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

RECEIVED 1 September 2015
REVISED 16 October 2015
ACCEPTED 9 November 2015
PUBLISHED ONLINE FIRST 17 January 2016

Ronilda Lacson[1,2], Ali S Raja[2,3], David Osterbur[2,4], Ivan Ip[1,2,5], Louise Schneider[2,5], Paul Bain[2,4], Carol Mita[2,4], Julia Whelan[2,4], Patricia Silveira[1], David Dement[1], Ramin Khorasani[1,2]

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 regulations[11] 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

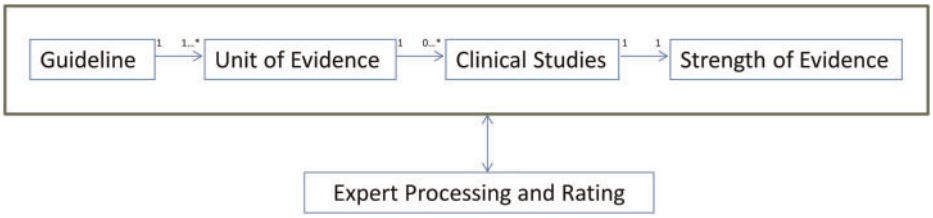
The annotation framework (Figure 1) was developed by an expert panel, taking into account current appropriate-use criteria for guiding medical imaging selection, as envisioned by PAMA. The panel consisted of clinicians in emergency medicine, internal medicine and radiology, and librarians and biomedical scientists with expertise in information retrieval, knowledge representation, and clinical study design.[35]

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.[36] 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 Evaluation[18,37,38] 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

Correspondence to Ronilda Lacson, Center for Evidence-Based Imaging, Department of Radiology Brigham and Women's Hospital, 20 Kent Street, 2nd floor, Boston, MA 02445, USA; rlacson@partners.org. For numbered affiliations see end of article.

Figure 1: Annotation framework for each unit of evidence.



Evidence-based Medicine (OCEBM) level of evidence grading system, 2009 version.[44] It is relatively simple to use and mimics the clinical decision-making approach. Unlike Grading of Recommendations Assessment, Development and Evaluation and 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,[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.[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."[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 evaluation). Based on a 64% baseline agreement for assessing strength of evidence for imaging in cerebrovascular diseases,[46] this sample size provided 80% power to detect a 25% increase in agreement with a two-tailed α value of 0.05. Each unit of evidence was assigned to at least two of four total curators, all medical librarians at the Countway Medical Library (Harvard Medical School, Boston, MA, USA). The annotation task was to grade strength of evidence utilizing both OCEBM and I statement (Appendix A). This entailed manually reviewing clinical studies listed in the bibliography of the published recommendation and in supplementary files (eg, evidence tables). For peer-reviewed articles, only the article itself was reviewed. In addition, the main annotation attributes for grading (Table 1) were fully captured for each unit of evidence.

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.[48] 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).[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 not 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.

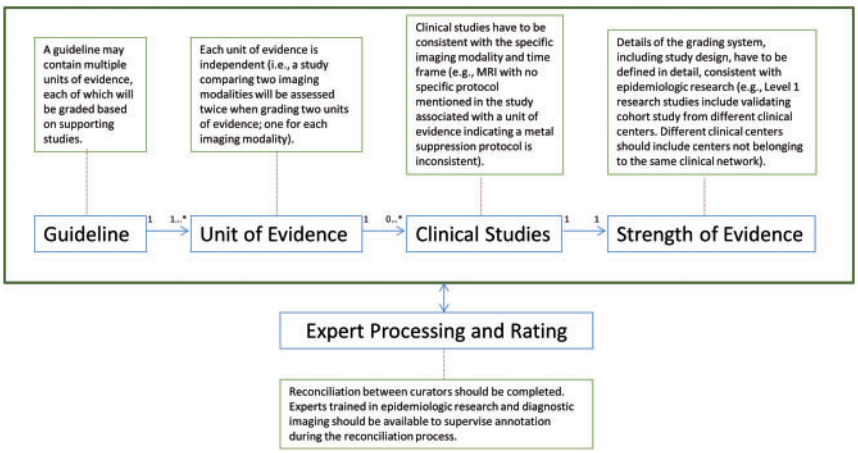
Agreement in Grading Strength of Evidence

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).[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

Table 1: Example of main annotation attributes and attribute definition

Attribute	Definition	Example
Guideline name	Label for a specific unit of evidence	Acute shoulder pain
Source	Source name, date, citation	American College of Radiology Appropriateness Criteria for acute shoulder pain, 2010[47]
Imaging modality	Radiologic imaging examination (with or without contrast)	Ultrasound
Body region	Anatomical region for radiologic imaging examination	Shoulder
Disease entity	Disease or sign/symptom	Acute shoulder pain
Indication (IF)	Criteria for assertion regarding performing an examination (or not)	[Prior radiograph performed] AND [Radiographs non-contributory] AND [Previous total shoulder arthroplasty] AND [Suspect rotator cuff tear]
Resultant action (THEN)	Assertion regarding performing an examination (or not)	Perform Ultrasound shoulder
Evidence grade (OCEBM)	OCEBM grade	2b
Evidence grade (USPSTF)	USPSTF grade	Non-I

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



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.[16] 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). This independent evaluation is necessary due to a disagreement identified early in the process; while clinical studies compare various imaging modalities for specific clinical criteria (eg, adults with shoulder pain), units of evidence assess specific modalities separately. Thus, a recommendation to perform shoulder ultrasound is independent from a recommendation to perform shoulder MRI. 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 2: Examples of disagreement between curators for grading level of evidence

Annotation framework	Cause of disagreement	Examples
Guideline	Disagreement in guideline message	No MRI for non-traumatic knee pain, when a clinical study supports MRI
Unit of Evidence	References have to be specific to a single unit of evidence	Recommend MRI for acute shoulder pain for a specific clinical variant, when clinical study is not specific for this variant
Clinical Studies	Recommended diagnostic exam is inconsistent with time frame or modality	MRI with dedicated metal suppression protocol for acute shoulder pain, whereas clinical study does not specify this protocol
	Study design is not clear	Inclusion criteria include MRI and CT with results analyzed in aggregate, whereas unit of evidence is only for MRI
Strength of Evidence	Differences in interpretation of OCEBM classification system for classifying units of evidence	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
Expert Processing and Rating	Errors in identifying study design	Level 3 (non-consecutive study without consistently applied reference standards) vs. Level 4 (poor reference standard)
	Errors when assessing negative recommendations	"No MRI" for initial evaluation of knee pain is mistakenly assessed as "Recommending MRI"

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.[44,52] 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[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://jamia.oxfordjournals.org/.

REFERENCES

1. Bates DW, Pappius EM, Kuperman GJ *et al.* Measuring and improving quality using information systems. *Stud Health Technol Inform.* 1998;52 (Pt 2):814–818.
2. Ip IK, Raja AS, Gupta A *et al.* Impact of clinical decision support on head computed tomography use in patients with mild traumatic brain injury in the ED. *Am J Emerg Med.* 2015;33(3):320–325.
3. Ip IK, Schneider L, Seltzer S *et al.* Impact of provider-led, technology-enabled radiology management program on imaging. *Am J Med.* 2013;126(8):687–692.
4. Raja AS, Ip IK, Prevedello LM *et al.* Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. *Radiology.* 2012;262(2):468–474.
5. Ip IK, Gershanik EF, Schneider LI *et al.* Impact of IT-enabled intervention on mri use for back pain. *Am J Med.* 2014;127(6):512–518.
6. Evans RS, Pestotnik SL, Classen DC *et al.* A computer-assisted management program for antibiotics and other antiinfective agents. *N Engl J Med.* 1998;338(4):232–238.
7. Kim SH, Cho SH. Educational and decision-support tools for asthma-management guideline implementation. *Asia Pac Allergy.* 2012;2(1):26–34.
8. Kucher N, Koo S, Quiroz R *et al.* Electronic alerts to prevent venous thromboembolism among hospitalized patients. *N Engl J Med.* 2005;352(10):969–977.
9. Thompson G, O'Horo JC, Pickering BW, Herasevich V. Impact of the electronic medical record on mortality, length of stay, and cost in the hospital and ICU: a systematic review and metaanalysis. *Crit Care Med.* 2015;43(6):1276–1282.
10. Singh H, Thomas EJ, Mani S *et al.* Timely follow-up of abnormal diagnostic imaging test results in an outpatient setting: are electronic medical records achieving their potential? *Arch Intern Med.* 2009;169(17):1578–1586.
11. Medicare and Medicaid programs; electronic health record incentive program–stage 2. Final rule. *Fed Regist.* 2012;77(171):53967–54162.
12. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ.* 2005;330(7494):765.
13. Johnston ME, Langton KB, Haynes RB, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. A critical appraisal of research. *Ann Intern Med.* 1994;120(2):135–142.
14. Langton KB, Johnston ME, Haynes RB, Mathieu A. A critical appraisal of the literature on the effects of computer-based clinical decision support

systems on clinician performance and patient outcomes. *Proc Annu Symp Comput Appl Med Care.* 1992;626–630.
15. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *JAMA.* 1998;280(15):1339–1346.
16. PUBLIC LAW 113 - 93 - PROTECTING ACCESS TO MEDICARE ACT OF 2014. CONGRESSIONAL RECORD, Vol.160 (2014) Public Law 113-93. 4-1-2014.
17. Archambault PM, Turgeon AF, Witteman HO *et al.* Implementation and evaluation of a Wiki involving multiple stakeholders including patients in the promotion of best practices in trauma care: the WikiTrauma Interrupted Time Series Protocol. *JMIR Res Protoc.* 2015;4(1):e21.
18. Djulbegovic B, Kumar A, Kaufman RM *et al.* Quality of evidence is a key determinant for making a strong guidelines recommendation. *J Clin Epidemiol.* 2015;68(7):727–732.
19. Li Y, Kong N, Lawley M *et al.* Advancing the use of evidence-based decision-making in local health departments with systems science methodologies. *Am J Public Health.* 2015;105 (Suppl 2):S217–S222.
20. Peterson JF, Bates DW. Preventable medication errors: identifying and eliminating serious drug interactions. *J Am Pharm Assoc.* 2001;41(2):159–160.
21. MacMahon H, Austin JH, Gamsu G *et al.* Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society. *Radiology.* 2005;237(2):395–400.
22. Silverman SG, Israel GM, Herts BR, Richie JP. Management of the incidental renal mass. *Radiology.* 2008;249(1):16–31.
23. Stiell IG, Clement CM, Rowe BH *et al.* Comparison of the Canadian CT Head Rule and the New Orleans Criteria in patients with minor head injury. *JAMA.* 2005;294(12):1511–1518.
24. Stiell IG, Lesiuk H, Wells GA *et al.* The Canadian CT Head Rule Study for patients with minor head injury: rationale, objectives, and methodology for phase I (derivation). *Ann Emerg Med.* 2001;38(2):160–169.
25. Stiell IG, Greenberg GH, McKnight RD, Wells GA. Ottawa ankle rules for radiography of acute injuries. *N Z Med J.* 1995;108(996):111.
26. Stiell IG, Greenberg GH, McKnight RD *et al.* Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. *JAMA.* 1993;269(9):1127–1132.
27. American College of Radiology. *ACR Practice Guideline for Communication of Diagnostic Imaging Findings.* Reston, VA; 2005.
28. Baker ME, Nelson RC, Rosen MP *et al.* ACR Appropriateness Criteria(R) acute pancreatitis. *Ultrasound Q.* 2014;30(4):267–273.
29. Henry TS, Kirsch J, Kanne JP *et al.* ACR Appropriateness Criteria(R) rib fractures. *J Thorac Imaging.* 2014;29(6):364–366.
30. Mosher TJ, Kransdorf MJ, Adler R *et al.* ACR Appropriateness Criteria Acute Trauma to the Ankle. *J Am Coll Radiol.* 2015;12(3):221–227.
31. Moy L, Newell MS, Mahoney MC *et al.* ACR Appropriateness Criteria stage I breast cancer: initial workup and surveillance for local recurrence and distant metastases in asymptomatic women. *J Am Coll Radiol.* 2014;11(12 Pt A):1160–1168.
32. Klem ML, Weiss PM. Evidence-based resources and the role of librarians in developing evidence-based practice curricula. *J Prof Nurs.* 2005;21(6):380–387.
33. Morrison RS, Krishnamurthy M. Customized library tutorial for online BSN students. *Nurse Educator.* 2008;33(1):18–21.
34. Robinson L, Hilger-Ellis J, Osborne L *et al.* Healthcare librarians and learner support: a review of competences and methods. *Health Info Libr J.* 2005;22 (Suppl 2):42–50.
35. Harvard Medical School. Harvard Medical School E-Library of Evidence. http://libraryofevidence.med.harvard.edu/. Accessed October 14, 2015.
36. Shwe M, Sujansky W, Middleton B. Reuse of knowledge represented in the Arden syntax. *Proc Annu Symp Comput Appl Med Care.* 1992;47–51.
37. Atkins D, Briss PA, Eccles M *et al.* Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. *BMC Health Serv Res.* 2005;5(1):25.
38. Gopalakrishna G, Mustafa RA, Davenport C *et al.* Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. *J Clin Epidemiol.* 2014;67(7): 760–768.
39. Saver BG. A piece of my mind. Should C be a passing grade for the USPSTF? *JAMA.* 2015;313(5):465–466.
40. Barton MB, Miller T, Wolff T *et al.* How to read the new recommendation statement: methods update from the U.S. Preventive Services Task Force. *Ann Intern Med.* 2007;147(2):123–127.
41. The U.S. Preventive Services Task Force. Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med.* 2009;151(10):716–236.
42. Berkman ND, Lohr KN, Ansari MT *et al.* Grading the strength of a body of evidence when assessing health care interventions: an EPC update. *J Clin Epidemiol.* 2015;68(11):1312–1324.
43. Berkman ND, Lohr KN, Ansari M *et al.* Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update. Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008-2013 Nov 18.
44. Oxford Centre for Evidence Based Medicine. Oxford Centre for Evidence-based Medicine-Levels of Evidence (March 2009). http://www.cebm.net/ocebm-levels-of-evidence/. 2015. Accessed March 15, 2015.
45. Harris RP, Helfand M, Woolf SH *et al.* Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med.* 2001;20 (Suppl 3):21–35.
46. Qureshi AI. A new scheme for grading the quality of scientific reports that evaluate imaging modalities for cerebrovascular diseases. *Med Sci Monit.* 2007;13(10):RA181–RA187.
47. Wise JN, Daffner RH, Weissman BN *et al.* ACR Appropriateness Criteria(R) on acute shoulder pain. *J Am Coll Radiol.* 2011;8(9):602–609.
48. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–174.
49. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull.* 1968;70(4):213–220.
50. Hripcsak G. Writing Arden Syntax Medical Logic Modules. *Comput Biol Med.* 1994;24(5):331–363.
51. Hripcsak G. Arden Syntax for Medical Logic Modules. *MD Comput.* 1991;8(2):76, 78.
52. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *Int J Prosthodont.* 2006;19(2):126–128.
53. Davino-Ramaya C, Krause LK, Robbins CW *et al.* Transparency matters: Kaiser Permanente's National Guideline Program methodological processes. *Perm J.* 2012;16(1):55–62.
54. Eikermann M, Holzmann N, Siering U, Ruther A. Tools for assessing the content of guidelines are needed to enable their effective use–a systematic comparison. *BMC Res Notes.* 2014;7:853.

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