Evaluating Complex Health Interventions:

A Guide to Rigorous Research Designs

June 2017

AcademyHealth

Introduction

Evaluation of clinical care, public health and social programs is essential to make judgements about a program and provide an estimate of the impact that can be attributed to the introduction of an intervention. Evaluations are also needed to improve the effectiveness of programs and inform evidence-based decision-making. Clinical care, public health and social interventions are often either complex (i.e. characterized as having several components influencing each other) and multifaceted or implemented within complex systems with multiple actors and contextual factors that can impact implementation. An evaluation design must respond to the stakeholders' needs and the program or intervention's features, including complexity.

Purpose of Guide

As the number, variety and complexity of innovations increase and the need to understand which ones are working, for whom, and under what circumstances grows across sectors, it is clear that there is no single "correct" evaluation design. This guide is aimed at program managers and other stakeholders implementing innovations in public health and community settings who are involved in evaluation but may not be evaluators themselves. The guide provides a framework to guide decision-making around appropriate designs to evaluate public health and other service interventions. This guide will also provide program managers and other stakeholders the information necessary to understand and assess the strengths, weaknesses and validity of an evaluation that has already been conducted. It aims to provide a range of approaches that could be used to enhance the rigor of evaluations thus improving the quality of the evidence upon which decisions are made and ultimately improving the public's health.

Types of Evaluation Designs

Evaluation designs can be divided in three broad categories:

- experimental,
- quasi-experimental and
- observational designs.

In an **experimental design**, the investigators actively manipulate who receives the intervention or service and who does not in order to determine the effect of the intervention. Participants are randomly assigned into one or more intervention groups or a control group. Examples of experimental designs include randomized controlled trials, randomized encouragement trials, staggered enrollment trials and factorial designs.

In a **quasi-experimental design**, a comparison group is used. However, randomization is not used. Controlled before and after studies, interrupted time series, multiple baseline and regression discontinuities are types of quasi-experimental designs.

In an **observational design**, the evaluator does not manipulate who receives the intervention or not. Instead, the evaluator only observes and does not intervene to find associations between the intervention and the outcome. Examples of observational studies include natural experiments, case-control and prospective and retrospective cohort studies.

In addition to these evaluation designs, statistical approaches such as difference in difference analyses can be used to adjust for the fact that in non-randomized experiments, the intervention may not explain all the difference between the interventions and the control group in terms of key outcome of interest because the two groups did not start at the same level at baseline.

Additional approaches can be used to evaluate programs, including qualitative and mixed methods. While quantitative approaches tend to focus on *whether* an intervention or service "worked" in achieving any number of outcomes (e.g. improved health outcomes, cost saving, etc.), **qualitative methods** can enhance evaluations by providing in-depth information on *how* a program is implemented, *why* it is yielding or not yielding expected results and explore variation in results across settings or contexts. They can therefore be used to identify issues related to the degree to which the intervention is implemented as intended (implementation fidelity). Qualitative approaches can also inform replication, spread and scale-up by providing information on the most important components of a program that are needed to successfully transfer the program to other settings, for instance. Qualitative methods include a variety of interview types (unstructured, semi-structured, structured), focus groups (i.e., group interviews), observations, and document or content analysis.

Mixed methods involve the integration of both quantitative and qualitative data collection and analysis in the evaluation. The combination of both approaches is expected to provide a better understanding than each approach separately, help overcome the weakness of a single approach and present different perspectives. In addition, evidence may be strengthened when one method confirms findings from the other method. Conflicting findings should lead to further investigation.

There is no **"one right"** evaluation design.

Considerations When Selecting a Design

While randomization allows investigators to assess the existence of a causal relationship between the intervention and the outcome

of interest, as opposed to just an association, randomization is not always feasible, practical or appropriate. There is no "one right" evaluation design; the evaluation design must be appropriate for the evaluation purpose or question (e.g., formative and/or summative), the nature of the intervention, the context in which it is implemented, the needs and/or expectations of the audience for the evaluation, and the availability of data.

Evaluation needs may change over time. The appropriate design will also depend on the characteristics of the intervention. For instance, quality improvement programs often start with an initial package of changes which participants test and adapt to their local context over time; this can have the effect of changing the intervention over time. Quality improvement evaluations need to take these iterative and context-specific features into account and include a comparator group whenever possible. Traditional, fixed evaluation protocol designs would not be appropriate, however much progress has been made in recent years in developing adaptive evaluation designs that are suited to improvement interventions and bring additional rigor.

Timeline and budget considerations may also influence the choice of an evaluation design. The specific considerations required for a given design will vary based on the evaluation question of interest and from setting to setting. In general, if the evaluation must detect a relatively small, but meaningful, difference (effect size) in the impact of an intervention, then the sample size will be increased, often resulting in an increased budget for data collection compared to large effect sizes. Additionally, if the evaluation is multi-site, additional data collection may be required with subsequent impact on the budget. Furthermore, costs are likely to be higher when secondary data (e.g., administrative data, electronic health record data, registries) are not available, and primary data (e.g., semi-structured interview, surveys) are required. However, the growing abundance of secondary data sources may enable more cost-efficient evaluation designs.[1] In relation to the timeline, consideration will also need to be given to the time it takes for the intervention to show an impact in the outcomes of interest.

Regardless of the approach selected, some key principles should be followed:

- Identify and engage the key stakeholders (e.g. funders, policymakers, community leaders, etc.) early in the process in order to define the evaluation questions and identify the most relevant outcomes and data collection approaches.
- Facilitate agreement between evaluators and implementers on a shared, explicit understanding of the theory of change–a comprehensive description of how and why a desired change is expected to occur in a specific context.
- Clearly define evaluation questions that respond to stakeholder's needs and priorities.
- Assess the attribution of the results to the program by establishing a counterfactual (a control group to assess what is likely to happen if the initiative is not introduced) whenever possible.
- Clearly define and prioritize measures and data required based on evaluation questions and available resources.
- Ensure evaluation results show the magnitude of the effect and the degree of confidence that exists in those results (e.g. by presenting confidence intervals).
- Ensure evaluation results are expressed in the most meaningful ways for the targeted stakeholders to enhance use in decision-making; (e.g. sometimes absolute risks are more informative than relative risks).
- Add qualitative methods to evaluation designs whenever possible to explore questions of how, why or the contexts and/or conditions under which policy, programs, or interventions are and are not successful. The Consolidated Framework for Advancing Implementation Science (CFIR) can be helpful. The CFIR describes an outer setting (economic, political, and social contexts) within which an organization resides, and an inner setting (structural, political, and cultural contexts) through which the implementation process will proceed.[2]
- When interpreting results, consider existing biases and confounding, and acknowledge limitations.
- Systematically promote the dissemination and use of evaluation results for decision-making. It is therefore essential for evaluators to be able to translate nuanced and complicated results to a policy or decision-making audience that may not fully understand complicated or complex statistical methods.

This guide aims to help the reader **make informed decisions** by providing information on the various trade-offs involved in the selection of an evaluation design.

Organization of Guide

The list of evaluation designs presented in this guide is not exhaustive but represents a mix of experimental, quasi-experimental and observational designs. Resources for more detailed information on a broader range of designs are listed in the "Further Readings" section. For each of the six designs presented, we provide a general description with a diagram to illustrate the design, two examples from the peer-reviewed literature of how the design was used to evaluate a specific health or social service, key strengths and weaknesses of the study design, timeline and budget considerations that

are specific to the design and policy implications and consideration for future use. We also include a flow chart to inform the selection of evaluation designs. This guide aims to help the reader make informed decisions by providing information on the various trade-offs involved in the selection of an evaluation design.

Further Readings on Evaluation Designs

Basu, S, Meghani, A, & Siddiqi, A. Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches. Annu Rev Public Health, 2017 Mar 20;38:351-370.

Campbell, DT, Stanley, JC, & Gage, NL. Experimental and Quasi-Experimental Designs for Research. Boston, MA: Houghton Mifflin, 1963.

McGlynn, EA, & McClellan, M. Strategies for Assessing Delivery System Innovations. Health Aff, 2017 Mar 1;36(3):408-416.

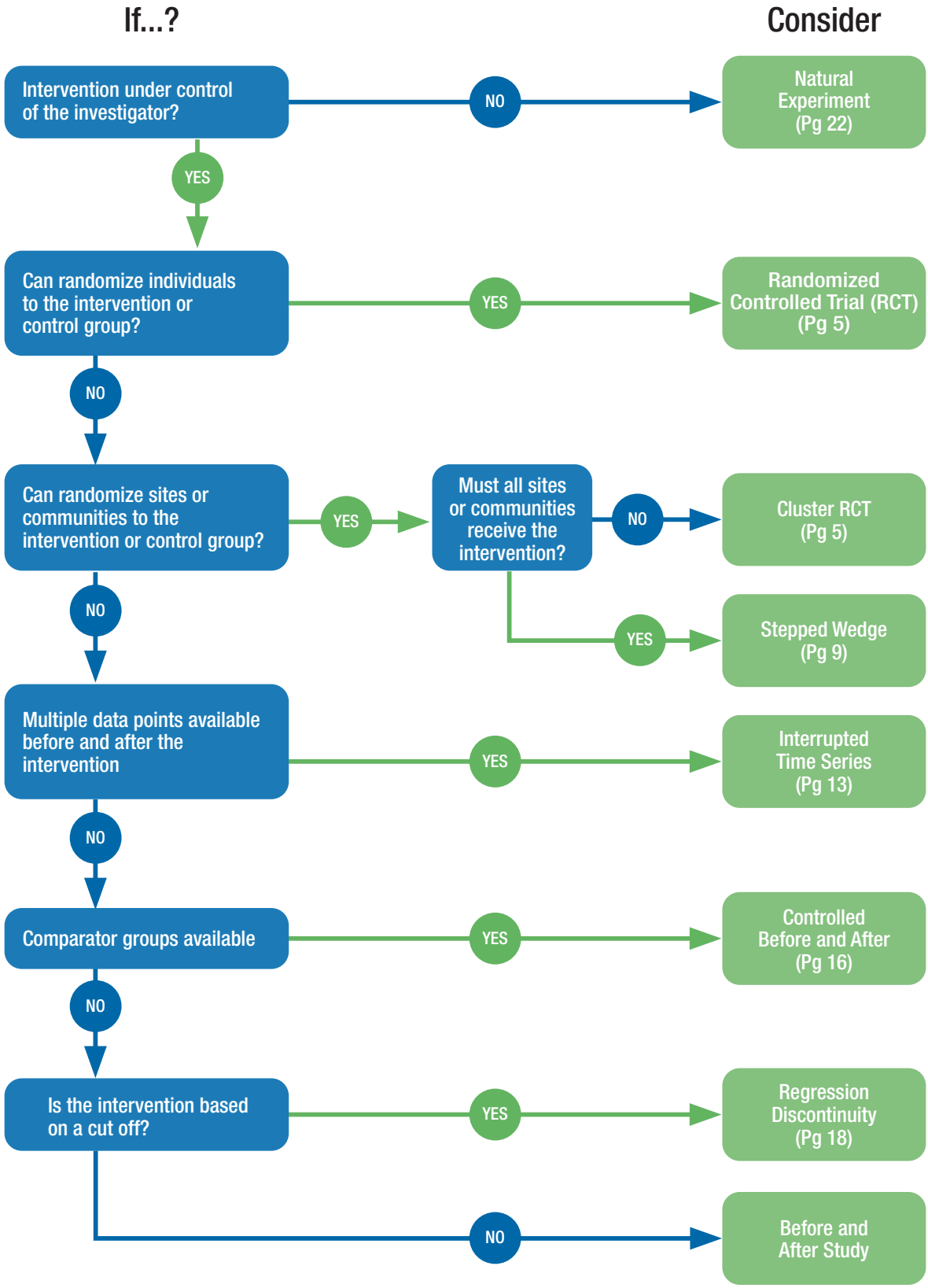
Raine, R, Fitzpatrick, R, Barratt, H, Bevan, G, Black, N, Boaden, R et al. Challenges, Solutions and Future Directions in the Evaluation of Service Innovations in Health Care and Public Health. Health Serv Deliv Res, 2016;4(16).

Rothman, K, Greenland, S, & Lash, TL. Modern Epidemiology, 3rd Edition. Philadelphia, PA: Lippincott Williams & Wilkins, 2008.

Selecting an Evaluation Design

This flow chart (Figure 1) can be used to guide the selection of evaluation designs among those presented in this guide. The flow chart focuses on a few key characteristics such as the availability of a comparison group, the feasibility of randomization at the individual or group level, etc. As emphasized in the introduction, other factors must also be considered in the selection of an evaluation design.

Figure 1. Flow Chart to Help Guide Choice of Evaluation Design



Experimental/Randomized Designs

Randomized Controlled Trial

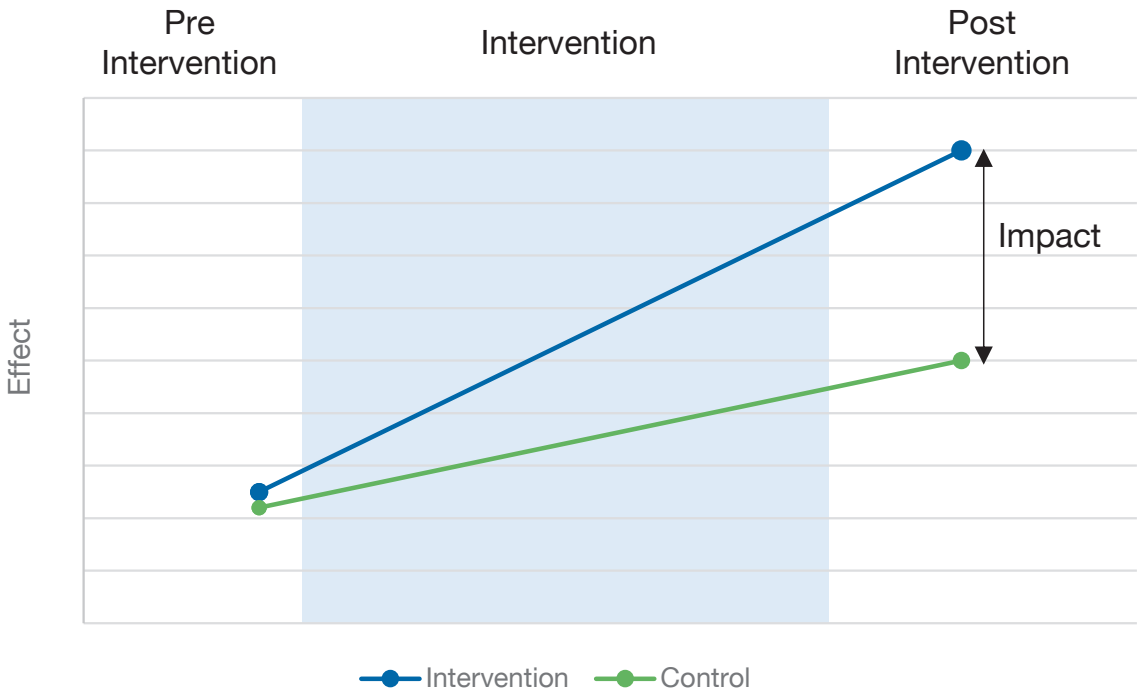
Individuals (individually randomized trials) or other units such as health facilities or communities (cluster/group randomized trials) are randomly assigned into one or more intervention groups or a control group to assess the effectiveness of an intervention.

The only expected difference between the control and intervention groups in a randomized controlled trial (RCT) is the outcome variable(s) being studied (Figure 2). Other factors that could possibly affect the outcome (potential confounders) are assumed to be evenly distributed across the groups; any significant differences between groups in the outcome of interest can therefore be attributed to the intervention and not to some other unidentified factor.

The rationale for conducting a RCT should be based on prior observational data and uncertainty about the intervention's effectiveness (the degree to which it has beneficial effects under real world circumstances).[3] RCTs should report variations in the nature and size of the effects across individuals or clusters, not just "average" effects in the study, to enhance learning from variation. The intervention should be amenable to randomization and the intervention's benefits should be expected to exceed the risks. For instance, evaluations of implemented policies or investments in infrastructures are not conducive to randomization. Similarly, it would be unethical to use randomization when studying the effect of exposure to known environmental hazards. Generally, when there is a standard intervention or service that is already being provided, this "standard of care" should be the control group rather than no intervention or service.

RCTs should report variations in the nature and size of the effects across individuals or clusters, not just **"average"** effects in the study, to enhance learning from variation.

Figure 2. Illustration of a Generic Randomized Controlled Trial Design



Example 1. A Web-Based Health Promotion Program for Older Workers: Randomized Controlled Trial

An individually randomized controlled trial funded by the National Institute of Health (NIH) was conducted by ISA Associates, a research organization, to determine the effectiveness of a web-based health promotion program called HealthyPast50 for workers 50 years and older. This program provides information and guidance to promote healthy aging, healthy diet, physical activity, stress management and quitting tobacco use. Participants were recruited from a global information technology company with multiples offices in the United States including locations in Massachusetts and California.

In this study, 278 workers were randomly assigned to HealthyPast50 or to a control group after baseline completion. Participants were not blinded to the group to which they were assigned. Self-reported data were collected from both the intervention and the control group on diet, physical activity, stress and tobacco use before the intervention and three months after the start of the intervention through an online survey.

This study found that participants in the intervention group had better self-reported outcomes than those in the control group in terms of: their belief in their ability to change their diet (adjusted difference: 0.16, 95% Confidence Interval (CI): 0.00, 0.31), mild exercise (1.03; 95% CI: 0.26, 1.81), planning healthy eating (0.17; 95% CI: 0.01, 0.33). The authors concluded that HealthyPast50 has the potential to contribute to improved diet and exercise in the short term.

This individually randomized trial was a suitable design for this study topic as there was some reasonable expectation that the HealthyPast50 program would be more beneficial than not receiving any guidance or support for healthy aging. It was ethical to use a control group with no services, as there was no current standard of care to address those issues in the target population. In addition, the use of an online survey and relatively short-term follow time did not require too many resources. Limitations of this particular study include the use of self-reported data which could lead to biases, a very short follow-up period that may not accurately capture the impact of the intervention over time, and a small number participants working at a global IT firm which may not be representative of the population as a whole.

Reference: Cook, RF, Hersch, RK, Schlossberg, D, & Leaf, SL. A Web-Based Health Promotion Program for Older Workers: Randomized Controlled Trial. Eysenbach G, ed. Journal of Medical Internet Research, 2015;17(3):e82.

Example 2. 12-month Outcomes of Community Engagement Versus Technical Assistance to Implement Depression Collaborative Care: A Partnered, Cluster, Randomized, Comparative Effectiveness Trial

A cluster randomized controlled trial was conducted in Los Angeles to compare the effectiveness of two approaches to provide depression collaborative care trainings to mental health, medical, and community-based agencies. The goal of the care was to increase depressed clients' mental health-related quality of life (MHRQL) and services use at 12 months. The intervention, community engagement and planning (CEP), supports a large number of community programs to collaboratively develop and implement a training plan to provide services for depression. The control, resources for services (RS), provides short-term trainings for collaborative care to individual organizations running community programs. Eligibility criteria for agencies to be included in the study were: serving at least 15 clients per week, having one or more staff, not focusing on psychotic disorders or home services.

A total of 133 potentially eligible programs were randomized into the RS group (65) or the CEP (68) group. Following randomization, 20 programs were determined ineligible by assessors blinded to the assignment, and an additional 18 declined to participate. Therefore, 95 programs were enrolled in the study, 46 in the RS arm and 49 in the CEP arm. Of the 1,322 eligible clients in those programs, 1,246 were enrolled in the study (606 for RS and 640 for CEP), and 981 completed the baseline survey, 759 completed the 6-month survey and 733 completed the 12-month survey. Reasons for differences in follow up included refusals,

Cont'd on next page

Example 2. 12-month Outcomes of Community Engagement Versus Technical Assistance to Implement Depression Collaborative Care: A Partnered, Cluster, Randomized, Comparative Effectiveness Trial (Cont'd)

illnesses or death, or incarceration. Data were collected through telephone surveys. The authors used a variety of statistical assumptions to adjust for missing data in the analyses.

In some analyses, CEP was associated with a decrease in poor MHRQL compared to RS at 6 months (OR=0.71; 95% CI: 0.55, 0.91) and 12 months (OR=0.77; 95% CI: 0.61, 0.97). CEP was also associated with less behavioral health hospitalization in the prior 6 months, at 6 months (OR=0.60; 95% CI: 0.37, 0.98) but not at 12 months (OR=0.70; 95% CI: 0.4, 1.22). The authors concluded that while CEP did not indicate an effect at 12 months, policymakers and communities should still consider this strategy given the lack of alternative approaches that have demonstrated higher effectiveness.

This was a suitable study design because there was a reasonable expectation that the benefits of the intervention outweigh the risk, as effectiveness had previously been demonstrated at 6 months using quasi-experimental designs. The control group received an alternative depression collaborative care trainings approach, hence eliminating concerns usually associated with withholding potentially beneficial services from the control group. In addition, cluster randomization was appropriate because it would have been logistically difficult to randomize individuals seen in the same facility to different interventions. The relatively short-term outcome measure led to a shorter study duration and lower costs as compared to a long-term outcome. However, it is not clear whether the impact was long lasting.

Reference: Chung, B, Ong, M, Ettner, SL, Jones, F, Gilmore, J, McCreary, M et al. 12-Month Outcomes of Community Engagement Versus Technical Assistance to Implement Depression Collaborative Care: A Partnered, Cluster, Randomized, Comparative Effectiveness Trial. Annals of Internal Medicine, 2014; 161(10 Suppl):S23-34.

Key Strengths

- Considered by many to be the gold standard for proof of effectiveness.
- A safety monitoring committee can be used to determine if the trial should be terminated early because preliminary results indicate that the intervention is effective or is unlikely to be effective.
- Appropriate randomization, sample size, and assessment of implementation fidelity can help ensure that the difference in outcomes between the intervention and control groups can be attributed to the intervention and not to other factors (i.e. bias/confounding).
- Randomization meets statistical assumptions needed to establish that the intervention caused the outcome of interest.
- Some questions are amenable to blinding so that participants and/or data collectors do not know which group is the intervention and which is the control.
- More than one intervention can be compared to a control group.
- Cluster randomized trials can be used when individual randomization is not possible or desirable.
- Pragmatic trials (trials which show effectiveness in real-life settings) are well suited for complex interventions and can inform clinical and policy decisions (versus explanatory trials which are conducted under optimal conditions).

Key Limitations

- Although RCTs are considered the gold standard for assessing the effectiveness of an intervention, they too can be subject to flaws in data collection, statistical analysis and interpretation.
- It may be challenging to construct an appropriate control group.
- Some questions/contexts are not amenable to RCTs: randomization may not be practical or ethical, and RCTs may be inefficient for rare outcomes or outcomes that take a long time to develop.
- RCTs suffer from limited external validity. Findings may not be generalizable to a broader population or different contexts.
- Because participants (at either the individual or group level) must agree to be part of the RCT, they may not be representative of the larger population.
- The control group may be inadvertently exposed to the intervention (i.e. contamination).
- Retention of participants may be different between intervention and control groups. Efforts to mitigate this using intent to treat analyses (including each participant who is randomized per their assigned group regardless of retention, compliance), can lead to more conservative estimates of effects, as participants who were not compliant and those who dropped out will still be considered to be part of the intervention group.

- External factors (e.g. in the policy or community environment) may produce major changes in trends over time that dilute the effects of the interventions being studied. For instance, an anti-smoking policy initiative may decrease the observed effect of a smoking cessation program.
- Cluster randomized trials require a large sample size and require more complicated statistical analysis.
- Fixed-protocol RCTs do not allow for adaptation of the interventions and their implementation based on field experience. This can be addressed by "adaptive trials", which include pre-specified time points during the trial when modifications in the protocol can be made. Adaptive trials also allow "pruning" of futile intervention arms to focus on interventions that are more likely to have an impact.

Timeline and Cost Considerations

- Measuring longer term outcomes can extend the length of the study and increase costs.
- RCTs require a data safety monitoring committee to avoid needlessly prolonging an intervention that is unlikely to be beneficial or prolonging a trial and withholding the intervention from the control group when the intervention is clearly beneficial. This may increase costs.
- Although RCTs can be costly, they can be conducted with reasonable costs when using existing data sources and when the outcomes of interest do not require a lengthy study. In addition, large simple trials (LST), which involve a large number of participants, limited data collection and an outcome that can be collected easily, are less costly than traditional RCTs and may permit rapid detection of effect.
- In many countries including the United States, RCTs must be registered so that data is accessible for scrutiny and further analysis, even if trials have a "negative" result. This may result in higher administrative costs.

The appropriate **role of RCTs in policy and program evaluations** across health and social policy sectors has been the subject of much debate.

Policy Implications and Considerations for Future Use

The appropriate role of RCTs in policy and program evaluations across health and social policy sectors has been the subject of much debate. In "Show Me the Evidence" Haskins and Margolis have argued that "claiming that RCTs are the best way to definitively establish causality does not imply that all other evidence has no value." Others such as Patton have argued that labeling RCTs as the "gold standard" does disservice to the field of evaluation as it implies that other methods are inferior and may lead evaluators to use RCTs even when not appropriate. He posits that RCTs may create an artificial environment and are not appropriate for context-specific interventions as they rely on standardized interventions. Patton has further argued that the gold standard for evaluations should be "methodological appropriateness." In a 2003 statement, the American Evaluation Association declared that "RCTs are not always best for determining causality and can be misleading. RCTs examine a limited number of isolated factors that are neither limited nor isolated in natural settings. The complex nature of causality and the multitude of actual influences on outcomes render RCTs less capable of discovering causality than designs sensitive to local culture and conditions and open to unanticipated causal factors." Due to these concerns, some have emphasized the importance of supplementing RCTs with other methodologies including historical (non-concurrent) controls, epidemiological and qualitative data.

Nonetheless, when well designed, and implemented, RCTs are considered the gold standard to assess the effectiveness of a wide range of interventions and services. Policymakers can use the results of RCTs to inform policies related to promoting uptake of interventions found effective. In doing so, policymakers will need to take care to address issues related to external validity (generalizability). They should consider whether interventions found effective in the context of the controlled setting of the trial are likely to be replicable in a variety of other contexts or settings.

Traditional RCTs primarily focus on internal validity, the ability to make causal inference for the intervention in a specified population. Thus, they are generally not well suited to establish generalizability (applying results of study to a larger population) or to evaluate the effectiveness of interventions that involve many components and actors and are likely to vary across individuals such as may be the case with complex interventions.

References and Further Reading

Abdul Latif Jameel Poverty Action Lab (J-PAL). https://www.povertyactionlab.org/research-resources/introduction-evaluations

American Evaluation Association, 2003. http://www.eval.org/p/cm/ld/fid=95

Buring, JE, Jonas, MA, & Hennekens, CH. "Large and Simple Randomized Trials". In Tools for Evaluating Health Technologies: Five Background Papers, BP–I–I-142, Ed, U.S. Congress, Office of Technology Assessment. Washington, DC: U.S. Government Printing Office, February 1995.

Chow, SC, & Chang, M. Adaptive Design Methods in Clinical Trials- A Review. Orphanet J Rare Dis, 2008 May 2;3:11.

Guta, SK. Intention to Treat Analysis: A Review. Perspect Clin Res, 2011 Jul-Sep; 2(3):109–112.

Ford, I, & Norrie, J. Pragmatic Trials. New England Journal of Medicine, 2016;375(5):454-463.

Haskins, R, & Greg, M. Show Me the Evidence: Obama's Fight for Rigor and Evidence in Social Policy. Brookings Institution Press, 2015.

Patton, MQ. Utilization-Focused Evaluation, 4th ed. Sage Publications, 2008.

Patton, MQ. Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use. Guilford Press, 2011.

Spieth, PM, Kubasch, AS, Penzlin, AI, Illigens, BM-W, Barlinn, K, & Siepmann, T. Randomized Controlled Trials – A Matter of Design. Neuropsychiatric Disease and Treatment, 2016;12:1341-1349.

Cluster Randomized Stepped Wedge Design

A cluster randomized stepped wedge design includes multiple phases; an initial phase where no clusters (geographically defined area, community, health facilities, schools etc.) receive the intervention and subsequent phases during which one or more clusters are randomized to be part of the intervention at regular pre-specified intervals or steps (Figure 3). By the end of the study, all clusters will have transferred to the intervention group.

There should be an expectation that the benefits of the intervention exceed the potential harm. This design is particularly suited for situations where:

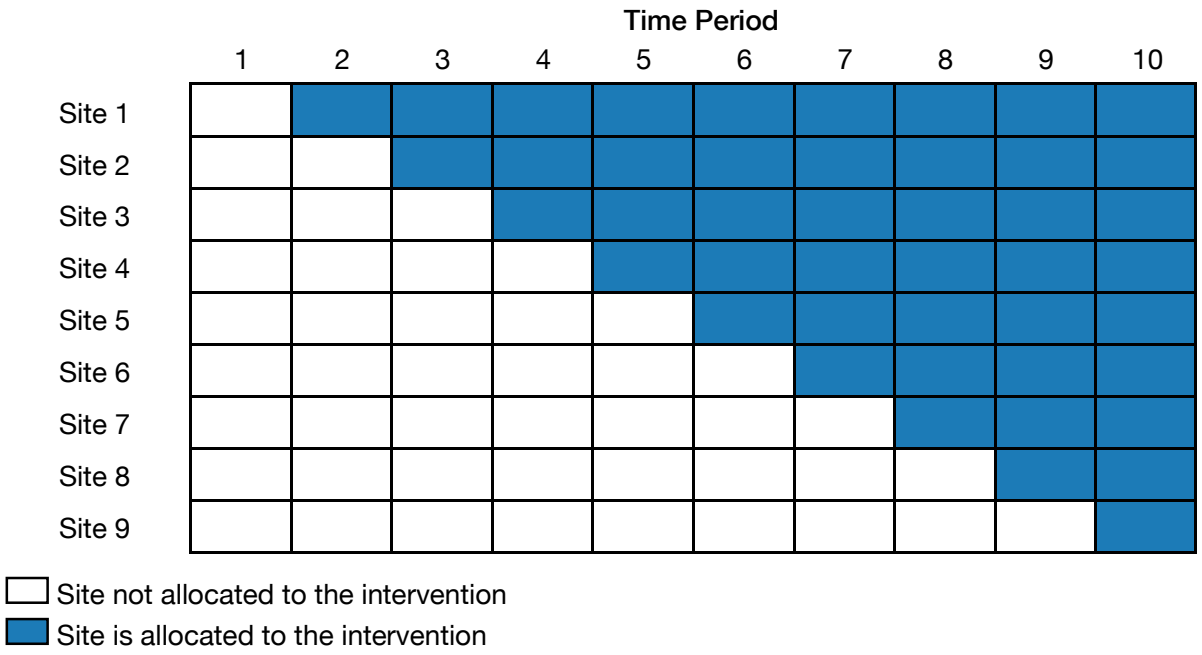
- policy makers/ providers/ users all want the intervention;
- it is unethical to have a group that does not receive the intervention;
- all participants cannot receive the intervention at the same time due to logistical, practical or financial considerations;
- there is a need to evaluate the effectiveness of programs or policies that are implemented in phases.

The cluster randomized stepped wedge design is a relatively new study design and has received increased attention because it allows the use of a rigorous evaluation design while addressing logistical, political and ethical considerations that might prevent a more traditional RCT.

One variation of the design involves starting the trial with a convenience sample or with sites that are available to join the intervention. While practical, this type of non-randomized design is subject to more bias.

The cluster randomized stepped wedge design **has received increased attention** because it allows the use of a rigorous evaluation design while addressing logistical, political and ethical considerations that might prevent a more traditional RCT.

Figure 3. Illustration of a Generic Stepped Wedge Design



Example 1. The Devon Active Villages Evaluation (DAVE) Trial of a Community-level Physical Activity Intervention in Rural South-West England: A Stepped Wedge Cluster Randomized Controlled Trial

Researchers at the University of Exeter in the United Kingdom (U.K.) used a cluster randomized stepped wedged design to evaluate the effect of a community-level physical activity intervention, the Devon Active Villages. A total of 128 clusters (villages) were randomized to receive the intervention in four time periods between April 2011 and December 2012. Computer generated numbers were used to determine the time during which each village joined the intervention. The number of villages which joined the intervention was as follows:

- April-June 2011 (22);
- September-November 2011 (36),
- April-June 2012 (35) and
- September-November 2012 (35).

The intervention consisted of providing villages with 12 weeks of at least three different types of opportunities for physical activities. Interestingly, the type of opportunities for physical activity presented to residents was tailored to each village's specific needs as residents were asked what activities they wanted the intervention to provide. All villages also received support for 12 months following the active phase of the intervention. During each data collection period, questionnaires were mailed to a random sample of households within each cluster and data collected from residents 18 or older. The response rate was 32.2%, and questionnaire responses from 4,693 adults in the intervention communities were compared to 5,719 adults in the control villages.

Findings showed that the intervention was not associated with increased likelihood of meeting the U.K. physical activity guidelines (adjusted OR =1.02, 95% CI: 0.88, 1.17). The intervention was only marginally associated with an increase in moderate to vigorous intensity activity per week. The adjusted mean difference before and after the introduction of the intervention was 171 (95% CI: -16, 358). The authors concluded that the lack of detected effect of the intervention may be due to low awareness of the intervention in the community and lack of residents' participation; only 4% of the residents reported participating in the intervention activities.

This design provided logistical, financial and ethical advantages. The stepped wedge design allowed the intervention to be delivered to a large population in several steps with relatively low costs. In addition, the entire population received the intervention. However, the response rate was low, and there is no information as to why the participation was extremely low. In addition, given that the type of activities presented was tailored to each village's needs, it is possible that differences in intervention effects may be due to differences in physical activities offered. Nonetheless, this evaluation provides an interesting example of adapting a randomized design to real-life considerations.

Reference: Solomon, E, Rees, T, Ukoumunne, OC, Metcalf, B, & Hillsdon, M. The Devon Active Villages Evaluation (DAVE) Trial of a Community-level Physical Activity Intervention in Rural South-west England: A stepped wedge Cluster Randomised Controlled Trial. Int J Behav Nutr Phys Act, 2014 Jul 18;11:94.

Example 2. A Structural Multidisciplinary Approach to Depression Management in Nursing Home Residents: A Multicenter, Stepped Wedge Cluster Randomized Trial

A stepped wedge design was conducted to assess the effectiveness of a structural approach to the management of depression among nursing home residents in four provinces in the Netherlands from May 2009 to April 2011. The intervention, Act in Case of Depression (AiD), consisted of implementing a two-step screening and diagnosis procedure, multidisciplinary treatment and monitoring the effect of the treatment. The treatment approach allowed nursing home staff to follow pathways for collaborative treatment including psychosocial interventions.

In the study, each nursing home was invited to enroll one dementia unit and one unit focused on physical health needs, and patients provided written informed consent to participate. Overall, 16 dementia units with 403 patients and 17 physical health units with 390 patients were enrolled in the study. Units were randomly assigned to each of five groups using computer generated numbers, and each of these groups joined the intervention at different time periods. The first group joined the intervention shortly following baseline data collection, the

Cont'd on next page

Example 2. A Structural Multidisciplinary Approach to Depression Management in Nursing Home Residents: A Multicenter, Stepped Wedge Cluster Randomized Trial (Cont'd)

other groups joined the intervention approximately every four months thereafter. By the end of the study all five groups were in the intervention.

Findings show that the intervention was associated with a decrease in the prevalence of depression in the physical health units –7.3%; 95% CI: –13.7%, –0·9%) but not in dementia units (0.6%; 95% CI: –5.6%, 6.8%). Physical health units had higher adherence to assessment procedures than dementia units (p=0.045), but adherence to treatment did not differ (p=0.745). The authors concluded that while a structural approach to depression management including systematic depression assessment can reduce depression in physical health units, depression screening needs to be addressed in dementia units as these units had lower adherence to screening.

The use of the stepped wedge design made it possible to manage at the beginning of the intervention a smaller number of clusters at a time. In addition, the intervention, which was believed to be more likely to be beneficial than harmful, was not withheld from any participating nursing home patients.

Reference: Leontjevas, R, Gerritsen, DL, Smalbrugge, M, Teerenstra, S, Vernooij-Dassen, MJ, & Koopmans, RT. A Structural Multidisciplinary Approach to Depression Management in Nursing-Home Residents: A Multicentre, Stepped-Wedge Cluster-Randomised Trial. The Lancet, 2013;381(9885):2255-2264.

Key Strengths

- Safety monitoring board can be used to adjudicate if the trial should be terminated early due to clear effectiveness or a low probability that effectiveness will be demonstrated by continuing the trial.
- Allows the use of a rigorous evaluation design while addressing logistical, political and ethical considerations that might prevent a more traditional RCT.
- Provides an alternative to the traditional RCT approach in situations where the traditional RCT design is not possible or practical, e.g. when individual randomization is not possible or desirable or where all individuals or groups need to receive the intervention.
- Within-cluster effects can be evaluated. Each cluster receives the intervention and is also a control. If there is a significant cluster effect, this increases the power of the evaluation to detect differences compared to a design where clusters are assigned to only a control or intervention group.
- Secular trends can be considered with participants joining intervention groups at different times.
- May have higher enrollment and retention than RCTs, as control clusters know they will ultimately receive intervention.
- Managing a smaller number of clusters at a time in the intervention group might be advantageous from a logistical point of view.
- All participants eventually receive the intervention. There are not ethical issues to withholding the intervention from participants if effectiveness has yet to be established.
- Can examine the importance of the timing and length of the intervention on the outcome of interest.
- Can be used to evaluate barriers to implementation of the intervention and to iteratively improve implementation in subsequent steps.

Key Limitations

- If more rapid results are a priority for the funder/key audience, then study duration may be a limitation, as stepped wedge designs can be longer than RCTs, especially if only one or a few clusters can join the intervention at a time.
- Retention may be lower in control groups that are waiting longer to be part of the intervention than in other groups.
- Preventing contamination between those receiving the intervention and those to receive the intervention may be particularly challenging.
- Initial control group has shorter follow up time.
- It is often difficult to conceal group assignment to clusters. This may be problematic as the estimated effect of the intervention may be overestimated in randomized trials when individuals know whether they are receiving the intervention.
- Requires frequent data collection. This may be problematic if appropriate secondary data sources are absent and primary data collection is required.
- When the same clusters begin as controls and transfer to interventions, it can be challenging to control for the fact that a strong predictor for one cluster at a point in time is its value in the preceding period.

- Secular trends in the outcome of interest need to be controlled for in the analysis.
- Requires fewer clusters but more participants than a cluster randomized trial.

Timeline and Cost Considerations

- Lengthier than many designs, including RCT.
- Longer duration may require more resources and higher costs; each step may require training and other resources. Primary data collection during each step may also require additional resources.
- Requires a data safety monitoring committee to avoid needlessly prolonging an intervention that is unlikely to be beneficial or prolonging a trial and withholding the intervention from the control group when the intervention is clearly beneficial.

Policy Implications and Considerations for Future Use

Although the stepped wedge design offers a pragmatic and often more ethical and logistically easier option for evaluating complex interventions, its use still requires careful consideration. Along with its increased use, there has been increasing controversy surrounding the use of stepped wedge designs. Concerns have been raised regarding the longer duration of the study and lack of the control group over the entire duration of the evaluation for interventions that have not been demonstrated as effective. There has also been increased discussion about the relative advantages of the stepped wedge versus traditional randomized control trials in terms of power. Recent publications have reported that the relative power of the stepped wedge design and RCT depend on the value of the intracluster correlations (the degree to which individuals within a cluster resemble each other in terms of the outcome of interest). The stepped wedge design is therefore believed to be more efficient for studies with process indicators while randomized clinical trials are more efficient for clinical outcomes.

In addition, the duration of the evaluation should be carefully considered. A review of published stepped wedge designs showed that 52% of those published did not report a significant effect on the key outcomes of interest. The authors argued that although it is possible that these studies and evaluations may not have had enough power to detect a difference, the lack of detected effect may be due to insufficient study durations.

Furthermore, given that all participants are to receive the intervention, some have stressed the importance of conducting intermittent analyses during the course of the study to determine whether the study should be stopped early. Despite all these concerns, it has been argued that the stepped wedge design remains an important option in the evaluation of health interventions, including complex interventions, as this design may be preferable to the alternative, which would be the absence of randomization.

References and Further Reading

Beard, E, Lewis, J, Copas, A, Davey, C, Osrin, D, Baio G et al. Stepped Wedge Randomised Controlled Trials: Systematic Review of Studies Published Between 2010 and 2014. Trials, 2015;16:353.

Campbell, MK, Fayers, PM, & Grimshaw, JM. Determinants of the Intracluster Correlation Coefficient in Cluster Randomized Trials: The Case of Implementation Research. Clin Trials, 2005;2:99–107.

Cousens, S, Hargreaves, J, Bonell, C, Armstrong, B, Thomas, J, Kirkwood, BR et al. Alternatives to Randomisation in the Evaluation of Public-health Interventions: Statistical Analysis and Causal Inference. J Epidemiol Community Health, 2011 Jul;65(7):576-81.

de Hoop, E, van der Tweel, I, van der Graaf, R, Moons, KG, van Delden, JJ, Reitsma et al. The Need to Balance Merits and Limitations from Different Disciplines When Considering the Stepped Wedge Cluster Randomized Trial Design. BMC Med Res Methodol, 2015;15:93.

Donner, A. An Empirical Study of Cluster Randomization. Int J Epidemiol, 1982;11:283–286.

Fan, E, Laupacis, A, Pronovost, P, Guyatt, GH, & Needham, DM. How to Use an Article About Quality Improvement, JAMA 2010;304:2279–87.

Handley, MA, Schillinger, D, & Shiboski, S. Quasi-Experimental Designs in Practice-based Research Settings: Design and Implementation Considerations. J Am Board Fam Med, September-October 2011;24:589-596.

Hemming, K, Haines, TP, Chilton, PJ, Girling, AJ, & Lilford, RJ. The Stepped Wedge Cluster Randomised Trial: Rationale, Design, Analysis, and Reporting. BMJ, 2015;350:h391.

Hemming, K, Girling, A, Martin, J, & Bond, SJ. Stepped Wedge Cluster Randomized Trials Are Efficient and Provide a Method of Evaluation Without Which Some Interventions Would Not be Evaluated. J Clin Epidemiol, 2013;66(9):1058–9.

Keriel-Gascou, M, Buchet-Poyau, K, Rabilloud, M, Duclos, A, & Colin, C. A Stepped Wedge Cluster Randomized Trial is Preferable for Assessing Complex Health Interventions. J Clin Epidemiol, 2014;67(7):831–3.

Kotz, D, Spigt, M, Arts, ICW, Crutzen, R, & Viechtbauer, W. Researchers Should Convince Policy Makers to Perform a Classic Cluster Randomized Controlled Trial Instead of a Stepped Wedge Design When an Intervention is Rolled Out. J Clin Epidemiol, 2012 Dec;65(12):1255-6.

Li, F, & Frangakis, CE. Designs in Partially Controlled Studies: Messages from a Review. Stat Methods Med Res, 2005;14:417–31.

Mdege, ND, Man, MS, Taylor (nee Brown), CA, & Torgerson, DJ. Systematic Review of Stepped Wedge Cluster Randomized Trials Shows That Design is Particularly Used to Evaluate Interventions During Routine Implementation. J Clin Epidemiol, 2011;64:936–48.

Mdege, ND, Man, MS, Taylor (nee Brown), CA, & Torgerson, DJ. There Are Some Circumstances Where the Stepped-wedge Cluster Randomized Trial is Preferable to the Alternative: No Randomized Trial at All. Response to the Commentary by Kotz and Colleagues. J Clin Epidemiol, 2012;65(12):1253–4.

Michael, G, Wason, JM, & Mander, AP. Stepped Wedge Cluster Randomized Controlled Trial Designs: A Review of Reporting Quality and Design Features. Trials, 2017 Jan 21;18(1):33.

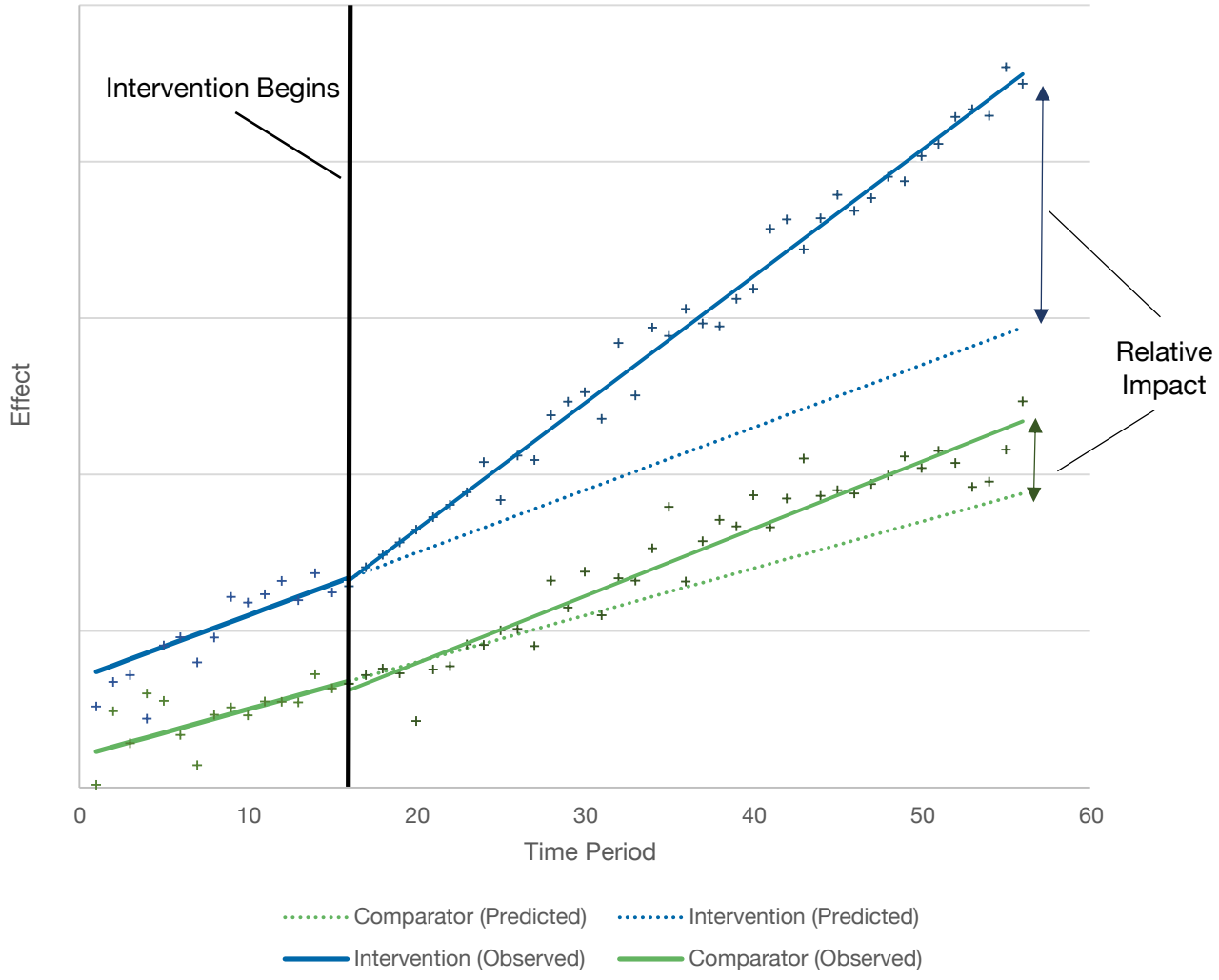
Quasi-experimental Designs

Interrupted Time Series Design

Interrupted time series (ITS) studies, a type of quasi-experimental design, are increasingly being used to evaluate the effectiveness of population-level health interventions. This design can be used when it is not considered possible to allocate the intervention randomly. Repeated data are collected over a series of time periods prior, during and after the introduction of an intervention. With sufficient collection of data over time, this can allow for simple pre-post test comparisons, and can allow for these comparisons to be adjusted for potential secular trends in the data before and after the introduction of the intervention (Figure 4).

This design is often used when the intervention is expected to have relatively quick and distinct effect, e.g. the uptake of health insurance. The design can focus on the introduction of an intervention in a single or group of settings, although is stronger when compared to changes in settings where the intervention was not introduced. To limit bias and confounding, *matching* of the comparator sites with the intervention sites, according to pre-specified patient and organizational characteristics, can be used. However, appropriate matching can often be hard to do, particularly when large scale interventions are being evaluated and it is hard to completely prevent exposure to the change amongst possible comparator sites.

Figure 4. Illustration of a Generic Interrupted Time Series Design



Example 1. Opioid Overdose Rates and Implementation of Overdose Education and Nasal Naloxone Distribution in Massachusetts: Interrupted Time Series Analysis

The overdose education and nasal naloxone distribution (OEND) program was introduced in Massachusetts in 2006 and was implemented among opioid users, their friends, families and related care staff. The investigators conducted an interrupted time series analysis to evaluate the impact of the OEND program on opioid-related mortality and acute care use in Massachusetts. The study used data on opioid-related mortality and acute care utilization from 19 Massachusetts communities from 2002 to 2009. The 19 communities were assigned to groups indicating no implementation, low enrollment and high enrollment in OEND. The investigators compared rates of opioid-related mortality and acute care utilization across the three groups, taking account of trends, prior to the introduction of OEND in 2006.

They found, relative to communities with no implementation, decreased opioid related mortality in low (adjusted ratio 0.75, 95% CI: 0.57, 0.91) and high enrollment communities (0.54, 95% CI: 0.39, 0.76). They found no significant difference in acute care utilization across the three groups. The investigators concluded that OEND is an effective intervention.

The assignment and uptake of the OEND program in these Massachusetts communities was not done on a randomized basis, meaning an observational study of some form was required. The investigators were able to make use of existing longitudinal outcomes related data from 2002 to 2009, covering the period before and after introduction of the program. By assigning the communities into no, low and high enrollment, the investigators were able to form reasonable comparator groups with which to assess changes in outcomes over time, and in particular prior to and after introduction of the program.

Reference: Walley, AY, Xuan, Z, Hackman, HH, Quinn, E, Doe-Simkins, M, Sorensen-Alawad, A et al. Opioid Overdose Rates and Implementation of Overdose Education and Nasal Naloxone Distribution in Massachusetts: Interrupted Time Series Analysis. BMJ, 2013 Jan 31;346:f174.

Example 2. Association Between Hospital Penalty Status Under the Hospital Readmission Reduction Program and Readmission Rates for Target and Non-target Conditions

As part of the Hospital Readmission Reduction Program (HRRP), in October 2012 financial penalties were imposed on hospitals with higher than expected readmissions for Medicare beneficiaries with acute myocardial infarction (AMI), congestive heart failure (CHF) and pneumonia. To assess the impact of the HRRP, the investigators used an interrupted time series analysis design to compare trends in readmission rates between hospitals subject to and those not subject to the penalty for program-specific and additional conditions.

The investigators used data on 48,137,102 hospitalizations of 20,351,161 Medicare beneficiaries admitted between January 2008 and June 2015. They identified 2,214 hospitals subject to a penalty and 1,283 not subject. The investigators identified three time periods:

- before announcement of the HRRP (January 2008 to March 2010),
- after announcement to implementation of the HRRP (March 2010 to October 2012) and
- after implementation of the HRRP (October 2012 to June 2015).

Before announcement of the HRRP, the investigators found that, across all hospitals, readmission rates were stable, with the exception of AMI, which reduced in nonpenalty hospitals. After the HRRP was announced, readmission rates declined more rapidly in all conditions studied in hospitals later subject to penalties relative to those not penalized, and this reduction was significantly greater in program-specific conditions (i.e. AMI, pneumonia and CHF). After implementation of the HRRP, the investigators found the rate of change for readmission rates plateaued for all conditions, with the exception of pneumonia which reduced in nonpenalty hospitals.

The investigators concluded that the HRRP was associated with greater reductions in readmission rates in penalized hospitals relative to nonpenalized hospitals. Moreover, within penalized hospitals, they also concluded that the reductions were greater among program specific conditions, relative to other conditions.

Cont'd on next page

Example 2. Association Between Hospital Penalty Status Under the Hospital Readmission Reduction Program and Readmission Rates for Target and Non-target Conditions (Cont'd)

The assignment of the hospitals to the penalty and nonpenalty groups was not randomized, meaning that a quasi-experimental design was required. The investigators were able to make use of longitudinal readmissions data from 2008 to 2015, covering a period before, and after introduction of the HRRP. In this study, whether the hospitals were in the penalty or nonpenalty group was related with the readmission outcome of interest. This might cause some to question whether bias occurred due to 'regression towards the mean', which occurs when a group is mistakenly identified as having high values of an outcome at an initial time point, whereas in fact the values are consistent with random variation. When observed over time, these outcomes are likely to reduce closer to the overall mean – hence 'regression to the mean'. However, the use of three time periods, and comparison of non-program specific conditions reduces the likelihood this source of bias occurred.

Reference: Desai, NR, Ross, JS, Kwon, JY, Herrin, J, Dharmarajan, K, Bernheim, SM et al. Association Between Hospital Penalty Status Under the Hospital Readmission Reduction Program and Readmission Rates for Target and Nontarget Conditions. JAMA, 2016 Dec 27;316(24):2647-56.

Key Strengths

- Takes into account underlying trends in the data–can examine trend before, during and after an intervention.
- With appropriate comparator groups, it can provide a strong quasi-experimental alternative to randomization.
- Can assess effect size, speed and sustainability of the intervention over time.
- Can provide an intuitive visual display of changes over time.

Key Limitations

- There can be challenges in identifying comparator groups that provide comparable data.
- It may be difficult to collect sufficient data points before and after the intervention to be able to detect a change in slope.
- Does not negate other events occurring at the same time as intervention.
- When using regression models, values towards the minimum and maximum range of the distribution exert the highest influence on the estimates provided. Interrupted time-series analysis creates a series of models, for each time period. In this situation, within each model, the maximum and minimum values lie at the start and end of the time periods, for example when an intervention starts. Consequently, the results of interrupted time series can be very sensitive to values at the start and end of time periods.
- The use of one data point, per time period means that adjusting for patient-level characteristics can be challenging. Often, a two-stage process has to be undertaken, where the analysis attempts to adjust for patient characteristics in the first step, and adjusted estimates are plotted over time, to form the time-series. This can limit the ability to fully understand the interaction of patient characteristics over time.

Timeline and Cost Considerations

- The requirement for multiple data points over time, before and after the intervention will:
  - Increase the resources required for data collection.
  - Increase the duration from when the intervention was administered to when the analysis will be undertaken and the results will be available.

Policy Implications and Considerations for Future Use

The interrupted time series design can align well with the evaluation of policy initiatives, especially when policies are introduced at a well-defined time point. In addition, as long as sufficient data exist or can be easily collected, questions relating to how long it takes for a policy to begin to achieve measurable impact can also be assessed. When stable aggregate-level data are available over time, population-level impact estimates can also be derived.

The interrupted time series design can allow for **exploration of variation at the site level**.

As with many study designs, considerations for wider policy implications need to take account how the intervention of interest interacted with a range of inner and outer contextual factors within the study setting, and how the intervention may interact with the context of new settings where it will be considered for further use. The interrupted time series design can allow for exploration of variation at the site level, but is more limited at the individual level, for example, by age, gender, ethnicity, income. Policymakers will need to account for such issues when making decisions on the results.

References and Further Reading

Bernal, JL, Cummins, S, & Gasparrini, A. Interrupted Time Series Regression for the Evaluation of Public Health Interventions: A Tutorial. International Journal of Epidemiology, 2016 Jun 9:dyw098.

Kontopantelis, E, Doran, T, Springate, DA, Buchan, I, & Reeves D. Regression Based Quasi-experimental Approach When Randomisation is Not an Option: Interrupted Time Series Analysis. BMJ, 2015 Jun 9;350:h2750.

Shadish, W, Cook, T, & Campbell, D. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin, 2002.

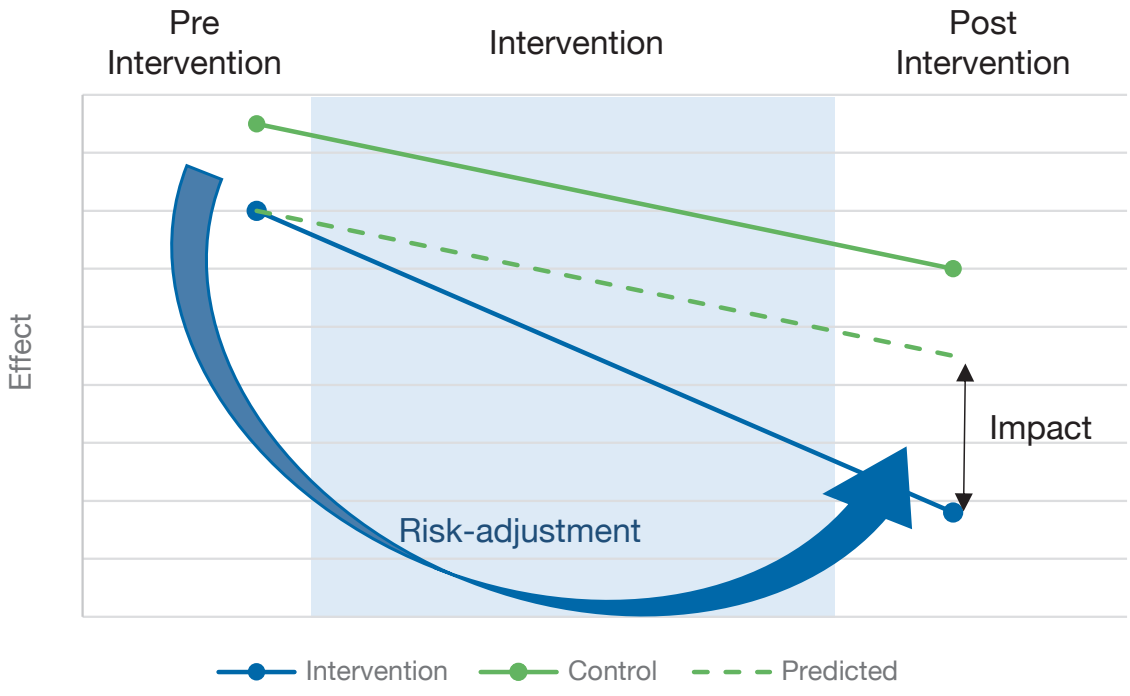
Controlled Before and After Design

Controlled before and after designs in the context of evaluating an intervention often compare the impact of an intervention that has been introduced in one or more settings or populations with selected settings or populations where it has not been introduced. This quasi-experimental design is used when it is not considered possible to allocate the intervention randomly. Studies compare outcomes in the intervention group, with those from a comparator group before and after the introduction of the intervention.

A variety of approaches can be used to adjust for potential confounding and bias, such as variation in patient and organizational factors likely to influence outcomes (Figure 5). These include *matching* the comparator sites with the intervention sites, so that they are similar in terms of patient or organizational factors that may impact on outcomes. Additionally, *risk-adjustment* or multiple regression techniques can be used to adjust the results for patient and organizational characteristics likely to influence the outcomes. Finally *propensity scoring*, can be applied to weight the data in a way that approximates samples with similar patient characteristics in the intervention and comparator groups.

In addition, a difference in difference approach is frequently used. Here, pre-intervention and post-intervention data on the outcome of interest are collected from the intervention and comparison groups. The difference between the two groups, pre-and post-intervention, can provide a relative estimate of the impact of the intervention.

Figure 5. Illustration of a Generic Controlled Before and After Design



Example 1. Multiple Component Patient Safety Intervention in English Hospitals: Controlled Evaluation of Second Phase

A controlled before and after study was used to evaluate the impact of a patient safety initiative in English hospitals. The study aimed to evaluate the impact of the Safer Patient Initiative in nine hospitals in 2007 on hand hygiene, adverse events, mortality, including mortality in intensive care, patient satisfaction and rates of hospital acquired infection.

The investigators identified a sample of matched control organizations to be used as comparators and determined that the intervention activities took place between 2007 and 2008. For the primary outcomes, they collected data before and after the intervention. This allowed the investigators to undertake a series of difference in difference analyses across the outcome measures.

Overall, they found a number of outcomes had improved in the intervention group between the pre-and post-intervention period. However, they also found that outcomes had improved in the control group. For example, for patient satisfaction, for the question "Overall, how would you rate the care you received?" respondents scored 80 pre-intervention, increasing to 84 post-intervention. In the control group, respondents' scores increased from 82 to 85, resulting in a difference in difference estimate of 1 (95% CI: -1, 3; p=0.292), suggesting no relative improvement in the intervention group.

The investigators concluded that although a number of outcomes were improving, the incremental impact of the initiative did not result in significant additional relative improvement.

The allocation of the intervention to sites was based on non-random selection criteria organized by the funder. Consequently, an observational approach was necessary. The selection of comparator sites and the use of data collected before, during and after introduction of the initiative allowed the investigators to estimate what was likely to happen without the introduction of the initiative.

Reference: Benning, A, Dixon-Woods, M, Nwulu, U, Ghaleb, M, Dawson, J, Barber, N et al. Multiple Component Patient Safety Intervention in English Hospitals: Controlled Evaluation of Second Phase. BMJ, 2011 Feb 3;342:d199.

Example 2. The Impact of Green House Adoption on Medicare Spending and Utilization

Following the passage of the Affordable Care Act, interest has grown in understanding the impact of a variety of nursing home delivery models. In this study, the investigators aimed to understand the impact of a specific model – the Green House (GH) on Medicare spending and utilization. The investigators described Green House as a culture change initiative that aims to offer a person-centered model of care. They described the three core tenets of the GH model as 1) consisting of small homes with 8–12 residents 2) empowering residents with greater control over their lives and care 3) eliminating the hierarchical nurse staffing structure often found in traditional nursing homes.

The investigators used data from Medicare claims and enrollment data from 2005 through 2010 and a resident-level assessment data set to estimate the impact of GH on Medicare acute hospital, other hospital, skilled nursing facility, and hospice spending and utilization.

The investigators identified 15 nursing homes that adopted the GH model and identified a group of 223 matched nursing homes that did not adopt the GH model. The matching was done by identifying nursing homes similar to each of the 15 receiving the intervention, according to 12 organizational factors, including ownership type, size and rural/urban location.

In the resulting dataset, comprising the 15 GH intervention nursing homes and 223 matched nursing homes, the investigators then applied a propensity scoring approach to weight the data in the subsequent analysis in a way that approximated intervention and comparison groups, with similar organizational characteristics.

The investigators applied a difference in difference approach to create a number of models comparing Medicare spending and utilization between GH intervention and comparator groups in the time period prior to introduction of the GH model relative to the time period post introduction. Overall the investigators did not detect any impact on Medicare spending and utilization in the GH intervention relative to the comparator groups. A sub-analysis revealed some savings within 12 nursing homes that had introduced the GH model sooner than three who had introduced it later.

Cont'd on next page

Example 2. The Impact of Green House Adoption on Medicare Spending and Utilization (Cont'd)

The investigators concluded that the nursing homes that adopted the GH model did not realize Medicare savings and suggested new approaches to align financial incentives may be required.

The investigators were not able to use randomization in their study design. Here they attempted to reduce bias and confounding by using a series of statistical approaches. First they identified a comparator group with similar organizational characteristics as the intervention group. Then they attempted to reduce bias and confounding at the patient-level by weighting the data according to a propensity scoring approach.

Reference: Grabowski, DC, Afendulis, CC, Caudry, DJ, O'Malley, AJ, & Kemper, P. The Impact of Green House Adoption on Medicare Spending and Utilization. Health Services Research, 2016 Feb 1;51(S1):433-53.

Key Strengths

- Not subject to ethical and practical constraints of randomization.
- Can be used in situations where it may not be possible to randomly assign the intervention.

Key Limitations

- The design is susceptible to bias and confounding which can be difficult or impossible to mitigate completely.
- Many factors may influence exposure to "intervention" or "control" group. These unobserved/unmeasured factors may be associated with the outcome of interest, leading to a biased estimate of the impact of the intervention.
- The reliance on existing data means that investigators may not have all needed data for the analyses.

Timeline and Cost Considerations

- Collecting data from comparator sites that are not part of the intervention will result in additional expenditure, unless the design uses existing, secondary data sources.
- The need to statistically adjust for bias is likely to increase the required sample size and thus the evaluation budget.

Policy Implications and Considerations for Future Use

Controlled before and after evaluation designs offer the potential to assess the impact of interventions or programs prospectively, when the option of randomization is not available. Care is required to select comparator groups to minimize confounding due to differences in case mix and wider contextual factors. Similarly, even when comparator groups appear to be well-matched, care will also be required when undertaking case mix adjustment of the outcome measures.

Quasi-experimental designs such as controlled before after designs are considered less robust than randomized studies by many in the health services research community. However, they can provide informative estimates of the impact of a program or initiative, and in turn, can provide valuable evidence to inform future policy decisions.

As with many designs, considerations for wider policy implications need to take account of how the intervention of interest interacted with a range of inner and outer contextual issues within the study setting, and how it may interact with the context of settings where it may be used.

References and Further Reading

Austin, PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 2011 May 31;46(3):399-424.

Eccles, M, Grimshaw, J, Campbell, M, & Ramsay, C. Research Designs for Studies Evaluating the Effectiveness of Change and Improvement Strategies. Quality and Safety in Health Care, 2003;12:47-52.

Iezzoni, LI. Risk Adjustment for Measuring Healthcare Outcomes, 4th ed. Health Administration Press, 1997.

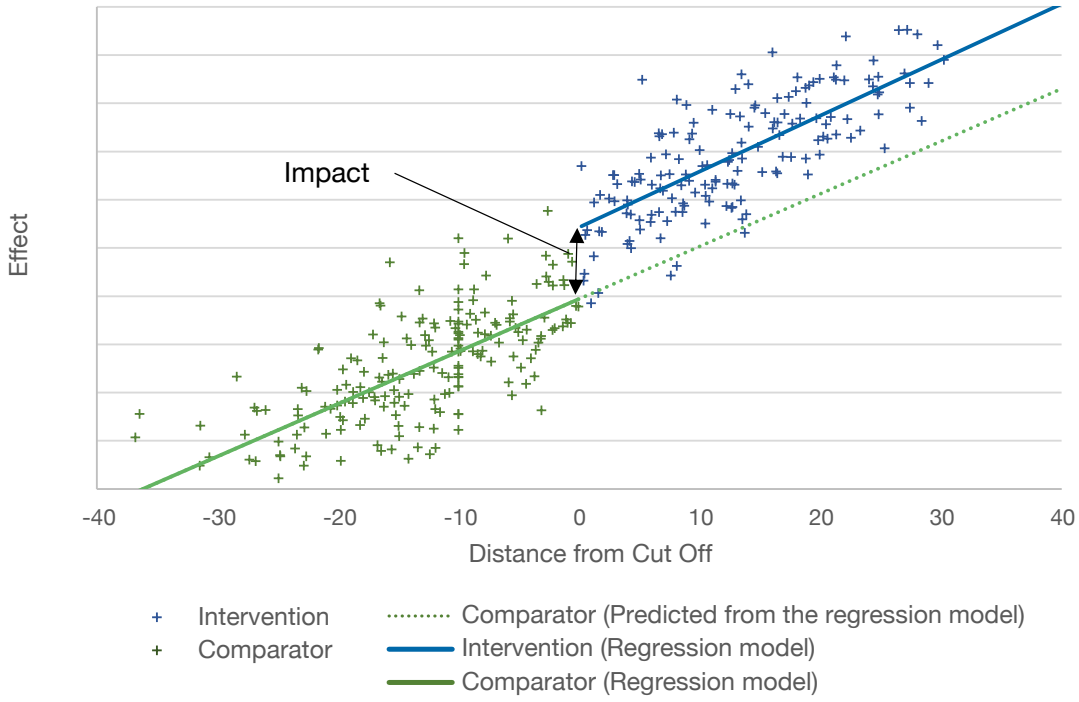
Regression Discontinuity Design

Regression discontinuity is a quasi-experimental design, increasingly used to assess policies or initiatives where inclusion in them is based on a cut-off in a continuous variable, for example the age of the person. The idea is that those people just below the cut off, and not included in the intervention, will be similar to those people just above the cut off, and included in the intervention. This level of similarity is considered good enough to allow those not included to act as a counter-factual to those included. Comparing subsequent outcomes between the two groups allows for an estimate of the relative impact of the policy or intervention (Figure 6).

Regression models statistically estimate the relationship between an outcome of interest and an independent variable holding other factors included in the model constant. At a high level, for an outcome of interest, regression techniques are used to model relationships in the data related to the variable used as a cut-off. The model is then used to predict likely outcomes above the cut-off. Differences in what was predicted by the regression model with the observed data from those just above the cut-off are used to provide an estimate of the impact of the policy or initiative.

Regression discontinuity was applied first in the 1960s by Thistlethwaite and Campbell to evaluate scholarship programs.

Figure 6. Illustration of a Generic Regression Discontinuity Design



Example 1. Effects of the Minimum Legal Drinking Age on Alcohol-Related Health Service Use in Hospital Settings in Ontario: A Regression–Discontinuity Approach

In Ontario, Canada, the minimum legal drinking age is 19 years. In this study, the investigators aimed to contribute to the policy evidence base by assessing the impact of the minimum legal drinking age (MLDA) legislation on alcohol related harm. They used a regression discontinuity design to compare youth just below the MLDA cut-off of 19 years, with youth just above this cut off.

They used data from all inpatient and emergency department events from 2002 to 2007, for patients aged 16 to 22 years. They selected patients presenting with a primary diagnosis of appendicitis as the control condition. To assess alcohol-related conditions, they used alcohol-use disorders, external injuries, suicides related to alcohol, suicides broadly defined, motor vehicle accidents and assault.

The investigators created age-groups based on the year and month of birth, to create a variable indicating months from the MLDA cut off of 19 years. For each outcome, the investigators applied a series of regression models using a variety of techniques to identify the best fitting model, including a variety of linear and non-linear approaches. The investigators applied what are increasingly becoming standard approaches to explore model-fit and sensitivity analysis.

Examining the coefficients of the regression models, the investigators found that compared to youth slightly below the MLDA, those just above had 10.8% (p=0.048) more alcohol-related inpatient and emergency department events and 51.8% (p=0.01) more alcohol-associated suicides.

The investigators concluded that young adults who gained legal access to alcohol used more hospital care for a variety of alcohol related issues. Moreover, they suggested the regression discontinuity design could be used by investigators to assess the impact of the minimum drinking age on additional health outcomes. In addition, they suggested the estimate of impact could also be used as the basis of a cost-benefit analysis, and by extension to cost-effectiveness analysis.

In this study, a population-wide policy had been enacted, which was specific to a particular sub-population. This resulted in no option to randomize individuals or sites, and no option to identify a contemporaneous independent comparator group. The regression discontinuity design allowed for the identification of a 'good enough' counterfactual. The regression analysis, although complex in execution, provided a means to compare observed outcomes against a reasonable predictor of outcomes.

Reference: Callaghan, RC, Sanches, M, Gatley, JM, & Cunningham, JK. Effects of the Minimum Legal Drinking Age on Alcohol-Related Health Service Use in Hospital Settings in Ontario: A Regression–Discontinuity Approach. American Journal of Public Health, 2013 Dec;103(12):2284-91.

Example 2. The Impact of Health Insurance for Children under Age 6 in Vietnam: A Regression Discontinuity Approach

In Vietnam, a national policy to provide health insurance to children under the age of 6 was introduced in 2005 with the aim of providing greater access to health care services. The investigators sought to assess the impact of this policy on health care utilization. They used a regression discontinuity design to compare health care utilization in children just above and just below the cut off of 6 years of age.

Using data from children aged 0 to 10 years from the Vietnam Household Living Standard Survey from 2006, 2008 and 2010, investigators examined a variety of utilization outcomes, including outpatient and inpatient visits.

The investigators assigned children to age in months, centered on 72 (6 years). For each measure of health care utilization, the investigators constructed a series of regression models.

The models also adjusted for a variety of demographic factors, including those related to household education and employment. The investigators also applied a series of sensitivity analyses to explore the robustness of their estimates.

From the resulting models, the investigators found that relative to children just above the cut off (not insured) those children slightly below the cut off (insured) had higher rates of inpatient visits (6.8%) and outpatient visits (21.7%).

The investigators concluded that the policy had led to increased health care utilization in children aged under 6 in Vietnam. Moreover, they suggested that public health insurance programs for children under age 6 may lead to improving service utilization in low- and middle-income countries.

Reference: Palmer, M, Mitra, S, Mont, D, & Groce, N. The Impact of Health Insurance for Children Under Age 6 in Vietnam: A Regression Discontinuity Approach. Social Science & Medicine, 2015 Nov 30;145:217-26.

Key Strengths

- Allows intervention assignment based on needs.
- In spite of a lack of randomization, can yield relatively robust estimates of effects when designed and analyzed properly, taking into account such threats to validity as cohort or period effects.
- The design avoids ethical issues related to allocation of participants to intervention or comparison groups.

Key Limitations

- Although the design makes intuitive sense, the statistical analysis is dependent on many assumptions that may not always be clear to stakeholders.
- An unbiased effect can only be obtained if the relationship between the cut-off variable and the outcome is correctly modelled. In particular, if the relationship is not linear, is the correct non-linear model used, for example, is the relationship best modelled using quadratic, log, exponential or another function?
- Less statistical power than randomized trials due to correlation between assignment and treatment variables.
- The design can be susceptible to contamination by other effects related to the same cut-off value. For example, in some settings the minimum alcohol drinking age could be similar to minimum driving age.

Timeline and Cost Considerations

- Frequently done using existing administrative or registry data, which require less resources than primary data collection, but may have other data quality issues such as completeness, accuracy and timeliness.
- The requirement for data to be available both prior to and after the intervention, often means the results are only available some time (often two years) after the intervention has been introduced.

Policy Implications and Considerations for Future Use

This design has particular value when policies targeted to a specific cut off in a variable of interest are being considered. The two examples provided above relate to age, however policies related to other demographic factors such as income, health status or screening results may be areas where a regression discontinuity design can be considered. The complex statistical analysis underpinning the study design means that data over a number of years (often two years) post implementation of a policy are likely to be required in order to achieve sufficient statistical power.

The design offers the potential to undertake the analysis in a single population, **which provides good internal validity**.

The design offers the potential to undertake the analysis in a single population, which provides good internal validity, but limited external validity. This means policymakers will need to consider carefully the likelihood the intervention will have similar impacts in other settings.

References and Further Reading

Imbens, GW, & Lemieux, T. Regression Discontinuity Designs: A Guide to Practice. Journal of Econometrics, 2008 Feb 29;142(2):615-35.

Thistlethwaite, D, & Campbell, D. Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment. Journal of Educational Psychology, 1960;51(6):309–317.

Observational Designs

Natural Experiment

Definitions of natural experiment vary in the literature. The commonality of the various definitions is that a natural experiment is a type of study in which an event or exposure was not planned or manipulated for the purposes of research or evaluation. Exposure to the intervention/policy/service occurred in natural circumstances and is outside the control of the investigator. Natural experiments may or may not include a control group. In the case of no control group, the outcome is assessed prior to and following the intervention/policy/service/other exposure (Figure 7). The outcome can also be assessed between two or more groups exposed to different interventions, policies, or services (Figure 8). Individuals, organizations, facilities, regions, districts, and countries where different interventions or policies have been implemented can be compared and analysis conducted using methods that attempt to make causal inferences.

This design is well suited for instances when:

- exposure to the intervention cannot be manipulated or assigned by the investigators;
- there is a need to understand the impact of large scale interventions and/or policies;
- there is a naturally occurring clearly defined prior exposure in a well-defined population.

A variety of statistical methods can be used to analyze data for natural experiments.

Figure 7. Illustration of a Generic Natural Experiment Design with no Control Group

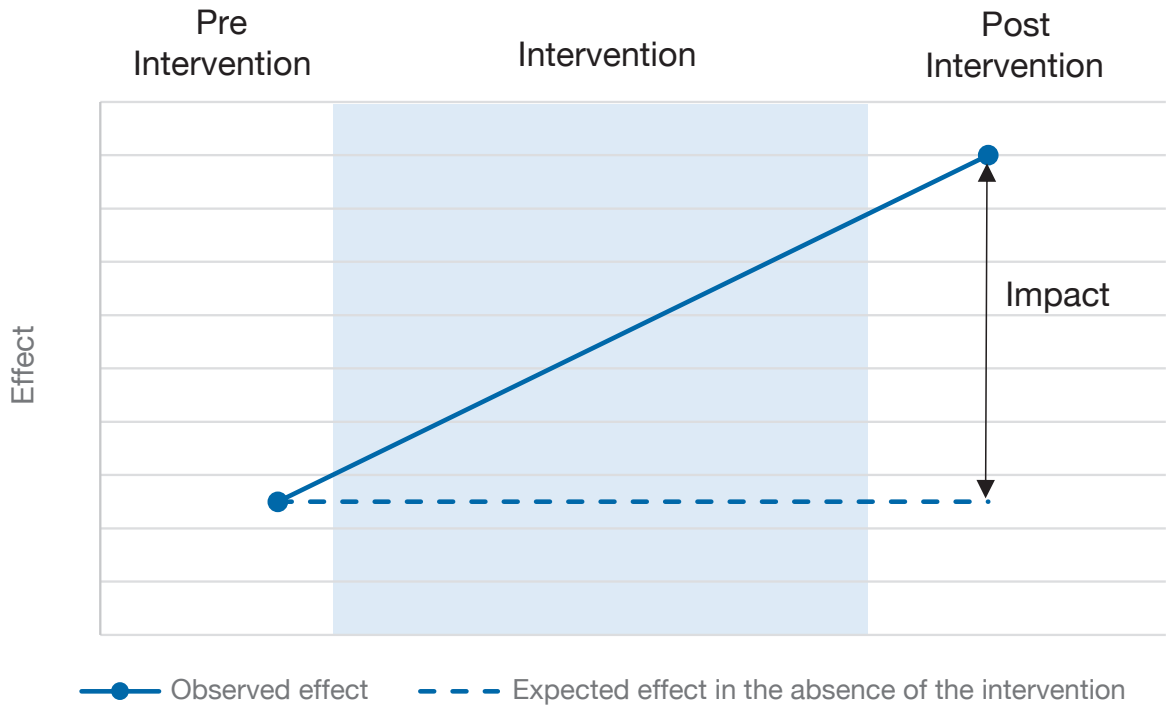
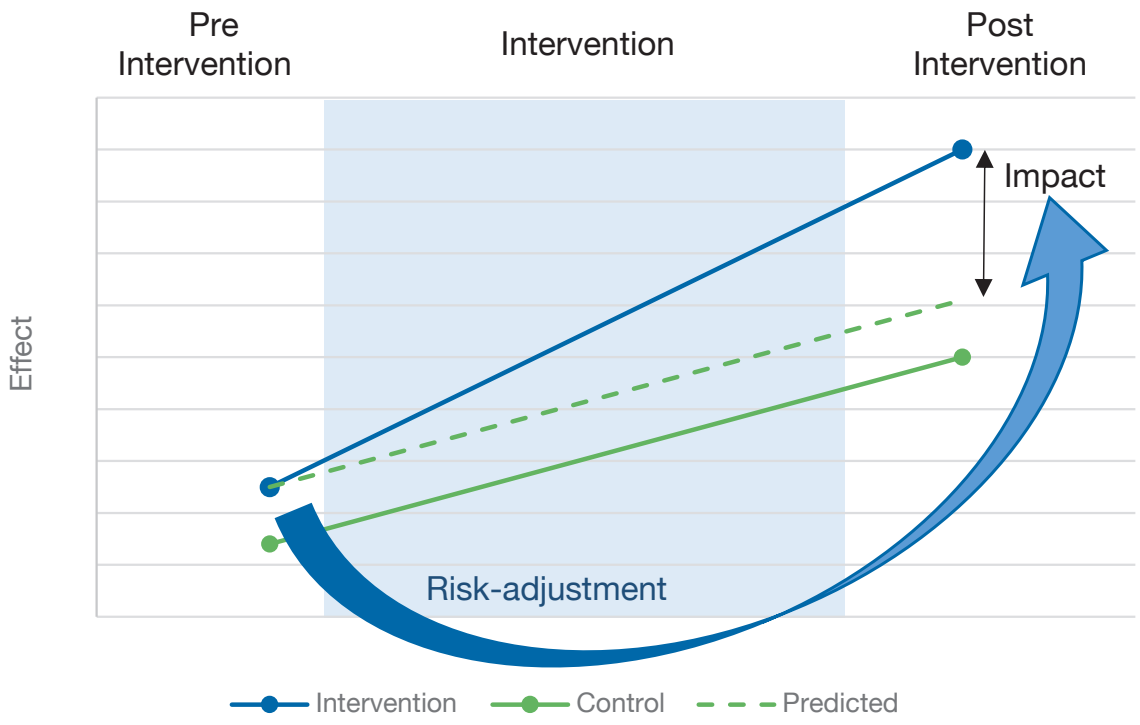


Figure 8. Illustration of a Generic Natural Experiment Design with Control Group



Example 1. Short Term Impact of Smoke-Free Legislation in England: Retrospective Analysis of Hospital Admissions for Myocardial Infarction

A natural experiment design with no control group was used to determine the impact of the introduction of smoke-free legislation in England on myocardial infarction hospital admissions. The smoke-free legislation was enacted on July 1, 2007. Hospital admission data were obtained from the National Health Service which provides routinely collected information on hospital episode statistics (HES). Data were obtained for all adults 18 or older living in England who were admitted in hospitals between July 1, 2002 and September 30, 2008 with a diagnosis of myocardial infarction. This period covers five years before the introduction of the legislation to 15 months following the introduction of the legislation. If the patient had multiple admissions for the same diagnosis within a 28-day period, only the first admission was counted. Regression analysis was used to determine whether emergency admissions for myocardial infraction changed after the introduction on the legislation, from July 2 onwards.

The investigators found a small reduction in the number of emergency admissions for myocardial infarction following the implementation of the legislation (-2.4%; 95% CI: -4.06%, -0.66%). This reduction, which varied by age and sex, translates to 1,200 less admissions in the first year. Admissions significantly decreased for men and women 60 and older but not for women younger than 60.

The authors concluded that these findings contribute to the evidence base for the effectiveness of smoke free legislation. The authors acknowledged, however, that the study suffers from important limitations such as the lack of availability of data on smoking. This means that it was not possible to determine how much of the reduction in admissions was due to a decrease in passive smoking because of the new law versus a decrease in active smoking.

This natural experiment with pre-and post-assessment is particularly adapted to evaluate the impact of the implementation of a nation-wide policy on health outcomes.

Reference: Sims, M, Maxwell, R, Bauld, L, & Gilmore, A. Short Term Impact of Smoke-Free Legislation in England: Retrospective Analysis of Hospital Admissions for Myocardial Infarction. BMJ, 2010;340:c2161.

Example 2. Medicaid Increases Emergency-Department Use: Evidence from Oregon's Health Insurance Experiment

A natural experimental design with a control group was used to determine the first-year effect of Medicaid coverage in Oregon on health care utilization and self-reported health. In 2008, Oregon implemented Medicaid expansion for its low-income and uninsured population. Due to limited resources, a lottery system was used to randomly select eligible individuals to apply for Medicaid coverage from the 89,824 people on the waiting list. Outcomes were compared between those who were selected by the lottery and those who were not selected to apply for Medicaid. Administrative data were extracted from hospital discharge records, credit report, and mortality records. Additional data were obtained from questionnaires mailed to all selected Medicaid recipients and the same number of non-selected individuals. Overall, 29,834 individuals were selected for the lottery and 45,088 were controls. Questionnaires were filled out by 29,589 in the treatment group and 28,816 in the control group.

Findings revealed that those selected by the lottery were 25% more likely to have insurance than those not selected in the first year after selection. Medicaid was associated with a statistically significant increase in health care utilization and improvements in self-reported mental and physical health, including a 10% increase in the probability of screening negative for depression compared to the control mean and 25% increase in the probability of reporting one's health as good, very good, or excellent. Medicaid was also associated with lower out-of-pocket expense and medical debt. Those in the Medicaid group were 10% less likely to have unpaid bills sent to collection than those in the control group ($p<0.0001$). The authors concluded that Medicaid is beneficial to this population and that this study provides important data for cost-benefit analysis for Medicaid expansion.

Although Medicaid coverage was randomly allocated, this was a natural experiment because Medicaid coverage was outside the investigators' control. The randomization process was not decided or designed by the evaluators. This was an implemented policy, the evaluators merely observed and analyzed the data. The use of a lottery system provided a unique opportunity for this natural experiment to mimic randomization and increase its potential for internal validity.

Reference: Taubman, SL, Allen, HL, Wright, BJ, Baicker, K, & Finkelstein, AN. Medicaid Increases Emergency-Department Use: Evidence from Oregon's Health Insurance Experiment. Science, 2014;343(6168):263-268.

Key Strengths

- Can be used in situations where it may be unethical or impractical to manipulate exposure.
- Not subject to the ethical and practical constraints of randomization.
- Convenient and often lower cost.
- Can evaluate the effect of an exposure on large populations.
- Can detect small effects due to the availability of data on large populations.
- Appropriate for rare outcomes or outcomes that take a long time to develop.
- Appropriate to evaluate policies and laws.
- Findings may be used to advocate for policies or changes in policies.

Key Limitations

- Natural experiments are observational studies. The investigators do not have control over the conditions of the study. Thus, this design is more susceptible to confounding and bias.
- Many factors may influence exposure to the "intervention" or "control" group. These unobserved/unmeasured factors may be associated with the outcome of interest, leading to a biased estimate of the impact of the intervention.
- Exposed group may be difficult to define or may change over time.
- Although investigators do not manipulate exposure, they still need to understand the process that leads to exposure versus non-exposure. To address these questions, qualitative and mixed methods may be used along with this design.
- The reliance on existing data means that the investigators may not have all needed data for the analysis.

Timeline and Cost Considerations

- Follow-up times depend on the outcome of interest.
- Can be conducted with low costs if secondary data sources are available, outcome is short to mid-term.

Policy Implications and Considerations for Future Use

Natural experiments provide a convenient and practical approach to evaluate laws, polices and other exposures beyond the control of

the investigators and which have been implemented in a real-world setting. However, because the conditions of exposure are not under the control of investigators, this design may be particularly prone to bias and confounding compared to quasi-experimental and experiment designs. This can be mitigated when there is a control group which is similar to the exposed group, and exposure mimics randomization as in RCTs but with lower costs. Policymakers can use findings from natural experiments to study the impacts of policies or policy changes and advocate for changes.

References and Further Reading

Craig, P, Cooper, C, Gunnell, D, Haw, S, Lawson, K, Macintyre, S et al. Using Natural Experiments to Evaluate Population Health Interventions: New Medical Research Council Guidance. Journal of Epidemiology and Community Health, 2012 Dec;66(12):1182-6.

Rubin, DR. For Objective Causal Inference, Design Trumps Analysis. Annals of Applied Statistics, 2008;2:808–40.

West, SG, Duan, N, Pequegnat, W, Gaist, P, Des Jarlais, DC, Holtgrave, D. et al. Alternatives to the Randomized Controlled Trial. American Journal of Public Health, 2008;98(8):1359–1366.

Glossary

Bias: Typically, there are three types of bias: 1) selection bias, 2) information bias and 3) bias due to confounding

Selection bias can occur when the group assigned to an intervention or to a comparison group is not representative of the intended populations. For example, when considering the impact of an intervention that aims to reduce blood pressure, intervention and comparator groups were established. However, in the course of the study, the intervention group may have systematically excluded older patients, whereas the comparator group included these patients. Any subsequent simple comparisons between these groups would have been subject to selection bias.

Information bias can occur when there are errors in the way measurements or data are collected. This especially matters if these errors differ between an intervention group and a comparator group. For example, when studying the impact of an intervention to reduce blood pressure, the method used to calculate blood pressure differs between the two groups.

Bias due to confounding can occur when the apparent impact of an intervention on an outcome is due to other factors. This may happen when a factor that has some impact on the outcome, varies between the intervention and comparator group. For example, the average age of those in the intervention group may be greater than those in the comparator group. Here, age may confound simple comparisons in the intervention and control group between comparisons of physical function, such as blood pressure. One can often control for confounding in the design of the evaluation or in the analysis.

Case-control design: The frequency of exposure to a given factor is retrospectively assessed and compared among individuals with the outcome of interest (cases) and those without the outcome of interest (controls). The odds ratio is calculated to determine whether there is an association between the exposure and outcome of interest.

Cohort design: A group of individuals exposed to a factor and a group who are not exposed to the risk factor are followed over time to assess the occurrence of an outcome of interest. The occurrence of the outcome in the exposed group is compared to that in the non-exposed group. The relative risk (incidence risk or incidence rate) is calculated to assess whether there is an association between the exposure and the outcome. Cohort studies can be **prospective** (participants are followed over time to determine the outcome) **or retrospective** (both the exposure and outcome have already occurred before participants were enrolled in the study).

Difference in difference analysis: analytical approach which addresses the fact that in non-randomized experiments, the intervention may not explain all the difference between the interventions and the control group in terms of key outcome of interest because the two groups did not start at the same level at baseline. In difference in differences analysis, the "normal" difference in the outcome of interest between the groups is calculated and the intervention effect is the difference between the observed outcome and the "normal" outcome. These estimates are derived from regression models.

Factorial designs: In a factorial study, two or more interventions and a control group are compared. Participants are randomized to each intervention independently. For instance, with a 2x2 factorial study, in the first randomization, participants are randomized to intervention 1 or to the control group and in the second randomization, the same participants are randomized to intervention 2. For a 2X2 factorial, this results in four groups: no intervention, intervention 1 only, intervention 2 only, intervention 1 and 2. The effect of each intervention and their interaction can be estimated.

Randomized encouragement trials: Type of RCT where individuals or groups are randomized to an intervention or control group. However, unlike a conventional RCT, participants are allowed to choose whether to receive the intervention. Participants assigned to the intervention group are encouraged to participate in the intervention or to select among specific intervention options. Controls are not offered the intervention.

Scale-up: No widely accepted definition "deliberate efforts to increase the impact of health service innovations successfully tested in pilot or experimental projects so as to benefit more people and to foster policy and program development on a lasting basis." (WHO ExpandNet Initiative definition)

Spread: Extend over a large area.

Staggered enrollment trials: Participants are randomized into an intervention or control group for a defined period of time. After this initial study period, control participants either transfer into the intervention group or are randomized a second time to either receiving the intervention or control. All participants will ultimately participate in the intervention.

Multiple baseline: The intervention is introduced to different subjects or settings at different points in time using one of three approaches: i. numerous components are included and components analysis used to determine which are the most effective; ii. components are added consecutively to the intervention until the desired effect is obtained, components may be substituted or modified if found not effective; iii. studying similar interventions at the same time in different settings.

About the Authors

Astou Coly, Ph.D., M.P.H, is senior improvement advisor with the USAID ASSIST Project at University Research Co., LLC. Gareth Parry, Ph.D., is senior scientist at the Institute for Healthcare Improvement.

Acknowledgements

The authors acknowledge the following individuals for their contributions to this report: Amanda Cash, Dr.P.H., M.P.H., senior advisor for Evaluation & Evidence, U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation; Kelly Devers, Ph.D., senior fellow, NORC at the University of Chicago; Don Goldmann, M.D., chief medical and scientific officer at the Institute for Healthcare Improvement; Michael Gluck, Ph.D., M.P.P., senior director of Evidence Generation and Translation at AcademyHealth; M. Rashad Massoud, M.D., M.P.H., FACP, director, USAID ASSIST Project and senior vice president at University Research Co., LLC; Rosalind Raine, Ph.D., FFPH, M.Sc., M.B.B.S., B.Sc., professor of health care evaluation, head of department of applied health research, UCL, and director of NIHR CLAHRC North Thames; and Lisa Simpson, M.B., B.Ch., M.P.H., FAAP, president and chief executive officer of AcademyHealth.

Suggested Citation

Coly, A., & Parry, G. "Evaluating Complex Health Interventions: A Guide to Rigorous Research Designs," AcademyHealth, June 2017.

Endnotes

1. Lauer, MS, & D'Agostino, RB Sr. The Randomized Registry Trial--The Next Disruptive Technology in Clinical Research? N Engl J Med, 2013 Oct 24;369(17):1579-81.

2. Damschroder, LJ, Aron, DC, Keith, RE, Kirsh, SR, Alexander, JA, & Lowery, JC. Fostering Implementation of Health Services Research Findings into Practice: A Consolidated Framework for Advancing Implementation Science. Implementation Science, 2009 Aug 7;4(1):50.

3. This design is also used in efficacy studies where an intervention is studied under ideal conditions with highly selected populations. This application is not in the scope of this guide.

AcademyHealth is a leading national organization serving the fields of health services and policy research and the professionals who produce and use this important work. Together with our members, we offer programs and services that support the development and use of rigorous, relevant, and timely evidence to increase the quality, accessibility, and value of health care, to reduce disparities, and to improve health. This report was made possible by generous support from the Robert Wood Johnson Foundation.