The Science of Improvement

Donald M. Berwick, MD, MPP, FRCP

In the early 1890s, Dr William Halsted developed radical mastectomy for breast cancer. Surgeons performed the Halsted procedure for more than 80 years even though there was little systematic evidence for its success. Then a new breed of scholars subjected the procedure to formal methods of evaluation unknown to Halsted.[1] The methods—randomized controlled trials (RCTs) principal among them—led to a surprise: radical mastectomy had no advantage over simpler forms of treatment.[2]

This is but 1 example of the hard-won victory of evidence over belief in medicine. The pioneers of the formal evaluation of medical practices raised questions that traditional practitioners did not welcome. But in time, formal evaluation prevailed.[3,4] The pioneers developed a hierarchy of evidentiary rigor relating the design of a study to the confidence that could be placed in the findings, from the lowly, nearly valueless anecdote to the royalty of evidence, the RCT.

Concurrently, a similar story of hard-won learning unfolded in the so-called quality movement. Scholars illuminated the scale and types of defects in the processes of care and the outcomes, including high rates of unscientific care,[5] inappropriate care,[6] geographic variations in practice,[7] latent disagreements among specialists,[8] and often-unrecognized medical injury to patients.[9] Like the pioneers of evidence-based medicine, students of medical quality were at first largely ignored, but no longer. In 1999 and 2001, the Institute of Medicine published 2 landmark reports on the evidence for quality failures[10,11] and called urgently for redesign of care systems to achieve improvements.

The story could end here happily with 2 great streams of endeavor merging into a framework for conjoint action: improving clinical evidence and improving the process of care. Instead, the 2 endeavors are often in unhappy tension.

Neither disputes that progress toward health care's main goal, the relief of illness and pain, requires research of many kinds: basic, clinical, systems, epidemiologic. The disagreement centers on epistemology—ways to get at "truth" and how those ways should vary depending on the knowledge sought. Individuals most involved in day-to-day improvement work fear that if "evidence" is too narrowly defined and the approach to gathering evidence too severely constrained, progress may be the victim. For example, the RCT is a powerful, perhaps unequaled, research design to explore the efficacy of conceptually neat components of clinical practice—tests, drugs, and procedures. For other crucially important learning purposes, however, it serves less well.

Recent controversies about the evaluation of rapid response teams provide a case in point. These controversies show the importance of adjusting research methods to fit research questions. Although only 10% to 15% of inpatients resuscitated outside intensive care units survive to hospital discharge, early warning signs are present in a large percentage of patients who ultimately experience cardiac arrest. Rapid response team systems bring expert clinicians to the bedsides of deteriorating patients before arrest occurs. In the mid 1990s, based largely on reports from Australian investigators, the Institute for Healthcare Improvement and others began introducing the concept to willing hospitals. Local experience strongly suggested that these systems often, although not always, were associated with improved outcomes, including reduced anxiety among nursing staff; increased interdisciplinary teamwork; decreased cardiac arrests outside of intensive care units; and, in some cases, declines in mortality.[12,13]

The evidence base took a turn in June 2005 with the publication of the Medical Early Response Intervention and Therapy (MERIT) Study,[14] a cluster randomized prospective trial that claimed to find no beneficial effect of these teams on several primary outcomes. Controversy has continued since then regarding the scientific evidence for rapid response systems.

In fact, the MERIT trial was not negative; it was inconclusive. The study team encountered an array of serious problems in execution, common in social science. For example, although the study's power calculation assumed a baseline rate of 30 events per 1000 admissions, the actual rate proved to be fewer than 7 events per 1000 admissions; thus, the study was effectively underpowered by 500%. Cross-contamination abounded; some control hospitals implemented rapid response protocols, and several study hospi-

Author Affiliation: Institute for Healthcare Improvement, Cambridge, Massachusetts.
Corresponding Author: Donald M. Berwick, MD, MPP, FRCP, Institute for Healthcare Improvement, 20 University Rd, Seventh Floor, Cambridge, MA 02138 (dberwick@ihi.org).

tals failed to do so. Variation among hospitals in outcome events was enormous—with a 95% confidence interval range of 4.37 events per 1000 admissions, 80% of the total event rates in both groups.

Nonetheless, some skeptics seized on the MERIT trial and a few other inconclusive experiments to urge caution in the spread of rapid response teams and criticized those who urge their adoption in locally suitable forms.[15,16] These critics refused to accept as evidence the large, positive, accumulating experience of many hospitals that were adapting rapid response for their own use, such as children's hospitals.[17]

How can accumulating local reports of effectiveness of improvement interventions, such as rapid response systems, be reconciled with contrary findings from formal trials with their own varying imperfections? The reasons for this apparent gap between science and experience lie deep in epistemology. The introduction of rapid response systems in hospitals is a complex, multicomponent intervention—essentially a process of social change. The effectiveness of these systems is sensitive to an array of influences: leadership, changing environments, details of implementation, organizational history, and much more. In such complex terrain, the RCT is an impoverished way to learn. Critics who use it as a truth standard in this context are incorrect.

In *Realistic Evaluation*,[18] Pawson and Tilley make a case for the improvement of evaluation. They argue strongly for methods that go beyond the classic "successionist" format of experimental design that dominates the usual toolkit of evidence-based medicine. They use the shorthand OXO to refer to such designs: observe a system (O), introduce a perturbation (X) to some participants but not others, and then observe again (O). Properly measured, the changes in outcome are, with a calculable degree of certainty, attributable to the perturbation.

Pawson and Tilley[18] assert boldly that when studies use the OXO paradigm to evaluate social programs (that include most system improvements in medicine), the result, in the aggregate, is almost always "a heroic failure, promising so much and yet ending up in ironic anticlimax. The underlying logic . . . seems meticulous, clear-headed and militarily precise, and yet findings seem to emerge in a typically non-cumulative, low-impact, prone-to-equivocation sort of way." Indeed, the assertion either that nothing works or that the results are inconsistent and more research is needed is a typical conclusion from classical OXO evaluations of quality-improvement efforts in health care, such as rapid response teams, chronic disease management projects, or improvement collaboratives.

Pawson and Tilley[18] suggest an alternative evaluation model, which they call CMO, context + mechanism = outcome. They write, "Programs work (have successful 'outcomes') only insofar as they introduce the appropriate ideas and opportunities ('mechanisms') to groups in the appropriate social and cultural conditions ('contexts')."

Why does the OXO model fail in this context? Pawson and Tilley[18] claim, "[E]xperimentalists have pursued too single-mindedly the question of whether a [social] program works at the expense of knowing why it works." Thus, although the OXO model seeks generalizable knowledge, in that pursuit it relies on—it depends on—removing most of the local details about "how" something works and about the "what" of contexts. It therefore reveals little about mechanisms or about factors that affect generalizability. Studying a few covariates, or using stratified designs, or probing for interactions can mitigate this loss, but these are inadequate tools for studying complex, unstable, nonlinear social change.

This is not news in health care evaluation.[19,20] Many have pointed out that there is, and ought to be, a strong relationship between what is studied and how it is studied. To study a linear, mechanical or natural, tightly coupled causal relationship most efficiently (for example, determining benefits of β-blockers for heart failure), an OXO design (such as an RCT) may be exactly correct. But with social changes—multicomponent interventions, some of which are interpersonal, all of which are nonlinear, in complex social systems—then other, richer, but equally disciplined, ways to learn (such as CMO designs) are needed.

Four changes in the current approach to evidence in health care would help accelerate the improvement of systems of care and practice. First, embrace a wider range of scientific methodologies. To improve care, evaluation should retain and share information on both mechanisms (ie, the ways in which specific social programs actually produce social changes) and contexts (ie, local conditions that could have influenced the outcomes of interest). Evaluators and medical journals will have to recognize that, by itself, the usual OXO experimental paradigm is not up to this task. It is possible to rely on other methods without sacrificing rigor. Many assessment techniques developed in engineering and used in quality improvement—statistical process control, time series analysis, simulations, and factorial experiments—have more power to inform about mechanisms and contexts than do RCTs, as do ethnography, anthropology, and other qualitative methods. For these specific applications, these methods are not compromises in learning how to improve; they are superior.

Second, reconsider thresholds for action on evidence. Embedded in traditional rules of inference (like the canonical threshold $P<.05$) is a strong aversion to rejecting the null hypothesis when it is true. That is prudent when the risks of change are high and when the status quo warrants some confidence. However, the Institute of Medicine report *Crossing the Quality Chasm*[11] calls into question the wisdom of favoring the status quo.

Auerbach et al[16] warned against "proceeding largely on the basis of urgency rather than evidence" in trying to improve quality of care. This is a false choice. It is both possible and wise to remain alert and vigilant for problems while

testing promising changes very rapidly and with a sense of urgency. A central idea in improvement is to make changes incrementally, learning from experience while doing so: plan-do-study-act.

Third, rethink views about trust and bias. Bias can be a serious threat to valid inference; however, too vigorous an attack on bias can have unanticipated perverse effects. First, methods that seek to eliminate bias can sacrifice local wisdom since many OXO designs intentionally remove knowledge of context and mechanisms. That is wasteful. Almost always, the individuals who are making changes in care systems know more about mechanisms and context than third-party evaluators can learn with randomized trials. Second, injudicious assaults on bias can discourage the required change agents. Insensitive suspicion about biases, no matter how well-intended, can feel like attacks on sincerity, honesty, or intelligence. A better plan is to equip the workforce to study the effects of their efforts, actively and objectively, as part of daily work.

Fourth, be careful about mood, affect, and civility in evaluations. Academicians and frontline caregivers best serve patients and communities when they engage with each other on mutually respectful terms. Practitioners show respect for academic work when they put formal scientific findings into practice rapidly and appropriately. Academicians show respect for clinical work when they want to find out what practitioners know.

The rhetoric and tone of comment on work in the field of day-to-day health care affect the pace of improvement. Academic medicine has a major opportunity to support the redesign of health care systems; it ought to bear part of the burden for accelerating the pace, confidence, and pervasiveness of that change. Health care researchers who believe that their main role is to ride the brakes on change—to weigh evidence with impoverished tools, ill-fit for use—are not being as helpful as they need to be. "Where is the randomized trial?" is, for many purposes, the right question, but for many others it is the wrong question, a myopic one. A better one is broader: "What is everyone learning?" Asking the question that way will help clinicians and researchers see further in navigating toward improvement.

Financial Disclosures: None reported.

Previous Presentation: Based on a presentation at the Sixth Great Ormond Street Hospital Lecture; September 26, 2007; London, England.

Additional Contributions: Paul Batalden, MD, Dartmouth Medical School; Frank Davidoff, MD, Institute for Healthcare Improvement; Thomas Nolan, PhD, Institute for Healthcare Improvement; Erika Pabo, BA, Harvard Medical School; and Jane Roessner, PhD, Institute for Healthcare Improvement, assisted in the development and refinement of this article.

REFERENCES

1. Fisher B. From Halsted to prevention and beyond: advances in the management of breast cancer during the twentieth century. *Eur J Cancer*. 1999;35(14):1963-1973.

2. Veronesi U, Saccozzi R, Del Vecchio M, et al. Comparing radical mastectomy with quadrantectomy, axillary dissection, and radiotherapy in patients with small cancers of the breast. *N Engl J Med*. 1981;305(1):6-11.

3. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J*. 1979;121(9):1193-1254.

4. Institute of Medicine. *Assessing Medical Technologies: Report of the Committee for Evaluating Medical Technologies in Clinical Use.* Washington, DC: National Academies Press; 1985.

5. Lembcke PA. Medical auditing by scientific methods illustrated by major female pelvic surgery. *JAMA*. 1956;162(7):646-655.

6. Brook RH. *Quality of Care Assessment: A Comparison of Five Methods of Peer Review.* Rockville, MD: Dept of Health, Education, and Welfare; 1973.

7. Wennberg J, Gittelsohn A. Small area variations in health care delivery: a population-based health information system can guide planning and regulatory decision making. *Science*. 1973;182(117):1102-1108.

8. Eddy DM, Billings J. The quality of medical evidence: implications for quality of care. *Health Aff (Millwood)*. 1988;7(1):19-32.

9. Brennan TA, Leape LL, Laird NM, et al. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. *Qual Saf Health Care*. 2004;13(2):145-152.

10. Kohn LT, Corrigan JM, Donaldson MS, eds. *To Err Is Human: Building a Safer Health System.* Washington, DC: National Academies Press; 1999.

11. Hurtado MP, Swift EK, Corrigan JM, eds. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academies Press; 2001.

12. Tibballs J, Kinney S, Oakley E, Hennessy M. Reduction of paediatric inpatient cardiac arrest and death with a medical emergency team: preliminary results. *Arch Dis Child*. 2005;90(11):1148-1152.

13. Esmonde L, McDonnell A, Ball C, et al. Investigating the effectiveness of critical care outreach services: a systematic review. *Intensive Care Med*. 2006;32(11):1713-1721.

14. Hillman K, Chen J, Cretikos M, et al; MERIT Study Investigators. Introduction of the medical emergency team (MET) system: a cluster randomized controlled trial. *Lancet*. 2005;365(9477):2091-2097.

15. Winters BD, Pham J, Pronovost PJ. Rapid response teams: walk, don't run. *JAMA*. 2006;296(13):1645-1647.

16. Auerbach AD, Landefeld CS, Shojania KG. The tension between needing to improve and knowing how to do it. *N Engl J Med*. 2007;357(6):608-613.

17. Sharek PJ, Parast LM, Leong K, et al. Effect of a rapid response team on hospital-wide mortality and code rates outside the ICU in a children's hospital. *JAMA*. 2007;298(19):2267-2274.

18. Pawson R, Tilley N. *Realistic Evaluation.* London, England: Sage Publications Ltd; 1997.

19. Davidoff F, Batalden P. Toward stronger evidence on quality improvement: draft publication guidelines: the beginning of a consensus process. *Qual Saf Health Care*. 2005;14(5):319-325.

20. Batalden PB, Davidoff F. What is "quality improvement" and how can it transform healthcare? *Qual Saf Health Care*. 2007;16(1):2-3.