Temple University
Beasley School of Law

Center for
Public Health Law
Research

RANDOMIZED TRIALS IN LEGAL EPIDEMIOLOGY

Harold Pollack

Helen Ross Professor, Crown Family School of Social Work, Policy and Practice, University of Chicago

Alida Bouris, PhD, MSW

Associate Professor, Crown Family School of Social Work, Policy and Practice, University of Chicago

Scott Cunningham

Professor, Baylor University

A Methods Monograph for the Center for Public Health Law Research Temple University Beasley School of Law

OCTOBER 2023

RANDOMIZED TRIALS IN LEGAL EPIDEMIOLOGY

Harold Pollack

Alida Bouris

Scott Cunningham

Summary

This chapter reviews the utility of randomized trials to study *policy candidates*, establish specific causal links and *mechanisms* of action, and *evaluate effects of actual laws* implemented in the real world. It begins by presenting the need for randomized trials to address selection bias and related challenges that arise in policy evaluation. It presents basic trial concepts such as intent-to-treat and treatment-on-the-treated effects using the example of a simplified intervention to reduce underage alcohol consumption. It explores strengths and limitations of randomized trials by discussing a successful experimental evaluation of a contingency management intervention to prevent prenatal smoking. We describe cluster-randomized trials, and discuss the importance of specifying mediating mechanisms and causal pathways in the proper interpretation of randomized trials. Finally, we present common critiques of randomized trials, and the use of mechanism experiments and hybrid implementation science designs to address limitations of "black-box" randomized trials, which do not address causal mechanisms of observed effects.

Learning Objectives

- Recognize the distinct uses of randomized trials in devising, implementing, and evaluating laws to improve public health.
- Understand the different types of randomized trials, and different units of randomization.
- Articulate challenges to internal validity and generalizability of experimental methods.
- Articulate distinctions between randomized trials of specific interventions or policy candidates and experimental law evaluations.

- Articulate the role of implementation science and hybrid trials to explore barriers, facilitators, and contextual factors.

Randomized controlled trials (RCTs) are regarded by many policymakers and researchers as the gold standard of policy evaluation. Trials such as the Oregon Health Insurance experiment demonstrate that expanded health coverage can significantly improve mental health, increase use of prescription medications for chronic conditions and reduce unmet dental needs, while also protecting individuals from adverse financial outcomes associated with illness and injury (Baicker, Allen, Wright, & Finkelstein, 2017; Baicker, Allen, Wright, Taubman, & Finkelstein, 2018). Conversely, the same trials provided disappointing findings regarding the power of expanded health insurance coverage to substantially affect key health outcomes such as blood pressure and debunked assertions that generous health insurance coverage would reduce emergency department use (Finkelstein, Taubman, Allen, Wright, & Baicker, 2016; Taubman, Allen, Wright, Baicker, & Finkelstein, 2014). Other trials have provided important, at-times chastening, results regarding promising models such as healthcare hotspotting (i.e. reducing costs of healthcare "superusers") (Finkelstein, Zhou, Taubman, & Doyle, 2020).

This chapter reviews key issues in randomized trials for public health policy and law. We begin by presenting the need for randomized trials. We describe some different purposes for randomized trials in public health law. We then present basic trial concepts such as intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects in the context of a posited intervention to reduce underage alcohol consumption. We then apply the same concepts in the context of another public health challenge – interventions to reduce the incidence of low infant birthweight by reducing smoking by pregnant patients. We show strengths and limitations of a successful randomized trial of contingency management interventions, and we compare trial results to those obtained through econometric analyses of state tobacco tax policies that also prevent prenatal smoking. We note the different levels of aggregation susceptible to experimental designs with a summary discussion of two cluster-randomized trials. We discuss the connection between cluster-randomized designs, stepped-wedge, and quasi-experimental designs, and discuss the importance of specifying mediating mechanisms and causal pathways in the proper interpretation of these study designs. We then present common critiques of randomized trials, and the use of mechanism experiments, before closing the chapter with a discussion of implementation science designs to address limitations of atheoretical "black-box" randomized trials.

Purposes of Randomized Trials

Randomized trials in legal epidemiology are useful tools for three distinct purposes. *Mechanism experiments* focus on one or a few specific causal links in a broader theory or longer causal chain hypothesized to be the reason a particular law has health effects. Such experiments can provide

convincing evidence that a particular policy approach appears sufficiently likely to be effective to be worth trying. Second, well-designed randomized trials are often feasible for specific *policy candidates* – programs and interventions that could be delivered, supported, or facilitated through laws, rules, and regulations. We describe two such policy candidates here: A hypothesized intervention to prevent underage drinking, and contingency management interventions whereby pregnant patients might receive immediate financial rewards as behavioral incentives to reduce tobacco use. An RCT of such interventions could clarify for policymakers whether insurers should be legally required to support such interventions, what might be expected (and what remains unknown) regarding the likely public health benefits if policymakers succeed in supporting the proliferation of such interventions.

There is an important distinction between RCTs on mechanisms and policy candidates, versus RCTs of actual laws and regulations. *Randomized trials of actual laws* to date are rare, but are an important third purpose where randomization as a study design element can be fruitfully used to advance the state of knowledge. Consider, we already know that increased safety belt use reduces automobile fatalities. We already know that driving under the influence of intoxicating substances increases road fatalities, and that underage drinking is associated with many public health harms. We already know that smoking by pregnant patients increases the risk of low infant birthweight and infant mortality. We might want to know more about effects of specific programs and other interventions that address these harms, but that knowledge alone is insufficient to improve use of law to scale those interventions.

Even if we know that a particular intervention is effective, we still need to know whether concrete legislation or regulations based on that intervention will meaningfully improve public health: If state Medicaid programs require insurers to cover prenatal smoking cessation services, will such policies actually improve population infant health? If states or counties intensify traffic enforcement or increase penalties for driving under the influence, does this actually reduce related automobile crashes and road fatalities? That is a question for true policy RCTs, which are informed by – but extend far beyond – rigorous evaluations of specific interventions in specific settings.

The Need for Randomized Trials: Concepts and Examples

Public health laws frequently seek to support, generalize or scale specific interventions that proved efficacious or effective in smaller scale randomized trials. In the absence of these prior trials, the benefits of promising program models often prove illusory or overstated, because initial studies faced methodological limitations that limited causal inference, that failed to illuminate plausible causal mechanisms for observed apparent benefits, or that prevented researchers from ruling out alternative explanations for observed effects. And, randomized intervention trials provide only one scientifically sound path through which one identifies candidates for public health policy. For example, regulatory policies to control particulate pollution may emerge from laboratory or engineering studies that shed light on potential health harms.

To gauge the basic risks of relying on observational studies to infer that an intervention caused a result, consider interventions designed to address underage alcohol consumption – a key public health concern. Suppose policymakers are intrigued by motivational interview (MI)-based interventions to reduce underage drinking and are contemplating changes to Medicaid reimbursement policies to cover such specific interventions, and changes to state curricular standards to encourage or require such interventions. Poignant testimonials by program participants and program staff suggest that participation in this intervention reduces underage drinking. Psychologists and social workers provide a strong conceptual explanation of why such programs might be effective, perhaps drawing on research showing the approach worked for other similar problems.

One might try to test these claims by comparing underage drinking patterns of youth who participated in the intervention to those of youth who didn't participate. This comparison is clearly vulnerable to selection bias. Most obviously, this approach may reflect *favorable selection,* whereby students who care more about health harms of alcohol use (or students whose parents care more about these issues) may seek out the intervention. Of course, many of these youth would have avoided underage drinking absent the intervention. Straightforward comparisons between participants and non-participants would thus overstate the intervention's causal impact. Controlling for observed student and family characteristics such as income and parental education might reduce such biases, but would not reliably eliminate them.

Alternatively, the program may exhibit *unfavorable selection*. This can happen if program implementors specifically recruit students they believe have greatest alcohol-related difficulties. Some of these factors might be observed and statistically addressed by researchers. For example, recruited students might have more school absences or lower test scores. Other relevant factors often cannot be fully observed by the researcher: recent social challenges in school, perhaps confidential contacts with school counselors whose records are not shared with the researchers. If this were the underlying dynamic, many intervention participants might still engage in underage drinking, but might have consumed even greater amounts without the intervention. Thus, straightforward comparisons of drinking patterns between participants and non-participants would likely understate the value of the intervention.

Controlling for observed student and family characteristics might reduce both kinds of biases. It will not eliminate them. An impressively complex non-experimental design might even make things worse, by making researchers or policymakers overconfident in the accompanying results. Robert LaLonde's (1986) famous re-analysis of experimental job-training data using non-experimental methods provides one chastening example. Common non-experimental econometric approaches yield results far from the experimental results. Even worse, these methods yielded tight confidence intervals that (incorrectly) excluded the experimental results.

A randomized trial may address these concerns, particularly in the evaluation of a specific policy candidate intervention. Suppose one performs an *encouragement-design* randomized trial of our intervention to reduce underage drinking. Here, researchers offer $10 video game gift cards to every high school senior assigned to the treatment group who actually attends the Saturday session. Students assigned to the control group can still attend the session, but would not get the gift card. Researchers keep track of the treatment "dose" each student receives, and track an adolescent drinking measure, say ounces of alcohol consumed in the past month.

In particular, let Z be a dummy variable signifying group assignment. Suppose that 55% of students assigned to the treatment group (Z=1) attend the prevention session, compared with 45% of those assigned to control (Z=0). Let's further suppose that mean monthly alcohol consumption is 61 ounces in the treatment group, and 63 ounces in the control group.

These deceptively simple results, shown in Table 13.1, help us to interpret and describe the effect of the posited Saturday intervention, and to present basic concepts of randomized trials.

	Total study population (n=1000)	Offered gamecards (n=500) (Z=1)	Not offered gamecards (n=500) (Z=0)
Proportion attending Saturday Session	500	275	225
Proportion not attending Saturday Session	500	225	275
Mean monthly alcohol consumption (ounces)	62 ounces	61 ounces	63 ounces

Table 13.1. Hypothetical encouragement trial to reduce underage alcohol consumption.

INTENT-TO-TREAT (ITT) VS. EFFECT OF TREATMENT ON THE TREATED (TOT)

As we frame this question, it becomes apparent that the "program effectiveness" has two complementary interpretations. Policymakers might ask one bottom-line question: How much, overall, might we reduce underage drinking by offering this voluntary Saturday intervention to everyone who is willing to participate? This is the *intent-to-treat* (ITT) effect. The ITT takes into account that not every patient or study subject actually takes-up the treatment. Perhaps the Saturday intervention is boring or unpleasant, is only offered in English within a multi-lingual student population, or is offered at a time that many students cannot attend due to work, school, or family obligations. Perhaps some students are simply uninterested in reducing their alcohol use, and thus see little value in participating.

Of course, we might also ask a different question: How much do we reduce underage drinking *among the students who actually attend?* This question is often labeled *the effect of treatment on the treated*, or TOT. This speaks to the specific value and effectiveness of the treatment for people who actually receive it.

TAKE-UP AND COMPLIANCE AS CENTRAL TO EFFECTIVENESS

ITT and TOT estimates often diverge in real-world trials. A new cancer drug might be powerfully effective for patients who actually take it. This is a critical accomplishment that speaks to the biological effect of the medication on this particular cancer. Yet the medication might have unpleasant or toxic side-effects, and thus low patient acceptability and correspondingly disappointing treatment benefits for the population of cancer patients one seeks to help.

The contrast between ITT and TOT also underscores a critical issue in many public health policies – the importance of exposure, awareness, engagement, and compliance in shaping the magnitude of effects of health laws or interventions. Take-up, compliance, and recruitment are central to program effectiveness. A particular preventive intervention can exhibit a strong TOT effect in a randomized trial. Yet that effect often does not hold up when implemented universally through legal incentives or mandates.

A recent intervention to provide supports for male pre-trial detainees leaving jail provides an illustrative case. The program aimed to prevent homelessness, reduce rearrests, and improve health outcomes among returning citizens who live with serious mental illness, substance use disorders, and related challenges. The intervention sought to improve these outcomes by creating immediate linkages to services and by offering participants a safe place to spend the night. The program significantly reduced risks of immediate rearrest among program participants. Yet program take-up was twice as high among released detainees aged 55 and older as among those ages 18–35. Given these patterns, it is plausible that the program can reduce homelessness, which is quite prevalent among older detainees. But it is much less likely to reduce violent re-offending, given low program take-up among younger offenders in the peak age-group associated with violent crime. Anecdotal evidence from program staff indicated that young men offered the intervention found that the central sales pitch – a place to stay for the night for those who might otherwise be homeless—to be a specifically stigmatizing message, and thus chose not to participate.

COVID-19 vaccination poses analogous take-up challenges, with more dramatic effects for population health and health disparities. An overwhelming body of evidence indicates that vaccination reduces infection risk, while dramatically reducing risks of serious illness, hospitalization, and death. Given this body of evidence, differential vaccination rates by political party, race-ethnicity, and other characteristics poses a substantial challenge to population health. The emergence of rural residents and political conservatives as key disparity groups poses a particular challenge to the public health community (Kirzinger et al., 2021). Existing RCTs suggest that culturally-competent messages delivered by trusted messengers is likely important to promote protective measures within affected communities (Breza et al., 2021a; Breza et al., 2021b; Torres et al., 2021). Randomized trials might explore the effects of different culturally-competent public health messaging strategies in eliciting compliance with work- or school-based vaccine mandates.

Our simplified underage drinking example provides a useful framework to present basic distinctions and nomenclature of the ITT and TOT effects. Under proper assumptions, within this simple framework of a binary intervention, the ITT effect estimate is the difference in drinking between the group assigned to treatment (offered gamecards) vs. the group assigned to the control (not offered gamecards), as in equation (1):

$$\beta_{ITT} = E[Ounces|Z = 1] - E[Ounces|Z = 0] \quad (1)$$

Here $Ounces_i$ is monthly alcohol consumption for participant i measured after random assignment, Z is an indicator for having been offered gamecards. Under these assumptions, the intent-to-treat estimate (ITT) is (63-61) = 2 ounces. That represents the effect of *inviting* students to the intervention. This provides one valuable metric of the likely population impact of this intervention – if the intervention could be scaled at the same level of effectiveness.

What about the TOT effect – that is, the causal effect of the intervention for students who actually attended? Under conventional assumptions, the observed two percentage point difference in alcohol consumption entirely arises from the 10 percentage point increase in attendance between the treatment and control groups. For this binary treatment framework, the classic Wald estimate provides a simple but intuitive approximation to the TOT given binary group assignment, by taking into account differences in intervention dose (i.e. attendance) across groups – how many people in each group actually received the intervention:

$$\beta_{Wald} = \frac{\overline{Ounces}_{Z=1} - \overline{Ounces}_{Z=0}}{\overline{Dose}_{Z=1} - \overline{Dose}_{Z=0}} = \frac{63 - 61}{0.55 - 0.45} = 20\ ounces. \quad (2)$$

Despite the modest 2-point ITT effect, considering attendance across groups suggests that the intervention exerts a surprisingly powerful influence – nearly one-third reduction in underage drinking – specifically for those students who actually attend. Efforts to improve program take-up – e.g. through provision of attractive incentives or even mandates if the intervention were fully imbedded in regular school curricula – might magnify the public health impact of this intervention, assuming that the program can be similarly effective among the new students one attracts as reach expands.

Notice that we *assume* that the entire reduction in alcohol consumption arises from the additional intervention provided to youth in the treatment group. There is no other benefit to being assigned to the treatment group, and there is no benefit to assignment to the control group. A

moment's thought calls to mind how these assumptions might be violated. Intervention participants might share program materials with their friends. Sharing materials with friends in the treatment group who prefer not to attend would complicate our interpretation of what this intervention was and how it actually worked. Treatment group participants or intervention implementors might also share materials with students in the control group. Alternatively, some control-group youth could have heard that this intervention seems to help. They and their parents might seek other, similar resources outside the school that the researcher never measured. If such crossover efforts were common, researchers would *understate* the benefits of the Saturday intervention, because some control group participants received benefits similar to the intervention.

Peer effects can also arise in more subtle ways. Suppose the intervention raises school-wide awareness about harms associated with underage drinking – and underscores to students that adults are concerned about the issue. One could imagine that many students in both the treatment and control groups – and others not in the study at all – might be more wary about holding parties where alcohol is served. That could reduce total underage drinking throughout the school.

Other challenges might lead us to overstate or understate the value of the intervention, or may limit generalizability of our results. Students in the treatment group develop personal relationships with program staff. They might conceal some drinking behaviors out of embarrassment or out of a desire not to disappoint program implementors. Researchers might have implicitly or explicitly excluded students from the initial randomization whom they believe might be disruptive or are unlikely to benefit from the intervention. As long as students are properly randomized, this would not undermine internal validity – the causal attribution. It would, however, plausibly undermine generalizability – the population to which the results apply. Another threat to generalizability stems from the virtues of the trial itself. An innovative intervention site might in any number of ways have greater resources or provide higher-quality services than are readily replicated at-scale within representative educational settings. Staff might have higher morale or superior management and training. The research intervention might be implemented within a setting with unusually strong supports to help it succeed, as a showcase intervention strongly supported by the high school principal and teaching staff.

Heterogeneous treatment effects pose additional challenges. We must consider *who* took up the program, and why that might matter if the treatment were more effective for some students than for others. This trial is particularly dependent on the effect of the intervention for students who enjoy video game coupons and thus are disproportionately induced to participate through this specific incentive. Our study design captures the *local average treatment effect* – the benefit accruing to students willing to participate in this intervention who are assigned to the treatment as opposed to the control group.

Similar issues arise in other settings and with other study designs. Suppose a state public health department seeks to reduce the incidence of fatal overdoses caused by the opioid epidemic. It

intervenes by distributing free naloxone kits to anyone who requests them, by passing a Good Samaritan law, and by taking other steps to facilitate harm reduction interventions. We might ask an ITT question: If we make naloxone legally available to everyone, how much does this reduce the state's overall opioid mortality? Alternatively, we might want to know what happens to fatal overdoses among those who actually accept the free naloxone. That is a question about the TOT.

There may be divergent answers to these questions. We know from myriad studies and long clinical experience that naloxone, when properly administered, sharply reduces opioid overdose mortality. There's no need for new information there. We have less information regarding the impact of laws and policies designed to expand naloxone use. Suppose few people who use drugs take up the offer for free naloxone. Maybe people accept the kits, but are not able to engage peers, friends, or loved ones to be present and to properly administer naloxone when overdose occurs. Maybe promotion materials are only available in English. Maybe the state does poor outreach. If take-up is low, the population health effect of offering free naloxone will be small, even though naloxone is powerfully effective when used correctly.

Experiments to explore the effectiveness of naloxone *laws* are quite different from experiments to examine effects of naloxone itself. There will be questions ranging from drug users' awareness of legal changes to whether or not pharmacists or physicians take advantage of expanded authority to dispense or prescribe the medication. As discussed below, a state might support an experimental test of the policy by implementing it through a *stepped-wedge* randomized design, whereby an initial group of randomly selected counties are first to roll it out, with other counties randomly selected to roll out the policy six months later, one year later, and so on.

However, the evaluation is designed, the ultimate impact of the naloxone policy will be shaped by the size and behavior of the group of service providers and people who use drugs who actually engage the program. If the program is powerfully protective for those who embrace naloxone, yet take-up is low, we will observe a policy effect analogous to the underage drinking intervention described above: An impressive TOT, accompanied by a disappointing population ITT effect because so few people embrace the intervention.

A REAL-WORLD EXAMPLE: CONTINGENCY MANAGEMENT FOR PRENATAL SMOKING CESSATION

Prenatal smoking provides a classic challenge, one at the boundary of clinical care and population health (Nighbor et al., 2020). Measures to address this challenge provide another illustration of the necessity and the limitations of randomized controlled trials to improve population health. Here randomized intervention trials may play an important role in state health insurance coverage mandates and in other regulatory policies.

Epidemiologists have long documented that tobacco use during pregnancy is associated with increased incidence of low birthweight, particularly low birthweight arising from fetal growth restriction. Such low birthweight is associated with infant mortality and other adverse outcomes.

Researchers have established clear biological pathways for some of these associations, and have documented large observed birthweight differences between infants born to pregnant smokers and those born to non-smokers (Lewandowska, Wieckowska, Sztorc, & Sajdak, 2020).

It is plausible that initiating smoking cessation interventions for pregnant patients can improve infant health. Yet for many reasons, large observed differences in health outcomes may overstate the direct causal impact of smoking during pregnancy. Some of the birthweight effect of prenatal smoking may arise during the preconception period if patients become pregnant in poor health.

Prenatal smoking is strongly negatively correlated with patient income and education (Higgins et al., 2009), correlated with maternal depression and other risk-factors (Yang & Hall, 2019; Yang et al., 2017) and correlated with other forms of substance use. For example, the majority of pregnant patients with opioid use disorders are also tobacco users (Isaacs et al., 2021). Many forms of co-occurring substance use are more deeply stigmatized than tobacco use, and are correspondingly less likely to be disclosed to clinicians or to researchers. If such substance use contributes to adverse birth outcomes and are more common among pregnant smokers, direct birthweight comparisons of infants born to pregnant smokers with infants born to non-smokers may overstate the causal role of prenatal smoking, (Noble et al., 1997; Vega, Kolody, Hwang, & Noble, 1993) and may correspondingly overstate potential health benefits of laws, policies, and clinical interventions designed to reduce prenatal smoking.

A randomized trial of a feasible smoking cessation intervention may help unpack some of these questions. Higgins and colleagues (2010) reported on the combined outcomes of three pertinent controlled trials. One hundred sixty-six pregnant patients were randomly assigned to either a contingency-management intervention arm (wherein patients earned vouchers exchangeable for retail items by abstaining from smoking) or a control arm (wherein patients earned vouchers independent of smoking status).

This RCT is well-suited to examine the specific effect of reduced smoking, because participants provided urine samples that allowed biometric verification of seven-day smoking status through cotinine tests. On this metric, researchers reported final-trimester smoking abstinence to be markedly higher in the intervention vs. the control group (34.1% vs. 7.1%). Compared with other smoking cessation interventions, this contingency-management intervention proved quite powerful to reduce smoking in this patient population. Such a trial may influence public policy through several routes, including state regulatory policies requiring insurance coverage for smoking cessation interventions and state requirements imposed on cigarette manufacturers regarding package warning labels.

	Contingency management group (n=85)	Non-contingent group (N=81)
Percent low birthweight	5.9%	18.5%
Percent preterm birth	5.9%	13.6%
Percent NICU admissions	4.7%	13.8%

Final trimester smoking abstinence	34.1%	7.1%

Table 13.2. Infant birth outcomes (drawn from Higgs, et al. 2010, Table 2).

As previously, let Z represent a binary indicator for group assignment, whereby Z=1 represent assignment to the contingency management intervention, and Z=0 represents assignment to controls. The incidence of low birthweight was significantly lower in the intervention vs. controls (5.9% vs. 18.5%). The *intent-to-treat* (ITT) effect of the contingency-management intervention on low birth weight was thus

$$\beta_{ITT} = E[LBW|Z = 1] - E[LBW|Z = 0] = 5.9\% - 18.5\% = -12.6\%. \quad (3)$$

That's arguably the most important bottom-line measure of the public health impact of this intervention. If we offered this same intervention to 1,000 pregnant smokers similar to the patients offered this intervention, we can anticipate that we would prevent 126 low-birthweight deliveries. Given the modest costs of contingency management interventions and the immediate benefit to both pregnant patients and to infants, prenatal smoking cessation ranks among the most cost-effective interventions in clinical care, with estimated costs of approximately $3,000 per pregnant patient who actually quits smoking due to the intervention (Mundt et al., 2021). Such findings provide a strong policy argument for states to require insurers to cover this service for pregnant smokers, and thus is a good example of *policy candidate* randomized trials.

Because this intervention specifically focused on cotinine-verified smoking cessation, it is plausible to believe that the reduction in low-birthweight incidence was entirely due to reduced smoking. However, we should also be open to the possibility that the intervention influenced birthweight through other channels. For example, the intervention may strengthen the therapeutic alliance between patients and providers. If so, this may lead patients to seek medical help more aggressively for other health concerns, or to reduce other risk behaviors. If so, the ITT would still provide an unbiased estimate of the intervention's treatment benefit. The policy arguments to cover this service would not be undermined or weakened. Yet our causal interpretation of the findings would be misplaced.

This intervention also shows the potential weaknesses of available strategies to estimate TOT effects, even within an excellent trial. In this case, the TOT corresponds to the estimated effect of smoking cessation on the probability of a low-birthweight outcome. If we posit that last-trimester smoking is the key behavioral parameter, we would be tempted to apply the standard Wald estimator to the available published data, to take into account actual smoking cessation:

$$\beta_{Wald} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\overline{Dose}_{Z=1} - \overline{Dose}_{Z=0}} = \frac{0.185 - 0.059}{0.341 - 0.071} = \frac{0.126}{0.27} = 0.467 \quad (4)$$

Taken at face value, these estimates would imply that for every 1,000 pregnant smokers who actually participate in this intervention, we prevent 467 low birthweight deliveries. Note, however, that this interpretation lacks face validity. After all, low birthweight prevalence within the *control group* of pregnant smokers is only 185 per 1,000. One possible explanation would be that prenatal smoking cessation interventions induce valuable reductions in smoking intensity among the 65.9% of the treatment group who *do not* achieve third-trimester complete abstinence from smoking. A second possibility is that smoking cessation interventions solidify the treatment group's connection to prenatal care, thus allowing clinicians to address other health challenges associated with greater risk of low birthweight.

Trials such as this could also face threats to generalizability and external validity. The intervention might be performed at an excellent medical center. The particular approach may be designed for rural non-Hispanic whites and may be less culturally-competent, and thus less effective (due to reduced take-up or reduced patient engagement) when offered to other patients. Policymakers would want to complement this study with other study designs, perhaps including a true policy experiment that includes random assignment when a policy is first implemented, to understand how best to design legal requirements for these services in the entire population.

Randomization At Different Units of Observation

Many randomized trials seek to evaluate specific interventions provided for individual students, patients or service recipients. These include familiar randomized trials in clinical care. The posited causal pathway resides in changing medical care, social services, or other treatments provided to individual participants, with corresponding changes in behaviors and circumstances at the individual level. Evidence from such trials can thus inform legal and regulatory policies that require or facilitate delivery of services and interventions found to be effective.

Even when a trial randomizes individual patients, contextual and institutional factors play fundamental roles, but are often not analyzed in the study. An intervention that proves attractive and accessible in one population may prove less so in another. As discussed below, the burgeoning field of implementation science has brought systematic attention and rigor to the effects of implementation facilitators and barriers as policymakers seek to generalize findings from a trial that serves a specific study population in a particular setting to other different populations in different settings (Eccles & Mittman, 2006; Hirschhorn, Smith, Frisch, & Binagwaho, 2020; Hoagwood, Purtle, Spandorfer, Peth-Pierce, & Horwitz, 2020; Proctor et al., 2011; Shelton, Cooper, & Stirman, 2018).

CLUSTER RANDOMIZATION

Randomized trials to improve population outcomes are often most useful when applied at higher units of aggregation. A recent cluster-randomized trial by Abaluck and colleagues (2021) exemplifies the importance and global reach of such methods. Within a year of the first documented

COVID-19 case, these authors conceived and executed a cluster-randomized trial of measures to increase mask wearing within 600 villages in rural Bangladesh, involving more than 342,000 adult participants. The authors demonstrated that public health measures could increase the prevalence of proper mask wearing from 13.3% in the control group to 42.3% in the intervention group. The authors demonstrated accompanying reductions in symptomatic COVID-19 prevalence, particularly among older adults in villages where surgical masks were distributed.

In similar fashion, Victor and colleagues (2011) performed a cluster-randomized trial of Black barbershops as intervention sites for pharmacy interventions to reduce systolic blood pressure among Black men with high systolic blood pressure (exceeding 140 mm Hg). Barber shops within the treatment group promoted follow-up with a specialty-trained pharmacist. Pharmacists met regularly with customers who frequented barber shops in the treatment group, prescribed anti-hypertensive medications, and sent progress notes to customers' primary care providers. Barber shops within the control encouraged patrons with high blood pressure to pursue healthier lifestyles and to attend medical appointments. Black male patrons with high systolic blood pressure in both treatment and control arms experienced improved systolic blood pressure relative to baseline (six-month reductions of 27.0 mm and 9.3 mm, respectively). Patrons who frequented barber shops in the treatment group were substantially more likely to reduce their systolic blood pressure (63.6% of participants below 130 mm Hg, compared with 11.7% among participants in the control.) These group differences appear to reflect greater use of anti-hypertensive prescription medication among treatment-group participants. Benefits provided to the control group are also noteworthy. These underscore the public health benefit of even modest measures (within the control group) by trusted community members. These findings also provide a useful reminder that RCTs can be properly designed to provide important health benefits to all participants.

Saltz and colleagues (2021) provide another valuable example, examining community-level interventions on alcohol-related crashes. This study examined effects of a bundle of actually-deployed public policy enforcement practices on important population health outcomes. Through a cluster-randomized trial of 24 small cities in California, these authors found that enforcement measures to reduce underage drinking, drunk driving, and other behaviors associated with harmful alcohol use induced a 17% reduction in single-vehicle nighttime crashes among drivers 15–30 years of age.

STEPPED-WEDGE DESIGNS FOR POLICY EXPERIMENTS

Using cluster randomization in stepped-wedge designs as a way to evaluate health effects of actually implemented laws provides strong causal inferences. A stepped-wedge design is a cluster-randomized trial in which the clusters are randomly divided into groups (Copas et al., 2015). All clusters initially receive status quo/usual care intervention for a given roll-out period. Then each group receives the evaluated intervention (crossing over from status quo/control to treatment) at a group-specific later time.

A very large literature now establishes the effectiveness of state laws requiring automobile safety belt use. However, decades ago when such laws were first being enacted, legislators might have designed the rollout of such laws to expeditiously provide a rigorous policy evaluation using a stepped-wedge design. For example, Illinois policymakers might randomly assign each of the state's 102 counties to one of three cluster-groups. Residents of Group 1 counties would be required to use safety belts six months after the legislation is passed. Residents of Group 2 counties would be required to do so one year later. Finally, residents of Group 3 counties would be required to do so 18 months after the legislation passed.

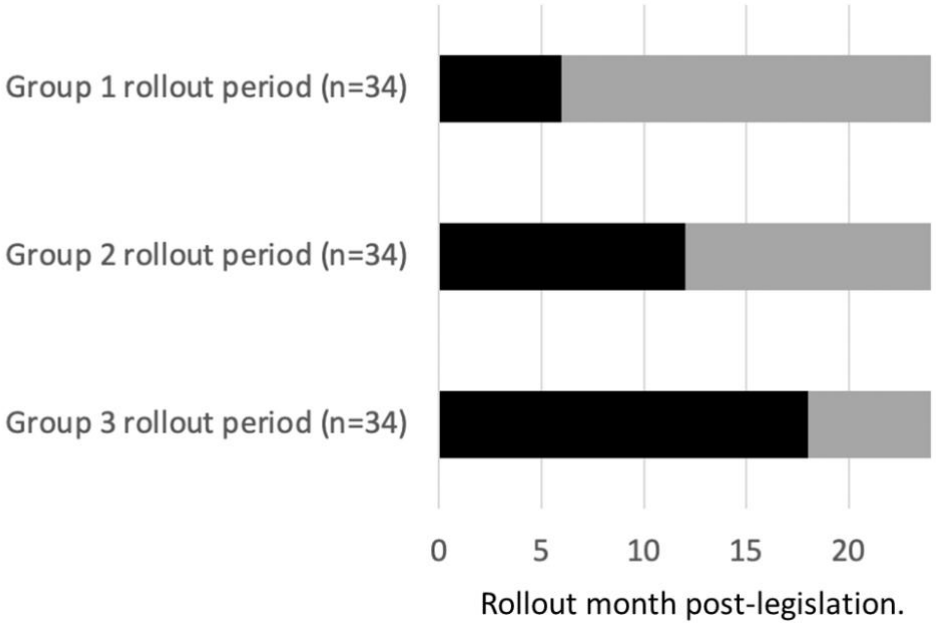

Figure 13.1. Stepped-Wedge Implementation Across 102 Illinois Counties.

Because counties are randomly assigned to the three groups, researchers can verify similar baseline prevalence and pre-intervention trends in road injuries and fatalities. They can then examine whether one can observe a break in these trends that match the new law's implementation. Researchers might explore mechanisms, for example by examining the rate of safety-belt-related road citations independent of crashes, and by collecting data on belt usage among individuals involved in crashes. Researchers might also explore contextual facilitators and barriers, for example by comparing effects across urban and rural counties, or by comparing across counties with high and low levels of safety belt law enforcement.

In like fashion, a county might employ a stepped-wedge design to evaluate clean indoor air smoking regulations. Here local officials might randomly divide workplaces and restaurants into multiple groups subject to focused enforcement in a staged roll-out, examining facility-specific

trends in respiratory complaints, worker sick days, and other outcomes in relation to the implementation of such regulations.

Notice that random assignment of jurisdictions addresses some common threats to the validity of quasi-experimental designs (see Chapter 14). Quasi-experimental evaluations of state policy efforts such as Medicaid expansion face the obvious challenge of policy endogeneity – the law is in part a result of other factors also affecting health outcomes. States with stronger commitment to public health may have been first to embrace such policies. If so, quasi-experimental approaches may overstate the public health benefits of early expansion because early-adopter states already had more favorable public health trends. Alternatively, early-adopter states may have included those facing new and alarming public health challenges such as the opioid epidemic. If so, quasi-experimental approaches may understate the public health benefits of early expansion, because early adopters turned to Medicaid expansion as a tool to address worsening trends. Given sufficiently large numbers of jurisdictions, stepped-wedge cluster-randomized designs help to address both challenges.

Notice in our safety belt example that roadside injuries and deaths are immediately proximate to the behaviors that one seeks to deter, and are readily observed. A stepped-wedge design is thus well-suited to evaluate such legal interventions. Much the same might be said regarding laws to deter driving under the influence of intoxicating substances, policies to facilitate naloxone distribution or jail-based opioid disorder treatment to reduce fatal overdose, and mandatory COVID-19 vaccination policies, all having a proximate effect on related hospitalizations (Bajema et al., 2021; Tenforde et al., 2021). A stepped-wedge design might also be effective in evaluating a law mandating insurance coverage for the contingency-management intervention discussed above to reduce prenatal smoking. State policymakers might require insurers to cover such services on a staggered basis across counties or other defined groups.

Stepped-wedge designs are more challenged to evaluate policies whose mechanisms, outcomes, and effects follow a more complex dynamic process. Suppose, for example, that one seeks to evaluate laws that restrict or discourage more general tobacco use. Such policies may take years to appreciably change chronic smoking behaviors. Moreover, many of the health effects one seeks to influence – for example asthma hospitalizations or smoking-related cardiovascular events – themselves unfold over years and are harder to directly link with the evaluated policies.

The internal validity of stepped-wedge designs is threatened when there are important spillover effects. Consider a county-level stepped-wedge design of interventions designed to reduce smoking through local excise tax increases. The ease with which smokers can purchase across county lines or underground-market sellers can smuggle cigarettes across county lines could undermine the research design.

MEDIATING MECHANISMS AND CAUSAL PATHWAYS

Both randomized trials and quasi-experimental studies are identified based on effects of interventions on "treatment compliers," whose circumstances or behaviors change as a result of the intervention. In the case of smoking cessation, the posited mechanisms are plausible and direct. In other cases, the pathways are more complex or diffuse.

Specifying such pathways is increasingly important to stakeholders and policymakers seeking to understand and apply scientific results from experimental trials, especially when generalizing from a trial on a specific problem in a specific situation to other opportunities for legal or regulatory intervention. The emerging discipline of implementation science speaks directly to these concerns.

Suppose that we conduct an RCT evaluating effects of small-group cognitive-behavioral therapies to reduce youth violence (Heller et al., 2017). One might reduce youth offending by imparting strategies young men can deploy to deescalate potentially violent confrontations. If this is the key mechanism, the intervention might generalize to broad health benefit. Properly implemented interventions based on this RCT would need to insure proper manualization based on the initial intervention, and proper training of staff to demonstrate fidelity to this initial model. Alternatively, the key mechanism may be less curriculum-dependent, and more dependent on young men building a strong therapeutic alliance with pro-social adults who operate the groups. If this is the key causal pathway, choosing staff with demonstrated ability to build strong relationships with young men may be the key variable and the appropriate focus of program implementation. Once again, the discipline of implementation science provides valuable attention to scale and sustainment in drawing proper policy insights from specific trials.

The National Institutes of Health have imposed increasingly stringent requirements on randomized trials, most obviously in requirements to pre-register trial outcomes and thereby reduce the risks of misleading incidental findings. NIH now also requires investigators to specifically identify and investigate causal mechanisms and pathways. Indeed, beyond exploratory and pilot studies, most NIH institutes will no longer fund or even review fully scaled "black-box" RCTs. As one institute explains:

> NIMH requires an experimental therapeutics approach … for the development and testing of therapeutic, preventive, and services interventions, in which the studies evaluate not only the clinical effect of the intervention, but also generate information about the mechanisms underlying a disorder or an intervention response. Studies … must clearly identify a target or mediator of the intervention being tested. A positive result will require that an intervention improves clinical outcomes and has a demonstrable effect on a target, such as a neural pathway, a key cognitive operation, interpersonal or contextual factor that is hypothesized to mediate the intervention's effect (Insel & Gogtay, 2014).

Such requirements have forced researchers to incorporate explicit logic models or explicit theoretical/conceptual frameworks for experimental and quasi-experimental work, in many cases bringing valuable rigor that is sometimes lacking in "black-box" randomized trials. This work is

especially important when policymakers wish to draw upon the same mechanisms identified in an RCT, but in support of a different legal or regulatory intervention. For example, an intervention trial might identify additional family resources as a key pathway to reduce criminal offending among affected youth. Such findings could assist policymakers to target child tax credits or other supports to families as a pertinent policy intervention. Another intervention RCT might identify reduced parental alcohol use as a key pathway to improved child outcomes. Such findings could inform state policy debates over alcohol regulatory policies.

This focus on mechanism has attracted criticism, as well. Researchers may identify one plausible clinical target that moves as a result of an intervention, but it may not in fact mediate the intervention's full causal effect. Some black-box RCTs have great clinical and policy value, even when researchers have poor understanding of multiple accompanying causal pathways.

In like fashion, a policy experiment to expand access to early-childhood education might improve developmental outcomes through multiple pathways that are hard to disentangle through a single study design. Identifying that a feasible public policy actually helps children might be a more important, immediately-actionable finding than the findings from a scientifically rigorous study which identified one precise causal mechanism, disconnected from a feasible public policy intervention.

Conversely, a fully informative or generalizable RCT is not always feasible. An RCT may be unethical, too costly, or too complex. For example, the RAND Health Insurance Experiment, the most famous randomized trial in the history of health economics, cost approximately $300 million in 2021 dollars, and took roughly a decade to complete (Frakt, 2010). With occasional exceptions, public health researchers and practitioners rarely have access to such resources for rigorous experimental analysis of public health challenges. Even if an RCT is technically feasible, explicit randomization of a valuable intervention may be socially or politically infeasible. No one was going to randomly assigning millions of seniors to Medicare or to a placebo intervention. Fortunately, randomized trials are not the only path to rigorous policy evaluation. In fact, a much more useful approach is to recognize that randomization is an excellent *design element* to incorporate whenever possible, but is only one of a dozen or more useful design elements enhancing the validity of causal inference. (See Chapter 14 for more on these issues.)

Critiques of Randomized Trials

The rising prominence of randomized control trials and experimental methods has sparked important critiques. Economist Angus Deaton laments that RCTs are wrongly regarded as the top of a hierarchy of research evidence, thus denigrating the importance of other empirical approaches that allow more explicit study of structural factors, and approaches that may bring greater external validity than is possible through an RCT. Deaton (2020) notes that

> RCTs are affected by the same problems of inference and estimation that economists have faced using other methods, as well as by some that are peculiarly their own … RCTs have no special status, they have no exemption from the problems of inference that econometricians have always wrestled with, and there is nothing that they, and only they, can accomplish …

Deaton and others note several ways RCTs are misinterpreted or overinterpreted: Evidence from one trial conducted in one population is uncritically assumed to apply to very different populations, even when there are strong reasons to expect heterogeneous treatment effects. Deaton notes that data acquired for randomized trials often include "gross outliers," missing data, and data quality challenges that can profoundly influence estimated program effects. Deaton suggests that one very costly pregnancy played an important role in findings from the RAND health insurance experiment (Manning et al., 1987). Such challenges are especially important in violence prevention research, when a single homicide can dominate economic valuations of program effects.

In related fashion, Mosley and colleagues (2019) express concern that an RCT-based "what works" analytic framework disvalues community-based knowledge, and often fails to develop proper understanding of organizational and community context essential to successful, sustainable interventions that earn legitimacy among service providers, affected communities, and other stakeholders. The rise of implementation science represents one effort to bridge these perspectives (Shelton et al., 2018).

Deaton (2020) also notes an important addition and caveat to critiques of RCTs: "Just as none of the strengths of RCTs are possessed by RCTs alone none of their weaknesses are theirs alone." Many pitfalls commonly associated with RCTs also arise in quasi-experimental studies and other designs. Given heterogeneous treatment effects, the results of a "natural experiment" study design cannot simply be generalized to policy innovations that serve very different populations. More subtle challenges also arise. Suppose one wishes to study the mortality effects of the Affordable Care Act's Medicaid expansion. There is some evidence that states which most eagerly and promptly embraced Medicaid expansion experienced more-favorable prior mortality trends in the years before Medicaid expansion took effect. If so, early quasi-experimental analyses may overstate Medicaid expansion's mortality effects (Kaestner, 2012; Miller, Johnson, & Wherry, 2021).

Heckman and Smith (1995) presented in succinct form many insights now widely presented by scholars worrying about the reification of randomized trials. These authors note the importance of "substitution bias," whereby members of an experimental control group find ways to obtain close substitutes for the studied treatment, and of "randomization bias," whereby individuals willing and able to voluntarily participate in an RCT differ from people likely to participate in a non-experimental intervention based upon the randomized trial. These authors cite one pediatric drug study (Kramer & Shapiro, 1984) that included an experimental and non-experimental component. The non-experimental component experienced a 4% refusal rate, compared with a 34% refusal rate within the randomized sub-trial.

Heckman and Smith (1995) note the widespread belief among social service providers that randomized evaluations are unethical, particularly when an RCT results in denial of services to control group subjects. Such critiques echo criticisms within the public health community, e.g. whether it is ethical to employ site-randomized trials to examine the efficacy of syringe support services (Leary, 1996). At minimum, such debates underscore the importance of ensuring that both intervention trials and policy experiments are conducted with cultural competence and legitimacy with all stakeholders, particularly when such experiments serve vulnerable or stigmatized populations.

STATISTICAL COMPLEXITIES IN RANDOMIZED TRIALS

Researchers and policymakers often hope that an RCT will allow simpler, and thus less-fragile or less-contestable statistical analysis. Such hopes are often dashed.

Differential attrition poses one challenge. Control group participants have less frequent contact and have less of a personal relationship with individuals operating a given study. They may be correspondingly more likely to be lost to follow-up. Such loss to follow-up may be ameliorated if researchers have access to administrative data such as birth and death certificates, arrest records, academic attendance, or grade data. The challenges are more significant if one relies solely on self-reported information or other subject data that requires continued study participation.

Differential attrition can bias the analysis of RCT data, overstating or understating benefits of an intervention. Suppose, for example, one conducts an RCT of a school-based intervention in which members of the treatment group receive valuable counseling to address substance use. The most vulnerable youth in the control group may change schools or leave the school district in an effort to obtain comparable services. Such a pattern might lead researchers to understate the benefits of the studied intervention. Conversely, treatment-group youth who do not benefit from the intervention may leave the school or drop out of the intervention. If researchers only examine data from the remaining students in the treatment group, they might easily overstate the benefits of the studied intervention.

Differences in data quality between treatment and control can create other biases. Individuals exposed to intensive safe-sex or crime prevention counseling may feel greater rapport to candidly describe their sexual practices or criminal behavior than would corresponding individuals not exposed to such messages. Alternatively, treatment group members may be reluctant to reveal information about risk-behaviors they know will disappoint researchers and program implementors. Biases can arise in more subtle ways, as well. For example, the research team may have greater opportunities for data cleaning or to resolve data discrepancies regarding members of the treatment group – particularly if such information is used for operational purposes.

NORMS, LONG-TERM INSTITUTIONAL CHANGES, AND EQUILIBRIUM EFFECTS

Randomized trials are implemented in particular institutional and social contexts, serving particular populations with a particular set of implementation resources. Understanding these contexts, populations, and resources is essential to properly interpret trial results and to understand potential challenges as one seeks to bring a particular intervention "to scale." Independent of any question of internal validity, randomized trials are often poorly suited to study effects of institutional changes, changes in social norms, and general equilibrium effects that might be induced if specific interventions were implemented on a broader scale. One can design an RCT to examine the effects of naloxone distribution and syringe support services within a particular population of people who inject drugs. It is harder to design an RCT to understand what drug-use practices would look like if states removed all legal barriers to naloxone distribution and syringe support services. Nor can an RCT explore changes in drug-use practices and peer norms if *all people who inject drugs* had access to these resources, and the ecology of drug use organically included these opportunities.

Burtless provides one provocative example in the context of targeted wage subsidies for disadvantaged workers (Burtless, 1985). In this trial, job seekers were given vouchers identifying them to prospective employers as eligible for a generous wage subsidy. This intervention trial was intended to investigate effects of an enhanced Earned Income Tax Credit (EITC) and related policies.

Contrary to prior hypotheses, workers provided with vouchers were significantly *less* likely to find employment than were job seekers who did not have the vouchers. This experiment was valuable in identifying ways that targeted subsidies can bring unintended and unforeseen harms. Such vouchers apparently carried a stigmatizing effect, leading employers to discriminate against voucher holders. Of course, an intervention that directly affects a small subset of the population is qualitatively different from a universal intervention affecting an entire population. Thus, such an experiment with a small subgroup cannot test the potential effects of broader policies such as an expanded EITC universally applied. A change in law would subsidize an entire population of low-income workers, avoiding potential stigmatizing effects that frequently arise for interventions serving specific small groups of disadvantaged individuals.

At an institutional level, randomized trials cannot fully examine effects of altered organizational practices induced by changes in public policy. This is an important point of dialogue between randomized trials and quasi-experimental designs. A randomized trial of a home visiting intervention, for example, supported by a two-year private foundation grant, sheds light on the immediate responses and outcomes for specific program beneficiaries within existing structures. This is important information as one seeks to study policy candidates for broader evaluation. Such an RCT can't create or test the kind of structural change that permanent Medicaid reimbursement for this service and other legal-policy changes might facilitate, and thus cannot rigorously address the full range of beneficial or deleterious consequences that might accompany a change in law.

In like fashion, suppose researchers conducted a randomized trial in 1980 – just prior to greatly expanded Medicaid benefits for prenatal, labor and delivery, and perinatal medical services – in which a treatment group of 1,000 low-income pregnant patients were provided free prenatal care and other medical services associated with pregnancy and neonatal care. One might have observed greater care utilization within this intervention group. One *would not* have been able to observe the effects of care delivered through the expanded network of prenatal and neonatal intensive care facilities implemented by providers in response to the permanent expansion of Medicaid to millions of pregnant women – an important set of institutional responses, induced by these policy changes, now regarded as critical to declining infant mortality (Chung et al., 2010; Currie & Gruber, 1997; Lorch, Rogowski, Profit, & Phibbs, 2021; Phibbs et al., 2007; Phibbs & Lorch, 2018). Indeed, such an individually-randomized RCT would have provided a misleading – or at-best incomplete – guide for policymakers seeking to understand the likely impact of such policies. A quasi-experimental study design would have provided more rigorous assessment of true policy effects (see Chapter 14).

Scaling interventions may alter economic and institutional contexts in other ways difficult to capture within a single RCT. One important challenge arises when bringing an intervention to scale stresses scarce resources, or in other ways alters pertinent wages and prices. A randomized trial may find that assigning precariously-housed individuals to case workers with one-half the normal caseload reduces homelessness. As such interventions are brought to scale, implementors need to hire less-proficient and less-experienced workers to handle the larger caseload, and may thus achieve less benefit (Jepsen & Rivkin, 2009). Interventions may also prove unexpectedly costly if service providers are required to raise wages to achieve the required staffing, which of course would alter trial-related assessments of cost-effectiveness.

Mechanism Experiments

Mechanism experiments offer one set of tools to address the above challenges (Ludwig, Kling, & Mullainathan, 2011). One might design an RCT with the aim of exploring a specific causal pathway rather than to directly scrutinize a feasible policy as a whole. Such intervention designs might be avowedly ill-suited to expand "at scale" to serve large populations. Their purpose is different: to scrutinize one particular causal mechanism or pathway that might facilitate development of other, more feasible public health policies. For example, many citizens and policymakers believe that the lack of local grocery stores offering affordable, nutritious food promotes obesity in Chicago's far south side and other "food deserts." One might test the hypothesis that food deserts play this important causal role by conducting an RCT in which treatment subjects have subsidized access to a service that makes possible the delivery of fruits and vegetables and other nutritious food items (Ludwig et al., 2011). One can then compare body mass index and other related outcomes between treatment and control subjects. If this rather extreme policy candidate fails to improve health outcomes, policymakers can reasonably conclude that policies that subsidize grocery stores that operate in food-desert areas, or regulations imposed on grocery chains requiring them to open

stores in current food deserts might bring other important community benefits, but this "mechanism" RCT indicates such policies are unlikely to improve these specific health outcomes tested.

One might hypothesize that financial factors and legal penalties influence individuals' willingness and ability to obtain COVID-19 vaccinations (Campos-Mercade et al., 2021). One might design an RCT in which study subjects are offered $500 if they agree to receive a vaccine. As an aside, this trial also underscores some significant challenges to mechanism-experimental designs in sensitive arenas. The $500 payment raises significant ethical concerns. Moreover, the large incentive may prove counterproductive, if it heightens concerns among study participants or others that any medical treatment requiring such payment must bring large accompanying risks. A second RCT could test effects of intensive enforcement of vaccine mandates where those unvaccinated are suspended without pay to evaluate effects of such disincentives on vaccine receipt. In short, assessing the role of financial factors and legal penalties in vaccination rates might involve a series of focused RCTs on particular components of the larger question.

Mechanism experiments can also be performed at the organizational level. Suppose one believes that outcomes among homeless individuals with substance use disorders are worsened because caseworkers, overwhelmed with large caseloads, cannot provide the individual attention required to effectively serve their most vulnerable clients. One might implement policies requiring social service agencies to maintain lower client-caseworker ratios to receive Medicaid reimbursement. As noted (and dismissed) above, one might also design a mechanism experiment in which 500 homeless individuals are assigned to caseworkers with half of the standard caseload, and 500 homeless individuals are assigned to a usual-care arm in which caseworkers have standard caseloads. This RCT might be operationally unrealistic or unscalable itself, but nevertheless may still provide valuable insights into the potential benefits of such regulatory policies.

The Burgeoning Role of Implementation Science

A traditional view of policy randomized trials draws heavily on the "pipeline" paradigm of randomized trials of new drugs and procedures in medical care. That is, one first establishes the efficacy of a new intervention under ideal (or at least accommodating) conditions. One then brings the intervention "to scale" in a broader population (Landes, McBain, & Curran, 2019). The burgeoning field of implementation science draws attention to the distinctive gaps between research and practice through the use of scientific methods to promote the uptake, implementation and sustainability of evidence-based practices, programs, and policies, with an ultimate goal of improving population health (Eccles & Mittman, 2006; Hirschhorn et al., 2020; Hoagwood et al., 2020). While a full discussion of the discipline is outside the scope of this chapter, we briefly review how greater engagement with implementation science approaches may improve the adoption, implementation, and sustainment of interventions within the field of legal epidemiology.

HYBRID EFFECTIVENESS-IMPLEMENTATION STUDY DESIGNS

Whereas traditional RCTs focus on intervention effectiveness, implementation science has developed hybrid trial designs that focus on both implementation and effectiveness (Curran, Bauer, Mittman, Pyne, & Stetler, 2012; Johnson et al., 2020; Landes et al., 2019). Such trials are increasingly influential within public health, with Curran and colleagues delineating three categories of hybrid study designs:

- *Hybrid type 1* designs primarily focus on testing clinical interventions and outcomes, but pay explicit attention to implementation processes, contextual barriers and facilitators, and needed program adaptations to local context (Landes et al., 2019; Pho et al., 2021). As Landes and colleagues describe, a hybrid type 1 trial often resembles a conventional RCT paired with a complementary process evaluation.
- *Hybrid type 2* designs give equal weight to clinical intervention and implementation related factors, including explicit measures of implementation outcomes. For example, one may have good reason to believe that smoking cessation counseling reduces smoking and improves health outcomes among patients in a diabetes clinic. A hybrid type 2 trial would scrutinize patient-level smoking status and health outcomes but would give equal weight in studying implementation strategies to understand when and how smoking cessation counseling is actually provided, and how staff might be supported in reliably executing such efforts.
- *Hybrid type 3* designs primarily focus on implementation itself and are secondarily focused on patient outcomes. For example, a state prison system might conduct a site-randomized trial in which staff were provided training materials that encourage and facilitate naloxone distribution for returning citizens who leave these facilities. Researchers might seek to track overdose reversals and related outcomes, but the main focus would be on implementation processes that influence whether naloxone is provided (Landes et al., 2019).

Kemp and colleagues (2019) recently expanded on this original typology, with an additional 12 hybrid designs. Although not yet widely used, they offer helpful additional perspectives for evaluating the relative weight of the intervention, implementation context, and the implementation strategies used to support implementation.

IMPLEMENTATION SCIENCE THEORIES, STRATEGIES AND OUTCOMES

Implementation science theories. Core to implementation science research is the use of different theories, models and frameworks (TMFs) to guide, understand and evaluate the implementation of a new program, practice or policy. In a seminal review, Nilsen (2015) identified five categories of TMFs: (1) process models: describing how to translate research to practice, (2) determinant frameworks: understanding the contextual barriers and facilitators that shape implementation outcomes, (3) classic theories: from disciplines outside implementation science, e.g., psychology, sociology, or organizational theory, used to evaluate different domains of implementation, (4)

implementation theories: developed by implementation scientists to understand or explain different domains of implementation; and (5) evaluation frameworks: specifying the relevant measures and metrics to assess implementation outcomes. Designed to be used before, during and after implementation, TMFs can be assessed with both qualitative (Hamilton & Finley, 2019) and quantitative methods (Smith & Hasan, 2020), and are essential for understanding the implementation context, selecting implementation strategies, and examining implementation outcomes.

Although Fulmer and colleagues (2020) argue that process models can strengthen the application of legal epidemiology in public health research, implementation science TMFs have not been widely used in legal epidemiology research. This does not mean that legal epidemiology scholars are not exploring knowledge translation processes or the contextual factors that shape the implementation of public health laws; rather, the discipline does not appear to have widely engaged with extant TMFs. One exception has been the use of Rogers' Diffusion of Innovations Theory (2010) – a classic TMF – and related concepts on the diffusion of innovations in the policy surveillance realm (Bae, Anderson, Silver, & Macinko, 2014; Burris, Hitchcock, Ibrahim, Penn, & Ramanathan, 2016; Komro et al., 2020; Politis, Halligan, Keen, & Kerner, 2014). More explicit engagement with TMFs in legal epidemiology will not only strengthen the field's ability to document the multiple factors that shape the implementation of public health laws, but also can strengthen the role of implementation science in improving population health (Damschroder et al., 2009; Feldstein & Glasgow, 2008; Nilsen, 2015).

Implementation strategies are the specific activities used to support the implementation, enforcement or delivery of an evidence-based intervention. Recent scholarship has provided greater conceptual clarity on the definition (Powell et al., 2015), selection (Powell et al., 2017) and reporting (Proctor, Powell, & McMillen, 2013) of such strategies. Because most of this scholarship has been developed within the context of health services research, additional work is necessary to fully specify implementation strategies for legal interventions.

Implementation scientists are also actively engaged in understanding *implementation pathways of influence* – the mechanisms of action through which an implementation strategy operates to achieve desired implementation outcomes (Boyd, Powell, Endicott, & Lewis, 2018; Lewis, Klasnja, et al., 2018; Lewis, Scott, & Marriott, 2018). Relatively little is known about the actual pathways through which implementation strategies affect change on implementation actors and organizational actors, especially within the context of legal interventions. Dual work on identifying and testing implementation strategies in the context of legal interventions would be welcome.

Finally, implementation outcomes are the "effects of deliberate and purposive actions to implement new practices, programs, and policies" (Proctor et al., 2011, p. 65). Two evaluation frameworks commonly used in implementation science are the Multilevel Implementation Outcomes Framework (Proctor et al., 2011) and the RE-AIM Framework (Glasgow, Vogt, & Boles,

1999). Although both frameworks are widely used within the field of implementation science, they have had limited application to date within legal epidemiology.

Table 13.3 lists the outcomes from each framework, their definitions and an example of how to operationalize each one in the context of a legal policy intervention on over-the-counter access to naloxone. For each framework, we also include a focus on equity, which has drawn significant attention in recent implementation science research (Baumann & Cabassa, 2020; Brownson, Kumanyika, Kreuter, & Haire-Joshu, 2021; Shelton, Adsul, & Oh, 2021; Snell-Rood et al., 2021).

Implementation Outcomes Framework (E. Proctor et al., 2011)		
Outcomes	**Definition**	**Research Questions**
Acceptability	"perception among implementation stakeholders that a given treatment, service, practice, or innovation is agreeable, palatable, or satisfactory" (p. 67)	How do key stakeholders view a state law permitting over the counter (OTC) access to naloxone? Do perceptions of acceptability differ among stakeholders who bear a disproportionately high burden of overdose deaths?
Adoption	"intention, initial decision, or action to try or employ an innovation or evidence-based practice" (p. 69)	To what extent do owners of local businesses, e.g., pharmacies, drug stores, community-based organizations (CBOs), other permitted settings, agree to provide OTC access to naloxone?
Appropriateness	"perceived fit, relevance, or compatibility of the innovation or evidence based practice for a given practice setting, provider, or consumer; and/or perceived fit of the innovation to address a particular issue or problem" (p. 69)	To what extent do key stakeholders perceive that a state law permitting OTC access to naloxone is appropriate for their community and will help to reduce overdose deaths? Do perceptions of appropriateness differ among stakeholders who bear a disproportionately high burden of overdose deaths?
Feasibility	"extent to which a new treatment, or an innovation, can be successfully used or carried out within a given agency or setting" (p. 69)	To what extent is legislation permitting OTC access to naloxone feasible within different settings (e.g., pharmacies, drug stores, CBOs, etc) in a state? To what extent is feasibility different in low and high resourced settings?
Fidelity	"degree to which an intervention was implemented as it was prescribed in the original protocol or as it was intended by the program developers" (p. 69)	To what extent are pharmacists or other permitted actors providing OTC naloxone as specified by the law?
Penetration	"integration of a practice within a service setting and its subsystems;" it also "can be calculated in terms of the number of providers who deliver a given service or treatment, divided by the total number of providers trained in or expected to deliver the service." (p. 70)	How many pharmacies in a given geographic area provide OTC access to naloxone?
Sustainability	"extent to which a newly implemented treatment is maintained or institutionalized within a service setting's ongoing, stable operations" (p. 70)	How many businesses continue to offer OTC access to naloxone within one year of the law coming into effect?

Equity	Extent to which access to the program is equitably distributed across the population. Extent to which the policy does not worsen outcomes, especially for marginalized populations.	How do we ensure that communities disproportionately impacted by overdose deaths have equitable OTC access to naloxone?
RE-AIM Framework (Glasgow et al., 2019; Glasgow, Vogt, & Boles, 1999; Shelton, Chambers, & Glasgow, 2020)		
Outcomes	**Definition**	**Research Questions**
Reach	At the individual level, the number, proportion, representativeness of individuals who participate in the intervention under study (Shelton et al., 2020)	How many people can access OTC naloxone? To what extent are the populations disproportionately impacted by overdose deaths able to access OTC naloxone?
Effectiveness	At the individual level, "the impact of an intervention on important health behaviors or outcomes, including quality of life (QOL) and unintended negative consequences; consider heterogeneity of effects" (Shelton et al., 2020, p. 4)	Does offering naloxone OTC reduce overdose deaths and affect other QOL outcomes? What are the unintended consequences of OTC naloxone? Which groups bear a higher burden of unintended negative consequences?
Adoption	At multiple levels, "the number, proportion, and representativeness of: (a) settings; and (b) staff/interventionists who deliver the program, including reasons for adoption or non-adoption across settings and interventionists." (Shelton et al., 2020, p. 4)	Two of four pharmacies in a rural county offer naloxone OTC. How do the adopting pharmacies/staff differ from those that do not offer naloxone? Did lower-resourced pharmacies adopt OTC naloxone to the same extent as higher-resourced pharmacies?
Implementation	"At multiple setting and staff levels, continued and consistent delivery of the EBI (and implementation strategies) as intended (fidelity), as well as adaptions made and costs of implementation." Shelton et al., 2020, p. 4)	To what extent did each adopting pharmacy implement OTC access to naloxone as described in the law? What adaptations did pharmacies make and why? Did all pharmacies have access to the resources necessary to successfully offer naloxone? What was the cost of offering naloxone OTC? What social-contextual factors shaped the implementation, including the social determinants of health?
Maintenance	"At the setting level, the extent to which a program or policy becomes institutionalized or part of the routine organizational practices and policies. At the individual level, maintenance has been defined as the long-term effects of a program on outcomes after a program is completed" (p .3, (Holtrop et al., 2021). Typically six months and one year after implementation, and an ongoing basis.	Which settings continued to offer OTC naloxone over time and why or why not? Which populations continue to be reached; do they continue to benefit or experience negative outcomes; why or why not? What factors shape sustainability low-and high-resource settings?

Conclusion

Randomized trials are more common and more influential than ever before in legal epidemiology and public health policy. Such research often produces more rigorous outcome evaluations than has ever been possible in public health, particularly when evaluated in terms of internal validity – the plausibility of inferring a causal effect. The growth of RCTs has produced other benefits as well. Not least of these benefits is an increase the proportion of social scientists who step out of the seminar and computer lab into the field, gaining tactile familiarity with public health challenges while performing intervention research (Blattman, 2016). Federal, state, and local policymakers are also more aware of the importance of strong research designs, and are increasingly willing to partner with researchers to perform experimental research.

Also critical have been more disciplined and systematic efforts to move beyond atheoretical "black-box" randomized trials. These include measures to rigorously establish specific causal mechanisms and pathways that underlay important public health interventions. Such innovations move the field beyond black box randomized trials that prove incapable of replication. More subtly, such innovations move the field beyond an arrogant "pipeline" model, in which researchers and practitioners regard policy innovation as the search for best-practice models that yield excellent results with strong internal validity with minimal attention paid to whether and how such interventions could truly be implemented, at-quality, on a broad scale, in diverse contexts and diverse populations. The rise of implementation science underscores new awareness of the importance of contextual factors and the need for well-designed experiments to better understand implementation processes and outcomes.

Newfound respect for experimental approaches brings new risks. One challenge arises from complex mutual dependencies between researchers and policymakers or organizational leaders, who control data and access to intervention sites. The realities of intervention research require long-term relationships. The spoken and unspoken exigencies of data use agreements within long-term relationships raise conflicting incentives for researchers and for policymakers alike. As researchers note the value of experimental methods, they must also guard against the loss of broader analytic and policy reflection than is typically engaged in experimental research. Experimental methods cannot directly explore changes in social norms, or large-scale institutional changes, such as broad long-term changes in medical practice associated with changes in Medicare policy. No experimental study could have documented the full health benefits of Medicaid expansion; nor could any randomized trial capture the downstream health harms associated with the Tuskegee experiment by reinforcing earned distrust of the American medical system among African-American men (Alsan & Wanamaker, 2018; Alsan, Wanamaker, & Hardeman, 2020). Nevertheless, randomized trials have significantly improved the quality of public health research.

No one methodological approach provides the gold standard, and RCTs are no exception. Many of the limitations of RCTs can be addressed by carefully designed quasi-experimental studies, to which we turn in the next chapter. As discussed there, quasi-experimental designs often face

surprisingly similar challenges, yet often have available tools to address questions that cannot currently be addressed through randomized trial designs.

Further Reading

Cunningham, S. (2021). *Causal Inference*. New Haven, Ct,: Yale University Press.

Kemp, C. G., Wagenaar, B. H., & Haroz, E. E. (2019). Expanding hybrid studies for implementation research: intervention, implementation strategy, and context. *Frontiers in Public Health*, *7*, 325.

Proctor, E. K., Powell, B. J., & McMillen, J. C. (2013). Implementation strategies: recommendations for specifying and reporting. *Implement Sci, 8*, 139. doi:10.1186/1748-5908-8-139.

References

Abaluck, J., Kwong, L. H., Styczynski, A., Haque, A., Kabir, M. A., Bates-Jefferys, E., . . . Mobarak, A. M. (2021). Impact of community masking on COVID-19: A cluster-randomized trial in Bangladesh. *Science*, *375*(6577), eabi9069.

Alsan, M., & Wanamaker, M. (2018). Tuskegee and the Health of Black Men. *Q J Econ, 133*(1), 407-455.

Alsan, M., Wanamaker, M., & Hardeman, R. R. (2020). The Tuskegee Study of Untreated Syphilis: A Case Study in Peripheral Trauma with Implications for Health Professionals. *J Gen Intern Med, 35*(1), 322-325.

Bae, J. Y., Anderson, E., Silver, D., & Macinko, J. (2014). Child Passenger Safety Laws in the United States, 1978–2010: Policy Diffusion in the Absence of Strong Federal Intervention. *Social Science & Medicine, 100*, 30-37.

Baicker, K., Allen, H. L., Wright, B. J., & Finkelstein, A. N. (2017). The Effect Of Medicaid On Medication Use Among Poor Adults: Evidence From Oregon. *Health Aff (Millwood), 36*(12), 2110-2114.

Baicker, K., Allen, H. L., Wright, B. J., Taubman, S. L., & Finkelstein, A. N. (2018). The Effect of Medicaid on Dental Care of Poor Adults: Evidence from the Oregon Health Insurance Experiment. *Health Serv Res, 53*(4), 2147-2164

Bajema, K. L., Dahl, R. M., Evener, S. L., Prill, M. M., Rodriguez-Barradas, M. C., Marconi, V. C., . . . Surie, D. (2021). Comparative Effectiveness and Antibody Responses to Moderna and Pfizer-BioNTech COVID-19 Vaccines among Hospitalized Veterans - Five Veterans Affairs Medical Centers, United States, February 1-September 30, 2021. *MMWR Morb Mortal Wkly Rep, 70*(49), 1700-1705.

Baumann, A. A., & Cabassa, L. J. (2020). Reframing implementation science to address inequities in healthcare delivery. *BMC Health Serv Res, 20*(1), 190.

Blattman, C. (2016). A lot of people think field experiments make scholars ask small questions, but I think they'll push us to answer the big ones. Retrieved March 27, 2022 from https://chrisblattman.com/2016/01/08/a-lot-of-people-think-field-experiments-make-scholars-ask-small-questions-but-i-think-theyll-push-us-to-answer-the-big-ones/

Boyd, M. R., Powell, B. J., Endicott, D., & Lewis, C. C. (2018). A Method for Tracking Implementation Strategies: An Exemplar Implementing Measurement-Based Care in Community Behavioral Health Clinics. *Behav Ther, 49*(4), 525-537.

Breza, E., Stanford, F. C., Alsan, M., Alsan, B., Banerjee, A., Chandrasekhar, A. G., . . . Duflo, E. (2021a). Doctors' and Nurses' Social Media Ads Reduced Holiday Travel and COVID-19 infections: A cluster randomized controlled trial in 13 States. *medRxiv*. doi:10.1101/2021.06.23.21259402

Breza, E., Stanford, F. C., Alsan, M., Alsan, B., Banerjee, A., Chandrasekhar, A. G., . . . Duflo, E. (2021b). Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: a cluster randomized controlled trial. *Nat Med, 27*(9), 1622-1628.

Brownson, R. C., Kumanyika, S. K., Kreuter, M. W., & Haire-Joshu, D. (2021). Implementation science should give higher priority to health equity. *Implementation Science, 16*(1), 28.

Burris, S., Hitchcock, L., Ibrahim, J. K., Penn, M., & Ramanathan, T. (2016). Policy surveillance: a vital public health practice comes of age. *Journal of Health Politics, Policy & Law, 41*(6), 1151-1167.

Burtless, G. (1985). Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment. . *Industrial and Labor Relations Review, 39*(1), 105–114.Buse, K., & Lee, K. (2005). *Business and global health governance*. London: London School of Hygiene & Tropical Medicine.

Campos-Mercade, P., Meier, A. N., Schneider, F. H., Meier, S., Pope, D., & Wengstrom, E. (2021). Monetary incentives increase COVID-19 vaccinations. *Science, 374*(6569), 879-882.

Chung, J. H., Phibbs, C. S., Boscardin, W. J., Kominski, G. F., Ortega, A. N., & Needleman, J. (2010). The effect of neonatal intensive care level and hospital volume on mortality of very low birth weight infants. *Med Care, 48*(7), 635-644.

Copas, A. J., Lewis, J. J., Thompson, J. A., Davey, C., Baio, G., & Hargreaves, J. R. (2015). Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials, 16*(1), 352.

Curran, G. M., Bauer, M., Mittman, B., Pyne, J. M., & Stetler, C. (2012). Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Med Care, 50*(3), 217-226.

Currie, J., & Gruber, J. (1997). *The technology of birth: health insurance, medical interventions, and infant health*. Cambridge, MA: National Bureau of Economic Research.

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci, 4*, 50.

Deaton, A. (2020). *Randomization in the tropics revisited: a theme and eleven variations.* Cambridge, MA: National Bureau of Economic Research.

Eccles, M. P., & Mittman, B. S. (2006). Welcome to implementation science *Implementation Science, 1*(1), 1-3.

Feldstein, A. C., & Glasgow, R. E. (2008). A practical, robust implementation and sustainability model (PRISM) for integrating research findings into practice. *Joint CommissionJournal on Quality and Patient Safety, 34*(4), 228–243.

Finkelstein, A. N., Taubman, S. L., Allen, H. L., Wright, B. J., & Baicker, K. (2016). Effect of Medicaid Coverage on ED Use - Further Evidence from Oregon's Experiment. *N Engl J Med, 375*(16), 1505-1507.

Finkelstein, A., Zhou, A., Taubman, S., & Doyle, J. (2020). Health Care Hotspotting - A Randomized, Controlled Trial. *N Engl J Med, 382*(2), 152-162.

Frakt, A. (2010). A little bit more about the RAND health insurance experiment. Retrieved March 27, 2022 from https://theincidentaleconomist.com/wordpress/a-little-bit-more-about-the-rand-health-insurance-experiment-2/.

Fulmer, E. B., Barbero, C., Gilchrist, S., Shantharam, S. S., Bhuiya, A. R., Taylor, L. N., & Jones, C. D. (2020). Translating Workforce Development Policy Interventions for Community Health Workers: Application of a Policy Research Continuum. *J Public Health Manag Pract, 26 Suppl 2, Advancing Legal Epidemiology*, S10-S18.

Glasgow, R. E., Harden, S. M., Gaglio, B., Rabin, B., Smith, M. L., Porter, G. C., . . . Estabrooks, P. A. (2019). RE-AIM Planning and Evaluation Framework: Adapting to New Science and Practice With a 20-Year Review. *Front Public Health, 7*, 64.

Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health, 89*(9), 1322-1327.

Hamilton AB, Finley EP. Qualitative methods in implementation research: An introduction. *Psychiatry Res.* 2019 Oct;280:112516. doi: 10.1016/j.psychres.2019.112516. Epub 2019 Aug 10. PMID: 31437661; PMCID: PMC7023962.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives, 9*(85-110).

Heller, S. B., Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. A. (2017). Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago. *Q J Econ, 132*(1), 1-54.

Higgins, S. T., Heil, S. H., Badger, G. J., Skelly, J. M., Solomon, L. J., & Bernstein, I. M. (2009). Educational disadvantage and cigarette smoking during pregnancy. *Drug Alcohol Depend, 104 Suppl 1*, S100-105.

Hirschhorn, L., Smith, J. D., Frisch, M. F., & Binagwaho, A. (2020). Integrating implementation science into covid-19 response and recovery. . *BMJ (Clinical research ed.), 369*.

Hoagwood, K. E., Purtle, J., Spandorfer, J., Peth-Pierce, R., & Horwitz, S. M. (2020). Aligning dissemination and implementation science with health policies to improve children's mental health. . *The American psychologist, 75*(8), 1130–1145.

Insel, T. R., & Gogtay, N. (2014). National Institute of Mental Health clinical trials: new opportunities, new expectations. *JAMA Psychiatry, 71*(7), 745-746.

Isaacs, K. R., Atreyapurapu, S., Alyusuf, A. H., Ledgerwood, D. M., Finnegan, L. P., Chang, K. H. K., . . . Washio, Y. (2021). Neonatal Outcomes after Combined Opioid and Nicotine Exposure in Utero: A Scoping Review. *Int J Environ Res Public Health, 18*(19), 10215.

Jepsen, C., & Rivkin, S. (2009). Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size,. *Journal of Human Resources, 44*(1), 223–250.

Johnson, A. L., Ecker, A. H., Fletcher, T. L., Hundt, N., Kauth, M. R., Martin, L. A., . . . Cully, J. A. (2020). Increasing the impact of randomized controlled trials: an example of a hybrid effectiveness-implementation design in psychotherapy research. *Transl Behav Med, 10*(3), 629-636.

Kemp, C. G., Wagenaar, B. H., & Haroz, E. E. (2019). Expanding Hybrid Studies for Implementation Research: Intervention, Implementation Strategy, and Context. *Front Public Health, 7*, 325.

Kirzinger, A., Sparks, G., Kearney, A., Stokes, M., Hamel, L., & Brodie, M. (2021). KFF COVID-19 Vaccine Monitor: November 2021. Retrieved March 27, 2022 from https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-november-2021/.

Komro, K. A., Dunlap, P., Sroczynski, N., Livingston, M. D., Kelly, M. A., Pepin, D., . . . Wagenaar, A. C. (2020). Anti-poverty policy and health: Attributes and diffusion of state earned income tax credits across U.S. states from 1980 to 2020. *PLoS One, 15*(11), e0242514.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review, 76*(4), 604-620.

Landes, S. J., McBain, S. A., & Curran, G. M. (2019). An introduction to effectiveness-implementation hybrid designs. *Psychiatry Res, 280*, 112513.

Lewandowska, M., Wieckowska, B., Sztorc, L., & Sajdak, S. (2020). Smoking and Smoking Cessation in the Risk for Fetal Growth Restriction and Low Birth Weight and Additive Effect of Maternal Obesity. *J Clin Med, 9*(11) 3504.

Lewis, C. C., Klasnja, P., Powell, B. J., Lyon, A. R., Tuzzio, L., Jones, S., . . . Weiner, B. (2018). From Classification to Causality: Advancing Understanding of Mechanisms of Change in Implementation Science. *Front Public Health, 6*, 136.

Lewis, C. C., Scott, K., & Marriott, B. R. (2018). A methodology for generating a tailored implementation blueprint: an exemplar from a youth residential setting. *Implement Sci, 13*(1), 68.

Lorch, S. A., Rogowski, J., Profit, J., & Phibbs, C. S. (2021). Access to risk-appropriate hospital care and disparities in neonatal outcomes in racial/ethnic groups and rural-urban populations. *Semin Perinatol, 45*(4), 151409.

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives, 25*(3), 17–38.

Ludwig, J., Sanbonmatsu, L., Gennetian, L., et al. (2011). Neighborhoods, obesity, and diabetes: A randomized social experiment. *New England Journal of Medicine, 365*(16), 1509–1519.

Mosley, J., Marwell, N., & Ybarra, M. (2019). How the "What Works" Movement is Failing Human Service Organizations, and What Social Work Can Do to Fix It. . *Human Service Organizations: Management, Leadership & Governance, 43*, 1-10.

Mundt, M. P., Fiore, M. C., Piper, M. E., Adsit, R. T., Kobinsky, K. H., Alaniz, K. M., & Baker, T. B. (2021). Cost-effectiveness of stop smoking incentives for medicaid-enrolled pregnant women. *Prev Med, 153*, 106777.

Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implement Sci, 10*, 53.

Noble, A., Vega, W. A., Kolody, B., Porter, P., Hwang, J., Merk, G. A., 2nd, & Bole, A. (1997). Prenatal substance abuse in California: findings from the Perinatal Substance Exposure Study. *J Psychoactive Drugs, 29*(1), 43-53.

Nighbor, T. D., Coleman, S. R. M., Bunn, J. Y., Kurti, A. N., Zvorsky, I., Orr, E. J., & Higgins, S. T. (2020). Smoking prevalence among U.S. national samples of pregnant women. *Prev Med, 132*, 105994.

Phibbs, C. S., Baker, L. C., Caughey, A. B., Danielsen, B., Schmitt, S. K., & Phibbs, R. H. (2007). Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *N Engl J Med, 356*(21), 2165-2175.

Phibbs, C. S., & Lorch, S. A. (2018). Choice of Hospital as a Source of Racial/Ethnic Disparities in Neonatal Mortality and Morbidity Rates. *JAMA Pediatr, 172*(3), 221-223.

Politis, C. E., Halligan, M. H., Keen, D., & Kerner, J. F. (2014). Supporting the diffusion of healthy public policy in Canada: the Prevention Policies Directory. *Online journal of public health informatics, 6*(2), e177-e177.

Powell, B. J., Beidas, R. S., Lewis, C. C., Aarons, G. A., McMillen, J. C., Proctor, E. K., & Mandell, D. S. (2017). Methods to Improve the Selection and Tailoring of Implementation Strategies. *J Behav Health Serv Res, 44*(2), 177-194.

Powell, B. J., Waltz, T. J., Chinman, M. J., Damschroder, L. J., Smith, J. L., Matthieu, M. M., . . . Kirchner, J. E. (2015). A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. *Implement Sci, 10*, 21.

Proctor, E. K., Powell, B. J., & McMillen, J. C. (2013). Implementation strategies: recommendations for specifying and reporting. *Implement Sci, 8*, 139.

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P. A., Aarons, G., Bunger, A., . . . Hensley, M. (2011). Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. . *Administration and policy in mental health, 38*(2), 65–76.

Saltz, R. F., Paschall, M. J., & O'Hara, S. E. (2021). Effects of a Community-Level Intervention on Alcohol-Related Motor Vehicle Crashes in California Cities: A Randomized Trial. *Am J Prev Med, 60*(1), 38-46.

Shelton, R. C., Adsul, P., & Oh, A. (2021). Recommendations for Addressing Structural Racism in Implementation Science: A Call to the Field. *Ethn Dis, 31*(Suppl 1), 357-364.

Shelton, R. C., Chambers, D. A., & Glasgow, R. E. (2020). An Extension of RE-AIM to Enhance Sustainability: Addressing Dynamic Context and Promoting Health Equity Over Time. *Front Public Health, 8*, 134.

Shelton, R. C., Cooper, B. R., & Stirman, S. W. (2018). The Sustainability of Evidence-Based Interventions and Practices in Public Health and Health Care. *Annu Rev Public Health, 39*, 55-76.

Smith JD, Hasan M. Quantitative approaches for the evaluation of implementation research studies. *Psychiatry Res*. 2020 Jan;283:112521. doi: 10.1016/j.psychres.2019.112521. Epub 2019 Aug 17. PMID: 31473029; PMCID: PMC7176071.

Snell-Rood, C., Jaramillo, E. T., Hamilton, A. B., Raskin, S. E., Nicosia, F. M., & Willging, C. (2021). Advancing health equity through a theoretically critical implementation science. *Transl Behav Med, 11*(8), 1617-1625.

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.

Tenforde, M. W., Self, W. H., Adams, K., Gaglani, M., Ginde, A. A., McNeal, T., . . . Network, I. a. O. V. i. t. A. I. I. (2021). Association Between mRNA Vaccination and COVID-19 Hospitalization and Disease Severity. *JAMA, 326*(20), 2043-2054.

Torres, C., Ogbu-Nwobodo, L., Alsan, M., Stanford, F. C., Banerjee, A., Breza, E., . . . Group, C.-W. (2021). effect of physician-delivered COVID-19 public health messages and messages acknowledging racial inequity on black and white adults' knowledge, beliefs, and practices related to COVID-19: A Randomized Clinical Trial. *JAMA Netw Open, 4*(7), e2117115.

Vega, W. A., Kolody, B., Hwang, J., & Noble, A. (1993). Prevalence and magnitude of perinatal substance exposures in California. *N Engl J Med, 329*(12), 850-854.

Victor, R. G., Ravenell, J. E., Freeman, A., Leonard, D., Bhat, D. G., Shafiq, M., . . . Haley, R. W. (2011). Effectiveness of a barber-based intervention for improving hypertension control in black men: the BARBER-1 study: a cluster randomized trial. *Arch Intern Med, 171*(4), 342-350.

Yang, I., & Hall, L. (2019). Factors related to prenatal smoking among socioeconomically disadvantaged women. *Women Health, 59*(9), 1026-1074.

Yang, I., Hall, L. A., Ashford, K., Paul, S., Polivka, B., & Ridner, S. L. (2017). Pathways From Socioeconomic Status to Prenatal Smoking: A Test of the Reserve Capacity Model. *Nurs Res, 66*(1), 2-11.