Assessing Strength of Evidence of Appropriate Use Criteria for Diagnostic Imaging Examinations

The Harvard community has made this article openly available. **Please share** how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1093/jamia/ocv194</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:34908222">http://nrs.harvard.edu/urn-3:HUL.InstRepos:34908222</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Assessing Strength of Evidence of Appropriate Use Criteria for Diagnostic Imaging Examinations

Ronilda Lacson1,2, Ali S Raja2,3, David Osterbur2,4, Ivan Ip1,2,5, Louise Schneider2,5, Paul Bain2,4, Carol Mita2,4, Julia Whelan2,4, Patricia Silveira1, David Dement1, Ramin Khorasani1,2

ABSTRACT

Objective For health information technology tools to fully inform evidence-based decisions, recommendations must be reliably assessed for quality and strength of evidence. We aimed to create an annotation framework for grading recommendations regarding appropriate use of diagnostic imaging examinations.

Methods The annotation framework was created by an expert panel (clinicians in three medical specialties, medical librarians, and biomedical scientists) who developed a process for achieving consensus in assessing recommendations, and evaluated by measuring agreement in grading the strength of evidence for 120 empirically selected recommendations using the Oxford Levels of Evidence.

Results Eighty-two percent of recommendations were assigned to Level 5 (expert opinion). Inter-annotator agreement was 0.70 on initial grading ($\kappa = 0.35$, 95% CI, 0.23-0.48). After systematic discussion utilizing the annotation framework, agreement increased significantly to 0.97 ($\kappa = 0.88$, 95% CI, 0.77-0.99).

Conclusions A novel annotation framework was effective for grading the strength of evidence supporting appropriate use criteria for diagnostic imaging exams.

Keywords: Clinical decision support system, clinical practice guidelines, evidence-based practice, diagnostic imaging

INTRODUCTION

Healthcare institutions are increasingly leveraging health information technology tools to improve care quality, enhance patient safety, and lower healthcare costs.1-15 Two recent federal regulations promote use of clinical decision support (CDS) – Meaningful Use regulations11 provide modest incentives for CDS adoption and the Protecting Access to Medicare Act (PAMA)16 mandates that, beginning in 2017, clinicians ordering covered diagnostic imaging (Computerized Tomography scans [CT], Magnetic Resonance Imaging [MRI], nuclear medicine, and Positron Emission Tomography scans [PET]) must consult specified appropriate use criteria through certified CDS mechanisms. These criteria must be evidence-based to the extent feasible, and reimbursement will depend on confirmation that evidence-based recommendations were consulted.

To be effective, recommendations delivered via CDS should be backed by high quality evidence.17-19 Frequent low-quality alerts likely provoke “alert fatigue.”20 However, while imaging-related clinical guidelines and recommendations are publicly available,21-31 information regarding their validity and quality of evidence is not. This is likely because grading clinical recommendations for strength of evidence requires a complex set of skills, particularly the ability to identify, obtain, and critically appraise relevant research publications. Medical librarians may be optimally positioned to perform these tasks and have previously assessed evidence-based materials for nursing curriculum development.32-34

OBJECTIVE

We aimed to create and evaluate a comprehensive and scalable annotation framework for grading the strength of evidence of appropriate use criteria for diagnostic imaging examinations.

METHODS

This study was ruled exempt from Institutional Review Board review.

Annotation Framework Development

The annotation framework (Figure 1) was developed by an expert panel, taking into account current appropriate-use criteria for guiding medical imaging selection, as envisioned by PAMA. The panel consisted of clinicians in emergency medicine, internal medicine and radiology, and librarians and biomedical scientists with expertise in information retrieval, knowledge representation, and clinical study design.

The primary unit of analysis was a unit of evidence, defined as an assertion regarding the appropriateness of utilizing a diagnostic imaging procedure for certain indications and contraindications, taken from a published recommendation, guideline, systematic review, or clinical decision rule. It consists of an “IF . . . THEN” statement wherein a single statement contains sufficient knowledge to make an independent assertion to perform an imaging procedure (eg, “THEN” phrase). The procedural orientation of the knowledge representation is rooted in the nature of appropriate use criteria – recommending an exam for a specific clinical situation – and is ideal for knowledge sharing. Each unit of evidence was abstracted from a single source and analyzed independently. However, each unit was allowed to have many (or no) clinical studies supporting it. Each study was reviewed to determine its type, and then graded for level of evidence.

Several evidence-based grading systems were considered: the Grading of Recommendations Assessment, Development and Evaluation18,37,38 the United States Preventive Services Task Force (USPSTF),39-41 and The Agency for Healthcare Research and Quality’s Strength of Evidence model.42,43 We chose the Oxford Centre for...
Evidence-based Medicine (OCEBM) level of evidence grading system, 2009 version.\textsuperscript{[44]} It is relatively simple to use and mimics the clinical decision-making approach. Unlike Grading of Recommendations Assessment, Development and Evaluation of Strength of Evidence, developed primarily to synthesize evidence to establish new recommendations, OCEBM allows busy clinicians to quickly assess evidence for implementation into practice. USPSTF grading, on the other hand, is designed to recommend a service for use in clinical practice.\textsuperscript{[40,45]} with a level of certainty regarding net benefit and considering professional judgment and patient preferences, which are difficult to quantify. Thus, we limit our use of USPSTF grading to the I statement, defined as current evidence that is insufficient to assess the balance of benefits and harms of the service.\textsuperscript{[45]} Our annotation framework grades each unit of evidence as I, defined, or non-I (ie, not insufficient). We also introduced a grade of non-scorable-contradicts, for evidence that is contradictory to that advocated in the corresponding clinical study.

We limited our study to the OCEBM grading system for diagnosis. Level 1 includes Systematic Review (SR) of Level 1 studies and Clinical Decision Rules tested in one (Level 1b) or more (1a) clinical centers. It also includes validating cohort studies with good reference standards (1b) and studies with findings whose specificity or sensitivity is so high to rule in/out a diagnosis (1c). Level 2 includes SR of Level 2 studies (2a), Clinical Decision Rules after derivation and exploratory cohort studies with good reference standards (2b). Level 3 includes SR of Level 3 studies (3a) which are either non-consecutive or have no consistently applied reference standards (3b). Level 4 includes case-control studies and those with poor or non-independent reference standards. Level 5 refers to expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles.”\textsuperscript{[44]}

Evaluating the Annotation Framework
We assessed the strength of evidence for a convenience sample of 120 empirically selected units of diagnostic imaging evidence to evaluate the framework from five sources – two professional society guidelines (American College of Radiology, American College of Physicians), local best practice from two healthcare organizations (Ottawa Civic Hospital, Brigham and Women’s Hospital) and a clinical study (Wells Criteria for Pulmonary Embolism) (Appendix B). These encompassed X-ray imaging (extremities), CT scanning (head, chest, and extremities), and MRI (head, spine, and extremities).

The annotation framework provides procedures for grading units of evidence (Figure 2).

Statistical Analyses
We calculated percentage agreement and kappa for five ordinal categories: Levels 1–5. Sublevels (eg, 1a, 1b, and 1c) were collapsed together for analyses. Percentage agreement measured exact agreement between curators for grading a unit of evidence, and kappa agreement measured inter-annotator agreement, taking into account the probability of agreement due to chance.\textsuperscript{[46]} Weighted kappa agreement was calculated based on a predefined linear weight matrix, with disagreements weighted based on the distance between levels of agreement (eg, Level 1 is closer to Level 2).\textsuperscript{[49]} We identified strategies for reconciling disagreements between experts based on the most common reasons for lack of agreement within the annotation framework. A weekly group discussion composed of at least one physician and other curators reconciled disagreements between librarians.

RESULTS
Data Sources
The selected guidelines included American College of Radiology appropriateness criteria for Acute Shoulder Pain, Minor Head Trauma, Knee Pain, and Colorectal Cancer Screening. Other guidelines and recommendations also included those for Ankle Pain and Pulmonary Embolism, and an American College of Physicians guideline for Low Back Pain (Appendix B). These encompassed X-ray imaging (extremities), CT scanning (head, chest, and extremities), and MRI (head, spine, and extremities).

Distribution of Strength of Evidence
A total 9/120 units of evidence were classified as Level 1 (8%), 7/120 as Level 2 (6%), 2/120 as Level 3 (2%) and the majority, 99/120, as Level 5 (82%). Expert opinion was limited to guidelines with no supportive studies. Rather, these often included studies that were not sufficient to support the specific unit of evidence (eg, I statement). A total 86/120 units of evidence (72%) were graded I; non-I and NS each had 17/120 (14%) units.

Agreement in Grading Strength of Evidence
Agreement of grades between curators for each unit of evidence was 84/120 (70%) before and 117/120 (97%) after discussion. Overall, initial inter-annotator kappa agreement was fair at 0.35 (95% CI, 0.23-0.48).\textsuperscript{[48]} After discussion and standardization, it increased to 0.88 (95% CI, 0.77-0.99). Weighted kappa for independent curators was 0.52 before discussion, indicating moderate agreement, and 0.92 after discussion. Table 2 enumerates the major causes of disagreement identified in grading. Although a significant amount of disagreement was due to human error (in identifying study design or missing study inclusion criteria), the
majority was due to clinical studies that provided insufficient support for the unit of evidence. This included studies that were inconsistent with specific clinical variants included in the recommendations, temporal attributes of the disease (eg, acute ankle pain), and specific protocols for the imaging modality (eg, metal suppression protocol).

Compliance with PAMA (public law 113-93) regulations will necessitate increasing reliance on evidence-based appropriateness use criteria for certain imaging examinations. Although guidelines are available for diagnostic imaging in specific clinical scenarios, the levels of evidence supporting these recommendations are not readily available. We developed a novel annotation framework for large-scale annotation of units of evidence that is comprehensive and scalable (http://libraryofevidence.med.harvard.edu).35

The framework can assess units of evidence from a range of sources. All the units of evidence are converted into single decision rules with recommendation for performing an imaging modality based on defined inclusion and exclusion criteria. These expressions are represented in a language based on Arden Syntax logic grammar for representing logical decision criteria.50,51

In addition, the annotation framework can capture grading disagreement, and contains procedures for reconciliation. In the future, a validating clinician will review the grading assignments (in lieu of group discussion).

The reconciliation of disagreements begins with a discussion of the underlying guideline, followed by a focused evaluation of the specific units of evidence (which are assessed independently). More importantly, if clinical studies suggest that both imaging modalities have similar accuracy for capturing a ligamentous tear but that ultrasound is less expensive, the recommendation to perform both exams will have equal evidence grading (ie, level of evidence). In our process, cost, experience, and availability are not considered in grading strength of evidence.

The OCEBM grading system poses another source of disagreement for curators. Although it specifies types of studies that would justify

---

**Table 1: Example of main annotation attributes and attribute definition**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guideline name</td>
<td>Label for a specific unit of evidence</td>
<td>Acute shoulder pain</td>
</tr>
<tr>
<td>Source</td>
<td>Source name, date, citation</td>
<td>American College of Radiology Appropriateness Criteria for acute shoulder pain, 201047</td>
</tr>
<tr>
<td>Imaging modality</td>
<td>Radiologic imaging examination (with or without contrast)</td>
<td>Ultrasound</td>
</tr>
<tr>
<td>Body region</td>
<td>Anatomical region for radiologic imaging examination</td>
<td>Shoulder</td>
</tr>
<tr>
<td>Disease entity</td>
<td>Disease or sign/symptom</td>
<td>Acute shoulder pain</td>
</tr>
<tr>
<td>Indication (IF)</td>
<td>Criteria for assertion regarding performing an examination (or not)</td>
<td>[Prior radiograph performed] AND [Radiographs non-contributory] AND [Previous total shoulder arthroplasty] AND [Suspect rotator cuff tear]</td>
</tr>
<tr>
<td>Resultant action (THEN)</td>
<td>Assertion regarding performing an examination (or not)</td>
<td>Perform Ultrasound shoulder</td>
</tr>
<tr>
<td>Evidence grade (OCEBM)</td>
<td>OCEBM grade</td>
<td>2b</td>
</tr>
<tr>
<td>Evidence grade (USPSTF)</td>
<td>USPSTF grade</td>
<td>Non-I</td>
</tr>
</tbody>
</table>

---

**Figure 2: Procedures for grading units of evidence to ensure consistency among multiple curators.**

A guideline may consist of multiple units of evidence, each of which will be graded based on supporting studies. Each unit of evidence is informed from (i.e., a study comparing two imaging modalities will be scored twice, each grading the units of evidence one for each imaging modality). Clinical studies have to be consistent with the specific imaging modality and evidence frame (e.g., MRI with no specific protocol mentioned in the study associated with a unit of evidence reducing a metal suppression protocol is inconsistent). Details of the grading system, including study design, have to be provided in detail, consistent with the clinical guidelines (e.g., Level 3 research studies include validating cohort study from different clinical centers, different clinical criteria should exclude centers not belonging to the same clinical network).

A guideline may consist of multiple units of evidence, each of which will be graded based on supporting studies. Each unit of evidence is informed from (i.e., a study comparing two imaging modalities will be scored twice, each grading the units of evidence one for each imaging modality). Clinical studies have to be consistent with the specific imaging modality and evidence frame (e.g., MRI with no specific protocol mentioned in the study associated with a unit of evidence reducing a metal suppression protocol is inconsistent). Clinical studies have to be consistent with the specific imaging modality and evidence frame (e.g., MRI with no specific protocol mentioned in the study associated with a unit of evidence reducing a metal suppression protocol is inconsistent). Details of the grading system, including study design, have to be provided in detail, consistent with the clinical guidelines (e.g., Level 3 research studies include validating cohort study from different clinical centers, different clinical criteria should exclude centers not belonging to the same clinical network).

---
certain levels of grading, the developers intentionally allowed decisions to upgrade or downgrade the level of evidence based on merits of the study design, believing that certain observational trials are sufficiently convincing to provide definitive evidence.\textsuperscript{34,52} We identified clinical studies that were validated in multiple centers, but belonging to the same practice network. We consider such studies to be Level 1b (ie, tested within one center), as opposed to 1a (ie, from different centers).

Expert grading relies heavily on human decision-making\textsuperscript{38,53,54} and is thus prone to human error. Typical causes of human error include erroneous documentation of grading or misunderstanding recommendations that are negated (e.g., do not perform chest CT is mistaken for a recommendation to perform the exam). The annotation framework clarifies a substantial amount of accidental mis-assignments and misunderstandings. Although clinicians and librarians provide complementary expertise in information management, investigative reasoning, and clinical assessment, there is need to precisely define various levels of the grading system, as well as elucidate some steps that are relevant to the grading process. The annotation framework addresses these steps in detail.

Limitations
The annotation framework relies on an expression language based on single decision rules, and will not generalize to multi-step decision support for which rules can be triggered by decisions/actions from previous states (as are necessary in some clinical guidelines). In addition, decision rules are represented using non-standard knowledge representation, albeit semi-structured, with a local dictionary. Next steps will include knowledge representation in a formal executable representation as well as integration with a standard terminology.

CONCLUSIONS
We developed an annotation framework for systematically grading recommendations regarding appropriate use of diagnostic imaging examinations. The framework captured all units of evidence extracted from various clinical sources, and could be used as the basis for a curated library of appropriate use criteria that would facilitate compliance with PAMA and help accelerate adoption of evidence into practice to optimize the return on substantial national investments in healthcare IT.

SUPPLEMENTARY MATERIAL
Supplementary material is available online at http://jama.oxfordjournals.org/.

REFERENCES
14. Langton KB, Johnston ME, Haynes RB, Mathieu A. A critical appraisal of the literature on the effects of computer-based clinical decision support

Table 2: Examples of disagreement between curators for grading level of evidence

<table>
<thead>
<tr>
<th>Annotation framework</th>
<th>Cause of disagreement</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guideline</td>
<td>Disagreement in guideline message</td>
<td>No MRI for non-traumatic knee pain, when a clinical study supports MRI</td>
</tr>
<tr>
<td>Unit of Evidence</td>
<td>References have to be specific to a single unit of evidence</td>
<td>Recommend MRI for acute shoulder pain for a specific clinical variant, when clinical study is not specific for this variant</td>
</tr>
<tr>
<td>Clinical Studies</td>
<td>Recommended diagnostic exam is inconsistent with time frame or modality</td>
<td>MRI with dedicated metal suppression protocol for acute shoulder pain, whereas clinical study does not specify this protocol</td>
</tr>
<tr>
<td>Strength of Evidence</td>
<td>Differences in interpretation of OCEBM classification system for classifying units of evidence</td>
<td>For validation studies, the study design has to be consistent with a prior exploratory study; a validation study has to be performed in an independent study setting</td>
</tr>
<tr>
<td>Expert Processing and Rating</td>
<td>Errors in identifying study design</td>
<td>Level 3 (non-consecutive study without consistently applied reference standards) vs. Level 4 (poor reference standard)</td>
</tr>
<tr>
<td></td>
<td>Errors when assessing negative recommendations</td>
<td>‘No MRI’ for initial evaluation of knee pain is mistakenly assessed as ‘Recommending MRI’</td>
</tr>
</tbody>
</table>

**AUTHOR AFFILIATIONS**

1. Center for Evidence Based Imaging, Department of Radiology, Brigham and Women’s Hospital, Boston, MA 02115, USA
2. Harvard Medical School, Boston, MA 02115, USA
3. Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
4. Countway Library of Medicine, Boston, MA 02115, USA
5. Department of Medicine, Brigham and Women’s Hospital, Boston, MA 02115, USA