ACR Appropriateness Criteria®
Evidence Document

The evidence cited in the Appropriateness Criteria (AC) topic narrative is evaluated in two ways. The first is the evidence Table (ET) which summarizes the evidence cited in the narrative and the second is the strength of evidence (SOE). The ET quantifies a source's quality based on the number of study quality elements described in that source. The Strength of Evidence (SOE) is an assessment of the amount and quality of evidence found in the peer reviewed medical literature for an appropriateness recommendation. A recommendation is defined as the appropriateness category assigned to performing a radiologic procedure for a specific clinical scenario.

Evidence Table

Development

The creation and revision of the ET is performed by ACR staff in order to apply the methodology consistently. It also alleviates some of the work burden placed on the topics' authors. Once an ET is constructed, the topic's author and panel members review the ET for completeness and validity.

The ET includes five components extracted from a source. These are the **reference** citation information, study **type,** number of **patients or events**, study **objective(s)**, and study **result(s)**.

- The study type designates the source's purpose and design. The purpose of **diagnostic studies** is to diagnose or assess patients by utilizing diagnostic tools while **therapeutic studies** assess the use of treatments and interventions in treating patients. Furthermore, diagnostic and therapeutic studies have different study quality elements that help assess the amount of bias that may be introduced, which may affect the results and conclusions of the study.
- There are four study design types: **experimental**, **observational**, **review/other**, and **meta-analysis**. Additional information for classifying these study types is found in the Appendix A. These broadly defined study designs contribute to the assessment of study quality. Well-designed and well-executed experimental studies typically are better at controlling biases and determining causality where other study types, like observational studies, may determine only when there is a relationship between events and outcomes. Because of the general inconsistency in the medical literature regarding study design names and definitions, the varying degrees of adherence to study design, the hybridization of study designs, and sometimes the lack of complete information in a source's study methodology to correctly assess the study design, a broad categorization of study design may identify important differences in study quality based on study design.
- The sixth component of the ET is **study quality**. A source's **study quality** is defined as described below.

Determining the Study Quality Category

ACR staff assesses each source that is included on the ET. The study quality categories and their definitions can be found in Table 1. The staff reviews the source to determine if specific quality components are described and meet the criteria for that component. The quality components were derived mainly from some of the concepts presented in the GRADE methodology[1]. Because some quality components may not apply to specific study designs, the study quality components required are specific to diagnostic, therapeutic, and/or meta-analysis studies. The study quality components are explained in Table 2.

Table 1. Definitions of Study Quality Categories

Study Quality Category Name	Study Quality Category Definition	Criteria for Diagnostic Studies	Criteria for Therapeutic Studies
Category 1	The study is well designed and accounts for common biases.	The source has all 8 diagnostic study quality elements present.	The source has 5 or 6 therapeutic study quality elements present.
Category 2	The study is moderately well designed and accounts for most common biases.	The source has 6 or 7 diagnostic study quality elements present.	The source has 3 or 4 therapeutic study quality elements present.
Category 3	The study has important study design limitations.	The source has 3, 4, or 5 diagnostic study quality elements present.	The source has 1 or 2 therapeutic study quality elements present.
Category 4	The study or source is not useful as primary evidence. The article may not be a clinical study, the study design is invalid, or conclusions are based on expert consensus. The study does not meet the criteria for or is not a hypothesis-based clinical study (eg, a book chapter or case report or case series description); *or* The study may synthesize and draw conclusions about several studies such as a literature review article or book chapter but is not primary evidence; *or* The study is an expert opinion or consensus document.	The source has 0, 1, or 2 diagnostic study quality elements present.	The source has zero (0) therapeutic study quality elements.
Good Quality Meta-analysis	The study design, methods, analysis, and results are valid and the conclusion is supported. (For a meta-analysis study quality to be "good", all eight elements must be answered Yes.)	n/a	n/a
Inadequate Quality Meta-analysis	The study design, analysis, and results lack the methodological rigor to be considered a meta-analysis study as defined.	n/a	n/a

Study Quality Components

Table 2. Definitions of Study Quality Components

ITEM	Component	Definition	Applies To
0D or 0T	**Statistical Measure**	If a journal article does not have a statistical comparison of results, it is not classified as a study and is categorized as Review/Other and assigned a Study Quality of 4 (not useful as primary evidence). The statistical measure component is not included in the total number of study quality components that determine the study quality. If there is no statistical measure, staff does not need to assess the article for any other study quality components. The statistical measure must compare the results of the treatments, interventions, or diagnostic tools. Measures that only relate to describing the study population or number of events do not meet the criteria for this element. EXAMPLES: Dx: Sensitivity/specificity, PPV/NPV, mean, median, Kappa, Pearson r, regression co-efficient, etc Tx: Odds ratios, survival rates/curves, hazard ratios, mean or median, etc.	Diagnostic Studies, Therapeutic Studies, or Combined Diagnostic / Therapeutic Studies

ITEM	Component	Definition	Applies To
1D or 1T	**Uncertainty Measure**	Some uncertainty measures are incorporated by the statistical measure. When one of these statistical measures is used for the study's results and there is no specific discussion of an uncertainty measure, the criteria for the uncertainty measure is fulfilled. For example, if the study states the PPV or NPV but does not explicitly state the uncertainty measure, it fulfills the criteria because an uncertainty measure can be calculated. EXAMPLES: Dx: Standard errors, confidence intervals, p-values, statistical comparison tests such as t-test, Fisher exact probability, Mann-Whitney U, etc. Tx: Standard errors, confidence intervals, percentiles, power calculations for sample size, etc.	Diagnostic Studies, Therapeutic Studies, or Combined Diagnostic / Therapeutic Studies
2D or 2T	**Prospective**	The study was designed prior to the data collection, such as before performing the intervention or comparing the index test and reference standard.	Diagnostic Studies, Therapeutic Studies, or Combined Diagnostic / Therapeutic Studies
3D	**Systematic Recruitment**	The study design must minimize selection bias through systematic recruitment or consecutive series methods Studies using representative samples must avoid selection bias because they may result in conclusions that are systematically different from the truth. Systematic recruitment refers to strategies to identify relevant patient characteristics which may impact the study outcomes and successfully recruit sufficient numbers to understand if any biases may be introduced.	Diagnostic Studies only
3D	**Consecutive Series**	The study design must minimize selection bias through systematic recruitment or consecutive series methods Studies using representative samples must avoid selection bias because they may result in conclusions that are systematically different from the truth. Consecutive series refers to the recruitment of everyone who is eligible for the study over a specified period of time. It assumes that those who are recruited are randomized by chance.	Diagnostic Studies only
4D	**Standard Of Reference**	The study identifies a standard of reference, typically it should be the gold standard or at least a standard that has been studied and compared to the gold standard. If a standard of reference is not identified, the comparison between two or more tests must identify which test is considered the standard.	Diagnostic Studies only
5D	**Reference Standard Applied**	The study must compare the standard of reference to the index standard for all subjects in the same way throughout the study.	Diagnostic Studies only
6D	**Independent Readers**	To reduce bias introduced by the ability of the readers of the diagnostic tests, at least two, independent readers/interpreters are required for the reference standard and each index standard.	Diagnostic Studies only
7D	**Index Test Results**	To reduce influence by the reference tests, the index test results must be interpreted without knowledge of the results of the reference standard. In the absence of a reference standard and when more than one type of imaging, pathologic or clinical tests are being compared, this element is fulfilled when the results of at least one of the test(s) has been interpreted without the knowledge of the results of the other test(s) in the study.	Diagnostic Studies only
8D	**Reference Standard Results**	To reduce influence by the index tests, the reference test results must be interpreted without knowledge of the results of the index standard. In the absence of a reference standard and when more than one type of imaging,	Diagnostic Studies only

ITEM	Component	Definition	Applies To
		pathologic or clinical tests are being compared, this element is fulfilled when the results of all of the test(s) has been interpreted without the knowledge of the results of the other test(s) in the study.	
0T	**Blinding**	There can be many qualifications that would affect the blinding element's contribution to study quality. The ACR Methodology Subcommittee recognizes blinding subjects to aspects of a study is important when those aspects can influence how participants report subjective measures (such as, pain levels experienced) that are used as outcomes of the study. However, blinding both researchers and study subjects is likely to be an important step in treatment studies that use measures that are more objective, such as those involving administration of medications versus placebos. Because it may not be ethically responsible to blind subjects or researchers in many treatment trials and because of the complexity of the blinding element and its contribution to study quality, the Subcommittee felt it could not be assessed for therapeutic studies consistently and objectively.	Therapeutic Studies only
3T	**Allocation Of Subjects**	Studies using representative samples must avoid selection bias because they may result in conclusions that are systematically different from the truth. To reduce selection bias, therapeutic studies often use control groups and intervention groups to understand the impact of an intervention. The term "control group" refers to a group that does not receive the primary intervention, therapy, or treatment being evaluated in the study. The control group may be comparable subjects who receive no intervention *or* who receive another intervention whose outcomes are accepted *or* may have been previously studied.	Therapeutic Studies only
4T	**Random Allocation Of Subjects**	Studies using representative samples must avoid selection bias because they may result in conclusions that are systematically different from the truth. This implies that each subject being entered into a trial has the same chance of receiving any of the possible interventions. It also implies that the probability that a subject will receive a particular intervention is independent of the probability that any other subject will receive the same intervention.	Therapeutic Studies only
5T	**Length Of Follow-Up**	The length of follow-up must be stated and assessed to determine its impact on the results. When length of follow up is present, staff will record it in the ET. If there is no information, the author must state the reason for the omission or provide analysis to correct for the lack of this measurement. EXAMPLES: Length of follow up, survival rates, re-occurrence rates, toxicity rates, or similar measures	Therapeutic Studies only
6T	**Disposition Of All Subjects**	The study should report and account for all of the subjects enrolled into the study. If there are no data, the author states the reason(s) why subjects did not complete the study or why they were excluded. EXAMPLES: Lost to follow-up, drop out	Therapeutic Studies only
1M	**Two Studies**	This is part of the Meta-Analysis assessment, even though it is definitional. The primary reason to include this is to exclude those studies that use the term meta-analysis to refer to reviews of the literature for a topic. Also, systematic literature reviews sometimes may use meta-analysis in their title which is NOT what we are referring to here. In order to compare studies there must be at least two studies in the analysis. This is similar to the Statistical Measure component in that if there are not two studies in the Meta-Analysis, it is not a Meta-Analysis Study.	Meta-Analysis Studies only
2M	**Clear Purpose**	The study should establish statistical significance with studies that have conflicting results, or to develop a more correct estimate of effect magnitude, or to provide a more complex analysis of harms, safety data, and benefits, or to examine subgroups with individual number that are not statistically significant	Meta-Analysis Studies only

ITEM	Component	Definition	Applies To
3M	**Prospective**	This mimics the requirement for the Prospective component for Diagnostic and Therapeutic Studies. The three items (Specific Hypothesis, Data Collection, and Analytic Strategies) should be specified prior to knowing the results from any of the primary studies.	Meta-Analysis Studies only
4M	**Inclusion/Exclusion Criteria**	The criteria for including or excluding studies for the meta-analysis must be determined before the literature search and analysis are completed.	Meta-Analysis Studies only
5M	**Appropriate For Questions Asked**	The criteria for including or excluding studies must be appropriate for the questions the meta-analysis study is asking.	Meta-Analysis Studies only
6M	**Pooled Statistic**	This is similar in concept to the uncertainty measure for diagnostic and therapeutic studies. Using inappropriate pooled statistics may lead to erroneous results or conclusion. When one of these statistical measures is used for the study's results and there is no specific discussion, the criteria for pooled statistic is fulfilled. As long as the pooled statistic can be calculated, the component is met.	Meta-Analysis Studies only
7M	**Study Designs**	Specific study designs are more appropriate for different types of research questions. The meta-analysis should only use the most appropriate study design in the analysis, though multiple analyses to account for the different study designs included in the meta-analysis may be appropriate.	Meta-Analysis Studies only
8M	**Study Results**	The study results of the different included studies should be comparable in both direction and magnitude of the effect. The analysis may determine that the direction or magnitude of the effect varies among the studies included but they have to assess similar results so they may be compared.	Meta-Analysis Studies only

Review of the ET

- It is the role of the panel to assess the quality of the ET and to question its assessment of study quality for the sources listed in the document. An initial review of the ET takes place before the first rating round. Panel members are expected to review the ET after any edits to the document and prior to the second rating round. Ratings are based on the evidence, which can be best interpreted when the rater understands the quality of the evidence supporting the recommendations.
- In the event a panel member (the inquirer) disagrees with the study quality assessment for any article in the ET, the inquirer will contact the ACR staff, report the disputed study quality element, and provide an explanation. Staff will re-evaluate the study quality assignment for any error and report the results to the inquirer. If a requested change to the study quality is warranted based on the accepted methodology, then the change to the study quality is made in the ET.
- If the inquirer is not satisfied with the result, a second level review involving the topic author, vice-chair, and possibly the chair may be initiated. The second level reviewers will recommend either to use the study quality recommended by staff or propose an alternative and provide an explanation.
- The proposed changes will be presented to the all members of the relevant panel for approval. If two thirds of the panel members approve, the study quality change will be made to the ET. If agreement cannot be achieved by the panel (i.e., less than two thirds of the panel agrees), the staff assignment made using the methodology will be the final study quality assessment.

Strength of Evidence

Development

The creation and revision of the SOE is performed by ACR staff in order to apply the methodology consistently. It also alleviates some of the work burden placed on the topics' authors. Once the SOE has been assessed for a recommendation, the topic's author and panel members will review it for completeness and validity.

Determining the Strength of Evidence Categories

The strength of evidence assessment uses an algorithm to determine the SOE category for the reference or group of references that are associated with a recommendation. The algorithm uses multiple factors in this assessment (for example, no disagreement of the panel's ratings for the recommendation; the number and quality of the studies associated with the recommendation; the directness of the studies to the specific recommendation; and the consistency of the studies supporting the recommendation to name the most prominent factors). The process is completed after the references have been assigned by the author to each recommendation. Staff reviews the studies for directness and consistency in relationship to the recommendation and designates the references as (direct and consistent, indirect and consistent, direct and inconsistent, or indirect and inconsistent). If there is only one reference assigned to a recommendation, consistency is not applicable. The algorithm is applied to determine the SOE category.

Table 3. Definitions of Strength of Evidence Categories

SOE Category Name	SOE Category Definition
Strong	Good quality studies are used to support the recommendation. All of the studies demonstrate similar estimates of effect and relate to the same or very similar clinical conditions.
Moderate	Good quality studies are used to support the recommendation. The studies may or may not demonstrate similar estimates of the effect and correlation to the same or very similar clinical conditions.
Limited	Poor quality studies (ie, reviews) are used to support the recommendation. The studies may or may not demonstrate similar estimates of the effect and relate to the same or very similar clinical conditions.
Expert Consensus	There is no literature to support the recommendation but the Panel has no disagreement regarding the recommendation.
Expert Opinion	The rating for the recommendation has disagreement as defined below.

Algorithm

Please go to Appendix B to view the algorithm.

Components

Determining SOE for a recommendation requires identifying the number of studies relevant to the recommendation, assessing a) their study quality b) the consistency of the findings among studies, and c) how closely the study addresses the recommendation. In those instances where there is only one strong or good study, additional, lower-quality studies may help supplement the overall strength of evidence assessment.

SOE Components and Definitions

Quality – Quality is assessed by looking at the study quality for each reference associated with a recommendation and determining the number of high or good quality studies. The concept applies if there are two or more good study quality combinations (ie, any combination of good quality meta-analysis study, high quality study [category1], or good quality study [category 2]) or one good study quality combinations with 3 or more supporting, lower-quality [category 3] studies.

Consistency – All of the studies for each reference associated with a recommendation demonstrate similar estimates of effect. Differences in the direction of effect, the size of the differences in effect, and the significance of the differences will guide the decision about whether important inconsistency exists. Unexplained inconsistencies may result in less reliable study conclusions.

Directness – The findings of all of the studies associated with a recommendation relate to the same or very similar clinical conditions / indications as described in the clinical scenario (ie, topic-variant description) and relate to the same imaging study as described in the procedure. Because many interventions have more or less the same relative effects across most patient groups, overly stringent criteria in deciding whether evidence is direct may not be applied.

Supporting Studies –Additional, MULTIPLE, LOW QUALITY [study quality category 3] studies may contribute to SOE. There must be at least three studies with conclusions that are at least indirectly related to the recommendation and are consistent with both the recommendation and any other evidence associated with the recommendation.

Disagreement/No disagreement – Disagreement is when the individual ratings for the recommendation have too much variation from the group median rating as defined by the IPRAS methodology from the RAND/UCLA Appropriateness User Manual [2].

* **Meta-analysis studies** –The quality of meta-analysis studies can vary enormously. It requires a meticulous analysis to determine its quality. The meta-analysis will be assessed to determine if it is "GOOD" quality (as defined above) and can contribute to the SOE for a recommendation.

Review of SOE

It is the role of the panel to assess the SOE and to question its assessment for the recommendations listed in the document. An initial review of the estimated SOE occurs before the first rating round (the focus is on the number and study quality of the references assigned to a recommendation assuming consistency, directness, and no appropriateness rating disagreement among the panel). Panel members are expected to review the SOE after any edits to the document and prior to topic finalization.

References

1. Schünemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ.* 2008;336(7653):1106-1110.

2. Fitch K. *The Rand/UCLA appropriateness method user's manual*. Santa Monica: Rand; 2001.

Appendix A

Study Type Categories

- *Experimental*

  Experimental studies create differences in the explanatory (independent) variable under controlled conditions and examine any resulting changes in the response (dependent) variable. These studies include methodologies that reduce the potential for bias, for example, randomization, blinding. An example is the randomized controlled trial.

  Characteristics of Experimental Design

  - True experiments have control and manipulation
  - Specifies an experimental group and control group
  - Test cause and measure effect

- *Observational*

  Investigators observe subjects and measure variables of interest (independent variables) without assigning treatments, interventions, or outcomes to the subjects. The treatment, intervention, or outcome that each subject receives is determined beyond the control of the investigator

  Characteristics of Observational Design

  - Investigator observes variables
  - Specifies cohorts (groups with similar characteristics of interest) or a case (groups with the variable of interest) and control (groups without the variable of interest) group.
  - Test association between variables but not causality.

- *Review/Other*

  Reviews or other studies are case reports, systematic literature reviews, clinical practice guidelines, consensus statements, book chapters, etc. These sources may not have a statistical measure that compares the results but include published literature that examines or reviews other studies, data, surveys, opinions, etc. and summarizes results or concludes outcomes.

  Other sources in this category may be studies that have descriptive statistics only that do not provide a result or outcome to the study, such as incidence or prevalence studies that only describe population or disease trends or patterns.

- *Meta-analysis*

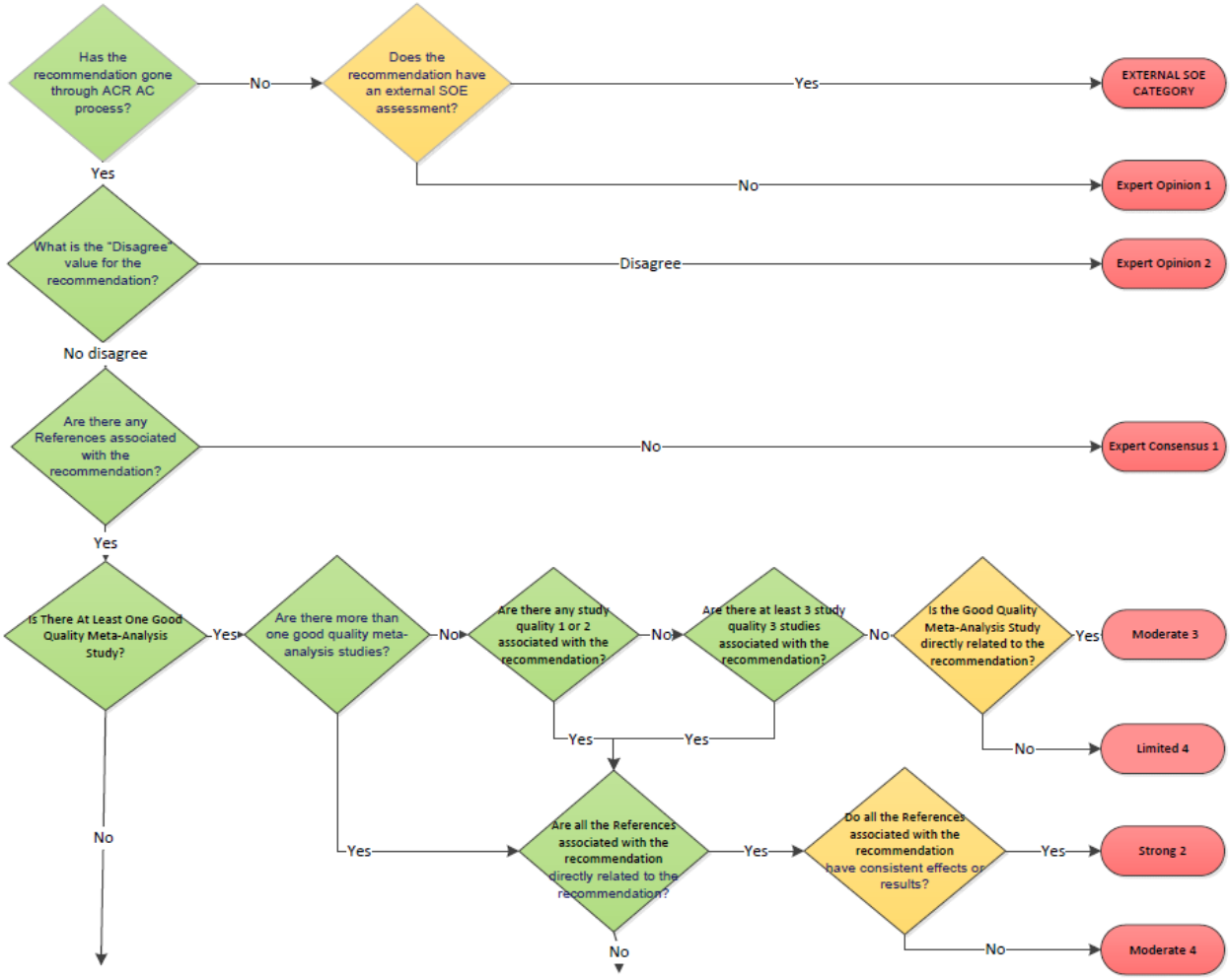
  Meta-analysis studies aggregate information in order to achieve a higher statistical power for the measure of interest, as opposed to a less precise measure derived from a single study. Other methods that do not create pooled samples using statistical methods such as systematic literature reviews and clinical practice guidelines are not included in the definition of meta-analysis. Meta-analysis studies are not multisite studies even when one of the studies in the meta-analysis is a multisite study because the same protocol may not be implemented in each study.

  Characteristics of Meta-analysis Design

  - Systematic review of literature
  - Pooled results
  - Provides a precise estimate of treatment effect or diagnostic performance

Appendix B

Strength of Evidence Category Assessment Algorithm



continued on next page

Strength of Evidence Category Assessment Algorithm (continued)