THE
MILBANK QUARTERLY
A MULTIDISCIPLINARY JOURNAL OF POPULATION HEALTH AND HEALTH POLICY

Perspective

Beyond Causality: Additional Benefits of Randomized Controlled Trials for Improving Health Care Delivery

MARCELLA ALSAN[*,†] and AMY N. FINKELSTEIN[†,‡]

[*]*Harvard Kennedy School;* [†]*J-PAL North America;* [‡]*Massachusetts Institute of Technology*

Policy Points:

- Policymakers at federal and state agencies, health systems, payers, and providers need rigorous evidence for strategies to improve health care delivery and population health. This is all the more urgent now, during the COVID-19 pandemic and its aftermath, especially among low-income communities and communities of color.
- Randomized controlled trials (RCTs) are known for their ability to produce credible causal impact estimates, which is why they are used to evaluate the safety and efficacy of drugs and, increasingly, to evaluate health care delivery and policy. But RCTs provide other benefits, allowing policymakers and researchers to: 1) design studies to answer the question they want to answer, 2) test theory and mechanisms to help enrich understanding beyond the results of a single study, 3) examine potentially subtle, indirect effects of a program or policy, and 4) collaborate closely to generate policy-relevant findings.
- Illustrating each of these points with examples of recent RCTs in health care, we demonstrate how policymakers can utilize RCTs to solve pressing challenges.

From Scurvy to Streptomycin to Social Policy

> Their cases were as similar as I could have them.
>
> —James Lind, 1753[1]

The Milbank Quarterly, Vol. 99, No. 4, 2021 (pp. 864-881)

With deep roots in clinical medicine, randomized controlled trials (RCTs) are a familiar tool to generate needed evidence in medicine. In 1753, James Lind conducted what is considered the first experiment resembling a modern controlled trial. While working on a ship, the surgeon noticed high mortality from scurvy among sailors. Lind then conducted a comparative controlled trial of the effect of various treatments to scurvy on 12 similarly sick sailors and found oranges and lemons to be the best treatment.[1] Randomized allocation, however, had to wait until the 20th century. The first RCT in medicine was conducted in 1946, by Austin Bradford Hill and his colleagues at the Medical Research Council (MRC), to evaluate streptomycin's effectiveness in tuberculosis.[2] Within a few decades, the US Food and Drug Administration (FDA) required drug producers to include RCT results in their drug applications.[3]

Compared to the long history and current prominence of RCTs in medical research, their use to improve health care delivery in the United States is more recent and, while gaining momentum, is still less prominent than in medicine. Since the 1960s, at least a few dozen RCTs of social policies in the United States have looked at health as an outcome,[4] but historically, RCTs were rarely used to evaluate innovations in health care delivery or health policy. Of course, there are well-known exceptions, such as the famous RAND Health Insurance Experiment in the 1970s and, more recently, the 2008 Oregon Health Insurance Experiment, but these exceptions seemed only to prove the rule.[5,6] For example, between 2009 and 2013, just 18% of studies of US health care delivery interventions used randomization, compared to 86% of drug studies and 66% of studies of nondrug medical interventions.[7]

There are, of course, practical reasons for the relative paucity of these RCTs, including cost and implementation challenges. Recently, however, researchers, practitioners, and policymakers have begun to find ways to overcome these oft-cited barriers to the widespread use of RCTs in health care delivery and to launch important RCTs evaluating health care delivery models and health policy options.[8]

Such RCTs are frequently heralded for their ability to produce clear and credible evidence of an intervention's causal effects. This is certainly a valuable aspect of an RCT. But RCTs do not have a monopoly on establishing causality, as evidenced by the plethora of compelling quasi-experimental studies using techniques such as regression discontinuity designs, instrumental variables, and event study methodologies (to name

just a few) that have been used successfully to estimate the causal impact of aptly named "natural experiments."[9]

Therefore, as capacity and enthusiasm build for implementing RCTs in health care delivery, we highlight in this article what we see as four other important benefits of RCTs. Specifically we discuss their ability (1) to answer the questions that practitioners and researchers *want* to study rather than what they *can* study with naturally occurring data; (2) to study a program's indirect, or "spillover," effects; (3) to test theory and uncover the mechanism behind why a program does or does not have an effect; and (4) to encourage valuable collaboration between researchers and implementing partners that can sharpen the questions asked and the hypotheses examined.

We begin by first reviewing the standard case for RCTs based on the clear and credible influence that they provide, as well as the important challenges for implementing such RCTs. We then turn to the meat of this article, which is a discussion of the additional potential benefits of RCTs. We illustrate these additional benefits with examples of RCTs in health care delivery published primarily in the last three years.

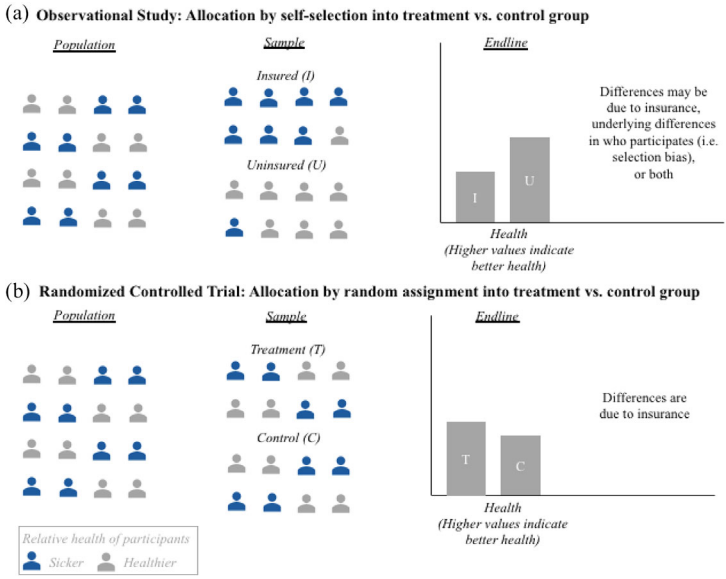
The Opportunity and Challenges for Conducting RCTs to Improve Health Care Delivery

> Having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity.
>
> —Austin Bradford Hill, 1952[10]

Opportunity

Several years after his successful streptomycin RCT in Britain, Austin Bradford Hill, in the preceding quotation, made the case to the Harvard Medical School faculty for randomized clinical trials, explaining how studies that allocated individuals to treatment randomly and without favor provided credible estimates of an average causal effect. His remarks make the now standard case for RCTs: random assignment can eliminate

Figure 1. Randomized Controlled Trials Can Make Credibly Estimating Causal Impact Easier: A Stylized Illustration of Health Insurance's Impact on Health [Color figure can be viewed at wileyonlinelibrary.com]



confounding factors that may bias the results of even well-controlled observational analyses.[11]

Consider the impact of health insurance on health. Figure 1 provides a schematic of two hypothetical, stylized studies of the health effects of insurance: an observational study and an RCT. Panel A illustrates the observational design: all individuals are offered insurance and some enroll. Health is subsequently compared between the insured and the uninsured. Since individuals choose whether or not to have insurance, baseline differences are likely between those who have insurance and those who do not, differences that the researchers may be unable to fully measure and control for. In our stylized example, those who choose insurance are sicker at baseline than are those who do not elect to obtain insurance, as would be expected by standard models of adverse selection in insurance.[12] On the right-hand side of Panel A are the two groups'

end-line differences in health. It shows that at the end line, the insured have worse health than do the uninsured. Did health insurance lead to worse health, or did it seem to worsen health because people who are especially sick are more likely to have health insurance?

Panel B investigates this same question using an RCT. In contrast to the observational study, the important feature of the RCT is that the allocation of insurance is randomized. Random assignment ensures that, on average, there are no systematic differences at baseline between those with and without insurance. As a result, as illustrated in the figure, the treatment group that is randomly allocated insurance has the same proportion of sicker individuals at baseline as does the control group that is randomly assigned to remain uninsured. The researchers therefore can be reasonably confident that any difference in health at the end line is due to the insurance coverage by itself, rather than to differences in the groups' underlying characteristics that may be correlated with the outcome of interest.

A recent real-world example illustrating the value of randomization for elucidating causal effects of health policy comes from a 2020 RCT of a care transition program known colloquially as "Hotspotting." Created by the Camden Coalition of Healthcare Providers, this program targets the "super-utilizers" of the health care system as a way to reduce spending and improve health. It provides patients with a team of community health workers, nurses, and social workers who visit them after discharge to coordinate their care and connect them to social services.[13] Observational studies of this program, and of similar programs, had found promising results, that the programs significantly reduced spending on health care.[14]

Observational studies can use statistical methods to try to account for bias that may exist, but sometimes the biases are very hard to correct for, as Dr. Paula Lantz explained in her *Milbank Quarterly* article on super-utilizer interventions.[14] Dr. Lantz and her coauthors' 2019 systematic literature review of 46 evaluations of interventions targeted at super-utilizers (most commonly, case management programs) warned that "methodological and study design weaknesses—especially regression to the mean—were widespread and call into question reported positive findings."[15] In other words, observational studies of super-utilizer programs are likely biased by regression to the mean, which in this case means the tendency for patients incurring unusually high costs at a particular point in time to move closer to the average over time. Indeed,

researchers selected by the Center for Medicare & Medicaid Innovation to evaluate this program using quasi-experimental methods concluded that they were unable to do so because of the difficulty identifying an appropriate, naturally occurring, comparison group.[16]

The 2020 study of the Hotspotting program began by looking at readmission rates in the treatment group—those patients who were randomly enrolled in the program. The results looked promising: patients in this group visited the hospital about 40% less often in the six months after the intervention. Unfortunately, however, this effect turned out to be entirely due to regression to the mean. Readmissions in the control group of individuals who were not offered participation in the program declined by the same amount. As a result, the findings from the RCT show that the program had no effect on readmissions.[16]

Challenges

The contrast between the RCT and the observational results underscores the importance of using a rigorous method to evaluate the program's causal effect. The RCT's findings also, however, highlight the inevitable limits of any one study, no matter how rigorous. The Hotspotting study, like any single study, did not provide all the answers. In particular, it was not able to address whether the program might affect other outcomes, such as patient self-efficacy and well-being, or whether the program's impacts might be different for other types of patients or in other settings. It also did not speak to *why* the intervention did not reduce readmissions, although we will discuss later how other RCTs have been successfully designed precisely to answer questions of mechanism.

In addition, as we emphasized at the outset of this article, RCTs are not the only compelling way to estimate causal impacts. That is fortunate, since for any given question of interest, an RCT may not be feasible or desirable. Chief among them are issues of cost, time, and ethics.

Administrative data can sometimes help surmount these challenges by reducing monetary costs. For example, two decades before the Hotspotting study just described, in an RCT of a similar care-transition program, researchers used telephone interviews with patients to obtain information on readmissions after discharge, the study's primary outcome.[17] In the more recent study, improved data systems allowed researchers instead to use existing hospital discharge data from the four

Camden hospital systems and the Camden Coalition Health Information Exchange database. This allowed them to measure readmissions at substantially lower cost and effort and with less risk of nonresponse bias.

But even when an RCT's monetary costs can be minimized, a related issue concerns their potential time costs. For example, in the Hotspotting RCT, it took researchers over three years to recruit and receive consent from the targeted 800 patients for the study, not to mention almost two years of planning, designing, and piloting the study.

Fortunately, not all RCTs take years. NYU Langone Health, for example, offers a model for implementing cheap, rapid-cycle, quality improvement RCTs. The researchers at Langone Health recently completed 10 RCTs in a year and concluded that this effort paid for itself by increasing the adoption of preventive care.[18]

A final important barrier to conducting RCTs is ethical concerns. In many settings, implementing an RCT would be unethical, for example, if the program of interest were already available for everyone it was designed to serve, or if resources were available to expand it to serve everyone. In such cases, alternative possible interventions may still offer equipoise. As Dr. Seth Berkowitz and Dr. Shreya Kangovi wrote, "Even though it doesn't take a randomized trial to know you should feed a hungry child, it may take one to know how best to do so."[19] In addition, when there are not enough resources to serve all eligible patients, a random lottery may be the fairest way to allocate limited slots, enabling health systems to learn in the process. In 2008, for example, the state of Oregon needed to allocate a limited number of Medicaid slots. As the then state director of Medicaid, Jim Edge, said at the time,

> We thought about other options, such as should we try to pick all of the sickest people or the kids or the people with cancer or heart disease. But the Feds won't allow that, and there's just no way to guarantee the fairness of that. Why would cancer be more deserving than heart disease?[20]

State officials and advocacy groups decided—without input from researchers—that random assignment was the best method to allocate the limited slots, and thus the Oregon Health Insurance Experiment was born.

When RCTs are ethical and practical, they provide several valuable benefits. The rest of this article discusses four of them.

RCTs Can Study the Questions You Want to Study, Rather Than the Ones Nature Permits

> Natural experiments can be used to obtain convincing evidence of causal effects or of the impacts of interventions. Perfect instances of these circumstances rarely occur in practice.
>
> —Peter Craig et al., 2012[21]

As already discussed, RCTs are just one method researchers can use to credibly estimate causal impacts of a program or intervention.[22] But a clear advantage of an RCT is that it allows researchers and partners an opportunity to design a study to investigate the questions they *want* answered, rather than the questions they *can* answer with naturally occurring variations.

As the preceding quotation from the Medical Research Council emphasizes, without RCTs, researchers are limited in what questions they can pursue by what natural experiments exist. RCTs empower researchers and partners to use randomization, when ethical and feasible, to study what they want to study, rather than what nature gives them.

One recent illustration comes from an RCT in 2019 that provided rigorous evidence on a topic notoriously challenging to answer using naturally available data: the impact of physician-patient race concordance on the patient's health behavior. This is difficult to study using observational data: Because most individuals choose their primary care doctor, selection already exists in concordant versus discordant dyads. In addition, given long-standing structural inequalities, many disadvantaged individuals do not even have a primary care doctor. The 2019 study overcame this challenge by randomizing patients to receive a racially concordant or discordant doctor and examining how this affected demand for preventive health care.

Specifically, the study focused on patient-provider concordance for Black men, the demographic group in the United States with the lowest life expectancy.[23] Existing correlative evidence suggests that racial concordance is associated with greater participation in care and adherence to treatment.[24] The researchers therefore hypothesized that the lack of diversity in the physician workforce might make it difficult for Black men to find a doctor who "looks like them" and thus would contribute to these disparities.

To examine this hypothesis, the researchers created a pop-up clinic and recruited participants from the surrounding Oakland (California) area. Once at the clinic, they randomly assigned patients to see either a Black or non-Black (white or Asian) physician. They found that Black male patients randomly assigned to Black physicians were 18 percentage points more likely to use preventive services like diabetes screenings and flu vaccines after interacting with their physician than were those assigned to non-Black physicians.[25] This study highlights how RCTs can be designed to help address seemingly intractable, systemwide challenges, like racial disparities in health, by breaking them down into answerable questions to identify root causes.

RCTs Can Test Theory and Unpack Mechanisms

> If researchers and policy makers continue to view results of impact evaluations as a black box and fail to focus on mechanisms, the movement toward evidence-based policy making will fall far short of its potential for improving people's lives.
>
> —Mary Ann Bates and Rachel Glennerster, 2017[26]

A common criticism of RCTs is that they produce "black box" studies that do not illuminate why a program did or did not have an impact. This critique can certainly apply to some RCTs, such as the Hotspotting study described earlier. But it also applies more broadly to any well-identified study of a causal effect—whether using randomized or quasi-experimental methods—that has insufficient variation to elucidate the drivers behind an estimate of impact. Moreover, unlike quasi-experimental studies, which can study only "naturally occurring" variations, RCTs can be designed specifically to see inside the black box or to test particular theories.

A 2010 RCT of the impact of immunization camps and incentives provides a classic example of how RCTs can be designed to reveal underlying structural barriers.[27] In this RCT, the nonprofit Seva Mandir aimed to increase children's immunization rates—which involved five courses of vaccines—in rural Udaipur, India. In most villages in this area, only 2% of the children had received all five courses of immunizations.

Researchers designed an experiment explicitly to distinguish between several possible reasons for these low rates. One possibility was a

supply-side problem: the clinics where families received immunizations were often closed. Another possibility was a demand-side problem: families might have difficulty going to the clinics five different times to complete the full course and might not have enough incentive to do so. These barriers are, of course, not mutually exclusive, and may interact in important ways.

To investigate this, the researchers randomly assigned communities to one of three groups: a control group, a treatment group that was provided with regularly scheduled immunization camps to investigate the supply hypothesis, and a treatment group that was provided with the camps and also with incentives (specifically 1 kilogram of lentils at each clinic visit and a set of plates after the fifth immunization course) to investigate the combined impact of the supply-and-demand hypotheses.

Just 50% of the control group received one course of immunizations, and only 6% received the full five courses. Both treatment groups had around 75% immunization rates after the first course. However, the camps and incentives group had 39% immunization rates for the full five courses, compared to 18% in the camps-only group. These results suggested that supply-side issues were a challenge but that they were not the only barrier. All together, the findings suggested that combining increased access with demand-side incentives could be an effective way to increase vaccination rates.

When RCTs help unpack mechanisms, they can be informative for policy even outside the study's specific location or population. As with all studies, assessing "external validity" or how the results can be applied in similar contexts, is challenging. In this case, we learned that in a setting in which immunization rates are low, providing reliable access can boost immunization rates but that by itself, access is not enough and incentives may be an important tool for making sure children complete the full immunization course. While these exact results may not translate across contexts, the idea that access is not a panacea is instructive in considering interventions in other contexts.

RCTs Can Credibly Examine the *Indirect* Effects of a Program or Policy

> RCTs have the ability to surprise you.
>
> —Esther Duflo, 2019

Esther Duflo, one of the 2019 Nobel laureates in economics, is famous for her work on experimental approaches to alleviating global poverty. As her preceding remark illustrates, RCTs can uncover surprising or subtle effects of a program or policy. In fact, her point is a corollary to the rigor of the evidence provided by RCTs. It facilitates using them to credibly examine potentially important indirect effects of a program, such as a policy's effects on untargeted populations.

It is very important from a policy perspective to be able to measure the effects of a policy on those not directly targeted by it, but estimating such "spillover" effects is notoriously difficult. In the canonical research design, which compares outcomes for directly targeted actors to outcomes for nontargeted actors, spillover effects cannot be identified. When a research design does permit the identification of spillovers, a skeptical reader may interpret the effects on nontargeted patients as evidence of a flawed research design rather than evidence of spillovers. For good reason, therefore, the bar for credibly identifying spillovers is high.

Well-conducted, large-scale RCTs can be designed to credibly determine the effects of spillovers, since random assignment mitigates concerns about how researchers estimate the effect. For example, to measure the spillover effect of bed nets ("herd immunity") against malaria, researchers can use a two-stage randomized design. Clusters can be randomly assigned to a treatment group in which some individuals will randomly receive the bed nets and a control group who does not receive them but has access to usual services. The difference between these two groups is the overall effect of the bed nets. Not everyone in the treatment group receives the bed nets. That is, within the treatment group clusters, people would be randomly assigned to receive the bed nets or to a control group. The difference between the control units in the treated versus control clusters provides a causal estimate of the spillover effect.[28]

Results from two recent nationwide Medicare policy RCTs conducted by the Centers for Medicare & Medicaid Services (CMS) show the substantial effects of Medicare policy on the treatment of non-Medicare patients, suggesting that an individual insurer's reforms can have important, broader effects on the system as a whole. In one case, CMS randomly assigned a Medicare payment reform for hip and knee replacement to 67 of 171 metropolitan statistical area (MSAs). The control MSAs continued to be reimbursed by Medicare under the status quo system. Evidence from the first two years of this five-year RCT indicates that the payment reform modestly reduced health care utilization among covered patients (those on Original Medicare), primarily by reducing discharges

to postacute care facilities by about 10%. But it also found that the payment reform had spillover effects on privately insured Medicare Advantage patients, even though their payment regime did not change. These spillover effects were of the same sign and magnitude as the directed effects on the targeted patients.[29–35]

Another study examined the spillover effects of a CMS warning letter on privately insured patients. This letter was sent to primary care physicians (PCPs) who prescribed significantly more of an antipsychotic drug to their Original Medicare patients than their peers. Each year, 2.8 million patients fill a subscription for quetiapine (brand name Seroquel), yet as many as 75% of these prescriptions are for uses not approved by the FDA.[36] CMS identified about 5,000 physicians who prescribed significantly more to their Medicare patients than their in-state peers, and randomly assigned half of them to receive letters stating that their prescribing to Original Medicare patients was extremely high relative to their peers and that they were under review by CMS. The strongly worded peer comparison letters from CMS reduced quetiapine prescribing for Original Medicare patients (the targeted group) by about 17% for the next two years. The letters also reduced the number of prescriptions to the PCP's *privately insured* patients (both those with Medicare Advantage and those with employer-sponsored insurance) by 12%. The researchers could not reject the hypothesis that the effects on targeted and nontargeted patients were the same.[37,38]

RCTs Can Facilitate Fruitful Collaborations with Implementing Partners

> The statistician, if he is to play his proper role in a clinical trial must be in it "up to his neck." … The statistically designed clinical trial is above all a work of collaboration between clinician and statistician, and that collaboration must prevail from start to finish.
>
> —Austin Bradford Hill, 1952[10]

In the then new realm of medical randomized trials, Austin Bradford Hill emphasized that proper statistical research is steeped in practice. What was true in the 1950s for medical trials is still true in the 2020s for RCTs on health care delivery. Unlike many clinical researchers, some social scientists (such as economists like ourselves) spend their careers up to their necks in data at their desks.

All research projects benefit from real-world experience and input from practitioners. Policy research in particular should involve key stakeholders and frontline providers, from the inception of the hypothesis to the interpretation of the results. Nothing prevents researchers from using observational data or quasi-experiments to engage with stakeholders, and of course, some do so. But RCTs typically *necessitate* researchers to confer with real-world providers to design and implement the study. In our own experience with both types of research, we have found that this "nudge" helps us sharpen the questions we ask and the hypotheses we test and better enables us to interpret the results.

Implementing partners often provide invaluable advice about both a study's design and the interpretation of it. For example, in the physician-patient racial concordance study discussed earlier, the implementing partner was instrumental in advising on recruitment and which messaging would and would not work to persuade patients to show up at the clinic. Owen Garrick, the president and chief operating officer of Bridge Clinical Research and a coauthor of the study, recommended recruiting participants from barbershops, a common location for health outreach among Black men. One of us (Alsan) literally drove around Oakland with Garrick to identify the barbershops in the area, which were not often on Google Maps. Furthermore, Bridge Clinical Research's focus groups with Black men had found that advertising injections (such as flu shots) may deter them from participating in a study. As such, the advertisements we used integrated this feedback, and the study recruitment, and ultimately the study, benefited.

In another example, the Hotspotting study discussed earlier, the intervention itself had been developed by the implementing partner based on the hypothesis that breaking down the silos to coordinating patient care could be successful in helping very complex patients. Not surprisingly, therefore, since the theory of change behind the intervention had originated from the partner, the partner provided invaluable input in how one of us (Finkelstein) should interpret and learn the most from the results of the study of the intervention's impact.

RCTs: A Many-Splendored Thing

Policymakers at federal and state agencies, health systems, payers, and providers are innovating constantly to provide higher-quality, more

efficient care to patients. Researchers and leaders in health care have recently made encouraging progress in overcoming ethical, logistical, and cost barriers to implementing RCTs to study how best to improve health care delivery and patients' experience.[8] This trend has been applauded, with people pointing to the widely known benefits of RCTS for producing credible evidence of causal effects.[39,40]

The COVID-19 pandemic has only heightened the need to identify effective and equitable ways to improve population health and the health care system. RCTs can help policymakers develop needed evidence. Although not every program or policy can or should be evaluated with an RCT, it is a powerful tool that should be considered when initiating or expanding health care delivery interventions.

In this article, we have described four *additional* (and perhaps less widely appreciated) benefits of RCTs from the perspective of social scientists. They put both researchers and policymakers in the driver's seat, allowing them to answer the questions that they want answered, rather than what they can answer with naturally occurring variations. They can help identify subtle, indirect effects of programs and policies. They also can be designed to uncover underlying mechanisms and explain why studies find particular results. Finally, the close collaborations between researchers and partners can help create more policy-relevant research.

References

1. Lind J. *A Treatise of the Scurvy, in Three Parts; Containing an Inquiry into the Nature, Causes, and Cure, of That Disease*. Cambridge, England: Cambridge University Press; 2014.
2. Hart P. A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s. *BMJ*. 1999;319(7209):572-573. https://doi.org/10.1136/bmj.319.7209.572/
3. Bothwell L, Podolsky S. The emergence of the randomized, controlled trial. *N Engl J Med*. 2016;375(6):501-504. https://doi.org/10.1056/nejmp1604635.
4. Courtin E, Kim S, Song S, Yu W, Muennig P. Can social policies improve health? A systematic review and meta-analysis of 38 randomized trials. *Milbank Q*. 2020;98(2):297-371. https://doi.org/10.1111/1468-0009.12451.
5. Newhouse J. *Free for All?*. Cambridge, MA: Harvard University Press; 1993.

6. Insuring the Uninsured. Abdul Latif Jameel Poverty Action Lab. https://www.povertyactionlab.org/sites/default/files/publications/Insuring_the_Uninsured.pdf. Published January 2014. Accessed October 14, 2020.
7. Finkelstein A, Taubman S. Randomize evaluations to improve health care delivery. *Science*. 2015;347(6223):720-722. https://doi.org/10.1126/science.aaa2362.
8. Finkelstein A. A strategy for improving U.S. health care delivery—conducting more randomized, controlled trials. *N Engl J Med*. 2020;382(16):1485-1488. https://doi.org/10.1056/nejmp1915762.
9. Angrist JD, Pischke J. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *J Econ Perspect*. 2010;24(2):3-30. https://doi.org/10.1257/jep.24.2.3.
10. Hill A. The clinical trial. *N Engl J Med*. 1952;247(4):113-119. https://doi.org/10.1056/nejm195207242470401.
11. Angus D. Optimizing the trade-off between learning and doing in a pandemic. *JAMA*. 2020;323(19):1895. https://doi.org/10.1001/jama.2020.4984.
12. Einav L, Finkelstein A. Selection in insurance markets: theory and empirics in pictures. *J Econ Perspect*. 2011;25(1):115-38. https://doi.org/10.1257/jep.25.1.115.
13. Finkelstein A, Zhou A, Taubman S, Doyle J. Health care hotspotting—a randomized, controlled trial. *N Engl J Med*. 2020;382(2):152-162. https://doi.org/10.1056/nejmsa1906848.
14. Lantz PM. "Super-utilizer" interventions: what they reveal about evaluation research, wishful thinking, and health equity. *Milbank Q*. 2020;98. https://doi.org/10.1111/1468-0009.12449.
15. Iovan S, Lantz PM, Allan K, Abir M. Interventions to decrease use in prehospital and emergency care settings among super-utilizers in the United States: a systematic review. *Med Care Res Rev*. 2020;77(2):99-111. https://doi.org/10.1177/1077558719845722.
16. Peterson G, Blue L, Kranker K, et al. Evaluation of the health care innovation awards (HCIAs): primary care redesign programs. addendum to third annual report. Princeton, NJ: Mathematica Policy Research; 2017. https://innovation.cms.gov/files/reports/hcia-primarycareredesign-thirdannrpt-addendum.pdf. Accessed December 14, 2020.
17. Naylor MD, Brooten D, Campbell R, et al. Comprehensive discharge planning and home follow-up of hospitalized elders: a

randomized clinical trial. *JAMA*. 1999;281:613-620. https://doi.org/10.1001/jama.281.7.613.

18. Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med*. 2019;381:1175-1179. https://doi.org/10.1056/NEJMsb1900856.
19. Berkowitz SA, Kangovi S. Health care's social movement should not leave science behind. *Milbank Q Opinion*. September 3, 2020. http://doi.org/10.1599/mqop.2020.0826.
20. Yardley W. Drawing lots for health care. *New York Times*. https://www.nytimes.com/2008/03/13/us/13bend.html. Published March 13, 2008. Accessed December 14, 2020.
21. Craig P, Cooper C, Gunnell D, et al. Using natural experiments to evaluate population health interventions: guidance for producers and users of evidence. *Medical Research Council*. 2012. https://mrc.ukri.org/documents/pdf/natural-experiments-guidance/. Accessed April 23, 2021.
22. Bor J, Moscoe E, Mutevedzi P, Newell M, Bärnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiol*. 2014;25(5):729-737. http://doi.org/10.1097/EDE.0000000000000138
23. Arias E, Xu J. *National Vital Statistics Report: United States Life Tables*, 2017. Washington, DC: Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_07-508.pdf. Accessed April 23, 2021.
24. Cooper-Patrick L. Race, gender, and partnership in the patient-physician relationship. *JAMA*. 1999;282(6):583. https://doi.org/10.1001/jama.282.6.583.
25. Alsan M, Garrick O, Graziani G. Does diversity matter for health? Experimental evidence from Oakland. *Am Econ Rev*. 2019;109(12):4071-4111. https://doi.org/10.1257/aer.20181446.
26. Bates MA, Glennerster R. *The generalizability puzzle*. Stanford Soc Innovation Rev. 2017. https://ssir.org/articles/entry/the_generalizability_puzzle. Accessed April 23, 2021.
27. Banerjee AV, Duflo E, Glennerster R, Kothari D. Improving immunisation coverage in rural India: a clustered randomised controlled evaluation of immunisation campaigns with and without incentives. *BMJ*. 2010;340:c2220.
28. Benjamin-Chung J, Arnold BF, Berger D, et al. Spillover effects in epidemiology: parameters, study designs and methodological considerations. *Int J Epidemiol*. 2018;47(1):332-347. https://doi.org/10.1093/ije/dyx201. PMID: 29106568; PMCID: PMC5837695.

29. Barnett M, Wilcock A, McWilliams J, et al. Two-year evaluation of mandatory bundled payments for joint replacement. *N Engl J Med*. 2019;380(3):252-262. https://doi.org/10.1056/nejmsa1809010.
30. Finkelstein A, Ji Y, Mahoney N, Skinner J. Mandatory Medicare bundled payment program for lower extremity joint replacement and discharge to institutional postacute care: interim analysis of the first year of a 5-year randomized trial. *JAMA*. 2018;320(9):892. https://doi.org/10.1001/jama.2018.12346.
31. Haas D, Zhang X, Kaplan R, Song Z. Evaluation of economic and clinical outcomes under Centers for Medicare & Medicaid Services mandatory bundled payments for joint replacements. *JAMA Intern Med*. 2019;179(7):924. https://doi.org/10.1001/jamainternmed.2019.0480.
32. The Lewin Group. CMS comprehensive care for joint replacement model: performance year 1 evaluation report. 2018. https://innovation.cms.gov/files/reports/cjr-firstannrpt.pdf. Accessed April 23, 2021.
33. Einav L, Finkelstein A, Ji Y, Mahoney N. Randomized trial shows healthcare payment reform has equal-sized spillover effects on patients not targeted by reform. *Proc Natl Acad Sci*. 2020;117(32):18939-18947. https://doi.org/10.1073/pnas.2004759117.
34. Wilcock A, Barnett M, McWilliams J, Grabowski D, Mehrotra A. Association between Medicare's mandatory joint replacement bundled payment program and post–acute care use in Medicare Advantage. *JAMA Surg*. 2020;155(1):82. https://doi.org/10.1001/jamasurg.2019.3957.
35. Meyers D, Kosar C, Rahman M, Mor V, Trivedi A. Association of mandatory bundled payments for joint replacement with use of postacute care among Medicare Advantage enrollees. *JAMA Network Open*. 2019;2(12):e1918535. https://doi.org/10.1001/jamanetworkopen.2019.18535.
36. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration. Pediatric postmarketing pharmacovigilance and drug utilization review: seroquel and seroquel xr. http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/PediatricAdvisoryCommittee/UCM494485.pdf. Published February 19, 2016. Accessed April 21, 2020.
37. Barnett M, Olenski A, Sacarny A. Common practice: spillovers from Medicare on private health care. NBER Working Paper Series. Working Paper 27270. May 2020. https://www.nber.org/papers/w27270.pdf. https://doi.org/10.3386/w27270.

38. Sacarny A, Barnett M, Le J, Tetkoski F, Yokum D, Agrawal S. Effect of peer comparison letters for high-volume primary care prescribers of quetiapine in older and disabled adults. *JAMA Psychiatry*. 2018;75(10):1003. https://doi.org/10.1001/jamapsychiatry.2018.1867.
39. Carroll A. Workplace wellness programs don't work well. Why some studies show otherwise. *New York Times*. August 6, 2018. https://www.nytimes.com/2018/08/06/upshot/employer-wellness-programs-randomized-trials.html. Accessed October 19, 2020.
40. Orszag P. The health-care breakthrough that wasn't. *Bloomberg Opinion*. January 8, 2020. https://www.bloomberg.com/opinion/articles/2020-01-08/hot-spotters-debunked-by-randomized-control-trial. Accessed October 19, 2020.

Funding/Support: None.

Acknowledgments: We are extremely grateful to Spencer Crawford for his assistance in drafting this article.

Conflict of Interest Disclosures: Both authors completed the ICMJE Form for Disclosure of Potential Conflicts of Interest. No conflicts were reported.

Address correspondence to: Marcella Alsan, Professor of Public Policy, Harvard Kennedy School, 79 John F. Kennedy St., Cambridge, MA 02138 (email: marcella_alsan@hks.harvard.edu)