SOUNDING BOARD

Creating a Learning Health System through Rapid-Cycle, Randomized Testing

Leora I. Horwitz, M.D., M.H.S., Masha Kuznetsova, M.P.H., and Simon A. Jones, Ph.D.

Last year at NYU Langone Health, we showed millions of best-practice alerts in the electronic health record system to prompt physicians to avoid adverse events and to promote guideline-based care. We called hundreds of patients to remind them that they were overdue for their annual physical examination. We made approximately 19,000 postdischarge telephone calls to patients in an attempt to reduce their risk of readmission. We sent thousands of letters to remind patients of unmet preventive care needs. In addition, we started a community health worker program in the emergency department to connect hundreds of high-risk patients to outpatient care. Collectively, these programs alone cost our institution more than a million dollars and used resources that potentially could have been used in other ways to improve care and outcomes. Until recently, we had no real idea whether any of these efforts were working.

Health care systems typically implement such interventions wholesale because they seem like good ideas. To our knowledge, they rarely formally evaluate the effectiveness of these interventions, let alone rigorously perform iterations of tests for improvement. At best, a hospital may track outcomes over time in the hope of seeing a benefit. However, such before-and-after analyses are typically limited by secular trends, selection biases, regression to the mean, loss to follow-up, lack of control groups, inconsistent implementation, different concurrently implemented interventions, and a host of other real-world challenges. Evaluations are rarely unbiased enough for personnel at health care systems to be confident that a program works; conversely, ineffective programs are routinely continued for years for lack of persuasive evidence that they are failing.

In January 2018, with seed funding provided by a hospital trustee, we began to upend this status quo and turn NYU Langone Health into a learning health system through rapid-cycle, randomized tests of existing systems-level programs (i.e., randomized quality-improvement projects). A learning health system is characterized by "continual improvement and innovation" with "new knowledge captured as an integral by-product of the delivery experience."[1] We now know with confidence that changing the text of a provider-targeted prompt to give tobacco cessation counseling in an office produces a significant increase in rates of medication prescriptions and that changing just a few sentences in telephone outreach scripts can both shorten telephone calls and increase rates of appointments for annual examinations. We have also learned that our postdischarge telephone calls have made no difference in rates of readmission or patient-experience ratings, that our appointment-reminder letters were completely ineffective, and that our community health worker program was inadvertently targeting patients who were unlikely to benefit (Table 1). Interestingly, the randomized quality-improvement projects have also uncovered unrecognized systems errors: for instance, the influenza vaccination alert was inappropriately being triggered in the operating room, and the algorithm used to identify patients with mental health disorders who were at high risk for visiting the emergency department included low-risk diagnoses such as nicotine dependence.

In just 1 year, we have completed 10 randomized quality-improvement projects, and the learning health system program has already shown that it can pay for itself through increased adoption of preventive services. The value of the program, however, lies beyond short-term quantifiable return on investment. By learning that many of the interventions we had regarded as routine are not working, we can iteratively test until they become effective, or, if appropriate, we can reassign staff to perform different interventions that are more effective. We think of studies that show

Table 1. Examples of Randomized Quality-Improvement Projects Performed to Date.

Target Outcome	Intervention (Project Period)	Randomization Method	Total No.	Result	Potential Effect if Improvement Were Applied Systemwide for 1 Yr	Potential Reasons for Ineffectiveness
Increasing influenza vaccination alert acceptance	Alternative best-practice alert vs. original alert for nurses (3 mo)	Performed at the patient level	1028 Hospitalizations, 3941 patient-days, 91,168 alerts	There was no difference in rates of alert acceptance	Not applicable	Nurses saw both versions of the alert; nurses expected to give vaccination only at discharge; the best-practice alert was being triggered at inappropriate times; patient refusal was not documented in the appropriate place to turn off the alert
Increasing smoking cessation	3 New versions of a best-practice alert (4 wk)	Performed at the practice level; stratified randomization according to practice size	15,353 Office visits	The version that indicated that smoking cessation was a billable service was the most effective	An additional 1842 patients would receive a prescription for smoking cessation medication; billings for smoking cessation counseling would total $373,000	Not applicable
Reducing preventive care gaps	Mailed reminder vs. no reminder (3 mo)	Performed at the patient level	3457 Patients	There was no difference in gap reduction	Discontinuation of mailed reminders would save $25,000	Uncertain
Increasing annual wellness visits	Alternative versions of telephone outreach scripts (6 wk per cycle)*	Performed at the patient level; script versions were alternated in weekly intervals	300–400 Patients per cycle	Incremental improvements were observed in each cycle	An additional 392 appointments would be scheduled; billings for visits would total >$30,000	Not applicable
Increasing acute care use	Community health worker vs. none for high-risk patients in the emergency department (6 mo)	Stratified according to odd vs. even medical-record numbers	2314 Patients	There was no difference in 30-day or 60-day rates of acute care use	Not applicable	Emergency department use and hospital admissions are multifactorial problems; targeted patients may not have been the most likely to benefit
Decreasing readmissions	Postdischarge telephone follow-up vs. none (3 mo+1 additional mo for readmission outcomes)	Stratified according to odd vs. even medical-record numbers	3267 Patients	There was no difference in 30-day rates of acute care use or patient-experience ratings	Not applicable	Readmission is a multifactorial problem; inpatient experience cannot be altered by a postdischarge telephone call
Increasing patient-reported outcome survey completion	Different text or poster messages in waiting room (2 wk per cycle)†	Performed at the patient level; message versions were alternated in weekly intervals	3000–3600 Office visits per cycle	There was no difference in rates of survey completion	Not applicable	Messages were insufficiently different; survey completion was dependent on time in the waiting room; patients were discouraged from future completion of surveys because physicians did not use the survey responses

* Four iterations of new scripts have been performed in three cycles to date.

† Two cycles have been performed to date (the first tested two different tablet-based messages, and the second tested two different poster-based messages).

no benefit not as failures but as successes in terms of identifying opportunities to improve care. This framing is crucial for the continued trust, support, and buy-in of the staff who partner with us to study their own practice. Our goal is to run dozens of such quality-improvement projects each year and ultimately to make randomization standard practice for the continual improvement of existing programs and the implementation of new programs. Areas of particular focus in the coming year include electronic health record–based alerts, which can improve quality and safety but may also increase burnout and alert fatigue, and care-coordination activities, which are resource intensive but have high potential for benefit if implemented effectively.

Controlled trials in clinical medicine date back to the scurvy trial by Lind in 1747.[2] Rapid-cycle, randomized tests (also called A/B tests) are routinely used by online media providers,[3] web designers,[4] and even some government agencies.[5] Yet they are virtually absent in health care. What does it take to bring randomization into health care operations?

First, frontline engagement is required. Our projects are not designed by external researchers who are unfamiliar with processes on the ground but are created and implemented by frontline staff in collaboration with our team — a measure that makes implementation seamless and low cost. An inspiration for this work is the model of research used by BetaGov, an organization that works with government agencies to perform randomized studies of interventions in the penal system and elsewhere.[5] The staff at the agencies ultimately become proficient at developing studies on their own.

Second, a judicious selection of programs to test is important. Because these are the quality-improvement analogues of pragmatic clinical trials, they share many of the same design constraints. Programs that make the most successful candidates for randomized quality-improvement projects have a high volume of events and have short-term outcomes that are already routinely captured. We collect no new data for these projects. Moreover, because the intent of the projects is to improve quality of care, we focus on comparing approaches to increase the adoption of accepted practice. Projects that are designed to test whether clinical interventions in themselves are effective or safe are not appropriate for this mechanism and should be performed as clinical trials.

Third, a support structure is crucial. This learning health system program is housed in the Center for Healthcare Innovation and Delivery Science at NYU Langone Health, which provides an experienced health care delivery scientist (L.I.H.), a project manager (M.K.), a project assistant, a data analyst, and a statistician (S.A.J.), with a total cost to the institution of less than $350,000 per year. Core infrastructure facilitates planning and avoids later problems. We have developed standard templates for study design (including problem analysis, strategy to manage change, proposed interventions, target population, definition of outcomes and potential unintended consequences, baseline outcome rate, anticipated number of observations per week, unit of randomization, blinding, and implementation of the randomization strategy) and for protocol submission to ClinicalTrials.gov. We have also created groups of practices, stratified according to size, that can be randomly assigned to interventions, saving us from repeated manual creation of intervention groups.

CHALLENGES

Selection of an appropriate randomization strategy is key and can materially influence the results. In our first project, we tested what we intended to be a more user-friendly version of a best-practice alert to encourage nurses to order influenza vaccination. Randomization was performed at the patient level largely for technical reasons: our electronic health record has built-in functionality that enabled us to display version A of the alert for one group of randomly assigned patients and version B of the alert for another group, but it has no functionality to randomly assign care providers to receive different alerts. We would have had to manually create a list of nurses to receive each alert, a process that is both exceedingly cumbersome and impractical, because new nurses are regularly hired. Moreover, our outcome (influenza vaccination) was evaluated at the patient level. This randomization strategy, however, turned out to be a mistake. We soon found that virtually every nurse had seen both versions of the alert by virtue of taking care of different patients. Accordingly, in the next alert-related project (smoking cessation), we performed randomization at the practice level instead.

Often, true randomization is not possible, in which case we revert to pseudorandomization. For instance, one of our system hospitals was beginning a new postdischarge telephone follow-up program and did not have enough staff to call every patient. They agreed to partner with us to randomly assign the patients who would or would not receive postdischarge telephone calls so that the effectiveness of the program could be determined. However, the electronic health record did not have functionality to randomly assign the patients included in the real-time list of discharges used by the callers. Instead, we applied a filter that simply removed all patients with even medical-record numbers from the list. This created a pseudorandom sample of patients; those with odd medical-record numbers received postdischarge telephone calls, and those with even medical-record numbers did not. Although this approach was not technically random, we confirmed that it created equally sized populations with similar demographic characteristics. Similarly, we performed a series of iterations of telephone outreach scripts for annual visits, successively comparing the most successful existing script with a new version. However, the callers found it confusing to switch scripts between calls. Instead, we first randomly assigned the patients to hear either script A or script B, and then the callers switched between using script A and script B in weekly intervals for several weeks, thus minimizing effects of secular trends while maintaining a pragmatic pseudorandomization scheme. In all cases, we avoid using randomization methods that would require the frontline staff to change their practice (e.g., by making them perform the randomization or enter data into a new database to track the randomly assigned patients), because it is impractical and undermines the embedded nature of the work.

ETHICAL CONSIDERATIONS

This work falls squarely into the challenging gray zone of quality improvement versus research. Before we began any projects, we discussed the learning health system program with our institutional review board, which determined that these projects meet the criteria for quality-improvement work (i.e., the projects are conducted by persons involved in the care of patients for the specific purpose of improving care at our local institution, positive results are promptly incorporated into practice, the projects involve minimal risk, the lessons we learn are likely to be specific to our culture and workflow and are not necessarily generalizable to other institutions, and the projects are intended to increase the provision or uptake of recommended practices to improve care or avoid harm).[6,7] Randomization alone does not define our projects as research. Institutional review boards at other institutions have made similar determinations for equivalent projects.[8]

Nonetheless, we take ethical considerations seriously.[9] We avoid the collection of personal identifiers, which are almost never necessary in an evaluation of effect. In projects that compare an intervention that is already in place with no intervention, we prioritize the interventions in which capacity constraints already prevent the intervention from being applied to all patients in order to avoid depriving patients of a potentially effective intervention without their consent. For instance, at baseline, our existing community health worker intervention was only enrolling 7% of eligible patients. Patient enrollment based on randomization instead of convenience did not change the number of patients who received the intervention, but it enabled a rigorous evaluation of the effect of the intervention and may even have helped to reduce bias in who was approached. Similarly, the institution did not have enough support staff to call every discharged patient at baseline; randomization allowed the same total number of calls to take place, but the calls were made to a sample that was selected without bias and could be evaluated.

We do not provide patients or clinicians with the opportunity to opt out of studies, because this is largely not feasible for wholesale systems interventions, nor is it ethically required for quality-improvement work.[9] However, we are exploring ways of publicizing the methods by which our institution is committed to rigorous, continuous improvement to our patients and staff so that they are aware of our approach. When we observe that an assigned intervention in one randomization group is superior to that in another, we make it the new standard, so that all patients may benefit from improved processes as we discover them. Although these are not clinical trials and are not federally funded, we voluntarily file a protocol with ClinicalTrials.gov before starting each project to maximize rigor and reproducibility and report the results of each project when it has been completed. Most important, we believe that it is our ethical obligation as employees at a health care institution to evaluate the effectiveness of our efforts to improve quality and avoid harm and to use the best methods available to improve the effectiveness of our processes so that we provide the maximum benefit to our patients.

NEXT STEPS

As our institution has gained experience with embedding randomized studies of systems interventions into routine practice, we have identified several areas that need improvement. To enable projects to run with less assistance, we need better infrastructure, such as prebuilt groups of practices that can be randomly assigned to interventions; standardized data extraction and analytic code, particularly for projects based on electronic health records; and reporting templates that would automatically generate final tables and figures. We have been using very basic study designs. Instead, we could begin to apply more sophisticated designs so that we can learn faster and either give all our patients access to a more effective process sooner or stop providing them with ineffective care, which diverts resources from potentially more valuable interventions. Factorial designs for evaluations of telephone outreach scripts and electronic health record alerts would enable us to test multiple iterations simultaneously. Adaptive trial designs would allow us to stop interventions early when futility or efficacy is shown, drop failing interventions earlier in projects with multiple study groups, adaptively enrich our study population to include those who are most likely to benefit, or shift our randomization ratio toward the more promising study-group assignment.[10,11] For instance, we could use a population-enrichment design to identify whether there are subgroups of patients for whom the community health worker intervention was less effective; the identification of such subgroups would enable us to stop recruiting these patients to avoid burdening them with a nonbeneficial treatment and to increase our ability to detect a benefit in other subgroups. Finally, we would like to find a means to disseminate our findings as rapidly and inexpensively as we conducted the studies. Although we plan to eventually report the results of most projects in peer-reviewed publications, such publication takes time, and some projects may be perceived as too incremental or local to warrant publication. Other means of dissemination, such as distributing preliminary results (i.e., "preprints"), posting basic findings on websites, or creating a quality-improvement study network for informal sharing, may be needed.

Health care institutions are facing increasing ethical, regulatory, and financial imperatives to improve care. Rapid-cycle, randomized quality-improvement projects are a potentially extremely effective, low cost, but underused tool in creating a learning health system that achieves the triple aim of providing better health and health care at lower cost.[12]

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Allen Thorpe for funding the learning health system program and NYU Langone Health for providing in-kind contributions, Felicia Mendoza for providing project assistance, the personnel at NYU Langone Medical Center Information Technology, and the many staff members and clinicians who collaborated on project designs and implementation.

From the Center for Healthcare Innovation and Delivery Science, NYU Langone Health (L.I.H., M.K., S.A.J.), and the Division of Healthcare Delivery Science, Department of Population Health (L.I.H., S.A.J.), and the Division of General Internal Medicine and Clinical Innovation, Department of Medicine (L.I.H.), NYU School of Medicine, New York.

1. Smith MD, Saunders R, Stuckhardt J, McGinnis JM, eds. Better care at lower cost: the path to continuously learning health care in America. Washington, DC: National Academies Press, 2012.
2. Lind J. A treatise on the scurvy in three parts, containing an inquiry into the nature, causes, and cure, of that disease; together with a critical and chronological view of what has been published on the subject. Edinburgh: A. Kincaid and A. Donaldson, 1753.
3. Bulik M. Which headlines attract most readers? New York Times. June 13, 2016 (https://www.nytimes.com/2016/06/13/insider/which-headlines-attract-most-readers.html).
4. Hanington J. The ABCs of A/B testing. Salesforce.com. July 12, 2012 (https://www.pardot.com/blog/abcs-ab-testing/).
5. BetaGov home page (http://betagov.org).
6. Finkelstein JA, Brickman AL, Capron A, et al. Oversight on the borderline: quality improvement and pragmatic research. Clin Trials 2015;12:457-66.
7. Baily MA. Harming through protection? N Engl J Med 2008;358:768-9.
8. Brenner AT, Rhode J, Yang JY, et al. Comparative effectiveness of mailed reminders with and without fecal immunochemical tests for Medicaid beneficiaries at a large county health department: a randomized controlled trial. Cancer 2018;124:3346-54.
9. Baily MA, Bottrell M, Lynn J, Jennings B. The ethics of using QI methods to improve health care quality and safety. Hastings Cent Rep 2006;36(4):S1-S40.
10. Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med 2018;16:29.
11. Etchells E, Ho M, Shojania KG. Value of small sample sizes in rapid-cycle quality improvement projects. BMJ Qual Saf 2016;25:202-6.
12. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27:759-69.

DOI: 10.1056/NEJMsb1900856