Getting Started with Critical Appraisals of Medical Literature

It Is Easier than You Think

by Sheri Ann Strite, MS and Michael E. Stuart, MD, Delfini Group, LLC

It is easier to get started critically appraising studies than many people believe. While it is true that many critical appraisal issues can be complex and that experience over time improves skills, it is also true that a few basic skills can make a meaningful difference between being misled by a low quality study and being able to identify sufficient threats to realize a study is not valid or clinically useful.

First, let’s focus on the goal: The purpose of using science in medicine is to help reduce medical uncertainty and increase predictability including the likelihood of benefit for which patients under what circumstances. With that in mind, let us look at some key myths that get in the way of critically reading research studies for validity and clinical usefulness.

Myth #1: I will get valid and clinically useful information if I restrict my reading to the right journals.

Not true. We estimate that — overall — fewer than 10 percent of studies published each week are valid and clinically useful. No journal seems to be immune. For example, one study found that 18-68 percent of the information in abstracts in six top-tier journals could not be verified in the body of the text.

Myth #2: I can’t do critical appraisal because I need to know a lot about statistics.

Not true. Usually the biggest bang for the buck is in identifying biases. Statistician colleagues of ours have agreed that bias trapping is, in many ways, most important for identifying high quality studies. Further, many statistical analyses are based on models and models are not “truth.” When we see that models have been used, we automatically add that fact to our “count” of potential...
Myth #3: It’s too hard to learn.
Do you understand the concept of the blind taste test in cooking?

Myth #3: It’s too hard to learn.
Do you understand the concept of the blind taste test in cooking? Many important issues in critical appraisal are as simple to understand. The real issue is that the majority of people are not aware that they need to read all published studies with a healthy skepticism because they are not aware that most published studies are so flawed that drawing conclusions about efficacy should not be done. As we’ve described previously (see Box 1. Critical Appraisal Matters), we have seen many people learn to critically appraisal studies in less than a day. We, ourselves, are self-taught. There are many wonderful resources available including those we make freely available on our website (www.delfini.org). In addition to the tools and tutorials available there, we have a Recommended Reading list. See Box 2. Critical Appraisal Resources for a few of our favorites.

Myth #4: It’s too much work.
It is a waste of time and inappropriate to focus on the results from lethally flawed studies. The easiest way to perform critical appraisal work is to become familiar with lethally flawed studies. In many cases, you will find one lethal threat, and you are done. If you find a sufficient number of threats and, therefore, feel uncertain about the conclusions of the study, you can discard that study. We might use such a study for safety (with cautious wording since the evidence supporting safety is often weak), but we would not use such a study to draw cause and effect conclusions for efficacy. Frequently, by hunting for key flaws in a study, we can reach a conclusion that the study is of uncertain validity and/or clinical usefulness in a few minutes (which is less time than it takes to read an article).

Here are some key considerations to help you be most efficient while applying critical appraisal consideration.

Should I bother with this article?
- If the results are true, will it change practice? (Ensure the endpoints and analysis groups were determined a priori, meaning ensure they were determined in advance of the start of the study.)
- Will addressing the important outcomes result in clinically meaningful events likely to benefit patients?
Meaningful benefit = size of benefit + areas of clinical significance to patients (morbidity, mortality, symptom relief, physical and emotional functioning and health-related quality of life).

Box 1. Critical Appraisal Matters
Critical appraisal matters. In a classic study, Chalmers et al demonstrated that a lack of concealment of allocation inflated the appearance of benefit, as did a lack of randomization, in studies with an outcome of mortality from acute myocardial infarction (MI). Studies that were randomized with concealment of the allocation sequence, on average, reported nonsignificant findings. As studies diminished in quality in these two dimensions, benefit was inflated as high as an absolute 10% in favor of the intervention. This translates into a number-needed-to-treat (NNT) of 10, which would be highly clinically significant if only it were true. Results of subsequently-published studies have supported the conclusion that bias tends to favor the intervention, inflating benefits by up to a relative 40% to 50%. These high rates of falsely inflated results have been demonstrated in studies in which methods such as generation of the randomization sequence, concealment of allocation, blinding, and assessing outcomes through statistical modeling are omitted or not done correctly.11–15


Box 2. Critical Appraisal Resources
Tools and tutorials freely available at www.delfini.org including Recommended Reading list. There are many textbooks available on the topics of clinical epidemiology and critical appraisal techniques. Here are some of our favorites:
The purpose of using science in medicine is to help reduce medical uncertainty and increase predictability.

Table 1. Initial checklist to help identify lethal flaws in published clinical trials of therapeutic interventions

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 treatment, 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
2. Lack of meaningful clinical benefit 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. Non-significant findings are reported, but the confidence intervals include clinically meaningful differences, which would result in a lethally threatened conclusion
   d. For non-inferiority and equivalence trials
      1. Lack of sufficient evidence confirming efficacy of referent treatment
      2. Inappropriate deltas (inferiority should be set at the smallest meaningful clinical benefit, equivalence should be set narrowly)
      3. 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 differences between groups resulting in a bias towards equivalence or non-inferiority, insufficient power, etc.)
3. Methods that increase chance findings
   a. Use of post hoc analyses (i.e., using study outcomes or research questions that are not determined in advance) to draw conclusions regarding cause and effect
   b. Subgroup analyses in which the subgroups are not determined in advance

Adapted from the Delfini Intervention Review Template. © Delfini Group, LLC. 2005-2008

- Example: If treatment is involved, will treating the problem result in clinically meaningful outcomes?
- How did they define success/failure, improvement/no improvement?
  Do you agree?
- Look at the boundaries of the confidence intervals (CIs) of valid studies to compare to your requirements for meeting clinical significance. See * below.

Is the study designed in a way that it can answer my clinical question?

- What is the study attempting to answer?
- Is its aim meaningful and appropriate?
- Is the study structured in a way that it can answer this question?

- Was choice involved in determining who got the therapy? If yes, this is an observational study and there is a high risk of being misleading when determining the efficacy of interventions.
- Was the method of determining the outcome measure reasonable?

How do I know the results are likely to be true?

- What could possibly explain the results other than the intervention? Could bias, confounding, or chance explain or affect the outcomes? Or can I rule them out and comfortably presume cause and effect?
- The burden of proof is on the intervention. Could anything advantage the intervention?
- Read your paper as a “clinician” AND as a validity detective! Bias tends to favor the intervention. Biases in areas such as randomization, concealment of allocation, blinding or determining results using models can inflate benefits by up to a relative 50 percent. Do not neglect these important considerations.
- Any difference between groups other than the area of investigation is an automatic bias. Frequently, lack of information on co-interventions or adherence will be suspect due to possible differences between groups.
- Missing data points are a big problem with many studies. Even non-differential loss can mean differences in prognostic variables. Take a quick look for ITT analysis and be sure to evaluate the imputation methods. Most analyses involve modeling and modeling requires unverifiable assumptions.
- For valid studies, consider what you judge to be a reasonable range for clinical significance – this need not be hard and fast. For statistically significant findings, is the confidence interval wholly within bounds for clinical significance? For non-significant findings, is the confidence interval wholly beneath your limit for clinical...
significance? A yes to these two questions means likely conclusive findings for valid studies. No, means findings are inconclusive.

How can I appraise an article quickly?

- Always use a checklist. Develop familiarity with the concepts of critical appraisal. With practice, you will swiftly develop ways to discard fatally flawed and clinically useless studies quickly. See checklists in Tables 1 & 2 for tips.


About the Authors

Delfini Group specializes in applied, practical evidence- and value-based approaches to clinical improvement and medical decision-making.

The Delfini Group consists of Michael Stuart MD and Sheri Ann Strite, MS who combine academic experience with decades of practical experience in health care, clinical quality health systems improvements, research, training and leadership.

They are internationally respected designers of evidence-based health care improvement methods and products, consultants, authors, scientific reviewers of medical technologies, clinical improvement project facilitators, clinical guideline facilitators, medical content developers and trainers with special expertise in critical appraisal of the medical literature.

To learn more about their expertise and medical literature review, visit www.delfini.org.

References


Table 2: Secondary checklist to help identify selected flaws within clinical trials of therapeutic interventions

This checklist can help to quickly identify selected flaws within a study. Consider the following when assessing the validity of trial results that were not deemed to be invalid using the initial validity checklist:

1. Was the study design appropriate to assess the research question? Was the research question useful?

2. If  composite endpoints were used, were they reasonable? And if  composite efficacy endpoints were used, were reasonable composite safety endpoints also used?

3. Could bias, confounders (known or unknown), or chance explain the study results?

4. Were the research questions, study population, outcomes, group assignment methods, study conduct methods, analysis methods, and measure of statistical significance prespecified and appropriate?

5. Were the groups included in the study appropriate, of appropriate size, concurrent, and similar in prognostic factors?

6. Were the methods for generating the group assignment sequence truly random? Did the sequencing avoid potential for anyone to affect the assignment to a study arm? Did the randomization remain intact throughout the study?

7. Were the allocation strategies concealed to prevent anyone from affecting assignment to a study arm?

8. Was the double-blindig adequately preserved throughout the study for the patients and all who worked with the patients or patients’ data?

9. Were reasonable interventions and reasonable comparator(s) used?

10. Was the study free of bias or differences between the groups (except the topic being studied, e.g., active agent vs placebo)? Considerations include intervention design and execution, co-interventions, concomitant medication use, adherence, inappropriate exposure or migration, crossover threats, protocol deviations, measurement methods, study duration, etc.

11. Were any missing data points or loss from randomization minimal or nonexistent, except in cases where an appropriate intention-to-treat analysis or other reasonable sensitivity analysis has been performed (e.g., a loss of approximately 5% of outcome data with a differential loss between the study groups, or approximately 10% without a differential loss)?

12. Were the research assessors blinded?

13. Was there a low likelihood that results were due to chance or that non-significant results were due to an insufficient number of study participants? This analysis requires a judgment call on statistical and clinical significance, including a review of the confidence intervals to determine whether the boundaries are outside of what you consider to be clinically significant. If the results are outside those boundaries, the findings should be considered inconclusive.

14. Did any use of modeling include only reasonable assumptions?

The phrase “intention-to-treat” (ITT) is often used by authors who have not actually performed an ITT analysis or who have performed it incorrectly. Evaluate whether outcomes are provided for all patients in the groups to which they were randomized and evaluate the method the researchers used to assign outcomes for missing values. You may be able to perform an ITT analysis by assigning outcomes for missing values that put the intervention or element of interest through a rigorous test. For example, you could test the P values by performing a worst-case scenario (also known as extreme case analysis). To do so, assign positive outcomes to missing controls and negative outcomes to missing study patients. If the results remain statistically significant and the other methodological considerations are acceptable, the study can be considered to have passed the worst-case analysis audit.

Adapted from the Evidence Grading Tool. © Delfini Group, LLC 2005–2008