How Systematic Review and Meta-Analysis Work

By Chuck Dinerstein, MD, MBA — June 23, 2017

My colleague, Josh Bloom, has written about flaws in IARC's recent systematic review of glycophates [1] and I thought perhaps this was a time to explain systematic reviews (SR) and their closely related kin, the meta-analysis (MA). Keeping up with the latest thinking and knowledge, even for specialists, is challenging. Textbooks, given the time necessary to write, edit and print are repositories for the accepted 'dogma.' Journals and conferences bring new information forward more quickly. But even then it is difficult for individuals to synthesize multiple articles and sources to 'know' what is both correct and useful. Archie Cochrane in his 1971 book, Effectiveness and Efficiency, was the first person to champion systematic reviews of the literature so that clinicians had evidence to base clinical decisions. Since then an entire industry has grown up around SRs. The audience has shifted from clinicians needing guidance to feeding news headlines and providing policy makers with summaries of clinical evidence.

Textbooks are narrative reviews; the chapter author provides a broad overview of generally accepted beliefs about etiology, progression, treatment, and results. While written by an expert what portions of the dogma are included and referenced is a personal decision susceptible to bias. Cochrane demonstrated this biases and offered the systematic review as a solution – gathering all the empirical evidence to answer a specific clinical question. It is impossible to collect all evidence, so evidence is chosen that most closely answers the question being asked. Selection bias, Cochrane's original concerns, returns. What to do? Using eligibility criteria to select evidence is thought to eliminate the choice bias of narrative reviews. But SR for all its trappings is a qualitative summary picking and choosing among the evidence for what is felt to be best. To further reduce choice bias we pursued quantitative measures. Meta-analysis applies statistical measures to systematic review's evidence – pooling multiple studies, increasing statistical power (the certainty of the findings), condensing a vast amount of information into a few quantitative results. Let us be fair, SR and MA do a lot of the heavy lifting in collecting and summarizing information; but as

scientific documents that should contain a hypothesis, a means of collecting data, the data itself and discussion (it is so hard to remove narration from science) of data's meaning. We are obligated, as information consumers, to assure ourselves that this methodology is followed.

1.**Hypothesis** - A clear clinical question is asked and answered. Typically, SR address cause, exposure, prognosis or treatment for specific problems or populations.

2. **Methodology** – The experimental variable is the literature chosen for review. Different people approach problems in differing and legitimate ways. "The elements that go into determining what's measured in a study are complex. They're related to the needs of good quality and/or feasible research – not necessarily the very specific questions we might have." To be systematic, all relevant information needs to be identified. Bias, judgment's inevitable companion can quietly slip in at this juncture

Reviewing only published materials in English, often facilitated by a computerized search excludes unpublished work, conference proceedings and non-English articles, the omission introducing publication bias. Publication, the size of the study, statistical significance are not criteria at this juncture, only the study's relevance to the question being asked and its design are relevant. Other biases lie within the selected papers. Positive results are published, negative results are not so that positive results may be overstated. And because each study has its methodology there may be selection bias in the group studied. Frequently surrogate outcomes and biomarkers are measures resulting in varying levels of detection and differences in the care of the groups outside the intervention may introduce performance and attrition bias. In demonstrating replicability and making bias apparent, there should be an identifiable, explicit search strategy.

For any particular question, many sources likely contain answers. So, in a second screening, we reduce the universe of studies selecting those of good quality. Quality, in this case, means that the chosen articles use appropriate methods to reach their conclusions. In essence, we are trying to prevent garbage in, garbage out. Generally, two or more SR authors will briefly identify papers by title, abstract and content to eliminate those that are inconsistent with the search protocol. These articles are, in turn, more rigorously reviewed to choose the articles to be studied. When reviewers disagree, a third tie-breaker is employed.

3. **Results** - For each selected study, its treatment effect (conclusion) is summarized along with confidence intervals. When outcomes are continuous, for example, how much weight is lost or gained, the measure is the difference in the mean values before and after treatment. For outcomes that are more yes no, odds ratios or relative risks express the measure. [1] Each measure has an associated confidence interval (CI) indicating the range that probably (usually 95% of the time) contains the actual value. It measures precision not accuracy; smaller CIs are more precise.

These individual measurements become the MA's data points and combined to give the MA's results. Measures are not combined equally, they are weighted. Studies with more 'information' have more gravitas because they are likely to be closer to the "true effect." Or are they? More 'weight' is usually given to studies with large sample sizes, but other criteria can be utilized such as the quality of the evidence – is it a randomized control or observational study? The key word is preference, bias, the enemy to be vanquished, can stealthily return as investigators weigh one

paper more heavily than another.
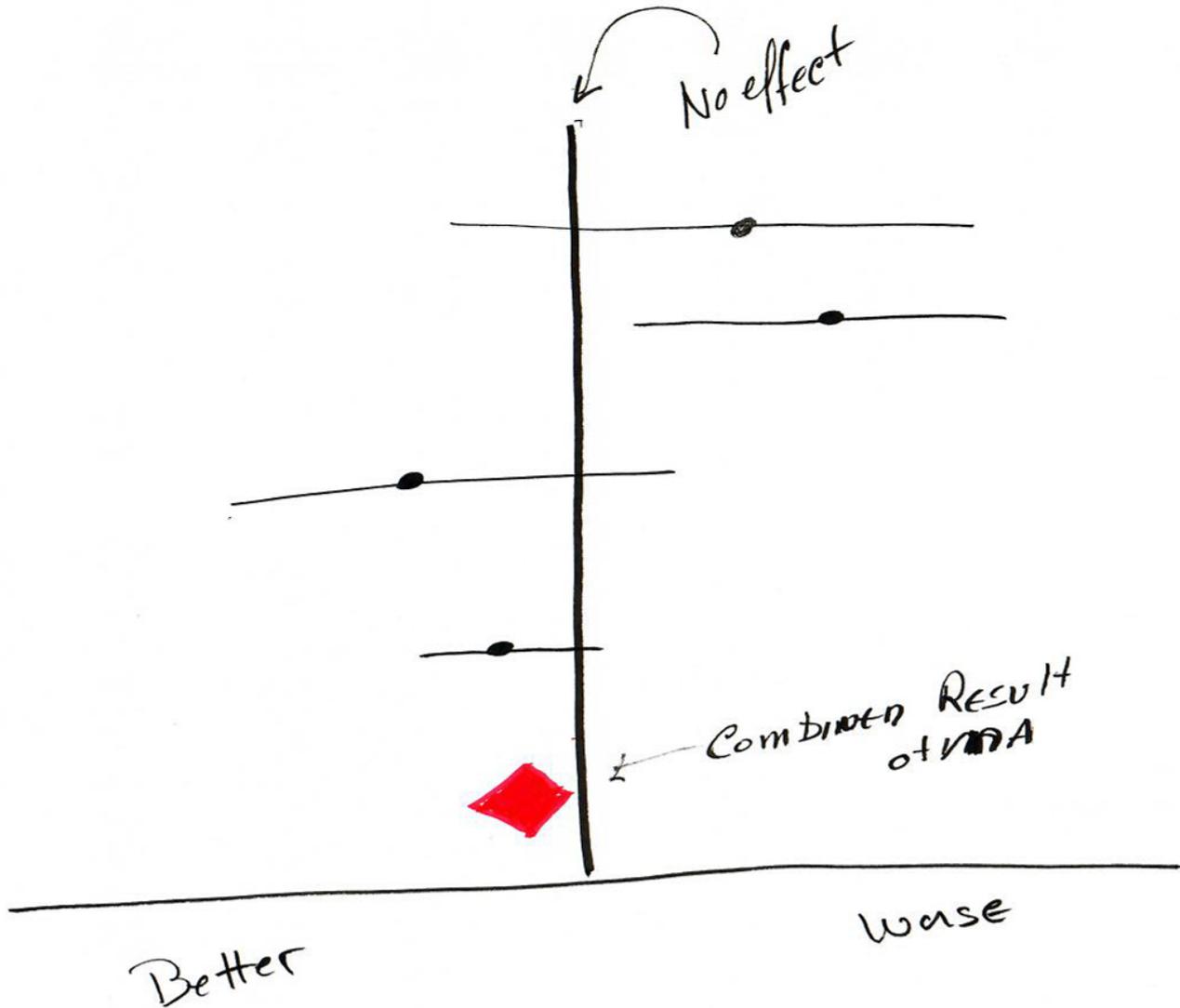
Each papers treatment effect along with the MA's summary is displayed graphically as a Forest Plot. Each study in the SR is represented by a line. The point is the measure and the line length the confidence interval. Usually, a diamond is used to represent the MA's weighted summary value, its location representing the measure, its width the MA's confidence in that result. Thinner and elongated suggests less certainty, big and squat more confidence. The lone vertical line represents no effect. Any of the individual study's line touching the vertical is not statistically significant.

It is inevitable that in gathering studies together they will vary, in the participants, intervention, outcomes or the methodology used in assessment. Pooling studies that shouldn't be combined is MA's Achilles tendon. Are the selected papers all apples or have a few oranges snuck in, creating heterogeneity of the data? For MA we want to be looking at similar studies. Heterogeneity can be, as you might expect, calculated statistically or you can use common sense to determine whether there are oranges in the bunch. You can make a judgment of whether the studies are similar enough concerning populations and outcomes to be considered together. One clue to the presence of heterogeneity is the degree the horizontal lines of the Forest plot overlap. Too little

overlap reflects the presence of heterogeneity. As with most things, we want some heterogeneity but not a great deal. Authors can address the problem of too much heterogeneity by not doing a meta-analysis, leaving only a systematic review. Frequently heterogeneity is treated by looking at sub-groups within the data. And that is OK except you need to remember because authors will rarely remind you, that the 'power' of the observation is diminished. The meta-analysis of 40 papers with 500,000 participants may now reduce to 3 articles concerning 10,000.

Sensitivity analysis demonstrates whether removing some of the chosen studies alters the outcome. When the outcome does not change, the MA is considered 'robust.' It is an appealing notion that SR and MA by combining many studies provide more authoritative information, but SR/MA is not a real experiment, it is a descriptive study with judgments at every juncture. One paper may carry more weight than many of the others and understanding that dominant study is necessary. And of course, if an article initially excluded alters the outcome, we never know.

4. **Discussion** - Authors utilize narrative putting the findings of the SR and MA in context, connecting the dots. Bias stealthily enters, as authors choose how those dots are connected, data is not always the decider. Scientists, and yes this is a generality, believe that answers can be found, it is a philosophical belief. So, finding no answer is both difficult to accept and publish. With this mindset, everything has an impact, even if obscure, tenuous or clinically insignificant; at least it can be published and tenure one paper closer.

And there is one further consideration, SR and MA have a shelf life. The papers underlying the analysis are fixed in time, and things change giving SR and MA expiration dates. They are more snapshots than movies. Despite the statistical pretense, It is exceedingly difficult for SR and MA to be unbias, we make judgments. Using the following checklist is a means of making sure what you are consuming is informationally nutritious and not merely fast food with empty calories.

Questions to consider when appraising a systematic review

- Did the review address a clearly focused question?
- Did the review include the right type of study?
- Did the reviewers try to identify all relevant studies?
- Did the reviewers assess the quality of all the studies included?
- If the results of the study have been combined, was it reasonable to do so?
- How are the results presented and what are the primary results?
- How precise are the results?
- Can the results be applied to your local population?
- Were all important outcomes considered?
- Should practice or policy change as a result of the evidence contained in this review? Can the results be applied to practice with assurances of benefit/harm/cost?

[1] Odds = # at endpoint/# not at endpoint. Risk = # at endpoint/total #. The odds ratio is odds of Rx/odds no Rx and same for relative risk.