A Benefit-Risk Assessment Framework for Development of Clinical Guidelines in Diagnostic Radiology

Maria Agapova

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Emily E. Devine, Chair
Louis P. Garrison
Larry Kessler

Program Authorized to Offer Degree:

Pharmaceutical Sciences

TABLE OF CONTENTS

University of Washington

Abstract

A Benefit-Risk Assessment Framework for Development of Clinical Guidelines in Diagnostic Radiology

Maria Agapova

Chair of the Supervisory Committee:

Associate Professor, Emily E. Devine

Department of Pharmacy

Background

The body of this dissertation focuses on benefit-risk assessment in diagnostic radiology guideline development. We look specifically at structuring and making more transparent the role of expert consensus in evidence-based practice guidelines like those developed by the American College of Radiology (ACR).

While several frameworks for structured benefit-risk assessment for pharmaceutical products are available, their applicability in guiding guideline development of diagnostic imaging is not well characterized.[2] The most prominent are the (BRAT) Benefit-Risk Action

Team and ProACT-URL (Problem, Outcomes, Alternatives, Consequence and Tradeoffs Uncertainty Risk and Linked decisions) frameworks. The six steps of the BRAT framework include, (1) defining the decision context (selecting stakeholder perspective and time horizon); (2) identifying benefit and risk outcomes visually with use of a value tree (one branch exploding all benefits and the other, all harms); (3) identifying data sources and creating a data source table; (4) tuning or customizing the framework, i.e. aligning outcomes in value tree criteria with outcomes for which data are available; (5) assessing the importance or weight of each outcome under consideration; (6) and displaying or communicating the metrics using visual and tabular formats. Likewise, the PrOACT-URL decision-making framework comprises five core elements (Problem, Outcomes, Alternatives, Consequences and Tradeoffs) and three elements relevant to evolving or volatile settings: Uncertainty, Risk attitude, and Linked decisions.[3]

The ACR expert panels follow a structured process in developing appropriateness criteria (AC) but do not employ a benefit-risk framework.[4-7] In many cases benefit-risk assessment is limited to comparing the diagnostic accuracy of a test against test procedural risks (e.g., exposure to radiation, invasiveness). Unlike endpoints of pharmaceutical product trials, the endpoint of diagnostic accuracy does not easily fit within the benefit-risk framework and needs to be translated into other endpoints. Endpoints that are easily identified as benefits or risks — effects of test information on the provider, on patient management, and on the patient — are rarely known. It is for this reason that the ACR AC are developed using a combined approach, relying in part on the body of evidence, and in part on expert consensus.[8] However, the ACR AC lack key framework elements that ensure the transparent contribution of expert consensus: making explicit relevant benefits and risks, their prioritization and respective data sources or lack thereof.

Independently, several frameworks for assessing the value of diagnostic imaging are available.[9-12] These frameworks are of limited value in this setting as few new studies are

conducted specifically to inform clinical guideline development. At the same time, there are calls for expanding the way imaging tests are valued.[13] Lee *et al.* proposed a three dimensional value framework (medical, psychic, and planning) for health technology assessment, proposing that the value of diagnostic tests may be underestimated in cost-effectiveness analyses that limit the scope to medical impacts.[9] Staub *et al.* promote inclusion of patient management measures as proxies for patient health outcomes.[14,15] Otero *et al*. discuss the lack of inclusion in cost-utility analyses of intrinsic value elements (non-clinical impacts of test information), similar to the Lee *et al.* psychic and planning dimensions but incorporating both provider and patient intrinsic value.[16] Bossuyt and McCaffery present a framework incorporating dimensions independent of clinical outcome (emotional, social, cognition, behavior) with incomplete overlap with Otero's and Lee's.[17]

Chapter 1. Toward a Framework for Benefit Risk Assessment in Diagnostic Imaging: Identifying Scenario-Specific Criteria

In the first paper, we work toward creating a unified framework incorporating elements of structured benefit-risk and elements of diagnostic imaging frameworks. To address the question of clearly defined criteria, we abstract from the literature measures of diagnostic imaging value and translate these into benefit-risk criteria (BRC). To further refine the BRC, we cross-reference our literature findings by surveying radiologist and non-radiologist perceptions of the benefits and risks in diagnostic radiology. Within the survey, we operationalize the initial broad list of BRC across four clinical use case scenarios. For each use case, we compare BRC selections between radiologists and non-radiologists. We arrive at thirty six criteria, organized into three domains: 1) those that account for differences among tests, attributable only to the test or device (n=17); 2) those that account for clinical management and provider experience effects (n=12); and 3) and those that measure distant, less direct effects of imaging tests on patients (n=7). Our results suggest that radiologist considerations do not dramatically differ from those of non-radiologists but

the addition of non-radiologist selections may help guideline developers reach the goal of a broader set of priorities. These results can inform future hypotheses and studies of effects of increased clinical diversity on guideline quality and adoption.

Chapter 2. A Proposed Approach for Quantitative Benefit-Risk Assessment in Diagnostic Radiology Guideline Development: The American College of Radiology Appropriateness Criteria Example

In the second paper, we continue building a framework for benefit-risk assessment with a critical appraisal of quantitative benefit-risk assessment (QBRA) methodology. As there is only limited guidance on method selection and it is not clear whether these methods are well suited to the clinical guideline development process, further exploration of the potential for benefit-risk methodology to meet the needs of the ACR AC process is warranted.[18-20] Thus, we review the benefit-risk methodology literature and propose several steps for selection of comparators and criteria. These steps include investigation of weak evidence and disagreement. We identify a set of benefit-risk methods addressing one or more of these needs and build a decision aid for selecting among these methods. Our results suggest there is opportunity to use multi-criteria decision analysis and incremental net health benefit methods for some decision problems the ACR faces when creating AC ratings. The process leading to the decision aid facilitates transparent contribution of expert opinion to ratings. Since the structure of the decision aid is based on clinicians' input but also requires the skills of methods experts, the decision aid represents a key component to uniting clinical experts and methodologists.

Chapter 3. Analytic Hierarchy Process for Prioritizing Imaging Tests in Diagnosis of Suspected Appendicitis

In the third paper, to complement inclusion of QBRA in the framework, we evaluate empirically one of the QBRA methods for use in guideline development. There is prevailing

skepticism that regulatory approval and clinical guideline decision-making is far too complicated, and too multi-dimensional for quantitative methods.[21] To investigate this position further, we compare the results of a multi-criteria decision analysis approach, analytic hierarchy process, evaluating computed tomography against magnetic resonance imaging and ultrasound for classic presentation of suspected appendicitis, to the ACR AC ratings. We ask those who participated whether the process is manageable, transparent, and improves shared-understanding of the decision problem. This is, to our knowledge, the first study to show that a quantitative method produces comparable results to ratings of ACR AC guidelines and that this QBRA was found, among study participants, to facilitate shared understanding and transparency.

Conclusions

Structured benefit-risk assessment promises to improve the transparency of the contribution of expert consensus to clinical guideline development in diagnostic radiology. In this body of work, we propose a step-wise process resembling existing benefit-risk frameworks, but tailored to the unique needs of diagnostic radiology and demonstrate the feasibility of these recommendations. Specifically, we provide a systematic approach to selection of benefits and risks in evaluating the value of diagnostic imaging. We show that diversity of participants in selection of BRC expands the decision problem and suggests that a comprehensive definition of value is best accomplished using a multi-disciplinary perspective. Likewise, we reduce the number of QBRA for ACR to consider and provide guidance for when and how to use QBRA. Lastly, we demonstrate that MCDA can add transparency to a process like the ACR AC guideline development. Future work will entail field testing the BRC, the decision aid for QBRA selection as well as the QBRA in a setting like the ACR AC.

Acknowledgements

Beth Devine deserves the highest praise for her enduring commitment to her students and especially for ensuring I succeed. The Pharmaceutical Research Outcomes and Policy Program is very fortunate to have a selfless but driven faculty member like Beth. Thank you for teaching me to have confidence in myself and in my ideas, to hold my head a little higher and walk a little brisker. I also want to recognize the high caliber expertise and grace of my committee members; for their willingness to share their knowledge with me and for their patience during frustrating moments. Thank you for continuously challenging me to improve as a researcher.

The colleagues at University of Twente, Maarten Izjerman and Henk Broekhuizen, shared their expertise in multi-criteria decision analysis with me and invaluable guidance for running a meeting with stakeholders. Thank you Lotte Stueten for connecting us. Key interactions with other experts such as David Kurth, at American College of Radiology and James Rawson at Georgia Regents University enriched my understanding of my research. Thank you to the University of Washington physicians, Matthew Thompson, Amber Sabbatini, Laura Mae Baldwin, Jerry Jarvik and Ken Linnau, for your time, honesty and for winning the interest of your colleagues to participate in my research activities.

I thank my colleagues and friends, Caroline, Katharine, Jean, Will, Justin, Richard and Floyd, for their time, engagement and appraisal of my work. Caroline, Katie, Preeti and Julia, I am so grateful that you always found time to listen.

My training in research began with a person I'll continue to emulate my entire career, my mentor at Blood Systems Research Institute, Dr. Custer. Dr. Custer shared his passion for improving policy and related to me as though I were already a seasoned researcher giving me opportunities to make my first decisions as a researcher. Thank you, Brian.

Michael, my husband, deserves my infinite gratitude for his endless supply of patience and enduring faith in me. Thank you for not allowing me to give up and for taking such good care of our baby girl so I could work. I will always cherish the nights you stayed up with me, programming algorithms and proofreading until your eyes watered. You brought into my life your parents—tireless cheerleaders and babysitting superstars.

Before I was lucky to encounter these inspirational people, there was only one, my mother. She has either suffered or rejoiced my every moment starting before I could talk and long after she could no longer understand what I write. Thank you, mom, for showing me what it means to have grit.

References

1. Cascade PN. The American College of Radiology. ACR Appropriateness Criteria project. *Radiology.* Jan 2000;214 Suppl:3-46.
2. Coplan PM, Noel RA, Levitan BS, Ferguson J, Mussen F. Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. *Clin Pharmacol Ther.* Feb 2011;89(2):312-315.
3. Hammond JS, Keeney RL, Raiffa H. *Smart Choices: A Practical Guide to Making Better Decisions*. Boston, MA: Harvard University Press; 1999.
4. American College of Radiology. Appropriateness Criteria. 2015; https://acsearch.acr.org/list. Accessed September 2, 2015.
5. ACR Appropriateness Criteria. ACR Appropriateness Criteria® Organization and Composition of Expert Panels. 2015; http://www.acr.org/~/media/ACR/Documents/AppCriteria/ETDevDiagnostic.pdf. Accessed September 4, 2015.
6. ACR Appropriateness Criteria. ACR ACR Appropriateness Criteria® Evidence Table Development — Diagnostic Studies. 2013; http://www.acr.org/~/media/ACR/Documents/AppCriteria/ETDevDiagnostic.pdf. Accessed September 4, 2015.
7. ACR Appropriateness Criteria. ACR ACR Appropriateness Criteria® Rating Round Information. 2015; http://www.acr.org/~/media/ACR/Documents/AppCriteria/RatingRoundInfo.pdf. Accessed August 4, 2015.
8. Bode FR, Cascade PN. The American College of Radiology appropriateness criteria. Will they be useful for us? *Chest.* Oct 1996;110(4):869-871.
9. Lee DW, Neumann PJ, Rizzo JA. Understanding the medical and nonmedical value of diagnostic testing. *Value Health.* Mar-Apr 2010;13(2):310-314.

10. Brook RH. *The RAND/UCLA Appropriateness Method.* Rockville, MD: Public Health Service, U.S. Department of Health and Human Services;1994.
11. Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care.* 1986;2(1):53-63.
12. Gazelle GS, Kessler L, Lee DW, et al. A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research. *Radiology.* Dec 2011;261(3):692-698.
13. Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. *Clinical chemistry.* Dec 2012;58(12):1636-1643.
14. Staub LP, Dyer S, Lord SJ, Simes RJ. Linking the evidence: intermediate outcomes in medical test assessments. *Int J Technol Assess Health Care.* Jan 2012;28(1):52-58.
15. Staub LP, Lord SJ, Simes RJ, et al. Using patient management as a surrogate for patient health outcomes in diagnostic test evaluation. *BMC Med Res Methodol.* 2012;12:12.
16. Otero HJ, Fang CH, Sekar M, Ward RJ, Neumann PJ. Accuracy, risk and the intrinsic value of diagnostic imaging: a review of the cost-utility literature. *Acad Radiol.* May 2012;19(5):599-606.
17. Bossuyt PMM, McCaffery K. Additional Patient Outcomes and Pathways in Evaluations of Testing. *Medical Tests-White Paper Series*. Rockville (MD)2009.
18. European Medicine Agency (EMEA) CHMP. Benefit-risk methodology project: Work Package 2 report: Applicability of current tools and processes for regulatory benefit-risk assessment. 2010; http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/document_listing/document_listing_000314.jsp&mid=WC0b01ac0580223ed6#section2. Accessed October 22, 2012.

19. Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch DW. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ISPOR risk-benefit management working group. *Value Health.* Aug 2010;13(5):657-666.
20. Mussen F, Salek S, Walker S. A quantitative approach to benefit-risk assessment of medicines - part 1: the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiol Drug Saf.* Jul 2007;16 Suppl 1:S2-S15.
21. *Structured Approach to Benefit-Risk Assessment in Drug Regulatory Decision-Making: Draft PDUFA V Implementation Plan.* Rockville, MD: FDA;2013.

Chapter 1. Toward a Framework for Benefit Risk Assessment in Diagnostic Imaging: Identifying Scenario-Specific Criteria

Maria Agapova, Brian W. Bresnahan, Ken Linnau, Louis Garrison, Mitchell Higashi,

Larry Kessler, Beth Devine

Target Journal: Academic Radiology

Word Count: 3,293

Tables: 5

Figures: 1

ABSTRACT

Abstract Word Count: 251 (including headings)

INTRODUCTION

It is unclear whether recent calls for expanding the way imaging tests are valued have reached clinical guideline development as the trade-offs in clinical guidelines are rarely made explicit. We describe initial steps toward the creation of a benefit-risk framework for diagnostic radiology.

METHODS

After a literature search to identify and collect risks and benefits relevant to diagnostic imaging tests, we performed an online survey of physicians to select these trade-off criteria. We operationalized the initial broad list of BRC across four clinical use case scenarios that vary in clinical decision context. The selected BRC were compared across clinical scenarios and between radiologists and non-radiologists.

RESULTS

Thirty-six BRC were identified in the literature and organized into three domains: 1) those that account for differences among tests, attributable only to the test or device (n=17); 2) those that account for clinical management and provider experience effects (n=12); and 3) and those that measure distant, less direct effects of imaging tests on patients (n=7). Participants selected twenty-two criteria from the initial list in the survey (9 -11 per case). Selected BRC differed only between fundamentally different clinical scenarios, and were similar within comparable cases and between radiologists and non-radiologists.

CONCLUSIONS

The results of our study suggest that there are differences in BRC across different clinical scenarios, and that imaging tests can be judged by more than diagnostic accuracy or potential for ionizing-radiation exposure. We propose a BRC list to ensure consistent examination of these differences among tests, even when evidence is sparse.

INTRODUCTION

Several frameworks for assessing the value of diagnostic imaging have been proposed.[1-4] There have also been calls for expanding the way imaging tests are valued.[5] Lee *et al.* proposed a three dimensional value framework (medical, psychic, and planning) for health technology assessment, proposing that the value of diagnostic tests may be underestimated in cost-effectiveness analyses that limit the scope to medical impacts.[1] Staub *et al.* promote inclusion of patient management measures as proxies for patient health outcomes.[6,7] Otero *et al*. discuss the lack of inclusion in cost-utility analyses of intrinsic value elements (non-clinical impacts of test information), similar to the Lee *et al.* psychic and planning dimensions but incorporating both provider and patient intrinsic value.[8] Bossuyt and McCaffery present a framework incorporating dimensions independent of clinical outcome (emotional, social, cognition, behavior) with incomplete overlap with Otero's and Lee's.[9]

Thus, the full spectrum of imaging test benefits and risks may include clinical, non-clinical, direct or indirect effects. Direct effects refer to those related to only the procedure (e.g., ionizing radiation exposure). When imaging tests are necessary but not sufficient to explain the benefit or risk, those effects may be labeled indirect (e.g. cancer-free survival, or consequences of incidental findings). While clinical outcomes may be captured in medical records or claims databases, non-clinical effects refer to the less documented cognitive, psychological, legal, and behavioral effects of the information provided by the test.[8] Since providers are the consumers of the information and the recipients of the bulk of indirect

effects of imaging tests, effects of imaging tests on providers may also be relevant.[10,11] To gain market approval, manufacturers of imaging tests are not required to capture the full spectrum of these effects and no economic incentives exist to collect evidence to inform clinical guidelines.[12,13]

In 2012, The US Food and Drug Administration (FDA) published a benefit-risk draft guidance for regulatory pre-market approval and *de novo* classification relevant to all medical devices.[14] The FDA recommended identifying benefits and risks and characterizing their magnitude, probability and duration; the number and aggregate effect of harmful events, uncertainty, disease characterization and patient tolerance and perspective on benefits. Revision of this draft guidance was released in 2015 with specific guidance for inclusion of patient preferences in benefit-risk assessments.[15] However, few imaging tests pass through these regulatory channels and consequently few are studied to this extent. More importantly, the frameworks provided are tailored to the regulatory, not clinical practice, setting.

In diagnostic radiology, appropriateness—defined by the American College of Radiology (ACR) as benefits outweighing risks at the population level—is currently assessed for imaging tests by panels, predominantly composed of radiologist experts. Tt is not clear whether ACR guidelines integrate the breadth of benefits and risks discussed. It is also not clear how selection of relevant benefits and risks may differ if as many non-radiologists as radiologists participated in the guideline development.

In this study, we work toward creating a unified framework incorporating elements of structured benefit-risk and elements of diagnostic imaging frameworks. We propose a comprehensive list of benefit and risk criteria (BRC) for use in diagnostic radiology guideline development. Using clinical scenarios, we investigate which trade-offs physicians consider, and whether those trade-offs differ by clinical scenario and clinical specialty.

METHODS

We conducted a literature search to identify and collect BRC relevant to the selection of diagnostic imaging tests. Using thematic synthesis, we organized and created themes from the data abstracted.[16] Starting themes used were the six levels of evidence first described in the Fryback and Thornbury hierarchical model of efficacy.[17] To ensure that the list of BRC abstracted from the literature was comprehensive and exhaustive, we administered an online survey to four groups of physicians. In the survey, we asked physician respondents to indicate the importance and frequency with which each criterion entered their decision-making process. Next, physicians were asked to select relevant BRC for specific clinical use cases and provide any additional criteria they felt were important.

Literature search

We performed a broad PubMed literature search for imaging, diagnostics and device effectiveness and safety measures using MeSH first-ordered terms: radiology, diagnostic imaging, diagnostic tests, and devices/diagnostic equipment, limiting by subheadings and combining with the following second-order terms: outcome, benefit, risk, quality of life, anxiety, decision-making, survival, incidental findings or adverse reactions. The search was limited to English language publications with an abstract, published from January 1st, 2011 thru December 31st, 2013. We excluded study designs that were not categorized as randomized clinical trials, cost-effectiveness, observational and meta-analysis studies or literature reviews. If literature reviews were located, we identified additional articles from the reference lists and also used Web of Science to search for any studies that may have cited the literature reviews.

Creation and design of survey

The list of BRC identified from the literature review was used as a start list. These criteria were then placed into a survey used to solicit additional BRC and arrive at a more comprehensive list. Pre-testing with clinicians guided development of survey structure, content, and choice of language.

We structured the survey as pairwise comparisons of imaging strategies. Participants were asked to consider each BRC independently and to select either the better strategy, or select that no strategy would be superior. For example, respondents selected computed tomography (CT) if CT offers an advantage over ultrasound (US) with respect to therapeutic success, US if the reverse was true, or marked "not applicable" if there was no difference between tests with respect to therapeutic success. Throughout the survey, text fields prompted respondents to provide additional BRC.

Selection of use cases

We selected clinical diagnoses and imaging tests to reflect both similarities and differences in several dimensions. First, we hypothesized that absolute appropriateness of a decision to test (e.g., whether or not to perform an MRI for low back pain) is conceptually different from relative appropriateness, choosing which test to use (e.g., CT or US for suspected appendicitis). To observe whether any differences in the use of the framework would be noted between the absolute and relative decision contexts, we selected use cases that represented each concept. Second, we reasoned that benefit-risk profiles may differ in some systematic way between technologies known to be overutilized (e.g., MRI for low back pain) [11] and those that are potentially underutilized (e.g., US for nephrolithiasis).[18,19] Third, it was of interest to us to explore benefit-risk profiles using examples of both chronic, non-specific disease, and acute, specific disease diagnoses. Case definitions were provided by the ACR AC (Table 1).

[Insert Table 1 here]

Survey respondent identification and recruitment

We sought University of Washington respondents from a targeted range of clinical specialties and backgrounds, including those with specialty radiology training and those without. However, radiologists rarely order imaging tests and interface with patients, thus we chose two additional groups of physicians, in primary care and emergency medicine specialties. Neuroradiologists and Emergency Department (ED) radiologists were selected as the gold standard for knowing the technical aspects of imaging tests. A key physician contact in each clinical specialty facilitated recruitment of colleague respondents. The University of Washington Institutional Review Board approval and respondent informed consent were obtained for the survey study.

Selection of BRC for specific use cases

We investigated the selection of BRC and the breadth of selections across four clinical use cases: lower back pain, chronic headache, lower quadrant pain-suspected appendicitis and acute onset flank pain-suspicion of stone disease. We also assessed consensus levels among survey respondents, and the inter- and intra-agreement within and across radiology and non-radiology specialties. Comparisons of criteria selections were made by clinical scenario and between radiologists and non-radiologists using nonparametric Wilcoxon signed-rank tests. Analyses were conducted using STATA, Version 9 (College Station, TX).

RESULTS

Literature search

After exclusion of non-relevant articles, review of the recent literature yielded one hundred and twenty nine studies that measured benefit and risk outcomes of devices and diagnostics

(Appendix Figure A1). A large proportion of studies (n=45; 35%) included one or more measures of diagnostic accuracy as the only measure of test performance. However, among remaining studies, a diverse list of criteria were assembled, and collapsed into thirty-three BRC. These were initially grouped according to the Fryback and Thornbury hierarchy model of efficacy (Appendix Table 1).[17]

Online survey responses

Forty eight University of Washington physicians participated in the survey about BRC of diagnostic imaging tests: primary care physicians (N=12), emergency physicians (N=16), neuroradiologists (N=8) and emergency radiologists (N=9). Responses to the survey added to the list, criteria not found in the literature search and pointed to the need for streamlining the definition of each criterion.

The finalized BRC domains and criteria list

We created three domains to fully represent the criteria that characterize benefits and risks of imaging tests in diagnostic radiology: 1) test specific features; 2) provider intrinsic value and patient management; and 3) patient intrinsic value and outcomes. Like the Fryback and Thornbury model, the three domains fall into an evidence hierarchy. The final lists of BRC, by domain, are provided in Tables 2 thru 4.

Test Specific Features (n=17)

Test specifications are not commonly thought of as potential benefits or risks. However, test-specific features are the basis for differentiating among tests and are also likeliest to have information available. These criteria adequately differentiate tests so that even two very similar tests can be differentiated based on these criteria.

[Insert Table 2 here]

Patient Management and Provider Intrinsic Value (n=12)

This group of benefits and risks are not exclusively influenced by imaging tests. These criteria represent either the outcomes influenced by complex interactions between the test information and the physician's judgment or the physician's intrinsic value, that is, the physician's perceived gains or losses from ordering the test. Criteria in this domain may overlap with intermediate patient outcomes, such as complications from a medical procedure. However, BRC in this domain are from the perspective of the clinician and the extent of these effects is assumed to cover the time the patient is under the care of the physician. High quality evidence of these benefits or risks, attributable to an imaging test, is unlikely to be available.

[Insert Table 3 here]

Patient Intrinsic Value and Outcomes (n=7)

The direct and indirect, clinical and non-clinical effects of imaging tests on patients are grouped in this domain. The criteria in this domain take the perspective of the patient. Clinical sequelae from an intervention are expressed broadly as effects on length and quality of life. When comparing two imaging tests, large incremental effects are likelier to exist among intrinsic criteria (e.g. comfort or burden) than on length and quality of life. The exception is the potentially large differential between tests in ionizing radiation dose, which makes relevant the clinical, indirect and distant outcome, ionizing radiation-induced cancers. Evidence of effects on survival and quality of life attributable to an imaging test is unlikely to be available in the published literature. Unknown differences in imaging test intrinsic value to patients may be traced and partially attributed to known differences in test-specific features.

[Insert Table 4 here]

Additional considerations

Survey respondents brought forth two considerations that may influence imaging utilization: healthcare idiosyncrasies and costs. Survey respondents identified healthcare workflow idiosyncrasies and sharing of decision-making authority as additional drivers of utilization. There are times when decision-making and thus, weighing of benefits and risks, is transferred fully or partially from the ordering physician to a third party. For example, a respondent stated, "A surgeon may require an imaging test before a patient is able to proceed to surgery or other intervention". An insurance carrier or a radiology benefit manager may require a sequence of other imaging tests before the test is regarded as optimum by the physician, and the patient becomes eligible for reimbursement. A survey respondent added that decision-making may be based on "the recommendation of a consultant or the radiologist".

Traditionally, costs are treated separately from benefits and risks. Survey respondents were asked whether insurance coverage and out-of-pocket costs impacted the decision to order tests. Forty-six percent of respondents indicated that insurance coverage was an important consideration. Of all entered text, the word "cost" appeared the most frequently (six respondents added costs to the BRC list). An emergency physician wrote, "Cost is a daily question". An emergency radiologist added to the survey response, "total healthcare costs." In response, costs from the perspective of the provider (weighing the financial risk of ordering a test that may not be covered by insurance) and the patient (the out-of-pocket risk of the copay, the co-insurance or the full payment associated with the test) were added to the BRC list.

Specific use cases

ED physicians and radiologists reached 80% consensus on nine BRC for each diagnostic use case they were presented; PCPs and neuroradiologists, on eleven BRC. These criteria are

listed in Table 5, grouped by use case and by radiologist and non-radiologist standing. Percent consensus is provided in appendix table A1 by clinical specialty. Twenty-two criteria were selected. Like cases had similar selections of criteria with slight overlap between the pairs of use cases. In all use cases, radiologists and non-radiologists selected missed cases, provider utility and potential confirmatory testing as relevant BRC.

[Insert Table 5 here]

Comparing percent agreement for relevance of each BRC within and between clinical specialties, we did not observe unanimous selection and unanimous non-selection of the criteria as would have been depicted by the clustering of points at the top right and left bottom corners of each plot in Figure 1. The observed dispersion pattern suggests lack of intra-group consensus on the relevance of each BRC. Comparing dissimilar clinical scenarios, we found significantly different selections, by Wilcoxon signed rank test: low back pain and chronic headache were compared to suspected appendicitis ($p=0.019$ and $p=0.022$), and suspected kidney stone ($p=0.008$ and $p=0.008$), respectively. At 80% consensus, non-radiologists selected on average, a higher number of criteria (n=12) than non-radiologists (n=9). Within clinical scenarios, using the Wilcoxon signed rank test, we did not see patterns of inter-group discordance between radiologists and non-radiologists (p-values not significant).

In the pair of use cases that compared a test to no test (low back pain and chronic headache), respondents were less likely to exclude criteria as visualized in the left lower corners of each plot in Figure 1. For example at 40% or lower consensus, fewer criteria were selected in the lowback (n=4) and chronic headache (n=5) use cases when compared tosuspected appendicitis (n=11) and kidney stone disease (n=11) use cases.

DISCUSSION

Using recent literature and surveys of physicians we found many effects of diagnostic imaging that may qualify as BRC. We observed that in both the literature and in qualitative responses there was a great deal of heterogeneity of terms and definitions. This highlighted the need for standardization of definitions for benefits and risks in diagnostic imaging.

We identified a class of criteria specific to imaging tests not previously considered in the literature, test-specific features. Identifying differences in test-specific features is especially useful if evidence for intermediate and long-term benefits and risks is sparse, of low quality, or missing, or if it is not possible to separate test-attributable effects from patient, disease, healthcare system and provider factors. Unlike intermediate and long-term outcomes, test-specific feature information is less influenced by practice variation and more readily known. If decision-makers identify ambiguities in judging technologies for a given criterion, it may be possible to decompose and define the criterion in terms of variation in one or more related test-specific criteria. Additional stratification of a criterion, redefinition and other modifications may be needed. For example, depth/breadth of visualization may divided further into mass shape, vascularization, region size.[20] Thus, the BRC are to be treated as the starting point for taking inventory of the differences among pairs of diagnostic strategies.

Selections of BRC differed among dissimilar and less so within similar pairs of clinical scenarios. The number of criteria selected was slightly higher for the chronic headache and lower back pain clinical scenarios, and consensus patterns suggest responders had a more difficult time excluding criteria in these two cases. These results point to the potential of group selection among BRC as a mechanism for describing the decision problem. Although at the 80% consensus level radiologists selected fewer BRC, their patterns of consensus did not differ from those of non-radiologists. These results suggest that radiologists and non-

radiologists generally agree about the relevant BRC but that non-radiologists perceive additional differences among testing strategies, primarily those related to patient management.

There were several limitations to our study. Our piloting of the BRC list involved a small number of use cases and small samples sizes of respondents from one institution. This limits the generalizability of the results. As patients are rarely aware of the full range of potential risks associated with medical interventions elicitation of novel BRC from patients was expected to have low yield. [21] Patient participation in selection of benefits and risks in future operationalization of the BRC will be important in ensuring the selection of benefits and risks is representative.

By creating a universal BRC list, we made some assumptions. The first is that the list is comprehensive and exhaustive. Although this assumption is likely not to hold, it is important to note that each BRC may be re-defined and shaped to fit the demands of a specific imaging test benefit-risk comparison. The second assumption is that no BRC are redundant or extraneous. The usefulness of this framework diminishes if only a subset of criteria is consistently relevant or if the ACR AC continues to exclude costs, ease of use and access to test from consideration.[22] Our results suggest costs are on the mind of physicians. This shift in awareness may be due to the movement by clinical societies, like the American Society of Clinical Oncology, to include cost-effectiveness in guideline development.[23] Further, research is needed to ascertain whether other criteria we included but the ACR exclude, ease of use and access to test, belong in clinical guideline development.

We also assumed the criteria that received the most votes ($\geq$80%) represented the most important differences between any two tests and that the observed variation in consensus represents uncertainty. Even if these assumptions do not hold there is value in performing the exercise of BRC selection from a set list. The group, not one or two individuals, select

criteria that enter the decision-making space. Also, by measuring variation in agreement, decision-makers are alerted to criteria that may require additional attention, reducing the chance that decision-makers implicitly judge imaging tests on criteria explicitly omitted from consideration.

We explored two use cases comparing a radiology procedure with no procedure and found that respondents had a difficult time excluding criteria. For example, in such comparison, while by definition, differences in test-specific features exist, selection of criteria from this domain may be less meaningful. However, as the list does include criteria relevant to absolute appropriateness (e.g., value of knowing), we showed that it is possible to use the BRC list, albeit less optimally, in this context.

While, to our knowledge, this is the first effort at standardizing benefits and risks in diagnostic imaging, there are several complementary guidance documents available. These include standards for reporting diagnostic accuracy as well as several frameworks for guiding outcomes selection for clinical trials.[4,5,24,25]

Direct and indirect effects of diagnostic imaging on patient outcomes are rarely well characterized but differences between imaging test features are well known. The results of our study suggest that there are differences in BRC across different clinical scenarios, and that imaging tests can be judged by more than diagnostic accuracy or potential for ionizing-radiation exposure. We propose a BRC list to ensure consistent examination of these differences among tests, even when evidence is sparse. This is the first step toward merging beliefs, expert opinion, and existing evidence, and documenting the guideline development process in a more transparent, reproducible manner.

ACKNOWLEDGEMENTS

This work was supported by the GE Healthcare Fellowship. The authors wish to acknowledge the expertise and support provided by Reed Johnson, Caroline Bennette and the survey participants for their participation.

REFERENCES

1. Lee DW, Neumann PJ, Rizzo JA. Understanding the medical and nonmedical value of diagnostic testing. *Value Health.* Mar-Apr 2010;13(2):310-314.
2. Brook RH. *The RAND/UCLA Appropriateness Method.* Rockville, MD: Public Health Service, U.S. Department of Health and Human Services;1994.
3. Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. *Int J Technol Assess Health Care.* 1986;2(1):53-63.
4. Gazelle GS, Kessler L, Lee DW, et al. A framework for assessing the value of diagnostic imaging in the era of comparative effectiveness research. *Radiology.* Dec 2011;261(3):692-698.
5. Bossuyt PM, Reitsma JB, Linnet K, Moons KG. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. *Clinical chemistry.* Dec 2012;58(12):1636-1643.
6. Staub LP, Dyer S, Lord SJ, Simes RJ. Linking the evidence: intermediate outcomes in medical test assessments. *Int J Technol Assess Health Care.* Jan 2012;28(1):52-58.
7. Staub LP, Lord SJ, Simes RJ, et al. Using patient management as a surrogate for patient health outcomes in diagnostic test evaluation. *BMC Med Res Methodol.* 2012;12:12.
8. Otero HJ, Fang CH, Sekar M, Ward RJ, Neumann PJ. Accuracy, risk and the intrinsic value of diagnostic imaging: a review of the cost-utility literature. *Acad Radiol.* May 2012;19(5):599-606.
9. Bossuyt PMM, McCaffery K. Additional Patient Outcomes and Pathways in Evaluations of Testing. *Medical Tests-White Paper Series*. Rockville (MD)2009.
10. American College of Radiology. About the ACR Appropriateness Criteria. 2015; http://www.acr.org/Quality-Safety/Appropriateness-Criteria/About-AC. Accessed December, 11, 2014.

11. Rao VM, Levin DC. The overuse of diagnostic imaging and the Choosing Wisely initiative. *Ann Intern Med.* Oct 16 2012;157(8):574-576.
12. Bresnahan BW, Garrison LP. Economic Issues in Diagnostic Imaging. In: Culyer T, ed. *Encyclopedia of Health Economics*. Oxford: Elsevier; 2014.
13. Agapova M, Devine EB, Bresnahan BW, Higashi MK, Garrison LP, Jr. Applying quantitative benefit-risk analysis to aid regulatory decision making in diagnostic imaging: methods, challenges, and opportunities. *Academic radiology.* Sep 2014;21(9):1138-1143.
14. Food and Drug Administration. Guidance for industry and Food and Drug Administration staff: factors to consider when making benefit-risk determinations in medical device premarket approval and de novo classifications. 2012; http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM296379.pdf Accessed October 22, 2012.
15. Food and Drug Administration. Patient Preference Information – Submission, Review in PMAs, HDE Applications, and De Novo Requests, and Inclusion in Device Labeling 2015; http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM296379.pdf Accessed August 22, 2015.
16. Guest G, MacQueen KM, Namey EE. *Applied thematic analysis*. Los Angeles: Sage Publications; 2012.
17. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making.* Apr-Jun 1991;11(2):88-94.
18. Westphalen AC, Hsia RY, Maselli JH, Wang R, Gonzales R. Radiological imaging of patients with suspected urinary tract stones: national trends, diagnoses, and predictors. *Acad Emerg Med.* Jul 2011;18(7):699-707.

19. Smith-Bindman R, Aubin C, Bailitz J, et al. Ultrasonography versus computed tomography for suspected nephrolithiasis. *The New England journal of medicine.* Sep 18 2014;371(12):1100-1110.

20. Hilgerink MP, Hummel MJ, Manohar S, Vaartjes SR, Ijzerman MJ. Assessment of the added value of the Twente Photoacoustic Mammoscope in breast cancer diagnosis. *Med Devices (Auckl).* 2011;4:107-115.

21. Hoffmann TC, Del Mar C. Patients' expectations of the benefits and harms of treatments, screening, and tests: a systematic review. *JAMA Intern Med.* Feb 1 2015;175(2):274-286.

22. ACR Appropriateness Criteria. ACR ACR Appropriateness Criteria® Rating Round Information. 2015; http://www.acr.org/~/media/ACR/Documents/AppCriteria/RatingRoundInfo.pdf. Accessed August 4, 2015.

23. Schnipper LE, Davidson NE, Wollins DS, et al. American Society of Clinical Oncology Statement: A Conceptual Framework to Assess the Value of Cancer Treatment Options. *J Clin Oncol.* Aug 10 2015;33(23):2563-2577.

24. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Fam Pract.* Feb 2004;21(1):4-10.

25. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ.* 2012;344:e686.

26. Podberesky DJ, Angel E, Yoshizumi TT, et al. Comparison of radiation dose estimates and scan performance in pediatric high-resolution thoracic CT for volumetric 320-detector row, helical 64-detector row, and noncontiguous axial scan acquisitions. *Acad Radiol.* Sep 2013;20(9):1152-1161.

27. Shekalaghe S, Cancino M, Mavere C, et al. Clinical performance of an automated reader in interpreting malaria rapid diagnostic tests in Tanzania. *Malar J.* 2013;12:141.

28. Novielli N, Cooper NJ, Sutton AJ. Evaluating the cost-effectiveness of diagnostic tests in combination: is it important to allow for performance dependency? *Value Health.* Jun 2013;16(4):536-541.

29. Batwala V, Magnussen P, Nuwaha F. Comparative feasibility of implementing rapid diagnostic test and microscopy for parasitological diagnosis of malaria in Uganda. *Malar J.* 2011;10:373.

30. Fowler JR, Maltenfort MG, Ilyas AM. Ultrasound as a first-line test in the diagnosis of carpal tunnel syndrome: a cost-effectiveness analysis. *Clin Orthop Relat Res.* Mar 2013;471(3):932-937.

31. Freixa X, Trilla M, Feldman M, Jimenez M, Betriu A, Masotti M. Right versus left transradial approach for coronary catheterization in octogenarian patients. *Catheter Cardiovasc Interv.* Aug 1 2012;80(2):267-272.

32. Pulcini C, Pauvif L, Paraponaris A, Verger P, Ventelou B. Perceptions and attitudes of French general practitioners towards rapid antigen diagnostic tests in acute pharyngitis using a randomized case vignette study. *J Antimicrob Chemother.* Jun 2012;67(6):1540-1546.

33. Rossi IA, D'Acremont V, Prod'Hom G, Genton B. Safety of falciparum malaria diagnostic strategy based on rapid diagnostic tests in returning travellers and migrants: a retrospective study. *Malar J.* 2012;11:377.

34. Loganathan AG, Chan MD, Alphonse N, et al. Clinical outcomes of brain metastases treated with Gamma Knife radiosurgery with 3.0 T versus 1.5 T MRI-based treatment planning: have we finally optimised detection of occult brain metastases? *J Med Imaging Radiat Oncol.* Oct 2012;56(5):554-560.

35. Than M, Cullen L, Aldous S, et al. 2-Hour accelerated diagnostic protocol to assess patients with chest pain symptoms using contemporary troponins as the only biomarker: the ADAPT trial. *J Am Coll Cardiol.* Jun 5 2012;59(23):2091-2098.

36. Margolis NE, Shaver CM, Rosenkrantz AB. Indeterminate liver and renal lesions: comparison of computed tomography and magnetic resonance imaging in providing a definitive diagnosis and impact on recommendations for additional imaging. *J Comput Assist Tomogr.* Nov-Dec 2013;37(6):882-886.

37. Rodrigo E, Lopez-Hoyos M, Corral M, et al. ImmuKnow as a diagnostic tool for predicting infection and acute rejection in adult liver transplant recipients: a systematic review and meta-analysis. *Liver Transpl.* Oct 2012;18(10):1245-1253.

38. Beaton A, Okello E, Lwabi P, Mondo C, McCarter R, Sable C. Echocardiography screening for rheumatic heart disease in Ugandan schoolchildren. *Circulation.* Jun 26 2012;125(25):3127-3132.

39. Simprini LA, Taylor AJ. Cardiac CT in women: clinical application and considerations. *J Cardiovasc Comput Tomogr.* Mar-Apr 2012;6(2):71-77.

40. Moschetti K, Muzzarelli S, Pinget C, et al. Cost evaluation of cardiovascular magnetic resonance versus coronary angiography for the diagnostic work-up of coronary artery disease: application of the European Cardiovascular Magnetic Resonance registry data to the German, United Kingdom, Swiss, and United States health care systems. *J Cardiovasc Magn Reson.* 2012;14:35.

41. Nance JW, Jr., Bamberg F, Schoepf UJ. Coronary computed tomography angiography in patients with chronic chest pain: systematic review of evidence base and cost-effectiveness. *J Thorac Imaging.* Sep 2012;27(5):277-288.

42. Wagner J, Aron DC. Incidentalomas: a "disease" of modern imaging technology. *Best Pract Res Clin Endocrinol Metab.* Feb 2012;26(1):3-8.

43. Tagliafico A, Succio G, Serafini G, Martinoli C. Diagnostic performance of ultrasound in patients with suspected brachial plexus lesions in adults: a multicenter

retrospective study with MRI, surgical findings and clinical follow-up as reference standard. *Skeletal Radiol.* Mar 2013;42(3):371-376.

44. Azeem N, Tabibian JH, Baron TH, et al. Use of a single-balloon enteroscope compared with variable-stiffness colonoscopes for endoscopic retrograde cholangiography in liver transplant patients with Roux-en-Y biliary anastomosis. *Gastrointestinal endoscopy.* Apr 2013;77(4):568-577.
45. Spencer JD, Bates CM, Mahan JD, et al. The accuracy and health risks of a voiding cystourethrogram after a febrile urinary tract infection. *J Pediatr Urol.* Feb 2012;8(1):72-76.
46. von Wagner C, Ghanouni A, Halligan S, et al. Patient acceptability and psychologic consequences of CT colonography compared with those of colonoscopy: results from a multicenter randomized controlled trial of symptomatic patients. *Radiology.* Jun 2012;263(3):723-731.
47. Mudrick DW, Cowper PA, Shah BR, et al. Downstream procedures and outcomes after stress testing for chest pain without known coronary artery disease in the United States. *Am Heart J.* Mar 2012;163(3):454-461.
48. Ma S, Kong B, Liu B, Liu X. Biological effects of low-dose radiation from computed tomography scanning. *Int J Radiat Biol.* May 2013;89(5):326-333.
49. Lin YK, Gettle L, Raman JD. Significant variability in 10-year cumulative radiation exposure incurred on different surveillance regimens after surgery for pT1 renal cancers: yet another reason to standardize protocols? *BJU Int.* May 2013;111(6):891-896.
50. Hao R, Zhang Q, Xu Z, et al. Magnetic navigation system and CT roadmap-assisted percutaneous coronary intervention: a comparison to the conventional approach. *J Invasive Cardiol.* Apr 2013;25(4):177-181.

51. Shahbazi-Gahrouei D, Baradaran-Ghahfarokhi M. Assessment of entrance surface dose and health risk from common radiology examinations in Iran. *Radiat Prot Dosimetry.* 2013;154(3):308-313.

52. Martin CJ, Huda W. Intercomparison of patient CTDI surveys in three countries. *Radiat Prot Dosimetry.* 2013;153(4):431-440.

53. Shah DJ, Sachs RK, Wilson DJ. Radiation-induced cancer: a modern view. *Br J Radiol.* Dec 2012;85(1020):e1166-1173.

54. Smith IR, Cameron J, Mengersen KL, Rivers JT. Evaluation of coronary angiographic projections to balance the clinical yield with the radiation risk. *Br J Radiol.* Sep 2012;85(1017):e722-728.

55. Norgaz T, Gorgulu S, Dagdelen S. A randomized study comparing the effectiveness of right and left radial approach for coronary angiography. *Catheter Cardiovasc Interv.* Aug 1 2012;80(2):260-264.

56. Durand DJ, Dixon RL, Morin RL. Utilization strategies for cumulative dose estimates: a review and rational assessment. *J Am Coll Radiol.* Jul 2012;9(7):480-485.

57. Koshy S, Thompson RC. Review of radiation reduction strategies in clinical cardiovascular imaging. *Cardiol Rev.* May-Jun 2012;20(3):139-144.

58. Kusmierek J, Plachcinska A. Patient exposure to ionising radiation due to nuclear medicine cardiac procedures. *Nucl Med Rev Cent East Eur.* 2012;15(1):71-74.

59. Grunheid T, Kolbeck Schieck JR, Pliska BT, Ahmad M, Larson BE. Dosimetry of a cone-beam computed tomography machine compared with a digital x-ray machine in orthodontic imaging. *Am J Orthod Dentofacial Orthop.* Apr 2012;141(4):436-443.

60. van Vlijmen OJ, Kuijpers MA, Berge SJ, et al. Evidence supporting the use of cone-beam computed tomography in orthodontics. *J Am Dent Assoc.* Mar 2012;143(3):241-252.

61. Barrett B, Stiles M, Patterson J. Radiation risks: critical analysis and commentary. *Prev Med.* Mar-Apr 2012;54(3-4):280-282.

62. Schoenhagen P, Thompson CM, Halliburton SS. Low-dose cardiovascular computed tomography: where are the limits? *Curr Cardiol Rep.* Feb 2012;14(1):17-23.

63. Ewer AK, Furmston AT, Middleton LJ, et al. Pulse oximetry as a screening test for congenital heart defects in newborn infants: a test accuracy study with evaluation of acceptability and cost-effectiveness. *Health Technol Assess.* 2012;16(2):v-xiii, 1-184.

64. Fennich N, Ellouali F, Abdelali S, et al. Stress echocardiography: safety and tolerability. *Cardiovasc Ultrasound.* 2013;11:30.

65. Winer JL, Liu CY, Apuzzo ML. The use of nanoparticles as contrast media in neuroimaging: a statement on toxicity. *World Neurosurg.* Dec 2012;78(6):709-711.

66. Firouzi A, Eshraghi A, Shakerian F, et al. Efficacy of pentoxifylline in prevention of contrast-induced nephropathy in angioplasty patients. *Int Urol Nephrol.* Aug 2012;44(4):1145-1149.

67. Neubauer A, Wolfsberger S. Virtual endoscopy in neurosurgery: a review. *Neurosurgery.* Jan 2013;72 Suppl 1:97-106.

68. Tay CM, Chang SK. Diagnosis and management of pancreaticopleural fistula. *Singapore Med J.* Apr 2013;54(4):190-194.

69. Jenssen C, Alvarez-Sanchez MV, Napoleon B, Faiss S. Diagnostic endoscopic ultrasonography: assessment of safety and prevention of complications. *World J Gastroenterol.* Sep 14 2012;18(34):4659-4676.

70. Pinto S, Pinto A, de Carvalho M. Phrenic nerve studies predict survival in amyotrophic lateral sclerosis. *Clin Neurophysiol.* Dec 2012;123(12):2454-2459.

71. Zondervan RL, Hahn PF, Sadow CA, Liu B, Lee SI. Body CT scanning in young adults: examination indications, patient outcomes, and risk of radiation-induced cancer. *Radiology.* May 2013;267(2):460-469.

72. Do CT scans cause cancer? For older men the risk from diagnostic CT scans is relatively small. *Harv Mens Health Watch.* Mar 2013;17(8):3.

73. Krille L, Zeeb H, Jahnen A, et al. Computed tomographies and cancer risk in children: a literature overview of CT practices, risk estimations and an epidemiologic cohort study proposal. *Radiat Environ Biophys.* May 2012;51(2):103-111.

74. Pauwels EK, Bourguignon MH. Radiation dose features and solid cancer induction in pediatric computed tomography. *Med Princ Pract.* 2012;21(6):508-515.

75. Saltzherr TP, Goslings JC, Bakker FC, et al. Cost-effectiveness of trauma CT in the trauma room versus the radiology department: the REACT trial. *Eur Radiol.* Jan 2013;23(1):148-155.

76. Zheng D, Huang X, Fan Y, et al. The effect of octreotide treatment on patients with pancreatic cancer who undergo endoscopic retrograde cholangiopancreatography (ERCP) with pancreatic duct stent placement. *Hepatogastroenterology.* Mar-Apr 2013;60(122):222-224.

TABLES

ACR AC Recommendation	Survey description of case
Magnetic resonance imaging compared to no testing	
Low back pain (NGC-8863) variant 1	A patient seeks diagnosis for non-specific, subacute lower back pain. The patient has had non-specific pain for more than 3 months but has no history of structural problems or trauma, leg pain or red flags.
Chronic headache, no new features (NGC-7779) variant 1	A patient seeks diagnosis for chronic uncomplicated headache. The patient is not experiencing new headache features, focal neurological deficits or red flags.
Ultrasound (US) compared to computed tomography (CT)	
Lower quadrant pain-suspected appendicitis (NGC-10146) variant 1	A patient arrives complaining of lower quadrant pain. Fever, leukocytosis and other signs point to a classic case of clinical appendicitis.
Acute onset flank pain-suspicion of stone disease (NGC-008476) variant 2	A patient arrives complaining of acute onset flank pain. The patient is having recurrent symptoms of stone disease.
NGC: National Guideline Clearinghouse	

Table 1: Description of clinical conditions presented in the survey

Table 2. Test Specific Features BRC

Criteria	Brief description
Missed cases	Type II error or as 1-NPV (the chance of having the condition among those that test negative)
False diagnoses	Type I error or 1-PPV (The chance of not having the condition among those that test positive)
Diagnostic accuracy consistency	Existence/extent of influence of patient characteristics on diagnostic accuracy
Inter-observer reading agreement	Proxy measure of image clarity and quality
Depth/breadth of anatomy visualization	Categorization of extent of anomaly characterization (e.g. size, shape, vascularization, shape)
Invasiveness/risk of adverse events	The number, or categorization of probabilities and/or severity of AE(s)
Contrast reaction potential	Probability or categorization of probability and /or severity of contrast reaction
Ionizing radiation dose	Measure of milliSievert dose or categorization of dose
Patient-specific exclusions	Measure of existence/extent categorization of exclusions (e.g. metal implants, BMI, age)
Failure/malfunction rate	Failure rate or categorization of the rate/manufacturer reputation
Patient preparation requirements	Number of minutes or categorization of relative wait times
Examination time	Number of minutes or categorization of relative wait times
Post-test observation time	Number of minutes or categorization of relative wait times
Decision support	Existence/extent of automated interpretation or characterization of function
Portability	Existence/extent of device portability
Ease of use	Categorization of dependence on skilled operator
Reimbursement potential	Categorization of relative potential for reimbursement

NPV: Negative Predictive Value; PPV: Positive Predictive Value; AE: Adverse Events; BMI: Body Mass Index

Table 3. Patient Management and Provider Intrinsic Value BRC

Criteria	Brief Description
Therapeutic/procedural success	Net counts/probability and severity/categorization of complications of medical treatment with/without test
Potential for additional confirmatory testing (Inconclusive/False positive results)	Existence/extent of confirmatory testing
Potential for incidental finding management	Existence /extent of repeat follow-up
Net unnecessary treatment (Test-prescribed or averted treatment)	Net counts/probability and severity/categorization of unnecessary treatments performed or averted based on test information
Access to test	Perceived relative access to test
Time to diagnosis	Net hours/days/weeks to diagnosis or extent of delay with/without test
Inpatient/outpatient healthcare visits	Net number of healthcare visits likely or extent of utilization with/without test
Time to discharge	Net hours/days/weeks to diagnosis with/without test
Provider confidence in test	Extent of confidence in test usefulness
Liability protections	Existence/extent of protection from liability afforded by test
Financial incentives	Existence/extent associated with test
Contribution of information to prognosis	Existence/extent of test information contributing to prognosis

Table 4. Patient Intrinsic Value and Outcomes BRC

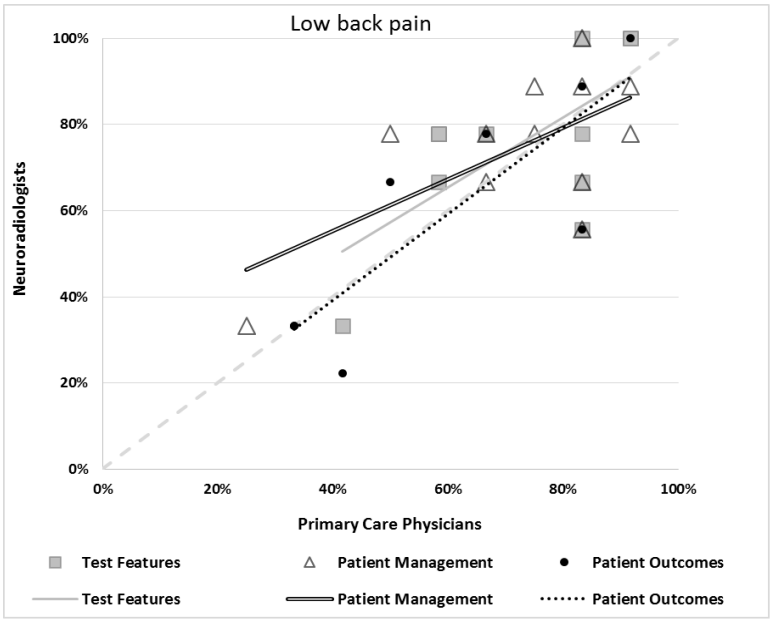
Criteria	Brief Description
Value of knowing	Value of knowing true test results. Decrease in perceived uncertainty (e.g. peace of mind, reassurance)
Disvalue of knowing	Disvalue of knowing false test results or learning of insignificant incidental findings. Increase in perceived uncertainty (e.g. anxiety, confusion, distrust)
Burden (time and money) to patient	Out of pocket, travel costs and work absenteeism: direct time and money costs of test
Patient comfort	Claustrophobia, fasting, physical discomfort and pain from test
Patient future compliance and behavior	Changes in behavior: measures of uptake or attrition of health visits/programs
Radiation-induced cancers	Count of expected cases/QALYs lost
Length/quality of life	Net incremental survival attributable to test/net QALYs
QALYS: quality-adjusted life-years	

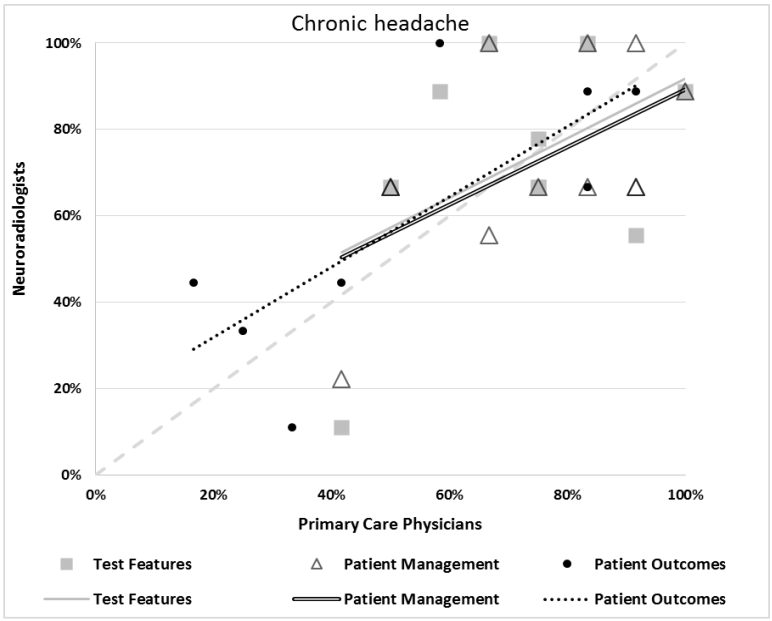
Table 5. Consensus Results

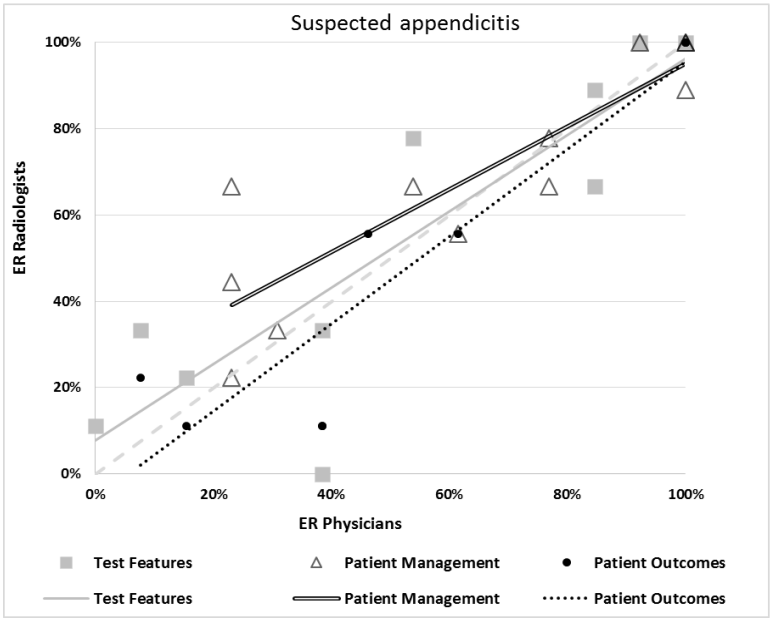
Criteria	≥80% radiologists	≥80% non-radiologists	>80% both groups
Missed cases			LBP, CH, SA, SSD
Provider utility			LBP, CH, SA, SSD
Potential for confirmatory testing			LBP, CH, SA, SSD
Test accuracy consistency	SA, SSD, CH	SA, SSD	SA, SSD
Radiation dose	SA, SSD	SA, SSD	SA, SSD
Radiation-induced cancers	SA, SSD	SA, SSD	SA, SSD
Portability	SSD		SA, SSD
False diagnoses	LBP, CH	LBP, CH	LBP, CH
Psychological impact	LBP, CH	LBP, CH	LBP, CH
Patient burden	LBP, CH	LBP, CH	LBP, CH
Unnecessary treatment	LBP, CH	LBP, CH	LBP, CH
Reimbursement potential	LBP, CH	LBP	LBP, CH
Incidental findings management		LBP, CH	LBP, CH
Invasiveness/risk of adverse events		SA, LBP	
Healthcare visits		SSD	
Value of knowing	CH	SSD	
Contrast reaction potential		LBP, CH	
Patient preparation		LBP, CH	
Liability protections		LBP	
Time to diagnosis	LBP	CH	
Contribution of information to prognosis	CH	SSD	SSD
Incidental findings management		SSD	SSD
SA: Suspected appendicitis; SSD: Suspected stone disese; CH: chronic headache; LBP: lower back pain			

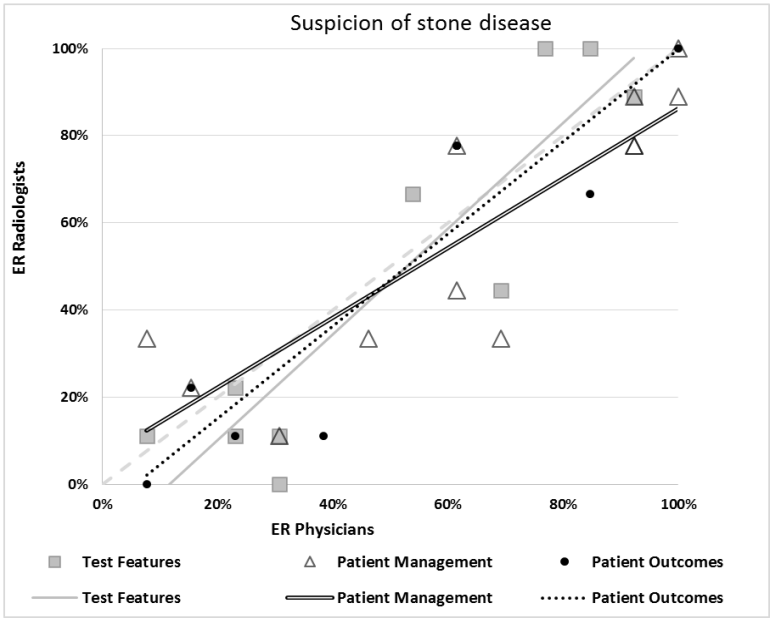
FIGURES

Figure 1. Inter and intra-agreement of selection of BRC among radiologists and non-radiologists grouped by domain









APPENDIX

Table A1. Literature search results mapped to the Fryback and Thornbury Hierarchical Model of Efficacy

Technical efficacy The design and technical performance of test Resolution of line pairs Modulation of transfer function change Gray scale range Amount of mottle Sharpness	Reliability of image quality[26] and performance[27] Performance dependencies (multiple tests relying on one another for effectiveness) [28] Patient wait/preparation time for test[29] Length of examination[30-32] Post-test observation time[33-35] Inconclusive results [8,36]
Diagnostic accuracy efficacy Yield of abnormal or normal diagnoses in case series Diagnostic accuracy Positive and negative predictive value Sensitivity and specificity in a clinical case ROC	Positive and negative predictive values (47 references) Positive likelihood ratio[37] Diagnostic odds ratio[37] Detection percentage[38] Accuracy in subpopulations[39]
Diagnostic thinking efficacy Impact on pretest probability Probability of learning incidental information	Provider confidence[6] Liability avoidance[8] Risk stratification[40,41] Probability of learning incidental information[42] Pre-test probability[43]
Therapeutic efficacy How helpful an imaging test is to patient management Unnecessary procedures avoided Pre-test/post-test therapy changes	Procedural success[44] Number of healthcare visits[45] Need for additional follow-up[46] Downstream procedures from incidental findings[47] Radiation exposure dose[26,39,48-62]

	Extent and severity of harms associated with unnecessary procedures[41]
Patient outcome efficacy Morbidity (procedures) avoided in QALYs Patient improvement with vs. without test Survival measured in QALYs Value of test information to patient (future planning and psychological impact) Cost per QALY saved with image information	Pain and other physical discomfort such as tight quarters[30,46] Psychological impact (e.g. anxiety, reassurance or peace of mind)[1,7,46,63] Future planning based on test information[1] Tolerability[64] Sequelae from contrast reagent reactions[65,66] Invasiveness[67-69] Patient compliance (downstream)[6] Morbidity and mortality benefit from improved treatment plan[34,70] Radiation-associated cancers[71-74] Non-institutionalized days alive [75] Quality of life[76]
Societal efficacy, the last tier of the hierarchy is not represented, as costs are not typically considered in benefit-risk analyses.	

Table A2. Percent consensus of criteria with >80% for each use case by clinical specialty

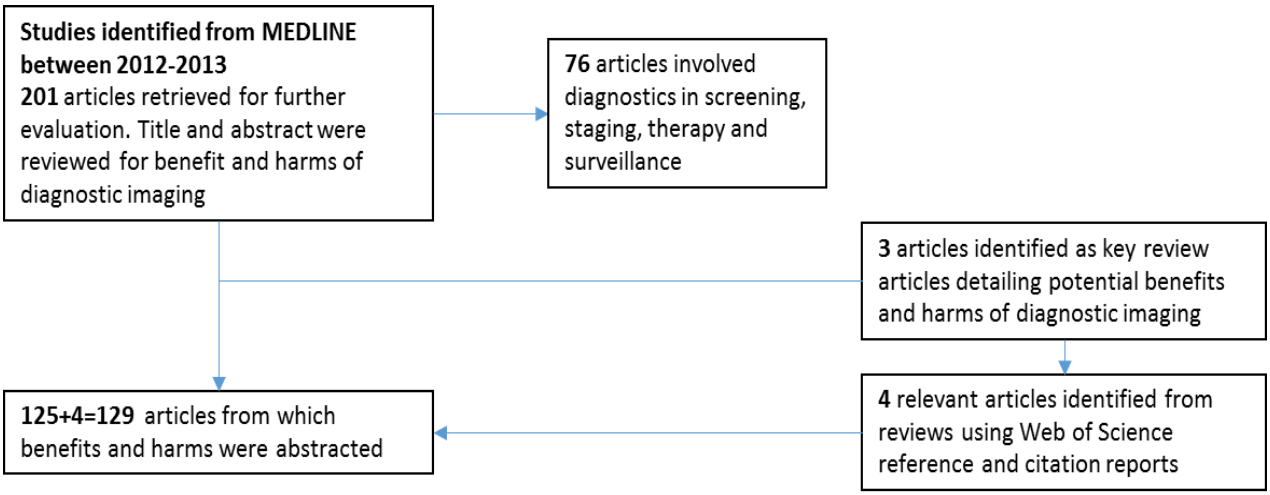
Appendicitis		ED phys	ED rad	Total
Non-radiologists	adverse events	85%	67%	77%
Both groups	contrast reactions	100%	100%	100%
	incidental findings management	92%	100%	95%
	missed cases	100%	100%	100%
	portability of the device	85%	89%	86%
	potential for confirmatory testing	100%	100%	100%
	provider confidence	100%	100%	100%
	radiation dose	100%	89%	95%
	radiation-induced cancers	100%	100%	100%
	test accuracy consistency	92%	100%	95%
Radiologists	No selections			
Kidney Stone Disease		ED phys	ED rad	Total
Non-radiologists	contribution to prognosis	92%	78%	**86%**
	healthcare visits	85%	44%	68%
	incidental findings management	92%	78%	**86%**
	value of knowing	85%	67%	77%

Both groups	missed cases	92%	89%	91%
	potential for confirmatory testing	100%	89%	95%
	provider confidence	92%	89%	91%
	radiation dose	100%	100%	100%
	radiation-induced cancers	100%	100%	100%
	test accuracy consistency	85%	100%	91%
Radiologists	portability of the device	77%	100%	**86%**

Lower Back Pain		PCP	Neurorad	Total
Non-radiologists	contrast reaction potential	83%	56%	71%
	incidental findings management	92%	78%	**86%**
	healthcare visits	83%	56%	71%
	liability protections	83%	67%	76%
	patient preparation	83%	78%	**81%**
	adverse events	83%	67%	76%
Both groups	false diagnoses	92%	100%	95%
	missed cases	83%	100%	90%
	patient burden	92%	100%	95%
	potential for confirmatory testing	83%	100%	90%
	provider confidence	83%	89%	86%
	psychological impact	83%	89%	86%
	Reimbursement	92%	100%	95%
	unnecessary treatment	92%	89%	90%

Radiologists	time to diagnosis	75%	89%	**81%**
Chronic headache		PCP	Neurorad	**Total**
Non-radiologists	contrast reaction potential	83%	67%	76%
	patient preparation	92%	56%	76%
	incidental findings management	92%	67%	**81%**
	time to diagnosis	83%	67%	76%
Both groups	false diagnoses	100%	89%	95%
	liability protections	100%	89%	95%
	missed cases	83%	100%	90%
	patient burden	92%	89%	90%
	potential for confirmatory testing	92%	100%	95%
	provider confidence	83%	100%	90%
	psychological impact	83%	89%	86%
	unnecessary treatment	92%	67%	81%
Radiologists	contribution to prognosis	67%	100%	**81%**
	reimbursement potential	67%	100%	**81%**
	test accuracy consistency	58%	89%	71%
	value of knowing	58%	100%	76%

Figure A1. A flowchart of included and excluded studies in literature search



Chapter 2. A Proposed Approach for Quantitative Benefit-Risk Assessment in Diagnostic Radiology Guideline Development: The American College of Radiology Appropriateness Criteria Example

Maria Agapova, Brian W. Bresnahan, Reed Johnson, Louis Garrison, Mitchell Higashi,

Larry Kessler, Beth Devine

Target Journal: Journal of Evaluation in Clinical Practice or BMJ Research Methods and Reporting

Word Count: 4,461

Tables: 4

Figures: 4

ABSTRACT

Abstract Word Count: 257 (including headings)

INTRODUCTION

By its nature, the contribution of expert opinion in evidence-based guideline development can be nontransparent and implicit. The American College of Radiology (ACR) develops evidence-based practice guidelines to aid appropriate utilization of radiological procedures. Panel members use expert opinion to weight trade-offs and consensus methods to rate appropriateness of imaging tests. Quantitative benefit-risk assessment (QBRA) methods hold promise for improving transparency of diagnostic radiology guideline development but their potential feasibility and effectiveness have not previously been assessed in this setting.

METHODS

We perform a critical appraisal of the QBRA literature and propose several steps for further exploring the clinical decision problem. These steps include investigation of weak evidence and disagreement. We identify a set of benefit-risk methods addressing one or more of these needs and build a decision aid for selecting among these methods.

RESULTS

Identifying specific benefit-risk criteria and developing a state-of-evidence matrix are key steps in improving transparency of ACR panel decision-making. We propose, that in the absence of both disagreement among decision makers and a weak evidence base, QBRA may not be needed. In the presence of disagreement but absence of a weak evidence base, MCDA is recommended. In the presence of weak evidence base INHB or a joint approach may be needed.

CONCLUSIONS

QBRA is a systematic and consistent approach that promises to increase the transparency of the ACR deliberative process. This investigation of the strengths and limitations of two QBRA methods guides decision-makers in choosing a useful QBRA method among incremental net health benefit, multi-criteria decision analysis, or a joint approach.

INTRODUCTION

Evidence-based guidelines and the ACR AC

The American College of Radiology (ACR) develops evidence-based practice guidelines to aid appropriate utilization of radiological procedures.[1] For a given diagnostic indication and for each imaging test in question, a panel of clinical experts reviews the evidence, and qualitatively assesses whether potential benefits outweigh potential risks, by a sufficient margin.[2] These guidelines are referred to as Appropriateness Criteria (AC).[3] Not all agree that the AC meet the definition of evidence-based practice (EBP) guidelines.[4] To adhere to the definition of EBP, the AC ought to represent integration of the best available evidence with clinician's expertise and patient choice.[5] AC evidence review is not systematic but "freestyle",[6] rarely includes non-radiology experts and excludes patient stakeholders. Nonetheless, the process offers regularly updated appropriateness ratings for hundreds of clinical scenarios. For guideline development, full inclusion of the three elements of EBP requires implementation of a complex, multi-step process, involving multiple stakeholders. The ACR AC can be viewed as a compromise between the full inclusion of EBP and breadth and timeliness of recommendations. In clinical practice, utilization of ACR AC is low and the effects of ACR AC on optimizing utilization of diagnostic imaging are poorly understood.[7,8]

Expert Opinion and Consensus in ACR AC Development

To rate appropriateness, developers of AC guidelines use expert consensus methods. Based on the RAND/UCLA Appropriateness Method, over a course of one to two modified Delphi voting rounds, panel members assign appropriateness scores (1-3 inappropriate; 4-6 equivocal; 7-9 appropriate).[1,2] Panel members assign equivocal ratings when it is unclear whether the test "may be appropriate". This category of appropriateness may be assigned for one or more of these reasons: 1) lack of consensus; 2) contradictory or unclear evidence; or 3) special circumstances. Individual panel members may rely on their expertise

to carry out several steps before entering the consensus-building process: 1) they may identify which benefits and risks apply to the clinical scenario and the tests in question; 2) apply individual preferences to weight the importance of relevant benefits and risks and; 3) judge performance of tests under high levels of uncertainty, using their personal risk tolerance to inform judgments. As these steps are implicit and done at the individual level, it is not clear how and to what extent expert opinion contributes to individual ratings of appropriateness and thereby affects expert consensus.

A more explicit approach based on quantitative measures of benefits and risks, Quantitative Benefit-Risk Assessment (QBRA), has the potential to structure and simplify the decision-making process leading to more transparent contribution of expert opinion in guideline development.[9] However, there are numerous QBRA approaches and the general consensus is that no single method is superior for all applications.[10-13] Mt-Isa *et al.* reviewed reviews of QBRA methodology,[14] and found that the Innovative Medicines Initiative Pharmacoepidemiological Research on Outcomes of Therapeutics by the European Consortium (IMI-PROTECT) Benefit-Risk Group Recommendations Report represented the most comprehensive review.[10] Stemming from this report, Hughes *et al*. recommend further testing of eleven QBRA methods in the context of pharmaceutical regulatory decision-making.[13]

A step common to QBRA is structuring the decision problem by identifying and defining relevant benefits and risks. In a previous report (Chapter 1), we developed benefit and risk criteria (BRC) specific for QBRA of diagnostic imaging tests. In brief, thirty-six BRC were identified, defined, and organized into three domains: 1) test-specific features; 2) patient management and provider intrinsic value; and 3) patient-intrinsic value and outcomes. We found that participants selected 22 of 36 BRC as relevant, across four ACR AC clinical scenarios, and for each of the four scenarios studied in that analysis, between eight and eleven criteria were selected.

The application of any specific QBRA method for a specific clinical question can be costly in terms of time and money, so its selection and use needs to be guided by a first step of deciding whether QBRA could be helpful and which QBRA method is preferred in terms of its effectiveness and efficiency. This analysis had three major aims: 1) identify which QBRA methods may facilitate ACR guideline development process; 2) identify opportunities for QBRA within ACR AC decision-making; and 3) where QBRA is recommended, make specific recommendations for specific QBRA methods.

METHODS

Development of a Decision Aid

We develop a multi-level decision aid to assist panel members in identifying opportunities for using QBRA and selecting among methods. In the first level, we propose a protocol for identifying when QBRA is needed based on the presence of equivocal ratings for one or more reasons (a) a lack of consensus or (b) issues with missing, unclear, or contradictory evidence but not for reason (c) special circumstances (this reason is outside the scope of this work). In the second level, we facilitate selection among methods. Within this level we propose a process characterizing the clinical scenario decision problem further than is currently performed by the ACR with respect to consensus-building or missing, unclear or contradictory evidence. Separately, we critically appraise QBRA methods with respect to controlling disagreement among panel members and missing, unclear or contradictory evidence. We develop decision rules that merge strengths of QBRA methods with the identified needs of the decision problem.

QBRA methods that may facilitate the ACR guideline development process

We further refine the recommendation list of IMI-PROTECT by targeting the role of QBRA for improvement of ACR AC development. For evaluation, we select methods for use in ACR AC if they included the following features:

1) Single measure of the benefit-risk balance: Panel members' single rating reflects weighting of multiple benefits and risks. QBRA methods that are capable of weighting benefits and risks, integrating benefits and risks and producing a composite benefit-risk metric may be particularly useful to AC development. To select among methods with respect to this feature, we use the IMI-PROTECT categorization of QBRA methods according to whether the method incorporated weighting of benefits and risks, whether benefits and risks were integrated, and whether the method produced a benefit-risk metric.

2) Sensitivity Analysis: Particularly in diagnostic imaging, evidence may be of poor quality, contradictory or not available. A desirable feature offered by some QBRA methods is their ability to further explore panel members' uncertainty about their judgments of performance of tests or importance of benefits and risks. We further refine the list by including only QBRA methods that assess uncertainty using at a minimum, deterministic sensitivity analysis.

RESULTS

Selected QBRA Methods

Four broad QBRA approaches from those recommended by the IMI-PROTECT project possessed a single measure of benefit-risk balance and were equipped to measure effects of uncertainty using sensitivity analyses: incremental net health benefit (INHB), multi-criteria decision analysis (MCDA), stochastic multicriteria acceptability analysis (SMAA), and discrete choice experiments (DCE) (Table 1). There are many derivative approaches within each of these methods. INHB has been generalized to incremental net benefit or incremental health benefit.[15] This method is commonly paired with decision-analytic modeling using the quality-adjusted life-year (QALY).[16,17] SMAA is often considered an extension of MCDA for decision problems in which decision-makers are unsure of effects across criteria or cannot or will not provide their preferences.[18] Several reviews of MCDA

consider DCE a type of MCDA.[19] Thus, we reduced the decision aid to consideration between two broad approaches to decision analysis: MCDA and INHB. We identified key differences between these two approaches, their strengths and limitations.

[Insert Table 1 Here]

INHB

INHB, as defined here, uses decision analytic modeling to synthesize available evidence and model consequences using data from several sources.[20,21] Within a benefit-risk context, a decision tree may be built representing possible benefit-risk consequence paths. Probabilities and the preferences for those events or health states populate the decision tree.[22] While INHB can be implemented using natural units, or other relative value metrics,[23] the quality-adjusted life-year (QALY) framework for integrating length with quality of life is a common metric for decision-analytic modeling.[24] Trade-offs among benefits and risks are measured incrementally as the sum of risks expressed as negative QALYs and benefits, expressed as positive QALYs. A framework for using these decision-analytic methods for BRA has been previously described.[15,16]

Strengths

IHNB offers structure to a decision problem modeling problem by estimating downstream clinical effects from those upstream with the use of probabilities.[25] By incorporating preference data from external studies (studies of similar health states) and linking measured outcomes to ones unmeasured via probabilities, INHB may offer an alternative to expert consensus when evidence is limited. Garrison *et al.* have also emphasized that the framework allows heterogeneous patient health state preferences, subgroup analyses, adjustment for patient risk, and value-of-information analysis.[16]

Expert opinion can be incorporated into a model by way of assumptions with regard to the probabilities of events or the consequences and related health state preferences for those

events. The use of expert opinion is transparent because model inputs that are assumptions are referenced as such. Scenario and sensitivity analyses can be performed to test independently the influence of probabilities and the preference for the consequences. Preferences for health states that inform the models are extracted from the literature and represent risk tolerance of the study population, not those of the decision-makers.

INHB can be performed using any natural unit (e.g., cases or hospitalizations averted). However, the QALY is a unifying measure that transforms benefits or risks into a single measure, capable of being added or subtracted, or at least compared. QALYs lost from potential risks can be subtracted from QALYs gained from potential benefits to produce an interpretable incremental net health benefit for grading technologies. Imaging tests with negative net QALYs can be labeled inappropriate and those with positive net QALYs, as appropriate. The mathematical properties of INHB, in which a margin can be measured between benefits and risks, aligns well with the ACR AC definition of appropriateness.[26] Garrison *et al.* have proposed that differences in relative uncertainty about the benefits and the risks should also be considered. A wider margin of benefits over risks may be called for when uncertainty about potential risks is higher than that about potential benefits.[27]

Limitations

Decision-analytic modeling cannot easily accommodate qualitative data inputs or unmeasurable but relevant factors. This presents a challenge for incorporation of expert opinion. While experts may have an understanding of relative differences among tests, it may be cognitively challenging for experts to accurately estimate an absolute probability, or a range of probabilities, for each test under evaluation. For example, experts may be able to qualitatively judge differences in indeterminate results or time to diagnosis but may struggle to assign numerical estimates in the form of probabilities. Problems can also arise when decision-makers provide estimates of their preferences in terms of utilities if those utilities are not elicited using valid methods.

Criteria not related to clinical outcomes or health states (e.g., financial incentives or examination time), may require additional assumptions to convert into QALYs or additional steps may be needed to merge patient and provider utilities. Because of these limitations, it is possible that not all relevant trade-offs can be simultaneously included in a single model. However, with sufficient resources and complexity, this limitation can be overcome. Increased complexity is at odds with the primary role of QBRA in this setting, which is to increase transparency. In order to perform INHB, some changes to the ACR AC process may be needed to accommodate additional literature review and analyst time external to the regular meetings.

MCDA

Multi-criteria decision-making (MCDA) originates in decision sciences as a framework for organizing conflicting criteria and comprises several criteria/attribute weighting methods ranging from ranking exercises to choice experiments, either intended to elicit stated or revealed preferences.[28,29] The first aim of MCDA is to help the decision-maker(s) take into consideration the important objective and subjective criteria about the decision problem using an explicit, rational, and efficient decision process. [28,30,31] The second aim of MCDA is to create consensus or "shared understanding of the issue".[28] IMI-PROTECT classified MCDA as not a singular method, but more broadly, as a quantitative framework. Although the IMI-PROTECT defines MCDA as the approach based on multi-attribute theory, MCDA comprises approaches from a number of divergent theories.[32] Likewise, two methods from the IMI-PROTECT inventory listed separately from MCDA—stochastic multi-criteria acceptability analysis (SMAA) and discrete choice experiments (DCE)—are commonly included in the MCDA family of methods.[19,33] We do not provide prescriptive guidance for choosing among MCDA methods due to multiplicity of measurement approaches, nomenclature inconsistencies, and lack of delineation among MCDA methods.[34,35] Instead, we selected three popular MCDA approaches to review.[34] Then, we distinguished differences among

these methods central to the three-fold role of MCDA for incorporation of expert consensus in ACR AC development: 1) judging weights of BRC; 2) scoring performance of tests (in presence of a weak evidence base); and 3) building consensus. We identified desirable elements across MCDA methods as those least burdensome to panel members.

Strengths

A desirable feature of MCDA for use in clinical guideline development is its ability to combine qualitative data, such as expert opinion, with quantitative data. Data from the literature are typically quantitative but in radiology, data can also be qualitative. For example, data can be presented as the degree of breadth and depth of visualization of specific abnormalities, or whether or not a test visualizes function in addition to morphology.[36]

MCDA may offer added value to the existing qualitative Delphi consensus-building approaches by helping participants to articulate values, applying the values rationally and documenting the results across alternative strategies.[37] Consensus is built by working together to decompose the decision problem into its smaller components.[38]

We subsequently show (Chapter 3) that MCDA steps are closely aligned with the existing ACR AC guideline development process and can be performed during meetings (Chapter 3). Thus, adoption of MCDA may be perceived as less disruptive to the existing ACR AC process when compared to INHB.

Limitations

With few exceptions, MCDA methods are additive, comprising positive coefficients and exclusively positive weighted sums.[30,39] Thus, it is not possible to measure net benefit. Instead, the weighted sum, or priority score, can provide a means to rank alternatives. This can be informative to decision-makers interested in comparative effectiveness of imaging tests. However, in order to distinguish appropriate from inappropriate imaging tests, a threshold is needed. For example, a threshold can be set based on a test or other

alternative's performance that is known to be appropriate *a priori* (gold standard) with relative appropriateness discerned by comparing performance of alternatives to the gold standard.

Most MCDA methods operate under the assumption that criteria are independent and do not measure linked effects over time. However, we and others have shown that in diagnostic imaging the presence of interrelated criteria is common (Chapter 1).[40] Indeed, eliciting preferences for related criteria can bias weights by way of double-counting.[41] As an example, in the diagnosis of suspected appendicitis, a potential downstream consequence of *missed cases* is lack of *procedural success* (defined as net number of perforated appendices). This relationship cannot be modeled in a causal way in MCDA. For example, weighting individually, *missed cases* and *procedural success*, especially if *procedural success* is the sole consequence of *missed cases*, results in two weights assigned to essentially one effect: *procedural success*. This problem can be remedied via employing INHB to model effects of missed cases as incremental changes in *procedural success*.

Process for Incorporating Expert Opinion

We propose a process for characterizing the clinical scenario decision problem further than is currently performed by the ACR, with respect to consensus-building or missing, unclear or contradictory evidence. The process involves the following decision aid inputs: selection of relevant benefit-risk criteria, state of the evidence matrix, and criteria linkages.

Decision Aid Inputs

Panel members need several pieces of information in order to use the decision aid. A substantial portion of information needed is already collected as part of the current guideline development process. However, some inputs may not be readily available. We described inputs into the decision aid in the context of the ACR AC process in Table 2.

[Insert Table 2 Here]

Relevant Benefit-Risk Criteria

We have previously shown that benefits and risks in diagnostic radiology are numerous and that the make-up of BRC varies across clinical scenarios (Chapter 1). In Chapter 1 of this dissertation we provide a framework for framing decisions in diagnostic radiology in terms of BRC. Steps for arriving at a list of relevant BRC are provided in the Appendix.

State-of-Evidence Matrix

An important step in structuring the contribution of expert opinion involves identifying any weaknesses in the evidence base. Once relevant BRC are selected, we propose summarizing the quality of the body of evidence in a matrix organized by BRC. Panel members review studies relevant to the clinical scenario and classify whether evidence is available, unclear, contradictory, or missing for each test and BRC (Figure 2). In later steps, panel members populate this matrix with estimates of effects to create a table of variables values (INHB) or a performance matrix (MCDA).[17,42] Weak evidence is defined by high levels of either unclear, contradictory, or missing classifications. Patterns indicative of a weak evidence base across entire columns or rows of this state-of-evidence matrix point to a need to re-consider the tests or criteria included in decision-making (Figure 2).

[Insert Figure 2 Here]

Criteria Linkages

The effects of tests have been previously described as distal to patient outcomes and more local to decisions of healthcare providers.[40] The effects of diagnostics on clinical decision-making may be short lived with few consequences or carry on longer term to significantly influence clinical management and patient outcomes.[43] As an example, we examined each of the test-specific features identified in our previous work (Chapter 1) and based on possible influences of each criterion created pairs with patient management BRC. We repeated this exercise a second and third time, pairing patient management and provider

intrinsic value BRC to patient outcomes and test-specific features directly with patient outcomes. From thirty-six criteria, we created roughly 80-paired criteria relationships. These we visualized using a multi-level mapping Sankey Diagram (Figure 3). Assignment of these linkages is specific to the clinical scenario and to the set of relevant BRC. The extent to which test effects influence patient management and outcomes may be informed by the literature review but, likely, this step will rely heavily on expert consensus.[44] This step is completed by determining whether it is possible to measure all relevant criteria using a single metric (e.g., the QALY).

[Insert Figure 3 Here]

Decision Aid Structure

We combined strengths and limitations of the methods above with the theory that decision problems vary in diagnostic radiology and developed a multi-level decision aid that identifies opportunities for INHB and MCDA methods to aid decision-making in clinical guideline development (Figure 1). The first level ascertains the potential need for QBRA and the second distinguishes whether the decision problem calls for INHB, MCDA, or a joint approach.

[Insert Figure 1 Here]

The decision aid comprises four decision rules (Table 3). The flow of decision rules is illustrated in Figure 4.

[Insert Table 3 Here]

[Insert Figure 4 Here]

Decision Rules

The decision rules apply to clinical scenarios comprising tests that received equivocal ratings for reasons of disagreement or evidence ambiguity, and can be summarized as follows: a)

in the absence of both disagreement and a weak evidence base QBRA may not be needed; b) in the presence of disagreement but absence of a weak evidence base, MCDA approaches are recommended; and c) in the presence of weak evidence base INHB or INHB with MCDA may be recommended, depending on the level of disagreement and the extent of linkages among BRC.

Benefit-risk assessment may not be needed if the following two conditions are met. First, there is no pattern of incomplete, contradictory, or missing evidence. The evidence base exists and is of acceptable quality for each of the tests across all relevant criteria. Panel members can use their judgment as the cut-off for a definition of a weak evidence base. Panel members may choose to make modifications (remove criteria or tests) to simplify the decision problem based on specific patterns of weak evidence base (Figure 2). Second, there is no presence of disagreement marked by patterns of high dispersion in individual ratings. The RAND/UCLA Appropriateness Method Manual prescribes judging excessive dispersion by comparing the difference between the interpercentile range of ratings and the interpercentile range adjusted for symmetry.[3]

When to Use INHB

INHB, as defined in this paper, is best suited to decision problems in which all BRC are interrelated and can be measured either using a single natural unit or in terms of QALYs. Differences in test–specific features may be used to model probabilities of intermediate outcomes, fully to patient outcomes, or any downstream criteria relevant to the decision problem. If not all relevant BRC have been incorporated in the model, MCDA methods may be used to weight the remaining BRC that are measured on different scales. Joint use of these methods has benefits. Performing INHB before MCDA may reduce the need for expert opinion to fill in gaps in evidence and reduces the number of criteria included in the MCDA. Panel members, for example may be weighting missed cases and procedural success among other criteria. Evidence exists for missed cases but little is known about differential rates of

procedural success. Rather than providing opinion about differences in procedural success among tests, probabilities of procedural success can be modeled based on information about missed cases. Of these two criteria, panel members consider only the intermediate outcome, procedural success, in the subsequent MCDA. This approach serves to simplify the MCDA task and reduces the risk of biasing weights by way of double counting related criteria.

Choosing among MCDA methods

We briefly describe the strengths and limitations MCDA approaches that appeared in at least two of the three reviews of MCDA for group decision making.[45-47] Multi-attribute value theory (MAVT), discrete choice experiments/conjoint analysis (DCE), and analytic hierarchy process (AHP) were judged based on considerations: scoring performance of tests (in presence of weak evidence base), judging weights of BRC and building consensus. In Table 4, we provide definition and identify the methods we consider least burdensome to panel members across these criteria. Although Stochastic Multi-Criteria Analysis met initial inclusion criteria (Table 1), this family of methods did not appear in reviews of group decision-making. Nonetheless, SMAA can be applied to several weight elicitation techniques and has been discussed in the context group decision making elsewhere.[18,33]

[Insert Table 4]

As we have mentioned, delineation of methods is often blurred. For example, merging features of AHP and MAVT is possible. For instance, an MCDA, Measuring Attractiveness by a Categorical Based Evaluation TecHnique (MACBETH), uses pairwise comparisons and performance levels to scale criteria weights.[48]

Lastly, a practical consideration may be popularity of a method. AHP and DCEs are more frequently used and cited in the literature.[19,49] Another practical consideration may be choosing a method with a wide offering of software packages. There are many software

solutions that accommodate the diversity of MAVT approaches and markedly fewer variations of the AHP or DCE approaches.[46,50] However, AHP is commonly used for consensus building, thus software packages that facilitate group decision-making are more widely available for AHP.[51] While choosing among approaches may be based on practical considerations, it is important to note these approaches have different theoretical underpinnings and corresponding strengths and limitations, discussed elsewhere.[52] Ultimately, decision-making that occurs on a repeating basis and in high volume, such as the development and revision of ACR AC, may benefit from adopting a single MCDA approach.[13]

DISCUSSION

We focused on two broad QBRA methods that show promise for improving the structure, transparency, and efficiency of the ACR guideline development process in diagnostic imaging. We developed a process to identify the extent and type of expert consensus contribution expected and matched these needs to specific QBRA methods based on their strengths and limitations by way of a decision aid.

There is growing interest in QBRA, and there are many references for benefit-risk methodology that review the theoretical foundations, statistical capabilities, and limitations of these approaches.[11-14] To our knowledge this is the first attempt to consider specifically how QBRA might improve in diagnostic radiology clinical decision-making. This analysis differs from prior reviews in that it examines the potential role of QBRA in improving the contribution of expert consensus in guideline development.

The criteria proposed here for using QBRA focus on addressing potential problems with expert opinion. But the exclusion of other QBRA methods does not imply that excluded methods cannot assist in guideline development. For example, meta-analytic methods have been used jointly with MCDA,[53] and probabilistic simulation modeling commonly

accompanies INHB.[15,54] And this decision aid does not address the third reason ACR names for an equivocal rating assignment, special circumstances. In the future, this process and decision aid could be adapted to characterize the special circumstances panel members encounter and link sources of these special circumstances to QBRA recommendations.

We appreciate that, due to time and budget constraints, the ACR AC are revised only every three years rather than on a continuous basis. We also appreciate that the process uses volunteer-provider panels and does not outsource analyses. These time and resource pressures may work against implementing QBRA. In response, we focused our review on less burdensome versions of approaches. However, QBRA methods can provide decision-makers a model for future panels to use and update, making for more efficient decision-making. Rather than recreating shared understanding among decision-makers, panel members can decide to add, delete, redefine, and reprioritize the set of benefits and risks. Performance of technologies can be re-calibrated based on new evidence and emerging technologies can be added to analyses. In the long-term, these features of QBRA are likely to offset the costs of including QBRA methods experts, as needed, in the ACR AC process.

The QBRA methods that we propose are intended to augment, not replace the existing consensus-building methods used in the development of ACR AC. INHB and MCDA are only as good as the data and expert opinion that inform them. Our work does not delve into the subject of reviewing evidence quality. Expert consensus is considered low quality evidence and high dependence on it can be problematic.[55] Guyatt *et al.* have proposed linking strength of recommendations to evidence quality.[56] Our proposed process leading to the use of the QBRA decision aid could fit nicely into redefining equivocal ratings to include weak and strong recommendations based on the extent of expert consensus contribution.

QBRA is a systematic and consistent approach that promises to increase the transparency of the ACR AC deliberative process. This work further extends investigations of the strengths and limitations of these methods from regulatory benefit-risk assessment to clinical

decision-making in radiology.[14,57] As the structure of the decision aid is based on clinician input but requires the skills of methods experts, the decision aid represents a key component to uniting clinical experts and methodologists.

REFERENCES

1. Cascade PN. The American College of Radiology. ACR Appropriateness Criteria project. *Radiology.* Jan 2000;214 Suppl:3-46.

2. ACR Appropriateness Criteria. ACR ACR Appropriateness Criteria® Rating Round Information. 2015; http://www.acr.org/~/media/ACR/Documents/AppCriteria/RatingRoundInfo.pdf. Accessed August 4, 2015.

3. Fitch K, Bernstein S, Aguilar MD, et al. *The RAND/UCLA appropriateness method user's manual*: Santa Monica, CA : RAND; 2001.

4. Blackmore CC, Medina LS. Evidence-based radiology and the ACR Appropriateness Criteria. *Journal of the American College of Radiology : JACR.* Jul 2006;3(7):505-509.

5. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ.* Jan 13 1996;312(7023):71-72.

6. Lavelle LP, Dunne RM, Carroll AG, Malone DE. Evidence-based Practice of Radiology. *Radiographics : a review publication of the Radiological Society of North America, Inc.* Oct 2015;35(6):1802-1813.

7. Sheng AY, Castro A, Lewiss RE. Awareness, Utilization, and Education of the ACR Appropriateness Criteria: A Review and Future Directions. *Journal of the American College of Radiology : JACR.* Oct 20 2015.

8. Bautista AB, Burgos A, Nickel BJ, Yoon JJ, Tilara AA, Amorosa JK. Do clinicians use the American College of Radiology Appropriateness criteria in the management of their patients? *AJR Am J Roentgenol.* Jun 2009;192(6):1581-1585.

9. Cross JT, Garrison LP, Jr. Challenges and opportunities for improving benefit-risk assessment of pharmaceuticals from an economic perspective. *OHE Briefing.* 2008.

10. Hughes D, Waddingham EDJ, Mt-Isa S, et al. *IMI-PROTECT Benefit-Risk Group RECOMMENDATIONS REPORT (Work Package 5): Recommendations for the*

methodology and visualisation techniques to be used in the assessment of benefit and risk of medicines: European Medicines Agency and Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium (PROTECT);2013.

11. Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch DW. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy-report of the ISPOR risk-benefit management working group. *Value Health.* Aug 2010;13(5):657-666.

12. Puhan MA, Singh S, Weiss CO, Varadhan R, Boyd CM. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med Res Methodol.* Nov 19 2012;12(1):173.

13. Mt-Isa S, Hallgreen CE, Wang N, et al. Balancing benefit and risk of medicines: a systematic review and classification of available methodologies. *Pharmacoepidemiol Drug Saf.* May 13 2014.

14. Mt-Isa S, Ouwens M, Robert V, Gebel M, Schacht A, Hirsch I. Structured Benefit-risk assessment: a review of key publications and initiatives on frameworks and methodologies. *Pharm Stat.* May 15 2015.

15. Lynd LD, Najafzadeh M, Colley L, et al. Using the incremental net benefit framework for quantitative benefit-risk analysis in regulatory decision-making--a case study of alosetron in irritable bowel syndrome. *Value Health.* Jun-Jul 2010;13(4):411-417.

16. Garrison LP, Jr., Towse A, Bresnahan BW. Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. *Health Aff (Millwood).* May-Jun 2007;26(3):684-695.

17. Kuntz KM, Tsevat J, Weinstein MC, Goldman L. Expert panel vs decision-analysis recommendations for postdischarge coronary angiography after myocardial infarction. *JAMA.* 1999;282(23):2246-2251.

18. Lahdelma R, Salminen P. SMAA-2: Stochastic Multicriteria Acceptability Analysis for Group Decision Making. 2001;49(3):444-454.

19. Marsh K, Lanitis T, Neasham D, Orfanos P, Caro J. Assessing the value of healthcare interventions using multi-criteria decision analysis: a review of the literature. *Pharmacoeconomics.* Apr 2014;32(4):345-365.

20. Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Medical decision making : an international journal of the Society for Medical Decision Making.* Sep-Oct 2012;32(5):667-677.

21. Habbema JD, Wilt TJ, Etzioni R, et al. Models in the development of clinical practice guidelines. *Annals of internal medicine.* Dec 2 2014;161(11):812-818.

22. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *BMJ.* 2011;342:d1766.

23. Johnson FR, Hauber AB, Ozdemir S, Lynd L. Quantifying women's stated benefit-risk trade-off preferences for IBS treatment outcomes. *Value Health.* Jun-Jul 2010;13(4):418-423.

24. Drummond M, Brixner D, Gold M, Kind P, McGuire A, Nord E. Toward a consensus on the QALY. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* Mar 2009;12 Suppl 1:S31-35.

25. Ollendorf DA, Blackmore CC, Lee JM. Toward evidence-based decisions in diagnostic radiology: a research and rating process for multiple decision-makers. *Acad Radiol.* Sep 2012;19(9):1049-1054.

26. Sistrom CL. The appropriateness of imaging: a comprehensive conceptual framework. *Radiology.* Jun 2009;251(3):637-649.

27. Garrison LP. Regulatory benefit-risk assessment and comparative effectiveness research: strangers, bedfellows or strange bedfellows? *Pharmacoeconomics.* 2010;28(10):855-865.

28. Belton VSTJ. *Multiple criteria decision analysis : an integrated approach*. Boston: Kluwer Academic Publishers; 2002.

29. Hunink MGMGPPSJE. *Decision making in health and medicine : integrating evidence and values*. Cambridge: Cambridge University Press; 2001.

30. Linkov I, Moberg E. Multi-criteria decision analysis environmental applications and case studies. 2012; http://public.eblib.com/choice/publicfullrecord.aspx?p=952014.

31. Guitouni A, Martel J-M. Tentative guidelines to help choosing an appropriate MCDA method. *European Journal of Operational Research.* 1998;109(2):501-521.

32. Figueira J, Greco S, Ehrgott M. *Multiple criteria decision analysis : state of the art surveys*. New York: Springer; 2005.

33. Tervonen T, Figueira JR. A survey on stochastic multicriteria acceptability analysis methods. *MCDA Journal of Multi-Criteria Decision Analysis.* 2008;15(1-2):1-14.

34. Adunlin G, Diaby V, Xiao H. Application of multicriteria decision analysis in health care: a systematic review and bibliometric analysis. *Health Expect.* Oct 18 2014.

35. Zopounidis C, Pardalos PM. *Handbook of multicriteria analysis*. Berlin; Heidelberg: Springer-Verlag; 2010.

36. Haakma W, Steuten LM, Bojke L, MJ IJ. Belief elicitation to populate health economic models of medical diagnostic devices in development. *Applied health economics and health policy.* Jun 2014;12(3):327-334.

37. Phillips L. *Chapter 19: Decision Conferencing* London: London School of Economics and Political Science 2006.

38. Phillips L, Bana e Costa C. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. *Annals of Operations Research.* 2007;154(1):51-68.

39. Yatsalo B, Didenko V, Gritsyuk S, Sullivan T. Decerns : A Framework for Multi-Criteria Decision Analysis. *International Journal of Computational Intelligence Systems.* 2015;8(3):467-489.

40. Bresnahan BW, Garrison LP. Economic Issues in Diagnostic Imaging. In: Culyer T, ed. *Encyclopedia of Health Economics*. Oxford: Elsevier; 2014.

41. Thokala P, Duenas A. Multiple criteria decision analysis for health technology assessment. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* Dec 2012;15(8):1172-1181.

42. Dodgson J, Spackman, M, Pearman, A and Phillips, LD,. *Multi-criteria analysis: a manual* London: London School of Economics;2009.

43. Staub LP, Lord SJ, Simes RJ, et al. Using patient management as a surrogate for patient health outcomes in diagnostic test evaluation. *BMC Med Res Methodol.* 2012;12:12.

44. Staub LP, Dyer S, Lord SJ, Simes RJ. Linking the evidence: intermediate outcomes in medical test assessments. *Int J Technol Assess Health Care.* Jan 2012;28(1):52-58.

45. Peniwati K. Criteria for evaluating group decision-making methods. *Mathematical and Computer Modelling.* 2007;46(7):935-947.

46. Hummel JM, Bridges JF, MJ IJ. Group decision making with the analytic hierarchy process in benefit-risk assessment: a tutorial. *Patient.* 2014;7(2):129-140.

47. Marsh K, Devlin N, Ijzerman MJ, Thokala P. MCDA for Health Care Decisions - Emerging Good Practices: Report 2 of the ISPOR MCDA Task Force. *TBD.* 2016;TBD(TBD).

48. Bana e Costa CA, Vansnick J-C. MACBETH ? An interactive path towards the construction of cardinal value functions. *International Transactions in Operational Research.* 1994;1(4):489-500.

49. Liberatore MJ, Nydick RL. The analytic hierarchy process in medical and health care decision making: A literature review. *European Journal of Operational Research.* 2008;189(1):194-207.

50. French S, Xu D-L. Comparison study of multi-attribute decision analytic software. *MCDA Journal of Multi-Criteria Decision Analysis.* 2005;13(2-3):65-80.

51. Ishizaka A, Nemery P. *Multi-criteria Decision Analysis : Methods and Software*. Somerset, NJ, USA: John Wiley & Sons; 2013.

52. Saaty TL. Axiomatic Foundation of the Analytic Hierarchy Process. *Management Science.* 1986;32(7):841-855.

53. Naci H, van Valkenhoef G, Higgins JP, Fleurence R, Ades AE. Evidence-based prescribing: combining network meta-analysis with multicriteria decision analysis to choose among multiple drugs. *Circ Cardiovasc Qual Outcomes.* Sep 2014;7(5):787-792.

54. Puhan MA, Yu T, Stegeman I, Varadhan R, Singh S, Boyd CM. Benefit-harm analysis and charts for individualized and preference-sensitive prevention: example of low dose aspirin for primary prevention of cardiovascular disease and cancer. *BMC Med.* 2015;13:250.

55. Evans D. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *J Clin Nurs.* Jan 2003;12(1):77-84.

56. Guyatt GH, Oxman AD, Kunz R, et al. Going from evidence to recommendations. *BMJ.* May 10 2008;336(7652):1049-1051.

57. Agapova M, Devine EB, Bresnahan BW, Higashi MK, Garrison LP, Jr. Applying quantitative benefit-risk analysis to aid regulatory decision making in diagnostic imaging: methods, challenges, and opportunities. *Academic radiology.* Sep 2014;21(9):1138-1143.

58. Habibzadeh F, Yadollahie M. Number needed to misdiagnose: a measure of diagnostic test effectiveness. *Epidemiology.* Jan 2013;24(1):170.

59. Til J, Dolan J, Stiggelbout A, Groothuis K, Ijzerman M. The Use of Multi-Criteria Decision Analysis Weight Elicitation Techniques in Patients with Mild Cognitive Impairment. *The Patient: Patient-Centered Outcomes Research.* 2008;1(2):127-135.

60. Lootsma FA. *Multi-Criteria Decision Analysis via Ratio and Difference Judgment.* Vol 29. Dordrecht, Netherlands: Kluwer Academic Publishers; 1999.

61. Saaty RW. The analytic hierarchy process—what it is and how it is used. *Mathematical Modelling.* 1987;9(3–5):161-176.

TABLES

Table 1. Inclusion Criteria for QBRA Decision Aid

Method Selection Definitions	Feature 1		Feature 2	Feature 1: Method weights and transforms different endpoints to a single scale to arrive at an integrated benefit-risk measure Feature 2: Method demonstrates the effects on the benefit-risk measure if weights or test effects are varied
IMI-PROTECT Definitions	Weighting	Integrated benefit-risk metric	Sensitivity Analysis	Comments
MCDA	✓	✓	✓	Several approaches possible for weighting
SMAA	✓	✓	✓	
NNH/NNT*	✓	✓		Feature 1: Some modifications to NNH and NNT include relative value, RV-NNH. Relative value incorporates preferences for multiple risks in order to create a composite NNH. No built-in mechanism for eliciting preferences (relative values).[12] Modification of NNT is utility and time-adjusted UT-NNT. This modification delivers a benefit-risk trade-off.[13] UT-NNT resembles Q-TWIST.

				Feature 2: Possible but not well-established in the literature.[13]
Impact numbers				Feature 1: Measures individual thresholds[13] for each effect, individually
Q-TWIST	✓	✓		Feature 2: Possible but not well-established in the literature[13]
QALY	✓	✓		Feature 2: Discrete metric, needs to be combined with a method
INHB	✓	✓	✓	Feature 1: Several approaches possible for weighting
PSM	✓		✓	Feature 1: Estimation technique, needs to be combined with a metric
ITC	✓		✓	Feature 1: Estimation technique, needs to be combined with a metric
DCE	✓	✓	✓	Feature 1:Several approaches possible for weighting

MCDA: Multi-criteria Decision Analysis; ; SMAA: Stochastic Multi-criteria Acceptability Analysis; NNH/NNT: Number Needed to Harm/Treat; Q-twist: Quality-adjusted Time Without Symptoms and Toxicity; QALY: Quality-Adjusted Life-Years; INHB: Incremental Net Health Benefit; PSM: Probabilistic Simulation Model; ITC: Indirect Treatment Comparisons; DCE: Discrete Choice Experiments

*Numbers needed to diagnose (NND) and misdiagnose (NNM) apply to diagnostic tests[58]

Table 2. ACR AC Current Process

Input	Publicly Available	Not Publicly Available May need to be collected
ACR AC Clinical scenario	The topic and variant describing a specific clinical indication for a diagnostic test	
Imaging tests to be evaluated	Imaging tests that appear in the ACR AC ratings	
Consensus ratings	Ratings assigned by panel consensus to each test	
Relevant benefit-risk criteria selection	*Benefit and risks as formal variables are not currently part of the ACR AC documentation*	A list of benefit and risk criteria relevant for this decision-problem
State of evidence matrix	*ACR AC panels collect and evaluate the quality of evidence but do not characterize the state of the evidence for each criterion and each test.*	A matrix characterizing unclear and missing evidence for criteria across tests
Criteria linkages	*Summary narrative of the ACR AC may implicitly link sensitivity and specificity to intermediate or patient outcomes*	Related criteria are linked to create a directed acyclic graph of the decision problem
Characterization of group consensus	*Disagreement among panel members may lead to equivocal ratings*	Report of dispersion for each rating
Italicized text indicates the portion of information publicly available		

ACR AC: American College of Radiology Appropriateness Criteria

Summary Narrative: discussion of the literature review

Equivocal ratings: ratings ranging from 4 to 6, interpreted as test "may be appropriate"

Dispersion: high levels of variation between median ratings and individual ratings

Table 3. Decision Aid Decision Rules

Decision Rule Questions	Decision Rule Answer: Yes	Decision Rule Answer: No
1. State-of-evidence matrix indicates patterns of weak evidence?	QBRA will be used to bridge evidence gaps as illustrated in the state-of-evidence matrix, BRA may also be needed to collect preferences for risks and benefits.	QBRA may be needed only to help panel members weigh benefits and risks.
2. Comparison of individual to consensus ratings indicates disagreement?	QBRA is needed to help panel members weight benefits and risks. MCDA is recommended	QBRA is not needed to help panel members weight benefits and risks.
3. Panel members have identified criteria linkages?	QBRA is needed to capture effects over time, INHB is recommended. Using MCDA may lead to double counting of related criteria	QBRA does not need to capture effects over time, MCDA may be used to capture weights and/or assign performance to tests, if applicable.
4. Criteria can be measured using a single endpoint?	INHB measures BRC using single endpoint (e.g., cases averted or the QALY)	INHB is recommended for criteria that can measured using a single endpoint. MCDA is recommended for weighting the importance of INHB results against remaining criteria.

Table 4. Factors to consider when choosing among MCDA methods for incorporation of expert opinion in decision-making

Considerations	MAVT SWING	DCE/Conjoint Analysis	AHP
Scoring performance using expert consensus[39] Desirable: low cognitive burden	Assumption about the absolute performance or normalized performance of test,	Assignment of test to a pre-specified level of an attribute	Assignment of relative performance on a ratio scale
The presentation of weighting tasks[39,59] Desirable: question is formulated intuitively	Scoring: assigning priority to criteria proportional to possible improvements in performance	Choice between two sets of multiple criteria, single levels of criteria	Rating of two criteria using multiple levels (1-9) of comparative preference
Group maintenance: leadership effectiveness[45] Desirable: building consensus	Some	None	High

Gray color indicates MCDA method performs less than desirably

(MAVT SWING) Multi-Attribute Value Theory using swing weights: Swing weighting is more advanced weighting approach in which differences in criteria scales are taken into account. Performance of alternatives is referenced to create hypothetical a worst and a best alternative. The swings in performance from best to worst inform criteria weights. The criteria with the most important improvement from worst to best is scored 100. Remaining criteria are scored, $0<x<100$, based on the relative importance of its swing. Thus, weights of an otherwise high-priority criterion

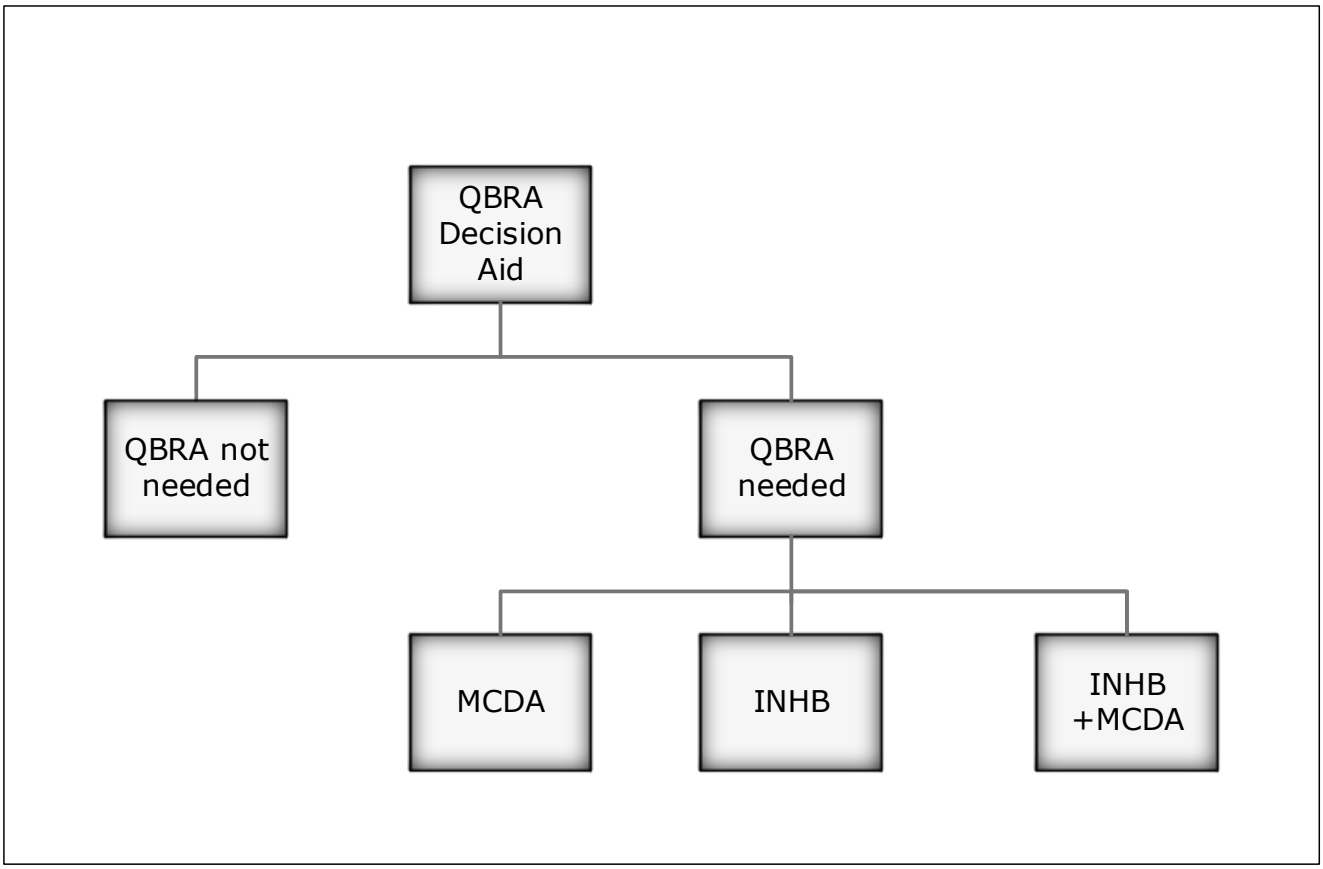
and a lesser criterion are calibrated, if for example, nominal differences in performance among alternatives define the former while large differences are present in the latter.

(DCE) Discrete Choice Experiments: DCE elicit preferences using hypothetical scenarios but rather than rating or ranking, the subject chooses the preferred scenario between two scenarios. Conjoint Analysis is a group of techniques that elicits from subjects preferences by formulating questions that require subjects to rank or rate hypothetical scenarios (sometimes cross-labeled as DCEs). Each scenario is a reasonable combination of attributes that represents a realistic medical intervention alternative and differs slightly in the levels of attributes.

(AHP) Analytic Hierarchy Process: AHP is an elicitation method for preferences in the MCDA group of methods. AHP does not provide direct comparison across criteria. Instead criteria are grouped into clusters of pairs for pairwise comparison.[60] Weights are elicited when criteria are compared pairwise using a scale (1-9) and then matrix algebra is employed, via the eigenvector.[61] After criteria are weighted pairwise, each alternative is pairwise weighted within each criteria and this matrix is multiplied by criteria weights.

FIGURES

Figure 1. Levels of the QBRA Decision Aid



QBRA: quantitative benefit-risk assessment
MCDA: multi-criteria decision-analysis
INHB: incremental net health benefit

Figure 2. Examples of State-of-Evidence Matrices

Matrix Example 1	TEST1	TEST2	TEST3	TEST4	Recommendations
BRC1					Removal of BRC1 may minimize extent of expert opinion required
BRC2					
BRC3					
BRC4					
Matrix Example 2	**TEST1**	**TEST2**	**TEST3**	**TEST4**	
BRC1					Removal of TEST3 may minimize the extent of expert opinion required
BRC2					
BRC3					
BRC4					
Matrix Example 3	**TEST1**	**TEST2**	**TEST3**	**TEST4**	
BRC1					Patterns consistent with a weak evidence base
BRC2					
BRC3					
BRC4					
Grey cells indicate evidence is missing, contradictory, or unclear; BRC: benefit-risk criterion					

Figure 3. Sankey Diagram Depicting Relationships among Criteria

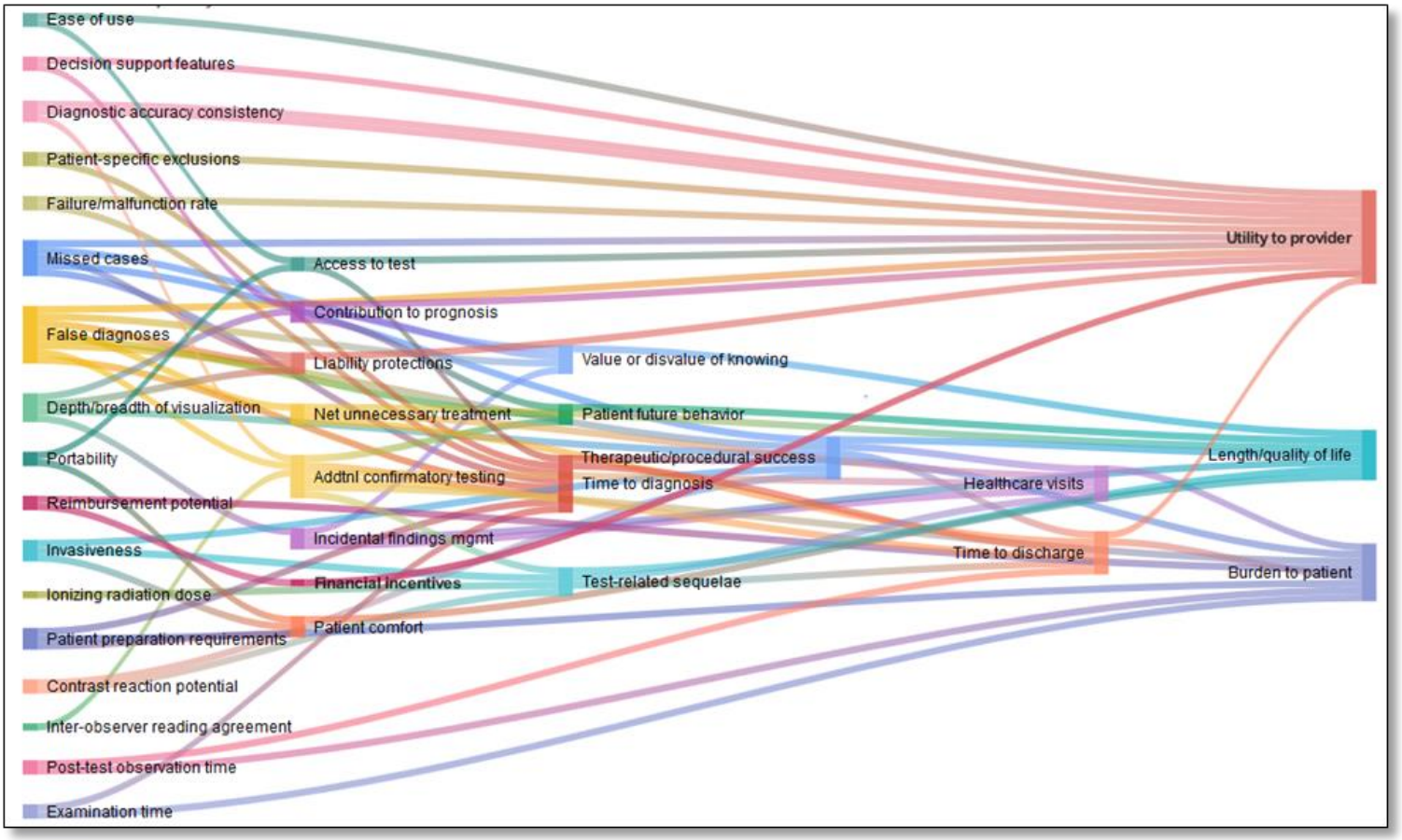
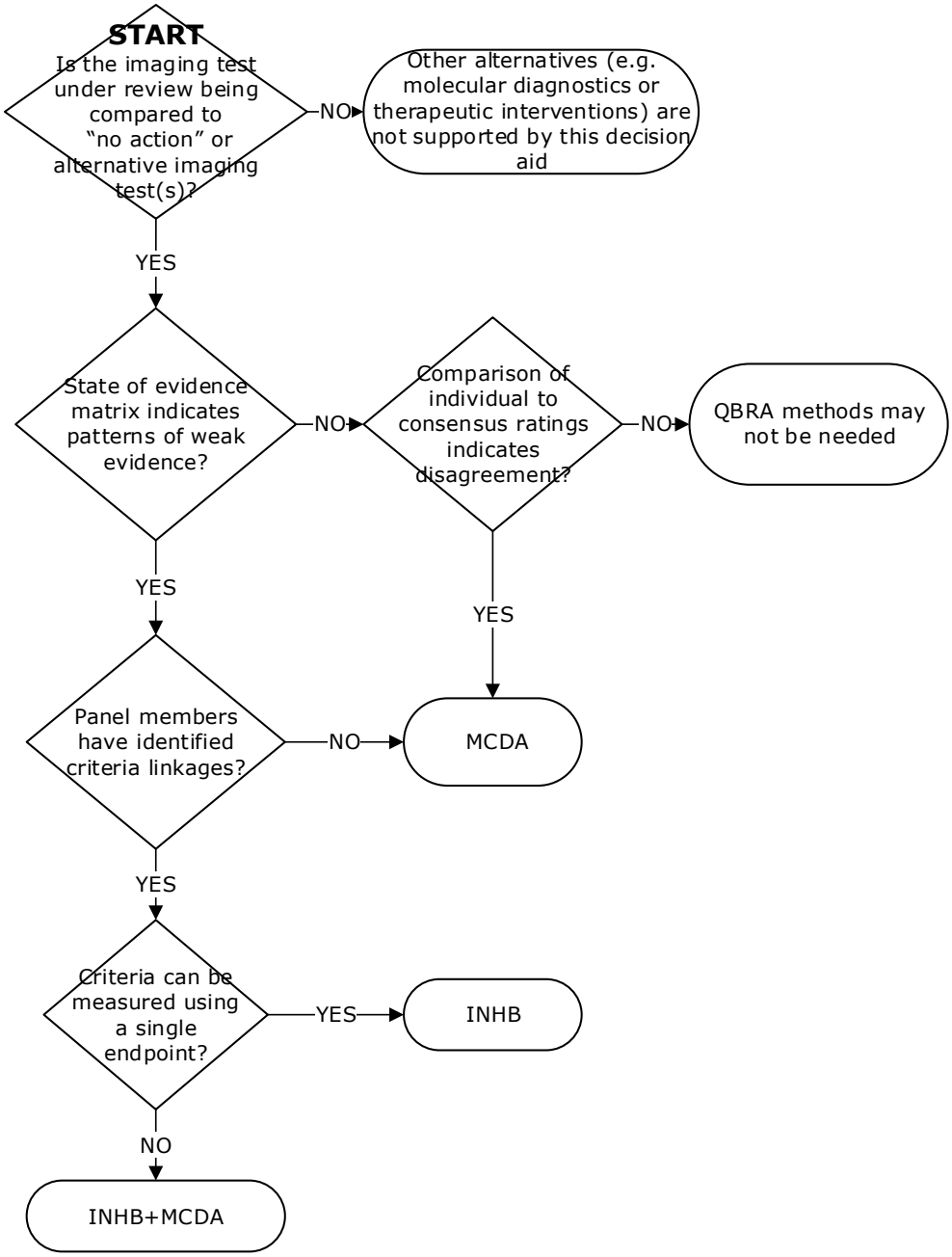


Figure 4. Decision Aid Flow Chart for QBRA Selection



QBRA: quantitative benefit-risk assessment
MCDA: multi-criteria decision-analysis
INHB: incremental net health benefit

APPENDIX

Consensus selection of relevant benefit-risk criteria

We propose a belief-based consensus process for selection of relevant benefit-risk criteria (BRC). Either the use of a pre-defined list of BRC (presented in Chapter 1 of this dissertation) or criteria drawn from a different source can be used. Figure 1 provides an illustrative example using the three domains and thirty-six BRC previously presented.

Step 1: Pairwise Comparisons based on Individual Beliefs

Individual panel members compare two tests at a time and select the criteria that they believe differ between the two tests. This step is repeated until all tests have been compared. Panel members vote on criteria that represent differences (for them) between tests; selection is not based on magnitude or clinical significance of differences.

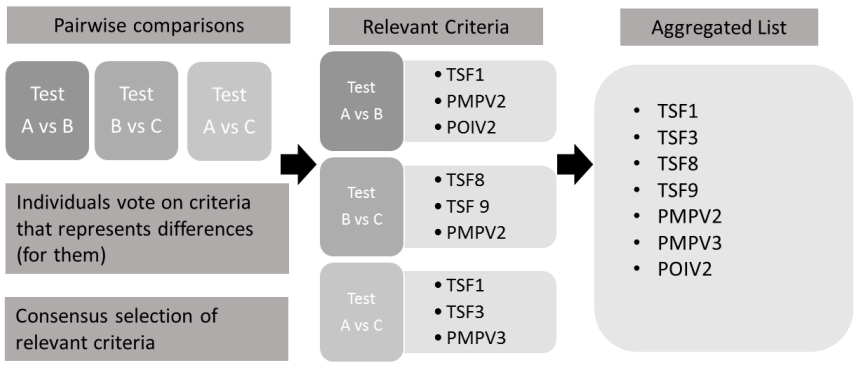
Step 2: Consensus Selection of Relevant Criteria

A member the panel or staff compiles criteria across all pairwise comparisons. Relevant criteria is defined as that meeting a consensus threshold (e.g., 80% of panel members voted for criterion) for at least one pairwise comparison.

Step 3: Group Review of the Aggregated List

A member of the panel or staff removes redundant criteria across pairwise comparisons to create a list of BRC aggregated across all pairwise comparisons. This list is presented to the panel. Panel members discuss the list, as a group decide whether the full list or an abbreviated list will be used to carry out the benefit-risk assessment. Panel members review each BRC to define and come to a shared understanding of the BRC definition.

Appendix Figure 1: Proposed process for selection of relevant BRC



TSF: Test-specific feature
PMPV: Patient-management provider intrinsic value
POIV: Patient outcome and intrinsic value

Chapter 3. Analytic Hierarchy Process for Prioritizing Imaging Tests in Diagnosis of Suspected Appendicitis

Maria Agapova, Brian Bresnahan, Louis Garrison, Mitchell Higashi, Larry Kessler, Kenneth Linnau, Beth Devine

Target Journal: Academic Radiology

Word count: 3,778

Tables: 1

Figures: 4

ABSTRACT

Abstract Word Count: 243 (including headings)

INTRODUCTION

Clinical guideline development, such as the American College of Radiology Appropriateness Criteria (ACR AC) process, asks experts to evaluate benefits and risks associated with imaging tests and to make complex decisions. The Analytic Hierarchy Process (AHP) decomposes complex decisions into structured smaller decisions, incorporates quantitative evidence and qualitative expert opinion and promotes structured consensus-building.

METHODS

We convened a mock ACR AC panel of Emergency Department radiology and non-radiology physicians to evaluate by multi-criteria decision analysis the relative appropriateness of imaging tests for diagnosing suspected appendicitis. Panel members selected benefit-risk criteria via an online survey and assessed contrast-enhanced computed tomography, magnetic resonance imaging and ultrasound using an AHP-based software. Participants were asked whether the process was manageable, transparent, and improved shared-understanding. Priority scores were converted to rankings and compared to the rank order of ACR AC ratings.

RESULTS

When compared to magnetic resonance and ultrasound imaging, participants agreed with the ACR AC that contrast-enhanced computed tomography is the most appropriate test. Contrary to the ACR AC ratings, study results suggest that magnetic resonance is preferable to ultrasound. When compared to non-radiologists, radiologists' priority scores reflect a stronger preference for computed tomography.

CONCLUSIONS

Results suggest that AHP may benefit the ACR AC guideline development process in identifying the relative appropriateness of imaging tests. Study participants addressed decision-making challenges at nominal expense in time and financial resources. With additional development, AHP may be a means for transparent inclusion of expert opinion in clinical guideline development.

INTRODUCTION

The American College of Radiology (ACR) publishes evidence-based guidelines for appropriate use of imaging tests. [1] ACR Appropriateness Criteria (ACR AC) are developed and revised every three years by panels composed of 10-16 volunteer ACR-members.[2,3] Panel members rate imaging tests on a scale that ranges from 1-9 (1-3 inappropriate, 4-6, equivocal; 7-9, appropriate) using the RAND/UCLA Appropriateness Method.[4] The ACR AC quantitative ratings represent qualitative reconciliation of benefits and risks into one measure that supports the ultimate decision of the panel of radiology experts. For a given clinical scenario, referred to as a topic variant, panel members assess the risks of each test against the benefits of performing the procedure. Expert opinion fills evidence gaps and supplements existing evidence.[3]

Decision scientists have shown that individuals struggle with complex decisions involving multiple objectives with uncertain trade-offs.[5] As the number of alternatives and criteria judgments increases, individuals' decision-making capabilities degrade.[6] In this context, we have identified that ACR AC expert panel members may face several challenges: 1) rating multiple imaging alternatives for any given clinical indication; 2) assignment of importance to multiple potential benefits and risks; 3) as volunteers, working with limited financial resources and time; 4) decision-making in an environment of high uncertainty with regard to benefits and risks across alternatives; and 5) given divergent views, arriving at a single metric representing the benefit-risk balance, or appropriateness. While the modified Delphi

consensus approach of the RAND/UCLA Appropriateness Method assists panel members with reaching consensus, decision support for decomposing complex decisions, individually or in groups, is lacking.

Multi-criteria decision analysis (MCDA) methods are one of many benefit-risk assessment approaches.[7,8] MCDA methods are particularly useful in organizing and weighting of multiple, often conflicting, criteria.[9] With underpinnings in operations research and decision theory, MCDA is now applied broadly in healthcare research.[10]

There are several reasons for choosing MCDA to facilitate ACR AC development. First, MCDA is well suited to meet the need in diagnostic imaging to merge quantitative, qualitative evidence and expert opinion. Second, MCDA assists participants in articulating values, applying the values rationally, and documenting the results across alternative strategies. Third, the steps in an MCDA are similar to steps taken by ACR AC panels (Figure 1). The added structure of MCDA methods promises to improve the efficiency of meetings to the extent that the marginal time the MCDA process consumes becomes negligible. Fourth, if decision-makers are unsure about their judgments, varying decision-makers' inputs in order to determine whether results are robust across a plausible range of values for one or more inputs can be assessed in sensitivity analyses. If decision-makers question criteria selection or definitions, the robustness of results to structural uncertainty can be investigated by conducting scenario analyses and varying criterion definitions or hierarchy structure. Fifth, when a lack of consensus requires additional Delphi consensus rounds, MCDA may offer an alternative approach to decomposing the decision problem and building consensus piecemeal. Two methods within MCDA—the Analytic Hierarchy Process (AHP) and Multiple Attribute Utility (Value) Theory Analysis (MAUT/MAVT)—are designed to facilitate group decision-making.[11]

While, in theory, MCDA methods appear well-suited for the problems faced by ACR AC, the feasibility of using MCDA in ACR AC guideline development has not been previously explored

or tested. Using the information gathered for ACR AC deliberations, we sought to explore whether MCDA analyses would yield comparable results while adding structure, transparency, and efficiency to the process. We selected for the use case an existing ACR AC clinical scenario: the diagnosis of lower quadrant pain, suspected appendicitis. In the United States, computed tomography (CT) is considered the gold standard for diagnosis of the classical presentation of suspected appendicitis in adults.[12] Magnetic resonance imaging (MRI) is a confirmatory test for equivocal findings by ultrasound or CT. Ultrasound is the first-line diagnostic for those most vulnerable to ionizing radiation exposure, children and pregnant women.[14] However, in campaigns to reduce patient exposure to ionizing radiation, some propose expanding this policy to all patients.[13,14] In Europe, standard practice is to use ultrasound first.[15] Further, some propose MRI as an alternative to CT when ultrasound findings are equivocal.[16] It is unclear which modality has the most favorable benefit-risk profile. The objective of this study was to use the AHP approach to assess the relative appropriateness of these diagnostic imaging tests in diagnosing suspected appendicitis.

METHODS

Selection and definition of clinical use case

Selection of an ACR AC clinical scenario was limited to four clinical scenarios previously studied in the context of benefit-risk determination in clinical guideline development. (Chapter 1) An ongoing debate about which modality is best for diagnosing suspected appendicitis made this use case the best candidate for investigating relative appropriateness using MCDA. We used the existing ACR AC case definition for appendicitis, last updated in 2013: Lower quadrant pain-suspected appendicitis (NGC-10146) variant 1: A patient arrives complaining of lower quadrant pain. Fever, leukocytosis and other signs point to a classic case of clinical appendicitis. The ten diagnostic modalities assessed by the ACR AC panel are listed on the ACR AC website.[17] We abstracted information from the following ACR AC supporting documentation: 1) ACR AC narrative written by a panel member serving as the

topic author describing the evidence base used in the decision-making process (Appendix C); 2) The evidence table of the studies cited in the narrative portion of the ACR AC including details of study design, summary of study results, and an evaluation of study quality.[18]

Selection of MCDA method

We chose AHP because pairwise comparisons are intuitive and cognitively less burdensome than using direct elicitation. We selected among software packages that featured real time voting, those that can be used on a participant's personal device and those that offered an affordable academic license. Scoring imaging tests in relative terms, a feature of AHP, also obviates the need to assign each imaging test a measure of performance: this is a strong advantage when limited evidence is supplemented with expert opinion.

Mock ACR AC panel participant recruitment

The recruitment population pool comprised physicians with a clinical specialty in emergency medicine, radiologists and non-radiologists, who had participated in a previously conducted, related study, and who had expressed interest in participating in the mock ACR AC Panel Activity. Participants received an invitation to participate via email letter. This study received approval from the University of Washington Institutional Review Board.

Selection of imaging technologies

From the ten technologies evaluated in the AC, three were selected for evaluation. With guidance from a radiologist (KL), we selected: (1) CT of the abdomen and pelvis with contrast (ceCT); (2) ultrasound of the abdomen; and (3) non-contrast MRI. In the US, ceCT and ultrasound are used most frequently to diagnose suspected appendicitis although the popularity of MRI, as an alternative to ceCT, is increasing.[16]

Pre-meeting survey

Before convening, participants completed an online survey containing a list of benefit-risk criteria specific to diagnostic imaging (Appendix B). Participants selected criteria they believed differed across each pair of imaging tests compared, and their votes were aggregated. A criterion was included in the analysis if it was selected by more than 75% of participants, or 50%, in the event no criteria within the domain reached the 75% selection threshold.

Estimates of effects were abstracted from studies identified in the ACR AC evidence table that accompanied the 2013 ACR AC review of the lower right quadrant pain topic. Pooled estimates of effects were preferred, if meta-analysis were available. Performance matrices were created in which criteria were listed in rows and imaging tests in columns (Appendix A). We created a separate performance table to accommodate the large volume of results for the criterion *missed cases*. We also presented positive and negative predictive values and numbers needed to diagnose. Criteria with no evidence were identified by blank fields.

Mock ACR AC meeting

Participants served as mock ACR AC panel members, and met once for three hours. A methods expert (MA) moderated the session. Participants reviewed the results of the survey and performance matrices, and were given time to discuss included/omitted criteria, definitions of criteria, and the structure of the AHP model.

In the latter half of the meeting, panel members performed the AHP activity using a web-enabled software package, TransparentChoice (Cheadle, United Kingdom). The voting was performed individually, using personal devices, but progression was synchronized and controlled by the moderator. Panel members were first presented with criteria in pairwise fashion and asked to select the more important criterion based on their preference, and to

assign a magnitude of relative priority on a scale from 1-9. Next, within each criterion, alternatives were scored using the same pairwise process and scale. The moderator reported the spread and geometric mean of the votes. The group was then given a chance to discuss and reach consensus. Individual and group votes were recorded.

The software generates weights of criteria, priority scores, inconsistency ratios, and sensitivity analyses. Weights of criteria are reported as local weights, across each level of hierarchy; and also as global weights for the lowest tier criteria. Priority scores are reported using consensus scores, geometric means of individual scores[19], and geometric means, stratified by clinical specialty.

We measured the inconsistency ratio to evaluate whether the assumption of transitivity holds across stacked sets of comparisons (e.g., for a given criterion, if test X performs 5-fold better than tests Y and Z, tests Y and Z should be considered equivalent in future comparisons).[20] A ratio of zero indicates perfect judgment, and a ratio of one indicates judgments akin to random selections. Sensitivity analyses were performed to investigate whether changes in criteria weights resulted in rank reversals.[21]

At the end of the session, panel members were provided with the results and answered several open-ended questions about their experience. We describe how criteria selections and ranks of technologies were similar or different to content of the ACR AC narrative and ratings, respectively.

RESULTS

Mock ACR AC panel participants

Nine University of Washington physicians (8 male, 1 female) participated in the online survey and mock ACR AC activity. Radiologists (44%) included 1 body imager and 3 emergency department (ED) radiologists. Non-radiologists (56%) were exclusively ED physicians. ED physicians were more experienced (mean=11 years, SD= 6.2) than ED

radiologists (mean=3 years, SD=2.5). One participant had previously served on a panel similar to the ACR AC.

Pre-meeting survey: selection of relevant criteria

All nine mock ACR AC panel members completed the survey and identified one or more differences in test-specific features among the three imaging tests under evaluation. Six respondents affirmed that differences likely extended to effects on patient management and four affirmed these differences extended to patient outcomes (Appendix A). Using the pre-defined consensus thresholds, seven test-specific features, six patient management, and one patient outcome criteria were selected. The finalized AHP hierarchy included two broad goals, two sub-goals and nine lower-tier criteria (Figure 2). The goals were: 1) decreasing time to diagnosis, and 2) decreasing potential risks to patient. Decreasing time to diagnosis was further stratified into a second-level tier: 1) increasing provider utility and 2) decreasing burden (time and money) to patient. The panel discussed the clustering of the lowest tier criteria denoted in italics (potential risks to patient: *missed cases*, *contrast reaction potential*, *ionizing radiation dose*; provider utility: *access to test*, *diagnostic accuracy consistency* and *patient-specific exclusions*; and burden to patient: *examination time*, *incidental findings management*, *potential for additional confirmatory testing*). Panel members defined provider utility as the question, "How useful is this test to me?"

Comparison of mock panel criteria selection to the ACR AC considerations

Participants' selection of benefit-risk criteria was similar to key words abstracted from the ACR AC narrative with a few exceptions. Participants did not select the criterion *contribution of information to prognosis*, but in the ACR AC narrative, CT was identified as superior to ultrasound in identification of complications. Negative appendectomies, were mentioned in the ACR AC narrative. When compared to clinical diagnosis, CT was associated with lower negative appendectomy rates. ACR AC narrative did not provide estimates of negative

appendectomy rates comparing ultrasound or MRI to clinical diagnosis. Mock ACR AC panel participants did not select the *net unnecessary treatment* criterion.

The data presented in the performance matrices also differed from the data presented in the narrative (Appendix A). First, the data were presented in a matrix, not in narrative form. In order to normalize data across studies, we provided positive and negative predictive values and numbers needed to diagnose to participants. Mock panel participants also reviewed results from two additional meta-analyses published before 2012 but that were not included in the ACR AC evidence table.

AHP voting results

Criteria Weights

Criteria weights for each tier of the hierarchy are provided in Figure 3. Panel members judged Goals 1 and 2 to be equally important. Within Goal 1, provider utility was judged more important than burden to patient (by roughly 3-fold). Within each sub-goal, minimizing potential risks to patient and burden (time and money) to patient, maximizing provider utility and participants weighted highest *missed cases*, *additional confirmatory testing* and *diagnostic accuracy consistency*, respectively. Radiologists placed heavier weight on the sub-goal *provider utility* of test, comprising *patient specific exclusions*, *diagnostic accuracy consistency* and *access to test*, than non-radiologists. When compared to radiologists, non-radiologists weighted highly, the criteria belonging to the minimizing potential risks to patients goals, comprising *ionizing radiation dose*, *missed cases* and *contrast reaction potential*.

Imaging Test Priority Scores

We present the AHP consensus and geometric mean priority scores of the group as well as the geometric mean scores stratified by clinical specialty in Figure 4. The consensus scores were used as the reference case in which ceCT (0.79) scored higher than non-contrast MRI

(0.71) and ultrasound (0.43). Results suggest that MRI is preferable to ultrasound by a greater proportion than ceCT is preferable to MRI. Although the geometric mean results were similar to consensus results, the differences between scores given CT and the other tests grew larger and smaller, between MRI and ultrasound (data not shown). Radiologists' priority scores reflect a stronger preference for ceCT, than non-radiologists (0.85 vs. 0.76, respectively) while non-radiologists scored MRI higher than radiologists (0.75 vs. 0.58, respectively). Across lower-tier criteria ceCT outperformed MRI in *time to diagnosis* while MRI scored slightly higher in decreasing potential risks to the patient. Ultrasound was superior in minimizing *incidental findings* and *patient specific exclusions* (Table 1).

Inconsistency Check and Sensitivity Analyses

For all but one criterion, *examination time* (0.13), the inconsistency ratio was below 0.1, indicating consistent voting behavior. Notably, lower inconsistency ratios in the geometric means group results indicate that movement toward consensus may have resulted in loss of consistency (data not shown). When stratified, the inconsistency ratio of this criterion dropped but was higher for non-radiologists (0.1) than radiologists (0.08). One-way variations in weights for higher and lower-tiered criteria revealed that ceCT and MRI reverse ranks (e.g., if the weight of the goal, decrease time to diagnosis, drops from 0.5 to 0.22 or less, MRI ranks first). There are no scenarios in which changes in criteria weights results in ultrasound ranking first.

Panel experience

We collected qualitative information during and after the mock panel activity. During the voting we observed that although the data that participants reviewed suggested MRI diagnostic accuracy is similar to ceCT (NND 1.06 and 1.07 compared to 1.10 and 1.14, respectively in Appendix A), participants scored MRI far lower in decreasing missed cases than ceCT (Table 1). During voting, a radiologist commented that the evidence suggested

very little difference in patient outcomes attributable to differences in sensitivities among tests. Based on this observation, we reviewed the pre-meeting survey results. Of note, only four of the nine survey participants answered that patient outcomes are likely to differ across diagnostic tests. Among those responders, the criterion, *length and quality of life* received zero votes and criterion *radiation-induced cancers* received one vote when ceCT was compared to ultrasound and to MRI (Appendix A).

Panel members unanimously agreed that weights assigned to criteria accurately represented their preferences and that the rankings of tests matched their expectations. Of note, the participants all agreed that the process helped them understand their values and priorities as well as how they may differ from others'. Participants referred to the experience of learning others' priorities as "eye-opening". Participants reported that this was an efficient way to treat the specifics of each modality separately. The participants liked the transparency and the iterative nature of the voting process.

DISCUSSION

Groups like the ACR AC panels face great decision complexity in developing practice guidelines for use in diagnostic radiology. MCDA methods have been developed to decompose complex decision problems. This study explored the use of AHP in judging appropriateness of three imaging tests for the diagnosis of classical presentation of suspected appendicitis**.**

The mock ACR AC panel members prioritized imaging tests differently than the ACR AC panel but these differences were minimal when converted to rankings. These differences may be linked to differences in selection of criteria between the ACR AC panel and the study participants. Participants selected BRC, *access to test*, which by policy, is excluded in ACR AC deliberations. MRI would have scored higher than ceCT had these considerations not borne weight in the analysis. These inconsistencies call into question what ought to be

considered when creating national guidelines and whether a panel of experts can reliably comply with those requirements. The appearance of the term "lesser clinical availability" in ACR AC summary of MRI evidence suggests that despite policy, these considerations played a role in ACR AC ratings.

Comparing ACR AC ratings to AHP results, study participants expressed lower confidence in ultrasound and higher confidence in MRI. Participants considered the probability that a ceCT is used, after an indeterminate ultrasound result to be a bigger burden to a patient than the marginal time and money costs of ceCT alone. Participants' judgments revealed an aversion to using multiple tests, in contradiction to the recent calls to use ultrasound as a primary test and reserve CT for equivocal ultrasound results.[22]

Although interpretation is limited by small sample sizes, stratified results suggest some differences between the rankings of radiologists and non-radiologists. These differences are attenuated by observed changes in individual votes during discussion. The observed reaction from both clinical specialties, with regard to the divergence in perspectives, suggests that guideline development relying predominantly on decision-makers with a primary training in radiology may be excluding considerations and preferences salient to non-radiologists using the guidelines. More research into the optimal composition for multi-disciplinary representation in guideline development is needed.

We noted in our group of study participants that expert opinion diverged from the body of evidence, possibly due to uncertainty. The criterion *missed cases* was supported by source data but not equally across imaging tests. Most participants reported that their experience with MRI in suspected appendicitis was limited. In turn, although the available evidence suggested little difference in missed cases between MRI and ceCT, participants scored ceCT much higher. Future research should consider implementing a process for showing explicitly when judgment overrides evidence and for what reason. If this reason is uncertainty in the

evidence base, more advanced analysis may be needed, such as probabilistic sensitivity analyses using group scores as distributions or fuzzy set theory.[23-25]

There were several limitations to our study. The first relates to our study population. There were notable differences between mock panel participants and the typical ACR AC panel. The mock panel members were fewer in number (n=9) than the range represented in ACR AC (i.e., 10 to 16) and represented radiologists and non-radiologists from one institution—University of Washington. All were well-acquainted and self-reported a positive rapport with their colleagues.

A related limitation is the exclusion of patient stakeholders. Although participants weighted criteria relevant to patients, patients did not serve as panel members. However, the hierarchical structure of AHP allows for separate groups of stakeholders to submit votes relevant to their preferences. Thus, it would be possible to involve patients in a post-hoc analysis.[26] In future reviews of this clinical scenario, patients could be asked to make trade-offs among *incidental findings*, *additional confirmatory testing* and *examination time*, *missed cases*, *contrast reaction potential*, and *ionizing radiation dose.*

Our comparisons of the ACR AC process to the study protocol are limited by a lack of experimental controls. We could not test the feasibility and experiences associated with following the ACR AC rating process in this study population. Conversely, the authors of the ACR AC ratings did not participate in this AHP analysis. Additionally, we could not run several AHP analyses to evaluated effects of changing structural components of the AHP model. The choice of AHP model structure has significant influence over how criteria are weighted, and changes in selection of criteria and structure are likely to influence results. The first structural component is the consensus threshold for selection of model criteria. Shifts in the consensus threshold result in changes to the composition of criteria, and subsequently a different model. The second component is deciding whether to have hierarchical or non-hierarchical weights. In order to reduce the number of comparisons,

hierarchies were created despite the criticism that hierarchical weights are 'steeper', that is, have higher weight ratios.[27] This highlights that MCDA approaches are vulnerable to manipulation, whether in biased selection of criteria, structuring of the model, or omission or inaccurate reporting of performance.[28] However, on the positive side, AHP leaves a traceable path for identification of possible manipulations and for transparent post-hoc modification of decision elements.

Our study is an early example of how AHP can be used to facilitate decision-making in diagnostic radiology. In screening, Hilgerink *et al.* used AHP to demonstrate the value of an emerging screening test for breast cancer, photoacoustic mammoscope.[29] Maruther *et al.* have shown successful use of AHP in comparing add-on therapies to metformin in diabetes medical decision-making.[30] The mock panel participants found AHP helpful in understanding each other's views and reaching consensus. However, differences between geometric mean and consensus priority scores suggest the group could not meet "in the middle" every time.

The mock panel reached similar conclusions to the ACR AC about the relative appropriateness of three imaging tests for diagnosis of suspected appendicitis. The results of this study suggest that it is both feasible and potentially very helpful to include quantitative benefit-risk assessment, specifically MCDA methodology, in the ACR AC guideline development process. By performing AHP, ACR AC panel members may address decision-making challenges at no or nominal expense in time and financial resources. As a result, priorities behind the appropriateness ratings become more transparent. Further development is needed to increase the transparency of the contribution of expert judgment to MCDA and in order to understand whether or when quantitative benefit-risk assessment is helpful across the spectrum of decision-complexity found in diagnostic radiology.

REFERENCES

1. Cascade PN. The American College of Radiology. ACR Appropriateness Criteria project. *Radiology.* Jan 2000;214 Suppl:3-46.

2. ACR Appropriateness Criteria. ACR Appropriateness Criteria® Organization and Composition of Expert Panels. 2015; http://www.acr.org/~/media/ACR/Documents/AppCriteria/ETDevDiagnostic.pdf. Accessed September 4, 2015.

3. American College of Radiology. About the ACR Appropriateness Criteria. 2015; http://www.acr.org/Quality-Safety/Appropriateness-Criteria/About-AC. Accessed December, 11, 2014.

4. Brook RH. *The RAND/UCLA Appropriateness Method.* Rockville, MD: Public Health Service, U.S. Department of Health and Human Services;1994.

5. Keeney RL, Raiffa H. *Decisions with Multiple Objectives: Performances and Value Trade-Offs*. New York: John Wiley and Sons; 1976.

6. Riabacke M, Danielson M, Ekenberg L. State-of-the-Art Prescriptive Criteria Weight Elicitation. 2012;2012.

7. Belton VSTJ. *Multiple criteria decision analysis : an integrated approach*. Boston: Kluwer Academic Publishers; 2002.

8. Hunink MGMGPPSJE. *Decision making in health and medicine : integrating evidence and values*. Cambridge: Cambridge University Press; 2001.

9. Linkov I, Moberg E. Multi-criteria decision analysis environmental applications and case studies. 2012; http://public.eblib.com/choice/publicfullrecord.aspx?p=952014.

10. Adunlin G, Diaby V, Xiao H. Application of multicriteria decision analysis in health care: a systematic review and bibliometric analysis. *Health Expect.* Oct 18 2014.

11. Dong Q, Saaty T. An analytic hierarchy process model of group consensus. *Journal of Systems Science and Systems Engineering.* 2014;23(3):362-374.

12. Smith MP, Katz DS, Lalani T, et al. ACR Appropriateness Criteria(R) Right Lower Quadrant Pain-Suspected Appendicitis. *Ultrasound Q.* Jun 2015;31(2):85-91.

13. Pare JR, Langlois BK, Scalera SA, et al. Revival of the use of ultrasound in screening for appendicitis in young adult men. *J Clin Ultrasound.* Jul 14 2015.

14. Krishnamoorthi R, Ramarajan N, Wang NE, et al. Effectiveness of a staged US and CT protocol for the diagnosis of pediatric appendicitis: reducing radiation exposure in the age of ALARA. *Radiology.* Apr 2011;259(1):231-239.

15. Karul M, Berliner C, Keller S, Tsui TY, Yamamura J. Imaging of appendicitis in adults. *Rofo.* Jun 2014;186(6):551-558.

16. Heverhagen JT, Pfestroff K, Heverhagen AE, Klose KJ, Kessler K, Sitter H. Diagnostic accuracy of magnetic resonance imaging: a prospective evaluation of patients with suspected appendicitis (diamond). *Journal of magnetic resonance imaging : JMRI.* Mar 2012;35(3):617-623.

17. American College of Radiology. Appropriateness Criteria. 2015; https://acsearch.acr.org/list. Accessed September 2, 2015.

18. ACR Appropriateness Criteria. ACR ACR Appropriateness Criteria® Evidence Table Development — Diagnostic Studies. 2013; http://www.acr.org/~/media/ACR/Documents/AppCriteria/ETDevDiagnostic.pdf. Accessed September 4, 2015.

19. Aczél J, Saaty TL. Procedures for synthesizing ratio judgements. *Journal of Mathematical Psychology.* 1983;27(1):93-102.

20. Hummel JM, Bridges JF, MJ IJ. Group decision making with the analytic hierarchy process in benefit-risk assessment: a tutorial. *Patient.* 2014;7(2):129-140.

21. Saaty TL. *The analytic hierarchy process: planning, priority setting, resource allocation*. New York: McGraw-Hill International Book Co.; 1980.

22. Poletti PA, Platon A, De Perrot T, et al. Acute appendicitis: prospective evaluation of a diagnostic algorithm integrating ultrasound and low-dose CT to reduce the need of standard CT. *European radiology.* Dec 2011;21(12):2558-2566.

23. Figueira J, Greco S, Ehrgott M. *Multiple criteria decision analysis : state of the art surveys*. New York: Springer; 2005.

24. Broekhuizen H, Groothuis-Oudshoorn CG, van Til JA, Hummel JM, MJ IJ. A Review and Classification of Approaches for Dealing with Uncertainty in Multi-Criteria Decision Analysis for Healthcare Decisions. *Pharmacoeconomics.* Jan 29 2015.

25. Millet I, Wedley WC. Modelling risk and uncertainty with the analytic hierarchy process. *Journal of Multi-Criteria Decision Analysis.* 2002;11(2):97-107.

26. Dolan JG. Involving patients in decisions regarding preventive health interventions using the analytic hierarchy process. *Health Expect.* Mar 2000;3(1):37-45.

27. Stillwell WG, Winterfeldt Dv, John RS. Comparing hierarchical and nonhierarchical weighting methods for eliciting multiattribute value models. *Manage. Sci.* 1987;33(4):442-450.

28. Bots PWGHJAM. Designing Multi-Criteria Decision Analysis Processes for Priority Setting in Health Policy. *JOURNAL OF MULTICRITERIA DECISION ANALYSIS.* 2000;9:56-75.

29. Hilgerink MP, Hummel MJ, Manohar S, Vaartjes SR, Ijzerman MJ. Assessment of the added value of the Twente Photoacoustic Mammoscope in breast cancer diagnosis. *Med Devices (Auckl).* 2011;4:107-115.

30. Maruthur NM, Joy SM, Dolan JG, Shihab HM, Singh S. Use of the analytic hierarchy process for medication decision-making in type 2 diabetes. *PLoS One.* 2015;10(5):e0126625.

TABLES

Table 1. Priority scores by goal and criterion

Goals		Criteria	Ultrasound	MRI	ceCT	Relative performance (US, MRI, ceCT)
Goal 1	Minimize time to diagnosis		0.1859	0.3493	0.4629	
	Goal 1a	Maximize provider utility	0.1364	0.2702	0.3659	
		Increase diagnostic accuracy consistency	0.0576	0.2304	0.2304	
		Maximize access to test	0.0349	0.0121	0.1006	
		Minimize patient-specific exclusions	0.0440	0.0277	0.0349	
	Goal 1b	Decrease burden to patient	0.0494	0.0791	0.0970	
		Minimize potential for additional confirmatory testing	0.0107	0.0679	0.0714	
		Minimize incidental findings management	0.0357	0.0083	0.0077	
		Minimize examination time	0.0030	0.0030	0.0179	
Goal 2	Minimize potential harms to patient		0.2454	0.3574	0.3231	
		Missed cases	0.0454	0.1651	0.3000	
		Minimize ionizing radiation dose	0.1000	0.0961	0.0116	
		Minimize contrast reaction potential	0.1000	0.0961	0.0116	

Relative performance, by goal, sub-goal, and criteria, is depicted in the far right column in the order: ultrasound, magnetic resonance imaging (MRI), contrast-enhanced computed tomography (ceCT). Dark grey bars distinguish the highest scoring technology.

FIGURES

Figure 1. Steps of the Multi-Criteria Decision Analysis (MCDA) and the American College of Radiology Appropriateness Criteria (ACR AC) processes

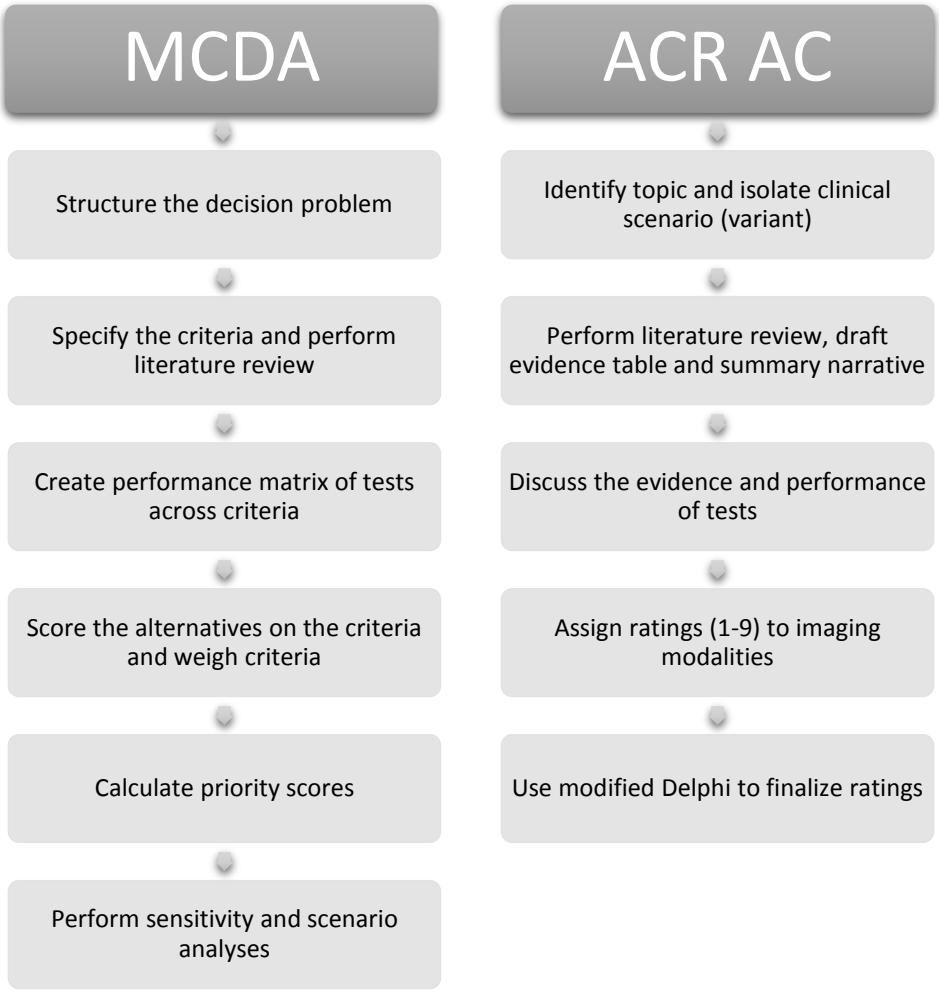
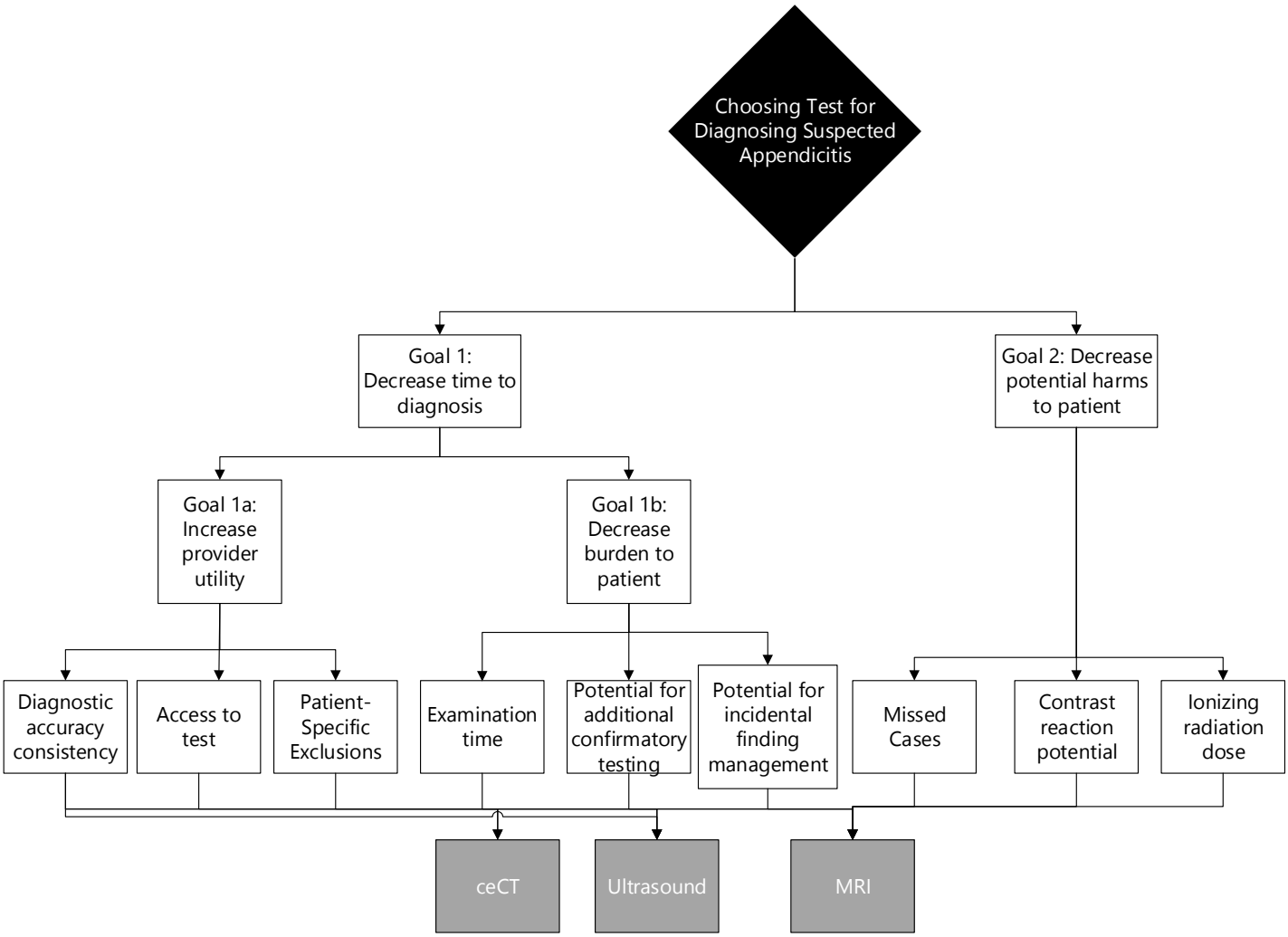


Figure 2. Analytic Hierarchy Model Hierarchy



Diamond black shape indicates the decision problem. White boxes denote goals (first level of hierarchy), sub-goals (second level of hierarchy) and criteria (lowest tier criteria). Square grey boxes denote imaging tests under evaluation.
ceCT: contrast-enhanced computed tomography
MRI: magnetic resonance imaging

Figure 3. Criteria weights of the Analytic Hierarchy Process (AHP) model

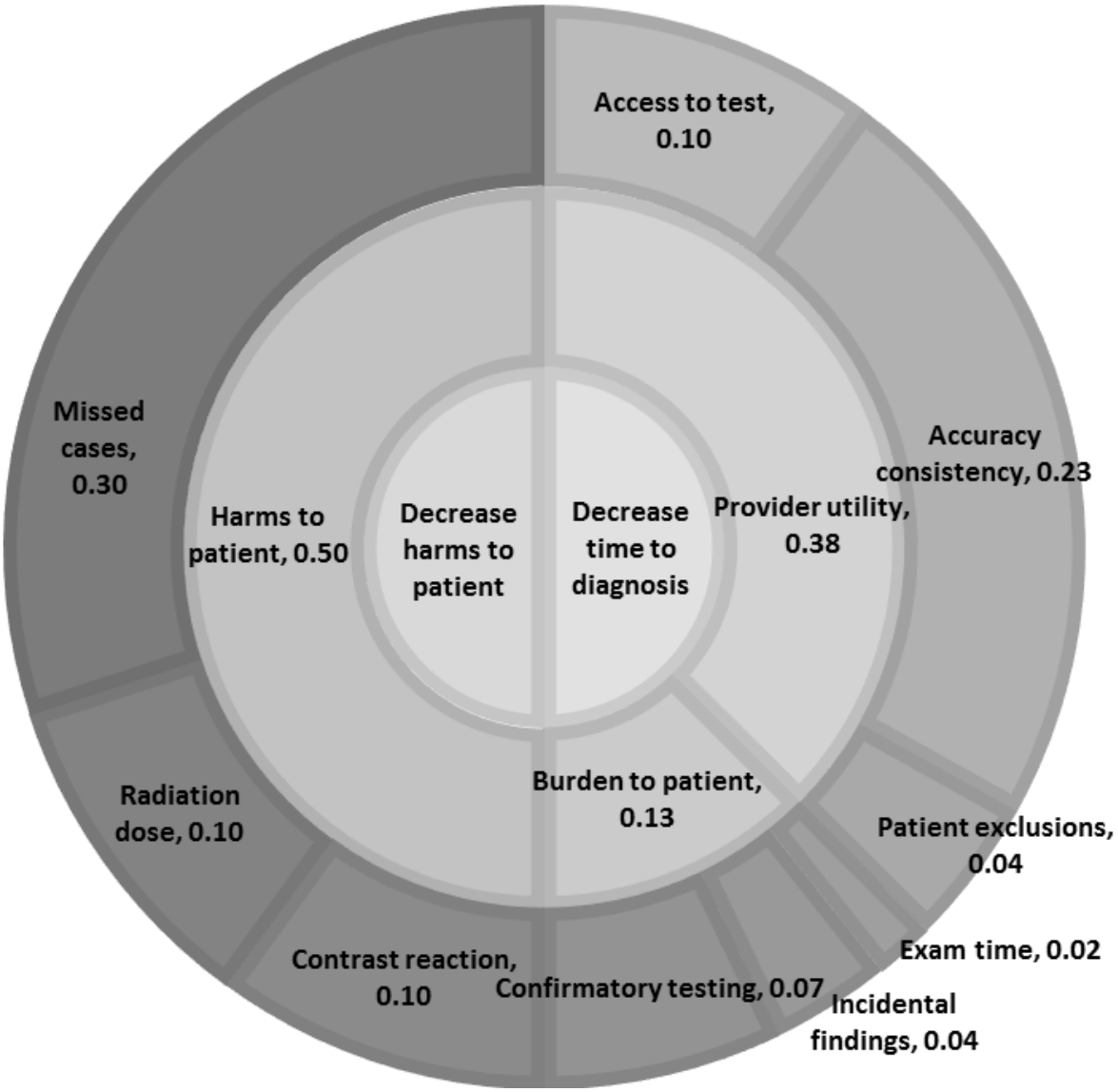
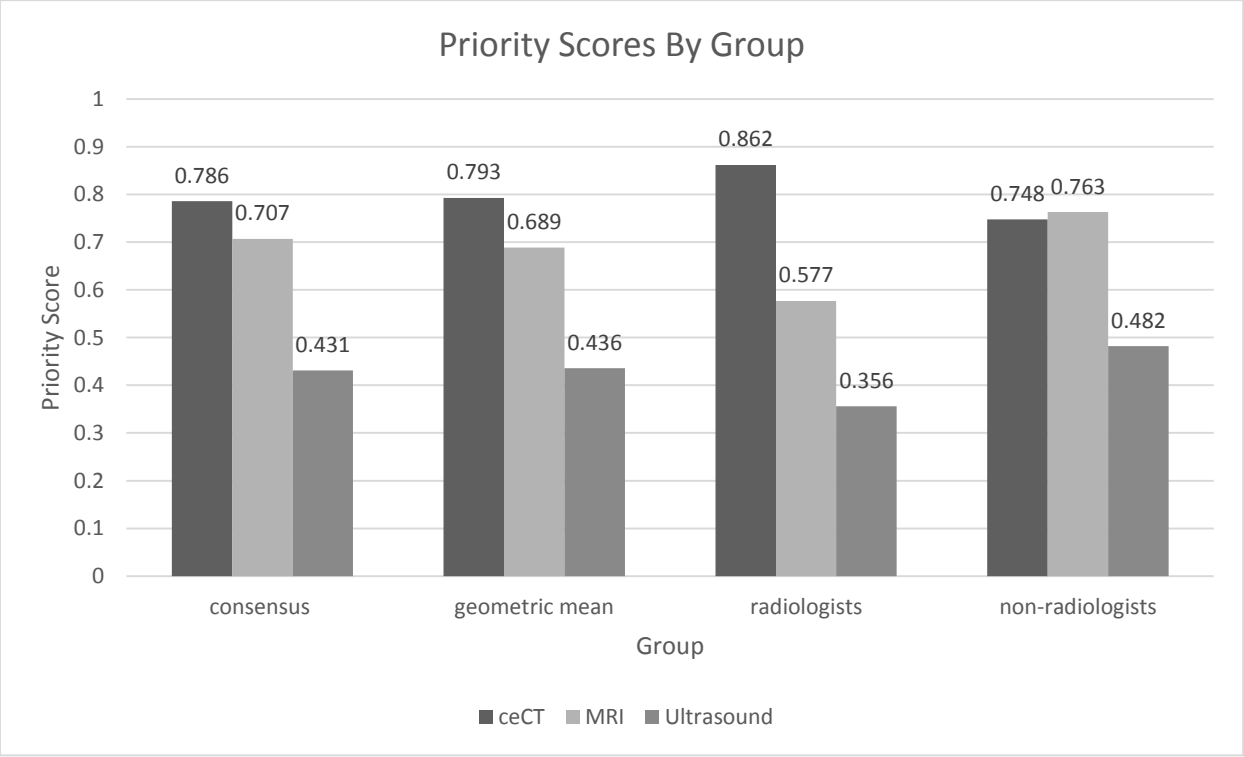


Figure 4. Priority Scores by Group



Priority scores of the three imaging tests in the reference case (consensus scores), based on the geometric mean of individual scores (geometric mean scores) and based on geometric mean scores stratified by clinical specialty (radiologist and non-radiologist priority scores).
ceCT: contrast-enhanced computed tomography
MRI: magnetic resonance imaging

APPENDIX A. TABLES

Table 1. Pre-meeting survey benefit-risk criteria selections

Criteria	Votes (Counts)			
	ceCT vs. MRI	ultrasound vs. MRI	ceCT vs. ultrasound	Max vote
Test Specific Features*	**9**	**9**	**9**	**9**
ionizing radiation dose	8	9	0	9
number of missed cases	8	1	6	8
contrast reaction potential	8	6	0	8
the number/extent of patient exclusions	4	8	7	8
examination time	4	8	7	8
diagnostic accuracy consistency	7	2	7	7
Patient management likely to differ? ǂ	**6**	**6**	**6**	**6**
access to test	1	4	6	6
potential for confirmatory testing	5	1	5	5
potential for incidental findings	5	2	5	5
time to diagnosis	2	4	4	4
time to discharge	4	3	3	4
provider utility	4	3	3	4
Patient outcomes likely to differ? ǂ	**4**	**3**	**4**	**4**
burden to patient	3	3	4	4

*>75% consensus needed, to be included as relevant criteria

ǂ >50% consensus needed, to be included as relevant criteria

Table 2. Performance matrix for missed cases criterion presented to participants

Year	Author (notes)	Pooled True Positives	Pooled False Negatives	Pooled True Negatives	Pooled False Positives			Percent Missed Cases			Percent False Diagnoses	Number Needed to Diagnose
	ceCT	**Disease +**		**Disease -**		**Pooled Sensitivity**	**95% CI**	**1-NPV**	**Pooled Specificity**	**95% CI**	**1-PPV**	
2008	van Randen (non-contrast studies removed)	167	8	160	160	0.97	(0.93, 1.00)	5%	0.91	(0.89, 0.93)	9%	**1.14**
2005	Weston (all studies; 2 non-contrast CT studies)					0.97	(0.95, 0.98)	3%	0.94	(0.92, 0.96)	6%	**1.10**
2004	Terasawa (3 non-contrast studies removed)	328	24	491	29	0.94	(0.92, 0.97)	5%	0.94	(0.91, 0.98)	8%	**1.14**
2001	*Bouillet (only IV-only study)*	*63*	*9*	*24*	*4*	*0.88*	*(0.78, 0.93)*	*27%*	*0.86*	*(0.69, 0.94)*	*6%*	***1.35***
2000	*Horton (non-IV CT), randomized design*	*37*	*1*	*11*	*0*	*0.97*	*(0.87, 1.00)*	*8%*	*1.00*	*(0.74, 1.00)*	*0%*	***1.03***
	ultrasound											
2008	van Randen	266	71	250	57	0.78	(0.67, 0.86)	22%	0.83	(0.76, 0.88)	18%	**1.64**
2005	Weston					0.87	(0.85, 0.89)	8%	0.93	(0.92, 0.94)	11%	**1.25**
2004	Terasawa	634	103	615	141	0.86	(0.83, 0.88)	14%	0.81	(0.78, 0.84)	18%	**1.49**
2000	*Horton, randomized design*	*23*	*2*	*3*	*1*	*0.92*	*(0.75, 0.98)*	*40%*	*0.75*	*(0.30, 0.95)*	*4%*	***1.49***
	MRI											
2010	Barger					0.97	(0.92, 0.99)		0.97	(0.94, 0.99)		**1.06**
2006	*Pedrosa (pregnant population, prevalence 0.09%)*					*1.00*	*(0.54, 1.00)*	*0%*	*0.94*	*(0.83, 0.98)*	*99%*	***1.07***

PPV: positive predictive value; NPV: negative predictive value; ceCT: Contrast-enhanced computed tomography; MRI: magnetic resonance imaging; CI: confidence interval

Table 3. Performance matrix of remaining criteria presented to participants

Criteria	ceCT	ultrasound	MRI	Source
Test Specific Features				
Diagnostic accuracy consistency				
Contrast reaction potential				
Ionizing radiation dose	Level 4 RRL	None	None	*Smith, et al. Ultrasound Q.* Jun 2015;31(2):85-91
Patient-specific exclusions				
Examination time				
Potential for additional confirmatory testing: number of nondiagnoses	Disease +: 8 Disease-: 14	Disease+: 14 Disease -: 9		Terasawa, et al. *Annals of internal medicine.* Oct 5 2004;141(7):537-546.
Potential for incidental finding management	23-78% of CT tests			Ozao-Choy, et al. *Am Surg.* Nov 2011;77(11):1502-1509
Access to test				
Time to diagnosis				
Time to discharge				
Provider utility				
Burden (time and money) to patient				

ceCT: contrast-enhanced computed tomography
ultrasound: ultrasound
MRI: magnetic resonance imaging
RRL: Relative Radiation Level (as defined by the ACR AC)
CT: any computed tomography imaging
Disease+: the number of nondiagnoses or inconclusive results in disease positive patients
Disease-: the number of nondiagnoses or inconclusive results in disease negative patients

APPENDIX B. SUPPLEMENTARY FILE

Survey presented to participants in preparation for the mock ACR AC MCDA activity

For suspected appendicitis, will patient outcomes likely differ if a CTc is ordered instead of an U/S? (e.g. patient comfort, future compliance and behavior, morbidity, survival and others)

◯ Yes
◯ No

Select the differences in patient outcomes between CTc and U/S. Check all that apply.

☐ Value of knowing true negative and positive results
☐ Disvalue of knowing false positive and negative results)
☐ Burden (e.g. out-of-pocket, travel and work absenteeism costs)
☐ Comfort (e.g. claustrophobia, physical discomfort/pain from test)
☐ Expected behavior changes (e.g. to future healthcare compliance, lifestyle)
☐ Potential for radiation-induced cancers
☐ Net incremental survival and quality of life attributable to test

---

CTc vs. U/S

Using your beliefs about the differences between computed tomography enhanced by at least IV-contrast (CTc) and graded compression ultrasound (U/S), answer the following questions.

For suspected appendicitis, CTc likely differs from U/S in: Check all that apply.

☐ Number of missed cases
☐ Number false diagnoses
☐ Diagnostic accuracy consistency within non-pregnant adults
☐ Level of inter-reader agreement
☐ Depth/breadth of visualization (e.g. anomaly size, shape, vascularization)
☐ Level of invasiveness/risk of adverse events
☐ Contrast reaction potential
☐ Ionizing radiation dose (e.g. relative radiation level)
☐ The number/extent of patient-specific exclusions (e.g. metal implants, BMI, age)
☐ Risk of failure/malfunction rate
☐ Patient preparation complexity (e.g. fasting, additional testing and procedures)
☐ Examination time
☐ Post-test observation time (e.g. processing of results, patient observation)
☐ Decision support (e.g. automated interpretation or function characterization)
☐ Portability
☐ Ease of use (e.g. dependence on operator skill)
☐ Reimbursement potential (e.g. Medicare fee for diagnostic test CPT code)

For suspected appendicitis, will management of a patient likely differ if a CTc is ordered instead of an U/S? (e.g. number of negative appendectomies, confirmatory testing, incidental findings, time to diagnosis and others)

◯ Yes
◯ No

Select the differences in patient management between CTc and U/S. Check all that apply.

☐ Therapeutic success (e.g. reductions in number of perforations)
☐ Potential for confirmatory testing
☐ Potential for incidental finding management
☐ Unnecessary treatment (e.g. negative appendectomy)
☐ Access to test
☐ Time to diagnosis
☐ Healthcare visits (e.g. test-related visits)
☐ Time to discharge if test negative
☐ Provider confidence in test (e.g. belief the test is useful for indication)
☐ Liability protection
☐ Financial incentive to provider
☐ Contribution of information to prognosis (e.g. staging of disease)

CTc: Contrast-enhanced computed tomography; U/S: ultrasound; MRI: magnetic resonance imaging

APPENDIX C. SUPPLEMENTARY FILE

RIGHT LOWER QUADRANT PAIN — SUSPECTED APPENDICITIS

Expert Panel on Gastrointestinal Imaging: Martin P. Smith, MD[1]; Douglas S. Katz, MD[2]; Max P. Rosen, MD, MPH[3]; Tasneem Lalani, MD[4]; Laura R. Carucci, MD[5]; Brooks D. Cash, MD[6]; David H. Kim, MD[7]; Robert J. Piorkowski, MD[8]; William C. Small, MD, PhD[9]; Stephanie E. Spottswood, MD[10]; Mark Tulchinsky, MD[11]; Vahid Yaghmai, MD, MS[12]; Judy Yee, MD.[13]

Summary of Literature Review

Introduction/Background

Relatively few comparative imaging studies evaluating right lower quadrant (RLQ) pain are available; most of the literature centers on the diagnosis of acute appendicitis (AA), the most common cause of acute RLQ pain requiring surgery [1]. For this reason, the focus of this narrative is on appendicitis and the accuracy of imaging procedures in diagnosing appendicitis, although consideration of other diseases is included.

In a few patients with AA, such as young men, imaging may not be necessary because the clinical presentation is sufficiently diagnostic to allow surgery [2,3]. Clinical prediction scores, such as the Alvarado score, have been used as a prediction rule for identifying patients with appendicitis; however, their accuracy is inferior to imaging and insufficient as a sole method for appendicitis evaluation [4]. In many published studies for appendicitis imaging, subjects with definitive clinical examination findings of appendicitis undergo operation without imaging. In the reported imaging studies, approximately 40% of imaged subjects on average had appendicitis and, in approximately 30% of subjects, another cause for RLQ pain was identified by imaging. Data on the overall effect of imaging on surgical treatment of appendicitis and patient outcome remain somewhat controversial, but growing evidence supports imaging use to reduce the negative appendectomy rate (NAR) [5-14].

Computed Tomography and Ultrasound

Computed tomography (CT) is the most accurate examination for evaluating patients without a clear clinical diagnosis of AA [15,16]. In a meta-analysis of 6 prospective studies through February 2006 of the accuracy of CT and ultrasound (US) in adolescents and adults, CT demonstrated superior sensitivity (91%; 95% confidence interval [CI], 84%–95%) and specificity (90%; 95% CI, 85%–94%) versus US (sensitivity, 78%; 95% CI, 67%–86%; specificity 83%, 95% CI, 76%–88%) [17]. The results of CT investigations were consistent across all studies and institutions, whereas US investigations demonstrated heterogeneity, suggesting greater dependence on operator skill [18]. The routine use of CT to evaluate for appendicitis has been shown to decrease overall costs by $447 to $1,412 per patient [13,19]. CT has been shown to decrease NAR from 16.7% to 8.6% in a meta-analysis of 20 studies with a broad range of 5,616 patients [20] and from 42.9% to 7.1% among 399 women aged 18–45 years at a single institution [21]. Accuracy of clinical diagnosis of the etiology of RLQ pain in women of childbearing age tends to be less accurate compared with adult men, thereby suggesting a lower threshold for imaging in this population [22]. In elderly patients with RLQ pain, the accuracy of clinical diagnosis also tends to be less accurate, and the increased risk of complications with AA in this population suggests a lower threshold for imaging with CT, as it has been shown to be highly accurate in depicting AA and its complications [23].

With the increased use of CT to evaluate for AA, concern has also increased about the effects of radiation exposure from CT, particularly since the majority of the population undergoing imaging for suspected AA is young or relatively young. A few studies have used algorithms with US as a first test to decrease the use of CT or have studied the use of CT with techniques that reduce the radiation dose while maintaining diagnostic accuracy. In 2 recent studies from Europe, diagnostic pathways used US as the primary modality after clinical evaluation by a surgeon. CT was reserved for cases where US was inconclusive [24] or negative [25]. These studies showed pathway sensitivity and specificity of 95.0% and 86.7% [24] with CT used in only 17.9% of cases, and 100% and

86% [25] with CT used in 39.7% of cases; all diagnostic errors in both studies were made in patients who underwent US only. Another European study of 183 patients first used an algorithm of US followed by low-dose (LD) CT when US was inconclusive, and then standard-dose (SD) CT when LDCT was inconclusive; 98.8% sensitivity and 96.9% specificity for diagnosing appendicitis were obtained with a 64% reduction in estimated dose compared to performing SDCT in all imaged patients [26]. A recent study comparing NAR between a SDCT (447 patients) and a LDCT technique (444 patients) in a routine university hospital emergency department (ED) showed no significant difference in NAR with the LD protocol using less than 25% of the estimated dose of the SDCT protocol [27]. In both studies, thin slices and multiplanar reformats were used to aid in diagnosis, which have also been shown in small studies to increase confidence in identifying the appendix [28-30].

When using CT, questions remain whether to use intravenous (IV) contrast, enteric contrast, both, or neither in the evaluation for AA. High accuracy has been reported for techniques using IV contrast as well as for those not using IV contrast (with or without enteric contrast), but few direct comparisons suggest higher accuracy when IV contrast is used [31]. A prospective study with 232 patients showed that non–contrast-enhanced CT (sensitivity, 90%; specificity, 86%) was inferior to rectal-only contrast (sensitivity, 93%; specificity, 95%) and IV and oral contrast (sensitivity, 100%; specificity, 89%) [32]. In lieu of individual patient contraindications to IV contrast, its use is recommended in evaluation of RLQ pain. However, if IV contrast is contraindicated, non–contrast-enhanced CT has been shown in 1 study of 300 patients to have a sensitivity of 96%, specificity of 99%, and accuracy of 97% [33] and in a meta-analysis of 7 studies with 1,060 patients to have a summary sensitivity of 92.7% (95% CI, 89.5%–95.0%) and specificity of 96.1% (95% CI, 94.2%–97.5%) [34].

The need for oral contrast when imaging suspected AA with CT, and particularly the need for rectal contrast, is less clear. In 1 prospective study, the use of rectal contrast has been shown to decrease ED stay by greater than 1 hour compared to oral contrast, without a significant difference in patient satisfaction or discomfort [35]. There is concern, however, that rectal contrast can be complicated by bowel perforation, with a cited number similar to barium enema of 0.04% [31]. One recent study showed similar sensitivity and specificity for detection of AA on 64-row multidetector CT (MDCT) with or without oral contrast performed with IV contrast [36]. Another recent study on 16-row MDCT showed no statistical difference either in sensitivity or specificity for detection of AA with or without oral contrast performed with IV contrast, and ED disposition was faster in the IV contrast only group [37]. In both of these studies, for the diagnosis of AA, sensitivity was 100%, and specificity was greater than 97% in the IV contrast only groups [36,37]. Another recent prospective study randomized patients to ingest or not ingest oral contrast; both groups then underwent IV unenhanced and enhanced standard dose MDCT with each study also using a simulated LD technique. The study determined that diagnostic correctness was more influenced by the reader than by the use of contrast medium or the LD technique [38]. With data from these and other studies, and the increased examination time, problems with patient tolerance, and potential increased radiation exposure from CT in patients with high-density enteric contrast, evidence is trending against the routine use of oral contrast, and particularly against the routine use of rectal contrast, for CT when IV contrast is used.

CT appears superior to US in identifying complications and in evaluating patients with periappendiceal abscess, especially when the abscesses become large [39]. CT results can be used to select therapeutic options other than immediate surgery, including antibiotic treatment with small abscesses and percutaneous drainage with well-defined or small, poorly defined abscesses [40]. Imaging-guided percutaneous drainage combined with antibiotics has been shown to be an effective initial treatment for AA complicated by perforation and abscess, followed by subsequent elective appendectomy or, in selected cases, conservative management [41]. High technical and clinical success rates have been shown with extraluminal appendicolith and large, poorly defined abscesses associated with repeat drainage and clinical failure in a recent study [42].

CT and US are effective in depicting alternative diagnoses for RLQ pain. In a large single-center study evaluating the diagnostic performance of MDCT for suspected AA, a cause of pain other than AA was established or suggested in a larger number of cases than AA (896 versus 675 in 2,871 patients) [43]. The range of diseases studied includes inflammatory bowel disease, infectious bowel disease, small-bowel obstruction, gynecological conditions, genitourinary conditions, and epiploic appendage, omental, and mesenteric inflammation.

Magnetic Resonance Imaging

At this time, few studies evaluate the value of magnetic resonance imaging (MRI) in the general population for AA. MRI is desirable due to its lack of ionizing radiation; however, its relative limitations include greater cost, longer acquisition time, and lesser clinical availability. A meta-analysis of 8 studies (5 retrospective) evaluating