Delfini Evidence Tool Kit
Evidence Grading, Wording Conclusions & Results Tables

Introduction to This Tool

This tool is intended as a supplement to critical appraisal checklists such as those available from Delfini and can help with tasks following appraisal such as—

- Grading the evidence including tips for identifying lethal flaws quickly
- Cautions regarding research findings generally and specific to efficacy and safety
- Reporting and assessing results
- Crafting conclusions

The tool provides advice and suggestions. Also included are tables that can be copied and pasted into individual study summary and critical appraisal checklists.

Key Points: Evidence Grading:
There are many evidence grading schemes available. When using any grading system, review the criteria used for each grade - these may vary even when the grade "name" is identical. Also, some grading schemes may assign misleading quality grades by inflating lower quality or invalid studies.

The Delfini grading system is designed to be easy to understand, easy to remember and flexible to apply. The concepts behind our grading system can be applied to individual studies or outcomes or conclusions from studies, systematic reviews, clinical recommendations, guidelines, etc. The scale is included in this document along with advice for evidence grading. General cautions about research are also included.

> **Timesaving Tip**: Tips about individual grades are included in order of efficiency of review, not in order of quality of grade. Utilizing Grade U tips first, for example, might help you identify easily studies that you can quickly eliminate from consideration.

Key Points: Results Reporting & Assessment
Effective critical appraisal requires assessing both validity and usefulness of results of valid studies. The table at the end of this document can be copied and pasted into validity assessments to perform the assessment of the usefulness of study results.

Key Points: Wording Conclusions
Many conclusions are misleading. This tool includes wording suggestions to help craft responsible and useful conclusion statements. A table is provided which can be copied and pasted into documents and then can be completed to provide concluding information for specific evidence reviews in a tabular format. Users may wish to add "cautions" to conclusion summaries.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Contents

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Delfini Validity & Usability Grading Scale for Summarizing the Evidence for Interventions

Due to complexities with studies of diagnostic tests, no recommendations for them are provided here.) All-or-none studies (observational) may be an exception and occur rarely

Grade of Usability	Strength of Evidence Advice Grades can be applied to individual studies, to conclusions within studies, a body of evidence or to secondary sources such as guidelines or clinical recommendations. General advice is provided below.
● **Grade A: Useful**	The evidence is strong and appears sufficient to use in making health care decisions—it is both valid and useful (e.g., meets standards for clinical significance, sufficient magnitude of effect size, physician and patient acceptability, etc.). Advice: Studies achieving this grade should be outstanding in design, methodology, execution and reporting and have successful study performance outcomes, providing useful information to aid clinical decision-making, enabling reasonable certitude in drawing conclusions. For a body of evidence: Several well-designed and conducted studies that consistently show similar results • For therapy, screening and prevention studies: RCTs. In some cases a single, large well-designed and conducted RCT may be sufficient; however, without confirmation from other studies results could be due to chance, undetected significant biases, fraud, etc. In such instance the study might receive a Grade A, but the Strength of the Evidence should include a cautionary note. • For natural history and prognosis: Cohort studies
⦿ **Grade B: Possibly Useful**	The evidence appears potentially strong and is probably sufficient to use in making health care decisions - some threats to validity were identified. Advice: Grade B studies should be very well designed and executed and meet most of the requirements that it takes to achieve a Grade A. Grade B evidence appears potentially strong and is probably sufficient to use in making health care decisions—some threats to validity have been identified. Studies achieving this grade should be of high quality and contain only non-lethal threats to validity and with sufficiently useful information to aid clinical decision-making, enabling reasonable certitude in drawing conclusions. Grade B is more frequent than Grade A, but is still a difficult grade to achieve For a body of evidence: The evidence is strong enough to conclude that the results are probably valid and useful (see above); however, study results from multiple studies are inconsistent or the studies may have some (but not lethal) threats to validity. • For therapy, screening and prevention studies: RCTs. In some cases a single, large well-designed and conducted RCT may be sufficient; however, without confirmation from other studies results could be due to chance, undetected significant biases, fraud, etc. In such instance the study might receive a Grade B, but the Strength of the Evidence should include a cautionary note.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

	▪ For natural history and prognosis: Cohort studies
O Grade B-U: Possible to uncertain usefulness	The evidence might be sufficient to use in making health care decisions; however, there remains sufficient uncertainty that the evidence cannot fully reach a Grade B and the uncertainty is not great enough to fully warrant a Grade U. Study quality is such that it appears likely that the evidence is sufficient to use in making health care decisions; however, there are some study issues that raise continued uncertainty. Health care decision-makers should be fully informed of the evidence quality.
○ **Grade U: Uncertain Validity and/or Usefulness**	There is sufficient uncertainty that caution is urged regarding its use in making health care decisions. • Uncertain Validity: This may be due to uncertain validity due to methodology (enough threats to validity to raise concern – our suggestion would be to **not** use such a study in most circumstances) or may be due to conflicting results. • Uncertain Usefulness: Or this may be due to uncertain applicability due to results (good methodology, but questions due to effect size, applicability of results when relating to biologic markers, or other issues). These latter studies may be useful and should be viewed in the context of the weight of the evidence. • Uncertainty of Author: If the author has reached a conclusion that the findings are uncertain, doing a critical appraisal is unlikely to result in a different conclusion. The evidence leaves us uncertain regardless of whether the study is valid or not. Critical appraisal is at the discretion of the reviewer. Advice: We end up assigning most studies a Grade U. As stated, we generally never use Grade U studies to inform efficacy decisions, but we will use Grade U evidence for safety, being very careful to describe that the evidence is of low quality

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Cautions

Critical Appraisal Matters: Exaggeration of Differences Between Outcomes in Intervention vs Control Groups (High Quality vs Low Quality RCTs): A Sampling of the Evidence

To see study citation and abstract, enter the PMID number in the PubMed search window.

Study Area of Concern	Relative Exaggeration	Reference (PMID)
Inadequate Generation of Randomization Sequence	17% to 75%	Juni (11440947), Kjaergard (11730399), Van Tulder (19770609)
Concealment of Allocation	14% to 73%	Schulz (7823387), Kjaergard (11730399), Moher (9746022), Juni (11440947)
Inadequate Double Blinding	4%% to 72%%	Schulz (7823387), Poolman (17332104), Kjaergard (11730399), Moher (9746022), Juni (11440947)
Loss of Data (Up to 38%)	2% to 35%	Vvan Tulder (19770609), Tierney (15561753), Nüesch (19736281), Canadian Orthopaedic Trauma Society (17200303)
Assessing outcomes through models	50% or greater	Lachin (11018568)

Cautions Regarding Research Findings Generally

- Endpoints should not be intermediate unless there is a body of evidence indicating a relationship between the endpoint and morbidity, mortality, symptom relief, emotional or physical functioning and health-related quality of life.
- Studies have shown that bias tends to favor the intervention, inflating benefits by a relative forty or fifty percent and sometimes higher. These high rates of falsely-inflated results have been found when study characteristics such as generation of the randomization sequence, concealment of allocation, blinding and assessing outcomes through statistical modeling are not utilized or are done incorrectly (references available from Delfini). Therefore, results may be distorted by threats to validity (see detailed critiques).
- Results in a single study that are not confirmed through other valid and useful studies may be due to chance or due to undiscovered or undetectable problems including lethal threats to validity, fraud, etc., and should be viewed with caution.

Cautions Regarding Efficacy Findings Generally

- Study biases frequently favor the intervention (and all studies have some kind of bias).
- The results of research within a research setting (efficacy) are usually better than the results we see in real practice (effectiveness).

Cautions Regarding Safety Findings Generally

Evaluating safety data is a complex process. Standards are often lower for using safety data than efficacy data and so there may be more uncertainty about the results. Therefore, conclusions about safety issues should be worded carefully so that information drawn from potentially flawed data regarding safety is not presented as if it is based on stronger evidence than actually exists. There are many cautionary tales about overzealous application of weak safety evidence that may, ultimately, have caused more harms to patients than if the agent had continued to be available.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.

© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

- Adverse events often occur infrequently, are often hard to find and usually not the topic of study.
- Systematic reviews of RCTs dealing with risks and harms should also be sought.
  - Keep in mind that risks and harms may be described in various ways in different studies.
  - There are potential limitations of RCTs and systematic reviews of RCTs that are not specifically focused on safety questions when the RCTs —
    - May not have reported or fully reported adverse events
    - May be of insufficient duration
    - May have relied upon small populations (eg, sampling error or power issues)
- Frequently, risks and harms are not prespecified as outcome measures in RCTs, increasing the likelihood of chance findings.
  - Advice is to look for patterns in multiple studies and to review whether several of the safety outcomes are biologically related or if there is a dose-response relationship, which lends support that these are true safety concerns and not a result of chance findings
  - Unless a study is powered for safety questions, lack of statistically significant differences may mean there is no difference or it may mean it is still unknown if there is a difference. Confidence intervals are useful in evaluating safety issues. For a valid study, the CI represents the range for which there is a 95% chance that the true answer lies. If the range includes a difference that is clinically meaningful, the study has not excluded the risk or harm.
    - Example: Authors report "The two groups did not differ in clinically relevant bleeding." PMID: 12049858. However, the CIs provide much more information: ARI, (95% CI) = 1.3, (0.3 to 2.9) and since the true difference in bleeding between the two groups could be as great as 2.9% (i.e., clinically relevant) the authors' conclusion is misleading.
- Authors sometimes utilize composite endpoints for efficacy, but single endpoints for safety endpoints. Consider looking at composites for risks and harms.
- It may be reasonable at times to use safety data from lower quality RCTs, because safety information from selected lower-grade RCTs may have greater validity and usefulness than observational studies or case reports.
- Risks and harms may not be detected until long after completion of RCTs through observational reports. Therefore, frequently long-term safety is unknown.
- It may be necessary to incorporate observational information into safety information if potentially significant risks are detected following the publication of an RCT. Such results should be regarded with caution and generally be considered "signals" as observations cannot demonstrate cause and effect and are highly prone to bias, confounding and chance.
- FDA post-marketing safety data may also be useful.
- As with efficacy data, safety data should be graded for quality and assigned an evidence grade. Although a study may be considered low quality overall, it may be sufficiently valid in one area, such as safety to include in a review; therefore, it may be worthwhile to grade individual study conclusions rather than to assign a grade to the overall study.
- Clinicians are urged to follow FDA recommendations.

See the **Delfini Searching Tool** for advice on search strategies for safety.

See **Conclusions & Wording Suggestions** > **Standard Safety Wording** for some wording tips.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.

© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Tips for Grading Primary Studies

These are general suggestions only. You need to apply judgment.

Grade U: Uncertain Usefulness

Grade U should be assigned when there is sufficient uncertainty about the accuracy of the estimates of effect resulting in an inability to comfortably draw conclusions from the research and in comfortably applying results. We end up assigning most studies a Grade U.

Lethal threats to validity or usability and other considerations warranting a Grade U generally include — and are not limited to — the following:

This checklist can help to quickly identify lethal flaws within a study. The trial could be considered invalid if any of the following exist:

1. Issues with study type
    a. Observational studies for efficacy of therapy, prevention, or screening interventions, unless the results are all-or-none results (standards are lowered for study quality when evaluating safety issues, but our advice is to take a net view and ensure that the wording of the conclusions is not misleading and that the strength of the evidence is described as being weak if that is the case)
    b. Case series (including reports using comparisons with historical controls or "natural statistics") unless the results are presented as all-or-none, which is extremely rare
    c. Claims of noninferiority or equivalence in a superiority trial
2. Methods that increase chance findings
    a. Use of post-hoc analyses (ie, using study outcomes, research questions or subgroups that are not determined in advance) to draw conclusions regarding cause and effect
3. Lack of meaningful clinical outcomes or other issues with outcomes
    a. For clinical questions, a lack of clinical significance (end points need to address direct and meaningful benefit with regard to morbidity, mortality, symptom relief, emotional or physical functioning, and/or health-related quality of life, or there needs to be other valid evidence that demonstrates a causal link between the study outcomes and a clinically significant outcome)
    b. Effect size is not clinically meaningful
    c. Nonsignificant findings are reported, but the confidence intervals include clinically meaningful differences
    d. For noninferiority and equivalence trials:
        i. Lack of sufficient evidence confirming efficacy of referent treatment
        ii. Inappropriate deltas (inferiority should be set at the smallest meaningful clinical benefit; equivalence should be set narrowly)
        iii. Significant biases or analysis methods that would tend to diminish an effect size (conservative application of intent-to-treat analysis, which would tend to diminish

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

differences between groups resulting in a bias towards equivalence or noninferiority, insufficient power, etc)

Grade B-U: Possible to Uncertain Usefulness

Grade B-U: We use Grade B-U when the study is not strong enough to warrant a B; but we are uncomfortable disregarding the study (Grade U).

Examples —

- Good study: Study may be well done, but many patients had to discontinue the study medication
- Not so good study: Flaws in the study prevent it from achieving a Grade B — however, it's not so flawed that it reaches Grade U.

Grade B-U Application:
Because of its use for clinical applications, B-U should be used conservatively. B-U is not a default grade. Rather, it should be used when the study is probably a B and the outcomes are highly likely to be true, but it doesn't quite comfortably reach a Grade B.

We use results of Grade B-U studies. Grade B-U means that we have greater uncertainty than if the study were Grade A or B — however, in contrast to grade U evidence we feel the evidence is strong enough to guide us clinically — and we would feel uncomfortable disregarding it.

Large discontinuation rates and statistical modeling are two frequent reasons for assignment of Grade B-U in otherwise well-designed and -executed studies.

Generally, studies with any of the issues below should not exceed a Grade B-U and may result in Grade U:

1. No blinding (open-label)
    a. No blinding with subjective measures or measures under the influence or control of those who are not blinded, and the outcomes for which could be influenced by lack of blinding may be most appropriately **Grade U**
2. Might attrition, including missing data, discontinuations or loss to follow-up, have resulted in distorted outcomes?

    Many researchers, biostatisticians and others struggle with this area—there appears to be no clear agreement in the clinical research community about how to best address these issues. There also is inconsistent evidence on the effects of attrition on study results. We, therefore, believe that studies should be evaluated on a case-by-case basis.

    The key question is, "Given that attrition has occurred, are the study results likely to be true?" It is important to look at the contextual elements of the study and reasons for discontinuation and loss-to-follow up and to look at what data is missing and why to assess likely impact on results. Attrition may or may not impact study outcomes depending, in part, upon the reasons for withdrawals, censoring rules and the resulting effects of applying those rules, for example.
3. Major problems such as too short study duration, inappropriate composite endpoints or use of composite endpoints for efficacy but not for safety, confounding, significant baseline differences, differences between groups besides element of interest, invalid measurement methods, lack of exposure (adherence issues) wholly inappropriate comparator or dosing or problematic combined endpoints. At times studies may be problematic due to lack of data comparing a new agent or intervention to placebo. For example in a study comparing a new COX-2 agent to an older NSAID, the newer COX-2 agent had a higher event rate for

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

thrombotic events and the authors claimed that the new agent did not increase thrombotic risk. They argued that the older agent was protective. Placebo information would have been useful in this situation.

4. Small sample size
5. Authorial uncertainty: Author has reached a conclusion that findings are uncertain and doing a critical appraisal is most likely not going to result in a different conclusion. The evidence leaves us uncertain regardless of whether the study is valid or not. Critical appraisal is at the discretion of the reviewer. These may be most appropriately assigned **Grade U.**

Review the Delfini Study Validity and Usability Tool (long or short version) for a list of critical appraisal considerations.

Grade A: Useful

Grade A should be rarely assigned to any study. ("Extra points" are not given for challenge or difficulty in answering the question. Authors should not be given extra points by second-guessing them.)

A study should not achieve a Grade A, at a minimum, if the following conditions have not been met:

1. Excellence in both design and execution along with clarity and completeness in reporting
2. Medium to large "n"
3. Representative and appropriate subjects
4. Details of randomization and concealment are adequately described and are valid methods
5. The subjects and all those working with the subjects are blinded, including assessors
6. Bias, confounding or chance have effectively been ruled out as possible explanations for study findings
7. Study subjects and controls have been treated identically excepting for the element of interest
8. Appropriate intervention (e.g., proper dosing) and appropriate comparator (e.g., including placebo or based on prior studies that demonstrate effectiveness against placebo)
9. Adequate duration
10. Measurement is objective
11. ITT analysis has been conducted (meaning analysis included all randomized patients and analyzes their outcomes in the groups to which they were assigned) plus puts the burden of proof on the intervention through the selected method of imputation of outcomes for missing data points
12. For clinical questions, meaningful benefit to patients in clinically significant areas

Grade B: Possibly Useful

Grade B is more frequent than Grade A, but is still a difficult grade to achieve. Grade B studies should be very well designed and executed and meet most of the requirements that it takes to achieve a Grade A.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

GRADE Grading System

GRADE (Grading of Recommendations Assessment, Development and Evaluation) is a system for summarizing and rating the quality of evidence and grading strength of recommendations in systematic reviews, health technology assessments (HTAs) and clinical practice guidelines.[1] There has been great interest in GRADE for several years and, because it is frequently referred to, we are summarizing the key points about how GRADE is applied to grade the quality of evidence and the strength of recommendations.

Following critical appraisal of the evidence, it is possible to apply the GRADE criteria to determine the rating of the evidence for an outcome across all studies based on study design, risk of bias, precision, consistency, directness and magnitude of effect.

Rating of the Evidence
Randomized controlled trials (RCTs) enter the system as high-quality evidence, and observational studies enter the system as low-quality evidence. Five factors (study limitations, imprecision, inconsistency of results, indirectness of results and publication bias) may lead to a rating down of the quality of evidence, and three factors (large magnitude of effect, dose response and confounder) which are likely minimize the effect size may lead to an uprating. Ultimately, the quality of evidence for each outcome falls into one of four categories from high to very low (high, moderate, low and very low) as shown in the table below.

GRADE distinguishes between quality assessment conducted as part of a systematic review and that undertaken in the process of guideline development. For systematic reviews, the rating of the quality of evidence reflects the degree of confidence that the estimates of the effect are correct. For guideline recommendations, the quality ratings reflect the degree of confidence that the estimates of effect are adequate to support a particular decision or recommendation.

Quality of Evidence Levels Definitions

- High: Very confident that the true effect lies close to the estimate of effect
- Moderate: Moderate confidence that the true effect lies close to the estimate of effect
- Low: Limited confidence that the true effect lies close to the estimate of effect
- Very Low: Very little confidence in the estimate of effect

GRADE Quality of Evidence Assessment Process and Rating

<table>
<tr><th>Study Design <u>At Entry</u> Into GRADE System</th><th>Quality of Evidence on Entry</th><th>Lower Category If…</th><th>Higher Category If…</th><th colspan="2"><u>Final</u> Quality of Evidence Rating (Select One)</th></tr>
<tr><td>RCT</td><td>HIGH</td><td rowspan="4">Risk of Bias
-1 Serious
-2 Very Serious
Inconsistency
-1 Serious
-2 Very Serious
Indirectness
-1 Serious</td><td rowspan="4">Effect Size
+1 Large
+2 Very Large
Dose Response
+1
All plausible confounders would reduce a</td><td>HIGH
++++</td><td></td></tr>
<tr><td rowspan="2">Observational Study</td><td rowspan="2">LOW</td><td>MODERATE
+++0</td><td></td></tr>
<tr><td>LOW
++00</td><td></td></tr>
<tr><td colspan="2"></td><td>VERY LOW
+000</td><td></td></tr>
</table>

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

	-2 Very Serious **Imprecision** -1 Serious -2 Very Serious **Publication Bias** -1 Serious -2 Very Serious	**demonstrated effect** +1 All plausible confounders would suggest a spurious effect when the results show no effect +1	

GRADE recommends creating a table or tables for outcomes that addresses the number of studies, design, limitations, inconsistency, indirectness, imprecision, publication bias, summary of findings, relative risk, absolute risk and a quality score of high to very low.

Some definitions:

- Consistency refers to the degree of similarity of effect sizes of included studies.
- Directness is the linkage between the intervention and health outcomes (e.g., some intermediate or surrogate outcome measures are strongly linked to health outcomes and some are not).
- Precision concerns the ability to draw a clinically useful conclusion from the confidence intervals. An imprecise estimate, for example, is one for which the confidence interval is wide enough to include clinically distinct conclusions (e.g., favoring both the interventions being compared).

Grading of Recommendations

Recommendations are tagged as strong or weak (alternative terms: conditional or discretionary) according to the quality of the supporting evidence and the balance between desirable and undesirable consequences of the alternative management options. Values, preferences and resource use are also considered in determining whether a particular recommendation will be strong or weak. GRADE recommends making strong recommendations when there is confidence that the desirable effects of adherence to a recommendation outweigh the undesirable effects. Weak recommendations indicate that the desirable effects of adherence to a recommendation probably outweigh the undesirable effects, but there is less confidence. The following are categories that have been described by GRADE [2]:

- Strong recommendation for using an intervention
- Weak recommendation for using an intervention
- Weak recommendation against using an intervention
- Strong recommendation against using an intervention

References

1. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr;64(4):401-6. Epub 2011 Jan 5. PubMed PMID: 21208779.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

2. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008 Apr 26;336(7650):924-6. PubMed PMID: 18436948

AHRQ-EHCP Grading System

Most groups who systematically review the medical literature have an approach to grading the quality of individual studies and also rating the quality for specific outcomes after evaluating the totality of evidence.

The AHRQ EHCP (the Agency for Healthcare Research and Quality and the Effective Health Care Program group grading methodology is worth knowing about.[1] AHRQ EHCP evaluates four domains when rating the overall strength of the evidence (SOE). These domains were selected after reviewing grading methodologies used by the U.S. Preventive Services Task Force (USPSF),[2] the GRADE working group,[3] and other evidence-based practice centers.[4,5]

Briefly, The AHRQ EHCP approach assesses the risk of bias, consistency, directness and precision for each outcome or comparison of interest after rating each study or key outcome from each study for bias (paraphrased, in some instances, below):

- Bias: each study is scored based on study design and methodology, and the aggregate of studies is rated for an overall "risk of bias" score. Aggregate risk of bias is scored as low, medium or high. The aggregate quality of studies is rated as good, fair or poor.
- Consistency (the degree of similarity of effect sizes of included studies) is scored as consistent, inconsistent or unknown/not applicable.
- Directness is the linkage between the intervention and health outcomes scored as direct or indirect (meaning intermediate or surrogate outcome measures which may or may not be valid measures for clinical usefulness).

[1] Owens DK, Lohr KN, Atkins D, et al. Grading the strength of a body of evidence when comparing medical interventions-Agency for Healthcare Research and Quality and the Effective Health Care Program. *J Clin Epidemiol*. 2009 Jul 10.

[2] Sawaya GF, Guirguis-Blake J, LeFevre M, Harris R, Petitti D; U.S. Preventive Services Task Force. Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Ann Intern Med*. 2007 Dec 18;147:871–5.

[3] Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924e6.

[4] West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/ Technology Assessment No. 47 (Prepared by the Research Triangle Institute- University of North Carolina Evidence-based Practice Center under Contract No. 290–97–0011). AHRQ Publication No. 02–E016. Rockville, MD: *Agency for Healthcare Research and Quality*; 2002.

[5] Treadwell JR, Tregear SJ, Reston JT, Turkelson CM. A system for rating the stability and strength of medical evidence. BMC Med Res Methodol 2006;6:52.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.

© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

- Precision concerns the ability to draw a clinically useful conclusion from the confidence intervals. An imprecise estimate, for example, is one for which the confidence interval is wide enough to include clinically distinct conclusions (e.g., favoring both the interventions being compared).

AHRQ EHCP—like GRADE—has additional domains which may be included when rating evidence:

- Dose-response associations (present or not present)
- Plausible confounding that would increase or decrease effect (present or absent)
- Strength of association (magnitude of effect)
- Publication bias (not necessary to formally score)

The AHRQ EHCP overall strength of evidence (SOE) for each outcome of interest includes three grades—high, moderate and inconclusive. For example, if the SOE is high, further research is unlikely to change confidence in the estimate of effect. If evidence is unavailable or does not permit a conclusion, the outcome in the AHRQ EHCP system is graded as inconclusive.

Delfini Modification

We use our usual grading system for individual studies (A, B, BU, U). When we use our modified AHRQ EHCP system for rating the overall level of evidence (LOE), we modify it by adding a fourth category—"borderline" to increase clarity as we believe that "moderate" is not precise enough to address evidence of borderline usefulness.

AHRQ EHCP and Delfini Evidence Grading Methodologies

AHRQ EHCP Evidence Grading and Strength of Evidence Methodology	Delfini Evidence Grading and Level of Evidence Methodology
For Each Outcome • Bias: Each study rated and an aggregate risk of bias level is selected from low/ medium/ high risk. The aggregate quality of the studies under consideration is rated as good/fair/poor • Consistency • Directness • Precision	Each study or outcome: A, B, BU, U for validity and usefulness
Overall SOE: high, moderate, low, insufficient	Overall LOE: high, moderate, borderline, inconclusive

Examples AHRQ EHCP Strength of Evidence (SOE) Ratings

Number studies; N	Risk of Bias	Consistency	Directness	Precision	SOE
Mortality					
1;80	RCT/Medium	Unknown	Direct	Imprecise	Insufficient
Improved Quality of Life					
6;265	RCTs/Low	Consistent	Direct	Precise	High SOE

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Conclusions & Wording Suggestions

The table below can be used to construct a concluding statement about the evidence:

Evidence Grade & Concluding Statement **Choose the applicable grade and strength of evidence statement (examples found in Results Table below).**	
Grade: **Evidence Statement:**	
Wording Conclusion Sample for Valid and Clinically Useful Studies We recommend **not** completing for Grade U studies for questions of efficacy. We recommend completing selectively for safety when safety outcomes are judged reasonably likely not to be due to chance.	
Consideration	**Research Outcomes** **Replace text and complete the consideration below, adding new tables as needed for different outcomes:**
Category/purpose of study	Efficacy, safety or other
Intervention	Include details such as dosing, means of administration, formulation, duration, etc.
Was demonstrated to be	Superior, equivalent, non-inferior, other [specify]
When compared with	Comparator
In the following clinically significant area(s)	Indication and endpoints
As measured by	Measurement instruments used
Within the following time period	Study duration (intervention to final outcomes measurements)
Study population	Describe patient population (inclusion and exclusion criteria and key baseline characteristics including demographic details such as ethnicity, gender and age and meaningful prognostic characteristics such as relevant medical histories), based on number of people studied, cared for in what setting and by what kind of provider
Results: Risk with and without treatment	Two x two table data
Results: Absolute measures	Formatted as: ARR%, 95% CI (CI to CI%), P=0.XXX.
Results: Number-needed-to Treat	Complete only for statistically significant results
Additional Considerations for Safety*	
Study type	Experiment, observational [specify exact type e.g., RCT, cohort, case report]
Safety issues were	Prespecified, not prespecified, other [specify]
Results of important safety issues reported as statistically significant, plus important safety issues that are reported as nonsignificant, but are actually "inconclusive" because of a potentially clinically relevant difference within the CI	Harms and results including confidence interval
Confirmation in additional studies	Yes, no [include references]

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Additional Considerations for Safety*
Consistency/Heterogeneity:
Limitations:
Other:
References:

* Although some safety issues may not be considered significant in studies with short-term follow-up, they may be determined to be clinically significant in studies with longer follow-up. Longer-term studies may also demonstrate additional safety issues not observed in shorter-term trials. Most long-term safety issues associated with newly approved drugs are unknown until after the drugs are marketed and patients experience adverse events. Clinicians are urged to adhere to FDA recommendations for therapy.

Evidence Tagging Options
(Be clear about whether you are referring to an overall level of evidence (LOE), a single study, a specific outcome, etc.)
Examples

- Sufficient evidence: Example "The evidence is sufficient to conclude that… (LOE: Moderate)"
- Insufficient evidence: Example "There is insufficient evidence to conclude that… (LOE: Inconclusive)"
- Inconclusive evidence (which can result from a high quality study)
    - Example: Based on this BU study of 670 subjects*, the evidence for a difference in bleeding rates between warfarin and enoxaparin is inconclusive based on a consideration of the rates and 95% CIs of study patients undergoing total hip replacement surgery; review of confidence intervals indicates that the difference in bleeding rates could have been as great as 2.4% favoring warfarin or up to 1.8% favoring enoxaprin. Reported point estimates for rates of major bleeding were 1.8% for warfarin patients vs 2.1% for enoxaparin patients, ARR= -0.3 favoring warfarin (95% CI -2.4 to 1.8)
- Individual study risk of bias ratings
    - high risk of bias
    - medium risk of bias
    - low risk of bias
- Overall LOE ratings
    - High
    - Moderate
    - Borderline
    - Inconclusive

Useful Phrases

- Flawed evidence
- Valid evidence
- Reliable evidence
- Evidence is conclusive
- Evidence is inconclusive
- Lethally threatened evidence
- Valid and clinically useful evidence
- Clinically meaningful - not clinically meaningful
- Meaningful clinical benefit

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.

© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Conclusions: Wording We Recommend be Avoided
We recommend avoiding language that is inaccurate, misleading or vague. Here are some examples of problematic language we often see.

Use of cause and effect language for observational studies or using language which is otherwise misleading or inaccurate --

1. "There is evidence that…"
2. "Recent evidence has shown that…" (as contrasted with, "Studies have reported that…")
3. "Studies have shown…"
4. "There is some evidence that…"
5. "Many physicians have found…"
6. "It appears that…"
7. "It may be that..."
8. "Research has shown that…"
9. "X has shown promise (in several studies)…"
10. "Overall there was no statistically significant difference between the groups, but a trend was noted…"
11. For safety issues, when results are not statistically significant, "There were no differences in adverse effects between groups…"

Use of misleading language for harms when there is no difference in harms between groups—
Unless a study is powered for harms, lack of statistically significant differences may mean there is no difference or it may mean it is still unknown if there is a difference. Confidence intervals are useful in evaluating harms. For a valid study, the CI represents the range for which there is a 95% chance that the true answer lies. If the range includes a difference that is clinically meaningful, the study has not excluded the harm.

> **Example**: Authors report: "The two groups did not differ in clinically relevant bleeding." PMID: 12049858. However, the CIs provide much more information: ARI, (95% CI) = 1.3, (0.3 to 2.9) and since the true difference in bleeding between the two groups could be as great as 2.9% (i.e., clinically relevant) the authors' conclusion is misleading.

Example of Wording
The evidence is insufficient to conclude that, in high risk women, the addition of MRI to mammographic screening reduces the need for mammography or ultrasound. (LOE: Inconclusive). Adding MRI will change treatment plans and result in more extensive surgery for some women (LOE: Borderline), but may not change incomplete excision rates or breast cancer recurrence rates (LOE: Inconclusive). We found no evidence that adding MRI to conventional screening in women at high risk of breast cancer will reduce mortality rates (LOE: Inconclusive).

Conclusions from Non-Significant Findings
Be cautious about making conclusions of no difference if no difference is potentially due to lack of finding a difference.

> **Examples of Wording**
> Significant differences between the groups were not detected. This could be due to an insufficient number of study subjects. Evaluating confidence intervals may provide more precision regarding the differences between the groups.
> OR

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.

© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Study has not confirmed that xxxx.

Other Wording Suggestions — Miscellany

ITT Analysis Verification

- True ITT analysis was not performed. Therefore, the reviewers recalculated the statistical significance, applying Fisher's exact test and based on numbers of patients or percentages reported, of all reported outcomes listed above to ensure they were statistically significant. The reviewers' goal was only to verify statistical significance.

  Results reported above are the authors' results which may be more favorable than they would be with true ITT analysis. Reviewers **did/did not** find a statistically significant result for _____ when missing values were assigned as [failed or success for which group].

Standard Safety Wording

Note
Delfini takes a conservative patient-centered approach. Because safety is difficult to assess and may never be fully understood even over time, our SOE conclusions for safety are almost always Borderline or Insufficient. We may conclude Sufficient Evidence for a specific safety outcome in an instance in which there is definitive causal information about the occurrence of a harm**. Standard Cautions** almost always apply.

Standard Cautions
Safety is often difficult to assess. Safety can only potentially be established with long-term follow-up.

Patients should be informed about known safety issues and the quality of the safety evidence even when the evidence is weak. Patients should also be informed that there may be unknown risks of adverse events from healthcare interventions.

Reports of no differences between groups should be viewed with caution because the population studied may have been too small for a true difference to be revealed. However, reports of adverse events might not, in fact, be due to the intervention.

Overall Value
Requires local comparison of current practice and cost and other value judgments to projected impacts of practice change.

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.

Results Table

Copy and paste the table below into your critiques.

Outcome	Study N (%)	Comparator N (%)	RRR/I, 95% CI (X to Y), P=	ARR/I 95% CI (X to Y), P=	NNT/H	Time Period	Clinically meaningful difference?
Outcome	Study N (%)	Comparator N (%)	RRR/I 95% CI (X to Y), P=	ARR/I 95% CI (X to Y), P=	NNT/H	Time Period	Clinically meaningful difference?
Outcome	Study N (%)	Comparator N (%)	RRR/I 95% CI (X to Y), P=	ARR/I 95% CI (X to Y), P=	NNT/H	Time Period	Clinically meaningful difference?

Use of this tool implies agreement to the legal terms and conditions at www.delfini.org.
© 2002-2015 Delfini Group, LLC. All Rights Reserved Worldwide.