EDITORIAL

Systematic reviews and meta-analysis: why might more clinical trials yield no greater precision?

Suppose you are caring for a preterm newborn who is hypotensive (for present purposes, assume this diagnosis can be made without controversy). Systemic hypotension has been associated with a variety of adverse outcomes. Your colleagues who work in the NICU across town have begun using corticosteroid treatment instead of volume expansion, followed by vasopressor/inotropic therapy, and you wonder about trying it for this newborn under your care. You check a systematic review (SR) at Cochrane Neonatal (http://www.cochrane.org/CD003662/ NEONATAL_corticosteroids-for-treating-hypotension-in-preterm-infants), reflecting four studies with 123 enrolled neonates. One study reported that compared with dopamine as primary treatment, persistent hypotension was more common in hydrocortisone treated infants. Two studies compared steroid against placebo, and found that persistent hypotension occurred less frequently in the steroid-treated infants. You determine that the study question posed by the SR was: 'What are the effectiveness and safety of corticosteroids used either as primary treatment of hypotension or for the treatment of refractory hypotension in preterm infants?' However, you are uncertain whether the SR has indeed answered the question. Moreover, you feel you have gained no more clarity about how to treat your hypotensive newborn, recognizing that your question was different from the study question. You also notice that the SR was published in 2011, so you wonder: perhaps a new randomized controlled trial (RCT) might help.

An SR is a study of studies with the objective of resolving a specific and answerable question.[1] An SR commonly includes meta-analysis, which applies specific research and analytical strategies to justify summarizing the results of several studies in a single effect estimate. In theory, incorporating all available data into an analysis improves the precision of that estimate because it increases the size of the sample—the number of subjects drawn from the applicable population.[2]

In this issue of the *Journal*, Hay *et al.*[3] evaluate the extent to which the most recent randomized trial impacts two aspects of an SR. One aspect is the precision of an effect estimate as measured by the width of the confidence interval (CI), that is, the uncertainty regarding the quantitative measure of the effect. The second is the clinical significance as measured by the distance of that quantitative effect measure from the null, a value denoting no difference—for example, in the case of relative risk, how far is the measured effect from a value of 1. Their intriguing and thought-provoking study found that additional RCTs usually did not contribute to substantially decreasing uncertainty around an effect estimate derived from antecedent RCTs, nor did it often substantially change the direction or magnitude of the effect. After enumerating many beneficial effects of our profession's strong reliance on 'comprehensive systematic review and meta-analysis of the randomized clinical trial literature,' the authors express some surprise that, in this light, they nonetheless 'found so little benefit at the margin of adding trials to these meta-analyses.' The authors describe a variety of factors that could contribute to the frequently meager added-value of another RCT to an existing series. These include small sample (underpowered), surrogate (inconsistent) outcomes, loss to follow-up and a wide variety of biases.[3] Additionally, studies can be clinically heterogeneous: for example, there may be differences across studies in details of enrolled patients or interventions.

That the findings of Hay *et al*[3] are surprising may be predicated on an important assumption: that more trials (observations) = less sampling error. The effect estimate, or point estimate, is a single number estimating the actual value for the population. Therefore, the point estimate varies with each sample, and thus a single estimate cannot indicate that variability. Consequently, analytical methods also estimate the size of the interval containing that variability, the CI, the range of values that are likely to contain the actual value. Moreover, the CI too varies across samples.[4] The CI takes into account the amount of variability in the sample estimate.[5] For a given sample, an analyst can compute CIs of varying widths, to reflect the chances the investigators wish to have of including the actual value. In particular, if several independent studies were done involving random samples from the same population and 95% CIs were computed for each, then on average 19 of every 20 (95%) of these CIs would include the actual population value; and 1 of every 20 would not.[4] The 95% cut point is common, but other cut points could be selected to suit investigative circumstances. The cut point value affects the width of the resulting interval, so for the same sample a 99% CI will be wider than a corresponding 95% CI. For a given percentile cut point, a more precise estimate will enclose a narrower range of values than one less precise. Ultimately, the precision of an estimate reflects the variability of the individual values in the sample.[5]

How might these statistical fundamentals account for the findings of Hay *et al.*?[3] As the authors detail in their Table 1, the median number of studies included in each of the three outcome categories was three or four; the median number of patients in the total SR ranged from 163 to 398; and the median sample size of the last study being evaluated for impact ranged from 57 to 101. Thus, the latest RCT added to the cumulative meta-analysis had substantial potential to decrease sampling error because its proportional contribution to the aggregate data set was large. However, the authors' analysis revealed that measures of sampling error often did not substantially decrease as the cumulative analysis proceeded. That each consecutive RCT usually did not substantially decrease sampling variability—did not improve the signal to noise ratio, so to speak—suggests that consecutive study samples were not drawn from a single population. Fine-grained details of each component study in a meta-analysis may differ sufficiently so that the meta-analysis, despite its stated objective, is actually attempting to resolve more than one precise study question. Component studies may not all be sampling from a single population. Since a CI is but an estimate of sampling error,[2] it appears the meager marginal benefit of additional trials signals a need for a more fine-grained and uniform investigative focus. In other words, one strategy to decrease uncertainty around a particular clinical trial effect entails subsequent research efforts specifically designed to enlarge the currently available sample by ensuring consistent sampling from the population of interest, the same one as was sampled by antecedent RCTs. Hay *et al.*[3] advise 'formal consideration of optimal information size.'

Let us return now to the opening scenario. It focused, as clinicians often do, on an individual patient. So a bit more discussion seems pertinent on applying to a particular patient the information contained in the precision of an interventional effect estimate computed from a study sample; or as Hay et al.[3] frame it in their title, the uncertainty surrounding an effect estimate. In general, RCTs provide the average answer to the question of what is the relationship between exposure and outcome. Blended within an average result may be patient outcomes better or worse than that, including patients who may have experienced extraordinary benefit or harm. The goal of integrating multiple studies in a meta-analysis is to compute a weighted average result.[6] It is crucial to appreciate that these research efforts therefore offer clinicians a relatively vague message about exactly what to expect for a particular patient. The research tells clinicians what to expect, on average, among a group of patients with an array of characteristics matching those in the study sample. Compounding this uncertainty, the average expectation itself is imprecise and uncertain. Among a (sufficiently large) sample of such patients a clinician cares for, the actual average effect will usually be some value within the CI. Closely related to this inferential consideration and perhaps more challenging yet, but beyond the present scope, is how to clearly define for an individual patient the notion of quantitative risk derived from RCTs and SRs, and how to frame it meaningfully.[7,8]

In light of all this, note that the estimated average effect and CI generally answer questions contained within a specific interrogative framework: 'Does it work?' An alternative framework, infrequently applied because it often entails great practical challenges, is to ask 'For whom might it work, and under what circumstances?'[9] Such an approach maps more closely to the priorities of minimizing sampling variation and providing answers more readily applied to an individual patient. Consequently, adding trials reflecting such design considerations to meta-analyses might yield greater benefit at the margin. Then again, with such trials perhaps we would discover we need fewer SRs.

CONFLICT OF INTEREST

The author declares no conflict of interest.

Joseph Schulman
California Department of Health Care Services, Sacramento, CA, USA
E-mail: joseph.schulman@dhcs.ca.gov

REFERENCES

1 Soll R. Meta-analysis in neonatal perinatal medicine. *NeoReviews* 2011; **12**(1): e8–e12.
2 Bland M. *An Introduction to Medical Statistics*, 4th edn. Oxford University Press: Oxford, UK, 2015.
3 Hay S, Kirpalani H, Viner C, Soll R, Dukhovny D, Mao W et al. Do trials reduce uncertainty? Assessing impact through cumulative meta-analysis of neonatal RCTs. *J Perinatol* 2017 (this issue).
4 Kirkwood BR, Sterne JAC. *Essential Medical Statistics*. 2nd edn.Blackwell Science Ltd: Oxford, UK.; 2003; 501 p.
5 Schulman J. *Managing Your Patients' Data in the Neonatal and Pediatric ICU: an Introduction to Databases and Statistical Analysis*. Blackwell: Oxford, UK, 2006.
6 Savitz DA. *Interpreting Epidemiologic Evidence: Strategies for Study Design and Analysis*. Oxford University Press: New York, USA, 2003.
7 Goodman SN. Probability at the bedside: the knowing of chances or the chances of knowing? *Ann Int Med* 1999; **130**(7): 604–606.
8 Davidoff R. Standing statistics right side up. *Ann Int Med* 1999; **130**(12):1019–1021.
9 Pawson R, Tilley N. *Realistic Evaluation*. Sage Publications, Ltd: London, UK, 1997.