Using Randomized Evaluations to Improve the Efficiency of US Healthcare Delivery

Amy Finkelstein
MIT, J-PAL North America, and NBER

Sarah Taubman
MIT, J-PAL North America

February 2015

Abstract: Randomized evaluations of interventions that may improve the efficiency of US healthcare delivery are unfortunately rare. Across top journals in medicine, health services research, and economics, less than 20 percent of studies of interventions in US healthcare delivery are randomized. By contrast, about 80 percent of studies of medical interventions are randomized. The tide may be turning toward making randomized evaluations closer to the norm in healthcare delivery, as the relevant actors have an increasing need for rigorous evidence of what interventions work and why. At the same time, the increasing availability of administrative data that enables high-quality, low-cost RCTs and of new sources of funding that support RCTs in healthcare delivery make them easier to conduct. We suggest a number of design choices that can enhance the feasibility and impact of RCTs on US healthcare delivery including: greater reliance on existing administrative data; measuring a wide range of outcomes on healthcare costs, health, and broader economic measures; and designing experiments in a way that sheds light on underlying mechanisms. Finally, we discuss some of the more common concerns raised about the feasibility and desirability of healthcare RCTs and when and how these can be overcome.

We are grateful to Innessa Colaiacovo, Lizi Chen, and Belinda Tang for outstanding research assistance, to Katherine Baicker, Mary Ann Bates, Kelly Bidwell, Joe Doyle, Mireille Jacobson, Larry Katz, Adam Sacarny, Jesse Shapiro, and Annetta Zhou for helpful comments, and to the Laura and John Arnold Foundation for financial support.

Table of Contents

Spotlights

I. Introduction and Summary

The US healthcare system has been characterized by high and rapidly rising healthcare costs. Healthcare spending has grown from about 5 percent of the economy in 1960 to 17.2 percent in 2012 (Centers for Medicare and Medicaid Services, 2012). In recent years, increasing attention has been paid to ways to slow the growth of medical expenditures. One particular area of focus is improving the efficiency of healthcare services, which has commonly been posed as the "triple aim": reducing costs, improving health and improving patient experience (Berwick, Nolan, & Whittington, 2008).

This paper focuses on opportunities for high-impact randomized evaluations of efforts to improve the efficiency of healthcare delivery in the United States. Randomized evaluations—also commonly referred to as randomized controlled trials, or RCTs, are routinely used to test new medical innovations, particularly new drugs. For good reason, they are widely considered to be the "gold standard" of evidence.

When RCTs are conducted in healthcare delivery, they have enormous influence, due to the simplicity, transparency, and credibility of their design. For example, two randomized evaluations of the impact of health insurance have had outsized influence in the voluminous academic literature and public policy discourse on health insurance reform. The RAND Health Insurance Experiment—a randomized evaluation of the impact of consumer cost-sharing conducted in the 1970s—is still widely held to be the gold standard of evidence for predicting the likely impact of health insurance design on medical spending. The Oregon Health Insurance Experiment—a 2008 randomized evaluation of Medicaid coverage for low-income uninsured adults—has dispelled some claims of leading public policy figures and healthcare experts—such as the claim that Medicaid had no proven benefits, or that Medicaid would reduce use of the emergency room—and has corroborated others—such as the claim that Medicaid increased use of healthcare, including primary care and preventive care, and increased healthcare spending. It has been the subject of numerous front-page articles and opinion pieces as well as the primary input into several government reports on the impact of expanding Medicaid under the Affordable Care Act (ACA).

RCTs are also unparalleled in their ability to produce surprising results that must be taken seriously. For example, we describe below a randomized evaluation of the effect of offering concurrent palliative care to metastatic lung cancer patients (Temel et al., 2010). It found that offering concurrent palliative care in addition to standard care improved self-reported quality of life, which is perhaps not surprising. More surprisingly, it also found that concurrent palliative care increased survival while reducing aggressive end-of-life care. A similar finding in a non-experimental setting would doubtless raise concerns about potential unobservable differences between those seeking concurrent palliative care and those not seeking concurrent palliative care that might be spuriously producing the result.

Unfortunately, there is only limited use of RCTs to test innovations in US healthcare delivery—such as changing who provides care and how, what standard of care is offered or incentivized, how payments are done, what information is conveyed and how, what cost patients face for receiving care, how to engage and coach patients in self care, and so on. Across top journals in medicine, health services research, and economics, less than 20 percent of studies of interventions in US healthcare delivery are randomized. By contrast, in top medical journals, about 80 percent of studies of US medical interventions are randomized.

Fortunately, the tide may be turning on RCTs in healthcare delivery. A number of recent policy reforms have given providers increasing "skin in the game," making them the residual claimant on cost savings or imposing financial penalties for low quality. This will likely create even greater incentives among healthcare providers, health insurers, large employers, and state and local governments to understand what is most effective at improving the efficiency for healthcare delivery. At the same time, a number of new developments have increased the feasibility of conducting RCTs on US healthcare delivery—including the increasing ability of using administrative data and information systems to conduct low-cost, high-quality RCTs with both quick turnaround on initial impact as well as the possibility of long-term follow up, and a number of new funding opportunities for RCTs on US healthcare delivery.

To encourage high-impact RCTs in this area, we discuss some of the optimal design features of RCTs on US healthcare delivery. We begin by surveying the nature of existing RCTs on US healthcare delivery, concentrating on the ones in top medical journals, where they are disproportionately to be found. We find that most existing RCTs on US healthcare delivery are relatively small, with a median sample size of around 500 patients. Most are patient-level randomizations testing a single (usually multi-faceted) intervention against the control (status quo). They tend to rely heavily on primary data collection—very few rely solely on pre-existing administrative data—and to examine effects over a time horizon of one year or less. Most focus on health outcomes, and only a minority looks at healthcare use or costs; even fewer look at both health outcomes and healthcare use or costs.

We suggest a number of design choices that can improve the feasibility and impact of RCTs on US healthcare delivery. We are of course cognizant that there are many challenges in conducting research of all kinds—including RCTs. Our primary conclusion that more RCTs are needed in field of healthcare delivery should not be overshadowed by the discussion of elements of ideal design. However, in many cases, these design elements are achievable. To emphasize this, we highlight through "spotlights" a few examples of RCTs on healthcare delivery interventions that achieved some of the objectives we outline, or illustrate the challenges faced by RCTs that did not.

Our main recommendations are:

1. **Take advantage of existing administrative data.** The field of healthcare delivery is ripe with detailed, reliable, and existing administrative data that enable researchers to study the impact of their intervention on a broad range of outcomes, including healthcare costs and health outcomes. Administrative data offer the potential to do high-quality, low-cost, rapid turnaround RCTs. They also make long-term follow up (including over the course of subsequent decades) much more feasible.

2. **Measure a wide range of outcomes.** Ideally an RCT would examine impacts on both healthcare costs and health outcomes, as well as on broader economic outcomes where relevant (such as employment or participation in other government transfer programs). Additionally they would try to measure any effects (positive or negative) of the intervention on non-targeted outcomes, as well as how the targeted behaviors evolve after the intervention has ended.

3. **Design RCTs to illuminate mechanisms.** Research may have the largest impact when it sheds light on the underlying mechanism behind the impact of a program. This is somewhat different than trials of medical technologies where it is often sufficient to know that a treatment is effective even if the underlying mechanism is not yet clear. In healthcare delivery, however, the interventions are often multi-faceted, specific to the original setting, and cannot be as easily reproduced as a drug or a device. This means a deeper understanding of the mechanism of action is very helpful to other institutions looking to adopt aspects of a successful program.

We consider some common concerns about the feasibility and desirability of RCTs on healthcare delivery. We discuss circumstances in which they are and are not likely important. Ethical concerns about rationing are important to consider, but for many healthcare delivery innovations logistical or financial capacity constraints create oversubscription in the program and hence require some form of rationing. In some such cases, randomization may actually be preferred on ethical grounds. Similarly, uncertainty about the impacts of a program can make randomization particularly appealing during a gradual phase in period. Time and cost considerations are also often cited as a concern with RCTs, but these are inherent costs of research—not of RCTs in particular—and we suggest some design elements that can reduce both costs and time lags.

Finally, to underscore our claims to the feasibility of RCTs, we conclude by suggesting a few specific examples of important questions in healthcare delivery that could be fruitfully answered via an RCT. We emphasize through these examples that RCTs can be used to provide important evidence on the impact not only of patient-level interventions but also of organization- or system-wide interventions, such as the impact of hospital management or provider payment reforms.

The importance of understanding how to improve efficiency in a sector that accounts for almost one-fifth of the economy and an even larger share of public sector budgets, and that literally affects the lives of every person in the country, is unquestioned. The value of RCTs to provide transparent and compelling evidence of the impact of interventions is unparalleled. This paper emphasizes the myriad opportunities to engage in low-cost, short turnaround, high-impact RCTs that can deepen our understanding of how to improve the efficiency of the US healthcare sector. We are optimistic that the use of RCTs in the study of US healthcare delivery will soon be closer to the norm rather than the exception, just as they are in drug trials and other medical interventions, and increasingly in other areas of social policy, such as US education policy and poverty reduction in developing countries.

II. The Value of Randomization

It is not always obvious what the effect of a given policy is. For example, what is the impact of covering the uninsured with health insurance? Comparisons of the insured and the uninsured often indicate that those with insurance are in worse health than those without insurance (see e.g. Finkelstein et al. 2012 Appendix). Would it be right to conclude therefore that health insurance makes individuals sicker? Or, more reasonably, are individuals who are in poor health more likely to seek out health insurance?

Of course, this inference problem is well understood, and sensible researchers can (and usually do) try to control for any *observable* differences between groups that they are comparing, which may be responsible for the differences in outcomes being studied. Richer data allows us to observe and adjust for more differences, but it remains the case that one set of people chose to (and was able to) obtain health insurance and the other set did not. Simply controlling for observables is unlikely to be sufficient. If we find two observationally identical groups of people, one of whom has health insurance and one of whom doesn't, we don't know *why* one observationally identical set of people has health insurance and the other doesn't, which creates lingering concerns about selection on *unobservables*. In other words, the reason that one group of observationally identical people has health insurance and the other does not may be due to differences in unobserved characteristics that are correlated with both insurance coverage and individuals' health or healthcare demand. For example, individuals with unobservably worse health (or private information that they are at risk of a disease) may be more likely to seek out insurance.

Random assignment solves this problem of inference. In randomized evaluations, individuals are selected to receive an intervention (such as health insurance in our example) based on a lottery. Those individuals who are not selected form a comparison group. Because the selection process is random, the two groups are similar in every respect, except that one group receives the program, while the other does not. Therefore if the group who receives the intervention has different outcomes (e.g. is more or less healthy, or uses more or less healthcare) after the program is implemented, we know that this difference was caused by the program. In other words, *by construction* of the random assignment, we know that the two groups on average will be identical except for the fact that one has health insurance and one does not. Therefore, we can attribute any subsequent differences between the insured and the uninsured to the effect of insurance per se. This clear attribution of the effects of the intervention is what gives random assignment its rightful place as the "gold standard" in the pantheon of scientific evidence.

The ability of randomized evaluations to clearly and credibly identify the impact of an intervention also means that they can generate surprising or unexpected results that must be taken seriously. For example, a randomized trial studied the effects of providing palliative care that was integrated with standard medical care to metastatic non-small-cell lung cancer patients in order to improve their quality of life. Unexpectedly, this study found that concurrent palliative care improved not only quality of life but also the length of life (See spotlight on Concurrent Palliative Care). A similar finding in an observational study would doubtless raise concerns about potential unobservable differences between those seeking concurrent palliative care and those not seeking concurrent palliative care which might be spuriously producing the result.

While their unparalleled ability to provide credible causal evidence is arguably the prime attraction of randomized evaluations, they have other important advantages as well. For example, there has been considerable progress in the last few decades in developing non-experimental approaches to causal inference based on research designs that try to approximate randomized controlled trials (Angrist & Pischke, 2010). Examples include sharp policy changes that affect some areas or groups but not others, or regression discontinuity designs that exploit sharp eligibility cut-offs. Continuing our health insurance example, Card et al. (2008, 2009) exploit the sharp discontinuity in health insurance coverage brought about by Medicare's age 65 eligibility threshold to study the impact of Medicare coverage at age 65 on healthcare use and health outcomes in a regression discontinuity framework. Yet these designs, despite their strengths and attractions, have their own limitations. In particular, in many cases, such quasi-experimental estimates can only capture very local estimates (in this example, the impact of Medicare at age 65). Relatedly, since the researchers do not design the variation, the intervention is often relatively "black box," which can make interpretation and extrapolation difficult; for example, as Card et al. (2008) explain, the estimated impact of Medicare at age 65 includes both the impact of covering previously uninsured with Medicare and the impact of changing the nature of coverage at 65 among those who previously had private insurance; this makes it difficult to disentangle the effects of the various mechanisms. By contrast, as we emphasize below, randomized evaluations allow researchers to choose the variation and to design the (potentially multiple) arms of the intervention to elucidate underlying mechanisms and test theories.

Spotlight on Surprising Results: Concurrent Palliative Care

A key benefit of randomized control trials is that the strength of the research design means that surprising and unexpected results can be taken seriously and yield major new insights. For example, aiming to improve quality-of-life of patients with a form of metastatic lung cancer, doctors at Massachusetts General Hospital ran a randomized trial of providing palliative care integrated with standard medical care (Temel et al., 2010). The trial produced the surprising result that concurrent palliative care not only improved quality-of-life, but also length of life.

The trial randomized 151 patients with metastatic non-small-cell lung cancer within 8 weeks of diagnosis. Individuals assigned to the treatment group received standard care. In addition, they were scheduled to meet with the palliative care team immediately (within 3 weeks) and monthly throughout the study. Individuals assigned to the control group received standard care and were scheduled to meet with the palliative care team if and when they requested it; in the first 12 weeks, around 15 percent did so. Quality-of-life was measured at 12 weeks in surviving patients using several measures, and those in the early palliative care group had greater quality-of-life. Although all 151 patients in the study died within 4 years, those assigned to the early palliative care group lived longer than the standard care group (median survival of 11.6 vs. 8.9 months; P = 0.02). The longer survival occurred even though individuals in the early palliative group were less likely to receive aggressive end-of-life care (33 percent vs. 54 percent, P = 0.05).

This example of a healthcare delivery intervention was found to have a substantial benefit on survival, with potential to also reduce costs, yet widespread adoption may be very difficult. Palliative care has been traditionally covered by hospice benefits requiring patients to forgo curative treatment, typically leading to very late use of palliative care when used at all. The Medicare hospice benefit requires patients to have less than six months to live and to choose hospice care instead of other Medicare-covered benefits to treat the terminal illness (Centers for Medicare and Medicaid Services, 2013). Currently, CMS is exploring covering early and concurrent palliative care through the Care Choice Model (Centers for Medicare and Medicaid Services, 2014c).

Back to section

III. Use of RCTs in Healthcare Delivery

Limited use to date

The medical profession recognizes the incomparable value of randomized evaluations. This is why, since the 1960s, the FDA has required RCTs demonstrating the safety and efficacy of new drugs prior to their approval (Junod, 2014). Even outside of the realm of pharmaceuticals, where trials are required, promising interventions in medicine are often tested using RCTs, and the results do not always support the observational inference.

However, randomized evaluations of interventions in US healthcare delivery—in contrast to medicine—are unfortunately rare. To study their prevalence, we conducted reviews of empirical papers studying the impact of an intervention that were published in top journals in economics, health services research, and medicine. We identified studies that investigated a potentially causal link or association between a treatment (or intervention) and an outcome. For each study, we recorded the broad topic (healthcare delivery, medicine, education, etc.) and whether the study design was a randomized controlled trial or not. Except where explicitly noted, we limited the analysis presented below to interventions that took place at least partly in the United States. The Appendix provides considerably more detail on our search and coding process (see Appendix A) as well as a list of the healthcare delivery RCTs identified by our review (see Appendix B).

Table 1 summarizes the results. In economics, we reviewed papers published from 2009-2013 in the *American Economic Review, Quarterly Journal of Economics, Journal of Political Economy, and Econometrica*. In health services journals, we reviewed papers published from 2009-2013 in *Health Affairs, Medical Care*, and *Milbank Quarterly*. In medicine, we reviewed papers from four randomly selected months per year from 2009-2013 for the journals *The New England Journal of Medicine, the Journal of the American Medical Association, Annals of Internal Medicine,* and *PLoS Medicine.* The underlying data generated from this review can be found here.

As can be seen in Table 1, the number of papers on US healthcare delivery varies greatly across the fields. This reflects the differences in volume across the fields, as well as our search parameters. The focus on the table is on the share of studies of US healthcare delivery interventions that are randomized. In economics, it was 15 percent. In health services, it was 3 percent. In medicine, it was 50 percent. Overall (accounting for the sampling differences across fields), an average of 18 percent of healthcare delivery intervention studies were randomized.

Table 1. Use of randomization in studies of US healthcare delivery, 2009-2013

	Number of studies	Number of RCTs	Percent randomized
Top Medical Journals (subsample)	62	31	50
Top Economics Journals	13	2	15
Top Health Services Journals	405	13	3
Adjusted Average*			18

*The average adjusts for the fact that we only reviewed 20 out of 60 months of medical journals.

We endeavored to find some benchmarks against which to compare this result. The benchmarks are naturally field-specific. Table 2 shows the results. The top panel compares the share of economics papers on US healthcare delivery interventions that are randomized to the share of economics papers on other empirical, applied microeconomics topics that are randomized. In US Education, 36 percent of studies were randomized. In international development, 46 percent are randomized. The bottom panel of Table 2 looks at medical journals, examining the share of papers on healthcare delivery interventions that were randomized and the share of medical interventions that were randomized. Fifty percent of the healthcare delivery studies used an RCT design, compared to 79 percent of the medical studies.[1] Within the medical studies, drug studies were very likely to be randomized (98 of 114 studies or 86 percent), but randomization was still common in evaluating other non-drug medical interventions (41 of 62 studies or 66 percent).

Table 2. Use of randomization by study topic

Topic	Number of studies	Number of RCTs	Percent randomized
Panel A: Top Economics Journals*			
US Healthcare Delivery	13	2	15
All Other US-based	192	14	7
Public	50	4	8
Industrial Organization	48	0	0
Labor	31	2	6
Education	22	8	36
Behavioral	12	2	17
Law and Crime	8	0	0
Health and Medicine	7	0	0
Information/Technology	5	0	0
Housing	5	0	0
Other	11	0	0
International development	37	17	46
Panel B: Top Medical Journals			
US Healthcare Delivery	62	31	50
Medical Treatment (US-based)	176	139	79

* Economics papers may be coded as having more than one topic and would contribute to each; the health and medicine group does not include healthcare delivery studies.

In addition to searching top journals in medicine, economics, and health services, we also reviewed projects done by domestic evaluation firms—specifically Abt Associates, Mathematica, MDRC, and RAND. These contract research organizations often work as the evaluators for government projects in health and other

[1] A potential concern with this analysis is that it is limited to top-tier publications only. Such publications may be the most likely to influence scholarship, public opinion and public policy. However, if there is publication bias either for or against healthcare delivery studies (or RCTs on these topics) in top journals, reviewing only these articles may provide a skewed picture of the general use of healthcare delivery RCTs. To address this, we looked at a stratified random sample of about 400 randomized trials registered at www.clinicaltrials.gov between 2006 and 2010. Of them, 17 percent were for healthcare delivery interventions, and 83 percent were for medical interventions. This is similar to the result in Table 2 concerning the share of randomized studies in top medical journals that are for healthcare delivery as opposed to medical interventions (18 percent), and therefore helps alleviate concerns that our analysis of top journals does not present a representative picture of the use of RCTs for studies of healthcare delivery.

fields, but the results of those evaluations may not always be published in academic journals. Our search of top academic journals is therefore missing an important component of domestic evaluation research. Therefore, for the year 2013, we reviewed all the publications (including journal articles, government reports, policy briefs, and others) listed on the websites of these four organizations. Table 3 below summarizes the number of studies done and the use of randomization in different topic areas. Randomization in the healthcare delivery studies by these organizations is about as common as it is in the top journals we reviewed (24 percent randomized compared to 18 percent in the journals). As a comparison, however, 43 percent of studies on other topics used randomized designs.

Table 3. Use of randomization by domestic evaluation firms: by study topic

Topic	Number of studies	Number of RCTs	Percent randomized
US Healthcare Delivery	41	10	24
All Other US-based	69	30	43
Education	40	15	38
Labor	11	4	36
Health and Medicine	8	4	50
Other	14	9	64

*Studies may be coded as having more than one topic and would contribute to each; the health and medicine group does not include healthcare delivery studies.

Possibility of a new era for RCTs

Despite the somewhat gloomy picture painted above on the paucity of RCTs on healthcare delivery, we are optimistic that the tide may be turning. Economic and policy developments have increased the demand of the public sector, insurers, employers and healthcare providers for credible evidence of the impact of different potential delivery interventions, while the feasibility of RCTs has been increased by the growing availability of administrative data and by funding opportunities.

Rising healthcare costs, as well as expanded public and private insurance coverage under the ACA, have increased the need for understanding what interventions can improve the efficiency of healthcare delivery. Rising healthcare costs place pressure on public sector budgets—federal, state, and local—as well as on insurance premiums and employer costs.

At the same time, a key trend in healthcare policy over the past decade has been to give healthcare providers increasing "skin in the game," making them the residual claimant on cost savings, or imposing financial penalties on them for low quality. This therefore creates even greater incentives for healthcare providers to understand what may most improve the efficiency of their healthcare delivery, for which, as discussed in Section II, RCTs can be invaluable.

For example, as part of the ACA, CMS began in 2012 to reduce Medicare payments to hospitals with readmission rates for specific conditions in excess of the expected rate based on the hospital's patient composition (Centers for Medicare and Medicaid Services, 2014b). The ACA also created Accountable Care Organizations (ACOs). These are new organizations comprised of a network of hospitals and providers that contract with CMS to provide care to a large bloc of Medicare patients and which are allowed to keep a fraction of the savings if they come in under their specified cost benchmarks while meeting their quality targets (Ginsburg, 2011). Such a program in principle incentivizes providers to reduce costs while maintaining quality of care, since they directly benefit financially from such cost reductions. Nor are these limited to the public sector. For example, in 2009, Blue Cross Blue Shield of Massachusetts, the state's largest commercial payer, introduced an Alternative Quality Contract in its health-maintenance organization (HMO) and point-of-service commercial enrollee populations, which is structured quite similarly to an ACO (Song et al., 2011).

This notion of "skin in the game" may be one reason why RCTs incentivizing individuals to invest in good health behaviors are done on employees at large firms; employers may perceive a "return on investment" in the form of lower healthcare costs (and hence health insurance premiums) and/or higher productivity workers (Baicker, Cutler, & Song, 2010). As the "Cadillac tax" in the ACA encourages employers to cap the generosity of their healthcare benefits, this will presumably only further increase their incentives to understand how to structure their health benefits to achieve the most "bang for the buck."

Another source of our optimism is that at the same time that there are increased incentives by key actors to undertake RCTs on healthcare delivery, a number of developments have increased the feasibility of conducting such RCTs. Chief among them is the increasing ability to use administrative data and information systems to conduct low-cost, rapid turnaround-time, high-quality RCTs (see Section V for more on this). At the same time, there are a number of new funding opportunities. The newly created Patient-Centered Outcomes Research Institute (PCORI) is providing an estimated $3.5 billion, and the latest round of Center for Medicare & Medicaid Innovation (CMMI) Health Care Innovation Awards provides approximately $1 billion in research grants. Only a fraction of the PCORI and CMMI grants will fund RCTs, of course, but there is also increased funding for RCTs in particular. For example, there is a recent National Institutes of Health (NIH) request for proposals on Low-Cost, Pragmatic, Patient-Centered Randomized Controlled Intervention Trials (Centers for Medicare and Medicaid Services, 2014a; National Institutes of Health, 2013a; Patient-Centered Outcomes Research Institute, 2014).

We therefore anticipate an increasing use of RCTs in healthcare delivery. With this optimism regarding the future, we now turn to a discussion of the characteristics of those healthcare RCTs that do exist (Section IV), followed by recommendations for some best practices going forward (Section V).

IV. Characteristics of Existing Healthcare Delivery RCTs

To examine the characteristics of existing high-profile RCTs in healthcare delivery—and consider potential areas for improvement—we focused on studies published in top medical journals between 2009 and 2013. As seen from Table 1, they have the highest percent of healthcare delivery interventions that are randomized. Because these also represent a large number of published papers in absolute terms, they account for nearly 85 percent of the RCTs on US healthcare delivery published across medicine, economics, and health services.

Table 4 summarizes the randomization and sample sizes of these 99 studies. Roughly half of the studies have sample sizes between 101 and 500, although quite a few are much bigger. The average sample size for these 99 US-based studies was 9,653 individuals, and the median sample size was 446 individuals. With around 450 individuals, a randomized trial has 80 percent power to detect an effect on a binary outcome that changes the probability by approximately 5-10 percentage points (depending on the prevalence of the outcome). For example, for an outcome happening 30 times out of 100 (say, re-hospitalization for a high-risk population), a study of around 450 individuals has 80 percent power to detect a decrease to approximately 21 out of 100 (a 30 percent decrease). Use of pre-randomization data to improve precision may allow for detection of smaller effects, but generally these studies will not have the power to detect more subtle effects.

Randomization was typically done at the patient level (79 percent of studies). Almost all of the studies with patient-level randomization used a standard protocol of recruiting, screening and consenting individual participants who were then randomized into the program or the control group. Only 4 studies used a more passive randomization, where being randomly assigned to the treatment group meant being offered (or auto-enrolled in) the program under study.

When randomization was done at the level of the practitioner or care setting (16 percent of studies), it was typically done within a single organization (such as a hospital system or integrated delivery system), although a handful of studies involved randomization across multiple participating hospitals. For the 5 studies randomized at the practitioner level, the number of randomized clusters ranged from 46 to 472, with an average of 155 (median of 87). For those studies, the outcome measurement and analysis were typically done at the practitioner level as well, so the sample sizes were the same as the number of clusters. For the 11 studies randomized by the care setting, the number of randomized clusters ranged from 8 to 84, with an average of 23 (median of 16). For those studies, the outcome measurement and analysis were typically done at the patient-level, so the sample sizes were much larger; more than half had over 5,000 patients and these studies averaged over 40,000 patients.

Table 4. Randomization and Sample Size

		Number of studies	Percent of studies
Randomization			
	Patient-level	78	79
	Practitioner-level	5	5
	Care-setting-level	11	11
	Other	5	5
Sample Size (Number of Observations)			
	1-100	7	7
	101-500	46	46
	501-1000	13	13
	1001-5000	15	15
	>5000	18	18

Table 5 provides more detail on the interventions tested by the 99 studies. Studies typically only tested a single intervention against a control condition; 73 of the studies (74 percent) included two arms (treatment and control). The interventions tested in the 99 studies varied widely. Most of the interventions focused on patients (86 percent) compared to targeting practitioners (14 percent). We categorized the interventions into some (mutually exclusive) general types; this was based on subjective groupings of similar studies. Interventions that supplemented typical medical care with fitness, diet or weight loss support were quite common (26 percent), as were interventions involving psychotherapy (9 percent). Also relatively common were interventions that attempted to help patients through patient education (7 percent) and group or peer support (4 percent).

There were a variety of interventions that targeted the healthcare delivery system more directly, including care coordination interventions (11 studies), changes to hospital protocols (9 studies), financial incentive interventions (4 studies), and home monitoring studies (4 studies). The home monitoring studies all used some version of self-testing or monitoring with information sent back to providers (Bosworth et al., 2009; Chaudry et al., 2010; Margolis et al., 2013; Matchar et al., 2010). The changes to hospital protocols range from changes in staffing procedures (Kerlin et al., 2013) to infection prevention in the ICU (Climo et al., 2013; Harris et al., 2013; Huang et al., 2013), to protected sleep time for interns (Volpp et al., 2012). Of the four studies which included financial incentives, two were tests of pay-for-performance to providers (Bardach et al., 2013; Petersen et al., 2013), one eliminated co-pays on preventive medications (Choudhry et al., 2011), and one involved direct payments to patients for smoking cessation (Volpp et al., 2009).

Table 5. Testing Interventions

	Number of studies	Percent of studies
Number of arms		
2 arms	73	74
3 arms	16	16
4+ arms	10	10
Intervention focused on		
Patients	85	86
Practitioners	14	14
Type of intervention		
Fitness, diet or weight loss	26	26
Care coordination	11	11
Hospital protocols	9	9
Psychotherapy	9	9
Patient education	7	7
Financial incentives	4	4
Group or peer support	4	4
Home monitoring	4	4
Screening	4	4
Training for providers	4	4
Timing of treatment	3	3
Other	14	14

Table 6 below summarizes the follow-up period and outcome measurement. Average length of follow-up was 600 days; median follow-up was 365 days. Follow-up was typically done by surveys, interviews, in-person measurements, and other primary data collection (85 percent). Of the 15 studies without primary data collection, 11 used medical records. An additional three studies relied exclusively on insurance claims data.[2]

Studies overwhelmingly focused on health, with 85 out of 99 studies (86 percent) measuring health outcomes. Measurement of costs was much less common; only 11 studies (11 percent) reported cost outcomes. Healthcare use, which can potentially be used to estimate costs, was reported in 31 studies (31 percent), of which 22 did not include cost directly.

To judge the impact of an intervention on healthcare efficiency, we would ideally know the changes in both healthcare use or costs and health outcomes. Of the 99 studies we reviewed, 29 studies (29 percent) reported both. There were 10 studies that reported neither, instead focusing on behavior or process.

[2] Only one study of the 99 did not collect primary data and did not use medical records or claims data, but instead relied exclusively on mortality records to measure outcomes (Coburn et al, 2012).

Measuring health without healthcare use or costs was common (57 percent); measuring healthcare use or costs, but not health outcomes was quite rare (4 percent).

Table 6. Measuring outcomes

	Number of studies	Percent of studies
Follow-up period		
< 30 days	9	9
31-90 days	8	8
91-180 days	18	18
181-365 days	33	33
366-730 days	17	17
> 730 days	14	14
Outcome data*		
Primary collection (survey, etc.)	84	85
Medical record	51	52
Insurance claims	3	3
Mortality records	8	8
Outcome measures*		
Health	85	86
Behavior	60	61
Healthcare use	31	31
Healthcare costs	11	11

*Studies may have used multiple sources of outcome data or measured multiple outcomes, so the total percent for both of those exceeds 100.

Overall, this review of recently published healthcare delivery RCTs suggests that such studies tend to be small (median sample size of less than 500), randomized at the patient level (79 percent), with interventions focused on patients (86 percent). Most (74 percent) tested a single, sometimes complex, intervention against a status quo control condition, rather than testing multiple approaches or variations. Follow-up was typically short (median follow-up time of a year or less), and relied heavily on primary data collection. Less than a third of studies (29 percent) looked at impacts on both health and use or costs. In Section V we discuss the value of administrative data, designs that elucidate mechanisms, longer-term follow-up, and examining a range of outcomes. In Section VII, we also discuss how RCTs at the practitioner or system level, rather than the patient, may in many cases be feasible and desirable.

V. Enhancing Feasibility and Impact of RCTs

In this section we discuss several recommendations for how RCTs can be designed to enhance feasibility and impact. Our discussion below assumes some basic elements of good design, such as adequate sample size to achieve balance through randomization and to detect meaningful effects. We highlight a handful of tools that seem underutilized and elements that seem particularly important in the setting of healthcare delivery; Glennerster and Takavarasha (2013) provide a much more thorough discussion of the practical design of randomized trials. We recognize, of course, the considerable challenges to always achieving "ideal" design. Nonetheless, articulating those ideals can be a useful guide to researchers and practitioners faced with key design decisions.

The value of administrative data

One challenge in running very large trials, especially historically, has been the cost and logistical challenges of collecting follow-up data for the purposes of the study. As we saw in Table 6, almost all existing RCTs in healthcare delivery involve primary data collection, whereas only about half used administrative data (and only 15 percent relied exclusively on them). Existing administrative data offer several key advantages as a complement or substitute for primary data collection.

A first advantage of administrative data over primary data collection is that it is usually far easier and cheaper to identify study participants in such data sources than to do primary data collection using telephone or mail surveys or in-person interviews from large numbers of participants. Identifying data sources and negotiating permissions to link to various administrative data can be costly in terms of initial set up time but is usually far cheaper. Better frameworks to share information about available data, application procedures, and data use agreements may help reduce some of these start-up time costs. Moreover, once the framework is in place, in many cases, administrative data can be available in real-time, or close to real time, such as through a hospital system's internal records (Finkelstein, Doyle, Taubman, Zhou, & Brenner, 2014). This can often be of great use to a provider—such as a hospital, an insurer, or government agency—who wants to know in real time whether the intervention is working, rather than to learn the results many years later when decisions about whether and what to maintain or scale up have already been made.

A second advantage of administrative data is that many existing administrative databases include a near-census of the relevant individuals. Such near-universal data coverage can be invaluable in guarding against potential bias from differential nonresponse or attrition to follow-up surveys across experimental treatments, which is a well-known concern for analysis of randomized evaluations (Ashenfelter & Plant, 1990). Of course, one must always be careful to guard against the possibility that the intervention itself may affect whether the subject appears in administrative data. For example, in studying the impact of an intervention using health records from medical providers, one must consider how complete the records will be: does the intervention impact whether the patient has contact with medical providers included in the records?

A third advantage of administrative data is that because the data are typically collected for a separate purpose than for the study, they are less likely than primary data to suffer from the possibility that respondents in different experimental arms skew their answers in favor of what they believe the researchers would like to hear, or what may be in their own interest given the intervention they are in. For example, participants in the treatment arm of the Negative Income Tax Experiment had an incentive to under-state their earnings on surveys, since it would increase their cash transfer; this led to over-statement of the impact of intervention on work effort (Greenberg & Halsey, 1983).

A fourth key advantage of administrative data is that the data in many cases may be more accurate and richer than data obtainable by surveys, especially for variables that are difficult to measure. For example, attempting to assess earnings from survey data can be extremely difficult, as people may be unsure of which sources of income to include or whether they should answer for themselves or their household, or may refuse to provide that information. State unemployment insurance agencies and the Social Security Administration routinely collect detailed earnings data, allowing much more precise measurement of employment and income. Administrative data can also provide data on issues that individuals may be reluctant to report, such as interaction with child protective services agencies (see Spotlight on Nurse-Family Partnership RCTs).

An example of the value of administrative data over survey data can be seen in the Oregon Health Insurance Experiment's study of the impact of covering uninsured low-income adults with Medicaid on emergency room use. This randomized evaluation found no statistically significant impact on emergency room use when measured in survey data, but a statistically significant 40 percent increase in emergency room use in administrative data (Taubman, Allen, Wright, Baicker, & Finkelstein, 2014). Part of this difference was due to greater accuracy in the administrative data than the survey reports; limiting to the same time periods and the same set of individuals, estimated effects were larger in the administrative data and more precise. The administrative data were also able to capture a longer time period of emergency department use (18 months compared with 12 months) and to provide considerably more detail on time of visit, diagnosis, and history of use, which would have been very difficult to obtain through surveys.

Finally, a fifth advantage of administrative data is that, because they are collected, used, and stored for reasons other than the study, administrative data can be very useful in following up on long-term outcomes. This can shed light on the long-term impact of an intervention, which can be a very important factor for policy decisions. We discuss this in the next subsection.

Of course, all of these various attractive characteristics of administrative data are relevant only to the extent that the outcomes of interest are measured in administrative data. There is no substitute for primary data collection, for example, if one wants to measure subjective well-being or patient knowledge of a topic. Increasingly, however, more and more data are being collected and stored electronically, by hospitals and healthcare providers, by insurers, by state and federal agencies, and by private companies. Examples of administrative data that provide a census or near-census of the relevant population include:

- State-level hospital and emergency room discharge data which provide a census of all visits and are available from many states (for a complete list see: Healthcare Cost and Utilization Project, 2013);
- Centers for Medicare and Medicaid Services data on all fee-for-service Medicare enrollees, as well as Medicaid claims data;
- The National Death Index or Social Security Death Master File;
- State-level data on economic outcomes such as earnings and employment (available through the state unemployment insurance records) and participation in various transfer programs such as cash welfare and food stamps.

Many of the large datasets from the private and non-profit sectors can also often provide valuable information, although, in such data, issues of selection bias require more attention. Studies of healthcare delivery in particular tend to rely on medical records—increasingly, electronic medical records—which can provide a relatively complete picture of healthcare use and health outcomes for all included individuals, but only if there is not a lot of care being provided outside of the healthcare system covered by the electronic medical record. The growth of Accountable Care Organizations and other organizational forms discussed in Section II, which make a set of providers fully "at risk" for the costs of a given set of patient, has the ancillary advantage of providing complete utilization data on such patients. More generally, insurance claims data provide information on care provided for a specific set of individuals across all settings.

Spotlight on Elements of Good Design: Nurse-Family Partnership RCTs

The Nurse-Family Partnership (NFP) is an early example of a rigorously tested healthcare delivery innovation that includes several elements of good design. The NFP trials relied on a combination of survey and administrative data and included very long-term follow-up of the participants. Additionally, there were several trials that tested the whether the program could be replicated in different settings and which tested variants of the program.

The NFP was first implemented in 1977 in Elmira, a semirural community of about 100,000 residents in New York with high rates of infant mortality and cases of child abuse and neglect (Olds, Henderson, Tatelbaum, & Chamberlin, 1988).

The NFP treatment involved nurse home visits for first-time mothers who were either unmarried, teenagers, or of low socioeconomic status. Specific objectives of the program included providing education about maternal diet, exercise, smoking/alcohol risks, and early newborn care and enhancing the support network by discussing participation of relatives, involvement of the baby's father, and accessing community resources. To test this intervention, 400 women were enrolled and randomly assigned to regular care, or to bimonthly nurse visits during pregnancy and for the first two years of the child's life.

The study examined the impact of the interventions not only on outcomes related to pregnancy and early childhood, but also on outcomes as many as nineteen years after the intervention. The study used a wide range of data sources including both primary data collection—interviews with study participants at multiple points in time, blood tests, cognitive and psychological testing—and also linked to administrative data such as medical records, school records, records for social services programs and records from Child Protective Services.

The study found significant impacts of NFP for both the mothers and children, which appeared early and continued through the latest (19-year) follow up. The results were often strongest for the highest-risk mothers, such as low-income, unmarried teenagers. Children in the intervention group had higher birth weight and, in early childhood, fewer emergency room visits and injuries requiring medical attention (Olds et al., 1988; Olds, Henderson, & Kitzman, 1994). For mothers, the intervention reduced subsequent pregnancies and increased maternal return to school and employment in the first four years post-partum (Olds et al., 1988). At the fifteen year follow-up, treatment was seen to significantly reduce verified Child Protective Service cases involving either the mother (as the perpetrator) or the child (as the victim) of violence (Eckenrode et al., 2000). At the nineteen year follow-up, girls born to mothers who participated in the NFP were significantly less likely to be arrested or convicted of a crime and had fewer children and less Medicaid use (Eckenrode et al., 2010).

Following the success of the initial NFP program, the program was expanded to two additional sites—Memphis, Tennessee in 1988 and Denver, Colorado in 1994. In contrast to the Elmira trial, the Memphis trial took place in an urban setting and enrolled primarily disadvantaged black women. The Memphis trial found effects of smaller magnitude, but the same direction as the Elmira trial (Kitzman et al., 2000; Olds et al., 2007; Olds et al., 2004). The Denver site included an additional treatment arm which consisted of the same NFP intervention, but delivered by "paraprofessionals" (high school graduates) trained to administer the program rather than by nurses with BSN degrees. This study found that the nurse-delivered intervention had similar effects as in the previous two trials, but that the program was not as successful when delivered by the paraprofessionals (Olds et al., 2004).

Back to section

Long-term follow up

The impacts of an intervention may well vary over the time period studied. The immediate impacts of an intervention—over the first few weeks, months or years—are, of course, important. They may be particularly relevant to providers and others responsible for the short-term costs and health of patients. We have already noted that the real-time nature of many existing administrative data sources can be useful for getting close to immediate results of an intervention. However, understanding the longer-run impacts of healthcare delivery interventions is naturally crucial for informed policy decisions; in many policy cases, it is arguably of greater interest than the more immediate results. Administrative data is also extremely useful for longer-term (e.g., decades-long) follow-ups.

Most of the randomized evaluations of healthcare delivery we reviewed measured outcomes for one year or less (see Table 6). This relatively short typical follow-up period is likely strongly related to another finding from our review: 85 percent of studies relied on primary data collection. Primary data collection is not only expensive, but makes very long-run follow ups more difficult, since it requires being able to locate and collect information on study participants decades after the intervention.

It can sometimes be much easier to link participants to administrative data to observe some long-term outcomes. One example where administrative data proved invaluable in tracking long-run outcomes of an intervention—and where these results differed strikingly from what would have been predicted from shorter-term follow ups—is the case of Project STAR, a randomized intervention in which approximately 11,500 students and teachers were assigned to different classrooms. Short run improvements in test scores from random assignment to higher quality classrooms in kindergarten through 3rd grade faded out by 4th through 8th grade, but when researchers linked the original study participants to their adult tax records, they found that random assignment to higher quality classrooms in kindergarten through 3rd grade increased earnings at ages 25-27 (Chetty et al., 2011) (See spotlight on Project STAR).

Moreover, the cost of long-term follow up in administrative data may be trivial compared to the intervention cost and/or initial data collection costs. For example, starting in 2007, the Social Security Administration conducted a randomized trial of providing accelerated access to Medicare for recipients of disability insurance. Typically, individuals deemed eligible for Social Security Disability Insurance (SSDI) start receiving payments immediately and gain health insurance coverage through Medicare after two years. In the AB Demonstration project, around 1500 newly approved beneficiaries were randomized into two groups: to the standard Medicare start date, or to immediate Medicare coverage (accelerated benefits). Results from primary data collection—specifically surveys of study participants during the two-year waiting period for Medicare—found that random assignment to health insurance increased healthcare use, reduced unmet medical need and out-of-pocket spending, and improved self-reported health (Michalopoulos Wittenburg, Israel, & Warren, 2012; Weathers & Stegman, 2012). Utilizing existing administrative data would allow for study of the longer-term impacts of the accelerated benefit, as well. In particular, now that both the treatment and control arms have passed through the two-year waiting period and are all covered by Medicare, it would be feasible to study in Medicare claims data what the longer-term effects of not having a two-year “coverage gap” are on healthcare use and health. This would require neither additional intervention costs nor primary data collection costs.

Spotlight on Value of Long-term Follow-up: Project STAR

Project STAR was a randomized intervention conducted at 79 Tennessee schools from 1985–1989. Some 11,500 students and their teachers were randomly assigned to attend either a small class (average size of 15 students) or a standard class (average size of 22 students). In general, students remained in their randomly assigned classes in kindergarten through 3rd grade, after which they all returned to standard-sized class rooms in 4th grade. The original analyses of this experiment offered encouraging results, finding that random assignment to a small class or to a higher quality classroom in the same school increased student test scores during the intervention (i.e. in kindergarten through 3rd grade). However, test scores are not an end in and of themselves. An improvement in test scores is of interest primarily because it may portend improvements in lifetime outcomes such as college attendance, employment and earnings. Here the picture looked bleaker, as follow-up beyond the time frame of the intervention (3rd grade) suggested that the initial improvements in test scores faded out by 8th grade. Since the ultimate question concerns the impact of educational experiences on adult outcomes, researchers subsequently linked the original STAR participants to administrative data on their tax returns from 1996–2008. This allowed them to study the impact of the experiment on participants' outcomes as late as ages 25-27. They found that being randomly assigned to a small class or to a higher quality classroom in the same school improved a variety of markers of adult success, including the probability of college attendance, of attending a higher quality college, and of home ownership and the quality of the neighborhood of residence. Moreover, they found statistically significant effects of a higher quality classroom on adult earnings. For example, they estimate that students randomly assigned to a classroom that is one standard deviation higher in quality earn 3 percent more at age 27. One potential reason for this long-run effect, despite the fading out of test score improvements, is improvements in non-cognitive skills (such as effort, initiative, and lack of disruptive behavior), which researchers found persisted through 8th grade, unlike the test score improvements (Chetty et al., 2011).

Back to section

Measuring a broad range of outcomes

A critical design question for any intervention study is what outcomes are studied. We have several recommendations.

1. Examine both healthcare costs (or use) and health outcomes.

As we saw in Section IV, randomized trials of healthcare delivery most commonly measure health or behavior as an outcome; measurement of healthcare use or costs is much less common—as is measurement of both. This is problematic, because, a complete evaluation of a program requires understanding both its costs and its benefits.

The primary purpose of the healthcare system is to protect and improve health. The "efficiency" of the healthcare system refers to the costs of these health improvements. Interventions that improve health while lowering costs improve healthcare efficiency. It is unrealistic, however, to expect that all interventions will simultaneously improve health and reduce costs. Interventions that improve health without increasing costs also improve the efficiency of the healthcare system, as do interventions that reduce costs without harming health. There are more difficult trade-offs to be made in other situations, such as those that improve health at some increase in costs. Whether such interventions are judged to improve the efficiency of the health system will depend on what societal value is placed on the health improvements relative to the cost increase.

To evaluate whether a program improves the efficiency of healthcare delivery, we therefore need to know both the cost of the program in terms of increased healthcare utilization and the impact of the program on health. This is not to say that programs are only of interest if they are cost-saving. Indeed many cost-saving programs might be of very little interest if they harm health. Programs will be of interest if they improve health without increasing costs, if they reduce costs without harming health, or if they improve health without increasing costs "too much," or vice versa. What is "too much" is a thorny and difficult issue that must be carefully considered and discussed in a specific context. Challenging trade-offs are involved in assessing programs that increase (or decrease) both health and cost. Those trade-offs, however, cannot be evaluated unless both health and cost outcomes are measured.

Finally, of course, a complete evaluation should consider the costs of the intervention per se. Some interventions—such as adopting checklists—can be relatively cheap while others—such as providing home visits—can be quite expensive, which is important to bear in mind in evaluating efficiency or cost-effectiveness.

2. Investigate potential non-health sector impacts

Many healthcare interventions may have indirect impacts on non-health outcomes. For example, interventions that improve patients' health and/or reduce their time spent with healthcare providers may potentially affect economic measures such as employment and earnings or participation in government transfer programs. Depending on the intervention, these may be important sources of costs and benefits, and ideally would be measured as well.[3]

As an example, the Oregon Health Insurance Experiment, a randomized evaluation of the impact of covering low-income uninsured adults with Medicaid, not only found that Medicaid increased healthcare use and improved self-reported health and depression, but also that it improved individuals' financial well-being; in particular, it reduced out-of-pocket medical spending and medical debt (Finkelstein et al., 2012). Measuring financial health outcomes involved linking administrative records to non-health sector outcomes (such as credit report data). Such linkages to administrative data on non-health sector outcomes seem quite rare; none of the 99 healthcare delivery RCTs in top medical journals we reviewed did such a linkage, although some included non-health sector outcomes in surveys.

3. Investigate potential external effects of intervention across space and time

In addition to investigating effects on a range of outcomes, it is also desirable to examine potential external effects of the intervention across space and time. What exactly these might be depends on the intervention. We discuss several here.

Multitasking. An important question is how an intervention that targets a specific behavior or outcome affects other, non-targeted behaviors or outcomes. For example, a natural set of instruments to explore is the impact of financial incentives to achieve targeted benchmarks; economic theory suggests that such incentives should increase efforts toward the targeted indicator (Holmstrom, 1979). But at the same time, one must consider the possibility that efforts allocated towards the targeted indicators may come at the expense of efforts on the non-targeted indicators. This is typically referred to as the multi-tasking problem (Holmstrom & Milgrom, 1991). Encouraging pediatricians to cover the importance of bike helmets in a 20-minute well child visit may seem helpful, but the net effect may actually be harmful if it means they spend less time covering the importance of seat belts. Of course the opposite may also be true: effort may be

[3] The converse may also be true: economic interventions may have important health consequences. For example, some of the most important long-run benefits from the Moving to Opportunity experiment – a randomized housing mobility experiment in which low-income families in public housing in distressed communities were offered an opportunity to move to private-market housing in less distressed neighborhoods – were improvements in mental and physical health, rather than on economic self-sufficiency (Ludwig et al. 2011b, 2013).

substituted toward non-incentivized outcomes if they are complements for the incentivized outcome. A full understanding of the impact of an intervention requires considering such potential spillover effects—either positive or negative—on non-targeted outcomes. In developing countries, several randomized evaluations aimed at targeting health improvements have investigated such potential spillovers (Bloom et al., 2006; Olken, Onishi, & Wong, Forthcoming).

Diminishing returns across interventions. In addition to spillovers from a targeted intervention to non-targeted behaviors or outcomes, there are also potential spillovers across interventions. For example, issues of information overload may result in the first information campaign to a population being more effective than a subsequent one. Similarly, "alarm fatigue" can kick in when there are too many systems attempting to gain the attention of clinicians. The first prompting system introduced to an electronic medical record may effectively change clinical behavior, but when most prescriptions or orders generate a prompt, clinicians will soon learn to click through without reading the warning. This is an important element to consider in design or to test for directly.

Transitory vs. permanent effects. A related issue is whether an intervention, if successful, can establish a new equilibrium practice or whether it must be continually administered. For example, consider an intervention which has an effect on changing the behavior of some actor (for example, by providing financial incentives for the behavior). If one follows the behavior of the study participants after the intervention ends (and the financial incentive is taken away) do they maintain their behavior at the new level? Revert back to their pre-intervention levels? To something in between or to something below even their pre-intervention level? For example, one early trial of paying people for smoking cessation found a significant difference in the initial quit rate (16.3 percent in the treatment group versus 4.6 percent in the control group), but the differences were not sustained at six months (Volpp et al., 2006). A later trial was designed to account for this with additional payments for sustained abstinence (Volpp et al., 2009), but the degree to which gains can be maintained beyond the intervention period remains unclear.

It is easy to say, of course, that a broad range of health sector and non-health sector outcomes should be measured. Carrying this out can pose challenges. One challenge is simply that of measurement. Fortunately, many of the outcomes we described are measurable at relatively low cost in administrative data, a point we discussed above. Another challenge may be that some event is important but rare. A study may therefore be powered up to detect the impact on one set of outcomes but not another. Consider, for example, a program that moves care for patients with diabetes from endocrinologists to nurses. Such a program will likely be cost-saving because the nurses' time is so much less expensive than the endocrinologists', and health outcomes such as glycemic control could be measured to assess whether health is harmed. There may, however, be rare bad outcomes (such as diabetic retinopathy or other complications of diabetes) that become more likely and which even a reasonably large study would not be powered to detect. This is analogous to the problem of rare but serious adverse effects of drugs, which often are not detected at the clinical trial phase.

Design for illuminating mechanisms

The best-designed trials in healthcare delivery will allow us to learn not only if a specific intervention works, but also why and how it works. This is somewhat different than trials of medical technologies where it is often sufficient to know that a treatment is effective, even if the underlying mechanism is not yet clear. In healthcare delivery, however, the interventions are often multi-faceted, specific to the original setting, and cannot be as easily reproduced as a drug or a device. This means a deeper understanding of the mechanism of action is very helpful to other institutions looking to adopt aspects of a successful program.

While challenging, there are several ways to design trials so that they will illuminate the mechanism as well as the effectiveness of an intervention. One approach is to include multiple treatment arms. We see that this approach is not the norm; Table 5 above showed that only about a quarter of healthcare delivery RCTs included multiple arms. Involving multiple arms can allow for testing of different "dosages" of the intervention or different components. It can also allow examination of the relative efficacy for different approaches. For example, is it better to change incentives, either financial (e.g., consumer cost-sharing, provider pay for performance) or non-financial (e.g., naming and shaming, ordeals and hassles), to provide information, to change the default option (Madrian & Shea, 2001), or to provide other "nudges" (Thaler & Sunstein, 2009)? The answer will of course depend on the context, but testing multiple approaches in different settings can help illuminate which situations may be best suited to which approaches. Another potential question is whether an actor should be given information or an incentive that is tied to his activities (inputs) or his results (outputs). Economic theory offers some general guidance (Holmstrom, 1982); this is a fruitful area for further empirical study.

In testing multiple treatment arms, it may also be interesting to assess the extent to which different instruments or interventions on different actors are complements or substitutes. For example, does the provision of information together with an incentive have more or less of an effect than the sum of their effects when done separately? Does an intervention on physicians and patients aimed at changing some aspect of healthcare choice have more or less of an effect when done together than the sum of their effects when done separately?

Another way to help inform our understanding of the mechanism—and hence the portability of an intervention to other contexts—is analysis of the impact of the intervention on different populations. For example, subgroup analysis can identify if the intervention is most effective on the very sickest patients (or those with a given history, or the youngest doctors, etc.), which can suggest hypotheses of how the intervention is working. The question of who responds most to the intervention is not limited to simple subgroup analysis, however. For example, comparing providing information (or changing incentives) for providers to doing the same for patients can shed light on the relative roles of the two groups in deciding on appropriate care. Similarly, care is often team-based, and testing different changes for the individuals on the team compared to the whole team illuminates the way the team works together.

More generally, interventions are ideally designed to shed light on theories for why they have the effect they do. This allows trial results to be put into a more general context. The design of the experiment can be crucial for maximizing what can be learned not only about one specific policy but also about broader theoretical mechanisms that can inform the design of more effective policies more generally.

As emphasized by Ludwig et al. (2011a), experiments designed to test or examine theories or "mechanisms" behind effects may often have greater policy value in terms of their ability to inform sensible policy design going forward than more "black box" (or A/B trial) policy evaluations that test the effect of a given intervention or policy directly. Most policy interventions are, naturally, bundles of a multi-faceted approach to a problem. Learning through a randomized evaluation *whether* a bundled / multi-faceted intervention does or does not have given effect is certainly a very useful piece of information. But without knowing *why* it has that effect, it may be difficult to successfully replicate in other contexts which will, by necessity, slightly differ in terms of e.g. the exact implementation or the target population.

Even interventions that directly test theories in settings that do not correspond to any (actual or contemplated) "real world policy" may provide useful guidance for sensible policy design in other contexts. For example, Ludwig et al. (2011a) describe how the "broken windows" policing policy—which encourages police to pursue minor crimes like vandalism on the theory that leaving it untouched gives the appearance that no one is paying attention to more serious crimes—has been experimentally tested not by randomizing the allocation of police who pursue vandalism, but rather by experimentally altering the presence of cars with broken windows in neighborhoods.

Another way to learn about mechanisms is by pooling results learned across a set of closely related studies, rather than designing from within a study. For example, there have been a number of RCTs investigating the impact of financial incentives on investments in health behaviors (See spotlight on Financial Incentives for Patient Health Behaviors). The evidence seems to suggest that even small financial incentives can have effects on investments in individual health behaviors such as weight loss, smoking, and drug adherence, although there are still open questions regarding the magnitude and permanence of these effects.

One corollary is that experiments that are actively designed by researchers will often shed more light on mechanisms than ones that come about for other reasons. An interesting contrast that helps illustrate this point is the contrast between two randomized evaluations of the impact of health insurance coverage. The 2008 Oregon Health Insurance Experiment came about due to the interests of Oregon policymakers to allocate a limited number of Medicaid slots in a manner that they perceived as fair, and was not designed to shed light on mechanisms. As a result, although the findings have received much attention, they leave much to be debated in terms of whether an alternatively designed Medicaid program could achieve most of the benefits found at the same cost, or more benefits at the same cost (see e.g., (Douthat, 2013)). By contrast, the RAND Health Insurance Experiment (Aron-Dine, Einav, & Finkelstein, 2013; Newhouse & the Insurance Experiment Group, 1993) was prospectively designed by researchers to shed light on the tradeoffs involved in setting cost-sharing. It used multiple arms to randomly vary the cost-sharing features of health insurance that individuals received. It has been widely used in policy and academic analysis for discussions of optimal cost-sharing designs.

Spotlight on Testing Mechanisms: Financial Incentives for Patient Health Behaviors

In the spirit of "the exception to every rule", one area where there have been a relatively large number of RCTs testing a specific mechanism is the area of financial incentives to patients designed to change their health behaviors. Trials exploring whether financial incentives can motivate individuals to adopt healthier habits have been conducted across a range of habits including smoking cessation, weight loss, and medication adherence (Kimmel et al., 2012; Kullgren et al., 2013; Long, Jahnle, Richardson, Loewenstein, & Volpp, 2013; Volpp et al., 2006; Volpp et al., 2009). Each study had differences in terms of how the incentives are structured, but they share the common practice of paying for specific actions or for meeting specific targets.

Typically, the financial incentives involved are relatively small – on the order of $100 – although some used incentives as high as $750. In the most straightforward design, payments are made to individuals for taking a specific action, such a completing a smoking cessation course, or for reaching a target, such as having a urine cotinine test showing no evidence of smoking (Volpp et al., 2006; Volpp et al., 2009). Variations on this include group incentives (Kullgren et al., 2013) or more complex payment structures such as lotteries (Kimmel et al., 2012; Volpp et al., 2008). These trials are typically small (~100 participants) and have examined relatively short-term effects (less than 1 year). Effectiveness of the incentive interventions can depend on the design of the incentive and on the time horizon.

So far, the evidence seems to suggest that financial incentives can change people's behavior for the duration of the intervention, but the gains are not always sustained beyond the end of the incentives. The exception is interventions on smoking cessation, where the changes appear more persistent. In one randomized trial testing financial incentives for completing a smoking cessation program and for quitting smoking (Volpp et al., 2009), there were significant differences in the quit rate immediately post-intervention (14.7 percent vs. 5 percent) and six months after the incentives ended (9.4 percent vs. 3.6 percent).

Back to section

Attention to scaling

An additional challenge in designing healthcare delivery trials is thinking about the scalability of the intervention itself. Often the intervention tested in the trials that are done is a very specific program created by a handful of motivated people in a specific setting. Even if the intervention is found to be effective, there remains a problem of whether it can be adopted at scale. Programs may not adapt well to different populations, different providers, or different settings. A clear understanding of the mechanisms by which the program had its impact can be very important in assessing the likelihood of success in other settings, as well as making sure the key elements of the program are maintained in other applications.

However, even if a successful intervention in one setting is judged appropriate for other settings, there are many potential reasons why it may not easily scale. At the very simplest level, it is important to understand whether the program can be replicated in other settings. One reason why replication may be difficult is if the intervention is a complex bundle of services, so that fidelity in implementing the same intervention at other sites can be difficult to achieve. This was, for example, part of the difficulty involved in replicating the successful employment and earnings results of the Center for Employment Training (CET) job training programs for high school dropouts in San Jose, California at other sites around the country (Miller, Bos, Porter, Tseng, & Abe, 2005). Another reason may be that the sites that initially adopt an intervention, or choose to evaluate it via an RCT, may have higher impacts of the intervention than average. Allcott (2012) presents evidence of such site selection bias in the context of the RCTs conducted on Opower energy conservation programs. In this program, residential electricity consumers are mailed "home energy reports" that compare their energy use to that of their neighbors and provide energy conservation tips. Using data on 111 randomized controlled trials of Opower programs involving almost 9 million US households, Allcott (2012) shows that the evidence from the first ten RCTs substantially over-predicts the efficacy of the next 101 sites, due to the fact that both the communities that initially sought out the intervention and the sites the program managers initially chose to target tended to be more responsive to the intervention.

In addition, even when a successful intervention can be replicated in multiple settings, diffusion and adoption of the intervention may be limited. The history of medicine is littered with examples of successful interventions that were slow to diffuse. The British Navy waited almost 200 years to take steps to increase vitamin C intake among its sailors, despite extremely high sailor mortality from scurvy and experimental evidence of the importance of vitamin C in preventing it (Berwick, 2003). Lister's 19th century discovery of the importance of sterile practice in preventing sepsis—a leading cause of death from surgery in his time —took a generation to be incorporated into routine medical practice (Gawande, 2013). In modern times, studies have found that it takes, on average, about 17 years for new knowledge generated from clinical trials to diffuse into medical practice (Agency for Healthcare Research and Quality, 2004). There is a large literature exploring the forces behind the often-slow diffusion of technologies and knowledge, not only in medicine but in many other fields as well. For example, Jack (2013) provides an extensive discussion of the slow diffusion of technologies in agriculture in much of the developing world, and interventions designed to reduce barriers to adoption.

As a result, even when RCTs have demonstrated the efficacy of an intervention aimed at improving the efficiency of healthcare delivery, they are often not adopted. For example, across a large number of RCTs, post-discharge follow-up of chronic heart failure patients involving home visits by nurses has been shown to reduce mortality and hospital readmissions (See spotlight on Reducing Re-admissions) (Takeda et al., 2012). Moreover, as discussed above in Section III above, hospitals are now being offered financial incentives if they can reduce readmissions. Still, it does not appear that there has been widespread adoption of the practices found successful through these randomized evaluations. As with agricultural technology, an important area for RCTs in healthcare delivery is to better understand barriers to adoption and the efficacy of different types of interventions (e.g. financial incentives, information campaigns, IT-based solutions, etc.) in encouraging the spread of adoption of successful healthcare interventions.

In some cases, dedicated effort can be effective in scaling-up an intervention found effective through a randomized evaluation. For example, a randomized evaluation of a primary school deworming project in Kenya showed that school-based deworming reduced the incidence of infection by 25 percentage points and school absenteeism by 25 percent at a cost of 50 cents per child per year (Miguel & Kremer, 2004). In response to these findings, a non-profit organization, Deworm the World, was launched in 2007 to promote and sustain implementation of deworming. Since then, an estimated 59 million people have been reached by school-based deworming efforts coordinated by Deworm the World and its partner organizations (Lab).

In other cases, the effort to scale up an intervention has been led by the government. For example, a large problem with redistribution programs in developing countries is making sure the transfer reaches the intended beneficiaries. A randomized evaluation of an identification card system to prevent "leakage" of a large targeted transfer program in Indonesia showed that the cards were effective in getting the subsidies to the intended households without generating social conflict between beneficiaries and non-beneficiaries, which was a main concern of the government in considering such a system (Banerjee, Hanna, Kyle, Olken, & Sumarto, 2014). As a result, the Government of Indonesia decided to scale up the identification card program and an estimated 65 million people have been reached by this improved distribution strategy (Abdul Latif Jameel Poverty Action Lab, 2014).

Spotlight on Scale Up Challenges: Reducing Re-admissions

One open question in healthcare delivery is how to organize care for individuals immediately following hospital care. Many patients end up back in the hospital relatively quickly after being discharged. Among Medicare beneficiaries, almost one-fifth are readmitted within 30 days and about one-third are readmitted within 90 days (Jencks, Williams, & Coleman, 2009). Patients with heart failure, a relatively common reason for initial hospital admission, are readmitted within 30 days about a quarter of the time. Fortunately, for patients leaving the hospital with chronic heart failure, many approaches to post-discharge care have been tested using randomized designs, and the accumulated evidence suggests that a case-management approach is successful at reducing hospital readmissions and subsequent mortality.

A recent Cochrane review and meta-analysis summarizes 25 randomized trials for post-discharge care for heart failure patients, involving over 6,000 patients total (Takeda et al., 2012). Six of the studies tested specialized “heart failure clinics” and found no significant reductions in hospital readmissions. Seventeen of the studies tested different case-management approaches, where patients were provided with additional follow-up through phone calls and home visits, typically provided by a nurse. Overall, a meta-analysis of these studies found the case-management approach substantially and significantly reduced hospital readmissions over both 6 months (OR 0.64; 95% CI: 0.46 to 0.88; P=0.007) and 12 months (OR 0.47; 95% CI: 0.30 to 0.76; P=0.002). More importantly, the case-management approach also reduced all-cause mortality over 12 months (OR 0.75; 95% CI: 0.57 to 0.99; P=0.011).

Despite this body of evidence for the benefits of a case-management approach for chronic heart failure patients leaving the hospital, adoption has remained relatively limited. Part of this may be natural hesitancy to implement an intervention until there is stronger evidence in its favor; in fact, a 2005 version of the same review found only non-significant reductions in mortality (Taylor et al., 2005). However, the potential harm to patients from the intervention seems limited, and it seems likely that the answer lies in financial and practical impediments to adoption.

Practically, adoption may appear rather complicated to a hospital. Hospitals have not typically provided this kind of post-discharge care, and unlike writing a prescription for a medication, this approach requires hiring, training and supervising staff in new roles. The above review included seventeen studies of case management, and there was quite a bit of variation in approach and results. Furthermore, many other studies of different variations and for differing conditions have been done under the rubrics of case management, care management, coordination of care, and transitional care, and some high profile ones found no improvements (McCall & Cromwell, 2011). All this complexity may discourage hospitals from implementing even a limited program for discharged heart failure patients, especially when combined with the financial impediments.

Financially, the case-management approach for heart failure patients leaving the hospital may reduce healthcare costs – particularly hospital readmissions – net of the intervention costs, but few of the studies actually measure costs of care as an outcome. Moreover, paying for post-discharge case management requires hospitals to incur a new up-front cost (for example, hiring specialty nurses to provide the care) likely without increased revenue. Indeed, since one effect of the intervention appears to be reduced readmissions, the hospital might also face reduced revenue. However, new penalties may change the financial calculations and encourage adoption. Starting in 2013, hospitals now face Medicare payment penalties for excess readmissions for three conditions, one of which is heart failure (Centers for Medicare and Medicaid Services, 2014b).

Back to section

Registering, pre-specifying, and disclosing

Prospectively registering a study, pre-specifying the planned analyses, and sharing data and other materials are all elements of good research practice. The benefits of such practices are not limited to randomized evaluations but may be particularly important in this setting. Analyses of randomized interventions appear to be at the forefront of adopting these practices, as well as at the forefront of advocating for them (Miguel et al., 2014).

Prospective public registering of the design and primary outcomes of a randomized evaluation helps create a public record of studies that is not subject to concerns about publication bias. Registration through a centralized public registry can also help build a valuable resource for meta-analyses and for information on current research practices. In medicine, the US National Institutes of Health runs a centralized, public registry of clinical trials using human participants around the world (www.clinicaltrials.gov). In 2012, the American Economic Association created such a centralized registry for randomized controlled trials in the social sciences (www.socialscienceregistry.org).

The International Committee of Medical Journal Editors (ICMJE) now requires all randomized trials that began recruitment after January 1, 2005 to have been prospectively registered as a condition of publication (De Angelis et al., 2004). In economics, it has become more commonplace in recent years for researchers conducting RCTs to write detailed pre-analysis plans, attempting to completely specify the exact and full set of analyses to be run and results to be produced (see e.g. Alatas, Banerjee, Hanna, Olken, & Tobias, 2012; Casey, Glennerster, & Miguel, 2012; Finkelstein et al., 2012; Olken et al., Forthcoming).

At the other end of the research process, producing clear and accessible public use data files from the randomized evaluation can enhance the total amount that can be learned from the effort. Such public use files encourage both important replication exercises as well as open up the possibility for additional analyses, both with the existing data and through linking the study participants to other pre-existing data to study both other effects of the intervention and potentially longer run outcomes than in the original study.

Of course, registration, pre-analysis and disclosure are not always feasible, nor always desirable. As discussed below, sometimes random assignment is designed for policy purposes (such as fairness) rather than research purposes, and in such situations there may be limits to the extent of prospective registration and pre-specification depending on when in the process researchers become involved. Pre-specifying the analysis that will be run has great value in terms of reducing (or at least making clear the extent of) data mining and multiple hypothesis testing but can come at a significant time and cognitive burden, especially in the case of a trial where there are a variety of possible initial results, each of which suggest a different subsequent line of analysis and exploration (Olken, Forthcoming). Finally, it is not always administratively feasible; public disclosure of data files is sometimes prohibited by data use agreements necessary to be able to conduct the original analysis.

VI. When to Use RCTs

We have discussed many advantages of conducting randomized evaluations to test healthcare delivery innovations, but there are, of course, many challenges as well. These range from ethical considerations, to practical ones about time and cost, to theoretical ones about what is learned from such trials. Understanding the challenges as well as the advantages is important in deciding when using a randomized evaluation is an appropriate tool.

Ethics of rationing

One commonly raised concern about randomized evaluations is whether such investigations are ethical. A standard framework for considering this question is that of clinical equipoise, asking if there is sufficient uncertainty about whether or not the intervention is beneficial. Of course, the hardworking individuals who have developed a program or intervention may well believe it will prove beneficial, just as the scientists who develop a new therapeutic compound may well believe it will prove effective. But if there is not compelling theory or evidence that would suggest it would be beneficial, there may well be clinical equipoise.

Randomized evaluations are particularly appropriate when programs are oversubscribed, are scheduled to be rolled out in a gradual fashion, or are initially tested with pilot programs. In healthcare delivery, it is common that as programs start, or even if they are long-standing, they are reaching only a small fraction of the individuals who might benefit. In these cases, where there are capacity constraints, random assignment to the program may actually be a more equitable way to allocate slots than the ad hoc ways previously used. In fact, sometimes these sort of random allocations are chosen even without an eye to the research advantages. Perhaps the most famous example is the Vietnam Draft lottery, which has been used to study the impact of military service on post-military mortality (Hearst, Newman, & Hulley, 1986) and lifetime earnings (Angrist, 1990). Various charter school networks use lotteries to select students (Foundation, 2014), and the random selection used in the Oregon Health Insurance Experiment was designed in conjunction with stakeholders to address concerns about fairness rather than to facilitate research (Allen, Baicker, Taubman, Wright, & Finkelstein, 2013) (See spotlight on Oregon Health Insurance Experiment).

Spotlight on Randomization for Fairness: Oregon Health Insurance Experiment

Sometimes randomization is chosen for reasons unrelated to research yet can still provide a valuable research opportunity. In 2008, Oregon decided to allocate limited slots in a previously closed Medicaid program by random selection from a waitlist, deeming this the fairest way to distribute the scarce slots. As this random assignment also created a rare research opportunity, the state government partnered with academics to study the impact of expanding public health insurance (Allen et al., 2013).

The Oregon Health Insurance Experiment followed 75,000 individuals who signed-up for the Medicaid waitlist. Of these, about 30,000 were selected in the lottery and mailed a Medicaid application; 10,000 eventually enrolled. The study included primary data collection—mail surveys, in-person interviews, anthropomorphic and blood pressure measurement, dried blood spots—and links to administrative data—hospital and emergency department records, credit reports, enrollment in social safety net programs, and employment and earnings records. The random assignment of the lottery allowed for the estimation of the casual impact of Medicaid coverage in the first 1-2 years (Baicker, Finkelstein, Song, & Taubman, 2014; Baicker et al., 2013; Finkelstein et al., 2012; Taubman et al., 2014).

Expanded Medicaid coverage increased use of healthcare services across the board—including hospital admissions, emergency department visits, doctor visits, prescription drugs, and perceived access to and quality of care. Medicaid decreased financial strain (including medical debt and collections) but did not significantly change employment or earnings. Expanded Medicaid coverage led to improvements in self-reported health and to a substantial reduction in depression. There were no statistically significant effects on measured physical health (blood pressure, cholesterol, and glycated hemoglobin).

These findings dispelled some claims of leading public policy figures and healthcare experts—such as the claim that Medicaid had no proven benefits, or that Medicaid would reduce use of the Emergency Room—and has corroborated others—such as that Medicaid increased use of healthcare, including primary care and preventive care, and increased healthcare spending. As there are currently important policy decisions being made about the expansion of Medicaid benefits, these findings generated considerable media attention and have subsequently been used in government reports on the likely effects of expansion (Council of Economic Advisers, 2014).

Back to section

Time and cost considerations

Another commonly raised concern about randomized evaluations is that they are prohibitively costly in time and money. It is important, however, not to confuse the time and monetary costs that are associated with some research projects with the *incremental* cost of doing that research as a randomized evaluation. Although doing interventions and collecting data often can be expensive, and prospectively following individuals often takes a long time, adding randomization need not, and often does not, add substantially to the costs, and in fact can reduce costs, particularly on the back end, where the analytical work is much simpler due to the randomized design. Moreover, the use of administrative data can not only dramatically reduce the monetary costs of evaluation, but can also reduce the time lags by allowing the researcher to get results essentially in real-time, including, for example, daily updates on any hospital re-admissions for study participants (Finkelstein et al., 2014).

An important issue in understanding the time and cost considerations for an RCT is when the randomization will be conducted. The standard in trials of new medical technologies (drugs or devices) is to recruit individual patients into the study, screen each for eligibility, obtain individual informed consent from each, and then only randomize those eligible individuals who agree to participate. In trials such as this, the burden of screening and recruiting individual patients means the sample sizes are often quite small. On the other hand, adherence to the assigned protocol is typically relatively high and attrition from the study is typically relatively low.

In the context of healthcare delivery randomized trials, there is often an alternative approach, where participants are randomized into being offered the program, rather than into the program itself. Study participants randomized into the treatment are then offered the program, screened for eligibility (if necessary), and then only take part in the treatment if they so choose. All individuals included in the random assignment are followed—including those who do not accept the offer of the intervention—often passively in administrative data, to study the outcomes. In trials such as this, large number of individuals can be randomly assigned, and there can be essentially no impact of the study on the administrators of the intervention, who simply receive a list of people to intervene on. However, take-up of the program (adherence to the assigned protocol) is typically low (See spotlight on Medicare Health Support Disease-Management Pilot). This low take-up or adherence in a trial does not interfere with obtaining estimates of causal effects (Angrist, Imbens, & Rubin, 1996). However, low take-up is detrimental to the statistical power of the study to detect effects of the intervention (see Appendix C).

For studies that are inherently small (when the intervention only applies to a limited set of people or there are capacity constraints), the gains in statistical power argue for enrolling participants and then randomizing them, as is typically done in medical trial.   Similarly, if follow-up requires active participation from individuals, randomizing after enrollment limits the sample to those most likely to participate. Enrolling after randomization may also be ethically required in situations where there is potential for more than minimal risk to participants, as each participant needs to give informed consent. In other situations, however, where there is only minimal risk to participants, and follow-up can be done passively in existing administrative records, an "encouragement design"—where individuals are randomly assigned to be offered or encouraged to use the intervention—may be ethically acceptable and may make a large-scale trial practically feasible. These "encouragement designs" can be used both to test the relative efficacy of different encouragements (financial or non-financial incentives, information, defaults or nudges) and to test the effects of the intervention itself.

A related time concern is whether the intervention must remain static for the length of the RCT. In practice, many healthcare delivery innovations are complex programs which are continually adapting and improving. This model of changing and tweaking the intervention poses challenges for any evaluation design—randomized or observational—as the evaluation will always be measuring the average effect of the intervention over the study period. To the extent that the intervention has very different effects early and late in the study as the result of adaptations, these differences may be obscured by examining only the average effect. This issue, however, is not linked to the randomized design. Randomization will always aid in identifying the causal impact of the program even when the program is changing over time. Sometimes, as in the case of drug trials, randomization is also linked to very narrowly specified protocols to clarify the intervention under study. This approach may not be as appropriate in healthcare delivery studies, where the interventions may be less easily forced into strict protocols.

Spotlight on Timing of Randomization: Medicare Health Support Disease-Management Pilot

Randomization need not occur prior to individuals agreeing to participate in a given program, as can be seen in the Medicare Health Support Disease-Management Pilot Program. To test using centralized call centers staffed by nurses to do disease management, Medicare randomly assigned which eligible beneficiaries were offered the support programs. The 242,417 Medicare beneficiaries who were included in the randomization were typically very sick and had, on average, more than one hospitalization annually and $15,000 in 2004 Medicare expenditures (McCall & Cromwell, 2011). If assigned to the intervention group, the beneficiaries were offered disease management support from one of eight private companies, and could choose to participate or not. In practice, participation among those offered the programs was quite high (over 75 percent for all eight companies), but intensity of participation varied substantially; the percentage of participants with more than 10 contacts with health support ranged from as few as 20 percent to as high as 80 percent.

The evaluation examined changes in the rates at which patients received certain medical screenings ("process-of-care" measures). It also analyzed the rates of hospitalizations and ER visits, both overall, and for a subset of conditions considered "ambulatory care sensitive" ("utilization of care" measures) (McCall & Cromwell, 2011). The study found that across the eight disease-management programs and five process-of-care measures, only 14/40 had any significant improvement for the intervention group. Moreover, these improvements were small in magnitude (1.3-3.1 percentage points). Only one company's intervention reduced the rate of growth of hospitalizations for any condition, and one other was able to reduce the rate of growth for hospitalizations for ambulatory care sensitive conditions.

This program shows the feasibility of randomizing very large numbers of beneficiaries prior to enrollment in the intervention program. Administrative data on Medicare beneficiaries was used to identify eligible candidates and to measure outcomes. The analysis was done on an intent-to-treat basis, identifying the effect of offering beneficiaries the health support. In this case, that may be the policy relevant parameter, but this could vary. It would also be possible to analyze the data using an instrumental variable approach to identify the effect of health support in those choosing to participate in the program when offered.

Back to section

Patient-level vs. organization or system level RCTs

As seen in Section IV, most RCTs on healthcare delivery randomize at the unit of the patient. For some questions, the necessary unit of randomization is larger. For example, a study of the efficacy of alternative interventions aimed at improving physician compliance with guidelines presumably needs to be randomized at the physician (rather than patient) level. Moreover, if physicians within a practice affect each other either through their behavior or through directly sharing information, such an intervention might in fact need to be done at yet a higher level of aggregation—the relevant physician "practice." Likewise, a study of the impact of improved electronic medical records might need to occur at the physician practice or hospital level. Efforts to randomize across these "larger" units—such as physician, physician practice, hospital unit, or hospital—can present logistical and financial challenges to achieving sufficient sample size. Coordination among multiple healthcare systems may well be needed to achieve sufficient sample size. Such cluster-randomization is not completely infeasible, however; some recent studies of ICU protocols to reduce hospital-acquired infections were randomized at the level of the ICU (Harris et al., 2013; Huang et al., 2013). More such efforts would be very valuable.

Of course, just because the intervention itself occurs at a large scale does not mean that the randomization must as well. New structures and organizations do not necessarily include all patients; more often they are created for some subset of patients (for example, those covered by a specific payer) and exist in parallel with the existing organizations and structures. At least in theory, then, some individuals could be randomized into the new structure or into difference incentives to enroll in the new system. We discuss this in more detail in Section VII below.

The converse is also true: even when the unit of intervention is small (e.g. a patient), the impact of the intervention may depend on the number of patients in the market (or relevant unit) affected. As a result, just because an intervention is at the patient level does not mean that the only relevant unit for randomization is the patient level. For example, the Oregon Health Insurance Experiment studied the impact of covering approximately 10,000 low-income uninsured adults in Oregon with Medicaid. This number, while large enough to measure the impact of Medicaid on many outcomes precisely, represents only a small fraction of the population of 4 million in Oregon, or even of the approximately 200,000 uninsured (Finkelstein et al., 2011). It therefore allows the researcher to detect effects of covering a given individual with health insurance, holding the general healthcare environment constant. The effects of a market-wide expansion in coverage of all or many of the uninsured may well be different (Finkelstein, 2007). They might be disproportionately smaller, for example, if patients encounter capacity constraints, such as a fixed supply of doctors and hospital beds, particularly in the short run. The effects of a market-wide expansion might also be disproportionately larger, if, for example, the market-wide change in insurance coverage causes changes in the general healthcare environment such as doctor practice style, or adoption of new medical technologies.

The challenges to designing a randomized evaluation that can capture effects of a sizable area-wide change in health insurance coverage (or some other policy) would presumably be substantial. In some contexts, however, researchers have been successful at designing studies to look at these broader effects. For example, Miguel and Kremer (2004) randomized de-worming treatments for children at the school (rather than student) level in order to capture the potential externalities on the education and health of untreated children at the school. Randomized evaluations can be designed to evaluate such spillover effects, by randomizing the *proportion* of individuals within the relevant unit that are assigned the treatment, as well as randomizing *which* individuals within the unit are assigned the treatment. Such designs have been effectively employed, for example, to study both the direct and indirect (displacement) impacts of job placement assistance on labor market outcomes for unemployed youth in France (Crepon, Duflo, Gurgand, Rathelot, & Zamora, 2013).

VII. Summary and Possibilities

Too few RCTs in US healthcare delivery are currently conducted. The value of more RCTs in the healthcare sector is very high and, fortunately, we see reasons to be optimistic that we will soon see many more RCTs in healthcare delivery; providers have increasing incentives ("skin in the game") to understand what is most effective at improving the efficiency of healthcare delivery, and the growing availability and sophistication of administrative data and information systems offer increasing opportunities for low-cost, high-quality RCTs. Attention to measuring a broad range of outcomes and to designing RCTs to test theories or uncover underlying mechanisms will further enhance their impact. Through our "spotlights," we have highlighted a few examples of RCTs on healthcare delivery interventions that have achieved some of these objectives.

We conclude by offering some questions on the efficiency of US healthcare delivery that could be explored through RCTs, along with some specific examples for each type of question. The discussion is not meant to be exhaustive or prescriptive. Rather it is meant to highlight the range of possible RCTs across interventions, actors, and outcomes. We emphasize in particular the potential for using RCTs to study the impact of interventions not only at the patient or provider level, but also the redesign of care to a specific population, and system-wide changes in practice. Some of these RCTs could be implemented quite easily with a single implementing partner (such as a hospital, pharmacy, employer, or insurer); others are more ambitious and require more coordination across multiple partners.

Choice of appropriate care

In almost every area of medicine, there is evidence that individuals do not receive all the care that is recommended while receiving care that is not recommended. The examples are too numerous to list; they run the gamut from concerns about underuse of preventive vaccines or medication for chronic conditions, to underuse and overuse of screening or invasive surgery. Physician professional societies, the US Preventive Services Task Force and other groups have long published guidelines for appropriate care; more recently under the Choosing Wisely initiative, professional societies have started publishing lists of commonly provided care that should be questioned (Choosing Wisely, 2014).

One can imagine a wide range of potential randomized interventions across patients and/or across providers to examine the impact of interventions designed to bring practice more in line with recommendations. These could use many of the tools discussed above: financial and non-financial incentives, information, defaults and nudges, decision support tools. Table 7 provides just a few examples of these kinds of potential RCTs.

Table 7. Choice of appropriate care

	Intervention	Randomization	Outcome
Flu vaccines	Provide information on the health benefits from flu vaccines Provide reminders or implementation prompt encouragements	Randomly assign patients or households. Alternatively, randomly assign at zip-code level	In claims data, examine impact on flu shot take up and downstream use of healthcare and health outcomes (for influenza, pneumonia, cardiovascular events, etc.).
Diagnostic imaging	Provide decision support information on the "appropriateness" of imaging when placing orders in Electronic Medical Record (EMR).	Randomly assign physicians	In EMR and claims data, examine impacts on scan rates, "appropriateness" rates, and downstream healthcare costs and health outcomes.
Colorectal cancer screening	Financial payment to individuals for getting recommended screening for colorectal cancer.	Randomly assign patients	In claims data, examine impacts on screening rates and downstream healthcare costs and health outcomes.

Tests of different approaches are most likely to arise in the context of a single medical condition and around a limited set of care choices and the results will therefore be most applicable to that condition and care choice. Ideally, however, through the accumulated evidence from multiple thoughtfully-designed studies, we can also learn some general principles about how we can likely best design systems that better encourage appropriate care. Moreover, any intervention that changes the likelihood of a specific treatment that is believed to be under- or over-used could then be used in an encouragement design to test the downstream effect of this treatment on patient health, healthcare utilization and healthcare costs.

Who provides care and where and how care is provided

As with questions about how best to encourage appropriate care, questions of how to provide the care efficiently are nearly limitless. For almost any area of medicine or any specialty, there are questions to consider of who should provide the care, where, and how it is provided. One very active area of randomized trials in this space is the provision of care management, accounting for 11 percent of the healthcare delivery RCTs we reviewed and including the Medicare Care Coordination Demonstration (Coburn et al., 2012; Peikes, Chen, Schore, & Brown, 2009). To name just a few other questions of these sort (also included in Table 8): can patients be safely and appropriately discharged earlier after caesarean section, does an in-hospital specialist consult improve post-discharge outcomes, and what are the impacts of providing patients with e-mail and phone access to primary care nurses and physicians? The specific options in the choice set may vary—depending on the condition, the relevant question may be whether care is provided by specialists vs. generalists, physicians vs. nurses or other allied health professionals, in hospital vs. outpatient, in home vs. in office, in person vs. by videoconference, etc. And for all of these, a natural set of questions is whether and when care in these different formats are substitutes or complements.

Table 8. Who provides care and where and how care is provided

	Intervention	Randomization	Outcome
Post-partum length of stay	Offer early discharge after Caesarian delivery with post-discharge home visit by a nurse and/or a lactation consultant.	Randomly assign patients	In EMR and claims data, examine impact on length of hospital stay, use of outpatient care, and post-discharge complications for mother and baby.
In-hospital pain specialist consultations	Do not discharge opioid-tolerant patients without an in-hospital consultation with a pain specialist	Randomly assign patients	In claims data, examine post-discharge use of pain medication, emergency department visits, and readmissions
Phone (or video) primary care visits	Offer "virtual" visits with primary care nurse practitioners and physicians	Randomly assign patients Alternatively, randomly assign primary care clinics	In claims data, examine impacts on use of virtual and standard primary care visits, and downstream healthcare costs and outcomes.

Efficient use of existing resources

Another important set of questions is whether with better practices, the healthcare delivery system could make more productive use of existing resources (be it doctors, nurses, hospital beds, operating rooms, or imaging machines). For example, can operations research techniques borrowed from manufacturing help to reduce the amount of time that expensive capital equipment is idle? Can more efficient scheduling and staffing algorithms reduce wait times for appointments and allow more patients to be seen? A related question is how to reduce wasteful and fraudulent behavior by healthcare providers, which is estimated to cost the US healthcare system nearly $200 billion each year, about $60 billion of which occurs within Medicare and Medicaid (Berwick and Hackbarth 2012).

Table 9 provides some examples of the types of RCTs that would fall into this category. Unlike many of the other topics, in this area, the intervention is more likely to be on a back-office system and may not be transparent to the patient and providers.

Table 9. Efficient use of resources

	Intervention	Randomization	Outcome
Outpatient appointment scheduling	Use machine-learning to predict no-shows to outpatient appointments and adjust number of patients booked	Randomly assign scheduling blocks	In administrative data, examine impact on patient wait times and physician time spent with patients
Emergency department staffing	Use queuing theory to adjust planned emergency department staffing	Randomly assign weeks	In administrative data, examine impact on probability of patients leaving without being seen, patient wait times, and use of staff
Outlier billing	Send letters to providers with unusual billing behavior	Randomly assign physicians or physician practices	In claims data, examine whether unusual billing by providers decreases, also spillover effects to other billing areas and if patients change providers

Insurance design

One area where there have been several large RCTs is insurance design and, in particular, the area of consumer cost-sharing. Evidence from randomized evaluations of cost-sharing in the RAND Health Insurance Experiment and the Oregon Health Insurance Experiment indicates that healthcare use rises when consumer out of pocket payments fall (Finkelstein et al., 2012; Newhouse & the Insurance Experiment Group, 1993). A recent large RCT even used this price-sensitivity to selectively adjust cost-sharing to encourage the use of recommended medications (Choudhry et al., 2011).

There is much less evidence on some of the contract and reimbursement insurance design ideas currently being explored. One such idea is the use of tiered or reference pricing systems which could be applied to coverage across providers or to coverage of different treatment options. Others include pre-authorization requirements and systems, and limited or high-value networks.

Table 10 gives some examples of potential RCTs in insurance and reimbursement design, including the use of reference pricing for CT scans, reimbursement rates for different types of care, and the extent of the provider network. Design of such RCTs presents some challenges, including often requiring coordination across multiple players (government, employers, hospital systems, providers). In addition, regulations may prevent employers or other payers from randomizing across individuals, even between actuarially equivalent plans. As a result, plans may need to be randomized at the employer-plan or employer level. The example of previous RCTs testing targeted consumer cost-sharing changes (Choudhry et al., 2011) and targeted physician pay-for-performance programs (Bardach et al., 2013; Petersen et al., 2013) offers a proof of concept that these types of RCTs may be feasible. Alternatively, one could consent individuals into participating in a study where they will be randomized into one of several different plans provided by the experiment, much as the RAND Health Insurance Experiment did over thirty years ago. Such ambitious trials may be needed to answer fundamental questions about the health insurance market.

Table 10. Insurance contract and reimbursement design

	Intervention	Randomization	Outcome
Reference pricing for CT scans	Cover only reference price for CT scans with patient responsible for balance	Randomly assign at employer level	In claims data, examine impact on use of CT scans, provider choice, and downstream healthcare costs and outcomes
Adjusting prices	Pay more for high-value care and less for low-value care	Randomly assign billing codes to readjustment	In claims data, examine impact on healthcare use, costs, and outcomes
Limited network plans	Cover only limited (high-value) network of providers	Randomly assign consenting individuals to different provided plans (actuarially equivalent)	In primary and administrative data, examine impact on healthcare use and costs, access, quality, health, and finances

System-wide innovations

An oft-heard concern about RCTs as a tool in improving the efficiency of healthcare delivery is that they are not feasible for the sort of system-wide innovations that may be needed. As a counterargument to this, we discuss in more detail two specific, distinct examples of system-wide issues and how RCTs might be used in those settings.

Hospital management

There is a great deal of variation across US hospitals in the quality of their management practices (Bloom, Genakos, Sadun, & Reenen, 2012) and in the health outcomes they produce with a given amount of inputs (Chandra, Finkelstein, Sacarny, & Syverson, 2013). There are many case studies on good management practices and cross-sectional evidence that better hospital management practices are associated with better patient clinical outcomes and better hospital financial outcomes (Bloom, Raffaella, & Van Reenen, 2014; Gawande, 2012; Leonhardt, 2009). A randomized trial of the introduction of better management practices on patient outcomes and hospital performance would tell us the direct causal impact of better management practices.

Such an RCT is wholly within the realm of possibility. Proof of concept has been provided by a successfully implemented RCT on management practices in textile plants. The researchers randomized the provision of free consulting on management practices to a randomly selected set of treatment plants and found that relative to the control plants, those who were randomized into the free management consulting services adopted better management practices and experienced increased productivity within a year through improved quality and efficiency and reduced inventory (Bloom, Eifert, Mahajan, McKenzie, & Roberts, 2013).

Payment reform

A major new theme in health policy is innovation in the payment structure for healthcare providers, including bundling payments for episodes of care and creating shared saving contracts. Because they are market-wide interventions, these innovations are often held up as an example of something that is hard to study through a randomized evaluation. Yet as these payment mechanisms evolve, the set of patients or conditions included is expanded and re-defined. As they expand to take on new blocks of risk (i.e. new groups of patients), one could randomize at the patient level which patients are included. Of course, as discussed in Section VI—and in Finkelstein (Finkelstein, 2007)—this would miss any impact of the payment reform that operates by providers modifying care for all patients (including those not covered by the payment reform). However it would still provide very useful information on the impact of alternate payment structures on health outcomes and healthcare use and costs that result from modifying care at the patient level.

More generally, even system-wide innovations in U.S. healthcare delivery often start as pilot or demonstration projects. This can provide opportunities to randomize at the patient level, as described above, or at the organization level. For example, CMS may run a demonstration project where they partner with a limited number of states, hospital systems, or other care organizations. When there is sufficient interest, it might be possible to use random selection to determine which of the well-qualified partners participate and to create a clear control group.

References

Abdul Latif Jameel Poverty Action Lab; Raskin: Improving Targeting Aid Distribution of Subsidized Rice; http://www.povertyactionlab.org/scale-ups/raskin-improving-targeting-and-distribution-subsidized-rice; August 22, 2014.

Abt Associates. 2013. Projects & Publications.

Agency for Healthcare Research and Quality; A Critical Analysis of Quality Improvement Strategies; http://www.ahrq.gov/research/findings/factsheets/quality/qgapfact/index.html; June 27, 2014.

Alatas, V., Banerjee, A., Hanna, R., Olken, B., & Tobias, J. 2012. Targeting the Poor: Evidence from a Field Experiment in Indonesia. American Economic Review, 102(4): 1206-1240.

Allcott, H. 2012. Site Selection Bias in Program Evaluation, Working Paper: NBER.

Allen, H., Baicker, K., Taubman, S., Wright, B., & Finkelstein, A. 2013. The Oregon Health Insurance Experiment: When Limited Policy Resources Provide Research Opportunities. Journal of Health Politics, Policy and Law, 38(6): 1185-1194.

Angrist, J. 1990. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. American Economic Review, 80(3): 313-336.

Angrist, J., Imbens, G., & Rubin, D. 1996. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association, 91(434): 444-455.

Angrist, J. & Pischke, J.-S. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. Journal of Economic Perspectives, 24(2): 3-30.

Aron-Dine, A., Einav, L., & Finkelstein, A. 2013. The RAND Health Insurance Experiment, Three Decades Later. Journal of Economic Perspectives, 27(1): 197-222.

Ashenfelter, O. & Plant, M. 1990. Nonparametric Estimates of the Labor-Supply Effects of Negative Income Tax Programs. Journal of Labor Economics, 8(1): S396-S415.

Baicker, K., Cutler, D., & Song, Z. 2010. Workplace Wellness Programs Can Generate Savings. Health Affairs, 29(2): 304-311.

Baicker, K., Taubman, S. L., Allen, H. L., Bernsterin, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., & Finkelstein, A. N. 2013. The Oregon Experiment -- Effects of Medicaid on Clinical Outcomes. New England Journal of Medicine, 368(18): 1713-1722.

Baicker, K., Finkelstein, A., Song, J., & Taubman, S. 2014. The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment. American Economic Review, 104(5): 322-328.

Banerjee, A., Hanna, R., Kyle, J., Olken, B., & Sumarto, S. 2014. Information is Power: Identification Cards and Food Subsidy Programs in Indonesia Working Paper: MIT.

Bardach, N., Wang, J., De Leon, S., Shih, S., Boscardin, J., Goldman, E., & Dudley, A. 2013. Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. Journal of the American Medical Association, 310(10): 1051-1059.

Berwick, D. 2003. Disseminating Innovations in Health Care. Journal of the American Medical Association, 289(15): 1969-1975.

Berwick, D., Nolan, T., & Whittington, J. 2008. The Triple Aim: Care, Health, and Cost. Health Affairs, 27(3): 759-769.

Bloom, E., King, E., Kremer, M., Bhushan, I., Clingingsmith, D., Loevinsohn, B., Hong, R., & Schwartz, B. 2006. Contracting for Health: Evidence from Cambodia: Brookings Institution.

Bloom, N., Genakos, C., Sadun, R., & Reenen, J. V. 2012. Management Practices Across Firms and Countries. Academy of Management Perspectives, 26(1): 12-33.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., & Roberts, J. 2013. Does Management Matter? Evidence from India. The Quarterly Journal of Economics, 128(1): 1-51.

Bloom, N., Raffaella, S., & Van Reenen, J. 2014. Does Management Matter in Healthcare?, Mimeo: Stanford.

Bosworth, H., Olsen, J., Neary, A., Orr, M., Powers, B., Adams, M., Svetkey, L., Reed, S., Li, Y., Dolor, R., & Oddone, E. 2009. Two Self-management Interventions to Improve Hypertension Control: A Randomized Trial. Annals of Internal Medicine, 151(10): 687-695.

Card, D., Dobkin, C., & Maestas, N. 2008. The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare. American Economic Review, 98(5): 2242-2258.

Card, D., Dobkin, C., & Maestas, N. 2009. Does Medicare Save Lives? Quarterly Journal of Economics, 124(2): 597-636.

Casey, K., Glennerster, R., & Miguel, E. 2012. Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan. Quarterly Journal of Economics, 127(4): 1755-1812.

Centers for Medicare and Medicaid Services; National Health Expenditures: 2012 Highlights; http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/highlights.pdf; June 30, 2014.

Centers for Medicare and Medicaid Services. 2013. Medicare Hospice Benefits.

Centers for Medicare and Medicaid Services; Health Care Innovation Awards Round Two; July 1, 2014.

Centers for Medicare and Medicaid Services. 2014b. Readmissions Reduction Program.

Centers for Medicare and Medicaid Services; Medicare Care Choices Model; http://innovation.cms.gov/initiatives/Medicare-Care-Choices/; June 24, 2014.

Chandra, A., Finkelstein, A., Sacarny, A., & Syverson, C. 2013. Healthcare Exceptionalism? Productivity and Allocation in the U.S. Healthcare Sector, Working Paper: National Bureau of Economic Research.

Chaudry, S. I., Mattera, J. A., Curtis, J. P., Spertus, J. A., Herrin, J., Lin, Z., Phillips, C. O., Hodshon, B. V., Cooper, L. S., & Krumholtz, H. M. 2010. Telemonitoring in Patients with Heart Failure. New England Journal of Medicine, 363(24).

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. Quarterly Journal of Economics, 126(4): 1593-1660.

Choosing Wisely; Choosing Wisely; http://www.choosingwisely.org/; August 24, 2014.

Choudhry, N., Avorn, J., Glynn, R., Antman, E., Schneeweiss, S., Toscano, M., Reisman, L., Fernandes, J., Spettell, C., Lee, J., Levin, R., Brennan, T., & Shrank, W. 2011. Full Coverage for Preventive Medications after Myocardial Infarction. New England Journal of Medicine, 365(22): 2088-2097.

Climo, M., Yoke, D., Warren, D., Perl, T., Bolon, M., Herwaldt, L., Weinstein, R., Sepkowitz, K., & Jernigan, J. 2013. Effect of Daily Chlorhexidine Bathing on Hospital-Acquired Infection. New England Journal of Medicine, 368(6): 533-542.

Clinicaltrials.gov; History, Policies, and Laws; http://clinicaltrials.gov/ct2/about-site/history#CongressPassesLawFDAMA; July 2, 2014.

Coburn, K., Marcantonio, S., Lazansky, R., Keller, M., & Davis, N. 2012. Effect of a Community-Based Nursing Intervention on Mortality in Chronically Ill Older Adults: A Randomized Controlled Trial. PLOS Medicine, 9(7): e1001265.

Council of Economic Advisers. 2014: Executive Office of the President of the United States.

Crepon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. 2013. Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment. Quarterly Journal of Economics, 128(2): 531–580.

De Angelis, C., Drazen, J., Frizelle, F., Haug, C., Hoey, J., Horon, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, J., Schroeder, T., Sox, H., & Van Der Weyden, M. 2004. Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. New England Journal of Medicine, 351(12): 1250-1251.

Douthat, R. 2013. What Health Insurance Doesn't Do, New York Times.

Eckenrode, J., Ganzel, B., Henderson, C. R., Smith, E., Olds, D. L., Powers, J., Cole, R., Kitzman, H., & Sidora, K. 2000. Preventing Child Abuse and Neglect with a Program of Nurse Home Visitation. Journal of the American Medical Association, 284(11): 1385-1391.

Eckenrode, J., Campa, M., Luckey, D. W., Henderson Jr., C. R., Cole, R., Kitzman, H., Anson, E., Sidora-Arcoleo, K., Powers, J., & Olds, D. 2010. Long-term Effects of Prenatal and Infancy Nurse Home Visitation on the Life Course of Youths: 19-Year Follow-up of a Randomized Trial. Archives of Pediatrics & Adolescent Medicine, 164(1): 9-15.

Engemann, K. & Wall, H. 2009. A Journal Ranking for the Ambitious Economist. Federal Reserve Bank of St. Louis Review, 91(3): 127-140.

Finkelstein, A. 2007. The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare. Quarterly Journal of Economics, 122(1): 1-37.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., & The Oregon Health Study Group. 2011. The Oregon Health Insurance Experiment: Evidence from the First Year. NBER Working Paper, 17190.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., & Group, O. H. S. 2012. The Oregon Health Insurance Experiment: Evidence from the First Year. The Quarterly Journal of Economics, 127(3): 1057-1106.

Finkelstein, A., Doyle, J., Taubman, S., Zhou, R., & Brenner, J. 2014. Health Care Hotspotting: A Randomized Controlled Trial (Analysis Plan).

Foundation, K.; Frequently Asked Questions; http://www.kipp.org/about-kipp/faq; July 3, 2014.

Gawande, A. 2012. Big Med, The New Yorker.

Gawande, A. 2013. Slow Ideas, The New Yorker.

Ginsburg, P. B. 2011. Spending to Save — ACOs and the Medicare Shared Savings Program. New England Journal of Medicine, 364(22): 2085-2086.

Glennerster, R. & Takavarasha, K. 2013. Running Randomized Evaluations: A Practical Guide: Princeton University Press.

Greenberg, D. & Halsey, H. 1983. Systematic Misreporting and Effects of Income Maintenance Experiments on Work Effort: Evidence from the Seattle-Denver Experiment. Journal of Labor Economics, 1(4): 380-407.

Harris, A., Pineles, L., Belton, B., Johnson, K., Shardell, M., Loeb, M., Newhouse, R., Dembry, L., Braun, B., Perencevich, E., Hall, K., & Morgan, D. 2013. Universal Glove and Gown Use and Acquisition of Antibiotic-Resistant Bacteria in the ICU: A Randomized Trial. Journal of the American Medical Association, 310(15): 1571-1580.

Healthcare Cost and Utilization Project; Data Organizations Participating in the SID; http://www.hcup-us.ahrq.gov/partners.jsp?SID; July 9, 2014.

Hearst, N., Newman, T. B., & Hulley, S. B. 1986. Delayed Effects of the Military Draft on Mortality: A Randomized Natural Experiment. New England Journal of Medicine, 314.

Higgins, J. & (editors), S. G. 2011. Cochrane Handbook for Systematic Reviews of Interventions The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Holmstrom, B. 1979. Moral Hazard and Observability. Bell Journal of Economics, 10(1): 74-91.

Holmstrom, B. 1982. Moral Hazard in Teams. Bell Journal of Economics, 13(2): 324-340.

Holmstrom, B. & Milgrom, P. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. Journal of Law, Economics, and Organizations, 7.

Huang, S., Septimus, E., Kleinman, K., Moody, J., Hickok, J., Avery, T., Lankiewicz, J., Gombosev, A., Terpstra, L., Hartford, F., Hayden, M., Jernigan, J., Weinstein, R., Fraser, V., Haffenreffer, K.,

Cui, E., Kaganov, R., & Lolans, K. 2013. Targeted versus Universal Decolonization to Prevent ICU Infection. New England Journal of Medicine, 368(24): 2255-2265.

Jencks, S., Williams, M., & Coleman, E. 2009. Rehospitalizations among Patients in the Medicare Fee-for-Service Program. New England Journal of Medicine, 360(14): 1418-1428.

Junod, S. W.; FDA and Clinical Drug Trials: A Short History; June 26.

Kerlin, M. P., Small, D., Cooney, E., Fuchs, B., Bellini, L., Mikkelsen, M., Schweickert, W., Bakhru, R., Gabler, N., Harhay, M., Hansen-Flaschen, J., & Halpern, S. 2013. A Randomized Trial of Nighttime Physician Staffing in an Intensive Care Unit. New England Journal of Medicine, 368(23): 2201-2209.

Kimmel, S. E., Troxel, A. B., Loewenstein, G., Brensinger, C. M., Jaskowiak, J., Doshi, J. A., Laskin, M.,

& Volpp, K. 2012. Randomized trial of lottery-based incentives to improve warfarin adherence. Am Heart J, 164(2): 268-274.

Kitzman, H., Olds, D., Sidora, K., Henderson, C., Hanks, C., Cole, R., Luckey, D., Bondy, J., Cole, K., & Glazner, J. 2000. Enduring Effects of Nurse Home Visitation on Maternal Life Course: A 3-Year Follow-up of a Randomized Trial. Journal of the American Medical Association, 283(15): 1983-1989.

Kullgren, J., Troxel, A., Loewenstein, G., Asch, D., Norton, L., Wesby, L., Tao, Y., Zhu, J., & Volpp, K. 2013. Individual- Versus Group-Based Financial Incentives for Weight Loss. Annals of Internal Medicine, 158(7): 505-514.

Lab, A. L. J. P. A.; School-based Deworming; http://www.povertyactionlab.org/scale-ups/school-based-deworming; August 22, 2014.

Leonhardt, D. 2009. Making Health Care Better, The New York Times Magazine.

Long, J., Jahnle, E., Richardson, D., Loewenstein, G., & Volpp, K. 2013. Peer Mentoring and Financial Incentives to Improve Glucose Control in African American Veterans: A Randomized, Controlled Trial. Annals of Internal Medicine, 156(6): 416-424.

Ludwig, J., Kling, J., & Mullainathan, S. 2011a. Mechanism Experiments and Policy Evaluations. Journal of Economic Perspectives, 25(3): 17-38.

Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G., Katz, L., Kessler, R., Kling, J., Lindau, S. T., Whitaker, R., & McDade, T. 2011b. Neighborhoods, Obesity, and Diabetes — A Randomized Social Experiment. New England Journal of Medicine, 365(16): 1509-1519.

Ludwig, J., Duncan, G., Gennetian, L., Katz, L., Kessler, R., Kling, J., & Sanbonmatsu, L. 2013. Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity. American Economic Review, 103(3): 226-231.

Madrian, B. & Shea, D. 2001. The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior. Quarterly Journal of Economics, 116(4): 1149-1187.

Margolis, K. L., Asche, S. E., Dehmer, S. P., Groen, S. E., Kadrmas, H. M., Kerby, T. J., Klotzle, K. J., Maciosek, M. V., Michels, R. D., O'Connor, P. K., Pritchard, R. A., Sekenski, J. L., Spel-Hillen, J. M., & Trower, N. K. 2013. Effect of Home Blood Pressure Telemonitoring and Pharmacist Management on Blood Pressure Control: A Cluster Randomized Trial. Journal of the American Medical Association, 310(1): 46-56.

Matchar, D. B., Jacobson, A., Dolor, R., Edson, R., Uyeda, L., Phibbs, C., Vertrees, J. E., Shih, M.-C., Holodniy, M., & Lavori, P. 2010. Effect of Home Testing of International Normalized Ratio on Clinical Events. New England Journal of Medicine, 363(17): 1608-1620.

Mathematica Policy Research. 2013. Publications.

McCall, N. & Cromwell, J. 2011. Results of the Medicare Health Support Disease-Management Pilot Program. New England Journal of Medicine, 365(18): 1704-1712.

MDRC. 2013. Publications.

Michalopoulos, C., Wittenburg, D., Israel, D. A. R., & Warren, A. 2012. The Effects of Health Care Benefits on Health Care Use and Health: A Randomized Trial for Disability Insurance Benefits. Medical Care, 50(9): 764-771.

Miguel, E. & Kremer, M. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Econometrica, 72(1): 159-217.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. 2014. Promoting Transparency in Social Science Research. Science, 343(6166): 30-31.

Miller, C., Bos, J., Porter, K., Tseng, F. M., & Abe, Y. 2005. The Challenge of Repeating Success in a Changing World: Final Report on the Center for Employment Training Replication Sites. National Institutes of Health; Funding Opportunity Announcement; http://grants.nih.gov/grants/guide/rfa-files/RFA-HL-14-019.html; July 1, 2014.

National Institutes of Health; National Institutes of Health Launches ClinicalTrials.gov Results Database,;

http://www.nlm.nih.gov/news/expanded_clinicaltrials.html; July 2, 2014.

Newhouse, J. P. & the Insurance Experiment Group. 1993. Free for All: Lessons from the RAND Health Insurance Experiment. Cambridge: Harvard University Press.

Olds, D., Henderson, C., Tatelbaum, R., & Chamberlin, R. 1988. Improving the Life-Course Development of Socially Disadvantaged Mothers: A Randomized Trial of Nurse Home Visitation. American Journal of Public Health, 78(11): 1436-1445.

Olds, D., Robinson, J., Pettitt, L., Luckey, D., Holmberg, J., Ng, R., Isacks, K., Sheff, K., & Henderson, C. 2004. Effects of Home Visits by Paraprofessionals and by Nurses: Age 4 Follow-Up Results of a Randomized Trial. Pediatrics, 114(6): 1560-1568.

Olds, D., Kitzman, H., Hanks, C., Cole, R., Anson, E., Sidora-Arcoleo, K., Luckey, D., Henderson, C., Holmberg, J., Tutt, R., Stevenson, A., & Bondy, J. 2007. Effects of Nurse Home Visiting on Maternal and Child Functioning: Age-9 Follow-up of a Randomized Trial. Pediatrics, 120(4): 832-845.

Olds, D. L., Henderson, C. R., & Kitzman, H. 1994. Does Prenatal and Infancy Nurse Home Visitation Have Enduring Effects on Qualities of Parental Caregiving and Child Health at 25 to 50 Months of Life? Pediatrics, 93(1): 89-98.

Olken, B. Forthcoming. Pre-Analysis Plans in Economics. Journal of Economics Perspectives.

Olken, B., Onishi, J., & Wong, S. Forthcoming. Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia. American Economic Journal: Applied Economics, Forthcoming.

Patient-Centered Outcomes Research Institute; How We're Funded; http://www.pcori.org/about-us/how-were-funded/; July 1, 2014.

Peikes, D., Chen, A., Schore, J., & Brown, R. 2009. Effects of Care Coordination on Hospitalization, Quality of Care, and Health Care Expenditures Among Medicare Beneficiaries. Journal of the American Medical Association, 301(6): 603-618.

Petersen, L., Simpson, K., Pietz, K., Urech, T., Hysong, S., Profit, J., Conrad, D., Dudley, R. A., & Woodard, L. 2013. Effects of Individual Physician-Level and Practice-Level Financial Incentives on Hypertension Care: A Randomized Trial. Journal of the American Medical Association, 310(10): 1042-1050.

RAND; Program Evaluation; http://www.rand.org/topics/program-evaluation.html?content-type=research+brief.

Song, Z., Safran, D. G., Landon, B., He, Y., Ellis, R., Mechanic, R., Day, M., & Chernew, M. 2011. Health care spending and quality in year 1 of the alternative quality contract. New England Journal of Medicine, 365(10): 909-918.

Takeda, A., Taylor, S., Taylor, R., Khan, F., Krum, H., & Underwood, M. 2012. Clinical service organisation for heart failure. Cochrane Database of Systematic Reviews.

Taubman, S. L., Allen, H. L., Wright, B. J., Baicker, K., & Finkelstein, A. N. 2014. Medicaid increases emergency-department use: evidence from Oregon's Health Insurance Experiment. Science, 343(6168): 263-268.

Taylor, S., Bestall, J., Cotter, S., Falshaw, M., Hood, S., Parsons, S., Wood, L., & Underwood, M. 2005. Clinical service organisation for heart failure. Cochrane Database of Systematic Reviews.

Temel, J., Greer, J., Muzikansky, A., Gallagher, E., Admane, S., Jackson, V., Dahlin, C., Blinderman, C., Jacobsen, J., Pirl, W., Billings, A., & Lynch, T. 2010. Early palliative care for patients with metastatic non-small-cell lung cancer. New England Journal of Medicine, 363(8): 733-742.

Thaler, R. H. & Sunstein, C. R. 2009. Nudge: Improving Decisions about Health, Wealth, and Happiness: Punguin Books.

Thompson Reuters. 2012a. 2012 Journal Citation Reports Social Sciences Edition: Thompson Reuters.

Thompson Reuters. 2012b. 2012 Journal Citation Reports Science Edition: Thompson Reuters.

Volpp, K., Gurmankin Levy, A., Asch, D., Berlin, J., Murphy, J., Gomez, A., Sox, H., Zhu, J., & Lerman, C. 2006. A randomized controlled trial of financial incentives for smoking cessation. Cancer Epidemiology, Biomarkers & Prevention, 15(1): 12-18.

Volpp, K., Troxel, A., Pauly, M., Glick, H., Puig, A., Asch, D., Galvin, R., Zhu, J., Wan, F., DeGuzman, J., Corbett, E., Weiner, J., & Audrain-McGovern, J. 2009. A Randomized, Controlled Trial of Financial Incentives for Smoking Cessation. New England Journal of Medicine, 360(7): 699-709.

Volpp, K., Shea, J., Small, D., Basner, M., Zhu, J., Norton, L., Ecker, A., Novak, C., Bellini, L., Dine, J., Mollicone, D., & Dinges, D. 2012. Effect of a Protected Sleep Period on Hours Slept During Extended Overnight In-hospital Duty Hours Among Medical Interns. Journal of the American Medical Association, 308(21): 2208-2217.

Volpp, K. G., John, L. K., Troxel, A. B., Norton, L., Fassbender, J., & Loewenstein, G. 2008. Financial incentive-based approaches for weight loss: a randomized trial. JAMA, 300(22): 2631-2637.

Weathers, R. & Stegman, M. 2012. The effect of expanding access to health insurance on the heath and mortality of Social Security Disability Insurance beneficiaries. Journal of Health Economics, 31(6): 863-875.

Appendix A: Literature Searches

Frequency of randomized trials vs. non-randomized studies of healthcare delivery

To understand how commonly RCTs are used to test healthcare delivery innovations, we conducted a review of papers on healthcare delivery published from 2009-2013 in the top journals in three fields where we expected studies of healthcare delivery interventions to be most common: medicine, health services research, and economics. In economics, we reviewed all papers published in *American Economic Review, Quarterly Journal of Economics, Journal of Political Economy, and Econometrica.*[4] In health services journals, we reviewed all papers published in *Health Affairs, Medical Care*, and *Milbank Quarterly.*[5] In medicine, we reviewed papers published in the *New England Journal of Medicine, the Journal of the American Medical Association, Annals of Internal Medicine,* and *PLoS Medicine.*[6] Because of the high volume of publications in these journals in medicine, we randomly selected 4 months per year, for a total of 20 months.

In all cases, we limited our review to inference studies that investigate a potentially causal link or association between an intervention (treatment) and an outcome. In addition, we further excluded those inference studies not taking place at least partly in the US or for which the location was unknown (with the exception of the economics literature review, where we included an explicit comparison to development studies).

Research assistants reviewed the online table of contents and abstracts to identify inference studies. Where necessary, they also reviewed the full article. For all inference studies, the research assistants recorded the title, authors, journal, and year in a spreadsheet, whether the design was a randomized controlled trial, and the topic field (such as healthcare delivery, medical technology, education, etc.). We provide more detail on each of the literature searches below.

Medicine

In medicine, we reviewed all articles in the *New England Journal of Medicine, the Journal of the American Medical Association, Annals of Internal Medicine,* and *PLoS Medicine* from four randomly selected months in the years 2009-2013. In total, there were 925 articles published in the four journals in the twenty months considered. Two research assistants split the work of reviewing the 925 abstracts. A randomly selected subsample of 60 abstracts was double-coded. For the 60 abstracts reviewed by both, they disagreed on the classification as inference on 3 articles and on the classification as healthcare delivery vs. medical on 2 articles.

We limit our discussion to inference studies that investigate a potentially causal link or association between a treatment and an outcome and that considered either a medical or healthcare delivery innovation (N=450). We further excluded those inference studies not taking place at least partly in the United States (N=203) or where the location of the study was not available (N=9). This left 238 US-based inference studies in either healthcare delivery or medical innovations.

[4] These economics journals have been ranked as the top journals for "ambitious economists" (Engemann, K. & Wall, H., 2009). We excluded the non-peer reviewed "Papers and Proceedings" issue of the *American Economic Review.*

[5] These health services journals are the top three ranked by impact factor (Thompson Reuters. 2012a.).

[6] These four journals are among the top six journals for general and internal medicine ranked by impact factor (Thompson Reuters. 2012.b.). We excluded the other two journals in the top six (*The Lancet,* and *British Medical Journal*) after a preliminary investigation of 4 months of articles found no papers on US healthcare delivery published in these British journals.

Health Services Research

For the years 2009-2013, we reviewed all the abstracts of the papers in *Health Affairs*, *Medical Care*, and the *Milbank Quarterly.* A single research assistant reviewed the titles and abstracts of the over 2200 articles published in those journals in those years. This review identified 548 empirical inference studies that investigate a potentially causal link or association between a treatment and an outcome (N=548). These were coded by location of the study, as being on healthcare delivery or not, and as being randomized control trials or not. Restricting to US-based healthcare delivery studies left 405 articles.

Economics

For the years 2009-2013, we reviewed all the abstracts of the papers in *American Economic Review*, *Quarterly Journal of Economics*, *Journal of Political Economy*, and *Econometrica*. A single research assistant reviewed the titles of abstracts of the roughly 1200 articles published in those journals in those years. This review identified 314 applied microeconomic papers making causal inferences. This excludes laboratory-based studies. These inference studies were coded by location of the study, general topic, and as being randomized controlled trials or not. There were 197 studies based at least partly in the US on healthcare delivery and other topics. We also use data on the 31 international development papers as a further comparison group.

Publications from evaluation firms

We reviewed the publication databases at MDRC, Mathematica Policy Research (MPR), Abt Associates, and the program evaluations database at RAND (Abt Associates, 2013; Mathematica Policy Research, 2013; MDRC, 2013; RAND, 2013). We restricted our search to publications from 2013 of work that took place in the US. We limit our discussion to inference studies that investigate a potentially causal link or association between a treatment and an outcome. We further exclude publications that only discussed aspects of the implementation of a study but did not discuss the results of the intervention itself. When a single study resulted in multiple publications, we count that study once. We do, however, allow studies to contribute to multiple topic areas when appropriate.

Potential publication bias

Many medical journals require that randomized studies that prospectively and randomly assign participants to an intervention have their basic information recorded in a registry before they begin. The Food and Drug Modernization Act of 1997 required the registration of all trials of new and experimental drugs (Clinicaltrials.gov, 2014). By 2005, the International Committee of Medical Journal Editors (ICMJE) had announced that it would require all trials to be registered as a condition of publication (De Angelis et al., 2004). Run by the US National Institutes of Health, www.clinicaltrials.gov is the largest single registry of clinical trials (National Institutes of Health, 2013b).

To estimate how many registered randomized controlled trials are for healthcare delivery innovations, we consider all studies whose information was first received by clinicaltrials.gov from January 1, 2006 and December 31, 2010.[7] We downloaded links to study descriptions for all 17,947 randomized "interventional studies" which were US-based. Studies were divided in the registry into nine categories according to the type of intervention tested: Drug, Behavioral, Procedure, Device, Biological, Dietary Supplement, Radiation, Genetic and Other. We randomly selected 50 studies from each category (except for "Genetic" and

[7] We use earlier years for the search in clinicaltrials.gov than for the search in the top medical journals because study registration occurs much earlier for a given study than the publication of findings. We use a 5-year period beginning in 2006 because that was the first full year for which study registration is required for subsequent publication in most journals.

"Radiation" studies, which had only 25 and 36 studies, respectively) to be catalogued: a total of 411 studies.[8] We recorded whether the studies were for medical interventions (such as drugs and devices) or for healthcare delivery innovations. To calculate statistics based on this stratified random sampling, we use weights equal to the average probability of being sampled divided by the probability of being sampled in the individual intervention category.

Characteristics of randomized trials of healthcare delivery

To identify randomized trials of US healthcare delivery innovations, we reviewed articles published in 2009-2013 in four top medical journals the *New England Journal of Medicine, the Journal of the American Medical Association, Annals of Internal Medicine,* and *PLoS Medicine*. We used the Cochrane Highly Sensitive Search Strategy (Higgins & (editors), 2011) to search PubMed for randomized studies published in these journals during those years. This search returned 1784 results, and we identified 99 as RCTs testing healthcare delivery innovations in the United States. The citations for the 99 included studies are included below.

Two research assistants split the work of reviewing the titles and abstracts returned by the search. For articles identified as healthcare delivery RCTs, they reviewed the full article as well. The research assistants recorded the author, journal, and year of the article and coded a variety of measures on the size and length of the trial, the randomization, the intervention, the data sources, and the outcomes measured. A total of 10 studies were double-coded by both research assistants. This identified several discrepancies in the coding procedures, and variables were recoded using new agreed-upon procedures.

[8] A minority of randomized interventions (7 percent) had two different study types listed. In these cases, studies were classified based on the first listed intervention type.

Appendix B: List of Healthcare Delivery RCTs

From Top Medical Journals (2009-2013)

Adair, R., Wholey, D. R., Christianson, J., White, K. M., Britt, H., & Lee, S. (2013). Improving chronic disease care by adding laypersons to the primary care team: a parallel randomized trial. Ann Intern Med, 159(3), 176-184. doi: 10.7326/0003-4819-159-3-201308060-00007

Allen, K. D., Oddone, E. Z., Coffman, C. J., Datta, S. K., Juntilla, K. A., Lindquist, J. H., Bosworth, H. B. (2010). Telephone-based self-management of osteoarthritis: A randomized trial. Ann Intern Med, 153(9), 570-579. doi: 10.7326/0003-4819-153-9-201011020-00006

Arriaga, A. F., Bader, A. M., Wong, J. M., Lipsitz, S. R., Berry, W. R., Ziewacz, J. E., Gawande, A. A. (2013). Simulation-based trial of surgical-crisis checklists. N Engl J Med, 368(3), 246-253. doi: 10.1056/NEJMsa1204720

Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Smith, J. (2013). The Oregon experiment--effects of Medicaid on clinical outcomes. N Engl J Med, 368(18), 1713-1722. doi: 10.1056/NEJMsa1212321

Bakitas, M., Lyons, K. D., Hegel, M. T., Balan, S., Brokaw, F. C., Seville, J., Ahles, T. A. (2009). Effects of a palliative care intervention on clinical outcomes in patients with advanced cancer: the Project ENABLE II randomized controlled trial. Jama, 302(7), 741-749. doi: 10.1001/jama.2009.1198

Bardach, N. S., Wang, J. J., De Leon, S. F., Shih, S. C., Boscardin, W. J., Goldman, L. E., & Dudley, R. A. (2013). Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. Jama, 310(10), 1051-1059. doi: 10.1001/jama.2013.277353

Bednarek, P. H., Creinin, M. D., Reeves, M. F., Cwiak, C., Espey, E., & Jensen, J. T. (2011). Immediate versus delayed IUD insertion after uterine aspiration. N Engl J Med, 364(23), 2208-2217. doi: 10.1056/NEJMoa1011600

Blumenthal, J. A., Babyak, M. A., O'Connor, C., Keteyian, S., Landzberg, J., Howlett, J., Whellan, D. J. (2012). Effects of exercise training on depressive symptoms in patients with chronic heart failure: the HF-ACTION randomized trial. Jama, 308(5), 465-474. doi: 10.1001/jama.2012.8720

Bosworth, H. B., Olsen, M. K., Grubber, J. M., Neary, A. M., Orr, M. M., Powers, B. J., Oddone, E. Z. (2009). Two self-management interventions to improve hypertension control: a randomized trial. Ann Intern Med, 151(10), 687-695. doi: 10.7326/0003-4819-151-10-200911170-00148

Carling, C. L., Kristoffersen, D. T., Montori, V. M., Herrin, J., Schunemann, H. J., Treweek, S., Oxman, A. D. (2009). The effect of alternative summary statistics for communicating risk reduction on decisions about taking statins: a randomized trial. PLoS Med, 6(8), e1000134. doi: 10.1371/journal.pmed.1000134

Chaudhry, S. I., Mattera, J. A., Curtis, J. P., Spertus, J. A., Herrin, J., Lin, Z., Krumholz, H. M. (2010). Telemonitoring in patients with heart failure. N Engl J Med, 363(24), 2301-2309. doi: 10.1056/NEJMoa1010029

Cherkin, D. C., Sherman, K. J., Kahn, J., Wellman, R., Cook, A. J., Johnson, E., Deyo, R. A. (2011). A comparison of the effects of 2 types of massage and usual care on chronic low back pain: a randomized, controlled trial. Ann Intern Med, 155(1), 1-9. doi: 10.7326/0003-4819-155-1-201107050-00002

Chlan, L. L., Weinert, C. R., Heiderscheit, A., Tracy, M. F., Skaar, D. J., Guttormson, J. L., & Savik, K. (2013). Effects of patient-directed music intervention on anxiety and sedative exposure in critically ill patients receiving mechanical ventilatory support: a randomized clinical trial. Jama, 309(22), 2335-2344. doi: 10.1001/jama.2013.5670

Choudhry, N. K., Avorn, J., Glynn, R. J., Antman, E. M., Schneeweiss, S., Toscano, M., Shrank,

W. H. (2011). Full coverage for preventive medications after myocardial infarction. N Engl J Med, 365(22), 2088-2097. doi: 10.1056/NEJMsa1107913

Church, T. S., Blair, S. N., Cocreham, S., Johannsen, N., Johnson, W., Kramer, K., Earnest, C. P. (2010). Effects of aerobic and resistance training on hemoglobin A1c levels in patients with type 2 diabetes: a randomized controlled trial. Jama, 304(20), 2253-2262. doi: 10.1001/jama.2010.1710

Climo, M. W., Yokoe, D. S., Warren, D. K., Perl, T. M., Bolon, M., Herwaldt, L. A., Wong, E. S. (2013). Effect of daily chlorhexidine bathing on hospital-acquired infection. N Engl J Med, 368(6), 533-542. doi: 10.1056/NEJMoa1113849

Coburn, K. D., Marcantonio, S., Lazansky, R., Keller, M., & Davis, N. (2012). Effect of a community-based nursing intervention on mortality in chronically ill older adults: a randomized controlled trial. PLoS Med, 9(7), e1001265. doi: 10.1371/journal.pmed.1001265

Curtis, J. R., Back, A. L., Ford, D. W., Downey, L., Shannon, S. E., Doorenbos, A. Z., Engelberg, R. A. (2013). Effect of communication skills training for residents and nurse practitioners on quality of communication with patients with serious illness: a randomized trial. Jama, 310(21), 2271-2281. doi: 10.1001/jama.2013.282081

Daumit, G. L., Dalcin, A. T., Jerome, G. J., Young, D. R., Charleston, J., Crum, R. M., Appel, L. J. (2011). A behavioral weight-loss intervention for persons with serious mental illness in psychiatric rehabilitation centers. Int J Obes (Lond), 35(8), 1114-1123. doi: 10.1038/ijo.2010.224

Davis, C. L., Pollock, N. K., Waller, J. L., Allison, J. D., Dennis, B. A., Bassali, R., Gower, B. A. (2012). Exercise dose and diabetes risk in overweight and obese children: a randomized controlled trial. Jama, 308(11), 1103-1112. doi: 10.1001/2012.jama.10762

Digenio, A. G., Mancuso, J. P., Gerber, R. A., & Dvorak, R. V. (2009). Comparison of methods for delivering a lifestyle modification program for obese patients: a randomized trial. Ann Intern Med, 150(4), 255-262.

Dobscha, S. K., Corson, K., Perrin, N. A., Hanson, G. C., Leibowitz, R. Q., Doak, M. N., Gerrity, M. S. (2009). Collaborative care for chronic pain in primary care: a cluster randomized trial. Jama, 301(12), 1242-1252. doi: 10.1001/jama.2009.377

Duncan, P. W., Sullivan, K. J., Behrman, A. L., Azen, S. P., Wu, S. S., Nadeau, S. E., Hayden, S. K. (2011). Body-weight-supported treadmill rehabilitation after stroke. N Engl J Med, 364(21), 2026-2036. doi: 10.1056/NEJMoa1010790

Dykes, P. C., Carroll, D. L., Hurley, A., Lipsitz, S., Benoit, A., Chang, F., Middleton, B. (2010). Fall prevention in acute care hospitals: a randomized trial. Jama, 304(17), 1912-1918. doi: 10.1001/jama.2010.1567

Ebbeling, C. B., Feldman, H. A., Chomitz, V. R., Antonelli, T. A., Gortmaker, S. L., Osganian, S. K., & Ludwig, D. S. (2012b). A randomized trial of sugar-sweetened beverages and adolescent body weight. N Engl J Med, 367(15), 1407-1416. doi: 10.1056/NEJMoa1203388

Fan, V. S., Gaziano, J. M., Lew, R., Bourbeau, J., Adams, S. G., Leatherman, S., Niewoehner, D. E. (2012). A comprehensive care management program to prevent chronic obstructive pulmonary disease hospitalizations: a randomized, controlled trial. Ann Intern Med, 156(10), 673-683. doi: 10.7326/0003-4819-156-10-201205150-00003

Flynn, K. E., Pina, I. L., Whellan, D. J., Lin, L., Blumenthal, J. A., Ellis, S. J., Weinfurt, K. P. (2009). Effects of exercise training on health status in patients with chronic heart failure: HF-ACTION randomized controlled trial. Jama, 301(14), 1451-1459. doi: 10.1001/jama.2009.457

Foster, G. D., Wyatt, H. R., Hill, J. O., Makris, A. P., Rosenbaum, D. L., Brill, C., Klein, S. (2010). Weight and metabolic outcomes after 2 years on a low-carbohydrate versus low-fat diet: a randomized trial. Ann Intern Med, 153(3), 147-157. doi: 10.7326/0003-4819-153-3-201008030-00005

Franklin, M. E., Sapyta, J., Freeman, J. B., Khanna, M., Compton, S., Almirall, D., March, J. S. (2011). Cognitive behavior therapy augmentation of pharmacotherapy in pediatric obsessive-compulsive disorder: the Pediatric OCD Treatment Study II (POTS II) randomized controlled trial. Jama, 306(11), 1224-1232. doi: 10.1001/jama.2011.1344

Garber, J., Clarke, G. N., Weersing, V. R., Beardslee, W. R., Brent, D. A., Gladstone, T. R., Iyengar, S. (2009). Prevention of depression in at-risk adolescents: a randomized controlled trial. Jama, 301(21), 2215-2224. doi: 10.1001/jama.2009.788

Gerber, J. S., Prasad, P. A., Fiks, A. G., Localio, A. R., Grundmeier, R. W., Bell, L. M., Zaoutis, T. E. (2013). Effect of an outpatient antimicrobial stewardship intervention on broad-spectrum antibiotic prescribing by primary care pediatricians: a randomized trial. Jama, 309(22), 2345-2352. doi: 10.1001/jama.2013.6287

Gitlin, L. N., Harris, L. F., McCoy, M. C., Chernett, N. L., Pizzi, L. T., Jutkowitz, E., Hauck, W. W. (2013). A home-based intervention to reduce depressive symptoms and improve quality of life in older African Americans: a randomized trial. Ann Intern Med, 159(4), 243-252. doi: 10.7326/0003-4819-159-4-201308200-00005

Gitlin, L. N., Winter, L., Dennis, M. P., Hodgson, N., & Hauck, W. W. (2010). A biobehavioral home-based intervention and the well-being of patients with dementia and their caregivers: the COPE randomized trial. Jama, 304(9), 983-991. doi: 10.1001/jama.2010.1253

Giugliano, R. P., White, J. A., Bode, C., Armstrong, P. W., Montalescot, G., Lewis, B. S., Newby, L. K. (2009). Early versus delayed, provisional eptifibatide in acute coronary syndromes. N Engl J Med, 360(21), 2176-2190. doi: 10.1056/NEJMoa0901316

Goodpaster, B. H., Delany, J. P., Otto, A. D., Kuller, L., Vockley, J., South-Paul, J. E., Jakicic, J. M. (2010). Effects of diet and physical activity interventions on weight loss and cardiometabolic risk factors in severely obese adults: a randomized trial. Jama, 304(16), 1795-1802. doi: 10.1001/jama.2010.1505

Green, B. B., Wang, C. Y., Anderson, M. L., Chubak, J., Meenan, R. T., Vernon, S. W., & Fuller, S. (2013). An automated intervention with stepped increases in support to increase uptake of colorectal cancer screening: a randomized trial. Ann Intern Med, 158(5 Pt 1), 301-311. doi: 10.7326/0003-4819-158-5-201303050-00002

Gregg, E. W., Chen, H., Wagenknecht, L. E., Clark, J. M., Delahanty, L. M., Bantle, J., Bertoni, A. G. (2012). Association of an intensive lifestyle intervention with remission of type 2 diabetes. Jama, 308(23), 2489-2496. doi: 10.1001/jama.2012.67929

Harris, A. D., Pineles, L., Belton, B., Johnson, J. K., Shardell, M., Loeb, M., Safdar, N. (2013). Universal glove and gown use and acquisition of antibiotic-resistant bacteria in the ICU: a randomized trial. Jama, 310(15), 1571-1580. doi: 10.1001/jama.2013.277815

Heisler, M., Vijan, S., Makki, F., & Piette, J. D. (2010). Diabetes control with reciprocal peer support versus nurse care management: a randomized trial. Ann Intern Med, 153(8), 507-515. doi: 10.7326/0003-4819-153-8-201010190-00007

Houston, T. K., Allison, J. J., Sussman, M., Horn, W., Holt, C. L., Trobaugh, J., Hullett, S. (2011). Culturally appropriate storytelling to improve blood pressure: a randomized trial. Ann Intern Med, 154(2), 77-84. doi: 10.7326/0003-4819-154-2-201101180-00004

Huang, S. S., Septimus, E., Kleinman, K., Moody, J., Hickok, J., Avery, T. R., Platt, R. (2013). Targeted versus universal decolonization to prevent ICU infection. N Engl J Med, 368(24), 2255-2265. doi: 10.1056/NEJMoa1207290

Jacobs, A. K., Normand, S. L., Massaro, J. M., Cutlip, D. E., Carrozza, J. P., Jr., Marks, A. D., Mauri, L. (2013). Nonemergency PCI at hospitals with or without on-site cardiac surgery. N Engl J Med, 368(16), 1498-1508. doi: 10.1056/NEJMoa1300610

Jakicic, J. M., Tate, D. F., Lang, W., Davis, K. K., Polzien, K., Rickman, A. D., Finkelstein, E. A. (2012). Effect of a stepped-care intervention approach on weight loss in adults: a randomized clinical trial. Jama, 307(24), 2617-2626. doi: 10.1001/jama.2012.6866

Katon, W. J., Lin, E. H., Von Korff, M., Ciechanowski, P., Ludman, E. J., Young, B., McCulloch, D. (2010). Collaborative care for patients with depression and chronic illnesses. N Engl J Med, 363(27), 2611-2620. doi: 10.1056/NEJMoa1003955

Kerlin, M. P., Small, D. S., Cooney, E., Fuchs, B. D., Bellini, L. M., Mikkelsen, M. E., Halpern,

S. D. (2013). A randomized trial of nighttime physician staffing in an intensive care unit. N Engl J Med, 368(23), 2201-2209. doi: 10.1056/NEJMoa1302854

Klevens, J., Kee, R., Trick, W., Garcia, D., Angulo, F. R., Jones, R., & Sadowski, L. S. (2012). Effect of screening for partner violence on women's quality of life: a randomized controlled trial. Jama, 308(7), 681-689. doi: 10.1001/jama.2012.6434

Kravitz, R. L., Franks, P., Feldman, M. D., Tancredi, D. J., Slee, C. A., Epstein, R. M., Jerant, A. (2013). Patient engagement programs for recognition and initial treatment of depression in primary care: a randomized trial. Jama, 310(17), 1818-1828. doi: 10.1001/jama.2013.280038

Kripalani, S., Roumie, C. L., Dalal, A. K., Cawthon, C., Businger, A., Eden, S. K., Schnipper, J. L. (2012). Effect of a pharmacist intervention on clinically important medication errors after hospital discharge: a randomized trial. Ann Intern Med, 157(1), 1-10. doi: 10.7326/0003-4819-157-1-201207030-00003

Kroenke, K., Bair, M. J., Damush, T. M., Wu, J., Hoke, S., Sutherland, J., & Tu, W. (2009). Optimized antidepressant therapy and pain self-management in primary care patients with depression and musculoskeletal pain: a randomized controlled trial. Jama, 301(20), 2099-2110. doi: 10.1001/jama.2009.723

Li, F., Harmer, P., Fitzgerald, K., Eckstrom, E., Stock, R., Galver, J., Batya, S. S. (2012). Tai chi and postural stability in patients with Parkinson's disease. N Engl J Med, 366(6), 511-519. doi: 10.1056/NEJMoa1107911

Long, J. A., Jahnle, E. C., Richardson, D. M., Loewenstein, G., & Volpp, K. G. (2012). Peer mentoring and financial incentives to improve glucose control in African American veterans: a randomized trial. Ann Intern Med, 156(6), 416-424. doi: 10.7326/0003-4819-156-6-201203200-00004

Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G. J., Katz, L. F., McDade, T. W. (2011). Neighborhoods, obesity, and diabetes--a randomized social experiment. N Engl J Med, 365(16), 1509-1519. doi: 10.1056/NEJMsa1103216

Margolis, K. L., Asche, S. E., Bergdall, A. R., Dehmer, S. P., Groen, S. E., Kadrmas, H. M., Trower, N. K. (2013). Effect of home blood pressure telemonitoring and pharmacist management on blood pressure control: a cluster randomized clinical trial. Jama, 310(1), 46-56. doi: 10.1001/jama.2013.6549

Matchar, D. B., Jacobson, A., Dolor, R., Edson, R., Uyeda, L., Phibbs, C. S., Lavori, P. (2010). Effect of home testing of international normalized ratio on clinical events. N Engl J Med, 363(17), 1608-1620. doi: 10.1056/NEJMoa1002617

McDermott, M. M., Ades, P., Guralnik, J. M., Dyer, A., Ferrucci, L., Liu, K., Criqui, M. H. (2009). Treadmill exercise and resistance training in patients with peripheral arterial disease with and without intermittent claudication: a randomized controlled trial. Jama, 301(2), 165-174. doi: 10.1001/jama.2008.962

McDermott, M. M., Liu, K., Guralnik, J. M., Criqui, M. H., Spring, B., Tian, L., Rejeski, W. J. (2013). Home-based walking exercise intervention in peripheral artery disease: a randomized clinical trial. Jama, 310(1), 57-65. doi: 10.1001/jama.2013.7231

McFall, M., Saxon, A. J., Malte, C. A., Chow, B., Bailey, S., Baker, D. G., Lavori, P. W. (2010). Integrating tobacco cessation into mental health care for posttraumatic stress disorder: a randomized controlled trial. Jama, 304(22), 2485-2493. doi: 10.1001/jama.2010.1769

Messier, S. P., Mihalko, S. L., Legault, C., Miller, G. D., Nicklas, B. J., DeVita, P., Loeser, R. F. (2013). Effects of intensive diet and exercise on knee joint loads, inflammation, and clinical outcomes among overweight and obese adults with knee osteoarthritis: the IDEA randomized clinical trial. Jama, 310(12), 1263-1273. doi: 10.1001/jama.2013.277669

Metsch, L. R., Feaster, D. J., Gooden, L., Schackman, B. R., Matheson, T., Das, M., Colfax, G. N. (2013). Effect of risk-reduction counseling with rapid HIV testing on risk of acquiring sexually transmitted infections: the AWARE randomized clinical trial. Jama, 310(16), 1701-1710. doi: 10.1001/jama.2013.280034

Mohr, D. C., Ho, J., Duffecy, J., Reifler, D., Sokol, L., Burns, M. N., Siddique, J. (2012). Effect of telephone-administered vs face-to-face cognitive behavioral therapy on adherence to therapy and depression outcomes among primary care patients: a randomized trial. Jama, 307(21), 2278-2285. doi: 10.1001/jama.2012.5588

Monson, C. M., Fredman, S. J., Macdonald, A., Pukay-Martin, N. D., Resick, P. A., & Schnurr, P. P. (2012). Effect of cognitive-behavioral couple therapy for PTSD: a randomized controlled trial. Jama, 308(7), 700-709. doi: 10.1001/jama.2012.9307

Morey, M. C., Snyder, D. C., Sloane, R., Cohen, H. J., Peterson, B., Hartman, T. J., Demark-Wahnefried, W. (2009). Effects of home-based diet and exercise on functional outcomes among older, overweight long-term cancer survivors: RENEW: a randomized controlled trial. Jama, 301(18), 1883-1891. doi: 10.1001/jama.2009.643

Nedeltcheva, A. V., Kilkus, J. M., Imperial, J., Schoeller, D. A., & Penev, P. D. (2010). Insufficient sleep undermines dietary efforts to reduce adiposity. Ann Intern Med, 153(7), 435-441. doi: 10.7326/0003-4819-153-7-201010050-00006

Olanow, C. W., Rascol, O., Hauser, R., Feigin, P. D., Jankovic, J., Lang, A., Tolosa, E. (2009). A double-blind, delayed-start trial of rasagiline in Parkinson's disease. N Engl J Med, 361(13), 1268-1278. doi: 10.1056/NEJMoa0809335

Peikes, D., Chen, A., Schore, J., & Brown, R. (2009). Effects of care coordination on hospitalization, quality of care, and health care expenditures among Medicare beneficiaries: 15 randomized trials. Jama, 301(6), 603-618. doi: 10.1001/jama.2009.126

Petersen, L. A., Simpson, K., Pietz, K., Urech, T. H., Hysong, S. J., Profit, J., Woodard, L. D. (2013). Effects of individual physician-level and practice-level financial incentives on hypertension care: a randomized trial. Jama, 310(10), 1042-1050. doi: 10.1001/jama.2013.276303

Powell, L. H., Calvin, J. E., Jr., Richardson, D., Janssen, I., Mendes de Leon, C. F., Flynn, K. J., Avery, E. (2010). Self-management counseling in patients with heart failure: the heart failure adherence and retention randomized behavioral trial. Jama, 304(12), 1331-1338. doi: 10.1001/jama.2010.1362

Rejeski, W. J., Ip, E. H., Bertoni, A. G., Bray, G. A., Evans, G., Gregg, E. W., & Zhang, Q. (2012). Lifestyle change and mobility in obese adults with type 2 diabetes. N Engl J Med, 366(13), 1209-1217. doi: 10.1056/NEJMoa1110294

Robbins, G. K., Lester, W., Johnson, K. L., Chang, Y., Estey, G., Surrao, D., Freedberg, K. A. (2012). Efficacy of a clinical decision-support system in an HIV practice: a randomized trial. Ann Intern Med, 157(11), 757-766. doi: 10.7326/0003-4819-157-11-201212040-00003

Rock, C. L., Flatt, S. W., Sherwood, N. E., Karanja, N., Pakiz, B., & Thomson, C. A. (2010). Effect of a free prepared meal and incentivized weight loss program on weight loss and weight loss maintenance in obese and overweight women: a randomized controlled trial. Jama, 304(16), 1803-1810. doi: 10.1001/jama.2010.1503

Rollman, B. L., Belnap, B. H., LeMenager, M. S., Mazumdar, S., Houck, P. R., Counihan, P. J., Reynolds, C. F., 3rd. (2009). Telephone-delivered collaborative care for treating post-CABG depression: a randomized controlled trial. Jama, 302(19), 2095-2103. doi: 10.1001/jama.2009.1670

Sacks, F. M., Bray, G. A., Carey, V. J., Smith, S. R., Ryan, D. H., Anton, S. D., Williamson, D. A. (2009). Comparison of weight-loss diets with different compositions of fat, protein, and carbohydrates. N Engl J Med, 360(9), 859-873. doi: 10.1056/NEJMoa0804748

Sadowski, L. S., Kee, R. A., VanderWeele, T. J., & Buchanan, D. (2009). Effect of a housing and case management program on emergency department visits and hospitalizations among chronically ill homeless adults: a randomized trial. Jama, 301(17), 1771-1778. doi: 10.1001/jama.2009.561

Safren, S. A., Sprich, S., Mimiaga, M. J., Surman, C., Knouse, L., Groves, M., & Otto, M. W. (2010). Cognitive behavioral therapy vs relaxation with educational support for medication-treated adults with ADHD and persistent symptoms: a randomized controlled trial. Jama, 304(8), 875-880. doi: 10.1001/jama.2010.1192

Saitz, R., Cheng, D. M., Winter, M., Kim, T. W., Meli, S. M., Allensworth-Davies, D., Samet, J. H. (2013). Chronic care management for dependence on alcohol and other drugs: the AHEAD randomized trial. Jama, 310(11), 1156-1167. doi: 10.1001/jama.2013.277609

Schmitz, K. H., Ahmed, R. L., Troxel, A., Cheville, A., Smith, R., Lewis-Grant, L., Greene, Q. P. (2009). Weight lifting in women with breast-cancer-related lymphedema. N Engl J Med, 361(7), 664-673. doi: 10.1056/NEJMoa0810118

Schmitz, K. H., Ahmed, R. L., Troxel, A. B., Cheville, A., Lewis-Grant, L., Smith, R., Chittams, J. (2010). Weight lifting for women at risk for breast cancer-related lymphedema: a randomized trial. Jama, 304(24), 2699-2705. doi: 10.1001/jama.2010.1837

Schoen, R. E., Pinsky, P. F., Weissfeld, J. L., Yokochi, L. A., Church, T., Laiyemo, A. O., Berg, C. D. (2012). Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. N Engl J Med, 366(25), 2345-2357. doi: 10.1056/NEJMoa1114635

Schwartz, L. M., Woloshin, S., & Welch, H. G. (2009). Using a drug facts box to communicate drug benefits and harms: two randomized trials. Ann Intern Med, 150(8), 516-527.

Sequist, T. D., Fitzmaurice, G. M., Marshall, R., Shaykevich, S., Marston, A., Safran, D. G., & Ayanian, J. Z. (2010). Cultural competency training and performance reports to improve diabetes care for black patients: a cluster randomized, controlled trial. Ann Intern Med, 152(1), 40-46. doi: 10.7326/0003-4819-152-1-201001050-00009

Shaukat, A., Mongin, S. J., Geisser, M. S., Lederle, F. A., Bond, J. H., Mandel, J. S., & Church, T. R. (2013). Long-term mortality after screening for colorectal cancer. N Engl J Med, 369(12), 1106-1114. doi: 10.1056/NEJMoa1300720

Shorr, R. I., Chandler, A. M., Mion, L. C., Waters, T. M., Liu, M., Daniels, M. J., Miller, S. T. (2012). Effects of an intervention to increase bed alarm use to prevent falls in hospitalized patients: a cluster randomized trial. Ann Intern Med, 157(10), 692-699. doi: 10.7326/0003-4819-157-10-201211200-00005

Song, J., Ratner, E. R., Wall, M. M., Bartels, D. M., Ulvestad, N., Petroskas, D., Gelberg, L. (2010). Effect of an End-of-Life Planning Intervention on the completion of advance directives in homeless persons: a randomized trial. Ann Intern Med, 153(2), 76-84. doi: 10.7326/0003-4819-153-2-201007200-00003

Stanley, M. A., Wilson, N. L., Novy, D. M., Rhoades, H. M., Wagener, P. D., Greisinger, A. J., Kunik, M. E. (2009). Cognitive behavior therapy for generalized anxiety disorder among older adults in primary care: a randomized clinical trial. Jama, 301(14), 1460-1467. doi: 10.1001/jama.2009.458

Stockwell, M. S., Kharbanda, E. O., Martinez, R. A., Vargas, C. Y., Vawdrey, D. K., & Camargo, S. (2012). Effect of a text messaging intervention on influenza vaccination in an urban, low-income pediatric and adolescent population: a randomized controlled trial. Jama, 307(16), 1702-1708. doi: 10.1001/jama.2012.502

Subak, L. L., Wing, R., West, D. S., Franklin, F., Vittinghoff, E., Creasman, J. M., Grady, D. (2009). Weight loss to treat urinary incontinence in overweight and obese women. N Engl J Med, 360(5), 481-490. doi: 10.1056/NEJMoa0806375

Sullivan, C., Sayre, S. S., Leon, J. B., Machekano, R., Love, T. E., Porter, D., Sehgal, A. R. (2009). Effect of food additives on hyperphosphatemia among patients with end-stage renal disease: a randomized controlled trial. Jama, 301(6), 629-635. doi: 10.1001/jama.2009.96

Temel, J. S., Greer, J. A., Muzikansky, A., Gallagher, E. R., Admane, S., Jackson, V. A., Lynch, T. J. (2010). Early palliative care for patients with metastatic non-small-cell lung cancer. N Engl J Med, 363(8), 733-742. doi: 10.1056/NEJMoa1000678

Thornton, J. D., Alejandro-Rodriguez, M., Leon, J. B., Albert, J. M., Baldeon, E. L., De Jesus, L. M., Sehgal, A. R. (2012). Effect of an iPod video intervention on consent to donate organs: a randomized trial. Ann Intern Med, 156(7), 483-490. doi: 10.7326/0003-4819-156-7-201204030-00004

Tulsky, J. A., Arnold, R. M., Alexander, S. C., Olsen, M. K., Jeffreys, A. S., Rodriguez, K. L., Pollak, K. I. (2011). Enhancing communication between oncologists and patients with a computer-based training program: a randomized trial. Ann Intern Med, 155(9), 593-601. doi: 10.7326/0003-4819-155-9-201111010-00007

Villareal, D. T., Chode, S., Parimi, N., Sinacore, D. R., Hilton, T., Armamento-Villareal, R., Shah, K. (2011). Weight loss, exercise, or both and physical function in obese older adults. N Engl J Med, 364(13), 1218-1229. doi: 10.1056/NEJMoa1008234

Volpp, K. G., Shea, J. A., Small, D. S., Basner, M., Zhu, J., Norton, L., Dinges, D. F. (2012). Effect of a protected sleep period on hours slept during extended overnight in-hospital duty hours among medical interns: a randomized trial. Jama, 308(21), 2208-2217. doi: 10.1001/jama.2012.34490

Volpp, K. G., Troxel, A. B., Pauly, M. V., Glick, H. A., Puig, A., Asch, D. A., Audrain-McGovern, J. (2009). A randomized, controlled trial of financial incentives for smoking cessation. N Engl J Med, 360(7), 699-709. doi: 10.1056/NEJMsa0806819

Wadden, T. A., Volger, S., Sarwer, D. B., Vetter, M. L., Tsai, A. G., Berkowitz, R. I., Moore, R. H. (2011). A two-year randomized trial of obesity treatment in primary care practice. N Engl J Med, 365(21), 1969-1979. doi: 10.1056/NEJMoa1109220

Wang, C., Schmid, C. H., Rones, R., Kalish, R., Yinh, J., Goldenberg, D. L., McAlindon, T. (2010). A randomized trial of tai chi for fibromyalgia. N Engl J Med, 363(8), 743-754. doi: 10.1056/NEJMoa0912611

Wennberg, D. E., Marr, A., Lang, L., O'Malley, S., & Bennett, G. (2010). A randomized trial of a telephone care-management strategy. N Engl J Med, 363(13), 1245-1255. doi: 10.1056/NEJMsa0902321

Wing, R. R., Bolin, P., Brancati, F. L., Bray, G. A., Clark, J. M., Coday, M., Yanovski, S. Z. (2013). Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes. N Engl J Med, 369(2), 145-154. doi: 10.1056/NEJMoa1212914

Woloshin, S., & Schwartz, L. M. (2011). Communicating data about the benefits and harms of treatment: a randomized trial. Ann Intern Med, 155(2), 87-96. doi: 10.7326/0003-4819-155-2-201107190-00004

Young, S. D., Cumberland, W. G., Lee, S. J., Jaganath, D., Szekeres, G., & Coates, T. (2013). Social networking technologies as an emerging tool for HIV prevention: a cluster randomized trial. Ann Intern Med, 159(5), 318-324. doi: 10.7326/0003-4819-159-5-201309030-00005

From Top Economics Journals (2009-2013)

Charness, G., & Gneezy, U. (2009). Incentives to Exercise. Econometrica, 77(3), 909-931. doi: 10.3982/ecta7416

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J.P., Allen, H., Baicker, K., & Oregon Health Study Group. (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. Quarterly Journal of Economics, 127(3), 1057-1106. doi: http://qje.oxfordjournals.org/content/127/3/1057

From Top Health Services Journals (2009-2013)

Allen, H., Baicker, K., Finkelstein, A., Taubman, S., & Wright, B. J. (2010). What the Oregon health study can tell us about expanding Medicaid. Health Aff (Millwood), 29(8), 1498-1506. doi: 10.1377/hlthaff.2010.0191

Barnes, D. E., Palmer, R. M., Kresevic, D. M., Fortinsky, R. H., Kowal, J., Chren, M. M., & Landefeld, C. S. (2012). Acute care for elders units produced shorter hospital stays at lower cost while maintaining patients' functional status. Health Aff (Millwood), 31(6), 1227-1236. doi: 10.1377/hlthaff.2012.0142

Halpern, S. D., Loewenstein, G., Volpp, K. G., Cooney, E., Vranas, K., Quill, C. M., Bryce, C.

(2013). Default options in advance directives influence how patients set goals for end-of-life care. Health Aff (Millwood), 32(2), 408-417. doi: 10.1377/hlthaff.2012.0895

Krebs, E. E., Bair, M. J., Damush, T. M., Tu, W., Wu, J., & Kroenke, K. (2010). Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. Med Care, 48(11), 1007-1014. doi: 10.1097/MLR.0b013e3181eaf835

Leveille, S. G., Huang, A., Tsai, S. B., Allen, M., Weingart, S. N., & Iezzoni, L. I. (2009). Health coaching via an internet portal for primary care patients with chronic conditions: a randomized controlled trial. Med Care, 47(1), 41-47. doi: 10.1097/MLR.0b013e3181844dd0

Michalopoulos, C., Wittenburg, D., Israel, D. A., & Warren, A. (2012). The effects of health care benefits on health care use and health: a randomized trial for disability insurance beneficiaries. Med Care, 50(9), 764-771. doi: 10.1097/MLR.0b013e31825a8bfc

Mosen, D. M., Feldstein, A. C., Perrin, N., Rosales, A. G., Smith, D. H., Liles, E. G., Glasgow, R. E. (2010). Automated telephone calls improved completion of fecal occult blood testing. Med Care, 48(7), 604-610. doi: 10.1097/MLR.0b013e3181dbdce7

Muennig, P., Rosen, Z., & Wilde, E. T. (2013). Welfare programs that target workforce participation may negatively affect mortality. Health Aff (Millwood), 32(6), 1072-1077. doi: 10.1377/hlthaff.2012.0971

Mundt, M. P., & Zakletskaia, L. I. (2012). Prevention for college students who suffer alcohol-induced blackouts could deter high-cost emergency department visits. Health Aff (Millwood), 31(4), 863-870. doi: 10.1377/hlthaff.2010.1140

Newhouse, R. P., Dennison Himmelfarb, C., Morlock, L., Frick, K. D., Pronovost, P., & Liang, Y. (2013). A phased cluster-randomized trial of rural hospitals testing a quality collaborative to improve heart failure care: organizational context matters. Med Care, 51(5), 396-403. doi: 10.1097/MLR.0b013e318286e32e

Nietert, P. J., Tilley, B. C., Zhao, W., Edwards, P. F., Wessell, A. M., Mauldin, P. D., & Polk, P. P. (2009). Two pharmacy interventions to improve refill persistence for chronic disease medications: a randomized, controlled trial. Med Care, 47(1), 32-40. doi: 10.1097/MLR.0b013e3181808c17

Piette, J. D., Richardson, C., Himle, J., Duffy, S., Torres, T., Vogel, M., Valenstein, M. (2011). A randomized trial of telephonic counseling plus walking for depressed diabetes patients. Med Care, 49(7), 641-648. doi: 10.1097/MLR.0b013e318215d0c9

Veroff, D., Marr, A., & Wennberg, D. E. (2013). Enhanced support for shared decision making reduced costs of care for patients with preference-sensitive conditions. Health Aff (Millwood), 32(2), 285-293. doi: 10.1377/hlthaff.2011.0941

Zarkin, G. A., Bray, J. W., Aldridge, A., Mills, M., Cisler, R. A., Couper, D., O'Malley, S. (2010). The effect of alcohol treatment on social costs of alcohol dependence: results from the COMBINE study. Med Care, 48(5), 396-401. doi: 10.1097/MLR.0b013e3181d68859

<u>From Evaluation Firms (2013)</u>

Farrell, M. (2013). Connections Between TANF and SSI:Lessons from the TANF/SSI Disability Transition Project. Washington, D.C.

Filene, J. H., Snell, E., Lee, H., Knox, V., Michalopoulos, C., & Duggan, A. (2013). First Annual Report from the Mother and Infant Home Visiting Program Evaluation-Strong Start. New York, NY: MDRC.

Hossain, F., Baird, P., & Pardoe, R. (2013). Improving Employment Outcomes and Community Integration for Veterans with Disabilities. New York, NY: MDRC.

Michalopoulos, C., Manno, M. S., Warren, A., & Somers, J. (2013). Final Report on the Kaiser Permanente Colorado Coordinated Care Pilot Program. New York, NY: MDRC.

Michalopoulos, C., Wittenburg, D., Israel, D. A. R., Schore, J., Warren, A., Zutshi, A., Schwartz,

L. (2013). The Accelerated Benefits Demonstration and Evaluation Project. New York, NY: MDRC.

Moreno, L., Felt-Lisk, S., & Dale, S. (2013). Do Financial Incentives Increase the Use of Electronic Health Records? Princeton, NJ: Mathematica Policy Research.

Riccio, J., Dechausay, N., Miller, C., Nunez, S., Verma, N., & Yang, E. (2013). Conditional Cash Transfers in New York City: The Continuing Story of the Opportunity NYC—Family Rewards Demonstration. (2013). New York, NY.

Shapiro, R. (2013). Balancing Fidelity and Flexibility: Implementing the Gen.M Program in Texas. Princeton, NJ: Mathematica Policy Research.

Wagner, G., Lovely, P., & Schneider, S. (2013). Pilot Controlled Trial of the Adherence Readiness Program. Aids and Behavior, 17(9), 3059-3065.

Yuan, K., Hamilton, L. S., Stecher, B. M., & Pane, J. F. Evaluating the Effectiveness of Teacher Pay-for-Performance. (2013). New York, NY: RAND Corporation.

Appendix C: Timing of Randomization and Statistical Power

An important issue in understanding the time and cost considerations for an RCT is when the randomization will be conducted: after recruiting, screening, and enrolling patients in the study (as is the standard in drug trials), or at the point of offering individuals access to the program. In the main text we discussed some of the tradeoffs involved in these two approaches. Here we expand on the discussion of the substantial loss of statistical power associated with randomizing at the point of offering instead of after enrollment.

In particular, if randomization occurs at the point of offering, and only X percent of those offered take up the program, the sample size needed to detect a given effect size will be $(100/X)^2$ times the sample size needed to detect that effect size with 100 percent take-up (as is usually close to the case if randomization occurs after enrollment). This is because while statistical power increases with sample size, it does so at a diminishing rate. For the intent-to-treat effect of an intervention estimated using OLS regression, statistical power is proportional to the t-statistic. Since $t = \beta / \frac{\sigma}{\sqrt{n}}$ where β is the coefficient on the intervention and n is the sample size, multiplying β (the intent-to-treat effect) by X is equivalent in its impact on power to multiplying *n* (the sample size) by X^2.

Imagine two RCTs using a target population of 200 people, of whom 50 percent will be willing to participate. In Trial A, we randomize 100 people to the control group and 100 people to the treatment group. Of those 100 assigned to the treatment group, all are offered the intervention and 50 (50 percent) chose to participate. In Trial B, we first approach all 200 people as ask them to enroll. We find 100 people (50 percent) who choose to participate and randomize them. We are left with 50 people in the control group and 50 people in the treatment group.

In Trial A compared to Trial B, the sample size is double (200 vs. 100), but the average treatment effect is half as big (since only 50 of the 100 treatments in Trial A get the benefit). Because there are diminishing marginal returns to sample size on power, doubling the sample size does not result in statistical power to detect half the effect. Trial B is higher powered for the same true treatment effect despite being smaller. Put another way, because power declines linearly in the take-up rate while it increases only with the square root of sample size, if with 100 percent take up of the intervention in the treatment arm (and 0 in the control) one would need a sample size of n to have power to detect a given effect size, then with 20 percent take-up of the intervention in the treatment arm (and 0 in the control) one would need a sample size of 25n to have power to detect the same effect size of the intervention.