Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

COMMENTARY

Rationality and the generalization of randomized controlled trial evidence

Jonathan Fuller BMSc

MD/PhD Student, Faculty of Medicine, University of Toronto, Ontario, Canada

Address at which work was carried out: The Wilson Centre, 200 Elizabeth Street, 1ES-565, Toronto, Ontario, Canada M5G 2C4.

doi:10.1111/jep.12021

The randomized controlled trial (RCT) is the gold standard of evidence because it has the highest internal validity; that is, its design is better than any other at preventing various sources of systematic bias from confounding our judgement. The preceding assertion has been amply rehearsed. It is precisely what is meant by the RCT's placement above all other population study designs in the hierarchy of evidence for prevention and treatment decisions. The hierarchy has been the target of a battery of criticisms over the past two decades [1], yet it lives on in the most recent edition of the *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice* [2], and the randomized trial still rules. Recurrent among critiques of the RCT are concerns over its *external* validity [3–13]. The common thread to these arguments is that trials map poorly onto the reality of clinical medicine; the trial participants are not representative of patients routinely encountered by clinicians, and the ideal and pristine trial conditions are too dissimilar from the concrete and messy world of clinical practice.

Concerns about the representativeness of trial populations are neither exaggerated nor trivial. A systematic review of 283 RCTs published in major medical journals found that 81.3% of trials excluded patients due to common co-morbidities, 38.5% due to older age and 54.1% due to concurrent use of commonly prescribed medications [14]. Medical co-morbidities and concurrent medication use served as exclusion criteria even more often among the subset of drug intervention trials. Yet, most patients with at least one chronic disease have multiple [15–17], and we often use multiple medications to manage multiple chronic diseases [18,19]. As the population ages, older patients, especially those with multiple chronic conditions, are becoming more common [20,21]. Older and multi-morbid patients are already the highest users of prescription medications in the community [22,23]. Against the backdrop of these concerns, a trial's recruitment strategies may further purify the participant pool by selecting particular kinds of patients among those who meet the eligibility criteria, resulting in baseline variables that fall within an even tighter range [3,4]. Thus, we test drugs in one population, only to use them in a very different population.

As medical orthodoxy dictates that treatment decisions are to be based on RCT evidence whenever feasible, it is necessary to ask: *what useful inferences can we make from RCT results?* This question, the question of generalization, is interested not in whether but in *how* we should draw conclusions from RCT efficacy results that are applicable to clinical practice. The importance of this matter cannot be understated. Finding and critically appraising clinical research evidence is a pointless exercise if we do not know how to establish its meaning for patients seen in everyday practice. Without a good approach, we are acting on faith, trusting in the evidence and the omnipotence of our therapeutics.

Over the past several decades, we devoted much energy to generating, reviewing and summarizing evidence. We have given far less attention to the issue of how to thoughtfully apply the evidence once we have it. That's fine if all we care about is that our clinical decisions are evidence-based, but not so good if we also want them to be well-reasoned. Let us not forget that evidence-based medicine (EBM) grew out of an interest in making medicine 'rational' [24], with the idea that rational clinical evaluations should be evidence-based. I agree with the uncontroversial statement that the best decision is supported, at least in part, by the best available evidence. Rationality, however, is constituted by reasoning, not evidence. Complete arguments are necessary for rational evaluations, arguments that begin with general evidence and end in a conclusion about a particular patient. In order to traverse these inferential gaps [25], medicine must address the issue of how to establish, as an intermediate premise, what the evidence has to say about the efficacy of an intervention for particular patients in a particular practice setting.

As Post *et al*. showed [26], relatively few authors have proposed a method for generalizing RCT efficacy results in the medical literature. Post *et al*. did a systematic review of the literature to identify possible strategies, and then grouped the 15 selected records into three categories based on the general approach taken. They considered all three approaches, with special reference to the empirical evidence favouring each, before endorsing one. What do Post *et al*. believe is at issue in generalizing a study's results and what did they find in their review?

For these authors, an approach to generalizing efficacy results must help us to decide whether to believe that the relative treatment effect, once deemed a valid approximation for the study population, would be seen in a certain practice population. The decision to treat is predicated on this prior belief. Post *et al*. dismissed the first general approach they identified, 'Checking eligibility criteria and applying trial findings only when patients one is interested in are sufficiently represented in the trial population' [26], as it is represented by only one publication. Meanwhile, many of the publications representing the second approach they describe include eligibility criteria as but one of several considerations for generalizability. We will return to the second approach momentarily.

The authors characterize approach number 3 as, 'Using observational studies that represent the target population to infer

© 2013 John Wiley & Sons Ltd,

whether trial results are generalizable' [26]. If your patients do not meet the eligibility criteria of the RCT, the strategy here is to generalize the treatment effect measured in the trial *if* efficacy was also demonstrated in an observational study that did include patients like yours. Post *et al*. rejected this approach on the grounds that observational studies are too untrustworthy, too susceptible to bias.

Their preferred approach is the second they identify: 'Checking in- and exclusion criteria and evaluating whether there are compelling reasons why the relative effect found in trial results should not be applied to the patient group' [26]. The EBM community has long advocated this method [27–29], so I will refer to it as 'the EBM approach'. Firstly, decide whether the target population to which you wish to generalize meets the eligibility criteria of the RCT(s). If so, then generalizing is not problematic. If not, then generalizing is usually still not a problem because relative effects are generally consistent across patient populations. On the odd occasion, however, the relative effect does vary, and as a check on generalization, one should search for 'compelling reasons' to reject the assumption that the effect estimate applies here. *Prima facie*, the approach seems reasonable. Let us examine it more closely, along with the central argument that props it up.

The general generalizability thesis

In the spirit of evidence-based argumentation, the authors support their claim to the cross-population stability of relative treatment effects with clinical research findings. Studies have found that the relative risk (RR) typically does not vary significantly among the RCTs included in a given meta-analysis [30,31]. However, within a meta-analysis, the RCTs chosen are often similar in eligibility criteria and in the trial protocol used [5]. These findings only demonstrate the reproducibility of the treatment effect among similar contexts rather than the generalizability of the treatment effect to dissimilar contexts. Another study found that the RR of major cardiovascular events did not vary by age among adults included in RCTs for antihypertensive medications [32]. Here, the evidence is limited to one type of intervention (antihypertensive medication) and one relevant prognostic risk factor (age), and does not support the claim to the nearly universal stability of treatment effects for all interventions and risk groups. Post *et al*. provided counterexamples [33,34] in which the relative effect did in fact vary by age for other interventions.

There is a serious problem with the body of evidence to which the authors point: it demonstrates consistency among RCTs, which as we have seen typically exclude older patients, patients with multiple chronic conditions and patients taking several medications concurrently – the very patients about whom we often wish to make extrapolations! The authors have not made a good case for the argument that relative effects are stable between trial populations and typical target populations. Furthermore, it is not only patient characteristics that can modify treatment effects, but also characteristics of the therapeutic environment. The trial environment often differs systematically from the typical clinical environment in relevant ways. Trial participants receive better-than-standard care and monitoring [12,13,35,36], which can confer benefits and avoid harms relevant to the primary outcome, and perhaps even modify the effect of therapy. Concurrent medication use, ubiquitous among those taking at least one medication, might also modify efficacy through drug interactions. Finally, more complex interventions (surgical, rehabilitative, etc.) might differ at a particular site compared with the trial site as a function of local resources, policies or expertise. Post *et al*. failed to provide convincing evidence that relative treatment effects are usually unmodified in diverse contexts that each constitute unique causal network. Stronger justification is required to hold up their ambitious claim to the general generalizability of RCT efficacy results.

Even if a target population meets the eligibility criteria for a study, we have not ruled out the possibility of systematic and relevant differences. The economic, social and ethnic diversity of trial participants is limited by the demographics of the local source from which a study enrols patients; such variables have been shown to influence response to treatment [37,38]. Recruitment strategies frequently result in there being systematic differences between patients who are recruited and those who are eligible but not recruited [3]. For instance, some trials have a run-in period in which they exclude patients who are unable to follow the study protocol, who experience adverse outcomes or who respond poorly to treatment [39]. Owing to non-random recruitment of eligible patients, the average participant in a trial may not be comparable to the average patient in a target population that satisfies the eligibility criteria. Thus, generalization may not be warranted even if a target population meets the in- and exclusion criteria.

Falsification

In the EBM approach, one evaluates limits to generalization that recognize the uniqueness of different patient groups. Limits are not considered in order to modify the effect estimate but to challenge the assumption that the effect estimate is accurate for the target population. One looks for known reasons to reject their belief in clinical efficacy for the target patients. What I have just described is essentially the rationality of *falsificationism*, famously championed by the philosopher of science Karl Popper as the logic for theory-testing [40]. To borrow Popper's language, one conjectures a theory and then seeks evidence that would refute the theory by demonstrating its falsity. The refuting evidence is incompatible with the truth of the theory, so if we believe the evidence, then we are *compelled* to reject the theory. In the EBM approach to generalization, one conjectures that the RR is accurate for the target population and then seeks to refute this assumption with compelling evidence.

What kind of refuting evidence is acceptable to their approach? Post *et al*. suggested that non-randomized studies cannot get the job done [26]. If a non-randomized study were to disagree with the RCT, even if participants in the non-randomized study were more similar to patients in the target population, the results are not believable enough to block generalization. (If another *randomized* study could be found that better represented the target patients, had equal validity and disagreed with the first study, then the generalizability of the first study would no longer be at issue; ostensibly, one would turn to the second RCT and question its generalizability instead.) The authors provide a typology of compelling reasons that might limit generalizability, borrowed from an earlier EBM paper by Dans *et al*. [28]. Included are 'biologic', 'social and economic' and 'epidemiologic' issues that are certainly worth considering. Dans *et al*. would have us ask whether physiological/

© 2013 John Wiley & Sons Ltd

pathophysiological processes or patient/provider compliance might diminish the treatment response, which is always a possibility when generalizing from the trial to the clinic. However, the standard of 'compellingness' only accepts as refuting evidence strong reasons that are incompatible with the generalization.

Because Post *et al*. did not provide examples, it is difficult to say exactly what reasons they would count as compelling, but the claim to the general generalizability of RCT efficacy results implies that compelling limitations will be rare. By that same claim, all generalization conjectures are automatically well-corroborated before they are even subjected to scrutiny – incidentally, a move that Popper would never accept. Generalization is a default first position in which our belief sits and pauses and, once there, cannot easily be unseated. The serious worry is that the EBM approach could promote unjustified generalizing, or *over-generalization*. Post *et al*. recommend 'that guideline panels deal with the issue of generalizability by accepting that results of randomized trials apply to wide populations unless there is a compelling reason to believe the results would differ substantially' [26]. In a forthcoming article [41], I analyse evidence-based guidelines recommending treatment with some of the most commonly prescribed medications. Indeed, guidelines generalize from RCT evidence to wide patient populations, usually inclusive of all adults with the target diagnosis. Reasoning around limits to generalization of RCT efficacy results is not explicit. Either authentic limits are not found, or they simply are not considered. Even if they are considered, it is difficult to justify recommending treatment for wide populations based on evidence of efficacy in trials enrolling narrow population samples. I do not find this justification particularly 'compelling'.

Crucially, if the reasons favouring the decision to generalize are not compelling, it is arbitrary to insist that the reasons opposing it need be. Used in this manner, the standard of 'compellingness' is a double standard. The only reason Post *et al*. offer in favour of generalizing is a general generalizability thesis that lacks support. If we set aside that thesis, we are apparently left with the RCT result itself, a poor voucher for generalizations [6]. A well-conducted RCT with a positive result provides compelling grounds to conclude that the treatment works in the trial population; but on its own, it provides weak grounds to conclude that the intervention works somewhere else [42,43]. Considering the known, convincing reasons 'why the relative effect found in trial results should not be applied' is not a good enough test to warrant acceptance of the generalization hypothesis. It is equally incumbent on the doctor to consider strong arguments for why the trial results *should* be applied.

Conclusion

The systematic review by Post *et al*. reveals that while there is no current consensus on the best approach to generalization of RCT efficacy results, many authors conceptualize the problem similarly as a matter of deciding whether to import the average value measured in a trial into a particular practice population. Different authors reflect on similar variables that could plausibly modify the treatment effect. The method of falsification emerges as a common theme, but so does the practice of considering non-randomized sources of clinical research evidence. We need more discussion as to what a generalized efficacy proposition should look like, as well as how we should use logic and evidence to scrutinize it. The debate around whether RCTs are 'the best' may be of great interest for clinical research, which must settle on what study designs to use, but can be a distraction to those who seek a philosophy of evidence-informed practice, which must tell us *how* to use RCTs *and other evidence* to inform medical judgement. Rationality is not concerned with the validity of studies but with the validity or strength of arguments.

Although the authors provide an important contribution by taking inventory of what the medical literature has to say regarding methods of generalization, they do not provide an adequate defence of the EBM approach. The approach is based on the assumption that RCT efficacy results are generally generalizable to commonly encountered situations, which is not supported by the empirical evidence the authors provide. The standard for reasons to 'falsify' the generalization hypothesis appears disproportionately high, compared to the sole reason for conjecturing the hypothesis, namely RCT average results. The EBM approach could influence guidelines and doctors to generalize too widely or too often, when more careful reasoning around warrants for generalization might temper their judgement. A more reasonable approach might conceivably lead to fewer judgements that treatments are efficacious in common but understudied patient cohorts. Such an alternative approach should not be seen as espousing therapeutic nihilism but simply therapeutic rationality, which might well offer a dose of therapeutic humility.

Acknowledgements

The author thanks Jacob Stegenga for feedback on an earlier draft of this article. The author is grateful for the support from the McLaughlin Centre, the Canadian Institutes of Health Research and the Wilson Centre for Research in Health Professions Education. The author has no financial interests to disclose.

References

1. Bluhm, R. & Borgerson, K. (2011) Evidence-based medicine. In Handbook of the philosophy of science: volume 16: philosophy of medicine (eds D. M. Gabbay, P. Thagard & J. Woods), pp. 203–238. Amsterdam: Elsevier.
2. Guyatt, G., Rennie, D., Meade, M. O. & Cook, D. J. (2008) Users' guides to the medical literature: essentials of evidence-based clinical practice, 2nd edn. New York: McGraw-Hill Medical.
3. Rothwell, P. M. (2005) External validity of randomised controlled trials: 'to whom do the results of this trial apply?'. *Lancet*, 365 (9453), 82–93.
4. Hampton, J. R. (2002) Size isn't everything. *Statistics in Medicine*, 21 (19), 2807–2814.
5. Bluhm, R. (2007) Clinical trials as nomological machines: implications for evidence-based medicine. In Establishing medical reality: essays in the metaphysics and epistemology of biomedical science (eds H. Kincaid & J. McKitrick), pp. 149–166. Dordrecht: Springer.
6. Cartwright, N. (2007) Are RCTs the gold standard? *BioSocieties*, 2 (1), 11–20.
7. Feinstein, A. R. & Horwitz, R. I. (1997) Problems in the 'evidence' of 'evidence-based medicine'. *American Journal of Medicine*, 103 (6), 529–535.
8. Black, D. (1998) The limitations of evidence. *Perspectives in Biology and Medicine*, 42 (1), 1–7.

© 2013 John Wiley & Sons Ltd

9. Campbell-Scherer, D. (2010) Multimorbidity: a challenge for evidence-based medicine. *Evidence-Based Medicine*, 15 (6), 165–166.
10. Fortin, M., Dionne, J., Pinbo, G. V., Gignac, J., Almirall, J. & Lapointe, L. (2006) Randomized controlled trials: do they have external validity for patients with multiple comorbidities? *Annals of Family Medicine*, 4 (2), 104–108.
11. Tinetti, M. E., Bogardus, S. T. & Agostini, J. V. (2004) Potential pitfalls of disease-specific guidelines for patients with multiple conditions. *New England Journal of Medicine*, 351 (27), 2870–2874.
12. Rawlins, M. (2008) De testimonio: on the evidence for decisions about the use of therapeutic interventions. *Lancet*, 372 (9656), 2152–2161.
13. Black, N. (1996) Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, 312 (7040), 1215–1218.
14. Van Spall, H. G. C., Toren, A., Kiss, A. & Fowler, R. A. (2007) Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *Journal of the American Medical Association*, 297 (11), 1233–1240.
15. Fortin, M., Bravo, G., Hudon, C., Vanasse, A. & Lapointe, L. (2005) Prevalence of multimorbidity among adults seen in family practice. *Annals of Family Medicine*, 3 (3), 223–228.
16. Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S. & Guthrie, B. (2012) Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet*, 380 (9836), 37–43.
17. Salisbury, C., Johnson, L., Purdy, S., Valderas, J. M. & Montgomery, A. A. (2011) Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study. *British Journal of General Practice*, 61 (582), e12–e21.
18. Guthrie, B., Payne, K., Alderson, P., McMurdo, M. E. T. & Mercer, S. W. (2012) Adapting clinical guidelines to take account of multimorbidity. *British Medical Journal*, 345, e6341.
19. Boyd, C. M., Darer, J., Boult, C., Fried, L. P., Boult, L. & Wu, A. W. (2005) Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *Journal of the American Medical Association*, 294 (6), 716–724.
20. Goulding, M. R., Rogers, M. E. & Smith, S. M. (2003) Public health and aging: trends in aging – United States and worldwide. *Journal of the American Medical Association*, 289 (11), 1371–1373.
21. Jadad, A. R., Cabrera, A., Martos, F., Smith, R. & Lyons, R. F. (2010) When people live with multiple chronic diseases: a collaborative approach to an emerging global challenge. Granada: Andalusian School of Public Health.
22. Qato, D. M., Alexander, G. C., Conti, R. M., Johnson, M., Schumm, P. & Lindau, S. T. (2008) Use of prescription and over-the-counter medications and dietary supplements among older adults in the United States. *Journal of the American Medical Association*, 300 (24), 2867–2878.
23. Hanley, G. E. & Morgan, S. (2009) Chronic catastrophes: exploring the concentration and sustained nature of ambulatory prescription drug expenditures in the population of British Columbia, Canada. *Social Science & Medicine*, 68 (5), 919–924.
24. Sackett, D. L. & Rennie, D. (1992) The science of the art of the clinical examination. *Journal of the American Medical Association*, 267 (19), 2650–2652.
25. Upshur, R. E. G. (2005) Looking for rules in a world of exceptions: reflections on evidence-based practice. *Perspectives in Biology and Medicine*, 48 (4), 477–489.
26. Post, P. N., de Beer, H. & Guyatt, G. H. (2013) How to generalize efficacy results of randomized trials: recommendations based on a systematic review of possible approaches. *Journal of Evaluation in Clinical Practice*, 19 (4), 638–643.
27. Oxman, A. D., Cook, D. J. & Guyatt, G. H. (1994) Users' guides to the medical literature. VI. How to use an overview. *Journal of the American Medical Association*, 272 (17), 1367–1371.
28. Dans, A. L., Dans, L. F., Guyatt, G. H. & Richardson, S. (1998) Users' guides to the medical literature. XIV. How to decide on the applicability of clinical trial results to your patient. *Journal of the American Medical Association*, 279 (7), 545–549.
29. Guyatt, G. & Rennie, D. (2001) User's guide to the medical literature. Chicago, IL: AMA Press.
30. Furukawa, T. A., Guyatt, G. H. & Griffith, L. E. (2002) Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. *International Journal of Epidemiology*, 31 (1), 72–76.
31. Deeks, J. J. (2002) Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, 21 (11), 1575–1600.
32. Turnbull, F., Neal, B., Ninomiya, T., *et al*. (2008) Effects of different regimens to lower blood pressure on major cardiovascular events in older and younger people: meta-analysis of randomised trials. *British Medical Journal*, 336 (7653), 1121–1123.
33. Hlatky, M. A., Boothroyd, D. B., Bravata, D. M., *et al*. (2009) Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. *Lancet*, 373 (9670), 1190–1197.
34. Pignon, J., P., le Maitre, A., Maillard, E. & Bourhis, J. (2009) Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17,346 patients. *Radiotherapy and Oncology*, 92 (1), 4–14.
35. Weiss, N. S., Koepsell, T. D. & Psaty, B. M. (2008) Generalizability of the results of randomized trials. *Archives of Internal Medicine*, 168 (2), 133–135.
36. Dekkers, O. M., von Elm, E., Algra, A., Romijn, J. A. & Vandenbroucke, J. P. (2010) How to assess the external validity of therapeutic trials: a conceptual approach. *International Journal of Epidemiology*, 39 (1), 89–94.
37. Fine, P. E. M. (1995) Variation in protection by BCG: implications of and for heterologous immunity. *Lancet*, 346 (8986), 1339–1345.
38. Materson, B. J., Reda, D. J., Cushman, W. C., *et al*. (1993) Single-drug therapy for hypertension in men: a comparison of 6 antihypertensive agents with placebo. *New England Journal of Medicine*, 328 (13), 914–921.
39. Pablos-Mendez, A., Barr, R. G. & Shea, S. (1998) Run-in periods in randomized trials: implications for the application of results in clinical practice. *Journal of the American Medical Association*, 279 (3), 222–225.
40. Popper, K. R. (1963) Conjectures and refutations. London: Routledge & Kegan Paul.
41. Fuller, J. (2013) Argumentation and rhetoric: how clinical practice guidelines think. *Journal of Evaluation in Clinical Practice*, 19 (3), 433–441.
42. Cartwright, N. (2011) A philosopher's view of the long road from RCTs to effectiveness. *Lancet*, 377 (9775), 1400–1401.
43. Cartwright, N. & Munro, E. (2010) The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice*, 16 (2), 260–266.

© 2013 John Wiley & Sons Ltd