FROM THE EDITOR'S DESK

Tips for Busy Clinicians to Gauge Conclusions from Clinical Trials

Jeffrey L Jackson, MD MPH[1], Allan Detsky, MD PhD[2,3], and Akira Kuriyama, MD, DrPH, PhD[4]

[1]Zablocki VAMC, Milwaukee, USA; [2]Institute of Health Policy Management and Evaluation and Department of Medicine, University of Toronto, Toronto, Ontario, Canada; [3]Department of Medicine, Sinai Health System and University Health Network, Toronto, Ontario, Canada; [4]Emergency and Critical Care Center, Kurashiki Central Hospital, Okayama, Japan.

J Gen Intern Med 37(1):1–3
DOI: 10.1007/s11606-021-07037-4

One of the problems clinicians face is knowing how much to trust conclusions drawn from the results of clinical trials. Trialists design studies to maximize the chance that they will detect differences in outcomes and label those differences as statistically significant. To achieve statistical significance, the difference either needs to be sufficiently large, the sample size needs to be large, or the groups compared need to have sufficiently small variation. Trialists can increase the chance of finding sufficiently large differences by selecting populations at high risk of the outcome and can decrease variation by choosing relatively homogenous populations. Clinical research is very expensive, both for patients and trialists. Because of ethical, economic, and pragmatic reasons, studies are powered to be barely statistically significant, based on estimates of the size of differences that are deemed clinically important (and to some extent expected) and variation. Design issues in the planning stages of trials such as desired power (and thus the sample size required to achieve that power) are fundamentally economic. The incremental benefit derived from not missing a true difference in outcomes (measured by estimating the value of health benefits for patients who will use an effective therapy) is balanced against the incremental cost of increasing the trial power and sample size (to look for small differences).[1] Missing a true difference means either people will not be offered effective therapies or the trial needs to be repeated.

The process of conducting clinical trials and evaluating their outcomes is both complicated and sophisticated. Clinicians often find themselves glossing over the methods and results sections of published reports in deciding whether or not study conclusions can be applied in their practice. However, there are a number of tips that can help busy providers decide how strong study evidence is.

1. Do not confuse statistical and clinical significance. The decision about whether to use a new intervention depends on many factors. How expensive is it? How difficult is it? How much harm does it cause the patient compared to the benefit? Studies that are sufficiently large, for example, databases, can find clinically trivial differences to be statistically different. On the other hand, studies that are too small may find large benefit from an intervention, yet fail to be statistically significant. Clinicians should ask whether the intervention differences will produce meaningful improvement in patient's lives.

2. Be aware of how the results are presented. Trialists present their results in the best possible light. For interventional trials, there are a number of different ways results can be presented: relative risk, relative risk reduction, absolute risk reduction, and number needed to treat. This can be confusing, both to clinicians and patients. Relative risk (RR) is the percentage of patients who have the outcome in one group divided by the percentage in the other group. For example, in a clinical trial to reduce myocardial infarctions (MI), if 4% of patients have an MI in the control group and 3% in the intervention group, the relative risk is 3%/4%=0.75. This is hard to explain to patients, "If you have the intervention, your chance of having a heart attack will be 75% of what it would be if you didn't have the intervention." Relative risk reduction is how much the risk is reduced and is calculated as 1-RR. This is easier to say to patients, "If you have the intervention, your risk of having a heart attack will be reduced by 25%." Both of these can overstate the difference. The absolute risk is the difference in risk between the two groups, in this case 4–3%. "If you have the intervention, your risk of having a heart attack will be reduced by 1%." The number needed to treat is the inverse of the absolute risk reduction, 1/0.01. "We would have to treat 100 patients to prevent 1 heart attack." Absolute risk reduction and number needed to treat may put "impressive" risk reductions into a more meaningful light. The FDA recommends that providers use absolute risks because patients are unduly influenced by relative risk, leading to suboptimal decisions.[2]

3. Be aware of how robust the results are. Trial results are robust if they are unlikely to change with differences in the

Published online July 29, 2021

study findings. The fragility index is one measure of this. If only a handful of patients in each group having a different outcome would change the statistical significance, one should worry that these results may not be robust. Of course, as Ho points out in her paper that appears in this issue of JGIM[3] and we underscore in our discussion of the economic consequences of sample size calculations above, if in the planning stages of a study, trialists guess the results they will get perfectly, an efficient trial will be right on the cusp of the exact sample size required to achieve statistical significance. For those trials, we should expect a very small fragility index. Indeed, some trials have sequential analyses with stopping rules to achieve exactly this result. Initial studies are carefully designed to show differences. If the study results are significant, but they are sensitive to small changes in patient outcomes, it is possible that the intervention will be found less effective or ineffective in future trials.[4]

4. Be aware of how much loss to follow-up occurred. A simple test is to ask how many patients were lost, compared to the reported benefit. For example, if the difference in outcome was 50%, and there was 5% loss to follow-up, the results are likely robust. However, if there is a 10% difference in outcome, but a 20% loss, there is reason to be concerned. In trials of tricyclics in the prophylaxis of headaches, it is not uncommon for 30–50% of patients to drop out, because of the side effect profile of tricyclics.[5] Patients who stay in the trial are more likely to benefit; otherwise, they would have dropped out because of how difficult the medication is to take. Intention to treat analyses can only partially deal with this problem.

5. How much are the trialists concealing? Most clinical trials collect a number of outcomes. If the methods report collecting 6 outcomes, but only 2 are presented in the results and both of those are significant, one should worry that the other 4 outcomes were not presented because they were not statistically significant.[6] For example, in headache trials, common outcomes include headache frequency, intensity, duration, and medication use. If the study collected all four, but only presents results for headache frequency, it is likely that the intervention had no effect on headache intensity, duration, or medication use. Trialists commonly underreport their outcomes and even change outcomes from primary to secondary.[7] In addition, if a commonly collected or an important outcome is not reported, one should be suspicious that the trialist is withholding.

6. Mind the axis. A common misleading trick is to stretch out the *y*-axis to make differences look larger than they are. With survival curves, look to see when they separate to see how long you will need to treat patient before they begin to see benefit. Also, beneath the *x*-axis is the number of participants still in the trial at that time point. Make sure there are sufficient numbers still in the trial at each time point to be trustable; if not, move back to a time where there are enough patients left. There are a several reasons why the numbers get smaller over time. Patients drop out or they could have the outcome of interest. However, for most trials, the biggest reason for the decline is that patients are enrolled over time. If a trial takes 2 years to enroll and follows patients for 5 years, participants enrolled on day 1 will potentially be followed for 5 years, while those enrolled in the 2nd year will only be in the trial for 3 years.

7. Be wary of artificial cut-points. Sometimes studies present findings that are not intuitively grouped. For example, "In our study the likelihood of response was greater in patients older than 53 than younger (RR: 1.8, 95% CI: 1.03–2.9)." The trialist likely selected 53 as the cut-point, rather than a more meaningful one (such as 65), because that age provided the greatest impact in their study.

8. Do not be fooled by dichotomous outcomes. Dichotomous outcomes can be misleading. If the outcome is rare in the control group, then an OR of 2.0, a doubling of outcome, is still rare, even with the intervention. Look to the baseline rate to assess whether the difference is clinically meaningful. Be suspicious if the trialist doesn't give you this information.

Clinicians are motivated to help patients make the best decisions. A strong evidence base helps make both clinicians and patients feel confident in decisions; they can accurately weigh the potential benefits, costs, and risks. However, the nature of research tends to bias the results towards benefit. Even interventions that are replicated and shown to be beneficial tend to have weaker effects than originally found. Hopefully these quick tips will help busy clinicians decide how much to trust the findings of and conclusions drawn from clinical trials.

Corresponding Author: *Jeffrey L Jackson, MD MPH; Zablocki VAMC, Milwaukee, USA (e-mail: jjackson@mcw.edu).*

REFERENCES

1. **Detsky AS**. Using economic analysis to determine the resource consequences of choices made in planning clinical trials. J Chron Dis. 1985. 38(9):753-765
2. **Fagerlin A** (2011). Chapter 7: Quantitative Information°. In: Fischhoff B, Brewer NT, Downs JS (eds.) Communicating Risks and Benefits: An Evidence-Based User's Guide, 57-61. Annapolis: Food and Drug Administration (FDA). http://www.fda.gov/downloads/AboutFDA/ReportsManualsForms/Reports/UCM268069.pdf
3. **Ho, AK**. The Fragility Index for Assessing the Robustness of the Statistically Significant Results of Experimental Clinical Studies. J Gen Intern Med.
4. **Dumas-Mallet, E. Smith A, Boraud T, Gonon F**. Poor replication validity of biomedical association studies reported by newspapers. 2017. PLoS ONE 12(2): e0172650. doi:https://doi.org/10.1371/journal.pone.0172650

5. **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;32(12):1351-1358.
6. **Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG**. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 5/26/2004 2004;291(20):2457-2465.
7. **Jackson JL, Kuriyama A**. From the Editors' Desk: Bias in Systematic Reviews-Let the Reader Beware. J Gen Intern Med. 2018 Feb;33(2):133-135. doi: https://doi.org/10.1007/s11606-017-4236-2.