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 k 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 k, 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 al17 and Saifi et al18 have reported rates of inappropriate nuclear stress testing of <10%, and a recent report by Doukky et al19 noted an inappropriate 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.4 Briefly, all 1097 consecutive patients undergoing nuclear stress testing during
WHAT IS KNOWN

• The 2009 Appropriate Use Criteria (AUC) for radio-
  nuclide imaging 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.

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.,\(^3\) 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 nonangular 
   chest pain for the purpose of determining the pretest probability 
   of CAD.\(^7\)
2. Symptoms of chest pain were classified as typical angina, atypi-
   cal angina, or noncardiac according to the Diamond and For-
   rester classification,\(^20\) which was then used to determine the
by Maldonado and Greenland. 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 Cohen $\kappa$ for each pairs of raters and for all noncardiologist raters jointly. The $\kappa$ 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. We also calculated the proportions of agreement for raters at different training levels and for all noncardiologist raters, using a bootstrap approach with 1000 replications of the entire sample with replacement, performed with the Stata program kapci.

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 (incorrectly identified as inappropriate) among those determined to be inappropriate by the cardiologist consensus, that were correctly classified as inappropriate by the noncardiologist rater. The $\kappa$ 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. 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.
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 \(\kappa\) ranging from 0.37 to 0.61. The overall \(\kappa\) for all 6 noncardiologist raters was 0.51 (95% CI, 0.45–0.55; Table 3). Unweighted and weighted Cohen \(\kappa\) values 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 26% to 87% and specificity ranging from 71% to 91%.

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

<table>
<thead>
<tr>
<th>Odds Ratio (95% Confidence Interval)</th>
<th>(P) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.98 (0.96–1.01)</td>
</tr>
<tr>
<td>Non-white</td>
<td>0.48 (0.19–1.23)</td>
</tr>
<tr>
<td>Medicaid/no insurance</td>
<td>0.78 (0.37–1.62)</td>
</tr>
<tr>
<td>Median zip code income</td>
<td></td>
</tr>
<tr>
<td>Lowest tertile</td>
<td>Ref</td>
</tr>
<tr>
<td>Middle tertile</td>
<td>1.22 (0.54–2.75)</td>
</tr>
<tr>
<td>Highest tertile</td>
<td>1.68 (0.70–4.05)</td>
</tr>
<tr>
<td>Asymptomatic</td>
<td>7.26 (3.50–15.07)</td>
</tr>
<tr>
<td>Aspirin</td>
<td>0.75 (0.37–1.54)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.90 (0.43–1.88)</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>0.41 (0.18–0.92)</td>
</tr>
<tr>
<td>Prior CAD</td>
<td>0.39 (0.11–1.41)</td>
</tr>
<tr>
<td>Prior revascularization</td>
<td>0.89 (0.17–4.66)</td>
</tr>
</tbody>
</table>

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

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

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. and Mehta et al. 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. has previously demonstrated modest agreement for AUC classification between 2 nurse practitioners,15 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
Future efforts will need to address the complexity of the AUC classification system, through steps such as consolidation of overlapping indications and further streamlined classification process, or through improved decision support such as that offered by the American College of Cardiology FOCUS initiative. 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 $\kappa$ 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 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 a 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 may be considered appropriate and inappropriate, 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. 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 imaging.

**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**


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

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
Copyright © 2015 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circoutcomes.ahajournals.org/content/8/1/23

Data Supplement (unedited) at:
http://circoutcomes.ahajournals.org/content/suppl/2015/01/06/CIRCOUTCOMES.114.001067.DC1

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Quality and Outcomes can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Quality and Outcomes is online at:
http://circoutcomes.ahajournals.org//subscriptions/
Supplemental Material
**Supplemental Table.** Inter-rater reliability of AUC classification, calculated as unweighted and weighted Cohen’s kappas.

<table>
<thead>
<tr>
<th></th>
<th>Unweighted Cohen's kappa (95% Confidence Interval)</th>
<th>Weighted Cohen’s Kappa (95% Confidence Interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intern 1</td>
<td>Intern 2</td>
</tr>
<tr>
<td>Cardiologists Consensus</td>
<td>0.71 (0.65-0.78)</td>
<td>0.67 (0.61-0.74)</td>
</tr>
<tr>
<td>Intern 1</td>
<td>-</td>
<td>0.61 (0.54-0.68)</td>
</tr>
<tr>
<td>Intern 2</td>
<td>-</td>
<td>0.58 (0.51-0.65)</td>
</tr>
<tr>
<td>Hospitalist 1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hospitalist 2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fellow 1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Fellow 2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>