How to Write a Critically Appraised Topic (CAT)

Gelareh Sadigh, MD, Robert Parker, ScD, Aine Marie Kelly, MD, MS, Paul Cronin, MD, MS

Medical knowledge and the volume of scientific articles published have expanded rapidly over the past 50 years. Evidence-based practice (EBP) has developed to help health practitioners get more benefit from the increasing volume of information to solve complex health problems. A format for sharing information in EBP is the critically appraised topic (CAT). A CAT is a standardized summary of research evidence organized around a clinical question, aimed at providing both a critique of the research and a statement of the clinical relevance of results. In this review, we explain the five steps involved in writing a CAT for a clinical purpose ("Ask," "Search," "Appraise," "Apply," and "Evaluate") and introduce some of the useful electronic resources available to help in creating CATs.

Key Words: Evidence-based medicine; evidence-based radiology; critically appraised topic; levels of evidence; literature search; systematic review.

Medical knowledge has expanded rapidly over the past 50 years. Many subcategories of disease, diagnostic testing, and treatment strategies are now known. Paralleling this improvement in medicine, the volume of scientific articles published has exploded and is doubling every 10 years (1). Therefore, evidence-based practice (EBP) and publications in this area have developed to help health practitioners keep up to date with the increasing volume of information to solve complex health problems (1).

One of the main formats for sharing information in EBP is the critically appraised topic (CAT). A CAT is a standardized summary of research evidence organized around a clinical question, aimed at providing both a critique of the research and a statement of the clinical relevance of results (2). In other words, CATs are not just abstracts of existing evidence. They critique the internal validity, external validity (generalizability), and statistical rigor (or methodology) of the best research evidence to date and summarize the results into a few pages (2,3). In contrast to systematic reviews, which are written by content and methodology experts, CATs may be more easily written by clinicians and practitioners (3). Critically appraised topics provide easy access to the scientific literature for clinicians who are either too busy to pursue the answer to a clinical problem among the mixed results from a search engine or do not have the specialized skill to critically appraise the literature and reach an appropriate conclusion (2).

Acad Radiol 2012; 19:872–888

From the Division of Cardiothoracic Radiology, Department of Radiology, University of Michigan Health System, University of Michigan B1 132G Taubman Center/5302, 1500 East Medical Center, Ann Arbor, MI 48109-5302 (G.S., A.M.K., P.C.); Department of Biostatistics, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109 (R.P.). Received August 23, 2011; accepted February 3, 2012. **Address correspondence to:** P.C. e-mail: pcronin@med.umich.edu

doi:10.1016/j.acra.2012.02.005

The main reason to produce a CAT is to answer an explicit clinical question arising from a specific patient encounter, and is the essence of EBP in that a health professional generates a clinical question from a real clinical situation, followed by finding and appraising the evidence, and finally applying it in clinical practice (1).

In this review, we start by explaining the steps involved in writing a CAT for a clinical purpose and introduce some of the available electronic CAT makers.

HOW TO WRITE A CAT?

Writing a CAT involves five steps along the five steps of evidence-based practice which can be summarized as "Ask," "Search," "Appraise," "Apply," and "Evaluate" (4). These steps are:

1. Asking a focused and answerable question that translates uncertainty to an answerable question
2. Searching for the best available evidence
3. Critically appraising the evidence for validity and clinical relevance
4. Applying the results to clinical practice
5. Evaluation of performance

Step 1: Ask an Answerable Question

The first step in writing a CAT is to formulate a well-built question regarding the clinical problem or knowledge gaps (decisions regarding patient's diagnostic workup, treatment, or intervention). To benefit patients and clinicians, such questions need to be both directly relevant to the patients' problems and phrased in ways that direct the search to relevant and precise answers. This involves taking a clinical question and changing its format so that the literature search is based

on this question (5,6). The question needs to be important to the patients' well-being, the clinicians' knowledge needs, of interest to the patient, clinician, or the learners, likely to recur in clinical practice, and answerable in the time available (7).

The majority of questions formulated to start a CAT are foreground questions, consisting of four components: 1) patient's problem of interest; 2) the main intervention (eg, a diagnostic test, or treatment) that is going to be compared with the existing reference standard; 3) the comparison intervention (diagnostic test or treatment) that is already identified; and 4) outcome of interest. These components can be abbreviated to PICO (patient, intervention, comparator, and outcome) (5). The finished question can be expressed in a single, clear and focused sentence (eg, "In patients with … how does … compare with … for the outcome(s) of …"). Sometimes there is no comparator intervention, and the question becomes PIO (8), or sometimes there is more than one comparator intervention or outcome. Some examples of foreground questions in radiology that generated a CAT and their PICO format are shown in Table 1 (6,9,10).

In diagnostic radiology, CAT questions may relate to the superiority of one imaging method over another in resolving clinical dilemmas and/or the power of imaging signs to reliably confirm or exclude a suspected disease processes. In interventional radiology, CAT questions are related to the short-, medium-, and long-term benefit/harm of new interventional techniques compared with older interventional methods or more invasive procedures.

Step 2: Search for the Best Current Evidence

The second step in writing a CAT is to perform a thorough search of the literature. To conduct a good search, one has to be familiar with the types and sources of information available, the levels of evidence and where to look for a particular type of evidence, and how to select articles with a high level of evidence. It is important when searching for evidence that search terms are referred back to the original PICO question (11). Examples of radiology CAT search strategies are shown in Table 2 (10,12–14). The process of searching for and finding the best current evidence therefore follows three key steps:

1. Identify terms to fit the PICO question
2. Search for secondary sources
3. Search for primary sources

Primary study designs. The goal of a primary research study, whatever design used, is to provide valid and generalizable data. Validity is internal to the study: the results are true for the population studied, and are not the result of bias or confounding. Generalizability (or applicability) is the ability to apply the results of the study to a broader population, hopefully including the population group of interest. Validity is a precondition for generalizability: if there are significant questions about whether the study results are valid, there is no information which can be applied to other populations.

The major factor affecting the validity of a study is bias. Bias means that the results of the study reflect other factors, in addition to and distinct from those that are formally being studied. As a simple example of a bias, if an investigator enrolled patients into a study and deliberately assigned those with a worse prognosis to the experimental arm to ensure that he would not be overestimating the potential benefit, the study would be biased. This assignment would minimize the observed benefit of the experimental treatment compared to the control group, biasing the results.

Compared to other research designs, the potential for bias is minimized in the randomized double-blind clinical (or controlled) trial, so it is the design most likely to provide valid data. As such, this design is considered as providing the best evidence on a question. This is because of two features mentioned in the name: randomization and double-blinding. Randomization is a process by which participants in the study are allocated after enrollment to either the intervention or the control group in a random manner. Implicit in this description is that no one knows which treatment the participant will receive until after the participant is enrolled. This eliminates any potential for the investigator or the patient to enroll into the study to receive a specific treatment, although most participants are likely to enroll in the hope of receiving the experimental treatment. This helps ensure that the intervention and control groups are similar in terms of both known and unknown prognostic factors. Double-blinding (sometimes called double-masking) is when both the participant and the outcome assessor do not know which treatment the participant is receiving. This reduces potential bias (generally toward an improvement) because of participant factors involved in knowingly receiving an experimental treatment, and potential bias if the assessor were determining outcome for a known treatment. Randomization also helps reduce the possibility of confounding, which is when an apparent treatment effect (or the lack of a treatment effect) is caused by another variable. Confounding requires that the variable be associated with both the treatment and with the outcome, but that it not be part of the mechanism of action of the treatment on the outcome. Randomization ensures that any variable related to outcome should on average be similar in the treatment groups. In addition, the temporal order is clear—the intervention in the randomized clinical trial precedes the outcome. Many textbooks have been written on clinical trials (15–17).

Other study designs, which collectively are called observational studies, have more potential for bias and thus run a greater risk of having validity problems than the randomized double-blind clinical trial. Even a randomized single-blind clinical trial (where the participant or the assessor but not both are blinded to treatment) is more prone to bias. Prospective cohort studies, where data are collected prospectively but participants are not randomized to exposure (or intervention) are generally considered the next best design in terms of validity, but suffer as do all the other designs mentioned in the following section, from potential confounding because

TABLE 1. Examples of CAT Questions in PICO Format

CAT Study Question	PICO Format Question	Patient or Problem	Intervention	Comparison Intervention	Outcome
Whether low-dose CTC would perform as well as optical colonoscopy for screening patients for CRC? (12)	Is low-dose CTC equivalent to optical colonoscopy in identifying >5-mm colonic polyps?	Screening population of patients with polyps	CTC	Optical colonoscopy	Accuracy in diagnosis of >5-mm colonic polyps/ colorectal cancer
Whether ultrasound performs better than MRI in the diagnosis of rotator cuff tears? (10)	In patients with rotator cuff tears, how does US compare to MRI for diagnosis?	Patients with rotator cuff tear	US	MRI	Accuracy in diagnosis of rotator cuff tears
Whether coronary CTA performs comparably to invasive coronary angiography for identifying potentially or probably hemodynamically significant native coronary artery disease, defined as luminal diameter stenosis of at least 50%? (13)	In patients with known or suspected coronary artery disease, how does coronary CTA compare with invasive coronary angiography for identifying ≥50% luminal diameter coronary artery stenosis?	Patients with known or suspected coronary artery disease	Coronary CTA	Catheter coronary angiography	Accuracy in diagnosis of ≥50% luminal diameter coronary artery stenosis
How does CTA perform in comparison with MRA in the detection of symptomatic carotid stenosis? (14)	In patients with symptomatic carotid stenosis, how does CTA compare to MRA for diagnosis?	Patients with symptomatic carotid artery stenosis	Carotid CTA	Carotid MR angiography	Accuracy in diagnosis of carotid artery stenosis
What is the current role of RFA and how does RFA compare to surgical resection for treatment of colorectal liver metastases? (9)	In patients with colorectal liver metastases how does percutaneous RFA compare with surgical resection or other ablative techniques?	Patients with colorectal liver metastasis	Percutaneous RFA	Surgical resection	Annual recurrence or mortality rate

CAT, critically appraised topic; CRC, colorectal carcinoma; CTA, computed tomography angiography; CTC, computed tomography colonography; MRA, magnetic resonance angiography; MRI, magnetic resonance imaging; PICO, patient or problem, intervention, comparison intervention, outcome; RFA, radiofrequency ablation; US, ultrasound.

TABLE 2. Examples of Search Strategies for the PICO of Questions

PICO Format Question	Patient or Problem		Intervention		Comparison Intervention		Outcome
Is low-dose CT colonography equivalent to optical colonoscopy in identifying clinically meaningful colonic polyps? (12)	Polyp *or* colorectal neoplasm *or* colonic neoplasm	**and**	CT colonography *or* colonography, computed tomographic **and** low-dose	**and**	Colonoscopy	**and**	Diagnosis *or* sensitivity and specificity
In patients with rotator cuff tears, how does US compare to MRI for diagnosis? (10)	Rotator cuff *or* shoulder *or* tendon *or* injury *or* pathology	**and**	Ultrasonography *or* US	**and**	Magnetic resonance imaging *or* MRI *or* contrast-enhanced MRI	**and**	Diagnosis *or* sensitivity and specificity
In patients with known or suspected coronary artery disease, how does coronary CT angiography compare to invasive coronary angiography for identifying ≥50% luminal diameter coronary artery stenosis? (13)	Coronary artery disease	**and**	Tomography, x-ray computed *or* tomography, x-ray computed/methods	**and**	Angiography *or* coronary angiography *or* coronary angiography methods	**and**	Coronary stenosis *or* diagnosis *or* sensitivity and specificity
In patients with symptomatic carotid stenosis, how does CTA compare with MRA for diagnosis? (14)	Carotid artery stenosis	**and**	Tomography, x-ray computed *or* tomography, x-ray computed/methods	**and**	MRI/MRA *or* MRI *or* contrast-enhanced MRI	**and**	Diagnosis *or* sensitivity and specificity
In patients with colorectal liver metastases how does percutaneous radiofrequency ablation compare with surgical resection or other ablative techniques? (9)	"Liver neoplasm" *or* "liver neoplasm/secondary"	**and**	Catheter ablation	**and**	Liver neoplasm/surgery	**and**	Efficacy *or* recurrence *or* mortality

CT, computed tomography; MRA, magnetic resonance angiography; MRI, magnetic resonance imaging; US, ultrasound.

the group exposed and the group unexposed may be different on other factors as well as the exposure being studied, leading to potential confounding (15–18). Retrospective cohort studies (in which some data are collected either from records from the past or from interviews about the past with participants) suffers from potential data and recall issues as well. In addition, the temporal order is less clear—the outcome may already have existed at the time of the exposure, especially in the retrospective cohort study (15–18). Case-control studies that compare current patients with a disease to a similar group without the disease potentially suffer from differential recall and other participant biases. In addition, the temporal order of events is even less clear cut than for a retrospective cohort study. In all the observational designs—and even in the single-blind randomized clinical trial—if the assessors are not blinded to the group then there are potential assessor biases as well. In addition, finding an appropriate control (nondiseased) population can be quite challenging for the case-control study, so multiple different types of control groups may be included to provide stronger evidence if results are consistent across them (15–20). Finally case series and other descriptive studies are just that—there are no comparative data involved, so are also subject to temporal bias when interpreted against historical data, where differences may be occurring only because treatment is generally better now than in the past.

Generalizability (or applicability) is separate from validity, although if there are concerns about the validity of the results (meaning that there are concerns about whether the results are correct), it is not clear why generalizability would be of interest. Generalizability means that the results apply—or at least are likely to apply—to a broad range of patients in addition to those in the study population. Generalizability is always a subjective question and can never be proven. However, the results of a study are likely to apply to a broad range of patients if the study itself involved a broad range of patients (eg, both genders, broad age range, relatively open enrollment), whereas generalizability is likely to be more limited if the population studied was very restricted (eg, only men ages 40–45 with stage IV disease and no comorbidities).

Levels of evidence. The Oxford Center for Evidence Based Medicine has graded study designs in a hierarchy of evidence (Table 3). With the help of this table, any retrieved article pertaining to therapy/prevention, etiology/harm, prognosis, diagnosis, differential diagnosis/symptom prevalence study, and economic and decision analysis can be rapidly assigned to a level of evidence (21).

The evidence pyramid. The available literature can be ranked using the "evidence pyramid" described by Haynes et al (22). The pyramid consists of four levels (4S): original studies, syntheses (systematic review) of evidence, synopses, and information systems, where original studies are at the base of the pyramid and information systems are at the top (Fig 1). The "4S" model (Fig 1a) for the organization of evidence-based information services begins with original studies at the foundation; syntheses (ie, systematic reviews, such as Cochrane Reviews) at the next level up; then synopses (very brief descriptions of original articles and reviews (including systematic reviews) such as those that appear in the evidence-based journals); and systems (such as computerized decision support systems that link individual patient characteristics to pertinent evidence) at the top (22,23).

The augmented "5S" model (Fig 1b) adds an additional layer to the model—namely, clinical topic summaries of evidence about all pertinent management options for a health condition, such as those included in Clinical Evidence, National Guidelines Clearinghouse, and the American College of Physicians' Information and Education Resource, which reside between synopses (succinct descriptions of an individual study or a systematic review) and systems (decision support services that match information from individual patients with the best evidence from research that applies) (24). In the augmented "6S" model (Fig 1c), there are synopses of studies in the second layer from the bottom and synopses of syntheses in the fourth layer from the bottom, which more accurately depicts their rigor (25). The evidence identified at higher levels of the pyramid is more compelling or comprehensive than that at lower levels. Therefore, when searching the literature, one should start the search looking for the highest level of evidence available and if an answer to a particular question is found at a higher level of evidence, then there is no need to search for evidence at lower levels (6,26). Most studies in the radiology literature still reside at the bottom or primary level (original studies). There are no resources specific to radiology at levels 5 and 6, and we will refer to the "4S" model.

Sources of evidence for primary literature. Primary literature is the first level of the evidence pyramid (studies) and consists of original articles (eg, randomized controlled trials, cohort, case-control, cross-sectional studies) with variable quality published in journals. These studies can found by searching medical search engines including PubMed (27), MEDLINE, EMBASE (28), ISI Web of Knowledge (29), MD Consult (30), and Google Scholar (6,8,31).

For the ordinary practitioner, PubMed (27) is perhaps the most-used electronic database for searching the primary literature. It is part of the Entrez series of databases. More than 20 million citations for biomedical literature are indexed in PubMed using medical subject heading (MeSH) terms, which provides a consistent way to retrieve information that may use different terminology for the same concepts (6,32). This is a very comprehensive way of identifying relevant articles. But if at the end of a search too many articles have been retrieved, limits can be applied to reduce the returns to a reasonable number using the *Limits* toolbar at the top of the PubMed home page screen. The search can be limited to items with abstracts, human studies, recent articles published (eg, articles within the past 2 years), English language, and so on. Adding a language limit (eg, English) can be done but should be reserved for the end stage of searching because it may cause loss of valuable articles (6).

TABLE 3. Designation of Levels of Evidence According to Type of Research

Level	Etiology	Diagnosis	Intervention	Prognosis
Ia	Systematic review with homogeneity* of Level Ib studies	Systematic review with homogeneity* of Level Ib studies	Systematic review with homogeneity* of Level Ib studies	Systematic review with homogeneity* of Level Ib studies
Ib	Randomized controlled trial	Validating† cohort study with good‡ reference standard	Randomized controlled trial	Cohort study with good (>80%) follow-up
Ic		Absolute SP-ins and SN-outs§		
IIa	Systematic review with homogeneity* of Level IIb studies	Systematic review with homogeneity* of Level IIb studies	Systematic review with homogeneity* of Level IIb studies	Systematic review with homogeneity* of Level IIb studies
IIb	Cohort study (prospective)	Exploratory† cohort study with good‡ reference standard	Cohort study (prospective)	Cohort study (retrospective)
IIc	Outcomes research	Outcomes research	Outcomes research	Outcomes research
IIIa	Systematic review with homogeneity*x of nonconsecutive cohort or case control studies	Systematic review with homogeneity* of Level IIIb studies	Systematic review with homogeneity* of case control studies	Systematic review with homogeneity* of case control studies
IIIb	Nonconsecutive cohort study or case control study	Nonconsecutive studies; or without consistently applied reference standard	Case control study	
IV	Poor-quality‖ cohort study or poor quality case control study or case series	Case series or poor or nonindependent reference standard	Poor-quality‖ cohort study or poor quality case control study or case series	Poor-quality‖ cohort study or poor quality case control study or case series
V	"Expert" opinion without critical appraisal	"Expert" opinion without critical appraisal	"Expert" opinion without critical appraisal	"Expert" opinion without critical appraisal

*A systematic review that is free of worrisome variations (heterogeneity) in the directions and degrees of results between individual studies. Not all systematic reviews with statistically significant heterogeneity need be worrisome, and not all worrisome heterogeneity need be statistically significant.

†Validating studies test the quality of a specific diagnostic test, based on prior evidence. Exploratory study collects information and trawls the data to find which factors are "significant."

‡Good reference standards are independent of the test, and applied blindly or objectively to applied to all patients. Poor reference standards are haphazardly applied, but still independent of the test.

§An "Absolute SpPin" is a diagnostic finding whose specificity is so high that a positive result rules-in the diagnosis. An "Absolute SnNout" is a diagnostic finding whose Sensitivity is so high that a negative result rules-out the diagnosis.

‖Cohort study that failed to clearly define comparison groups and/or failed to measure exposures and outcomes in the same (preferably blinded), objective way in both exposed and nonexposed individuals and/or failed to identify or appropriately control known confounders and/or failed to carry out a sufficiently long and complete follow-up of patients. Case-control study that failed to clearly define comparison groups and/or failed to measure exposures and outcomes in the same (preferably blinded), objective way in both cases and controls and/or failed to identify or appropriately control known confounders.

Sources of evidence for secondary literature. Secondary evidence comprises the upper three levels of the pyramid (syntheses, synopses, and information systems). Syntheses (ie, evidence-based reviews) consist of publications in which other authors have searched the literature on the topic in question and have appraised the retrieved articles. Systematic reviews with meta-analysis, CATs, and reviews that use EBP methodology are found on this level (6). These kind of articles can be found in databases and guidelines accessible through search engines and gateways such as PubMed Clinical Queries (8,27) as well as specific electronic databases such as the Cochrane Library (33), the Turning Research into Practice (TRIP) database (34), the National Institute of Clinical Excellence (NICE) (35), SUMSearch (36), Cumulative Index to Nursing and Allied Health Literature (CINAHL) (37), the National Guidelines Clearinghouse (NGC) (38), the National Library for Health (39), and the Scottish Intercollegiate Guidelines Network (SIGN) (6,32,40). There are no specific radiology resources available at this level (32).

Synopses consist of the current best evidence combined with clinical expertise from an expert in the area (6). A perfect synopsis would provide exactly enough information to support a clinical action, without needing to read all of the systematic reviews and articles behind it (22). They can be found in EBM sites such as the American College of Physicians (ACP) Journal Club (41) and Evidence based Medicine Online (42).

Finally, information systems are evidence-based clinical information systems that integrate and summarize all relevant

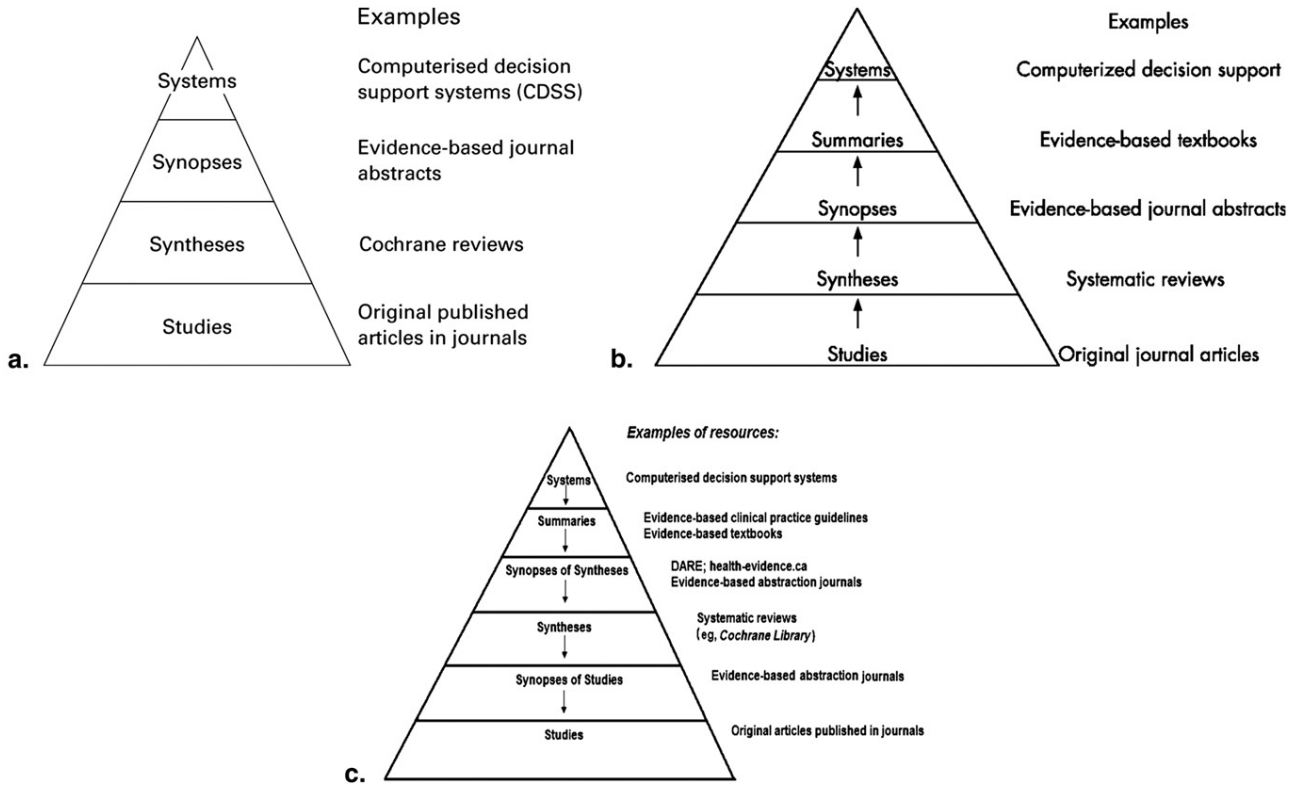

Figure 1. Evidence pyramid with levels of organization of evidence from research. **(a)** The "4S" levels (22,23). **(b)** The "5S" levels (24). **(c)** The "6S" levels of organization (25).

and important research evidence about a clinical problem and are regularly updated. Information systems are capable of linking a specific patient's problem through electronic medical records to the relevant information and research evidence on the subject. Readily available systems do not provide this function and cannot integrate with electronic medical records. The currently available systems include electronic textbooks such as Clinical Evidence (43) and Up to Date (6,22,44).

Step 3: Appraise the Literature

In this step, the retrieved articles are evaluated for their internal validity and external generalizability as well as their level of evidence. This process of evaluating searched articles is called "appraisal." If the search identifies secondary literature this should be appraised before the primary literature as secondary literature are higher level evidence. However, the majority of studies in the radiology literature still reside at the lowest or primary level (original studies) so we will review the appraisal of primary literature first.

Critically appraising primary literature. The first step in critical appraising primary literature is to evaluate how close the PICO of the study is compared to the PICO of the clinical question and if the study can assist with the clinical decision. Then the Methods section is reviewed to assess the internal validity of the study. This includes assessing if the intervention of interest is appropriately compared with the standard intervention and if efforts are made to eliminate or reduce confounding factors or potential sources of bias. Finally the generalizability of the evidence needs to be assessed and, most importantly, whether the population studied is relevant to your PICO. Review of the result section should be delayed until after you have decided that the article is methodologically sound, to avoid potentially being biased by the results, particularly when there are dramatic (small) *P* values. The review of the results section should focus both on whether the difference observed is clinically important and separately whether the results are statistically significant—the two do not necessarily go together.

A group of scientists and editors has developed the STARD (STAndards for the Reporting of Diagnostic accuracy studies) statement to improve the quality of reporting of studies of diagnostic accuracy. The statement consists of a checklist of 25 items and flow diagram that authors can use to ensure that all relevant information is present. Although this was designed to improve reporting of diagnostic accuracy studies, it is also a useful instrument for assessing the diagnostic literature (45). See the STARD checklist in Appendix Table 1.

Articles following the STARD standards allow for a thorough review for potential bias in the results. However, following the reporting standard is not sufficient to ensure the validity of a study. Because diagnostic accuracy studies

have several unique features in terms of design that differ from standard interventional studies (46), the quality assessment of (primary) studies of diagnostic accuracy included in systematic reviews (QUADAS) instrument was developed. See the QUADAS checklist in Appendix Table 2.

To summarize some of the main points of the QUADAS checklist, and add others of concern, the first validity concerns whether there is potential bias in interpreting test results by readers. Are test results read in a blinded fashion, or is information available about the results of one test when assessing the results of the second test? For example, selecting only patients positive on the gold standard test for an experimental test would tend to bias the reading of the experimental test. Were both tests interpreted with the same clinical data? Are the technical methods for each test sufficiently well-defined that they can be reproduced? As an additional point, are the methods in some sense "optimal," because suboptimal methods would tend to reduce the utility of a test? In addition, we would add are the readers appropriately experienced and familiar with both techniques to interpret results? If different readers are used for different techniques, this alone can bias the comparison because of differences in how each assesses results as positive or negative. If multiple readers have been used, are results similar for each reader in terms of the basic test characteristics (sensitivity and sensitivity)? Finally, how were uninterpretable/intermediate results handled for each test?

For interventional studies focusing on outcome, the first key issue is whether participants were randomized to intervention after enrollment in the study? If so, then on average potential patient characteristics that could confound (bias) the assessment are eliminated. If not, then patient characteristics—both known and unknown—could explain differences in outcome. Is there evidence that at least the known characteristics prognostic for outcome are similar between groups? Second, is the study double-blind, with both the participant and the assessor blinded to knowledge of the intervention? If not, there are potential biases due to either (or both) the patients' beliefs and the assessors' beliefs. This is especially problematic for subjective endpoints, but even "objective" endpoints (such as laboratory values) can be biased by patient or assessor beliefs. If the study is randomized and double-blinded, then issues about the comparability of ancillary treatments, study procedures, and follow-up should be minimized. If not, all these issues need to be considered as potential biases as well.

If the study appears to be internally valid, the final question to consider would be whether the results are generalizable. In particular, are the results relevant to your specific PICO? Note that even if a study is not truly generalizable, it might still be applicable to your PICO. Generalizability is always a subjective decision, but generally speaking a broad patient population covering the range of patients with the disease of interest makes a study generalizable to other similar settings.

Interpreting results in diagnostic radiology. The fundamental measures of performance of a diagnostic are a test's sensitivity and specificity (see Table 4 for detailed calculation of these and other measures described here). Sensitivity is the probability that the test will correctly identify someone with the disease as having the disease (47–50). Specificity is the probability that the test will correctly identify someone without the disease as not having the disease (47–50). Neither sensitivity nor specificity tells anything about the other characteristic. Although sensitivity and specificity are often thought of as absolute characteristics of a test, in reality both measures, but especially sensitivity, depend on the distribution of the severity of the disease in the population studied as well. If a population is studied that only includes extreme cases (e.g. cancers of >5 cm or <5 mm), then the reported sensitivity and specificity will appear higher than if a range of patients is studied. This occurs because some small masses detected by the gold standard test will be missed by the experimental test, decreasing sensitivity. Similarly, some small masses may have been correctly detected by the test but will be missed by the gold standard test because the gold standard itself is unlikely to be perfect. This means that the patients would be misclassified by the gold standard as not diseased, reducing the reported specificity of the experimental test. It is thus important that the results be based on a representative sample of all patients potentially with the disease, not just extreme examples.

A publication should also report the precision of these estimates. The precision of each estimate depends on the number of people involved in the estimate. For sensitivity, this is the number of people classified as having the disease (based on the gold standard), whereas for specificity this is the number of people classified as not having the disease. It is not the size of the total population that matters, but rather the size of the smaller group (diseased or not diseased) that matters for the overall utility of a study, because both sensitivity and specificity matter. Precision is conventionally reported in terms of a 95% confidence interval. Confidence intervals are calculated using specific procedures in such a way that if the procedures are followed in a large number of experiments, then the true value for the measurement being estimated will be within the calculated confidence interval 95% of the time (51–53).

The prevalence of the disease in the population being studied determines two other important characteristics: the positive predictive value (PPV) and the negative predictive value (NPV) of the test in the population studied. The positive predictive value is the probability that someone with a positive test actually has the disease (47,54). The negative predictive value is the probability that someone with a negative test does not have the disease (47,54). Both can be expressed as percentages. These characteristics, however, are specific to the population studied, and need to be determined in your clinical population.

Both the sensitivity and specificity are used in calculating the likelihood ratio (LR) for either a positive (LR+) or a negative (LR-) test (55). The LR+ is calculated by dividing the sensitivity of the test by 1-specificity (or 100-specificity, if

TABLE 4. Statistical Measures for Diagnostic and Therapeutic Literature

Measure	Meaning	Interpretation
Diagnostic literature		
Sensitivity	Patients with the disease with positive test results ("true positives")/patients with disease	Negative test result in a highly sensitive test can rule out the disease if prevalence is relatively low
Specificity	Patients without the disease with negative results ("true negatives")/ well people (no disease)	Positive test result in a highly specific test can rule in the disease if prevalence is relatively high
Positive predictive value (PPV)	Patient with true positive test results/ all patients with positive test result (true positives + false positives)	This value is affected by the prevalence in the population studied. If a test has a high PPV (>95%) in your population, one may confidently commence treatment
Negative predictive value (NPV)	Patients with true negative results/ all patients with negative test result (true negatives + false negatives)	This value is affected by the prevalence in the population studied. If a test has a high NPV (>95%) in your population, one May be able to safely withhold treatment
Likelihood ratio (LR)	Likelihood of a given test result in patients with disease/the likelihood of the same result in patients without disease	LR = 0 (negative LR): excludes the disease LR = infinity (positive LR): confirms the disease LR = 1 the test result is uninformative as the result is equally likely in the two groups
Therapeutic literature		
Relative risk (RR)	Risk of developing disease (or outcome) in the treatment group/risk of developing disease (or outcome) in the control group	RR = 1: no difference between groups RR <1 treatment reduces the risk of the disease RR >1 treatment increases the risk of the disease
Absolute risk (AR)	Difference in the rates of events calculated as the rate in the control group minus the rate in the experimental group	AR = 0 no difference between groups AR positive: treatment is beneficial AR negative: treatment is harmful
Number needed to treat	The number of patients needed to be treated for one of them to benefit from the treatment	
Number needed to harm	The number of patients needed to be treated for 1 patient to experience an adverse effect	

percentages are used) of the test. This formula is equivalent to how likely a positive result would be found in a patient if they had the disease divided by how likely a positive result would be found in a patient who did not have the disease (47,55). Similarly, the LR- is calculated as 1-sensitivity (or 100-sensitivity, if percentages are used) divided by the specificity of the test: how likely a negative test result would be found in a patient who had the disease divided by how likely a negative test result would be found in a patient who did not have the disease (47,55). These likelihood ratios are useful because they can be used to calculate the odds of having the disease after the test results are known for the specific patient (known as the posttest odds). This is done by multiplying the appropriate LR (either LR+ or LR-, depending on the test result) by the specific patient's pretest odds of having the disease (47,55,56). The pretest odds are calculated by dividing the patient's chance of having the disease by the chance of the patient not having the disease before the test. Thus, a pretest odds of 1 means that there is a 50:50 chance that the patient has the condition before doing the diagnostic test. The posttest odds can be converted into a percent or probability with the formula Odds/(1+Odds) (47,55,57).

The likelihood ratio can also be used to assess the utility of a test itself. Because LRs are calculated based on sensitivity and specificity of the test, they should be independent of the population studied (with the caveat that the population represents the full spectrum of disease as discussed previously). LRs can also be calculated for indeterminate results. Perhaps most usefully, they can be used to combine the results of multiple diagnostic tests to estimate the patient's odds of having the condition after the sequence of tests is completed.

Statistical measures for therapeutic studies. Results from therapeutic studies can be summarized in numerous different ways. Depending on the outcome being measured, the results can be summarized as a hazard ratio (HR; when comparing the time to event), as relative risk (RR; the ratio of the proportions with an adverse outcome at a specific time point) or as attributable risk (AR; the difference in proportions with an adverse outcome at a specific time point) (58,59). Other measures based on the AR are the number needed to treat (harm) (NNT or NNH; when there is a reduction [increase] in the event rate with treatment) (58,59). Possibly the NNT is the most important of these measures—it shows how many patients would need the intervention for one patient

to actually benefit from it. In therapeutic studies, one should also find either (or both) *P* values and confidence intervals on the estimated effects, to help indicate whether the effects are likely to be due to chance.

For all these outcome measures, there is a baseline rate reflecting no effect of the intervention. For both HR and RR, no effect would be a rate of 1 (so both the intervention and the control have the same rates), whereas for AR it would be zero so that the event rates are the same in the two groups. Using the definition of HR and RR in Table 4 (with rates calculated as the rate in the experimental group/rate in the control group), values less than 1 would indicate that the intervention has reduced the rate of the event. If AR is calculated in Table 4 as the rate in the control group minus the rate in the experimental group, a positive value would mean that the rate is lower in the experimental group than the control group.

These different measures may give very different impressions of the importance of the effect. For example an effect of 0.1 in the RR scale would seem to be very dramatic: the risk of the event is only 1/10 in the experimental group the risk in the control group. But this measure does not consider the magnitude of the risk in the control group. The RR would be 0.1 if the rates are 0.1% and 1% in the experimental and control groups, respectively, or if they are 5% and 50%. For this reason, the AR and its related measure NNT (which is 100/absolute value of the AR in percent) are probably much more relevant to the practicing physician. In the first example, the AR is 0.009% and one would need to treat about 111 people to see one person benefit (NNT = 100/0.9 = 111.1), whereas in the second example the AR is 45% and the NNT is approximately 2.2.

The *P* value is a measure of how likely results as extreme would occur, by chance, if all the assumptions of the *P* value calculations are correct (60). *P* values are based on the distribution of a test statistic, which is calculated as the ratio of the effect size divided by a measure of the variability of the effect size (60). *P* values are the probability of observing a more extreme test statistic when the null hypothesis is true (60). A small *P* value means that the results would be unusual if the null hypothesis is true. A very small *P* value occurs when the test statistic is large, which can happen because the effect is large, because the variability is small, or from a combination of the two. The variability in an experiment is a function both of the inherent noise in the experiment and the number of subjects in the experiment—the larger the experiment, the smaller the variability is likely to be—so that one can have dramatically small *P* values even if the effect itself is relatively small. There are several critical assumptions in *P* value calculations including that there is no bias in the results. This is why it is essential to evaluate the validity of the study when reviewing a study, and not focus solely on the reported results. By convention, *P* values less than .05 are considered statistically significant.

As mentioned previously, confidence intervals are calculated using procedures such that if the procedures are repeated numerous times than the true parameter value will be included in the confidence interval the specified percent of times. They provide an idea of the precision of the estimated effect (52). If confidence intervals are given without *P* values, if the 95% confidence interval excludes the baseline value (1 for RR or HR; 0 for AR) then the results would be considered statistically significant at $P < .05$ (52,60).

Critically appraising secondary literature. Systematic reviews and/or meta-analysis involve first; finding the evidence by framing objectives of the review and identifying the relevant primary literature; second, assessing study quality, and applicability of the primary studies to the clinical problem at hand; third, summarizing the evidence, qualitatively and if appropriate, quantitatively (ie, doing a meta-analysis), and fourth, clinical interpretation and application of findings and possible development of recommendations.

The main problems in secondary studies (not found in primary studies) are the identification and then selection of primary studies to be included in the review/meta-analysis. These problems are analogous to the identification of potential participants and then their recruitment in a research study. Ideally one wants to identify all potential participants and then recruit a representative sample.

There are two separate issues involved in identifying the literature. First is the need for a broad search to ensure that as much of the available literature relevant to the question is retrieved. This requires searches for the published literature, much as one would do for a CAT. But in addition, the authors of secondary studies need to find what are termed the "file drawer" studies—studies done but whose results have not been widely disseminated (published), often because the results are not statistically significant or are ambiguous. This reflects the preference of authors to work on manuscripts that have some positive results and journal editors to publish positive studies. By including such studies, the synthesis/meta-analysis will include all the relevant results, and not just those that show a significant effect. It should be clear in a quick review of the methodology how the authors of the secondary literature identified the relevant published literature and undertook to identify relevant but unpublished results.

The second major problem in doing a systematic review is the selection of studies to abstract for inclusion in the review. Not all studies will contain all the information needed, especially for a meta-analysis and sometimes studies will not be done as rigorously as one might wish. If some studies are excluded either for missing data or for quality reasons, it should be clear what the standard was for judging study quality and which studies were excluded for such reasons.

To help in the formal appraisal of the quality of the secondary literature, there are several instruments useful for quality assessment: QUOROM (quality of reporting of meta-analyses); PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses; Appendix Table 3), which has superseded the existing QUOROM Statement; AMSTAR (assessment of multiple systematic reviews; Appendix Table 4) (61–63); and AGREE (Appraisal of

Guidelines for Research and Evaluation (64), now updated to AGREE II; Appendix Table 5).

PRISMA (Appendix Table 3) is a reporting checklist, consisting of 27 headings and subheadings describing the preferred way to present each section of the review (61). The items include using a structured format for the abstract, describing the clinical problem, rationale for the intervention and explicit rationale for the review, conducting a complete search and mentioning any restrictions that were applied, describing the inclusion and exclusion criteria (with special regard to the population), intervention, outcomes and the study design, assessing validity and quality of included studies, explaining the process used for data abstraction (eg, duplicate abstraction), describing included study characteristics and how heterogeneity was assessed, describing the method for combining results, how the missing data was handled and if publication bias was assessed, providing a "trial flow" in the result section, presenting study characteristics, and quantitative data synthesis, and finally summarizing the key findings in the discussion section (61,62).

AMSTAR (Appendix Table 4) is a new instrument for evaluating the quality of systematic reviews (63). It consists of 11 items evaluating the presence of "a priori" design for the review article, performing duplicate study selection and data extraction, as well as a comprehensive literature search, using the status of publication as an inclusion criterion, providing the list of included and excluded studies and the characteristics of included studies, assessing the scientific quality of included studies, and using it in formulating the conclusion, use of methods for assessing heterogeneity, publication bias, and combining the results of studies, and finally clearly describing the potential sources of support (63).

In addition, there is also an instrument for assessing guidelines the Appraisal of Guidelines for Research and Evaluation (AGREE) instrument (64) (now updated to AGREE II, Appendix Table 5). Since its original release in 2003, the AGREE instrument has advanced the science of practice guidelines (PG) appraisal and quickly became the international gold standard for PG evaluation and development (64). PGs are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances. AGREE is a valid and reliable tool that can be applied to any PG in any disease area and can be used by health care providers, guideline developers, researchers, decision/policy makers, and educators (65). The AGREE II instrument can also serve as the foundational development framework when developing a new practice guideline. The AGREE II instrument has three goals: to assess the quality of clinical practice guidelines (CPGs), to provide a methodologic strategy for the development of guidelines, and to recommend how and what information should be reported in guidelines (65). AGREE II assesses CPGs based on six domains: 1) scope and purpose, 2) stakeholder involvement, 3) rigor of development, 4) clarity and presentation, 5) applicability, and 6) editorial independence (65). Each domain is scored using several items. There are a total of 23 items in all—to be scored on a 7-point Likert scale by at least two independent observers (65).

To reiterate, the main potential problem areas for any systematic review or meta-analysis are the identification and selection for articles included in the review. There needs to be a clear and transparent process for identifying the published literature and a process in place to identify relevant material that has not been formally published. This is particularly important because negative or inconclusive results often are not published. In addition, it should be clear what papers were and were not included in the review and why articles were excluded.

Results of a meta-analysis. In a meta-analysis, the objective is to summarize the results from a number of studies into an overall estimate of effect with the expectation that the overall result will be more precise than the results of any individual study. In addition, a meta-analysis can assess whether there is significant heterogeneity (variability or differences) between the studies, and, on occasion, attempt to identify some of the reasons for the heterogeneity. Results from the individual studies are combined into an overall measure by weighting the estimates in the individual studies either by study size or study precision. For diagnostic tests, two such measures would be needed (eg, sensitivity and specificity; LR+ or LR-; PPV or NPV) because no single measure accurately presents the tests performance for both diseased and healthy individuals. Results of the meta-analysis are traditionally displayed in a figure, called a forest plot (66). In a forest plot, by convention individual studies are represented by a black square and a horizontal line that correspond to the point estimate and 95% CI for the estimate, respectively. The size (area) of the black square reflects the weight of the study in the meta-analysis. The diamond at the bottom represents the combined or pooled estimate with its 95% CI (66).

There are several different ways in which the combination can be done, depending on whether there is evidence of substantial variability (heterogeneity) between the studies. There are several different ways of assessing heterogeneity of which Cochran's Q test (which assesses the statistical significance of the heterogeneity) (67,68), and the I^2 index (which quantifies the magnitude of heterogeneity between studies compared to the total variability both within and between studies) are probably the most important (68). The I^2 index ranges from 0 (no heterogeneity) to 1 (or 100%) when all variability is due to heterogeneity between studies (68). Even if there is not statistically significant heterogeneity, if the I^2 index is large (0.5), this suggests that the analysis needs to incorporate heterogeneity when assessing the overall precision of the results of the meta-analysis (68).

Summarizing the results of your literature review. After evaluating all the retrieved evidence, the article or articles that are most relevant to the current clinical question, have a higher level of evidence, and are more valid compared to other articles at the same level of evidence should be selected.

For diagnostic literature, the resulting data from the selected article(s) can be analyzed rapidly with the help of a spreadsheet

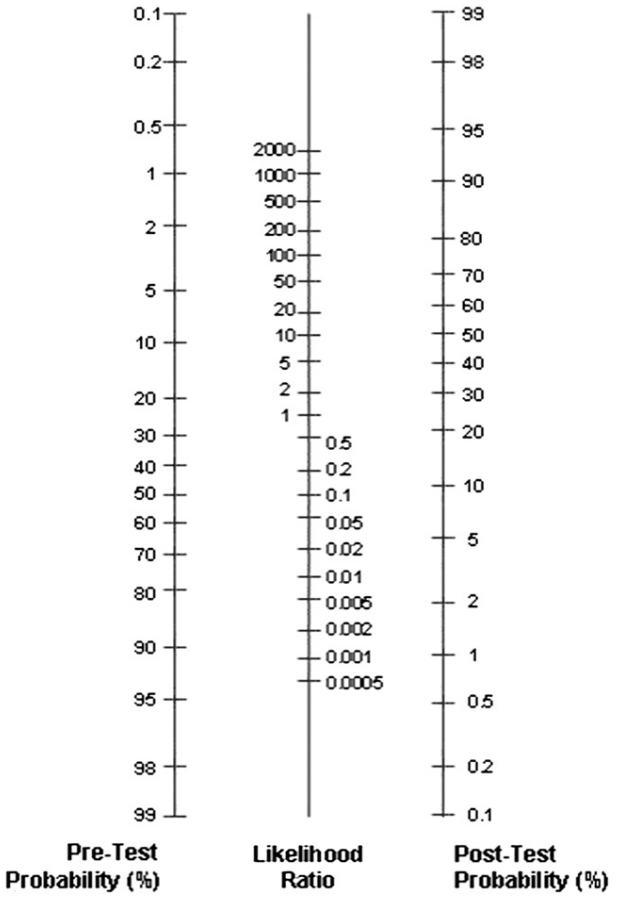

Figure 2. Likelihood ratio (Fagan) nomogram. Posttest probability is derived by drawing a straight line from the pretest probability vertical axis to the appropriate likelihood ratio and continuing the straight line to the vertical posttest probability axis. Where this line intersects the vertical posttest probability axis is the posttest probability (72).

program designed for this purpose. The Microsoft Excel version of this program can be downloaded from http://www.radiography.com/pub/ (47,69). This spreadsheet can simply be completed by entering the result data (number of patients with true positive or negative and false positive or negative, sensitivity and specificity, prevalence of the disease). The program then automatically draws a graph of conditional probabilities (Fig 3), which can be interpreted easily.

For therapeutic literature, the comparator and standard therapy would be evaluated for their efficacy, safety, recurrence rate, and survival rate.

Step 4: Apply

In this step, the results from the selected and appraised article(s) are interpreted with regard to the level of evidence of the article, its internal and external validity and the similarity of the PICO of article and PICO of the current clinical question. Then, the results will be applied to the patients' problem of interest. In step 4, what the test result means to an individual patient is assessed. At the "apply" stage, the information acquired is applied to improve clinical decision making. This is done by transforming the diagnostic test characteristics (sensitivity and specificity) to the likelihood of or probability of the patient having the disease. Clinicians work in terms of disease probabilities. Probability may be thought of as ranging from absolute certainty that the patient does not have the disease in question to absolute certainty that the patient does have the disease in question. Absolute certainty is rarely achieved in clinical medicine, so clinicians work in terms of probability thresholds. The action (or treatment) threshold is a level of probability above which the clinician is content to treat the patient for the disease in question, whereas the exclusion (no treat) threshold is a level of probability below which that disease is disregarded by the clinician. In between these two thresholds is a gray area.

The role of diagnostic imaging to try to move the probability assessment above the action threshold or below the exclusion threshold by converting the pretest probability (ie, the clinician's estimate of the patients probability of having disease given all available data before imaging) to the posttest probability (ie, the pretest probability is updated by using additional information, obtained from the imaging test result). To be useful, the diagnostic test results should move the probability estimate above or below the treat/no treat thresholds to aid clinical decision-making.

The results of clinical tests are usually used not to categorically make or exclude a diagnosis but to modify the pretest probability in order to generate the posttest probability. To apply the result of a diagnostic test to a patient, test sensitivity and specificity estimates need to be transformed into disease probability. This is achieved using Bayes theorem (70). The Bayes theorem is a mathematical relationship that allows the estimation of posttest probability. The Bayes theorem describes how to update or revise beliefs in the light of new evidence. Applied to diagnostic tests, the theorem describes how the result of a test (positive or negative) changes our knowledge of the probability of disease. This is done by combining the pretest odds of the disease (eg, estimated from clinical experience, local disease prevalence) with the likelihood ratio of the test (calculated from the test sensitivity and specificity). Therefore, pretest odds (based on clinical information) are as important as the strength of the diagnostic test in determining and interpreting the test result. In routine clinical practice, there are two ways of using the Bayes theorem to estimate posttest odds by direct mathematical calculation or using graphical methods (70). An advantage of using graphical methods is that they can be applied directly to pretest probabilities to obtain posttest probabilities, whereas direct calculation requires first converting pretest probability to pretest odds, and then later converting posttest odds to posttest probability. There are two ways to visually calculate a posttest probability; either one can use the Fagan's nomogram using the positive or negative likelihood ratios (Fig 2) or uses the graph of conditional probabilities using the pretest probability and the test result (Fig 3) (47,70,71).

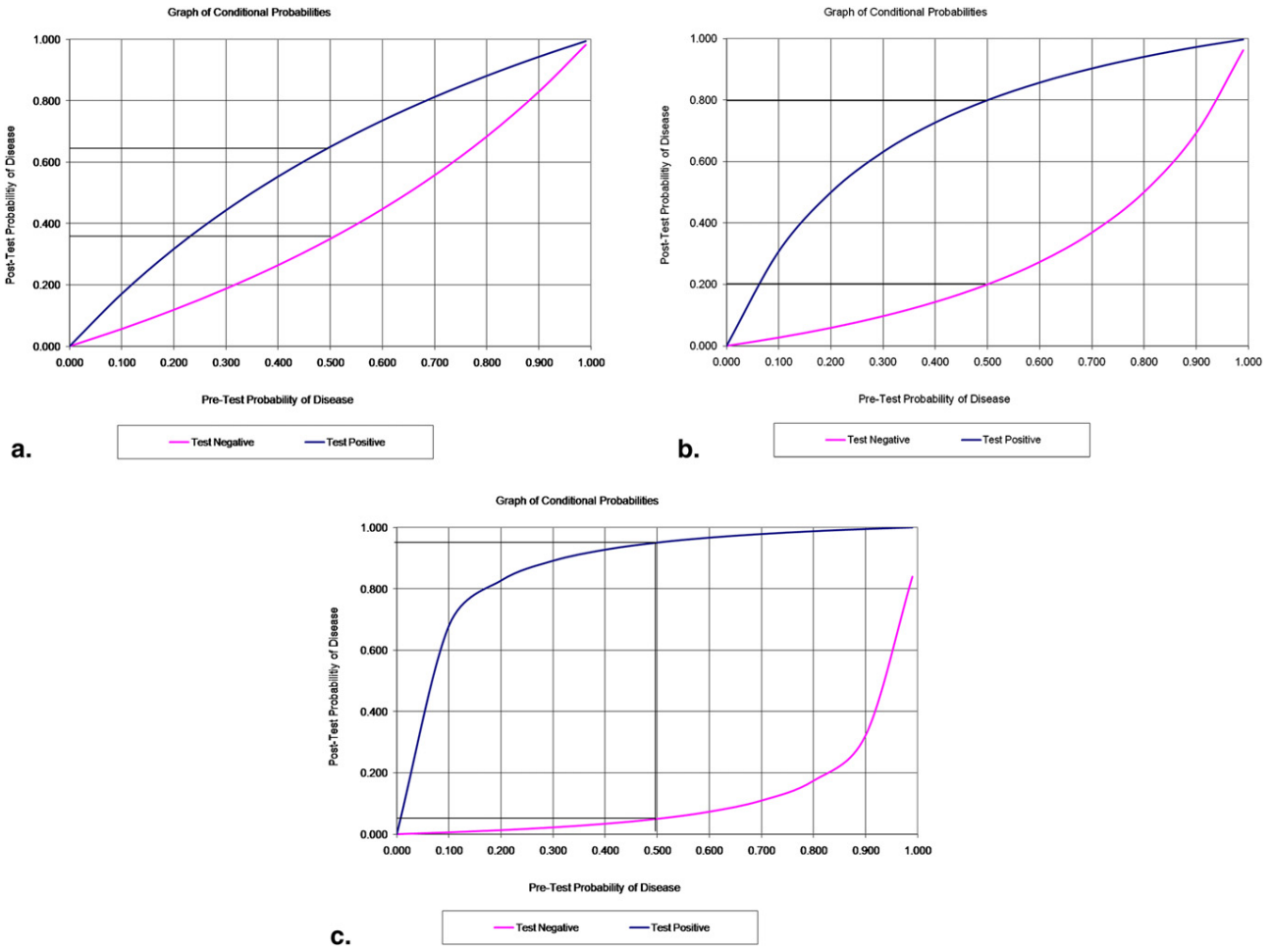

Figure 3. Use of graph of conditional probabilities to achieve clinical resolution. *Blue line* indicates a positive test result, pink line indicate a negative result. Posttest probability for a positive result is derived by drawing a vertical line up to the *blue curved line* and then across to the y-axis. Posttest probability for a negative result is derived by drawing a vertical line up to the *pink curved line* and then across to the y-axis. **(a)** A weak diagnostic test (sensitivity 65%, specificity 65%). **(b)** A moderate diagnostic test (sensitivity 80%, specificity 80%). **(c)** A strong diagnostic test (sensitivity 95%, specificity 95%).

Fagan's nomogram is a paper-based method of obviating the calculations and is as effective as the graph of conditional probabilities (72). The likelihood ratios for positive and negative results of the test in question must be known to use the nomogram (70). The graph of conditional probabilities can be quickly used to estimate the posttest probability of disease in any individual patient with a positive or negative test result. Using the conditional probability approach with its potential for graphic representation, not only can one see the overall value of a test but one can use the data over the whole spectrum of individual patient risks, regardless of the disease prevalence in the local population. Figure 3 shows how the posttest probabilities are altered given a positive or negative test result compared to a pretest probability of 0.5 using graphs of conditional probabilities for a weak diagnostic test (sensitivity 65%, specificity 65%; Fig 3a), for a moderate diagnostic test (sensitivity 80%, specificity 80%; Fig 3b), and for a strong diagnostic test (sensitivity 95%, specificity 95%; Fig 3c).

For interventional radiology, it is important to assess if the individual patient or group of patients is similar to the study population. One should estimate whether the patient is much more likely, just as likely, as or less likely to be helped/harmed than a typical study patient. Then, one should correct the NNT and/or NNH to reach an adjusted NNT and/or NNH for the individual patient.

The pros and cons of the test or treatment are then weighed for the target patient or group of patients. Availability of the diagnostic test or treatment in question, ease of access, and cost compared to the standard test or treatment; and how its benefits (eg, change in pretest probability for diagnostic tests or change in survival for treatments) compare to the costs or availability may be discussed. In this section a conclusion is obtained from the whole search and appraisal process with regard to the clinical problem and the applicability of the test or treatment would be discussed.

Step 5: Evaluate

In this section, how the information gained is applied to or may alter clinical practice is evaluated while acknowledging that ongoing developments in imaging technology provide ever-improving accuracy in diagnosis. It is important to evaluate technical parameters and generations of equipment used for any given imaging technique described. The limitations associated with the available literature on the subject of interest should be discussed and additional better quality studies or secondary studies recommended if necessary.

Helpful tools. The websites for the Centre of Evidence-based Medicine in Oxford, the Centre of Evidence-based Medicine in Toronto, and evidence-based radiology.net all have useful resources (26,73,74). They provide information and direct links to many websites that can be used to search for secondary and primary literature. Examples include the following (the web links for them are included in Table 5).

Good secondary sources

- **Level 3: Evidence-based reviews**
  - Search engines
    - The Cochrane Library
    - DynaMed
    - SumSearch
  - Databases
    - TRIP database (Turning Research in to Practice)
  - **Guidelines**:
    - US
      - National Guidelines Clearinghouse
    - UK
      - National Library for Health
      - NICE (National Institute of Clinical Excellence)
      - SIGN (Scottish Intercollegiate Guidelines Network)
    - Other
      - Canadian Medical Association
      - New Zealand Guidelines Group
- **Level 2: Synopses**
  - Structured abstracts:
    - EBM Online
    - ACP Journal Club
- **Level 1: Information Systems**
  - Evidence-based summaries:
    - Bandolier
    - Clinical Evidence
    - Up to Date
- **Other**
  - Systematic reviews:
    - Cochrane Library
  - To search several of the databases simultaneously:
    - TRIP database (Turning Research in to Practice) at www.tripdatabase.com

Good primary sources

- Free searching
  - PubMed
    - PubMed Clinical Queries
    - Google Scholar
- Subscription-based searching
  - Ovid
    - Ovid tutorial
  - Knowledge Finder
  - RSNA Index to the Imaging Literature

The Centre of Evidence-based Medicine in Oxford has critical appraisal worksheets to help appraise systematic reviews, diagnostic studies, and randomized control trials. They also have calculators for an all-purpose 2 × 2 table, an interactive nomogram that generates posttest probabilities from likelihood ratios, and a confidence interval calculator. They have explanations and examples of pretest probability, SpPIn and SnNOut, likelihood ratios, and NNTs. In addition, there is CATmaker, a computer-assisted critical appraisal tool, which helps one to create CATs for articles about therapy, diagnostic tests, prognosis, etiology/harm, and systematic reviews of therapy. Table 5 includes the website links to several useful resources.

CATmaker

A computer-assisted critical appraisal tool has been designed to facilitate the process of creating CATs and help clinicians save time. This software is designed by Oxford Center for Evidence-based Medicine and is available at their website (26). CATmaker assists with the process by carrying out the important clinical calculations, storing appraisals as well as the search strategies that lead to them and generating files that can be formatted with word processors and restored.

The evidence-based radiology website contains spreadsheets to calculate sensitivity and specificity (with confidence intervals), positive and negative predictive value, and positive and negative likelihood ratios and how these affect the pretest and posttest probability and conditional probabilities (73). It also generates graphs of conditional probability.

Using CATs to teach evidence-based medicine/imaging. Evidence-based medicine principles should be introduced as early on as possible in the student's curriculum, at a medical student level. Reviewing and critically appraising a CAT including reviewing guidelines and appropriateness criteria such as the American College of Radiology (ACR) Appropriateness Criteria can be used to teach EBM principles. In a study that introduced two focused sessions on evidence-based imaging during the required radiology core clerkship at their institution, the authors found that the ACR Appropriateness Criteria are a valuable resource for teaching evidence-based imaging to medical students, and a majority of students indicated that they plan to use this resource in the future (75).

For current trainees, already at resident level, EBM also needs to be introduced. Learning evidence-based medicine, biostatistics, epidemiology, and critical appraisal skills in

TABLE 5. Helpful Web Sites

Name	Link
PubMed, US National Library of Medicine, and National Institutes of Health	http://www.ncbi.nlm.nih.gov/sites/entrez
Medline/Ovid SP	http://gateway.ovid.com/
EMBASE	http://www.embase.com/
Google Scholar	http://scholar.google.com/
ISI Web of Knowledge	http://apps.isiknowledge.com/UA_GeneralSearch_input.do?product=UA&search_mode=GeneralSearch&SID=4BKm4mOcfDnofbKjHBI&preferencesSaved=
MD Consult	http://www.mdconsult.com/php/120885574–2/homepage
Cochrane Collaboration	http://www.cochrane.org
The Cumulative Index to Nursing and Allied Health Literature	http://www.ebscohost.com/cinahl/
National Institute of Clinical Excellence	http://www.nice.org.uk/
Scottish Intercollegiate Guidelines Network	http://www.sign.ac.uk/
SUMSearch search engine	http://sumsearch.uthscsa.edu/
The National Library for Health	http://www.library.nhs.uk/Default.aspx
National Guidelines Clearinghouse	http://www.guideline.gov/
PubMed Clinical Queries	http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml
The American College of Physicians Journal Club	http://www.acpjc.org/
BMJ Evidence Based Medicine	http://ebm.bmj.com/
Bandolier electronic journal	http://www.medicine.ox.ac.uk/bandolier/aboutus.html
The Turning Research Into Practice Database	http://www.tripdatabase.com/index.html
UpToDate	http://www.uptodate.com/home/index.html
Dynamed Clinical Reference Tool	http://www.ebscohost.com/dynamed/
Physicians Information and Education Resource	http://pier.acponline.org/index.html
BMJ Clinical Evidence	http://clinicalevidence.bmj.com/ceweb/index.jsp
The Database of Abstracts of Reviews of Effects	http://mrw.interscience.wiley.com/cochrane/cochrane_cldare_articles_fs.html
The Cochrane Central Register of Controlled Clinical Trials (CENTRAL)	http://www.mrw.interscience.wiley.com/cochrane/cochrane_clcentral_articles_fs.html
The National Library of Medicine	http://www.nlm.nih.gov/
Netting the Evidence	http://www.shef.ac.uk/scharr/ir/netting/
The Center for Evidence Based Radiology at the Brigham and Women's hospital	http://www.brighamandwomens.org/cebi/default.aspx
Centre for Reviews and Dissemination at the University of York	http://www.york.ac.uk/inst/crd/
The Centre for Health Evidence at the University of Alberta	http://www.cche.net/
The Centre for Evidence Based Medicine at Oxford	http://www.cebm.net/
The Centre for Evidence Based Medicine at the University Health Network, Toronto	http://www.cebm.utoronto.ca/practise/evaluate/index.htm#top
The Blue Cross and Blue Shield Association Technology Evaluation Center	http://www.bcbs.com/blueresources/tec/
BestBETs Best Evidence Topics	http://www.bestbets.org/
CAT crawler	http://www.bii.a-star.edu.sg/achievements/applications/catcrawler/cat_search.asp
Centre for Evidence Based Radiology	http://www.evidencebasedradiology.net

general fall under practice-based learning and improvement (76). Practice-based learning and improvement is one of the Accreditation Council for Graduate Medical Education's six general competencies that is particularly relevant to training today's residents in diagnostic radiology, both because of the growth of medical imaging and the cost associated with this imaging (77). An introduction to the tools and process of evidence-based radiology (EBR) provides a forum for teaching and promoting practice-based learning in a diagnostic radiology curriculum. A resident journal club facilitates the introduction of evidence-based radiology into a diagnostic radiology resident education program (77). There are regular EBR journal clubs at several institutions (76,77). Critically appraised topics are an excellent way to teach EBM principles at a journal club, initially reviewing and critically appraising a CAT, and later assisting trainees writing a CAT.

With rapidly growing health care and imaging costs, it has become necessary for radiologists to play an integral role in assisting referring physicians in making cost effective decisions with respect to imaging studies for their patients (78). Current radiologists and future generations of radiologists must be able to provide evidence-based decision support, or at the very least, know where to find the necessary information and CATs are an excellent resource.

CONCLUSION

Medical knowledge has expanded greatly over the past 50 years, coupled with an increasing complexity, which can be overwhelming to the everyday clinical practitioner. Paralleling these improvements in medicine, the volume of scientific articles published has exploded. Evidence-based practice and publications in this area have developed to help practitioners keep up to date with the increasing volume of information to solve complex health problems.

A user friendly format to share EBP information is the CAT, which is a standardized summary of research evidence organized around a clinical question, aimed at providing both a critique of the research and a statement of the clinical relevance of results. Critically appraised topics provide easy access to the scientific literature for busy clinicians facing a clinical question with insufficient time to assess the vast (and mixed) results from a search engine or who lack the specialized skills to critically appraise the literature and reach an appropriate conclusion themselves. The main reason to produce a CAT is to answer an explicit clinical question arising from a specific patient encounter, and is the essence of bottom up EBP in that a health professional generates a clinical question from a real clinical situation, followed by finding and appraisal of the evidence, and finally applying it in clinical practice.

In this review, the need for and the usefulness of CATS has been outlined, the steps involved in the performance of and the writing up a CAT for a clinical purpose have been explained, and available electronic resources have been introduced to and summarized for the reader.

ACKNOWLEDGMENT

Funded in part by the GE-RSNA Radiology Educational Scholarship Award.

REFERENCES

1. Dawes M. Critically appraised topics and evidence-based medicine journals. Singapore Med J 2005; 46:442–448. quiz 449.
2. Fetters L, Figueiredo EM, Keane-Miller D, et al. Critically appraised topics. Pediatr Phys Ther 2004; 16:19–21.
3. Wendt O. Developing critically appraised topics (CATs). Presented at the American speech-language hearing association, division 12: augmentative and alternative communication (DAAC), 7th annual conference, San Antonio, January 2006. Available at: http://www.edst.purdue.edu/aac/Developing%20Critically%20Appraised%20Topics.pdf. Accessed May 3, 2010.
4. Dawes M, Summerskill W, Glasziou P, et al. Sicily statement on evidence-based practice. BMC Med Educ 2005; 5:1.
5. Glasziou P, Salisbury J. EBP step 1: formulate an answerable question. Evidence based practice workbook, 2nd ed. Malden, MA: Blackwell Publishing; 2007. p. 21–38.
6. Staunton M. Evidence-based radiology: steps 1 and 2—asking answerable questions and searching for evidence. Radiology 2007; 242:23–31.
7. Feussner JR, Matchar DB. When and how to study the carotid arteries. Ann Intern Med 1988; 109:805–818.
8. Kelly AM. Evidence-based radiology: step 1—ask. Semin Roentgenol 2009; 44:140–146.
9. McGrane S, McSweeney SE, Maher MM. Which patients will benefit from percutaneous radiofrequency ablation of colorectal liver metastases? Critically appraised topic. Abdom Imaging 2008; 33:48–53.
10. Kelly AM, Fessell D. Ultrasound compared with magnetic resonance imaging for the diagnosis of rotator cuff tears: a critically appraised topic. Semin Roentgenol 2009; 44:196–200.
11. El-Maraghi RH, Kielar AZ. CT colonography versus optical colonoscopy for screening asymptomatic patients for colorectal cancer a patient, intervention, comparison, outcome (PICO) analysis. Acad Radiol 2009; 16:564–571.
12. El-Maraghi R, Kielar A. Low-dose computed tomographic colonography versus optical colonoscopy: a critically appraised topic. Semin Roentgenol 2009; 44:191–195.
13. Czum JM. Coronary CT angiography for coronary artery stenosis: a critically appraised topic. Semin Roentgenol 2009; 44:188–190.
14. Petrou M, Foerster BR. Relative roles of magnetic resonance angiography and computed tomographic angiography in evaluation of symptomatic carotid stenosis: a critically appraised topic. Semin Roentgenol 2009; 44: 184–187.
15. Meinert CL. An insider's guide to clinical trials. Oxford, UK: Oxford University Press; 2011.
16. Friedman LM, Furberg C, DeMets DL. Fundamentals of clinical trials. Basel, Switzerland: Birkhäuser; 1998.
17. Piantadosi S. Clinical trials: a methodologic perspective. Hoboken, NJ: John Wiley and Sons; 2005.
18. Rosenbaum PR. Observational studies. Berlin, Germany: Springer; 2002.
19. Schlesselman JJ, Stolley PD. Case-control studies: design, conduct, analysis. Oxford, UK: Oxford University Press; 1982.
20. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - the analysis of case-control studies. Lyon, France: IARC Sci Publ; 1980. p. 5–338.
21. Levels of evidence. Oxford Center for Evidence-Based Medicine Web site. http://www.cebm.net/index.aspx?o=1025. Accessed May 3, 2010.
22. Haynes RB. Of studies, summaries, synopses, and systems: the "4S" evolution of services for finding current best evidence. Evid Based Ment Health 2001; 4:37–39.
23. Haynes RB. Of studies, summaries, synopses, and systems: the "4S" evolution of services for finding current best evidence. Evid Based Nurs 2005; 8:4–6.
24. Haynes B. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. Evid Based Nurs 2007; 10:6–7.
25. Dicenso A, Bayley L, Haynes RB. Accessing pre-appraised evidence: fine-tuning the 5S model into a 6S model. Evid Based Nurs 2009; 12:99–101.
26. The Centre for Evidence Based Medicine at Oxford website. http://www.cebm.net/index.aspx?o=1016. Accessed on July 23, 2010.
27. NCBI Pubmed Web site. US National Library of Medicine. http://www.ncbi.nlm.nih.gov/PubMed. Accessed May 3, 2010.
28. EMBASE. EMBASE Biomedical Answers website. http://www.embase.com/home. Accessed May 3, 2010.
29. ISI Web of Knowledge website. www.isiknowledge.com. Accessed May 3, 2010.
30. MD Consult website. www.mdconsult.com. Accessed May 3, 2010.
31. Google Scholar Beta. Google website. http://scholar.google.com. Accessed May 3, 2010.
32. Kelly AM. Evidence-based radiology: step 2—searching the literature (search). Semin Roentgenol 2009; 44:147–152.
33. The Cochrane Collaboration website. http://www.cochrane.org. Accessed May 3, 2010.
34. TRIP database. http://www.tripdatabase.com. Accessed May 3, 2010.
35. NICE. National Institute for Health and Clinical Excellence website. http://www.nice.org.uk. Accessed May 3, 2010.
36. SUMSearch. University of Texas Health Science Center at San Antonio website. http://sumsearch.uthscsa.edu. Accessed May 3, 2010.
37. CINAHL. Cumulative Index to Nursing and Allied Health Literature website. http://www.ebscohost.com/cinahl. Accessed May 3, 2010.

38. National Guidelines Clearinghouse. http://www.guideline.gov. Accessed May 3, 2010.
39. The National Library for Health. www.library.nhs.uk. Accessed May 3, 2010.
40. SIGN. Scottish Intercollegiate Guidelines Network. http://www.sign.ac.uk. Accessed May 3, 2010.
41. American College of Physicians Journal Club website. http://www.acpjc.org. Accessed May 3, 2010.
42. Evidence-Based Medicine Online website. BMJ Publishing Group. http://ebm.bmj.com. Accessed May 3, 2010.
43. Clinical Evidence website. BMJ Publishing Group. http://clinicalevidence.bmj.com/ceweb/index.jsp. Accessed May 3, 2010.
44. Up to Date website. http://www.uptodate.com/home/index.html. Accessed May 3, 2010.
45. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138:W1–W12.
46. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003; 3:25.
47. Maceneaney PM, Malone DE. The meaning of diagnostic test results: a spreadsheet for swift data analysis. Clin Radiol 2000; 55:227–235.
48. Mayer D. Essential evidence based medicine. Cambridge, UK: Cambridge University Press; 2004.
49. Sackett D, Strauss S, Richardson W, et al. Evidence-based medicine: how to practice and teach EBM. London: Churchill-Livingstone; 2000.
50. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ 1994; 308:1552.
51. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986; 292:746–750.
52. Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. CMAJ 1995; 152:169–173.
53. Halligan S, Altman DG. Evidence-based practice in radiology: steps 3 and 4—appraise and apply systematic reviews and meta-analyses. Radiology 2007; 243:13–27.
54. Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994; 309:102.
55. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004; 329: 168–169.
56. McGinn T, Jervis R, Wisnivesky J, et al. Tips for teachers of evidence-based medicine: clinical prediction rules (CPRs) and estimating pretest probability. J Gen Intern Med 2008; 23:1261–1268.
57. Chang PJ. Bayesian analysis revisited: a radiologist's survival guide. AJR Am J Roentgenol 1989; 152:721–727.
58. Barratt A, Wyer PC, Hatala R, et al. Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. CMAJ 2004; 171:353–358.
59. Jaeschke R, Guyatt G, Shannon H, et al. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. CMAJ 1995; 152:351–357.
60. Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 1. Hypothesis testing. CMAJ 1995; 152:27–32.
61. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009; 6:e1000100.
62. Moher D, Cook DJ, Eastwood S, et al. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999; 354: 1896–1900.
63. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009; 62:1013–1020.
64. The AGREE Collaboration. Appraisal of Guidelines for Research & Evaluation (AGREE) Instrument. www.agreecollaboration.org. Accessed on 2/25/2011.
65. Dans AL, Dans LF. Appraising a tool for guideline appraisal (the AGREE II instrument). J Clin Epidemiol 2010; 63:1281–1282.
66. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001; 322:1479–1480.
67. Deeks JJ AD, Bradburn M. Statistical methods for examining heterogeneity and combining results from several studies in metaanalysis. In: Systematic reviews in health care: meta-analysis in context. Oxford, UK: BMJ Books; 2001. p. 285–312.
68. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003; 327:557–560.
69. Downloadable excel spreadsheet to generate graphs of conditional probability. http://www.radiography.com/pub/. Accessed on July 23, 2010.
70. Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. Acta Paediatr 2007; 96:487–491.
71. Cronin P. Evidence-based radiology: step 3—critical appraisal of diagnostic literature. Semin Roentgenol 2009; 44:158–165.
72. Fagan TJ. Letter: nomogram for Bayes theorem. N Engl J Med 1975; 293: 257.
73. Evidence Based Imaging website. http://www.evidencebasedradiology.net/index.html. Accessed on July 23, 2010.
74. The Centre for Evidence Based Medicine at Toronto Website. http://ktclearinghouse.ca/cebm/. Accessed on July 23, 2010.
75. Dillon JE, Slanetz PJ. Teaching evidence-based imaging in the radiology clerkship using the ACR appropriateness criteria. Acad Radiol 2010; 17: 912–916.
76. Kelly AM, Cronin P. Setting up, maintaining and evaluating an evidence based radiology journal club: the University of Michigan experience. Acad Radiol 2010; 17:1073–1078.
77. Heilbrun ME. Should radiology residents be taught evidence-based radiology? An experiment with "the EBR Journal Club". Acad Radiol 2009; 16:1549–1554.
78. Logie CI, Smith SE, Nagy P. Evaluation of resident familiarity and utilization of the ACR musculoskeletal study appropriateness criteria in the context of medical decision support. Acad Radiol 2010; 17:251–254.