FROM THE EDITORS' DESK

From the Editors' Desk: Bias in Systematic Reviews—Let the Reader Beware

Jeffrey L. Jackson, MD, MPH^{1,2} and Akira Kuriyama, MD, MPH³

¹Zablocki VAMC, Milwaukee, WI, USA; ²Medical College of Wisconsin, Milwaukee, WI, USA; ³Department of General Medicine, Emergency and Critical Care Center, Kurashiki Central Hospital, Kurashiki, Japan.

J Gen Intern Med 33(2):133–5

DOI: 10.1007/s11606-017-4236-2

© Society of General Internal Medicine (outside the USA) 2017

Systematic reviews are relied upon by clinicians and policymakers as high-quality evidence for decision-making. In some hierarchies of evidence quality, systematic reviews are ranked at the top, higher than randomized controlled trials. A properly conducted systematic review that is based on high-quality articles provides very strong evidence; policymakers recognize this value and solicit many such reviews. Busy clinicians rely on well-executed systematic reviews to quickly synthesize the literature and guide them in managing patients. However, there are important inherent weaknesses that can limit the quality of systematic reviews and can lead to erroneous conclusions. Consumers of systematic reviews should approach them with a healthy sense of skepticism. Unfortunately, many of these weaknesses may not be obvious to the various stakeholders who routinely invest their trust in such reviews.

The first bias one encounters is inherent in the research enterprise. Generally, researchers design studies to demonstrate maximum effect; hence they are careful about the population selected, the interventions tested, and the outcomes assessed. For example, interventions are often tested in populations that are at high risk for experiencing the outcome in question and are likely to respond to the intervention. Not only can this make results difficult to generalize, but it often results in studies that overstate the benefit; subsequent studies frequently find less benefit. This is particularly common when the pool of available studies consists of small, single-center investigations, trials that are particularly likely to show large effects.¹ While usually not intentional, the combination of careful selection of population and outcomes can distort clinical research and obscure truth. This inherent bias in the research enterprise cannot be fixed; however, if there are a sufficient number of studies, conducted across different settings and populations, this problem can be partially ameliorated with systematic reviews.

Systematic reviews are vulnerable to a number of biases. First, the question may not be well-defined. A marker of a high-quality systematic review is the inclusion of a purpose

statement in which the specific research question is clearly articulated. Specificity is key: what are they willing to accept as evidence, what population, what intervention, what duration, what doses, what outcomes and when? Second, the search strategy should be broad enough to reassure readers that no studies were left out. Researchers commonly limit their search to English-language articles. However, there is no difference between the quality of articles in English and those in other languages, and limiting the search to English-only should be a red flag to readers.² The bibliographies of retrieved articles should also be reviewed. All search strategies should include input from an experienced medical librarian.

Third, the quality of included articles must be carefully assessed. The higher the quality of articles that provide the bulk of the systematic review's conclusions, the more confident readers can be in those conclusions. Readers should be wary of reviews that are based on observational data or uncontrolled clinical trials. A statistical package will analyze whatever data the analyst feeds it, and will provide an estimate of an overall effect and a 95% confidence interval. This can mislead readers into believing that our understanding of the effect is more concrete than it is. Pooling of observational data can lead to answers that differ materially from the results of randomized controlled trials. Readers should also be wary of systematic reviews built on low sample numbers, either few studies or small samples within each study. The number of studies necessary to produce believable results is an area of ongoing debate. A recent article discussed the benefits of Trial Sequential Analysis (TSA), a method that allows meta-analysts to determine the likelihood that the review contains either a type I error (finding a difference that does not exist) or a type II error (reporting no difference when in fact one exists).³ TSA theory is similar to the principles governing stopping rules for interventional trials. It takes the total sample size and the size of the effect, and establishes boundaries. These boundaries provide confidence that the sample size is large enough to avoid type I or II errors. We strongly recommend routinely including them in meta-analyses.

Fourth, very careful attention must be paid to the issue of publication bias. Most readers recognize that studies with non-significant results face greater publication hurdles. This is a well-recognized phenomenon; protocols with non-significant results are both less likely to be published and, when published, more likely to encounter significant delays than studies with significant

findings.⁴ Reviews that have recent stop dates are likely to miss negative trials (that generally take longer to be published). A more insidious and equally important form of publication bias is the practice of selective reporting: among all outcomes collected in a particular study, only those that are significant are likely to be reported. For example, fibromyalgia affects multiple domains of patient function. While studies of fibromyalgia interventions collect outcomes in multiple domains (pain, energy, sleep, trigger points, functioning), many trials report only those outcomes that change and not those that do not respond to treatment.⁵ Sometimes the “primary” outcome reported in clinical trials is not the primary outcome the study was intended to assess, the outcome originally stated in the study protocol; instead, it is the outcome with the greatest response to the intervention.⁶ This can lead readers to believe that the intervention’s impact on outcomes is more impressive than it actually is. Increasingly, journals are requiring authors to include a statement in their manuscript as to whether the reported primary outcome is the same as that originally specified in the protocol, and to provide information, including online access to the data, for all outcomes assessed, not just those reported in the article.

The inadequate use of quality measures for included trials is also an issue. Most systematic reviews formally assess study quality, but many do nothing with that information and do not use it to inform their conclusions. Better-quality systematic reviews will rate the strength of their conclusions based in part on the quality of the included studies. An excellent, formal method is the GRADE [Grading of Recommendations, Assessment, Development and Evaluations] approach to rating the quality of evidence.⁷ Sophisticated reviewers should demand a formal assessment; the absence of such an approach should make the reader wary of the reviewer’s conclusions.

Another source of bias is the way the data is analyzed and reported. Not uncommonly, the meaning of an outcome measure on patient quality of life is not clear. For example, tricyclic antidepressants have been reported to have a significant impact on reducing migraine headache, with a standardized mean difference of 1.29 compared with placebo—a “large” effect.⁸ However, when re-analyzed as the reduction in the actual number of headaches experienced, the reduction is from 21 headaches a month to 16, a reduction of about 5 headaches a month. From a clinical perspective, this “significant” effect represents minimal improvement from the perspective of the patient experiencing migraines.⁹ The response from the National Institutes of Health (NIH) to this common problem was to develop a set of patient-centered scales, the Patient-Reported Outcomes Measurement Information System (PROMIS), that could be used across studies to standardize outcome assessment.¹⁰ The NIH is also conducting research to define clinically meaningful differences and is encouraging researchers to use these measures in addition to measuring disease-specific outcomes.

Another common problem in systematic reviews is carelessness regarding outcome time points. Many systematic reviews pool outcomes from individual trials based on those experienced at the trial conclusion. This can result in the

pooling of trials of vastly different durations: 4-week trial outcomes are pooled with findings after 4 months and even 4 years. Analysts must carefully note the time of the outcome and pool only those outcomes from similar time points. Heterogeneity is another important quality marker for systematic reviews. Of course, if there were no heterogeneity between studies, if all reported the same effect, there would be no reason to conduct a systematic review—the pooled effect would be the same as that reported in all the trials (though the confidence interval would be narrower). Some heterogeneity is important in order to explore the possible source of the difference in outcomes between studies. Since most systematic reviews do not have access to patient-level data, the ability to determine the source of heterogeneity is limited. This is complicated by the narrow range of certain independent variables seen in studies. For example, if the analyst is hoping to explore whether gender is one reason for differences in outcome, but the proportion of women in individual studies ranges only from 45 to 55%, this range restriction makes it difficult to find possible differences. Most authors of systematic reviews blithely tell readers that “we found no difference in outcome by sex,” without alerting the reader that the range restriction was too narrow to show differences that may actually exist. Another problem occurs if the heterogeneity between studies is too great to instill confidence that the pooled results are a meaningful approximation of truth. Merely using a “random effects” approach to deal with this problem does not correct the underlying fundamental problem of trying to combine apples and oranges. “Fixed effects” approaches will often provide a narrower confidence interval. Readers should be very wary of analysts who use a fixed effects approach and have a *p* value that is barely significant. Most likely, if they had used random effects models, the outcome would not have been different. In general, the use of random effects will constitute the more conservative approach, and should be the default model for analyzing data. When the statistical heterogeneity is small, random and fixed effects models provide the same results. Alternating models according to outcomes may create the potential for multiplicity and spurious findings.

The *Journal of General Internal Medicine* is committed to reducing these sources of bias as much as possible. First, to help ensure a robust pool of articles for conducting systematic reviews, we will publish quality research, regardless of whether the results are significant, and we will publish both in the same timeline. Second, we will require authors to include in their methods the primary outcome specified in their protocol, and unless this is a secondary analysis and the primary outcome has been reported elsewhere, we will ask authors to provide an explanation for the change. Third, unless the author is planning additional papers from the data set, we will ask researchers to provide all outcomes assessed, not just those that are statistically significant. Finally, we will work with our authors to promote the reporting of clinically meaningful outcomes, or at least including an explanation of what would constitute a clinically meaningful change. With respect to

reducing bias in our published systematic reviews, we have assembled a team of general internists who are expert meta-analysts, with an impressive track record of publications. We will examine each review submitted to our journal in an effort to reduce common sources of bias in systematic reviews.

Well-executed, appropriately meticulous systematic reviews can provide great benefit to the research and clinical communities. They can highlight areas of strength and weakness in the evidence base. Poorly conducted reviews can set science back and can mislead policymakers and clinicians.

Corresponding Author: Jeffrey L. Jackson, MD, MPH; Medical College of Wisconsin, 5000 W. National Ave, Milwaukee, WI 53208, USA (e-mail: jjackson@mcw.edu).

REFERENCES

1. **Dechartres A, Boutron I, Trinquart L, Charles P, Ravaud P.** Single-center trials show larger treatment effects than multicenter trials: evidence from a meta-epidemiologic study. *Ann Intern Med.* 2011;155(1):39–51.
2. **Moher D, Fortin P, Jadad AR, et al.** Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet.* 1996;347(8998):363–6.
3. **Wetterslev J, Jakobsen JC, Gluud C.** Trial Sequential Analysis in systematic reviews with meta-analysis. *BMC Med Res Methodol.* 2017;17:39.
4. **Stern JM, Simes RJ.** Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ.* 1997;315:640–5.
5. **O'Malley PG, Balden E, Tomkins G, Santoro J, Kroenke K, Jackson JL.** Treatment of fibromyalgia with antidepressants: a meta-analysis. *J Gen Intern Med.* 2000;15(9):659–66.
6. **Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG.** Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA.* 2004;291:2457–65.
7. **Balshem H, Helfand M, Schünemann HJ.** GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol.* 2011;64(4):401–6.
8. **Jackson JL, Cogbill E, Santana-Davila R.** A comparative effectiveness meta-analysis of commonly prescribed drugs for the prophylaxis of migraine headache. *PLOS One.* 2015;10(7):e0130733. <https://doi.org/10.1371/journal.pone.0130733>.
9. **Jackson JL, Mancuso JM, Nickoloff S, Bernstein R, Kay C.** Tricyclic and tetracyclic antidepressants for the prophylaxis of frequent episodic or chronic tension-type headache in adults. *J Gen Intern Med.* 2017. <https://doi.org/10.1007/s11606-017-4121-z>.
10. **Rose M, Bjorner JB, Becker J, Fries JF, Ware JE.** Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidemiol.* 2008;61(1):17–33.