How to Critically Appraise the Clinical Literature

Paul Cronin, MD, MS, James V. Rawson, MD, Marta E. Heilbrun, MD, MS, Janie M. Lee, MD, MSc, Aine M. Kelly, MD, MS, MA, Pina C. Sanelli, MD, MPH, Brian W. Bresnahan, PhD, Angelisa M. Paladin, MD

Recent efforts have been made to standardize the critical appraisal of clinical health care research. In this article, critical appraisal of diagnostic test accuracy studies, screening studies, therapeutic studies, systematic reviews and meta-analyses, cost-effectiveness studies, recommendations and/or guidelines, and medical education studies is discussed as are the available instruments to appraise the literature. By having standard appraisal instruments, these studies can be appraised more easily for completeness, bias, and applicability for implementation. Appraisal requires a different set of instruments, each designed for the individual type of research. We also hope that this article can be used in academic programs to educate the faculty and trainees of the available resources to improve critical appraisal of health research.

Key Words: Appraisal; cost-effectiveness assessments; diagnostic test accuracy; guidelines; medical education; recommendations; screening; systematic review and meta-analysis; therapy.

This article is the second in a series of two articles that review how to report and how to critically appraise research in health care. The first article reviewed the reporting of and the available guidelines on how to report health research. In this article, critical appraisal of screening studies and diagnostic test accuracy studies, therapeutic studies, systematic reviews and meta-analyses, cost-effectiveness studies, recommendations and/or guidelines, and medical education studies is discussed as are the available instruments to appraise literature.

Recent efforts have been made to standardize both the reporting and the appraisal of clinical health research including clinical guidelines. Many of the presentations at the joint Radiological Alliance for Health Service Research/Alliance of Clinical-Educators in Radiology session at the 2013 Association of University Radiologists annual meeting highlighted appraisal instruments available. By having standard formats for reporting research findings and guidelines, these studies can be appraised more easily for completeness, bias, and applicability for implementation. Appraisal requires a different set of instruments, each designed for the individual type of research. Quality Assessment of Diagnostic Accuracy Studies (QUADAS) was initially published in 2003 and has been used to evaluate the quality of diagnostic accuracy studies (1,2). Recent revisions produced QUADAS-2 (3). Similarly, the Assessment of Multiple Systemic Reviews (AMSTAR) was published in 2007 and uses 11 yes or no questions to assess the methodological quality of systematic reviews (4). The Appraisal of Guidelines for Research and Evaluation (AGREE) instrument evaluates the process of practice guideline development and the quality of reporting (5).

HOW TO APPRAISE LITERATURE

Screening Studies and Diagnostic Test Accuracy Studies

There is no specific Enhancing the QUAlity and Transparency Of health Research network (EQUATOR) or other recommendations for appraising screening studies. However, when reading a screening study, we suggest considering the following questions:

- How was the study designed?
- What are the results?
- Will these results help me care for my patients?

How Was the Study Designed?. In appraising screening study design, the following details should be identified: the screening strategies that were compared and the similarity of the groups at the start of the study. Frequently, the comparator for a screening test is "No screening". This study design best demonstrates the maximal benefit from screening. However, if a diagnostic test is the current screening standard, a study design comparing the

Acad Radiol 2014; 21:1117–1128

From the Division of Cardiothoracic Radiology, Department of Radiology, University of Michigan Hospitals, Ann Arbor, MI 48109 (P.C., A.M.K.); Departments of Radiology (J.V.R.) and Imaging (J.V.R.), Medical College of Georgia, Georgia Regents University, Augusta, Georgia; Department of Radiology, University of Utah School of Medicine, Salt Lake City, Utah (M.E.H.); Department of Radiology, University of Washington, Seattle, Washington (J.M.L., B.W.B., A.M.P.); and Department of Radiology, Weill Cornell Medical College/New York Presbyterian Hospital, New York, New York (P.C.S.). Received March 25, 2014; accepted May 3, 2014. **Address correspondence to:** P.C. e-mail: pcronin@med.umich.edu

http://dx.doi.org/10.1016/j.acra.2014.05.004

current and emerging screening tests would better demonstrate the incremental benefit, or comparative effectiveness, of the new technology. For example, digital breast tomosynthesis is an emerging screening test for which early reports indicate both improved sensitivity and specificity compared to standard digital mammography (6–9). This study design highlights the additional cancer detection rate achievable with the new technology relative to the current clinical standard.

It is also important to note whether the groups being compared were similar at the start of the study. The best study design approach to comparing a screening test in similar groups of patients is randomization of patients (10). Randomization allows for even distribution of known and unknown confounding factors and helps to isolate differences in outcome related to the screening test.

What Are the Results?. Before recommending a screening test, both its benefits and its potential harms should be identified. When considering the benefit of screening, the most important outcome is reduction of disease-specific mortality. However, it takes many years of follow-up to establish this benefit at a level of statistical significance. Many studies will instead report intermediate outcomes thought to predict reduced disease-specific mortality, such as the ability of screening tests to detect tumors at smaller sizes and cancer at earlier stages.

It is also important to consider whether known biases are accounted for (10,11). Two important biases are lead-time bias and length bias. *Lead-time bias* can occur when a new screening test detects disease earlier, without affecting disease-specific outcome. When survival is measured from the time of diagnosis, survival with screening can appear spuriously prolonged. Lead-time bias can be minimized when survival is measured from the time of enrollment, regardless of intervention arm. *Length bias* is the overestimation of survival among screened groups because of the disproportionate identification of slowly progressing disease.

Overdiagnosis or overtreatment is a potential harm related to screening. It occurs when screening detects asymptomatic disease that would not have become clinically apparent over an individual's lifetime or when screening results in treatment of disease that would not have shortened an individual's life expectancy. Overdiagnosis or overtreatment occurs more frequently when older populations with higher competing mortality risks undergo screening.

Will These Results Help Me Care for My Patients?. To assess whether a study's results might be applied in one's clinical practice, it is important to note the study's inclusion and exclusion criteria, whether the results are similar or different for specific subgroups and whether the technology of interest has changed since the study was published. These details will help determine how similar a study population is to one's own patient population and how the results may apply to individual patients with varying characteristics.

Although the objective of the STAndards for the Reporting of Diagnostic accuracy studies (STARD) initiative is to improve the accuracy and completeness of reporting of studies of diagnostic accuracy, appraising a study of diagnostic test accuracy with the STARD checklist also allows readers to assess the potential for bias in the study (internal validity) and to evaluate its generalizability (external validity) (12). The STARD checklist and flow diagram are discussed in detail in the previous article [How to Report a Research Study, pages 1088–1116 (13–28)].

There is a tool designed to assess the quality of primary diagnostic accuracy studies, the QUADAS tool. The QUADAS instrument was first developed in 2003 as an assessment tool to be used in the context of reviews of diagnostic test accuracy studies (Appendix Table 1) (1,2). This has been updated in 2010, and QUADAS-2 is the current version that is recommended for use in systematic reviews to evaluate the risk of bias and applicability of primary diagnostic accuracy studies (Appendix Table 2). QUADAS-2 differs from the original tool and is more like a guideline to be customized for each review (3). It consists of four key domains covering patient selection, index test, reference standard, and flow of patients through the study and timing of the index test(s) and reference standard (1,3). At present, no formal evaluation of the tool has been conducted.

Therapeutic Studies

Therapeutic trials in radiology including percutaneous and endovascular interventions have specific issues that introduce bias to results reporting. These include impossible or partial blinding, clustering, the experience of providers and centers, and both patient and provider willingness to undergo randomization (29). When appraising therapeutic literature, it is critical that the reader be aware of the intrinsic limitations in study design created when it is not possible to randomize patients, blind providers, and patients to study arms or hypotheses. Critical appraisal should assess if descriptions of the experimental treatment and comparator have been reported and in sufficient detail (30). In addition, there should also be appraisal of descriptions of how, why, and when treatment is modified to help demonstrate sufficient separation of study arms, as study arm crossover may limit the conclusions that can be drawn from interventional trials (31). Although rarely included, it is recommended that trials describe in detail the volume of experience of providers, as patient outcomes are directly related to experience (30,32). These may have an impact on study external validity or generalizability. Although the Extended Consolidated Standards of Reporting Trials (CONSORT) statement is designed to facilitate trial reporting, using the checklist when reading or reviewing trial literature is recommended (33). Following the checklist enables the reader to assess whether the authors and designers of the study accounted for potential sources of bias when designing a trial and interpreting the results. One example would be a selection bias that could be introduced based on the characteristics of the providers offering the treatment. For example, a high-volume center would be expected to recruit more patients. If the study intervention is done in