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. The methods—randomized controlled trials (RCTs)—principal among them—led to a surprise: radical mastectomy had no advantage over simpler forms of treatment.

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. The pioneers developed a hierarchy of evidence 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, inappropriate care, geographic variations in practice, latent disagreements among specialists, and often unrecognized medical injury to patients. 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 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 joint 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 con-
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 ref-
used to accept as evidence the large, positive, accumulat-
ing 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 sys-
tems, 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 inter-
vention—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 con-
text 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 per-
turbation (X) to some participants but not others, and then
observe again (O). Properly measured, the changes in out-
come are, with a calculable degree of certainty, attributable
to the perturbation.

Pawson and Tilley18 assert boldly that when studies use
the OXO paradigm to evaluate social programs (that in-
clude most system improvements in medicine), the result,
in the aggregate, is almost always “a heroic failure, prom-
ising so much and yet ending up in ironic anticlimax. The
underlying logic . . . seems meticulous, clear-headed and mili-
tarily precise, and yet findings seem to emerge in a typi-
cally 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 re-
sponse teams, chronic disease management projects, or
improvement collaboratives.

Pawson and Tilley18 suggest an alternative evaluation model,
which they call CMO, context + mechanism = outcome. They
write, “Programs work (have successful outcomes) only in-
solar as they introduce the appropriate ideas and opportuni-
ties (‘mechanisms’) to groups in the appropriate social and cul-
tural conditions (‘contexts’).”

Why does the OXO model fail in this context? Pawson
and Tilley19 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 relation-
ship between what is studied and how it is studied. To study
a linear, mechanical or natural, tightly coupled causal rela-
tionship most efficiently (for example, determining ben-
everts 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 inter-
personal, all of which are nonlinear, in complex social sys-
tems—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 medi-

cal journals will have to recognize that, by itself, the usual
OXO experimental paradigm is not up to this task. It is pos-
sible 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 quali-
tative methods. For these specific applications, these meth-
ods are not compromises in learning how to improve; they
are superior.

Second, reconsider thresholds for action on evidence. Em-
bedded 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 Cross-
ing the Quality Chasm11 calls into question the wisdom of
favoring the status quo.

Auerbach et al12 warned against “proceeding largely on
the basis of urgency rather than evidence” in trying to im-
prove quality of care. This is a false choice. It is both pos-
sible 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 persuasiveness 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 everything 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 Osmund 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; Elisa P. Benefield, BA, Harvard Medical School; and Jane Reesner, PhD, Institute for Healthcare Improvement, assisted in the development and refinement of this article.

REFERENCES