follows: (1) determining the certainty that a segment is abnormal, (2) assigning the score to each of the 17 segments, and (3) using our expert system to modify the score consistent with all the information available for that segment which we call a smart score.

Step 1: determining certainty of segment abnormality. A certainty factor (CF) is determined ranging from − 1 to + 1 for each of the 17 LV segments (− 1 = definitely no count reduction (normal), + 1 = definitely count reduction, and the range from − 0.2 to + 0.2 means the presence of any finding that is equivocal or indeterminate). This CF determination of segment abnormality first calculates the % abnormal probability (P_s) for each segment[13] whether a patient's normalized perfusion distribution (relative blood flow) is lower than that of the normal distribution redeveloped from a previously reported group of normal low likelihood (LLK) patients.[14,15] Since the relative blood flow is extracted in terms of number of counts and these counts vary depending on the injected dose, patient size, LV size, and instrument sensitivity (SN), these count distributions for each voxel segment c_{vs} have to be normalized both by the maximal voxel count uptake (C_{max}) over the entire LV, and by the total

number of LV voxels in each segment (V_s). The normalized count density (n) for each voxel in segment s is given by

$$n_{vs} = [100c_{vs}]/[V_s C_{max}].$$

The value of a cumulative distribution function over all voxels in segment s is given by Γ^{pt}_{ns} as the sum of all normalized count densities for patient pt:

$$\Gamma^{pt}_{ns} = \sum_{v} n_{vs} : \ \Gamma^{pt}_{ns} = 0 \quad \text{for all } n_{vs} > n_{vs} \quad (\Gamma^{pt}_{ns} = 100).$$

Thus, for example, the value of Γ^{pt}_{ns} at 50% in segment 2 in Figure 1 is found by finding the 50 in the x-axis to reach the patients red distribution, the value that you read 55% from the y-axis is Γ^{pt}_{ns}—this represents percentage of the total number of voxels in segment 2 which are $\leq$ an n_{vs} of 50%. In Figure 1, the red distributions are the normalized cumulative count value stress distributions for each of the 17 segments of the patient shown in the polar map. Note that the patient's distribution (red) is set to zero after it reaches 100%. This was done to increase the $[[\Gamma^{pt}_{ns} - \Gamma^{nl}_{ns}]$ difference and thus the discriminatory power of P_s.

The white distributions Γ^{nl}_{ns} are the cumulative distribution functions from all normal patients used to create this specific nonparametric normal database. The probability P_s, is then determined for each of the 17 LV segments whether a patient's tracer distribution is lower than that of the normal distribution as

$$P_s = 100 \sum_{n} [\Gamma^{pt}_{ns} - \Gamma^{nl}_{ns}]/\Gamma^{pt}_{ns}.$$

Note that P_s is a function of n_{vs}. Also, note that to determine the probability P_s, we are summing over all available n's (i.e., all available samples of *normalized* count values) that is equivalent to summing all n's from 0% to 100%. These P_s are converted to CFs by a transformation from [0, 100] $\rightarrow$ [$-$ 1, 1] using Shannon's information theory.[16] In this information approach, CF is obtained by using a transformation function between percent (P_s) of a segment being abnormal and uncertainty $U = (1 - \text{CF})$ as

$$U = -\sum_{i} P_{si} \log_2 P_{si},$$

where i is the potential number of states: in this case 2, normal and abnormal. For example, in Figure 1, for segment 6, $P_6 = .89$ (or 89%), hence $U = -(.89 \log_2 .89 + .11 \log_2 .11) = .50$, and therefore CF is abnormal as $1 - .5 = .5$, consistent with this hypoperfused (abnormal) segment. For segment 8, on the other hand, the patient's distribution (red) is inside the normal distribution (white), and thus, the CF obtained is negative, which indicates that the segment is normally perfused. This allows CFs to range from $-$ 1 to + 1. CFs are calculated for each segment and for each quantitative parameter used as input to the AIsR. This is a nonparametric approach as no assumptions are made as to the properties of the normalized count distribution (usually incorrectly approximated as Gaussian).

Step 2: assigning a score to each of the segments. This step converts the CF value for each segment into a score (0 to 4, Figure 2). All segments with a normal CF ($< -.2$) are given a score of 0. The score for each abnormal (CF $> .2$) or equivocal ($-.2 <$ CF $< .2$) segment depends on two parameters: (1) the type of distribution (stress, rest perfusion; perfusion reversibility; AC vs non-AC, supine vs prone, stress, rest thickening, thickening reversibility) and (2) the magnitude of the parameter (% uptake for perfusion, % thickening for thickening). These CF settings were done at three different levels (modes) of SN/specificity (SP) settings: (1) high SP, where an equivocal CF in the AIsR was set to normal; (2) high SN, where an equivocal CF in the AIsR was set to abnormal; and (3) tradeoff SN/SP, where the lower half of the equivocal CF range ($-$ .2 to 0) was set to normal and the upper half (0 to .2) to abnormal.

A set of scores is determined for each segment in each distribution and then are merged into one set of results for stress perfusion, rest perfusion, reversibility perfusion, stress thickening, and rest thickening. The merger takes place such that the most normal score for each segment in each distribution is retained. For example, if the scores for segment 16 in the stress perfusion distribution is a 2 for non-AC, -and a 0 for AC (or prone) the combined score retained is a 0.

Step 3: determining smart-scores and AIsR generation. Here all sets of scores from step 2 are used as input to our expert system. This is a Bayesian inference engine forward chaining our MPI knowledge base of interpretation and reporting heuristic rules, similar to our previous reports[8,9] following the well-established expert system methods.[17] This expert system uses these input scores to determine the certainty of the location, size, shape, and reversibility of both the perfusion defects and thickening abnormalities to infer the certainty of the presence and vascular location of CAD. This information is then transmitted to the AIsR in natural language text. One main difference between our current expert system and our previous one[9] is that now all information for each segment is weighted to modify each segmental score during this iteration and the AIsR follows ASNC guidelines for reporting.[18] Thus, for example, a segment that exhibits a fixed perfusion defect in the non-AC distributions is more certain to be fixed if it is also fixed in the AC distributions and even more certain if the segment is thickening abnormally. Once all perfusion and function smart-scores (Figure 2A inset) and pertinent prespecified data elements [example LVEF, transischemic dilatation (TID), etc.] along with their CF values are determined, they are exported as a highly structured object which is then imported by the AIsR. These exported data elements are mapped onto the existing data entry fields within the AIsR. When the user begins generating the report, all of the mapped input entry fields are automatically prepopulated including the smart-scores data generated by our expert system.

All the natural language text is conditionally generated by the reporting module of the system. In brief, take, for example, the results in Figure 3 and the AIsR report in Figure 4A. Specifically consider the conclusion in both figures ''the apical lateral segment is completely reversible.'' Before reaching the report, the nonparametric statistics combined with the expert system portion of the AIsR has determined CFs for each possible state (categories). In this case of apical lateral

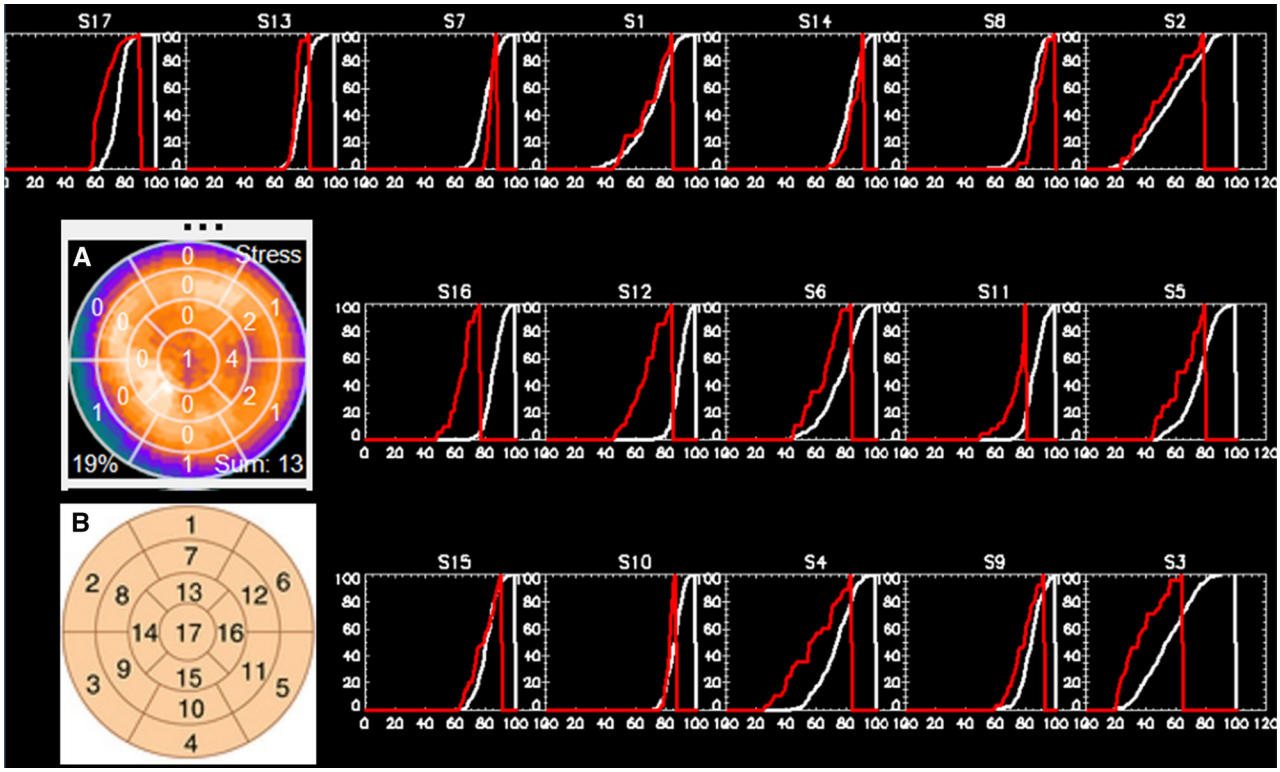

Figure 1. 17-segment results from a patient with LCX vessel disease. Color polar map inset (**A**) shows the myocardial perfusion distribution for a female patient with LCX vessel disease with the 17-segment model with scores superimposed. The 17 plots correspond to the 17-segment model (**B**) with the LAD segments on the top, LCX in the middle, and RCA in the bottom rows. The *x*-axes are the normalized count values, and the *y*-axes are the normalized voxel frequencies with those count values. The white distributions are the averaged normalized cumulative distributions from 20 female patients with low likelihood of CAD. The red distributions are the normalized cumulative count value distributions for the patient shown in the polar map. Note that red distributions to the left of the white normal ones represent increasing certainty of abnormality. Also note how well behaved is the shape of each of the patient's segmental distributions even though it represents a small portion of the LV from just one patient.

segmental reversibility, it has determined a CF that the segment is completely reversible, another CF that it is partially reversible, another CF that it is minimally reversible, and another CF that it is fixed. The natural language generator reads these states and chooses the one with the highest CF as the condition to report, in this case completely reversible.

Statistical Analysis

All studies were classified as normal (definitely normal or probably normal) or abnormal (definitely abnormal or probably abnormal) based on the report describing the presence of one or more stress perfusion defects. To test the primary hypothesis the methodology previously reported by us to test for noninferiority was used.[14] The difference between two population proportions from a single sample[19] was used to test if there were differences in reporting agreements between AIsR-expert to independent-expert. If AIsR findings are equivalent to expert findings, the expected difference between the AIsR findings agreement to independent-expert agreement is zero. The primary analysis tested the null hypothesis of equivalence of AIsR-expert agreement to independent-expert agreement (no agreement rate reduction) vs inferiority (a reduction of $> 0\%$). A 95% confidence interval (CI) for the difference between AIsR-expert agreement rates to independent-expert agreement rate was calculated and the null hypothesis rejected if the upper limit was below 0% with a corresponding one-tail *P* value less than .05. Interobserver agreement between AIsR findings and expert findings for all 1,000 MPI studies was measured using percent agreement (accuracy) and Cohen's κ value. McNemar's test was used to test the statistical differences in accuracy in the 1,000 MPI studies between each of the three SN/SP modes. To test whether there were differences between the MPI studies from the 1,000 patients and the 100-patient cohort as to the prevalence of CAD, ischemia, and AIsR agreement rate, the Medcalc χ^2 comparison of proportion

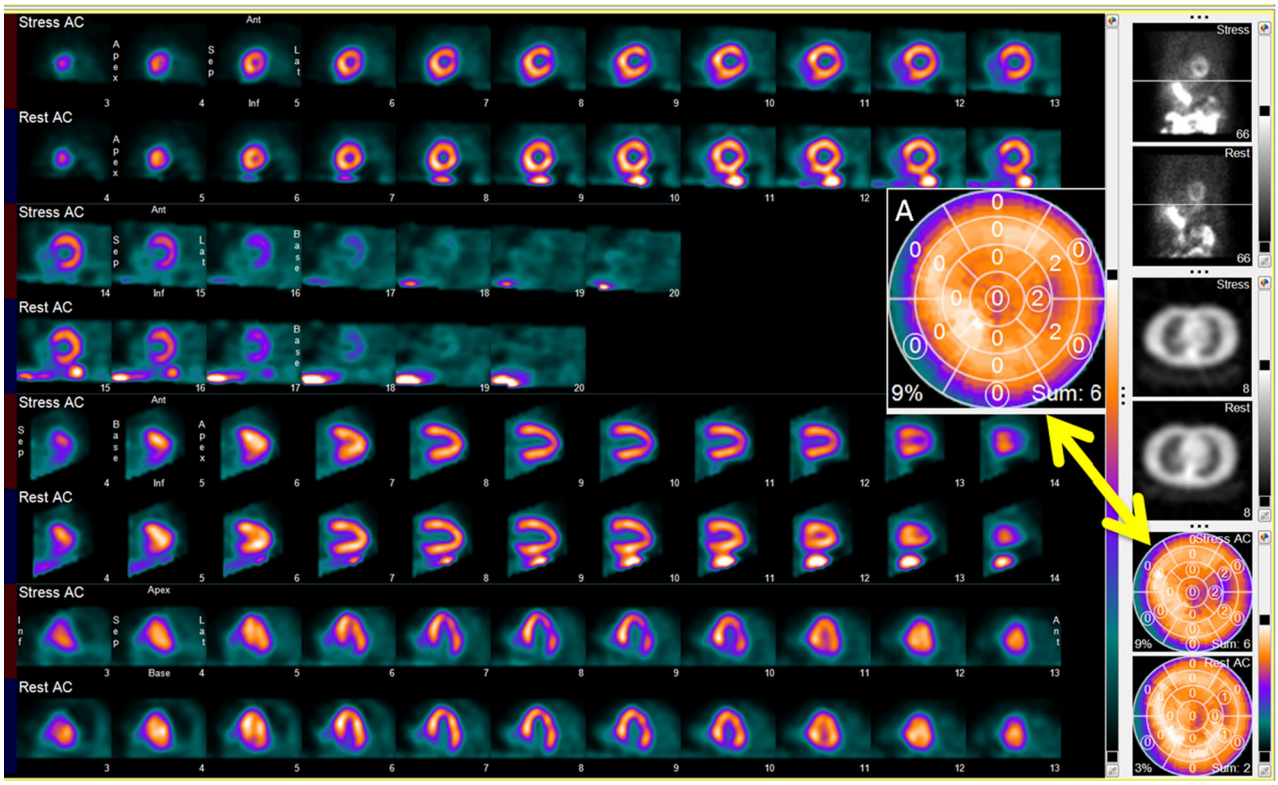

Figure 2. Combined slices/polar map displaying the patient with reversible lateral wall perfusion defect from Figure 1. Stress (top)/rest (bottom) SPECT attenuation-corrected slices, rotating projections, transmission slices, and 17-segment smart-scores. Note three contiguous segments in the lateral wall of the stress polar maps each with a score of 2 (SSS = 6) corresponding to 9% of the LV hypoperfused. Also note that circles around the stress perfusion scores (inset **A**) signify that the original scores in Figure 1A were modified by the expert system.

was used. A $P < .05$ was considered significant for all comparisons.

RESULTS

Interobserver Analysis

The human experts' reporting of the 100-patient subgroup resulted in 17 patients with CAD and 83 without. Of the 17 patients diagnosed with CAD 9 were reported to be ischemic. The breakdown of stress hypoperfusion by vascular territory in the 17 CAD patients were as follows: 8 LAD, 10 LCX, and 5 RCA. The breakdown of reversible ischemia by vascular territory in the 9 ischemic patients were: 6 LAD, 5 LCX, and 1 RCA. The overall agreement rates, *P* values, agreement differences, and 95% CI for each of the validated reported categories are shown in Table 2. At the high SP level, there were no statistical differences in the agreements between the AIsR findings/impressions compared to the experts' findings/impressions when compared vs the independent (10th) reader findings/impressions vs the experts in reporting the same studies. The finding of no statistical difference was true for the reporting of CAD ($P = .33$) or ischemia ($P = .37$). There were statistical differences for the tradeoff SN/SP level (CAD $P = .01$, ischemia $P = .03$) and even more differences for the high SN level (CAD $P = < .001$, ischemia $P = < .001$). At the high-SP level the 95% CI is above 0% for all categories (i.e., the AIsR findings are not inferior to the human expert reports) whereas they are below zero at four of eight categories at the tradeoff level and all eight categories for the high-SN levels.

AIsR Agreement with Experts

The nine human experts reporting of the 1,000-patient population resulted in 247 patients with CAD and 753 without. Of the 247 patients diagnosed with CAD, 120 were deemed ischemic. The breakdown of stress hypoperfusion by vascular territory in the 247 CAD patients revealed 135 LAD, 103 LCX, and 85

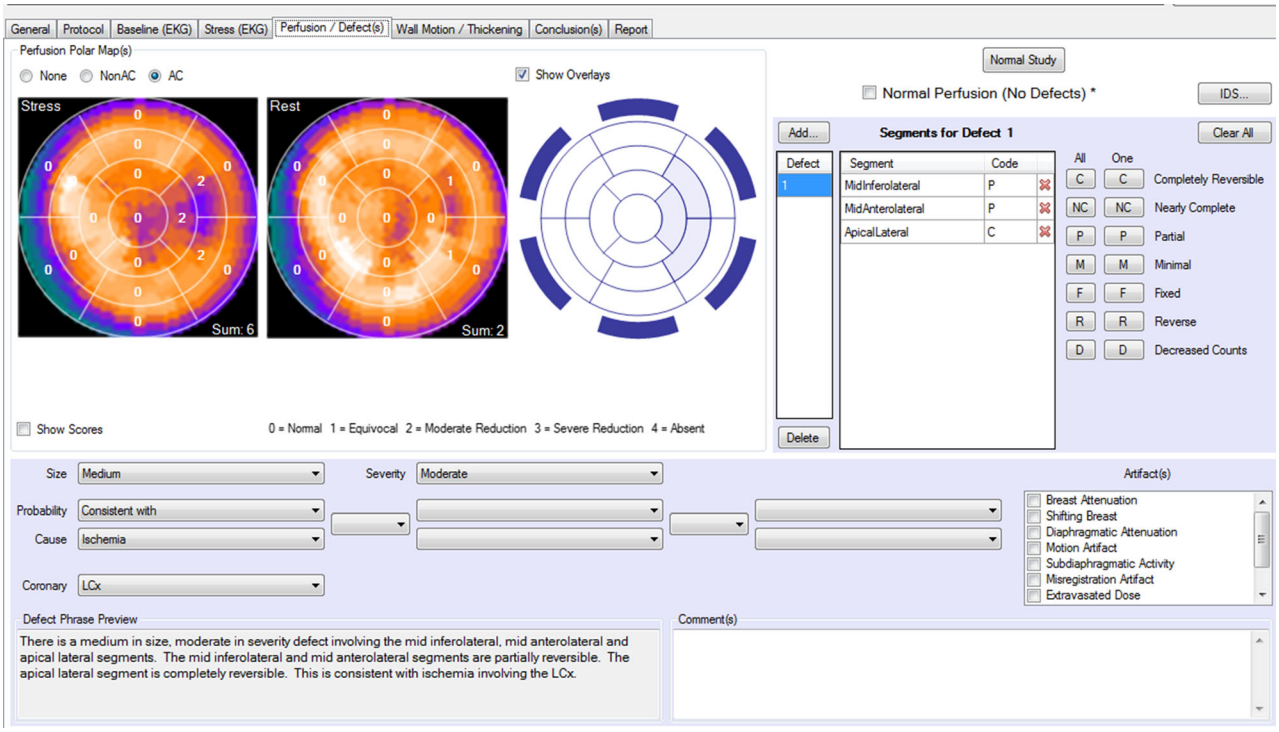

Figure 3. Automatically generated AIsR perfusion subreport of patient from Figure 2. Note concordance with the oblique slices and smart-scores. All drop-down arrows indicate a parameter that can be modified by the nuclear cardiology expert before it reaches the final report (not used for this validation).

RCA. These included 194 patients with single-vessel disease, 169 with double-vessel disease, and 117 with triple-vessel disease. The breakdown of reversible ischemia by vascular territory in the 120 ischemic patients revealed 61 LAD, 63 LCX, and 28 RCA. There were no significant differences between the 100-patient cohort used to test the noninferiority of AIsR vs expert and the 1,000-patient study group used to determine agreement rates between AIsR and experts. The categories tested were prevalence of CAD (347/1,000 vs 27/100; $P = .11$), prevalence of ischemia (120/1,000 vs 9/100; $P = .37$), agreement rate for CAD (820/1,000 vs 85/100; $P = .45$), and agreement rate for ischemia (880/1,000 vs 89/100; $P = .77$). All statistical comparisons were done using AIsR's high-SP mode.

Figure 2 depicts images and smart-scores in a female patient with reversible defects in the LCX coronary territories with the corresponding smart-reports shown in Figures 3 and 4A. Figure 4B shows the findings and impressions of the actual clinical report.

Figure 5 shows agreement results of AIsR-experts for the entire 1,000 patient group using the reported expert clinical read as the reference and compared for the three levels of SN/SP. These agreements are shown with regard to detection of stress-induced hypoperfusion and stress-induced ischemia. Note that for both the CAD and ischemia category, the high SP level yielded the highest accuracy and SP across global and regional results. These accuracies were determined to be statistically significant across all comparisons for global and regional hypoperfusion and reversibility. Table 3 shows percent agreement, κ agreement values between the AIsR and the experts' impressions of CAD and ischemia in the 1,000 MPI studies. These κ values ranged from 32.3 to 51.9 corresponding to a range from fair to moderate agreement as might be expected in the variation of clinical reports amongst nine different experts.

DISCUSSION

We developed and validated the diagnostic performance of an MPI natural language reporting system that utilizes nonparametric relative perfusion and function quantification as input to our expert system to interpret the study and generate the report. This is the first study

A Automatically generated findings and impressions excerpted from the AIsR report.

Nuclear Perfusion Findings:
There is a medium in size, moderate in severity defect involving the mid inferolateral, mid anterolateral and apical lateral segments. The mid inferolateral and mid anterolateral segments are partially reversible. The apical lateral segment is completely reversible. This is consistent with ischemia involving the LCx.

Nuclear Wall Motion Findings:
Post stress: The ejection fraction calculated at 57%. The diastolic volume calculated at 106 mL and systolic volume calculated at 45 mL. At rest: The diastolic volume calculated at 107 mL and systolic volume calculated at 55 mL.

Impressions:
- Abnormal myocardial perfusion study
- Low risk study

B Patient's actual clinical findings and impressions excerpted from the clinical report.

```
FINDINGS:
The study shows no significant patient motion in stress and rest
images and the quality is adequate. Left ventricular size is normal.

There is an 8%, small, reversible perfusion defect within the lateral
wall.

Gated tomographic images demonstrate normal wall motion and wall
thickening with a left ventricular ejection fraction of 50 % at rest
and 60 % at stress.

IMPRESSION:
1. There is an 8% reversible perfusion defect within the lateral wall.

2. Left ventricular ejection fraction is 50 % at rest and 60 % at
stress
```

Figure 4. Findings and impressions extracted from AI-structured report (**A**) and actual excerpts of the clinical report (**B**) for the MPI study shown in Figures 1 to 3. Note concordance in the presence and the location of hypoperfusion associated with ischemia.

that compares automatically generated MPI natural language reports to actual clinical reports.

Our results show that the reporting of CAD (hypoperfusion at stress) and ischemia (reversibility at rest) from our automatically generated AIsR is not statistically inferior from that of experts when a high-SP mode is used (i.e., equivocal = normal) and the reporting of other experts is used as the reference standard. Importantly this high-SP mode yielded the highest accuracy in our extensive population. It should not be surprising that AIsR best agreed with the experts in the high-SP mode since this indicates the human image interpretation trend being adjusted to the drop in the prevalence of abnormal studies to 25% at our institution (also in this population) similar to trends reported by others[20] and reported as low as 9% at other major institutions.[21] These findings are also consistent with those reported from a meta-analysis of 49,000 patients demonstrating diagnostic performance for referral bias corrected MPI (similar to echocardiography) of 99% SP and 38% SN (from 69%, 85% uncorrected, respectively).[22]

Strength of the Approach

This is the first report showing full integration between an image analysis system and structured reporting: to serve a critical need in modern imaging practice. Although the best agreement existed when the

Table 2. Agreement between automated smart-report results and human experts at three different sensitivity/specificity modes (n = 100).

High-specificity	**CAD**	**LAD**	**LCX**	**RCA**
%Agree: AIsR:expert	85	95	92	93
%Agree: Ind:expert	83	90	94	89
P value	.33	.07	.24	.14
Δ Agreement	.052	.05	− .02	.04
95% CI	− .07 to .11	− .01 to .11	− .08 to .04	− .03 to .11
High-specificity	**Ischemia**	**LAD**	**LCX**	**RCA**
%Agree: AIsR:expert	89	95	96	98
%Agree: Ind:expert	90	94	94	99
P value	.37	.33	.21	.28
Δ Agreement	− .01	.01	.02	− .01
95% CI	− .07 to .05	− .03 to .05	− .03 to .07	− .04 to .02
Tradeoff SN/SP	**CAD**	**LAD**	**LCX**	**RCA**
%Agree: AIsR:expert	74	82	77	89
%Agree: Ind:expert	83	90	94	89
P value	.01	.02	< .01	.5
Δ Agreement	− .09	− .08	− .17	.00
95% CI	− .17 to − .01	− .16 to − .003	− .25 to − .09	− .07 to .07
Tradeoff SN/SP	**Ischemia**	**LAD**	**LCX**	**RCA**
%Agree: AIsR:expert	83	91	89	97
%Agree: Ind:expert	90	94	94	99
P value	.03	.12	.06	.08
Δ Agreement	− .07	− .03	− .05	− .02
95% CI	− .14 to − .0007	− .08 to .02	− .11 to .01	− .05 to .007
High-sensitivity	**CAD**	**LAD**	**LCX**	**RCA**
%Agree: AIsR:expert	61	65	63	73
%Agree: Ind:expert	83	90	94	89
P value	< .001	.03	< .001	.001
Δ Agreement	− .22	− .25	− .31	− .16
95% CI	− .31 to − .13	− .35 to − .15	− .41 to − .21	− .26 to − .06
High-sensitivity	**Ischemia**	**LAD**	**LCX**	**RCA**
%Agree: AIsR:expert	64	76	72	88
%Agree: Ind:expert	90	94	94	99
P value	< .001	< .001	< .001	< .001
Δ Agreement	− .26	− .18	− .22	− .11
95% CI	− .35 to − .17	− .25 to − .10	− .31 to − .13	− .17 to − .05

high-SP mode was selected, this choice is easily modified to a high-SN level (or tradeoff level) when the AIsR is used to report on patients from a high-risk population such as diabetes. Newly reported here is the determination and use of our 17-segment smart-scores. This novel scoring uses a nonparametric normalized

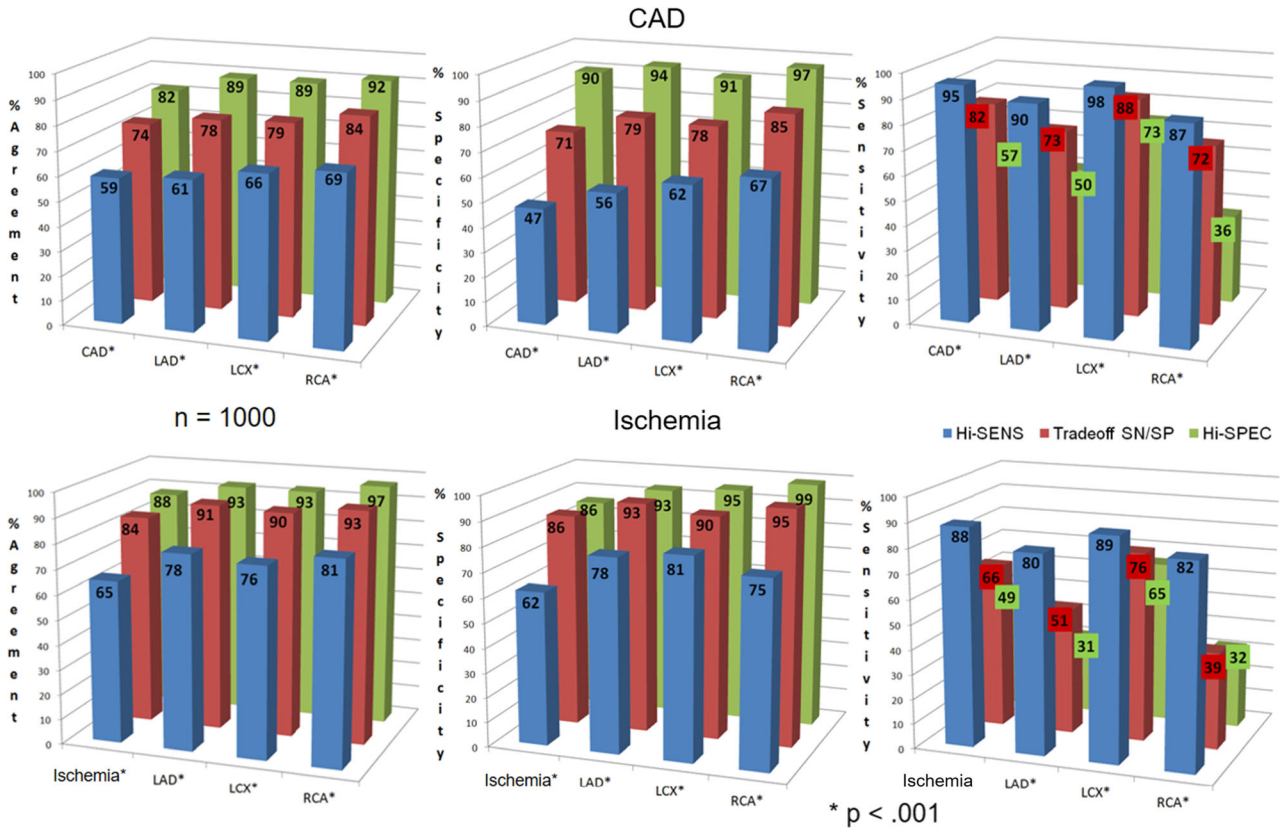

Figure 5. Diagnostic performance of the AI-structured report in reporting stress-induced hypoperfusion as indicative of CAD (top row) and reversibility at rest as indicative of ischemia (bottom row). Results for the modes: high specificity (green bars); sensitivity (SN)-specificity (SP) tradeoff, (red bars); and high sensitivity (blue bars) results are shown for agreement (i.e., accuracy: left column), specificity (middle column), and sensitivity (right column) (*$P < .001$). The labels CAD and ischemia in the abscissa of each graph refers to global findings regardless of vascular territory.

Table 3. Agreement, κ, and 95% CI results for the automated AIsR using high-specificity mode and the human experts reports as reference standard (n = 1000)

High-specificity	**CAD**	**LAD**	**LCX**	**RCA**
%Agree: AIsR:expert	82	89	89	92
κ	48.7	47.7	51.4	40.3
95% CI	42.0 to 55.4	38.7 to 56.7	42.8 to 59.9	27.6 to 53.0
High-specificity	**Ischemia**	**LAD**	**LCX**	**RCA**
%Agree: AIsR:expert	88	93	93	97
κ	43.6	32.3	51.9	36.9
95% CI	34.0 to 53.2	16.8 to 47.9	40.7 to 63.1	14.2 to 59.3

count distribution applied to information theory to generate a certainty of abnormality. This certainty for each segment is modified according to all the available perfusion and function information for that segment including rest, stress, changes between stress and rest, AC and non-AC images, and prone images. Although not validated here, the diagnostician is allowed to change manually any of the scores that in turn would

modify the report if needed. Importantly, as previously reported,[23] the expert system tracks all steps in generating the report as a justification, which may be used by the diagnosticians to decide whether they agree or not with the findings or impressions in the report. This is an important benefit of expert systems over conventional neural net or machine-learning approaches. Another benefit of the expert system approach used here is that, compared to other AI approaches, only the 40 normal patients used for database generation were needed to train the system as most of the training comes from the cumulative experience of the experts.

Comparison of AIsR to PERFEX

As described in the ''Methods'' section, we had previously developed and validated a decision support expert system to assist NC physicians with the image interpretation process.[8,9] There are several differences between that system (PERFEX) and the one reported here. PERFEX divided the LV into 32 segments; AIsR uses the standard 17-segment system. PERFEX depended on Gaussian distributions and statistics to determine normality and abnormality criteria; AIsR uses nonparametric statistics. PERFEX did not use the global or regional functional information to reach its conclusions; AIsR integrates the functional information into all its conclusions. PERFEX did not use its conclusions to modify the ECTb results; AIsR uses its knowledge base and the available quantitative information to modify the original segmental scores into smart-scores. If AC was performed, PERFEX would provide a separate interpretation for the AC study and one for the non-AC study; AIsR integrates both into one set of scores and one conclusion. If there was, also a prone study performed, AIsR would also integrate it. This integration takes place by trying to mimic in the code how human experts use the information. Before the integration is done AIsR determines segmental scores separately for each of the diagnostic categories considered: stress perfusion, rest perfusion, reversibility, and thickening. After these individual scores are determined, AIsR integrates the information into a meta-analysis module. Therefore, if an MPI study had AC, non-AC, and prone studies performed, AIsR would use the most normal score for that segment. If the same segment exhibited reversibility, AIsR would then modify the score using Bayesian statistics and the strength of the information (i.e., how much reversibility was present). Similarly, if the same segment exhibited abnormal thickening, then AIsR would again modify the score using the same approach as the one with reversibility. Perhaps the most obvious difference between PERFEX and AIsR is that AIsR propagates its conclusions into a sR.

Reference Standard

Since AI systems have to be ''trained'' and validated with both input images and accepted output interpretations, the question of what to use as the reference standard often arises. Use of invasive coronary angiography or clinical outcome as the gold standard for training and validating is often mentioned for an MPI AI system as attractive goal, but it misses the point of these systems, that is, to interpret studies with the same level of expertise as experts. Moreover, using invasive catheterization as a gold standard is biased by the referral pattern of abnormal MPI studies to catheterization as well as by the discrepancies in comparing physiologic results to anatomic ones. Outcome is certainly an important measure, but in MPI, coronary angiography and outcomes as gold standards are confounded by the fact that the scan interpretation (e.g., ischemia or no ischemia) has a major impact on the referral to the catheterization lab or the clinical outcome (intervention vs observation); consequently, these gold standards are biased. Simply stated, the interpretation of the study affects the treatment, and the treatment affects the outcomes thus biasing the outcomes as a reference standard. Thus, the practice of using interpretation of the MPI studies by experts is an acceptable approach that has been used by other researchers and ourselves.[9,24]

Limitations

First, all the data used for this evaluation were obtained retrospectively from one center. Second, we had to extract manually the needed diagnostic information from the clinically dictated reports to use as the reference standard. Third, all the clinical reporting was performed by Emory experts. Although these experts were trained at different institutions, it could be argued that over time, they tended to read similarly and perhaps different from readers from other institutions. Fourth, although the AIsR uses standardized reporting guidelines, we did not compare the size and severity of the hypoperfused or reversible areas between the experts and the AIsR, but only studied whether these were present and if so in which vascular territory. This is because in part when the clinical reports were generated reporting guidelines were not being strictly applied by the experts. Fifth, we also chose not to report here the clinical reporting agreements as to functional variables. Although these functional parameters were used in the generation of the smart-scores, these variables are quantitative and straightforward in how they are usually reported and therefore not compared for simplification. Sixth, although we have previously integrated patients' clinical information with their imaging results in order

to improve diagnostic accuracy,[25] this was not attempted here, as it would require either manual input and/or EMR interfaces with hospital systems that now would limit the applicability of this AIsR. Seventh, the agreement in reporting between the AIsR in the high-SP mode and our clinicians reflects the current reduced prevalence of disease (25%) of our patient referral pattern. In other scenarios (such as other countries) where the prevalence of disease is much higher than 25%, different results could have been obtained. This is the rationale that motivated us for allowing the AIsR to switch easily between modes such as high SN and SN/SP tradeoff mode. Finally, although the use of AC is not a limitation but an attribute that reduces the complexity of image interpretation, results of applying our approach to a large study population without AC (or prone imaging) cannot be predicted by the present study.

NEW KNOWLEDGE GAINED

Nonparametric statistics can be used to determine certainty that a regional parameter of LV perfusion and/or function is abnormal. Due to apparent reduced prevalence of CAD in populations of patients undergoing MPI, automated diagnostic systems agreement with experts improves when set to analyze images at high-SP settings.

CONCLUSIONS

Automatic sRs from computer-assisted interpretation of rest/stress myocardial perfusion SPECT studies by an AI expert system when operating at a high-SP level statistically agree with the interpretations of NC experts and exhibit diagnostic accuracy consistent with that of experts when their clinical reports are used as the reference standard.

Acknowledgments

This work was supported by the NHLBI Grant Number R42HL106818. The authors acknowledge Emory University Hospital Nuclear Cardiology Diagnosticians for allowing the use of their clinical MPI reports as well as Archana Kudrimoti for data mining the data warehouse for the clinical data reported.

Disclosures

EVG, CDC, RF, and JLK receive royalties from the sale of the Emory Cardiac Toolbox and/or Smart Report described in this article. The terms of this arrangement have been reviewed and approved by Emory University in accordance with its COI practice. CDA and CDC are employees of or consultants to Syntermed. All other authors report no conflicts of interest.

References

1. Fujita H, Katafuchi T, Uehara T, Nishimura T. Application of neural network to computer-aided diagnosis of coronary artery disease in myocardial SPECT Bull's-eye images. J Nucl Med. 1992;33:272–6.
2. Porenta G, Dorffner G, Kundrat S, Petta P, Duit-Schedlmayer J, Sochor H. Automated interpretation of planar thallium-201-dipyridamole stress-redistribution scintigrams using artificial neural networks. J Nucl Med. 1994;35:2041–7.
3. Hamilton D, Riley PJ, Miola UJ, Amro AA. A feed forward neural network for classification of bull's-eye myocardial perfusion images. Eur J Nucl Med. 1995;22:108–15.
4. Lindahl D, Lanke J, Lundin A, Palmer J, Edenbradt L. Improved classifications of myocardial bull's-eye scintigrams with computer-based decision support system. J Nucl Med. 1999;40:96–101.
5. Haddad M, Adlassnig KP, Porenta G. Feasibility analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams. Artif Intell Med. 1997;9:61–78.
6. Arsanjani RA, Xu Y, Dey D, Fish M, Dorbala S, Hayes S, et al. Improved accuracy of myocardial perfusion SPECT for the detection of coronary artery disease using a support vector machine algorithm. J Nucl Med. 2013;54:549–55.
7. Arsanjani RA, Xu Y, Dey D, Vahistha V, Nakanishi R, Hayes S, et al. Improved accuracy of myocardial perfusion SPECT for detection of coronary artery disease by machine learning in a large population. J Nucl Cardiol. 2013;20:553–62.
8. Ezquerra N, Mullick R, Cooke D, Krawczynska E, Garcia E. PERFEX: An expert system for interpreting 3D myocardial perfusion. Expert Syst Appl. 1993;6:459–68.
9. Garcia EV, Cooke CD, Folks RD, Santana CA, Krawczynska EG, De Braal L, et al. Diagnostic performance of an expert system for the interpretation of myocardial perfusion SPECT studies. J Nucl Med. 2001;42:1185–91.
10. Douglas PS, Hendel RC, Cummings JE, Dent JM, Hodgson JM, Hoffmann U, et al. ACCF/ACR/AHA/ASE/ASNC/HRS/NASCI/RSNA/SAIP/SCAI/SCCT/SCMR 2008 health policy statement on structured reporting in cardiovascular imaging. JACC. 2009;53:76–90.
11. Hansen CL, Richard A. Goldstein (Co-chairs): Myocardial perfusion and function: Single photon emission computed tomography. ASNC guidelines for nuclear cardiology procedures. J Nucl Cardiol. 2007;14:e39–60.
12. Garcia EV, Faber TL, Cooke CD, Folks RD, Chen J, Santana C. The increasing role of quantification in nuclear cardiology: The Emory approach. J Nucl Cardiol. 2007;14:420–32.
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. Esteves FP, Raggi P, Folks RD, Keidar Z, Askew JW, Rispler S, et al. Novel solid-state-detector dedicated cardiac camera for fast myocardial perfusion imaging: Multicenter comparison with standard dual detector cameras. J Nucl Cardiol. 2009;16:927–34.
15. Esteves FP, Galt JR, Folks RD, Verdes L, Garcia EV. Diagnostic performance of low-dose rest/stress Tc-99m tetrofosmin myocardial perfusion SPECT using the 530c CZT camera: Quantitative vs. visual analysis. J Nucl Cardiol. 2014;21:158–65.
16. Shannon EC, Weaver W. The mathematical theory of communication. Chicago: University of Illinois Press; 1949.
17. Shortliffe EH. Computer-based medical consultations: MYCIN. Amsterdam: Elsevier Scientific Publishing Company; 1976. p. 264.

18. Tilkemeier PL, Cooke CD, Grossman GB, McCallister BD, Ward RP. Standardized reporting of myocardial perfusion and function. J Nucl Cardiol. 2009. https://doi.org/10.1007/s12350-009-9095-8.
19. Dunn OJ. Basic statistics: A primer for the biomedical sciences. New York: Wiley; 1977. p. 116–9.
20. Chhabra L, Ahlberg AW, Henzlova MJ, Duvall WL. Temporal trends of stress myocardial perfusion imaging: Influence of diabetes, gender and coronary artery disease status. Int J Cardiol. 2016. https://doi.org/10.1016/j.ijcard.2015.09.020.
21. Rozanski A, Gransar H, Hayes SW, Min J, Friedman JD, Thomson LEJ, et al. Temporal trends in the frequency of inducible myocardial ischemia during cardiac stress testing: 1991 to 2009. JACC. 2013;10:1054–65.
22. Ladapo JA, Blecker S, Elashoff MR, Federspiel JJ, Vieira DL, Sharma G, et al. Clinical implications of referral bias in the diagnostic performance of exercise testing for coronary artery disease. J Am Heart Assoc. 2013;2:e000505. https://doi.org/10.1161/jaha113.000505.
23. Garcia EV, Taylor A, Manatunga D, Folks R. A software engine to justify the conclusions of an expert system for detecting renal obstruction on ^{99m}Tc-MAG3 scans. J Nucl Med. 2007;48:463–70.
24. Taylor A, Hill A, Binongo J, Manatunga A, Halkar R, Dubovsky EV, Garcia EV. Evaluation of two diuresis renography decision support systems designed to determine the need for furosemide in patients with suspected obstruction. AJR. 2007;188:1395–402.
25. Garcia EV, Taylor A, Folks R, Manatunga D, Halkar R, Savir-Baruch B, et al. iRENEX: A clinically informed decision support system for the interpretation of ^{99m}Tc-MAG3 scans to detect renal obstruction. Eur J Nucl Med Mol Imaging. 2012;39:1483–91. https://doi.org/10.1007/s00259-012-2151-7.