Rigorous evaluations of evolving interventions: can we have our cake and eat it too?

Robert E Burke,[1] Kaveh G Shojania[2]

[1]Research and Hospital Medicine Sections, Denver VA Medical Center, Denver, Colorado, USA
[2]Department of Medicine, Sunnybrook Health Sciences Centre and the University of Toronto, Toronto, Ontario, Canada

Correspondence to
Dr Robert E Burke, Research and Hospital Medicine Sections, Denver VA Medical Center, Denver, CO 80220, USA; Robert.Burke5@va.gov

Accepted 29 January 2018
Published Online First 9 February 2018

► http://dx.doi.org/10.1136/bmjqs-2017-007342

To cite: Burke RE, Shojania KG. *BMJ Qual Saf* 2018;**27**:251–254.

The years immediately following the widespread interest in patient safety[1] and then healthcare quality[2] saw considerable debate between pragmatically oriented improvers and research-oriented evaluators[3–6] —or between 'evangelists' and 'snails' as one longtime observer characterised the two groups.[7] Too often, enthusiastic improvers ('evangelists') relied on simple pre-post designs within a single context leading to erroneous claims of efficacy.[8] In contrast, research-oriented investigators ('snails') and journals pushed for ever more rigorous designs including randomised trials, potentially at the cost of discouraging many improvers without this training and leading to slower development and deployment of effective interventions.[9 10] Many clinicians, quality improvement (QI) experts and researchers are thus caught in a quandary: how best to evaluate a candidate QI intervention? How can we best balance the pragmatic needs of improvement—including the frequent need to refine the intervention or its implementation—with the requirement of most traditional evaluative designs, which typically require a static intervention?

We believe this question is one of the most important issues to consider when developing a QI intervention and is often not considered carefully enough—either by snails or evangelists. Decisions about when and how to evaluate potentially promising interventions can have crucial implications for the future of the intervention and the patients it could affect.

TWO RECENT EXAMPLES OF IMPROVEMENT INTERVENTIONS EVALUATED USING TRADITIONAL DESIGNS

In this issue of *BMJ Quality and Safety*, Swaminathan and colleagues[11] present a rigorous evaluation of the Michigan Appropriateness Guide for Intravenous Catheters (MAGIC) QI intervention, intended to reduce adverse events stemming from the insertion of peripherally inserted venous central catheters (PICC). PICCs have become ubiquitous as a substitution for a central intravenous line when patients need longer term central intravenous access, but clinicians often order them unnecessarily or order inappropriate types—for example, a double-lumen PICC when a single-lumen PICC would work just as well and carry a lower risk of complications. The authors implemented MAGIC at a single intervention hospital and used data from nine contemporaneous controls drawn from a QI collaborative in the state of Michigan (all 10 sites participate in the collaborative).

The MAGIC intervention included computerised decision support at the time of ordering and a much larger role for PICC nurses to regulate appropriate PICC placement. Training was also delivered for PICC nurses and ordering providers. Outcomes included rates of inappropriate PICC use and device-related adverse events. The intervention achieved a statistically significant but relatively small decrease in the rate of inappropriate PICC use at the intervention site after adjustment for measurable potential confounders (incidence rate ratio 0.86; 95% CI 0.74 to 0.99, P=0.048). Fewer adverse events occurred at the intervention hospital, but this reduction largely reflected fewer catheter occlusions. Rates of venous thrombosis and infection rates remained unchanged, though prior work by the authors has shown low rates for both of these complications (5.2% and 1.1%, respectively, for thrombosis and infection).[12]

Some might characterise these results as disappointing—bordering on a 'negative trial'. The authors (understandably) regard the intervention as having achieved some success and probably hope to refine the intervention further and test it in other hospitals in this collaborative. We do not seek to debate this point. Our interest here lies in discussing the tension in QI between the need to refine interventions, especially early in their development, and the desire to conduct rigorous, compelling evaluations to demonstrate their impact.

Consider a second example, in which Westbrook and colleagues evaluated a bundled intervention to reduce nurse interruptions during medication administration using a cluster randomised controlled trial (RCT).[13] For every 100 medication administrations, nurses on intervention wards experienced 15 fewer non-medication-related interruptions compared with control wards. Using results from their previous work on the risk of adverse drug events with interruptions during medication administration,[14] the authors themselves acknowledged that the observed reduction in interruptions would likely achieve little benefit for patients. Moreover, the nurses hated wearing the 'do not interrupt' vests, which constituted a core feature of the intervention.

WHY THESE TWO EXAMPLES?

What both interventions share, in addition to their small to modest impacts, is the use of rigorous, traditional evaluation paradigms—one an interrupted time series combined with contemporaneous controls (about as rigorous a non-randomised design as possible) and the other a cluster RCT. Yet, the disappointing effect sizes raise the question: did these interventions need further refinement before subjecting them to rigorous evaluation?

In the case of the MAGIC, the authors had a reasonable idea for an intervention, but much less prior work to inform the precise ingredients or implementation. In the example of the cluster RCT of a bundled 'do not interrupt' intervention, substantial prior work (not just by these authors) had explored this type of intervention. Thus, the investigators did not plan modifications to the intervention or its implementation strategy. Consequently, it made sense to randomise wards to a fixed intervention or to usual care and focus on evaluating the impact. That said, it obviously occurred to Westbrook *et al*[13] that the nurses might not like wearing the vests, since they solicited this feedback as part of their results. Thus, they might have anticipated the need for making some changes to the intervention.

We recognise that hindsight is 20–20, and that the increasing use of controlled before-and-after studies, interrupted time series, and RCTs to evaluate improvement interventions represent a welcome advance compared to the simple before-after study, which has nothing to recommend it yet remains woefully common. On the other hand, we question the degree to which these traditional designs provide the appropriate balance between rigour and the need to refine interventions, since these evaluative designs presume a 'fixed' or unchanging intervention. None offer an obvious way for investigators to modify the intervention in response to implementation challenges or a disappointing effect size. More adaptive versions of these traditional evaluative designs do in fact exist, and we believe are underused.

HAVING OUR CAKE AND EATING IT, TOO: RIGOROUS EVALUATIVE DESIGNS WHICH PERMIT REFINEMENTS TO THE INTERVENTION

Plan-Do-Study-Act (PDSA) cycles offer the most familiar and accessible way to iteratively improve an intervention, but this approach has two main challenges: (1) it requires specialised expertise and effort to do well,[15–17] and (2) improvers may struggle to successfully pair this with a rigorous evaluative design. However, pairing high-quality PDSA cycles with run charts or statistical process control (SPC) charts can be a powerful and rigorous approach.[18 19] We have published many studies using SPC, but a particularly compelling example involved an 'electronic physiological surveillance system' implemented in two hospitals in England.[20] Nurses recorded patients' vital signs using a handheld device, which then algorithmically determined if another set of vitals should be done sooner, whether care should be escalated to a Rapid Response Team, and displayed the required timing of the response automatically on the same handheld device. Understandably, given the complexity of the intervention, the investigators permitted flexibility in aspects of the roll-out at the two hospitals with built-in cycles for improving fidelity to the intervention. Importantly, the authors could track how often the intervention was used at each site. This allowed them to powerfully demonstrate statistically significant decreases in mortality rates *that occurred in temporal association with successful deployment of the intervention*.[20]

For larger multisite studies, 'stepped wedge' designs offer an appealing option[21 22] that will seem familiar to those accustomed to PDSA cycles. In their most rigorous form, they are cluster randomised trials, except that randomisation determines not whether a site receives the intervention, but rather when. The total number of sites is randomised into sequential cohorts, with all cohorts eventually implementing the intervention and each providing control data in the meantime. This design preserves the benefits of randomisation and also allows investigators to identify implementation challenges during the first cohort, and make adjustments if needed before moving on to the second cohort. Stepped-wedge trials can also increase the power to detect differences in smaller samples, since each site provides both control and intervention

data, with no need to divide the total N into mutually exclusive intervention and control sites. We have also published so-called 'naturalistic' stepped-wedge studies,[23 24] which potentially introduce bias due to the lack of randomisation (eg, if the sites most eager to implement and best poised to succeed make up the first cohort), but still represent a reasonably rigorous design.

Consider how MAGIC might have benefited from using a stepped-wedge design (randomised or not). Instead of roughly 1000 intervention patients and 6000 controls, in the stepped wedge, the study could be thought of having 7000 controls and 7000 intervention patients—substantially strengthening their power to find important differences in patient outcomes that might be of most interest but have low baseline prevalence rates (such as thrombosis and infection). Similarly, the authors could build in time to evaluate their intervention at different clinical sites, deriving important insights into how to make the intervention generalisable beyond their single intervention hospital. All the hospital sites were already part of the collaborative and collecting PICC-related data; the additional effort may have been small relative to the real benefits in terms of rigour of the evaluation.

For interventions that are complex or multicomponent, where the greatest risk consists of unsuccessful implementation, an important alternative is an adaptive or Sequential Multiple Assignment Randomized Trial 'SMART' design.[25] In this randomised design, intervention sites or clusters that do not successfully implement the health system intervention after a certain period of time are again randomised—to receive a different method of support for implementation of the intervention. Sites or units that successfully implement the intervention at the outset do not receive any additional intervention. The timing and types of support can be prespecified, and again offer insight into generalisability of the intervention or key ingredients to successful implementation at different study sites. Rather than waiting until the end of the trial to discover some sites had struggled to implement the intervention, much less evaluate its effectiveness, steps built into the trial can 'catch' struggling sites and intervene to support them in successfully implementing the intervention—which is, after all, what is needed to evaluate the intervention's value. This can be thought of having the effect of reducing the difference between 'intention to treat' and 'per-protocol' analyses for complex health system interventions.

This 'SMART' design might well have benefited the evaluation of the intervention reported in our second example by Westbrook *et al*.[13] Their 'do not interrupt' bundle consisted of multiple elements: nurses wearing the 'do not interrupt' vest; interactive workshops with nurses about the intervention; standardised education sessions with clinical staff as well as patients to remind them not to interrupt the nurses except for serious or urgent concerns; and the use of reminders posted on the unit.[13] While there was high fidelity for wearing the vest among nurses on the intervention units, adherence to other aspects of the bundled intervention was not measured or not reported. One can easily imagine low adherence to or impact from these other elements of the intervention, given conflicting priorities on busy inpatient wards. The authors could have built in times where fidelity to the bundled intervention was measured on intervention wards, and wards with low fidelity received additional support to make sure the intervention was deployed as intended, allowing us to understand its value and gaining important information along the way about how best to implement it. Westbrook *et al* are not alone in struggling with this issue, which is surprisingly common among multicomponent QI interventions that are subsequently tested in large cluster RCTs.[26]

Adaptive study designs provide rigour in evaluation, allow flexibility to modify the intervention and demand that improvers stop at intervals to measure progress. This does not just provide critically important context for generalising to other systems, but also allows evaluation of the intervention in its optimal form (ie, at its maximum effectiveness)—and if the benefit remains disappointingly small, allows improvers to move on and pursue other paths.

The rigour by which QI interventions are evaluated is increasing, with journals increasingly pushing for randomised designs in order to be considered for publication.[10] This debate has gone on for some time, with broad agreement that evaluations such as simple before-after studies are not sufficiently rigorous,[9] especially retrospective ones, where the evaluation has clearly come as an afterthought. However, we believe this focus on randomisation misses a critical point: that QI interventions often require some refinement—of the intervention itself or the implementation strategy—and thus need flexible, adaptive evaluations. PDSA cycles paired with SPC charts, stepped-wedge designs and SMART trials can all provide rigorous evaluations that allow for such flexibility. Moreover, the refinements allowed by these designs provide valuable lessons learnt which will assist in further dissemination of effective interventions. The balance between flexibility and rigour provided by adaptive designs may provide the best balance between 'snails' and 'evangelists' and more rapidly identify and deploy effective system-level interventions, which remain in sadly short supply.

Contributor REB and KGS both contributed to the conception of the paper; they both critically read and modified subsequent drafts and approved the final version. They are both editors at *BMJ Quality & Safety*.

Funding Dr. Burke is supported by a VA Career Development Award. Views expressed are not necessarily those of the US Department of Veterans Affairs.

Competing interests None declared.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES

1 Institute of Medicine. *To err is human: building a safer health system*. Washington, DC: The National Academies Press, 2000.
2 Institute of Medicine. *Crossing the quality chasm: a new health system for the 21st century*. Washington, DC: The National Academies Press, 2001.
3 Shojania KG, Duncan BW, McDonald KM, *et al*. Safe but sound: patient safety meets evidence-based medicine. *JAMA* 2002;288:508–13.
4 Leape LL, Berwick DM, Bates DW. What practices will most improve safety? Evidence-based medicine meets patient safety. *JAMA* 2002;288:501–7.
5 Berwick DM. The science of improvement. *JAMA* 2008;299:1182–4.
6 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.
7 Davidoff F. Systems of service: reflections on the moral foundations of improvement. *BMJ Qual Saf* 2011;20(Suppl 1):i5–10.
8 Shojania KG, Grimshaw JM. Evidence-based quality improvement: the state of the science. *Health Aff* 2005;24:138–50.
9 Shojania KG. Conventional evaluations of improvement interventions: more trials or just more tribulations? *BMJ Qual Saf* 2013;22:881–4.
10 Grady D, Redberg RF, O'Malley PG. Quality improvement for quality improvement studies. *JAMA Intern Med* 2017.
11 Swaminathan L, Flanders S, Rogers M, *et al*. Improving PICC use and outcomes in hospitalised patients: an interrupted time series study using MAGIC criteria. *BMJ Qual Saf* 2018;27:271–8.
12 Chopra V, Smith S, Swaminathan L, *et al*. Variations in peripherally inserted central catheter use and outcomes in michigan hospitals. *JAMA Intern Med* 2016;176:548–51.
13 Westbrook JI, Li L, Hooper TD, *et al*. Effectiveness of a 'Do not interrupt' bundled intervention to reduce interruptions during medication administration: a cluster randomised controlled feasibility study. *BMJ Qual Saf* 2017;26:734–42.
14 Westbrook JI, Woods A, Rob MI, *et al*. Association of interruptions with an increased risk and severity of medication administration errors. *Arch Intern Med* 2010;170:683–90.
15 Taylor MJ, McNicholas C, Nicolay C, *et al*. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. *BMJ Qual Saf* 2014;23:290–8.
16 Leis JA, Shojania KG. A primer on PDSA: executing plan-do-study-act cycles in practice, not just in name. *BMJ Qual Saf* 2017;26:572–7.
17 Reed JE, Card AJ. The problem with plan-do-study-act cycles. *BMJ Qual Saf* 2016;25:147–52.
18 Fretheim A, Tomic O. Statistical process control and interrupted time series: a golden opportunity for impact evaluation in quality improvement. *BMJ Qual Saf* 2015;24:748–52.
19 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.
20 Schmidt PE, Meredith P, Prytherch DR, *et al*. Impact of introducing an electronic physiological surveillance system on hospital mortality. *BMJ Qual Saf* 2015;24:176–7.
21 Hemming K, Haines TP, Chilton PJ, *et al*. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015;350:h391.
22 Kullgren JT, Krupka E, Schachter A, *et al*. Precommitting to choose wisely about low-value services: a stepped wedge cluster randomised trial. *BMJ Qual Saf* 2017. doi:10.1136/bmjqs-2017-006699. [Epub ahead of print].
23 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.
24 Franklin BD, Reynolds M, Sadler S, *et al*. The effect of the electronic transmission of prescriptions on dispensing errors and prescription enhancements made in English community pharmacies: a naturalistic stepped wedge study. *BMJ Qual Saf* 2014;23:629–38.
25 Kilbourne AM, Almirall D, Eisenberg D, *et al*. Protocol: Adaptive Implementation of Effective Programs Trial (ADEPT): cluster randomized SMART trial comparing a standard versus enhanced implementation strategy to improve outcomes of a mood disorders program. *Implement Sci* 2014;9:132.
26 Kane RL, Huckfeldt P, Tappen R, *et al*. Effects of an intervention to reduce hospitalizations from nursing homes: a randomized implementation trial of the INTERACT program. *JAMA Intern Med* 2017;177:1257–64.