Conventional evaluations of improvement interventions: more trials or just more tribulations?

Kaveh G Shojania*

Department of Medicine, Sunnybrook Health Sciences Center and the University of Toronto Centre for Quality Improvement and Patient Safety, Toronto, Ontario, Canada

Correspondence to
Dr Kaveh G Shojania, Sunnybrook Health Sciences Centre, Room H468, 2075 Bayview Avenue, Toronto, ON, Canada M4N 3M5; kaveh.shojania@sunnybrook.ca

*Editor, BMJ Quality and Safety

Linked

▶ http://dx.doi.org/10.1136/bmjqs-2012-001244

To cite: Shojania KG. *BMJ Qual Saf* Published Online First: [*please include* Day Month Year] doi:10.1136/bmjqs-2013-002377

Debates over the degree to which standards of evidence and methods from traditional clinical research can or should apply to quality improvement (QI) have recurred over the past 10 years.[1–4] When, if ever, do we need a randomised controlled trial (RCT) demonstrating benefit to decide that an intervention has worked? Can we recommend QI interventions for widespread adoption even without supportive RCTs? On one side of the debate, some have argued that QI and the RCT are like oil and water—never the twain shall mix. Certainly, many have argued, we should not presume that RCTs represent the gold standard for evidence in QI.

On the face of it, the report by Mate *et al*[5] supports this oil and water view of RCTs and QI interventions. The authors report their struggles conducting a pragmatic, multisite RCT of a complex intervention to reduce perinatal transmission of HIV in KwaZulu-Natal Province, South Africa. The intervention included socioadaptive strategies,[6 7] such as engaging local health system leaders, securing a commitment to the aims of the project, and providing participating health centres with the tools to perform data-driven improvement cycles. It also promoted specific best practices for key steps in the prevention of perinatal transmission of HIV (eg, increasing the proportion of women receiving early antenatal care that includes HIV counselling and testing, increasing the proportion of mothers with low CD4 counts who receive treatment, and so on). The authors initially planned to evaluate this complex intervention using an equally complex study design—a step-wedge, cluster RCT involving 48 clusters of clinics (for a total of 222 individual clinics) in three waves of intervention and control sites; hence, the 'step-wedge' label.

It will come as no surprise to most readers that this double dose of complexity—from the intervention itself and the trial design—overwhelmed all parties involved. Citing frustration among participants, tensions between implementation ideas at different levels of the health system and other very legitimate sounding factors, investigators abandoned the planned RCT. They opted instead for a mixed-methods evaluation that included qualitative analysis and time series, with statistical process control charts and interrupted time series methods. (The initially intended intervention itself also underwent simplification.)

The need to abandon this trial seems particularly supportive of the oil and water view of RCTs and QI, since the planned trial contained two pragmatic characteristics generally believed to facilitate the evaluation of QI interventions. Randomisation occurred at the level of clinics (hence, 'cluster RCT') rather than individual patients. And, the intervention occurred in waves, accommodating the reality of organisations having differing timelines for implementing new interventions.[8] This step-wedge design has found successful applications in other QI settings, including an evaluation of rapid response teams[9] and the central line bundle to reduce catheter-associated bacteraemia.[10]

So, what does this case illustrate: are some interventions simply too complex for RCTs? Do low-resource settings make RCTs impractical? Or, perhaps the problem is context.[11 12] Maybe important variations in local context prohibit evaluation using a randomised design.

The first two questions are easy to address, since multiple examples of complex RCTs in low-resource settings

exist. In one example, the intervention's complexity rivals that seen in any QI study: investigators randomised poor, rural villages to an intervention involving microfinance loans, education about gender equity issues and an HIV training curriculum.[13] The trial showed mixed effects, with no reduction in HIV incidence but a 55% reduction in intimate partner violence. Thus, the cluster randomised design did not undermine the intervention even in this explicitly low-resource setting (in fact, in South Africa, like the intervention reported by Mate *et al*[5]). Another recent example involved cluster randomisation of 98 healthcare zones in Ghana to evaluate the impact of a complex intervention involving training of community-based surveillance volunteers.[14] These volunteers identified pregnant women in their community in order to make two home visits during pregnancy and three in the first week postpartum to promote key newborn care practices, assess babies for concerning signs and make referrals to formal healthcare services as necessary. While the trial achieved only a small, non-significant reduction in neonatal mortality, it achieved multiple significant improvements in the uptake of key practices.

Concerns about context as a barrier to RCTs are also relatively straightforward to address. Despite much emphasis on context as a unique consideration in QI, the challenges of accounting for important variations in trial participants exist routinely confront clinical trialists. As Bond *et al*[15] replied to a critique of the Medical Research Council framework[16] for the evaluation of complex interventions:

> Imagine an intervention whose effects vary within and between individuals and depend on subtle interactions between deliverers and recipients, and in which exposure is uncertain. Given this complexity, who would contemplate conducting a randomised controlled trial? In fact, all these issues must be dealt with in drug or therapeutic trials, as well as in more obviously complex interventions.[15]

The situation with contextual factors in QI thus resembles that seen for important patient characteristics in conventional clinical trials. Variations in clinical characteristics such as comorbid conditions, different genetic predispositions and socioeconomic factors may greatly impact the effects of a medication. One hopes that randomisation balances out the factors we do not yet know about and we can always choose to stratify the randomisation according to key factors we do know about. In fact, conventional clinical trials must also sometimes deal with what amount to contextual factors, although clinical trialists tend to call them 'centre effects'. Some centres have greater expertise in delivering the treatment of interest (eg, a new surgical procedure), have relevant infrastructure such as relevant consulting services or attract patient populations with characteristics not easily taken into account with simple risk adjustment.

While clinical trialists do not account for centre effects as often as they should,[17] the point remains that conventional clinical trials face the same types of variations—across patients and centres—that we in QI ascribe to context.[11 12 18 19] If an intervention might plausibly have different effects in settings with certain leadership styles, levels of patient safety culture, sophistication of clinical informatics infrastructure or any of the numerous other possible elements of context,[11] investigators can explicitly balance the study groups for some factors (perhaps the ones they can best measure) and let randomisation take care of the rest.

That said, when we really believe that contextual factors play important roles in an intervention's effectiveness, it may be premature to conduct an RCT. Better to first identify the key contextual factors, so that modifiable ones can be addressed as part of the intervention and unmodifiable ones can be avoided through the trial's exclusion criteria. ('Organisations that did not have such-and-such features in place were not invited to participate.')

Once we establish that a given intervention requires a certain safety culture, concrete commitments by leadership, frontline engagement or various other contextual factors, investigators can address these factors in the implementation plan. If we forgo characterising these factors and rush to disseminating the intervention widely (as has often occurred), then we have to expect that the intervention will not work in many places. An RCT would undoubtedly show this. The problem, then, lies not with the RCT, but with the developmental stage of the intervention—dissemination occurred too soon.

TOO COMPLEX VERSUS TOO SOON

Rather than inferring from the report by Mate *et al*[5] that RCTs present insurmountable challenges for complex interventions or low-resource settings, one could argue that the attempted RCT simply occurred too soon. Drug studies occur in stages, progressing from in vitro studies of basic pharmacodynamics and small pilot studies through to dose-finding studies before moving on to efficacy trials that compare one or two specific doses with a placebo in order to assess benefits and harms. Similarly, QI interventions require development in stages, with many interventions, especially complex ones, requiring numerous pilot and quasi-experimental studies before an RCT becomes appropriate. The UK Medical Research Council has provided guidance on the evaluative stages in the evolution of complex interventions in the form of several frameworks that have themselves evolved over time.[16 20 21]

Well before the current interest in QI, early leaders in clinical research articulated the challenge of evolving interventions and other 'contraindications' for the conduct of RCTs.[22] Some present purely practical

challenges (relative contraindications, so to speak) and can be overcome with sufficient resources to recruit more patients or extend the observation period. These practical challenges apply to many primary prevention trials in public health settings (requiring large numbers of patients and observation periods) and interventions where several alternate treatments already exist (requiring multiple arm trials). However, other situations present more fundamental challenges to the conduct of an RCT.

An intervention that is unstable—in the sense of still evolving—constitutes precisely such a challenge that makes an RCT 'contra-indicated'.[22] In clinical research, this situation arises most often with interventions involving procedures and medical devices. A promising surgical technique has shown benefit in observational studies, but the details of the technique continue to undergo refinement. Conducting an RCT prematurely will serve little purpose. If negative, proponents of the procedure will respond: "This trial was designed and initiated X years ago, before we recognized the importance of including such-and-such in the treatment." Similarly with diagnostic imaging technologies: advocates will argue that the resolution (and thus, diagnostic utility) of the images generated by machines of today so far surpass the ones of several years ago that the null result of the RCT does not apply to current practice.

A QI intervention that has only been sketched out in broad terms and still needs to undergo iterative refinement clearly represents an unstable intervention. If an RCT occurs before the rapid cycle improvement process of optimizing concrete details of the intervention and the implementation strategy, then the trial amounts to a test of the general iterative improvement cycle approach, not any more specific intervention. Evaluating the Plan Do Study Act (PDSA) approach to improvement with an RCT has some merit. But, then the burden of 'maturity' or completeness for the intervention then shifts from the specifics of how best to prevent perinatal HIV transmission (in a low-resource setting) to the details of how best to engage participants in such an endeavour, how best to train them in the methods of improvement and provide mentorship for them (in a low-resource setting).

We have ideas and somewhat informed recommendations about these details, but none so empirically established that we feel confident that a negative trial demonstrates that the method does not work. In response to a negative RCT, we would simply say "They should have provided a more intensive training, more frequent contact with improvement coaches etc." Either way—viewed as a specific approach reducing perinatal HIV transmission or a more general strategy for supporting the development of locally tailored solutions to a shared improvement goal—the intervention undertaken by Mate *et al*[5] seems to have still been under evolution and not yet ready for an RCT.

WHEN ORGANISATIONS BECOME THE PATIENTS

Many of the reasons commonly cited for not wanting to conduct an RCT (including some of those reported by Mate *et al*[5]) resemble the reasons patients hesitate to participate in clinical trials. Many patients understandably want to end up in the intervention group so they can receive a potentially beneficial treatment for their cancer, chronic pain, progressive dementia and so on. Certainty, no patient relishes having to undergo frequent blood work or other periodic assessments if they are just receiving a placebo. With due respect to Mary Poppins, a spoonful of sugar may make the medicine go down, but most patients want more than just the sugar. They want the shot at improvement that the active medication offers, not just the contribution to science that their receiving a sugar pill (ie, placebo) will bring.

Yet, the biomedical community advocates that patients accept the need for randomisation. "We really don't know if this new treatment will work. The only way we can know is if some patients receive placebo." Maybe we need to do a better job in QI of telling ourselves just these sorts of things. A particular hospital wants to be in the intervention group because its staff or senior management are very enthusiastic about reducing surgical complications, shortening emergency department wait times, lowering hospital readmission rates or whatever the case may be. They do not want to delay implementing the study intervention. They certainty do not want to collect monthly data if they will just be in the control group. How does this differ from patients' wishes to receive the active treatments in a clinical trial and not undergo repeated clinical or laboratory assessments if they are just receiving placebo?

CONCLUSION

Even those who would like to see more RCTs conducted in QI (myself included) recognise the need for well-executed and well-reported improvement work using non-randomised designs. Even if we never had any RCTs, we could make a lot of progress just by avoiding simple before–after studies: "Last year our audit showed X% compliance with the given bundle, checklist or guideline; this year the audit shows Y%; $p=0.05$."

As many have advocated, using run charts or statistical process control,[25] often applied in conjunction with iterative cycles of improvement (eg, the plan–do–study–act model) provides a very informative and robust approach both to developing and evaluating an intervention. Too often those who come from a clinical research tradition have skipped over this approach and rushed to an RCT before the intervention was mature, before answering such basic questions as: was the decision support intervention sufficiently user-friendly? Did participants look at the audit and feedback reports they received? Could the case managers

reach patients by phone? Applying a rigorous evaluative design before optimising the intervention serves no purpose. Yet, RCTs still have an important role in the evaluation of complex interventions, especially when we want to advocate for their widespread implementation (eg, as with surgical checklists, medication reconciliation and rapid response teams, to name just a few examples). Without an RCT, we can have no idea what effects the intervention has in a range of institutions.

Many of the challenges we regard as unique to QI in fact exist in clinical research and have been recognised for decades.[22] In some cases, we need to choose the right time for an RCT (once the intervention is sufficiently mature). In other cases, we may need to adopt alternative designs, such as step-wedge randomisation, to accommodate the realities of implementing complex interventions in the midst of other institutional activities or adaptive randomisation to minimise the number of sites assigned to the control group. But, we also have to remember that many of the misgivings we feel as providers of healthcare asked to participate in RCTs of improvement interventions echo those made by patients all the time. Our enthusiasm for assignment to active treatment carries no more weight than theirs. Rather than so often avoiding multi-site RCTs in QI, we may just need to find the right spoonful of sugar for ourselves when we end up in the control group.

Acknowledgements Dr Shojania receives salary support from the University of Toronto Faculty of Medicine and Sunnybrook Health Sciences Centre. He also gratefully acknowledges the comments of Dr David Stevens and Dr Sanjay Saint on an earlier draft of this editorial.

Competing interests None.

Provenance and peer review Not commissioned; internally peer reviewed.

REFERENCES

1 Leape LL, Berwick DM, Bates DW. What practices will most improve safety? Evidence-based medicine meets patient safety. *JAMA* 2002;288:501–7.
2 Shojania KG, Duncan BW, McDonald KM, *et al*. Safe but sound: patient safety meets evidence-based medicine. *JAMA* 2002;288:508–13.
3 Berwick DM. The science of improvement. *JAMA* 2008;299:1182–4.
4 Auerbach AD, Landefeld CS, Shojania KG. The tension between needing to improve care and knowing how to do it. *N Engl J Med* 2007;357:608–13.
5 Mate KS, Ngidi WH, Reddy J, *et al*. A case report of evaluating a large-scale health systems improvement project in an uncontrolled setting: a quality improvement initiative in KwaZulu-Natal, South Africa. *BMJ Qual Saf*. Published Online First: 30 Nov 2012 doi:10.1136/bmjqs-2012-001244
6 Bosk CL, Dixon-Woods M, Goeschel CA, *et al*. Reality check for checklists. *Lancet* 2009;374:444–5.
7 Pronovost PJ. Navigating adaptive challenges in quality improvement. *BMJ Qual Saf* 2011;20:560–3.
8 Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Medical Research Methodolog* 2006;6:54.
9 Priestley G, Watson W, Rashidian A, *et al*. Introducing Critical Care Outreach: a ward-randomised trial of phased introduction in a general hospital. *Intensive Care Med* 2004;30:1398–404.
10 Bion J, Richardson A, Hibbert P, *et al*. 'Matching Michigan': a 2-year stepped interventional programme to minimise central venous catheter-blood stream infections in intensive care units in England. *BMJ Qual Saf* 2013;22:110–23.
11 Stevens DP, Shojania KG. Tell me about the context, and more. *BMJ Qual Saf* 2011;20:557–9.
12 Kaplan HC, Provost LP, Froehle CM, *et al*. The Model for Understanding Success in Quality (MUSIQ): building a theory of context in healthcare quality improvement. *BMJ Qual Saf*. 2012;21:13–20.
13 Pronyk PM, Hargreaves JR, Kim JC, *et al*. Effect of a structural intervention for the prevention of intimate-partner violence and HIV in rural South Africa: a cluster randomised trial. *Lancet* 2006;368:1973–83.
14 Kirkwood BR, Manu A, ten Asbroek AH, *et al*. Effect of the Newhints home-visits intervention on neonatal mortality rate and care practices in Ghana: a cluster randomised controlled trial. *Lancet* 2013;381:2184–92.
15 Bond L, Craig P, Egan M, *et al*. Evaluating complex interventions. Health improvement programmes: really too complex to evaluate? *BMJ* 2010;340:c1332.
16 Craig P, Dieppe P, Macintyre S, *et al*. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008;337:a1655.
17 Tangri N, Kitsios GD, Su SH, *et al*. Accounting for center effects in multicenter trials. *Epidemiology* 2010;21:912–13.
18 Taylor SL, Dy S, Foy R, *et al*. What context features might be important determinants of the effectiveness of patient safety practice interventions? *BMJ Qual Saf* 2011;20:611–17.
19 Ovretveit JC, Shekelle PG, Dy SM, *et al*. How does context affect interventions to improve patient safety? An assessment of evidence from studies of five patient safety practices and proposals for research. *BMJ Qual Saf* 2011;20:604–10.
20 Campbell M, Fitzpatrick R, Haines A, *et al*. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694–6.
21 Campbell NC, Murray E, Darbyshire J, *et al*. Designing and evaluating complex interventions to improve health care. *BMJ* 2007;334:455–9.
22 Feinstein AR. An additional basic science for clinical medicine: II. The limitations of randomized trials. *Ann Intern Med* 1983;99:544–50.
23 Liang Y, Carriere KC. Stratified and randomized play-the-winner rule. *Stat Methods Med Res* 2008;17:581–93.
24 Rosenberger WF, Huc F. Maximizing power and minimizing treatment failures in clinical trials. *Clin Trials* 2004;1:141–7.
25 Perla RJ, Provost LP, Murray SK. The run= chart: a simple analytical tool for learning from variation in healthcare processes. *BMJ Qual Saf* 2011;20:46–51.