Can Physicians Identify Inappropriate Nuclear Stress Tests?

An Examination of Inter-Rater Reliability for the 2009 Appropriate Use Criteria for Radionuclide Imaging

Siqin Ye, MD, MS; LeRoy E. Rabbani, MD; Christopher R. Kelly, MD; Maureen R. Kelly, MD; Matthew Lewis, MD; Yehuda Paz, MD; Clara L. Peck, MD; Shaline Rao, MD; Sabahat Bokhari, MD; Shepard D. Weiner, MD; Andrew J. Einstein, MD, PhD

Background—We sought to determine inter-rater reliability of the 2009 Appropriate Use Criteria for radionuclide imaging and whether physicians at various levels of training can effectively identify nuclear stress tests with inappropriate indications.

Methods and Results—Four hundred patients were randomly selected from a consecutive cohort of patients undergoing nuclear stress testing at an academic medical center. Raters with different levels of training (including cardiology attending physicians, cardiology fellows, internal medicine hospitalists, and internal medicine interns) classified individual nuclear stress tests using the 2009 Appropriate Use Criteria. Consensus classification by 2 cardiologists was considered the operational gold standard, and sensitivity and specificity of individual raters for identifying inappropriate tests were calculated. Inter-rater reliability of the Appropriate Use Criteria was assessed using Cohen κ statistics for pairs of different raters. The mean age of patients was 61.5 years; 214 (54%) were female. The cardiologists rated 256 (64%) of 400 nuclear stress tests as appropriate, 68 (18%) as uncertain, 55 (14%) as inappropriate; 21 (5%) tests were unable to be classified. Inter-rater reliability for noncardiologist raters was modest (unweighted Cohen κ, 0.51, 95% confidence interval, 0.45–0.55). Sensitivity of individual raters for identifying inappropriate tests ranged from 47% to 82%, while specificity ranged from 85% to 97%.

Conclusions—Inter-rater reliability for the 2009 Appropriate Use Criteria for radionuclide imaging is modest, and there is considerable variation in the ability of raters at different levels of training to identify inappropriate tests. (***Circ Cardiovasc Qual Outcomes*. 2015;8:23-29. DOI: 10.1161/CIRCOUTCOMES.114.001067.)**

Key Words: ■ coronary disease ■ radioisotopes

Use of cardiovascular imaging has increased dramatically over the past decade,[1–4] leading to concerns that many of the nuclear cardiology tests being performed may have inappropriate indications, offer limited clinical value, while also increasing medical costs and patient radiation exposure.[2,3,5] To promote appropriate use of nuclear cardiology testing, the American College of Cardiology Foundation and other professional societies jointly released in 2005 and updated in 2009 the Appropriate Use Criteria (AUC) for radionuclide imaging (RNI).[6,7] However, although several studies have shown that nuclear stress tests with inappropriate indications are commonly performed,[8–14] efforts for broad application of the AUC to reduce inappropriate nuclear stress testing have frequently been unsuccessful.[9,15,16] Moreover, reported rates of inappropriate nuclear testing have ranged widely between studies. For example, while studies by Gibbons et al[17] and Saifi et al[18] have reported rates of inappropriate nuclear stress testing of <10%, and a recent report by Doukky et al[14] noted an inappropriateness testing rate that was as high as 45%.

Editorial see p 4

One potential barrier for the effective implementation of the AUC to reduce inappropriate nuclear stress testing is the complexity of the classification process, which can lead to marked classification disagreements between different raters.[8] To inform the future use and improvement of the 2009 AUC for RNI, it is critical to understand the extent to which such disagreements occur and their impact on the identification of appropriate and inappropriate nuclear stress tests. We, therefore, undertook a thorough investigation of the inter-rater reliability of the 2009 AUC for RNI for raters at different levels of training, by performing additional analysis of nuclear stress tests included in the COlumbia Nuclear CardiOlogy Radiation Dose (CONCORD) study.[5]

Methods

Study Sample

Details of the CONCORD study have been published previously.[5] Briefly, all 1097 consecutive patients undergoing nuclear stress testing during

Received June 6, 2014; accepted November 13, 2014.

From the Department of Medicine, Columbia University Medical Center and New York-Presbyterian Hospital.

The Data Supplement is available at http://circoutcomes.ahajournals.org/lookup/suppl/doi:10.1161/CIRCOUTCOMES.114.001067/-/DC1.

Correspondence to Siqin Ye, MD, Columbia University Medical Center, 622 W, 168th St, PH 9–320, New York, NY 10032. E-mail sy2357@cumc.columbia.edu

WHAT IS KNOWN

- The 2009 Appropriate Use Criteria (AUC) for radionuclide imagining sought to promote the appropriate utilization of nuclear cardiology testing.
- However, inter-rater reliability for the 2009 AUC may limits its effectiveness and has not been well characterized.

WHAT THE STUDY ADDS

- In this retrospective analysis of the COlumbia Nuclear CardiOlogy Radiation Dose (CONCORD) trial, we found that the inter-rater reliability for the 2009 AUC is modest between different physicians.
- We also found considerable variation in the ability of individual raters to identify inappropriate nuclear stress tests.
- These findings can inform the use of the AUC in clinical practice and can guide future interventions to apply effectively the AUC.

the first 100 days of 2006 (January 1–April 10) at Columbia University Medical Center, New York, NY, were identified through query of the electronic health records. Of these, 400 patients were randomly chosen as the study sample for the present analysis, with a separate 40 patients randomly chosen as the training sample to standardize AUC classification for different raters (as described below). As part of the CONCORD study, patient demographic data, including age, sex, race, insurance coverage, and zip code, were obtained through querying the electronic health records. Median annual household income in individual patient's zip code, a surrogate for socioeconomic status, was obtained using the 1999 US Census Bureau data.[19] For the present analysis, we further abstracted additional medical covariates from the electronic health records, including risk factors such as hypertension, hyperlipidemia, diabetes mellitus, and tobacco use; use of medications such as aspirin, β-blockers, statins, and other antihypertensive medications; history of prior coronary artery disease (CAD), myocardial infarction, percutaneous coronary intervention, or coronary artery bypass grafting; and results of the nuclear stress test. Symptoms at time of nuclear stress testing were also abstracted. Specifically, chest pain was classified as typical angina, atypical angina, and noncardiac chest pain, and other signs and symptoms including dyspnea, palpitations, and abnormal ECG were captured as potential ischemic equivalents, all in accordance with the 2009 AUC for RNI.[7] Complete radiation dosimetric data for all procedures was recorded[5]; its relationship to appropriateness will be evaluated in a separate report. The study was approved by the Institutional Review Board of the Columbia University Medical Center, and informed consent was waived because of the retrospective nature of this research.

AUC Classification and Assumptions

For AUC classification, the hierarchical flowchart outlined in the 2009 AUC for RNI was followed strictly.[7] In addition, similar to the method previously described by Gibbons et al,[8] we made several assumptions to standardize the application of the AUC. These include:

1. In accordance with the 2009 AUC for RNI definition for what should be considered angina equivalent, patients undergoing tests for symptoms other than chest pain, such as dyspnea, were defined as symptomatic and have atypical angina or nonanginal chest pain for the purpose of determining the pretest probability of CAD.[7]
2. Symptoms of chest pain were classified as typical angina, atypical angina, or noncardiac according to the Diamond and Forrester classification,[20] which was then used to determine the pretest probability of CAD in conjunction with age and sex, in accordance with recommendations contained in the 2009 AUC for RNI.[7]
3. Because of the retrospective design of the present study, for patients with missing data that were not captured in our electronic health records but that are needed for AUC classification (eg, results from tests performed at other facilities, such as prior cardiac imaging or lipid levels), their nuclear stress tests were considered as unable to be classified by the raters.
4. For a nuclear stress test that can be classified with >1 indication that have the same appropriateness category, the indication with the smallest numeric value is assigned as the AUC classification. An example is preoperative evaluation for patients undergoing noncardiac, intermediate-risk surgery who has both moderate to good functional capacity and no clinical risk factors. Nuclear stress testing in this setting qualifies for indications 41 and 42 of the AUC, both inappropriate. In this analysis, they are assigned indication 41.
5. For preoperative nuclear stress testing, risk classification of planned noncardiac surgeries (ie, as vascular, intermediate risk, or low risk) was performed using the 2007 American College of Cardiology/American Heart Association Guidelines on Perioperative Cardiovascular Evaluation and Care for Noncardiac Surgery,[21] as recommended by the 2009 AUC for RNI.[7]

Raters Recruitment and Training

For this analysis, 8 individual raters were recruited: 2 board-certified/eligible cardiologists not affiliated with the nuclear cardiology laboratory (L.E.R. and S.Y.), 2 first-year cardiology fellows (M.L. and S.R.), 2 internal medicine hospitalists (M.R.K. and C.L.P.), and 2 internal medicine interns (C.R.K. and Y.P.). The 2 cardiologists separately assigned AUC classification for 20 nuclear stress tests of the training subset, met to reconcile discrepancies, then repeated the process for the other 20 tests of the training sample. They then separately assigned AUC classifications for the 400 nuclear stress tests in the study sample and reconciled all discrepancies; a third cardiologist (A.J.E.) was available to adjudicate in case reconciliation was not possible.

For the other raters, the training process began with a 30-minute orientation session explaining the 2009 AUC for RNI, with specific emphasis on the classification process described above. In addition, to maximize the standardization of different raters, detailed instructions were provided on which data sources within the electronic health records should be used to carry out the rating process, and all raters were provided with both printed and electronic versions of the AUC document. Each rater then separately completed 2 training sessions. These sessions composed of assigning AUC classification to 20 nuclear stress tests from the training sample, followed by a one-on-one in-person feedback session with one of the cardiologists (S.Y.) to review discrepancies with the cardiologist consensus and to reinforce the classification process. After the 2 training sessions, each rater independently completed AUC classifications for the 400 nuclear stress tests from the study sample, assigning to each test 1 of 67 possible AUC indications or, when key data were missing, to the category unable to classify. For each test that could be classified, the indication assigned by each rater was then mapped to its corresponding appropriateness category of Appropriate, Uncertain, or Inappropriate.

Statistical Analysis

For this analysis, the consensus AUC classification of the 2 cardiologists was considered the operational gold standard. Baseline characteristics of patients were compared across appropriateness categories, using χ^2 tests for categorical variables and the Kruskal–Wallis rank test for age (because of skewed distribution). Subsequently, multivariable analysis to identify predictors of inappropriate nuclear stress testing was performed using logistic regression, with having an inappropriate indication being the binary dependent variable. In the multivariable model, potential predictors of inappropriate testing was selected using a cutoff of $P<0.20$ in univariable analysis, as suggested

Table 1. Baseline Characteristics, by Appropriate Use Categories

	Total (n=400)	Unable to Classify (n=21)	Appropriate (n=256)	Uncertain (n=68)	Inappropriate (n=55)	*P* Value*
Age, y, mean (SD)	61.5 (13.8)	61.2 (10.7)	61.0 (13.9)	66.1 (12.4)	58.3 (14.7)	0.017
Female, n (%)	214 (54%)	11 (52%)	133 (52%)	37 (54%)	33 (60%)	0.75
Non-white, n (%)	296 (74%)	14 (67%)	200 (78%)	42 (62%)	40 (73%)	0.042
Medicaid/no insurance, n (%)	128 (32%)	2 (10%)	91 (36%)	21 (31%)	14 (25%)	0.058
Tertile of median zip code income, n (%)						
Lowest tertile	150 (38%)	6 (29%)	105 (41%)	23 (34%)	16 (29%)	0.044
Middle tertile	117 (29%)	3 (14%)	79 (68%)	21 (31%)	14 (25%)	
Highest tertile	133 (33%)	12 (57%)	72 (28%)	24 (35%)	25 (45%)	
Angina symptoms, n (%)						
Typical angina	44 (11%)	2 (10%)	33 (13%)	8 (12%)	1 (2%)	<0.001
Atypical angina	130 (33%)	5 (24%)	96 (38%)	23 (34%)	6 (11%)	
Noncardiac chest pain	70 (18%)	0 (0%)	55 (21%)	7 (10%)	8 (15%)	
No chest pain	120 (30%)	11 (53%)	49 (19%)	22 (32%)	38 (69%)	
Unable to classify	36 (9%)	3 (14%)	23 (9%)	8 (12%)	2 (4%)	
Asymptomatic status, n (%)	87 (22%)	9 (43%)	30 (12%)	20 (29%)	28 (51%)	<0.001
Medication use, n (%)						
Aspirin	196 (49%)	10 (48%)	131 (51%)	38 (54%)	18 (33%)	0.068
β-Blocker	163 (41%)	8 (38%)	111 (43%)	28 (42%)	16 (29%)	0.27
Statins	165 (41%)	10 (48%)	109 (43%)	30 (44%)	16 (29%)	0.25
Anti-hypertensives	232 (58%)	10 (48%)	157 (61%)	37 (54%)	28 (51%)	0.31
Risk factors, n (%)						
Hypertension	293 (73%)	14 (67%)	195 (76%)	51 (75%)	33 (60%)	0.085
Hyperlipidemia	215 (54%)	14 (67%)	136 (53%)	39 (57%)	26 (47%)	0.44
Diabetes mellitus	126 (32%)	8 (38%)	87 (34%)	21 (31%)	10 (18%)	0.13
Tobacco use	56 (14%)	3 (14%)	39 (15%)	6 (9%)	8 (15%)	0.60
Prior CAD, n (%)	121 (30%)	6 (29%)	92 (36%)	16 (24%)	7 (13%)	0.004
Prior MI, n (%)	71 (18%)	5 (24%)	50 (20%)	11 (16%)	5 (9%)	0.26
Prior PCI, n (%)	55 (14%)	1 (5%)	44 (17%)	7 (10%)	3 (5%)	0.049
Prior CABG, n (%)	29 (7%)	0 (0%)	28 (11%)	0 (0%)	1 (2%)	0.002
Prior revascularization, n (%)	77 (19%)	1 (5%)	65 (25%)	7 (10%)	4 (7%)	0.001

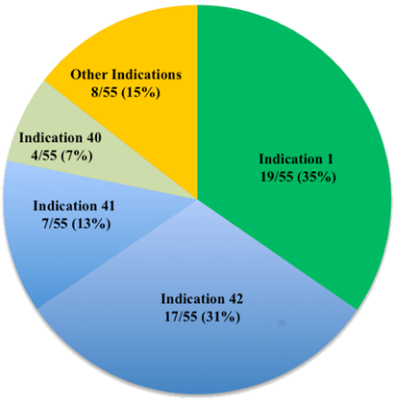
Prior revascularization includes having either prior percutaneous coronary intervention or coronary artery bypass grafting. Asymptomatic status is defined as the absence of chest pain and other signs and symptoms that could be considered as ischemic equivalents, including dyspnea, palpitations, and abnormal ECG. Patient is considered asymptomatic if there was no reported chest pain, dyspnea, palpitations, or abnormal ECG. CABG indicates coronary artery bypass grafting; CAD, coronary artery disease; MI, myocardial infarctions; and PCI, percutaneous coronary intervention.

*P values were calculated for comparison across different appropriate use categories; χ^2 tests were used except for age, where Kruskal–Wallis rank test was used because of skewed distribution.

by Maldonado and Greenland.[22] Furthermore, to avoid overfitting the model, we used the variable being asymptomatic to capture symptoms and the variables prior CAD and prior revascularization to capture previous cardiac history and revascularization status.

We tabulated the most common indications for appropriate and inappropriate testing both for the consensus cardiologist rating and for other individual raters. We determined the proportion of tests with results that were normal or probably normal by each appropriateness category. To determine inter-rater reliability, we calculated unweighted κ for each pairs of raters and for all noncardiologist raters jointly. Weighted Cohen κ for each pairs of raters was also calculated using a weight of 0 for disagreements in which one rater determine a test to be inappropriate while another determined the test to be appropriate, and a weight of 0.5 for all other disagreements. The weighting scheme is designed to account for the fact that the distinction between appropriate and inappropriate indications is likely to be more important than other kinds of disagreements (for example, between appropriate and uncertain indications). The 95% confidence intervals (CIs) for unweighted and weighted Cohen κ was derived for each pairs of raters, using a bootstrap approach with 1000 replications of the entire sample with replacement, performed with the Stata program kapci.[23] The κ statistic is conventionally interpreted as representing excellent inter-rater agreement when its value is >0.75, modest inter-rater agreement when its value is 0.40 to 0.75, and poor agreement when its value is <0.40.[24] We also calculated the proportions of agreement for raters at different training levels and for all noncardiologist raters, as well as the proportions of specific agreement for appropriate and inappropriate indications using the same groupings of raters. The 95% CIs for all proportions were estimated through the asymptotic (Wald) method.

To describe the ability of individual raters to identify inappropriate tests, we also performed a validity analysis examining the sensitivity and specificity of each noncardiologist rater for identifying these tests as inappropriate. In this context, sensitivity is defined as the proportion of nuclear stress tests with inappropriate indications (according to the cardiologist consensus) that were correctly classified



Most Common Indications for Inappropriate Tests
1. Detection of CAD: Symptomatic; Evaluation of Ischemic Equivalent (Nonacute) Low pretest probability of CAD; ECG interpretable AND can exercise **(I)**
42. Risk Assessment: Preoperative Evaluation for Noncardiac Surgery Without Active Cardiac Conditions; Intermediate-Risk Surgery No clinical risk factors **(I)**
41. Risk Assessment: Preoperative Evaluation for Noncardiac Surgery Without Active Cardiac Conditions; Intermediate-Risk Surgery Moderate to good functional capacity (greater than or equal to 4 METS) **(I)**
40. Risk Assessment: Preoperative Evaluation for Noncardiac Surgery Without Active Cardiac Conditions; Low-Risk Surgery Preoperative evaluation for noncardiac surgery risk assessment **(I)**

Figure 1. Most common indications for inappropriate tests, by frequency. **Left**, frequency of most common inappropriate indications; **right**, descriptions of most common inappropriate indications. CAD indicates coronary artery disease.

as inappropriate by an individual rater. Similarly, specificity is calculated as the proportion of nuclear stress tests that do not have inappropriate indications (according to the cardiologist consensus) that were classified by an individual rater into a category other than inappropriate. For all statistical tests, a *P* value of 0.05 was considered statistically significant, and all analyses were conducted using Stata software, version 12.0 (Stata Corp, College Station, TX).

Results

Of the 400 patients included in this analysis, the mean (SD) age was 61.5 (13.8) years, and 214 (54%) were female. Other baseline characteristics are as shown in Table 1. The most frequent indications for nuclear stress testing in our sample are indications 55 (evaluation of ischemia in symptomatic patients after percutaneous coronary intervention or coronary artery bypass grafting; 61/15%), 8 (possible acute coronary syndrome, no ECG changes, low-risk Thrombolysis in Myocardial Infarction score, negative troponins; 56/14%), 31 (preoperative evaluation for vascular surgery, no clinical risk factors; 46/12%), 3 (evaluation of ischemia, intermediate pretest probability, ECG interpretable and able to exercise; 41/10%), and 4 (evaluation of ischemia, intermediate pretest probability, ECG uninterpretable or unable to exercise; 22/6%). The 2 cardiologists classified 256 (64%) of 400 tests as appropriate, 68 (18%) as uncertain, 55 (14%) as inappropriate; 21 (5%) tests were not able to be classified by the raters. Of the 55 nuclear stress tests classified as inappropriate, 47 (85%) had indications 1, 40, 41, or 42 (Figure 1).

In univariable analysis, there were significant differences across appropriateness categories for age, race, and median zip code income. Types of angina symptoms, being asymptomatic, having prior history of CAD, and having percutaneous coronary intervention, coronary artery bypass grafting, or revascularizations also differed across appropriateness categories. In the multivariable logistic regression model (Table 2), being asymptomatic (odds ratio, 7.26; 95% CI, 3.50–15.07; $P<0.001$) and having diabetes mellitus (odds ratio, 0.41; 95% CI, 0.18–0.92; $P=0.03$) independently predicted inappropriate nuclear stress testing.

Inter-Rater Reliability of AUC Classification

Individual raters identified 61% to 70% of nuclear stress tests as appropriate and 11% to 23% of nuclear stress tests as inappropriate. The most common appropriate and inappropriate indications were also broadly similar for individual raters. Reliability was modest between raters at the same level of training, with unweighted κ ranging from 0.37 to 0.61. The overall κ for all 6 noncardiologist raters was 0.51 (95% CI, 0.45–0.55; Table 3). Unweighted and weighted Cohen κs for all pairs of raters are as presented in the Table in the Data Supplement. The proportion of agreement for raters at the same level of training ranged from 0.66 to 0.79 and was 0.74 (95% CI, 0.73–0.75) for all 6 noncardiologist raters. The proportion of specific agreement was higher for appropriate indications than for inappropriate indications (Table 3). For validity of AUC rating, there was marked variation in the sensitivity and specificity of different raters for the identification of inappropriate tests compared with the cardiologist consensus, with sensitivity ranging from

Table 2. Predictors of Inappropriate Nuclear Stress Tests in Multivariable Logistic Regression Model

	Odds Ratio (95% Confidence Interval)	*P* Value
Age	0.98 (0.96–1.01)	0.15
Non-white	0.48 (0.19–1.23)	0.13
Medicaid/no insurance	0.78 (0.37–1.62)	0.50
Median zip code income		
Lowest tertile	Ref	Ref
Middle tertile	1.22 (0.54–2.75)	0.63
Highest tertile	1.68 (0.70–4.05)	0.25
Asymptomatic status	7.26 (3.50–15.07)	<0.001
Aspirin	0.75 (0.37–1.54)	0.44
Hypertension	0.90 (0.43–1.88)	0.79
Diabetes mellitus	0.41 (0.18–0.92)	0.03
Prior CAD	0.39 (0.11–1.41)	0.15
Prior revascularization	0.89 (0.17–4.66)	0.89

Prior revascularization includes prior percutaneous coronary intervention or coronary artery bypass grafting. Asymptomatic status is defined as the absence of chest pain and other signs and symptoms that could be considered as ischemic equivalents, including dyspnea, palpitations, and abnormal ECG. CAD indicates coronary artery disease.

Table 3. Inter-Rater Reliability and Proportion of Agreement (Overall, and Specific for Appropriate and Inappropriate Tests), by Training Level and for All Noncardiologist Raters

Rater Groupings (Number of Raters)	Unweighted κ (95% CI)	Proportion of Agreement, All Indications (95% CI)	Proportion of Specific Agreement for Appropriate Indications (95% CI)	Proportion of Specific Agreement for Inappropriate Indications (95% CI)
Interns (2)	0.61 (0.54–0.68)	0.79 (0.75–0.83)	0.88 (0.85–0.91)	0.52 (0.43–0.62)
Hospitalists (2)	0.63 (0.55–0.70)	0.81 (0.77–0.85)	0.87 (0.84–0.89)	0.57 (0.48–0.65)
Fellows (2)	0.37 (0.30–0.46)	0.66 (0.61–0.71)	0.76 (0.72–0.80)	0.48 (0.40–0.55)
All Noncardiologist raters (6)	0.51 (0.46–0.55)	0.74 (0.73–0.75)	0.83 (0.82–0.84)	0.52 (0.50–0.54)

47% (fellow 1) to 82% (hospitalist 1) and specificity ranging from 85% (fellow 1) to 97% (intern 1; Figure 2).

Discussion

In our application of the 2009 AUC for RNI to nuclear stress tests performed at a single academic medical center in 2006, we found that ≈15% of the tests examined were performed for inappropriate indications, with a small number of indications capturing a majority of these tests. Furthermore, we also found that inter-rater reliability for AUC classification was only modest and that there was considerable variation in the ability of different raters to identify accurately tests with inappropriate indications despite standardized training.

Our results on the prevalence, make-up, and findings of nuclear stress tests with inappropriate indications are also broadly consistent with prior studies. Similar to our findings, Gibbons et al[8] and Mehta et al[13] both used the earlier 2007 AUC document to identify 13% to 14% of nuclear stress tests performed at academic medical centers as having inappropriate indications. In contrast, other studies have reported both higher and lower proportions of nuclear stress tests with inappropriate indications, likely reflecting differences in practice settings and institutional, geographical, and temporal differences in ordering patterns.[9–12,14] Previous studies also support our finding that a small number of inappropriate indications explained a majority of inappropriate nuclear stress tests that were performed.[8–10,13] Nonetheless, as we applied the 2009 AUC for RNI to a time period before its widespread adoption, it is possible that the prevalence and characteristics of nuclear stress tests with inappropriate indications in the current era may differ from our subgroup analysis of the CONCORD study. It is important to note, however, that the differences in era should not affect our finding that individual providers may have difficulty applying the AUC to identify inappropriate tests. Although Gibbon et al[8] has previously demonstrated modest agreement for AUC classification between 2 nurse practitioners,[17] our study is the first to examine extensively inter-rater reliability of the AUC for physician raters with different clinical backgrounds. The substantial disagreements between AUC classifications of different raters despite standardized training highlight potential challenges for using the AUC at point-of-care to guide appropriate test ordering, especially as there is considerable disagreement and variable sensitivity for different raters applying the AUC to identify tests with inappropriate indications.

There may be several explanations for the suboptimal inter-rater reliability of the AUC observed in our study. The assignment of AUC ratings to individual tests is a cognitively complex task that involve many steps, such as determination of past history of cardiac testing, assessment of pretest probability for CAD, and identification of the AUC indication that best describes the clinical scenario at hand. Potential errors can occur at each step of the process and can potentially cascade to result in incorrect AUC classification. The rating process could also be influenced by heuristic biases that have been well described in medical decision-making literature.[25,26] For instance, a recent encounter by a rater with a young patient admitted for myocardial infarction could affect perceptions of risk and appropriateness, leading to deviations from the AUC document as the result of availability bias. Finally, it is possible that the training sessions provided in our study was not of sufficient duration or intensity to ensure that raters conform to the recommended AUC rating process.

Our study also has implications for the effective use of AUC in clinical settings. The substantial variation between different raters' abilities to identify inappropriate nuclear stress tests suggest a potential explanation for why AUC interventions that relied on judgment of appropriateness by individual providers did not reduce inappropriate nuclear stress testing.[15]

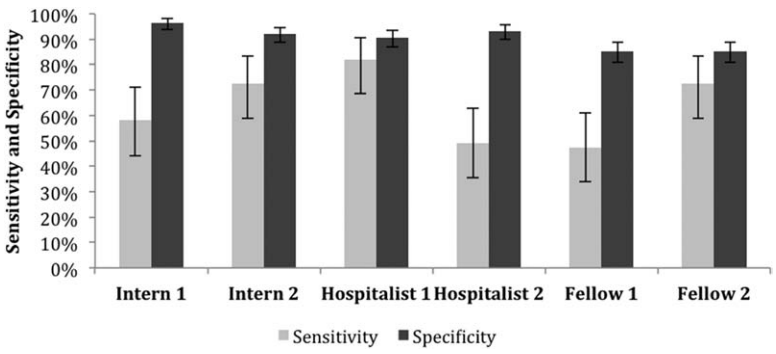


Figure 2. Sensitivity and specificity of individual raters for identifying inappropriate nuclear stress tests. Error bars represent 95% confidence intervals.

Future efforts will need to address the complexity of the AUC classification system, through steps such as consolidation of overlapping indications and further streamlining of the classification process, or through improved decision support such as that offered by the American College of Cardiology FOCUS initiative.[18] Ultimately, in order for AUC documents to realize successfully their mission of promoting appropriate use of medical resources, more research is also needed to address the usability of AUCs and to determine the most effective and efficient approaches to implementing them.

There are several limitations to our study. We used the κ statistic to assess inter-rater reliability of the AUC, which can be affected by the baseline rates of appropriate and inappropriate tests. The retrospective design of our study may affect the accuracy of clinical data collected and could impact appropriateness determination, and the affiliation of individual raters with the academic medical center studied could also introduce bias. However, our findings on the prevalence and make-up of tests with inappropriate indications are broadly consistent with prior literature. Furthermore, our demonstration of only modest inter-rater reliability between different raters is more likely an intrinsic characteristic of the AUC classification system and is unlikely to be affected by the retrospective nature of our study or by rater affiliations. The number of raters involved in our study is modest and limits our ability to assess the effect of training level on inter-rater reliability. We also cannot rule out the possibility that our raters may not be representative of all potential users of the 2009 AUC for RNI. Future research will need to confirm our findings with broader sample of raters and to determine most effective approaches to training clinicians to identify inappropriate tests. Finally, we used the terminology of appropriate, uncertain, and inappropriate as set forth in the 2009 AUC document and used the consensus between ratings of 2 cardiologists as the operational gold standard for our analysis. Prior studies, however, have suggested that valid differences in opinions may exist for what should be considered appropriate and inappropriate,[27–29] contributing to a recent AUC Methodology Update that has changed the terminology of AUC classifications for subsequently issued AUC criteria to appropriate, may be appropriate, and rarely appropriate.[30] This change in terminology should not affect our overall findings, and we agree with the implication of the change that for some nuclear stress tests, appropriateness classification may not fully capture their clinical utility or lack thereof.

In conclusion, in this retrospective analysis of the appropriate use of nuclear stress tests at an academic medical center, we characterized the pattern of inappropriate testing and demonstrated the difficulty that individual clinicians may have in using the AUC to identify inappropriate tests. These findings identify an important barrier for successful implementation of AUCs and can inform future interventions that promote the appropriate use of cardiovascular imagining.

Sources of Funding

Dr Ye was supported by a National Institutes of Health (NIH) training grant (T32 HL007854-16), by an American College of Cardiology Foundation/Merck Research Fellowship Award, and by a NIH K23 career development award (K23 HL121144). Dr Einstein was supported by a NIH KL2 institutional career development award (KL2 RR024157), by a Herbert Irving Associate Professorship, as a Victoria and Esther Aboodi Cardiology Researcher, by the Louis V. Gerstner, Jr. Scholars Program, and by the Lewis Katz Cardiovascular Research Prize for a Young Investigator.

Disclosures

Dr Einstein has received research grants for other work from GE Healthcare, Philips Healthcare, and Spectrum Dynamics. The other authors report no conflicts.

References

1. Iglehart JK. The new era of medical imaging–progress and pitfalls. *N Engl J Med*. 2006;354:2822–2828. doi: 10.1056/NEJMhpr061219.
2. Fazel R, Krumholz HM, Wang Y, Ross JS, Chen J, Ting HH, Shah ND, Nasir K, Einstein AJ, Nallamothu BK. Exposure to low-dose ionizing radiation from medical imaging procedures. *N Engl J Med*. 2009;361:849–857. doi: 10.1056/NEJMoa0901249.
3. Rozanski A, Gransar H, Hayes SW, Min J, Friedman JD, Thomson LE, Berman DS. Temporal trends in the frequency of inducible myocardial ischemia during cardiac stress testing: 1991 to 2009. *J Am Coll Cardiol*. 2013;61:1054–1065. doi: 10.1016/j.jacc.2012.11.056.
4. Lucas FL, DeLorenzo MA, Siewers AE, Wennberg DE. Temporal trends in the utilization of diagnostic testing and treatments for cardiovascular disease in the United States, 1993-2001. *Circulation*. 2006;113:374–379. doi: 10.1161/CIRCULATIONAHA.105.560433.
5. Einstein AJ, Weiner SD, Bernheim A, Kulon M, Bokhari S, Johnson LL, Moses JW, Balter S. Multiple testing, cumulative radiation dose, and clinical indications in patients undergoing myocardial perfusion imaging. *JAMA*. 2010;304:2137–2144. doi: 10.1001/jama.2010.1664.
6. Brindis RG, Douglas PS, Hendel RC, Peterson ED, Wolk MJ, Allen JM, Patel MR, Raskin IE, Hendel RC, Bateman TM, Cerqueira MD, Gibbons RJ, Gillam LD, Gillespie JA, Hendel RC, Iskandrian AE, Jerome SD, Krumholz HM, Messer JV, Spertus JA, Stowers SA; American College of Cardiology Foundation Quality Strategic Directions Committee Appropriateness Criteria Working Group; American Society of Nuclear Cardiology; American Heart Association. ACCF/ASNC appropriateness criteria for single-photon emission computed tomography myocardial perfusion imaging (SPECT MPI): a report of the American College of Cardiology Foundation Quality Strategic Directions Committee Appropriateness Criteria Working Group and the American Society of Nuclear Cardiology endorsed by the American Heart Association. *J Am Coll Cardiol*. 2005;46:1587–1605. doi: 10.1016/j.jacc.2005.08.029.
7. Hendel RC, Berman DS, Di Carli MF, Heidenreich PA, Henkin RE, Pellikka PA, Pohost GM, Williams KA; American College of Cardiology Foundation Appropriate Use Criteria Task Force; American Society of Nuclear Cardiology; American College of Radiology; American Heart Association; American Society of Echocardiology; Society of Cardiovascular Computed Tomography; Society for Cardiovascular Magnetic Resonance; Society of Nuclear Medicine. ACCF/ASNC/ACR/AHA/ASE/SCCT/SCMR/SNM 2009 Appropriate Use Criteria for Cardiac Radionuclide Imaging: A Report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, the American Society of Nuclear Cardiology, the American College of Radiology, the American Heart Association, the American Society of Echocardiography, the Society of Cardiovascular Computed Tomography, the Society for Cardiovascular Magnetic Resonance, and the Society of Nuclear Medicine. *J Am Coll Cardiol*. 2009;53:2201–2229. doi: 10.1016/j.jacc
8. Gibbons RJ, Miller TD, Hodge D, Urban L, Araoz PA, Pellikka P, McCully RB. Application of appropriateness criteria to stress single-photon emission computed tomography sestamibi studies and stress echocardiograms in an academic medical center. *J Am Coll Cardiol*. 2008;51:1283–1289. doi: 10.1016/j.jacc.2007.10.064.
9. Hendel RC, Cerqueira M, Douglas PS, Caruth KC, Allen JM, Jensen NC, Pan W, Brindis R, Wolk M. A multicenter assessment of the use of single-photon emission computed tomography myocardial perfusion imaging with appropriateness criteria. *J Am Coll Cardiol*. 2010;55:156–162. doi: 10.1016/j.jacc.2009.11.004.
10. Nelson KH, Willens HJ, Hendel RC. Utilization of radionuclide myocardial perfusion imaging in two health care systems: assessment with the 2009 ACCF/ASNC/AHA appropriateness use criteria. *J Nucl Cardiol*. 2012;19:37–42. doi: 10.1007/s12350-011-9467-8.
11. Carryer DJ, Askew JW, Hodge D, Miller TD, Gibbons RJ. The impact of ordering provider specialty on appropriateness classification. *J Nucl Cardiol*. 2012;19:285–290. doi: 10.1007/s12350-011-9459-8.

12. Gupta A, Tsiaras SV, Dunsiger SI, Tilkemeier PL. Gender disparity and the appropriateness of myocardial perfusion imaging. *J Nucl Cardiol*. 2011;18:588–594. doi: 10.1007/s12350-011-9368-x.
13. Mehta R, Ward RP, Chandra S, Agarwal R, Williams KA; American College of Cardiology Foundation; American Society of Nuclear Cardiology. Evaluation of the American College of Cardiology Foundation/American Society of Nuclear Cardiology appropriateness criteria for SPECT myocardial perfusion imaging. *J Nucl Cardiol*. 2008;15:337–344. doi: 10.1016/j.nuclcard.2007.10.010.
14. Doukky R, Hayes K, Frogge N, Balakrishnan G, Dontaraju VS, Rangel MO, Golzar Y, Garcia-Sayan E, Hendel RC. Impact of appropriate use on the prognostic value of single-photon emission computed tomography myocardial perfusion imaging. *Circulation*. 2013;128:1634–1643. doi: 10.1161/CIRCULATIONAHA.113.002744.
15. Gibbons RJ, Askew JW, Hodge D, Kaping B, Carryer DJ, Miller T. Appropriate use criteria for stress single-photon emission computed tomography sestamibi studies: a quality improvement project. *Circulation*. 2011;123:499–503. doi: 10.1161/CIRCULATIONAHA.110.975995.
16. Lin FY, Dunning AM, Narula J, Shaw LJ, Gransar H, Berman DS, Min JK. Impact of an automated multimodality point-of-order decision support tool on rates of appropriate testing and clinical decision making for individuals with suspected coronary artery disease: a prospective multicenter study. *J Am Coll Cardiol*. 2013;62:308–316. doi: 10.1016/j.jacc.2013.04.059.
17. Gibbons RJ, Askew JW, Hodge D, Miller TD. Temporal trends in compliance with appropriateness criteria for stress single-photon emission computed tomography sestamibi studies in an academic medical center. *Am Heart J*. 2010;159:484–489. doi: 10.1016/j.ahj.2009.12.004.
18. Saifi S, Taylor AJ, Allen J, Hendel R. The use of a learning community and online evaluation of utilization for SPECT myocardial perfusion imaging. *JACC Cardiovasc Imaging*. 2013;6:823–829. doi: 10.1016/j.jcmg.2013.01.012.
19. US Census Bureau. Download Center: American FactFinder. http://factfinder.census.gov/servlet/DownloadDatasetServlet?_lang=en. Accessed July 1, 2009.
20. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. *N Engl J Med*. 1979;300:1350–1358. doi: 10.1056/NEJM197906143002402.
21. Fleisher LA, Beckman JA, Brown KA, Calkins H, Chaikof EL, Fleischmann KE, Freeman WK, Froehlich JB, Kasper EK, Kersten JR, Riegel B, Robb JF. ACC/AHA 2007 guidelines on perioperative cardiovascular evaluation and care for noncardiac surgery: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (writing committee to revise the 2002 guidelines on perioperative cardiovascular evaluation for noncardiac surgery). *Circulation*. 2007;116:e418–e500.
22. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol*. 1993;138:923–936.
23. Reichenheim ME. Confidence intervals for the kappa statistic. *Stata Journal*. 2004;4:421–428.
24. Fleiss JL. *Statistical Methods for Rates and Proportions*. New York, NY: John Wiley & Sons; 1973.
25. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med*. 2003;78:775–780.
26. Choudhry NK, Soumerai SB, Normand SL, Ross-Degnan D, Laupacis A, Anderson GM. Warfarin prescribing in atrial fibrillation: the impact of physician, patient, and hospital characteristics. *Am J Med*. 2006;119:607–615. doi: 10.1016/j.amjmed.2005.09.052.
27. Lin FY, Rosenbaum LR, Gebow D, Kim RJ, Wolk MJ, Patel MR, Dunning AM, Labounty TM, Gomez MJ, Shaw LJ, Narula J, Douglas PS, Raman SV, Berman DS, Min JK. Cardiologist concordance with the American College of Cardiology appropriate use criteria for cardiac testing in patients with coronary artery disease. *Am J Cardiol*. 2012;110:337–344. doi: 10.1016/j.amjcard.2012.03.026.
28. Chan PS, Brindis RG, Cohen DJ, Jones PG, Gialde E, Bach RG, Curtis J, Bethea CF, Shelton ME, Spertus JA. Concordance of physician ratings with the appropriate use criteria for coronary revascularization. *J Am Coll Cardiol*. 2011;57:1546–1553. doi: 10.1016/j.jacc.2010.10.050.
29. Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE. The reproducibility of a method to identify the overuse and underuse of medical procedures. *N Engl J Med*. 1998;338:1888–1895. doi: 10.1056/NEJM199806253382607.
30. Hendel RC, Patel MR, Allen JM, Min JK, Shaw LJ, Wolk MJ, Douglas PS, Kramer CM, Stainback RF, Bailey SR, Doherty JU, Brindis RG. Appropriate use of cardiovascular technology: 2013 ACCF appropriate use criteria methodology update: a report of the American College of Cardiology Foundation appropriate use criteria task force. *J Am Coll Cardiol*. 2013;61:1305–1317. doi: 10.1016/j.jacc.2013.01.025.

Circulation
Cardiovascular Quality and Outcomes

Can Physicians Identify Inappropriate Nuclear Stress Tests?: An Examination of Inter-Rater Reliability for the 2009 Appropriate Use Criteria for Radionuclide Imaging

Siqin Ye, LeRoy E. Rabbani, Christopher R. Kelly, Maureen R. Kelly, Matthew Lewis, Yehuda Paz, Clara L. Peck, Shaline Rao, Sabahat Bokhari, Shepard D. Weiner and Andrew J. Einstein

Circ Cardiovasc Qual Outcomes. 2015;8:23-29; originally published online January 6, 2015;
doi: 10.1161/CIRCOUTCOMES.114.001067
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1941-7705. Online ISSN: 1941-7713

Supplemental Material

Supplemental Table. Inter-rater reliability of AUC classification, calculated as unweighted and weighted Cohen's kappas.

	Unweighted Cohen's kappa (95% Confidence Interval)					
	Intern 1	**Intern 2**	**Hospitalist 1**	**Hospitalist 2**	**Fellow 1**	**Fellow 2**
Cardiologists Consensus	0.71 (0.65-0.78)	0.67 (0.61-0.74)	0.65 (0.58-0.72)	0.54 (0.46-062)	0.40 (0.33-0.48)	0.48 (0.40-0.56)
Intern 1	-	0.61 (0.54-0.68)	0.59 (0.51-0.66)	0.56 (0.49-0.63)	0.44 (0.36-0.51)	0.42 (0.34-0.50)
Intern 2		-	0.58 (0.51-0.65)	0.58 (0.50-0.65)	0.41 (0.34-0.49)	0.52 (0.44-0.59)
Hospitalist 1			-	0.63 (0.55-0.70)	0.44 (0.36-0.52)	0.46 (0.38-0.54)
Hospitalist 2				-	0.53 (0.45-0.61)	0.48 (0.40-0.55)
Fellow 1					-	0.37 (0.30-0.46)
Fellow 2						-
	Weighted Cohen's Kappa (95% Confidence Interval)					
	Intern 1	**Intern 2**	**Hospitalist 1**	**Hospitalist 2**	**Fellow 1**	**Fellow 2**
Cardiologists Consensus	0.69 (0.61-0.76)	0.65 (0.58-0.72)	0.64 (0.57-0.72)	0.50 (0.43-0.60)	0.36 (0.26-0.44)	0.48 (0.39-0.56)
Intern 1	-	0.57 (0.48-0.65)	0.55 (0.46-0.63)	0.52 (0.44-0.61)	0.39 (0.30-0.48)	0.40 (0.32-0.50)
Intern 2		-	0.55 (0.47-0.63)	0.54 (0.45-0.62)	0.37 (0.29-0.46)	0.53 (0.45-0.61)
Hospitalist 1			-	0.56 (0.47-0.64)	0.37 (0.28-0.46)	0.45 (0.36-0.54)
Hospitalist 2				-	0.47 (0.38-0.56)	0.44 (0.35-0.52)
Fellow 1					-	0.34 (0.25-0.43)
Fellow 2						-