Definitions and Criteria for Signals for Updating Systematic Reviews

This document presents supplementary material for Shojania KG, Sampson M, Ji J, Ansari MT, Garritty C, O'Rourke K, Rader T, Moher D. Updating systematic reviews, performed by the University of Ottawa Evidence-based Practice Center under Contract No. 290-02-0021 with the US Agency for Healthcare Research and Quality.

This project included an empric definitions below reflect the criteria applied to the new trials for each of the cohort of 100 quantitative systematic reviews evaluated.

A. Qualitative signals for need to update

We defined signals for two categories of qualitative signals for potential changes in evidence (i.e., for the need to update the original meta-analysis) in terms of their level of importance.

Potentially invalidating change in evidence: one would no longer want clinicians to act upon the results of the original review; an agency or organization that supported the production of the original review would want to retract the review until it could be updated. Examples of such changes include: high quality new evidence that suggests conclusions opposite to those in the original review; high quality new evidence suggests a degree of harm that would completely undermine use of the therapy; or, a head-to-head trial data show that the treatment evaluated in the original review is substantially inferior to another treatment. The specific operational details for each of the three criteria for potentially invalidating changes in evidence are provided below. Importantly, the designation *'potentially invalidating'* refers to the recommendations for clinical practice implied by the original meta-analysis, not the methods or conduct of the meta-analysis itself.

Major change in evidence: the conclusions of the original review have not been overturned or superseded, but new evidence clearly has the potential to affect clinical decision making. Examples of such changes include: new evidence that suggests the therapy does not work in certain patient populations; new evidence that affects how the therapy must be delivered in order to confer the benefit suggested in the original review (e.g., duration of treatment or in conjunction with other co-treatment); evidence about harm that would not completely undermine use of the therapy, but would clearly affect the decision to recommend therapy for at least some patient populations; changes in conclusion that fall short of 'opposite' but to those in the original review; high quality new

Qualitative signals were detected using explicit criteria for comparing the language used to characterize findings in the original meta-analysis with descriptions of findings in new meta-analyses that addressed the same topic, new 'pivotal trials', new clinical practice guidelines, or new editions of major textbooks (e.g., UpToDate). Pivotal trials were defined as trials that had a sample size at least three times the previous largest trial or were published in one of the 5 top general medical journals (*New England Journal of Medicine*, *Lancet*, *Journal of the American Medical Association*, *Annals of Internal Medicine*, and the *British Medical Journal*) based on a ranking by journal impact factor.

Specific types of qualitative signals are defined below.

Criteria for Signals of Potentially Invalidating Changes in Evidence

A1. **Opposing findings:** Pivotal trial, meta-analysis including at least one new trial, practice guideline (from major specialty organization or published in peer-reviewed journal), or recent textbook (e.g., UpToDate) characterizes the treatment in opposite terms to those in the cohort review: e.g., definitely effective → ineffective or vice versa (i.e., ineffective → effective). We operationalized 'opposite' as described at the end of this section.

We included guidelines and textbooks as sources of qualitative criteria because our definition of pivotal trial sets a very high bar. For example, we have not included any high impact specialty journals. The only way for a trial not published in one of the top 5 general medical journals to count as a pivotal trial would be for it to have a sample size at least three times that of the previous largest trial. To minimize our overlooking important new evidence while still permitting the efficiency of narrow searches for pivotal trials, we included guidelines and textbooks as sources of qualitative signals for changes in evidence. If new evidence has appeared that is judged of sufficient quality to inform recommendations in practice guidelines or textbooks, then it seems reasonable to call attention to these recommendations as signals for the need to update the original systematic review.

A2. **Substantial harm:** Pivotal trial, meta-analysis including at least one new trial, practice guideline, recent textbook calls into question the use of the treatment on the basis of harm (i.e., the treatment would no longer be recommended because risks outweigh benefits). A new result for harm that does not undermine use altogether, but has clear potential to affect clinical decision making would count as a 'major change' (criterion A6, 'Important caveat', as defined below).

A3. **Superior new treatment:** Pivotal trial, systematic review including at least one new trial, practice guideline, or recent textbook characterized another treatment as significantly superior to the one evaluated in the original meta-analysis (based on efficacy or harm)—to the point that it would be preferred in most settings.

Criteria for Signals of Major Changes in Evidence

A4. Important changes in effectiveness short of 'opposing findings': Pivotal trial, new meta-analysis, more recent practice guideline, or recent textbook does not contradict the previous review, but characterizes benefit in substantially different terms (e.g., therapy previously characterized as "promising", "likely beneficial" or similar description and now characterized as definitely beneficial.) This criterion is defined below in greater detail in the explanation of 'Operational definition of changes in conclusions.' Importantly, no attempt was made to distinguish between varying descriptions of "possibly effective." Characterizations such as "may be effective," "promising," "trends towards "effectiveness," and other similar phrases or concepts were all

categorized as "possibly effective." Thus, this criterion captured substantive differences in the characterization of treatment effects, not merely semantic differences.

A5. Expansion of treatment: Pivotal trial, new meta-analysis, more recent practice guideline, or recent textbook has expanded of the role of the treatment (e.g., the treatment has now been shown to be of benefit in children or the elderly; or benefit now shown to apply to primary prevention of disease, not just secondary prevention).

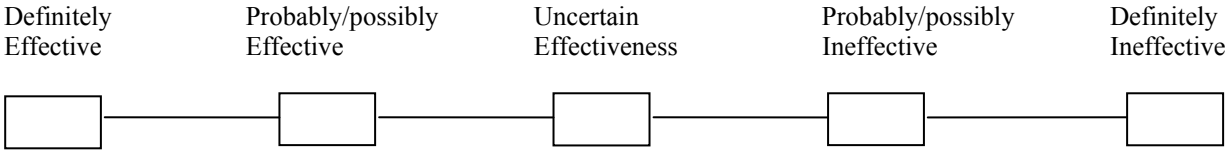
A6. **Important caveat**: Pivotal trial, new meta-analysis, more recent practice guideline, or recent textbook adds an important caveat, about the patient populations who benefit, way in which treatment has to be delivered in order to derive benefit, sustainability of benefit (e.g., benefits on short term outcomes, but not long-term ones), or increases in harm that are not sufficient to undermine use altogether, but would clearly affect the decision to recommend treatment for at least some patient populations.

A7. Opposing findings from discordant meta-analysis or non-pivotal trial: The treatment has been characterized in sufficiently different terms to the cohort review that disagreement would have met criteria for 'opposing findings' (criterion A1) except the source was not a pivotal trial, new-meta-analysis, or more recent practice guideline, or recent textbook—rather, it was a discordant meta-analysis or trial indexed in *ACP Journal Club*. ('Discordant meta-analysis' was defined as one that reached different conclusions than the original meta-analysis, despite effectively covering the same search period.)

We included this criterion because our definition of pivotal trial sets a very high bar, including only the top 5 general medical journals and trials with sample sizes at least three times the size of the previous largest trial. This criterion allows other sources of evidence to count as qualitative singles, without allowing any new trial with different results than in the previous systematic review to count as a signal for updating.

Operational definition of changes in conclusions

Labels such as 'effective' and 'ineffective' do not capture the distinction between trends towards effectiveness or uncertainty in the face of conflicting results or major limitations of the existing evidence. On the other hand, attempting to capture such nuances runs the risk of regarding semantic or stylistic differences between different authors. To balance these concerns, we consider conclusions about effectiveness in terms of a 5-point scale as shown below.



For systematic reviews that focused on adverse effects of treatment, we replaced effective/ineffective with 'harmful/not harmful.

In the interest of having qualitative signals of changes in evidence with high specificity, we did not attempt to make distinctions between statements of 'probable' or 'possible' benefit (or lack of benefit). We assigned descriptions such as 'promising,' 'possibly', 'probably', 'maybe', 'likely' into the same category. While still subjective, our labels are thus quite conservative—we are distinguishing firm or confident results, from trends, and equipoise or complete uncertainty (the middle position).

We defined 'opposite' conclusions (criterion A1 for potentially invalidating changes in evidence) as a movement of at least two positions on the above scale and 'important changes in effectiveness short of opposing findings' (criterion A4 for major changes in evidence) as a movement of one position on this scale. A movement of two positions generally includes movements from benefit to lack of benefit (or vice versa), but also includes movements from uncertain to definite conclusions about effectiveness. In this context, it is important to emphasize that we were careful not to equate summary conclusions (e.g., in article abstracts) of the type "the evidence does not permit definite conclusions" with 'complete uncertainty.' In many such cases, the results reported in the trial or meta-analysis indicated a trend, but the authors regarded the trend as inconclusive, on statistical or methodological grounds. Such cases were judged as 'possible' benefit (or lack of benefit, depending on the results). We reserved 'completely uncertain' for cases in which the authors clearly regarded the evidence as not indicating towards either benefit or lack of benefit. Thus, we regarded that a change from a definite or confident conclusion to compete uncertainty (or vice versa) would represent a potentially invalidating change in evidence. For example this would include a change from there being no basis on which to recommend a treatment to its being definitely recommended (or vice versa).

B. Quantitative signals of changes in evidence

We performed updated meta-analyses that combined the results from new trials with the meta-analytic result reported in the original review. To count as a quantitative signal, an outcome explicitly identified as a primary outcome in the original meta-analysis or any mortality outcome had to meet one of the criteria below. To count as a primary outcome, we required use of the word 'primary' or 'main.' Even in cases where those words used, we discounted such outcomes if authors stated that they had more than 3 such outcomes (on the grounds that more than 3 undermines the concept of 'primary').

B1. Change in statistical significance: at least one of the 95% confidence limits lies on a different side of the line of no effect (i.e. odds ratio or relative risk=1, risk difference=0). This criterion captures whether a result that was statistically significant in the original systematic review is now not statistically significant or vice versa—a previously non-significant result has become statistically significant.

To avoid counting trivial or 'borderline' changes in statistical significance as quantitative signals for updating, we required that at least one of the two results (i.e., the original and updated meta-analyses) have a p-value outside the range of 0.04 to 0.06. In other words, we excluded cases in which the original systematic review reported a borderline result and the updated result is also borderline but happens to lie in the other side of the line of no effect. For instance, a change from p=0.041 to p=0.059 would not count as a quantitative signal to update, nor would the converse change (from p=0.059 to p=0.041).

B2: **Change in effect size of at least 50%:** the new result indicates a *relative change* in effect size of at least 50%. For example, if RRRnew / RRRold <=0.5 or RRRnew / RRRold >=1.5, where RRR is the relative risk reduction. Thus, if the original review has found RR=0.70 for mortality, this implies RRR of 0.3. If the updated meta-analytic result for mortality were 0.90, then the updated RRR would be 0.10, which is less than 50% of the previous RRR. In other words the reduction in the risk of death has moved from 30% to 10%. The same criterion applied for odds ratios (e.g., if previous OR=0.70 and updated result were OR=0.90, then the new reduction in odds of death (0.10) would be less 50% of the magnitude of the previous reduction in odds (0.30). For risk differences and weighted mean differences, we applied the criterion directly to the previous and updated results (e.g., RDnew / RDold <=0.5 or RDnew / RDold >=1.5).

B3: **New and old point estimates differ significantly**. This test was operationalized by applying a Z-test to determine if the difference between in the original meta-analytic result and the updated result is statistically significant. No examples of this criterion were found, so it is not mentioned in the main results.