An Introduction to Systematic Reviews

The Lifeblood of Medical Decision-Making

By Guest Editors: Sheri Ann Strite and Michael E. Stuart, MD, Delfini Group, LLC

This is our second editorial in California Pharmacist about evidence-based approaches to evaluating the medical literature. Our first editorial outlined how pharmacy student writers for the California Pharmacist Evidence-based Medicine (EBM) reviews approach critical appraisal of primary studies. In this editorial, we discuss secondary studies, which represent a "round-up" of primary studies into an evidence synthesis to summarize what is known about a healthcare issue to help inform medical decision-making.

First a reminder that EBM can be defined as "a set of principles and methods intended to ensure that — to the greatest extent possible — medical decisions, guidelines, and other types of policies are based on and consistent with good evidence of effectiveness and benefit."[1] The reason for basing decisions on good medical evidence is that evidence-based decisions made by clinicians, patients and others are more likely to result in predictable outcomes than if decisions are based on clinical experience or other factors. Unfortunately, it is not widely known that most of the published medical literature and, thus, the evidence for effectiveness, is seriously flawed.

We consider the hallmarks of evidence-based practice to be these:

1) When seeking information on a topic, a systematic search is conducted for science and science-based information using evidence-based searching and filtering techniques.

2) All sources of information to guide medical decision-making are critically appraised, using science-based principles, for validity and usefulness.

3) Any conclusions drawn from the science are carefully crafted to be as valid as possible.

4) Methods used and reporting are transparent so that the work can be evaluated for quality, replicated and updated.

5) Clinical information sources are updated when significant new information becomes available and such information is periodically sought.

Reviews can be useful for background reading, but should not be relied upon to inform medical decision-making as they may present misleading information about efficacy of interventions and can potentially add significantly to the cost and quality problems we face in health care — and usually leaning toward overuse of interventions.

ABOUT SYSTEMATIC REVIEWS

What Is a Systematic Review?

A systematic review is a secondary study — meaning it utilizes original research (primary studies) to summarize results of more than one study in an attempt to establish a body of scientific evidence to answer a research question — and it does so conforming to an established formal methodology. Meta-analyses are a subtype of systematic reviews, the end product for which is a statistically computed outcome, either through combining study results or pooling raw study data, whereas a systematic review which is not a meta-analysis may be qualitative or may aggregate outcomes in a non-statistical way. If prepared using formal methods for systematic reviews, a monograph or a drug class review can be considered a type of systematic review.

The medical literature, however, is replete with another type of secondary study which is called a "review," sometimes referred to as an "overview" or a "narrative review." Reviews are problematic because they lack some or all of the necessary components of systematic reviews; thereby, presenting opportunities for bias. Reviews tend to be written by "experts." However, studies have shown that in many specialty areas, experts differ significantly in estimates of benefit from various interventions, begging the questions: who is actually expert, and how can one tell?[2, 3] Reviews can be useful for background reading, but should not be relied upon to inform medical decision-making as they may present misleading information about efficacy of interventions and can potentially add significantly to the cost and quality problems we face in health care — and usually leaning toward overuse of interventions.

Table 1. Features of Quality Systematic Reviews

Item	Feature	Remarks
1	**Research Question** Objectives of the review are determined in advance of doing the review and utilize clearly stated and clinically meaningful questions.	**A priori Questions** "In advance" is referred to as **a priori**. Sound scientific methodology frequently requires that certain items (eg, research question, outcomes of interest, populations for analysis, etc.) are determined **a priori** because such an approach helps reduce opportunities for bias and the likelihood of chance findings. **Question Focus** It is important to evaluate whether the research objective will capture the right information for such considerations as population, condition, intervention or exposure and important outcomes. For interventions, focus must be clinically meaningful when considering efficacy. **Clinically Meaningful Questions** Clinically meaningful questions are those that address outcomes of importance to patients specifically in the areas of morbidity, mortality, symptom relief, emotional and/or physical functioning and/or health-related quality of life.
2	**Study Selection** Explicit, documented and appropriate selection criteria (inclusion and/or exclusion) are chosen in advance for included studies that are sufficiently similar to each other (homogenous).	**A priori Criteria** Selection criteria must be determined **a priori** to help reduce opportunities for a biased approach. **Criteria Transparency** At a minimum, criteria should specify study type (eg, randomized controlled trial (RCT), cross-sectional, cohort, etc.), population, methods, inventions or exposures and outcomes **Criteria for Study Type** Study type needs to be appropriate to the clinical question. For example, if this is a question of therapy, screening or prevention and observational studies are used to answer questions of efficacy, Delfini suggests not using the review because of the high potential for being misled by the results. (See **Table 2** for help in matching study types to clinical questions.) **Criteria for Homogeneity** Sufficiently similar means similar in methods, population, intervention or exposures or characteristics, follow-up period, outcomes, etc. **Criteria for Patient Population** Is the population appropriate for this question?
3	**Search** A comprehensive and systematic search is employed to identify all valid and clinically useful scientific information that should be included in the summary outcomes.	**Search Strategy Considerations** Consider whether the search strategy appears to be well-thought out and reasonable to answer the research question. Is it documented? Is it comprehensive? You should be able to replicate the search which is important for your ability to assess the quality of the systematic review and important for your ability to update it. **Strategy** Needs to include search terms, sources, filters used and dates covered Needs to include a search from the National Library of Medicine Textbooks are generally not considered to have relevant scientific information. If they are used, this may be an indicator of a poor quality systematic review.
4	**Critical Appraisal of Included Studies** Studies considered for inclusion in the systematic review are critically appraised for validity and clinical usefulness using a sound and rigorous methodology. Only valid and clinically useful studies are utilized for drawing conclusions of efficacy.	**Critical Appraisal Quality Assessment** What is the quality of included studies? Did the authors use an explicit and quality method for determining validity of individual studies? Is there more than one author appraising the studies? This is a strong preference as much judgment is required in critical appraisal. And, if yes, how were disagreements resolved? *NOTE:* The Jadad Scale is frequently employed by reviewers for determining study quality. The Jadad Scale is referred to as a "validated" scoring system; however, it is **not** a good measure of study quality. If the Jadad Scale is used, is there some assurance that the reviewers went beyond the Jadad Scale criteria to critically examine the studies so that only valid and clinically useful studies are used to draw conclusions about efficacy, for example? We highly recommend that you perform an audit even of the most "trusted" sources or ones with the best reputations as quality varies widely. See the section: **Steps and Tips for Working with Secondary Studies and Secondary Sources.**

continued

Table 1. Features of Quality Systematic Reviews

Item	Feature	Remarks
	Missing Data Reasonable choices are made by reviewers for addressing missing data.	**Missing Data Considerations** Consider how loss to follow-up is handled and whether it done appropriately. Losses of approximately 5 percent or greater can impact study outcomes in significant ways. Differential losses are another concern and generally reflect something different between the groups which could be an indicator of a bias. Reviewers should make analysis choices that do not give an advantage to the intervention under study.
	Combinability of Studies & Summarizing Results Quality (and transparent) methods are employed for analyzing and summarizing results. If results are combined, they are combined in a reasonable and appropriate manner.	**Combinability Assessment** If results were combined, were the authors explicit about how they did so and did they employ quality methods? (For example: Were authors explicit about how they summarized the data such as in percentages or ratios?; Did authors make reasonable choices for grouping or stratifying outcomes of interest using such variables as age, duration of treatment, dosage, etc?) Did more than one author extract and combine data? This precaution reduces the risk of errors. **Homo-/heterogeneity** If results of the studies were combined, such as in a meta-analyses, did the authors apply tests of homogeneity/heterogeneity to assure that the variation between studies is due to chance (ie, p-value >.05, similar point estimates, overlapping confidence intervals, etc.)? However, this test is susceptible to problems depending upon the number of trials combined. Ideally a test for inconsistency is run *I2* statistic which reports percent of total variation due to heterogeneity instead of chance: [*I2* 0-25% is good, to 50% moderate, to 75% not good]. Fixed-effects model assumes each study as the same treatment effect. Random-effects model assumes effects of treatment vary around an overall average treatment effect. Random-effects model is more conservative and should be used for studies with greater inconsistencies. **Weighting** If weighting was employed, was a reasonable approach taken? Weighting is generally used to favor larger studies or higher quality studies and reduce potential bias from smaller studies or those of lower quality. Be aware, however, that larger studies are not necessarily higher quality so both size and quality need to be considered, and weighting from flawed studies could distort results. Consider sensitivity analyses where results of higher quality studies are compared with lower quality studies.
	Author's Discussion Robust discussion is provided including an assessment of limitations and with conclusions that are supported by the evidence.	**Discussion Considerations** Systematic reviews can be complex and often, studies may vary in key areas such as populations, outcomes, comparators, dosing, formulations, measurement instruments, duration and so forth. As such, authors might be well advised to have performed multiple sensitivity analyses, considering "what if's" (ie, do the results still hold when smaller studies are removed, for different populations, etc.?). And the discussion should describe any such testing, detail differences in studies and their implications, describe study limitations, and so forth. **Conclusions** Surprisingly, we frequently find in otherwise well-done systematic reviews, a failure to draw conclusions supported by evidence. For example, a review may have resulted in a yield of no valid information to answer a clinical question — yet the authors present an efficacy conclusion. These reviews should be considered to be uncertain. **Safety** It is acceptable to report on safety concerns; however, authors should be cautious in representing safety data as stronger than it is if it is based on low quality RCTs or observational data.
	Transparency The process is transparent and includes information about the specific studies utilized in the review	**Accessibility & Replicability Potential** Is sufficient detail provided that enables a through quality assessment of the review and such that the review could be replicated?

Advantages of Systematic Reviews

A well-done systematic review can be considered the lifeblood of medical information because it provides a summary measure of outcomes based on the scientific weight of valid and clinically useful studies. This is important, at the very least, because results of a single study which passes a critical appraisal screen could still be due to chance or even fraud. Further, systematic reviews may be able to answer questions not answerable by small studies because of power issues. Meaning a study population can end up being too small to show a statistically significant difference

even if there is one. By pooling data from multiple studies, a meta-analysis might reveal that there is a true difference since the population may be sufficiently increased. Systematic reviews are considered so important that they have been a recent focus of the Institute of Medicine (IOM).[4]

Disadvantages of Systematic Reviews

Systematic reviews can be complex to do and, often, greater complexity may result in greater opportunities for error and bias. Even well-done systematic reviews can only be informative to the extent that solid original studies are available and that there are sufficient studies similar enough to each other that they are combinable (the degree of similarity between studies is referred to as homogeneity). Frequently, studies may vary in key areas such as populations, outcomes, comparators, dosing, formulations, measurement instruments, duration and so forth, which may render the studies different enough that they cannot be combined to answer a clinical question.

Further, systematic review quality is wholly dependent upon the inclusion of valid and clinically useful studies. All too frequently, critical appraisal is done poorly or is not done at all — the result of which is the inclusion of flawed studies which can mislead. Estimates vary, but depending upon the flaw, distortion of study results may be as high as a relative 50 percent difference or more in outcomes.[5] Unfortunately, it is often difficult to tell if critical appraisal has been done well. Some systematic reviewers rely solely on use of the Jadad Scale[6] which we and some other experts think insufficient because it neglects many important critical appraisal considerations.[7] Others mention they perform critical appraisal, but do not provide sufficient information to judge the adequacy of their approach. We advocate auditing the review and provide guidance for doing so below. Also, systematic reviews become out-of-date with the inception of new research and must be updated.

What Characterizes a Quality Systematic Review?

As we have stated, a well-done systematic review uses a formal methodology to combine and analyze data from multiple studies, consisting of the following features.

STEPS AND TIPS FOR WORKING WITH SECONDARY STUDIES

Searching Tips

An efficient way to save time is to utilize sources that are more likely to be of higher quality, such as reviews from The Cochrane Collaboration. Recommended sources can be found at http://www.delfini.org/delfiniWebSources.htm.

If these sources do not provide a review relevant to your question, search the National Library of Medicine at www.pubmed.gov. You can specify systematic reviews by selecting the Limits tab and selecting Systematic Reviews under Subsets. Meta-analyses appear to be included in Systematic Review yields; however, you can check this or limit your search solely to meta-analyses using the Type of Article

Table 2. Match Your Clinical Question to Study Design Type
Seek out original studies or systematic reviews of the study types below to appraise for validity and usefulness.

Question Topic	Study Design Features	Study Type
Treatment, screening or prevention – What happens when you do something to someone?	Experiment – Randomize appropriately and compare groups	RCT
Diagnosis – Does it do what it is supposed to do & does it improve care?	Need data to calculate sensitivity/specificity/predictive values – independent, blind comparison with a "gold standard"	RCT/ Cross-sectional
Natural history – What happens when you follow natural course of a condition?	Not an experiment – follows what actually happens over time	Cohort
Prognosis – What happens when you follow the natural course of a condition?	Not an experiment – follows what actually happens over time	Cohort

box under Limits. You can also pre-specify these choices automatically by setting up your own tabbed yield windows through My NCBI which is accessible from the PubMed homepage. A tool providing more advice on searching is freely available at http://www.delfini.org/delfiniTools.htm.

Critical Appraisal of the Systematic Review

The next step is to critically appraise the review to determine if it meets the requirements for quality systematic reviews. **Table 1** can serve as an appraisal guide and checklist. There is also a free tool available specifically designed for critically appraising systematic reviews, complete with sample quality answers and problem answers, which is freely available at http://www.delfini.org/delfiniTools.htm.

Another efficiency is to determine if someone else has already critically appraised the review. In PubMed, review all published commentaries. Also, search for the review in the free Database of Abstracts of Reviews of Effects (DARE) at http://www.york.ac.uk/inst/crd/crddatabases.htm. DARE identifies potential systematic reviews and assesses them for methodological quality against a set of inclusion criteria and summarizes the results. If DARE finds significant problems with a review, be forewarned that their language is subtle. If they state, "Use with caution," the review should probably not be used to address questions of efficacy.

We would like to emphasize the need for a checklist or tool to be used by all health professionals when "reading" studies. A checklist helps to guide the process and serves as an important reminder about critical study elements that are otherwise not addressed in the article (if not addressed, they become an automatic threat to validity).

Quality Issues

Quality of any source — even those with the best potential or reputation for quality — is variable. Here are our most conservative suggestions for working with any source:

- The secondary study or source needs to be critically appraised using a tool appropriate for that purpose — PLUS the science used by the secondary study or source needs to be evaluated for validity and clinical usefulness.
- It is highly recommended that users review the methods used by the authors of the secondary study or source for critical appraisal considerations.
- Jadad scoring is insufficient to be relied upon, and conclusions from such systematic reviews should not be used unless reviewers demonstrate strong critical appraisal skills which they have applied as discerned through study discussion, and the review passes an audit.
- The most conservative approach is to review all studies considered to be of quality — and compare your outcomes to that of the review.
- A less conservative approach is to audit a sampling of included studies.

How to Audit a Secondary Source

- Of the included studies, critically appraise one or more original studies identified as high quality and one or two of the lowest quality.

- If these pass, it is probably reasonable to assume that the rest of the studies are of sufficient validity and clinical usefulness.
- A review that does not pass a critical appraisal review might still be usable as a foundation if the search strategy and criteria for excluded studies is sound.
- If yes, critically appraise all studies selected for inclusion, discarding any not meeting a rigorous critical appraisal screen for validity and clinical usefulness.

Grading the Systematic Review

Just as we assign a grade to a primary study or to individual study conclusions, we also assign a grade to systematic reviews and/or their conclusions. **Table 3** provides information about the Delfini Grading Scale. More information on evidence grading can be found in the Delfini Evidence Grading, Wording Conclusions & Results Table Tool at http://www.delfini.org/delfiniTools.htm.

Updating the Review

Also, be reminded that systematic reviews become out-of-date with the inception of new research. To be informed, readers need to perform a search and critical appraisal of relevant studies published after the search date (not the publication date) of the systematic review.

FINAL ADVICE & RECOMMENDATIONS

We would like to again emphasize the need for a checklist or tool to be used by all health professionals when "reading" studies. A checklist helps to guide the process and serves as an important reminder about critical study elements that are otherwise not addressed in the article (and, if not addressed, then become an automatic threat to validity). It may seem reasonable to carry a short checklist in your head; however, while helpful to have such a heuristic, we have found that threats to validity are often missed when a written checklist is not used. If utilized to document threats, a checklist can also facilitate evidence grading since this is easiest to accomplish when looking at a summary list of study concerns. Help in performing critical appraisals, tools and advice for performing every phase of evidence- and value-based clinical improvements is freely available at www.delfini.org.

References

1. Eddy DM. *Evidence-based medicine: a unified approach.* Health Aff (Millwood). 2005 Jan-Feb;24(1):9-17.

2. Eddy DM. *Evidence-Based Medicine: A Unified Approach. Health Affairs* 24 no.1 (2005):9-17. PMID: 15647211

Table 3. Delfini Validity & Usability Grading Scale for Summarizing the Evidence for Interventions

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 healthcare 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, execution and reporting with 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, prevention and diagnostic 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 healthcare decisions - some threats to validity were identified Advice: Studies achieving this grade should be of high quality in design, execution and reporting with non-lethal threats to validity and with sufficiently useful information to aid clinical decision-making, enabling reasonable certitude in drawing conclusions. 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, prevention and diagnostic 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. Also for diagnosis, valid studies assessing test accuracy for detecting a condition when there is evidence of effectiveness from valid, applicable RCTs. For natural history and prognosis: Cohort studies
Grade B-U: Possible to uncertain usefulness	The evidence might be sufficient to use in making healthcare 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 healthcare decisions; however, there are some study issues that raise continued uncertainty. Healthcare 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 healthcare 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. Uncertain Validity and Usefulness: This is a combination of the above. 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.

3. Eddy DM. The Challenge. JAMA. 1990;263:287-290. PMID: 2403607

4. Institute of Medicine (IOM). 2008. *Knowing what works in health care: A roadmap for the nation. Washington, DC: The National Academies Press.* PREPUBLICATION COPY: UNCORRECT-ED PROOFS

5. References available from Delfini Group, LLC upon request at www.delfini.org.

6. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. *Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clinical Trials.* 1996 Feb;17(1):1-12. PMID: 8721797

7. Jüni P, Altman DG, Egger M. *Systematic reviews in health care: Assessing the quality of controlled clinical trials.* BMJ 2001; 323: 42-46