M50
MATHEMATICA
Policy Research

White PAPER

BY KARA CONTREARY, KATHARINE BRADLEY, AND SANDRA CHAO

Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations

June 2018

This white paper was prepared on behalf of the Centers for Medicare & Medicaid Services (CMS) as part of the national evaluation of Medicaid section 1115 demonstrations (contract number: HHSM-500-2010-00026I). In 2014, the Center for Medicaid and CHIP Services within CMS contracted with Mathematica Policy Research, Truven Health Analytics, and the Center for Health Care Strategies to conduct an independent national evaluation of the implementation and outcomes of Medicaid section 1115 demonstrations. As part of the evaluation, Mathematica provides technical assistance focused on states' demonstration evaluation designs and reports. This paper is a guide intended to support states and their evaluators by describing best practices in causal inference.

CONTENTS

TABLES

FIGURES

I. INTRODUCTION

Many states use section 1115 Medicaid demonstrations to implement eligibility and coverage reforms focused on non-disabled adults with incomes up to 133 percent of the federal poverty level (FPL). Recent reforms of this type include (1) use of features common in commercial health coverage, like required monthly payments or beneficiary health accounts, (2) premium assistance to help people buy plans through the federal Marketplace and encourage them to transition to commercial coverage, and (3) community engagement requirements that are intended to increase employment levels and promote transitions from Medicaid to employer-sponsored insurance.

These demonstrations must be rigorously evaluated to understand whether they achieved their goals and to realize their full value as policy experiments. As part of all section 1115 demonstrations, states are required to contract with independent evaluators to perform interim and final evaluations. The Centers for Medicare & Medicaid Services (CMS) supports states in the evaluation process by giving them tailored guidance on the required evaluation content and feedback on their developing evaluation plans. State Medicaid agencies still have many choices to make as they develop requests for evaluation proposals from potential contractors, guide the execution of evaluations, and interpret the evidence presented to them.

This guide, which uses examples from recent reforms for adult Medicaid beneficiaries, is intended to support demonstration states by describing best practices in causal inference. In this context, “causal inference” is the process of determining whether a demonstration policy (also called the treatment) is responsible for an observed outcome. Establishing an association between treatment and outcome variables is relatively straightforward, requiring only that they move reliably in the same or opposite directions. Establishing causation—that is, confidence that a change in treatment *caused* observed changes in outcomes—is much more difficult. Yet a primary goal of demonstration evaluations is to determine whether particular state Medicaid policies cause changes in outcomes such as health care access, utilization, and costs, and—in the case of some eligibility and coverage policies—the uptake of commercial coverage.

Although there are already many academic guides about causal inference, this guide is designed to be a concise reference for state Medicaid agencies and their evaluation contractors. It was informed by state-based evaluations of eligibility and coverage demonstrations, but much of

Section 1115 Medicaid Demonstrations

Medicaid is a health insurance program that serves low-income children, adults, individuals with disabilities, and seniors. Medicaid is administered by states and is jointly funded by states and the federal government. Within a framework established by federal statutes, regulations and guidance, states can choose how to design aspects of their Medicaid programs, such as benefit packages and provider reimbursement. Although federal guidelines may impose some uniformity across states, federal law also specifically authorizes experimentation by state Medicaid programs through section 1115 of the Social Security Act. Under section 1115 provisions, states may apply for federal permission to implement and test new approaches to administering Medicaid programs that depart from existing federal rules yet are consistent with the overall goals of the program, likely to meet the objectives of Medicaid, and budget neutral to the federal government.

the information presented here is also relevant to other types of section 1115 demonstrations.[1] In the following sections, we offer suggestions intended to help states create a framework for their evaluations by identifying outcomes, measures, hypotheses, and research questions (Section II), establish valid counterfactuals (Section III), check the feasibility of causal inference (Section IV), use a variety of research methods to interpret results (Section V), and check the robustness of findings (Section VI). Along the way, we offer examples relevant to eligibility and coverage demonstrations, note common pitfalls, and make practical suggestions for evaluations.

II. CREATE A FRAMEWORK FOR THE EVALUATION

This section is a step-by-step approach to planning a state evaluation. Careful planning positions states to make evaluation choices that support causal inference. This process begins with a clear statement of goals and ends with a set of research questions that inform all evaluation activities.

A. Articulate the goal of each demonstration policy

CMS and states typically work together to establish overarching policy goals for each section 1115 demonstration. These goals are important guideposts for both the implementation and evaluation of the demonstrations, but they might be quite general and not necessarily reflect all of the objectives of specific demonstration policies. For example, CMS and a state may set the overarching goals of improving beneficiaries' health outcomes and lowering the cost of their care. There could be two specific policies authorized in the demonstration's special terms and conditions: an incentive for beneficiaries to obtain preventive care, and a graduated co-payment structure for emergency department (ED) visits, with non-emergency visits having a higher co-payment than emergency visits. In this case, the first step in the evaluation process is to state that the goal of the beneficiary incentive is to increase the use of preventive care, ultimately improving health outcomes and lowering costs, and the goals of the graduated ED co-payment are to encourage appropriate use of care and also to lower costs.

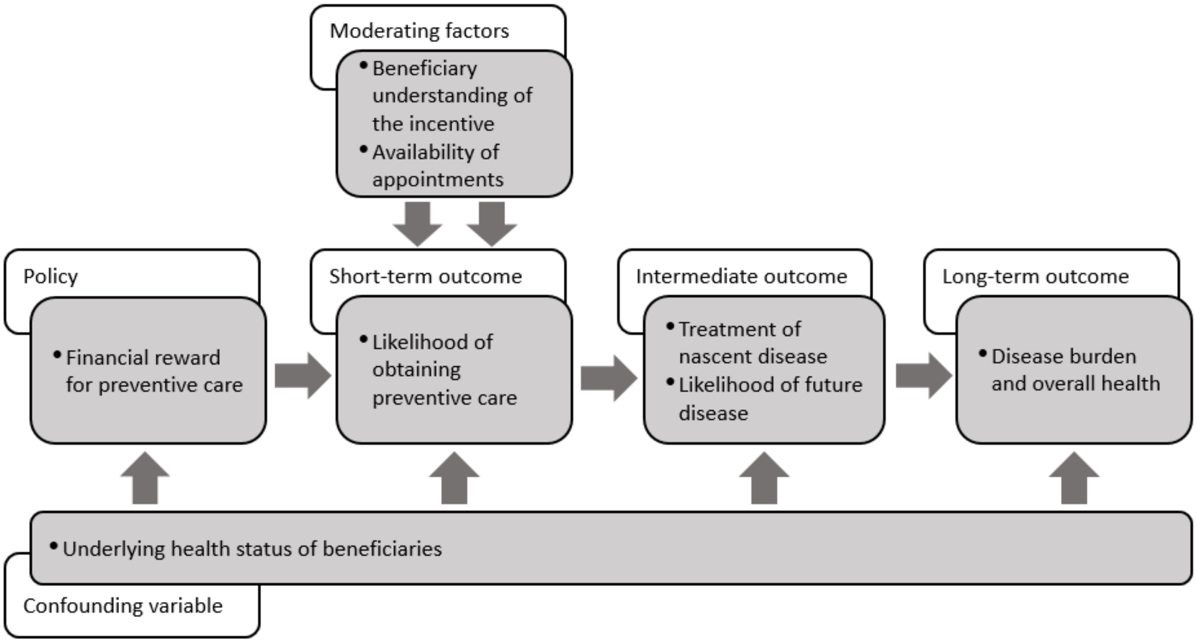
Section 1115 demonstrations often consist of multiple interventions intended to achieve different goals, such as increasing use of preventive care, familiarizing beneficiaries with the principles of private insurance, and lowering costs. Formally stating the goal(s) of each policy clarifies its expected contribution to the overall demonstration goals and informs the evaluation design, helping to ensure that evaluations will generate evidence about each specific policy. Generating evidence about each policy not only helps states understand how well each part of the demonstration is working, but it also informs the design of new demonstrations by other states.

[1] This guide is a companion to a related guide on selecting comparison groups for rigorous evaluations of all types of section 1115 demonstrations, by Reschovsky et al. (2018).

B. Use a logic model or driver diagram to identify outcomes and causal pathways

Logic models and driver diagrams help states and their evaluators identify each step in the causal pathway between a demonstration policy and its goal.[2] Logic models should depict short-term, intermediate, and long-term outcomes for each policy, as well as those factors that could moderate the relationship between the treatment and one or more outcomes. Identifying short-term and intermediate outcomes is especially valuable for assessing policies that may not have a measurable effect on outcomes for a long time, or when long-term outcomes are difficult to measure. States should also include potential confounding variables in their logic models—variables that are thought to influence both the treatment and the outcome and can bias evaluation results if they are not controlled for. In addition, if the policy could have unintended or adverse consequences, states should consider those consequences and include them as outcomes in the logic model. Figure II.1 is a simplified example of a logic model.

Figure II.1. Example logic model for a policy incentivizing beneficiaries to get preventive care


Moderating factors
• Beneficiary understanding of the incentive
• Availability of appointments
Policy
• Financial reward for preventive care
Short-term outcome
• Likelihood of obtaining preventive care
Intermediate outcome
• Treatment of nascent disease
• Likelihood of future disease
Long-term outcome
• Disease burden and overall health
• Underlying health status of beneficiaries
Confounding variable

C. Decide which outcomes to focus on for the evaluation

Which outcomes in Figure II.1 are measurable? In general, states should select outcomes that can plausibly change in the time frame under examination. Process measures such as receipt

[2] Both logic models and driver diagrams depict a theory of change that supports evaluation design. Driver diagrams typically focus on factors that must change in order to achieve a policy goal. The Centers for Medicare and Medicaid Innovation's Learning and Diffusion Group (Centers for Medicare & Medicaid Services 2013) provides a helpful description of the process for developing a driver diagram. "Logic model" is a more general term, and logic models have a less prescribed form; we use the term here to underline our focus on measurable policy outcomes. In practice, these diagrams can serve similar purposes in evaluation planning.

of care can be useful outcomes to study. In addition, some health outcomes that respond relatively quickly to treatment (such as diabetes control) may be promising outcome measures. Measures should be avoided if (1) changes the demonstration makes in beneficiaries' outcomes would likely occur after the demonstration's approval period ends, (2) such outcomes would be difficult to measure, or (3) the likely effect of confounding variables would be large. In the example shown in Figure II.1, measuring the long-term and intermediate outcomes would be challenging. A more feasible approach would be to focus on the short-term outcome: the likelihood that beneficiaries get preventive care.

D. Develop a hypothesis for each outcome

After identifying the outcomes for each policy, states and evaluators should develop a hypothesis about each demonstration policy's expected effect on the outcome(s) selected for the evaluation. The Special Terms and Conditions for each demonstration typically include hypotheses, but states might want to elaborate on or add to them. Hypotheses lend clarity to evaluations by articulating evaluators' plans for determining whether the policy is working as intended. Likewise, hypotheses guide interpretation of the eventual evaluation results, as well as the related assessment of whether policies are contributing to the overarching demonstration goals as planned. Using the example of Figure II.1, the expected effect of the incentive is that the likelihood of obtaining preventive care increases.

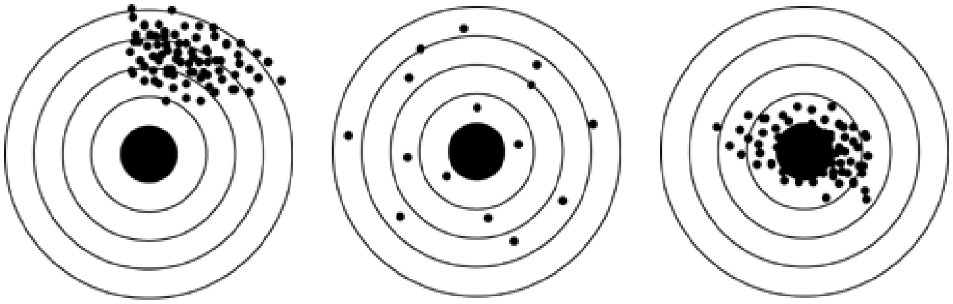
E. Consider data availability and choose valid, reliable measures for each outcome

Next, the state must choose one or more appropriate measures for the selected outcome. For example, looking again at Figure II.1, the state might choose to measure increases in use of preventive care by assessing receipt of wellness exams. In addition to choosing measures that can plausibly change within the demonstration period, states and evaluators should carefully consider data availability and the reliability and validity of selected measures.

Data availability is an important concern for evaluations of eligibility and coverage reforms, particularly for interventions that involve more than one administrative entity or have disenrollment penalties for beneficiaries who do not adhere to program requirements. If different entities (for example, state Medicaid programs, third-party administrators, other state agencies, and health plans) collect data on beneficiaries, it is important to be able to link those data sources to form a comprehensive sense of how beneficiaries respond to demonstration policies. If a demonstration disenrolls beneficiaries for noncompliance with demonstration requirements (for example, monthly payments or community engagement activities), it may be important to collect data on those who leave the sample to understand their long-term health, coverage, and employment outcomes. Given that health care utilization data are often difficult to obtain for non-Medicaid enrollees, keeping track of beneficiaries who exit Medicaid may require investing in longitudinal surveys.

Measure reliability (see Figure II.2) refers to how consistently a measure reflects the intended outcome. In the context of section 1115 demonstrations, states and evaluators should be particularly alert to whether measures are likely to work the same way (1) for different subgroups of beneficiaries within a demonstration group, (2) for demonstration and comparison groups, and (3) at different time periods. For example, questions on a beneficiary survey may not be reliable measures of self-reported health status or access to care, because different groups of beneficiaries—such as those who are medically frail and those who are not—may use different standards to evaluate their health, and their answers would therefore not be comparable. States using such measures could mitigate this problem by triangulating data sources, controlling for sources of variation, or conducting subgroup analyses.

Figure II.2. Difference between reliability and validity



Reliable but not valid Valid but not reliable Valid and reliable

Source: Columbia Center for New Media Teaching and Learning.

Validity, in this context "construct validity," means the degree to which a measure reflects the intended idea or state of the world.[3] A number of existing measure sets have already been tested and validated, making them good sources for measures of quality and access to care that have high construct validity (Box 1). Using relevant measures from these sources can save states the effort of developing their own quality measures, which may or may not have equivalent construct validity.

Box 1. Existing measure sets that have been tested and validated and are relevant for Medicaid populations

Medicaid Adult Core Set: https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html

Medicaid Child Core Set (relevant for 19- and 20-year old demonstration beneficiaries and pregnant women): https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/child-core-set/index.html

Healthcare Effectiveness Data and Information Set (HEDIS): http://www.ncqa.org/hedis-quality-measurement

Consumer Assessment of Healthcare Providers and Systems (CAHPS): https://www.ahrq.gov/cahps/index.html

Existing measure sets are unlikely to contain all the measures needed for state evaluations, however. States often use Medicaid claims and enrollment data, national household surveys, and state-specific beneficiary surveys to create other needed measures of enrollment, utilization, and access to care, and must take care that all such measures are both valid and reliable. State-specific beneficiary surveys can be critical data sources for section 1115 demonstration evaluations because they can offer measures of beneficiaries' understanding of and experience with demonstration-specific policies, but there are many potential pitfalls involved in the design of survey instruments (Box 2). For example, questions to probe a beneficiary's understanding should be written to allow clear interpretation of responses. The (hypothetical) question, "Are you aware that preventive services have no co-payment, and you may receive a reward if you use them?" will not yield a valid measure of beneficiary knowledge because there is no way to know

[3] Construct validity is sometimes discussed in terms of the absence of measurement bias.

whether a "yes" answer relates to the lack of co-payment, the reward, or both. States interested in fielding beneficiary surveys should choose evaluation contractors with experience in developing neutral, informative survey questions for beneficiaries. Supplementing beneficiary surveys with national survey data could expand the evaluation's evidence base, although national surveys are also subject to known weaknesses such as undercounting Medicaid enrollment.

Box 2. Common pitfalls in beneficiary surveys

Confusingly worded questions may produce measures that suggest limited understanding of demonstration policies, when in reality beneficiaries did not understand the questions.

Double-barreled questions ask about more than one aspect of a policy but only allow one answer, preventing accurate interpretation of responses.

Leading questions, or questions that may lead beneficiaries to avoid admitting ignorance about a policy, produce biased measures of beneficiaries' understanding and preferences.

Questions about sensitive topics such as risky health behaviors may lead beneficiaries to give socially acceptable responses, resulting in social desirability bias.

F. Articulate research questions

Finally, states and evaluators should articulate research questions that are related to each outcome measure. Answers to primary research questions address the hypotheses about the effects of the policy. However, policymakers are also interested in obtaining information to help them decide whether any observed effects on outcomes of interest are caused by the demonstration policy and to better understand the demonstration's impact. States should therefore also develop subsidiary research questions about mediating factors, subgroup effects, or other issues that will help them explore and address the primary question. Designating questions as primary or subsidiary can help an evaluator structure a group of related research questions. An example follows.

- *Hypothesis:* The incentive for preventive care will cause more demonstration beneficiaries to seek preventive care.
- *Primary research question:* Are beneficiaries with the incentive more likely to have wellness exams than other beneficiaries?
  - *Subsidiary research question:* Are there differences between key demographic subgroups in their likelihood of having wellness exams?
  - *Subsidiary research question:* Do beneficiaries understand the incentive?
  - *Subsidiary research question:* Is the receipt of wellness exams influenced by beneficiaries' access to primary care providers?

III. ESTABLISH THE COUNTERFACTUAL

After developing a set of research questions, the next step in the evaluation design process is planning the analytic approach to answering each question. Ideally, analytic approaches will establish causation, which requires states and evaluators to compare what actually happened to what *would have happened* in the absence of the demonstration—the latter is called the "counterfactual."

Consider again a beneficiary living in a state whose Medicaid program has an incentive for preventive services. If the beneficiary has a wellness visit, the state would like to determine whether the incentive played a role in her decision. If it were possible to observe her decision to receive a wellness visit in an environment that is identical in every way except for the incentive, we would know whether the incentive was the causal factor in her receipt of care.[4] The difference in use of preventive visits under these two regimes would be the "treatment effect." Unfortunately, it is impossible to observe the counterfactual for a given individual or set of individuals. Instead, evaluators must compare outcomes for two otherwise similar sets of beneficiaries who are and are not exposed to the intervention, with the latter group representing the counterfactual. In this section, we describe how different analytic methods and comparison group characteristics allow evaluators to establish counterfactuals in section 1115 demonstration evaluations.

A. The counterfactual in analytic methods

Analytic methods for causal inference can be grouped into two broad categories: experimental and non-experimental. Methods fall into one of the two categories based on (1) the amount of control the evaluator has over which beneficiaries receive treatment, and (2) the amount and type of data available for the evaluation.

The gold standard approach to establishing causality is a randomized controlled trial (RCT), an experimental design in which study participants are randomly assigned to either a treatment group or a control group. Because assignment is random, the only difference between the treatment group and the control group (the counterfactual) is the exposure to the demonstration, making it possible to infer that differences in outcomes were caused by the demonstration. Randomizing access to an entire demonstration may create uncertainty about which policies drive outcomes, however, so RCTs may be best suited to test different applications of a single demonstration policy, such as different incentive amounts. This ensures accurate identification of the policy lever that is influencing observed outcomes.

RCTs require robust data systems to support randomization and link to all included data sources. States may also have concerns about the appropriateness of randomizing benefits or incentives among a population of individuals who nominally have equal rights to benefits.

Because of the challenges associated with RCTs, evaluations of section 1115 demonstrations typically involve non-experimental methods,[5] which can support causal inference if they are conducted properly. These approaches involve identifying a comparison group of beneficiaries who are not subject to the demonstration, but are otherwise similar to the demonstration group. A key task of the evaluator is to select a comparison group that constitutes a valid counterfactual for the demonstration group (see Reschovsky et al. [2018] for an extensive discussion of comparison group selection).

[4] The impossibility of observing the treated and untreated outcomes for the same individual was called the "fundamental problem of causal inference" in Paul Holland's seminal 1986 paper. See http://zmjones.com/static/causal-inference/holland-jasa-1986.pdf.

[5] Some non-experimental causal inference methods, particularly those that attempt to replicate a treatment and control group comparison, are referred to as "quasi-experimental," although researchers are divided about precisely which methods fall under this label.

B. Characteristics of an ideal comparison group

As noted, an ideal comparison group represents what would have happened to beneficiaries in the demonstration group if they had never been exposed to the demonstration. The comparison group must therefore consist of individuals who are (1) similar to the demonstration group in their observable characteristics, (2) not exposed to the intervention, and (3) exposed to the same policy environment. In addition, evaluators must be able to calculate relevant outcome measures for the comparison group. We discuss each of these issues in turn.

- **Similar observable characteristics.** It is important to ensure that demonstration and comparison groups are as similar as possible in terms of health status, income, and other potentially confounding characteristics that can be observed. At a minimum, evaluators should check and report the balance between the two groups on members' observable characteristics, and may need to consider using matching techniques to improve balance. Matching and propensity score methods are increasingly common, and may be useful in defining a comparison group that is similar enough to the demonstration group.[6] If demonstration and comparison groups differ in meaningful ways (for example, if they are drawn from mutually exclusive income categories), statistical controls (see Section IV) may be inadequate as a solution. Evaluators should therefore be cautious when interpreting findings.
- **Unexposed to the intervention.** Comparison group beneficiaries who are exposed to the intervention no longer represent a counterfactual to the demonstration group. This phenomenon is called "spillover," "diffusion," or "contamination," depending on the circumstances. For example, diffusion might occur if providers change their practice patterns in response to a demonstration policy that incentivizes preventive care, and then apply the changes to their patients regardless of whether they are part of the demonstration group. An example of spillover would be a demonstration in which beneficiaries pass information about the demonstration to their friends and family members, potentially causing those friends and family members to change their behavior in similar ways to the demonstration group. In these cases, the comparison group's outcomes would be influenced in the same way that demonstration group outcomes were. As a result, evaluators might fail to find an effect of the policy even when one exists. Evaluators should consider the potential for spillovers in selecting comparison groups.
- **Exposed to reference policy environment.** For every evaluation, an important part of establishing the counterfactual involves defining the policy environment that beneficiaries would be exposed to if the demonstration were not taking place. For evaluations that are meant to discern whether outcomes under the demonstration differ from outcomes under standard Medicaid coverage, the comparison group must be covered under standard Medicaid. Comparing outcomes to those of a different type of section 1115 demonstration or to a group with no health insurance coverage may not yield the correct information about the demonstration's effect, although this depends on the specific research question. Furthermore, not only should members of the comparison group have coverage consistent

[6] For a summary and explanation of methods, see E.A. Stuart, "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, 2010, pp. 1–21. doi:10.1214/09-STS313.

with the counterfactual implicit in the research question, but they should resemble the demonstration group in terms of other policies and conditions that can influence the outcomes of interest. For example, comparison groups used in evaluations of community engagement policies should have unemployment rates comparable to those of the demonstration group.

- **Relevant measures can be calculated.** No matter what data evaluators use to calculate outcome measures, similar data must be available for the demonstration and comparison groups. Ideally, data from the same source(s) can be used, because different data sources can have different types and directions of bias. Comparing administrative data for one group to survey data for another group is unlikely to support reliable inference.

C. Common comparison groups in section 1115 evaluations

Table III.1 summarizes strengths, drawbacks, and strategies designed to help overcome the drawbacks for three common comparison groups in evaluations of eligibility and coverage demonstrations focused on non-disabled adults. (See Reschovsky et al. [2018] for a more detailed discussion of comparison group selection and how it relates to evaluation design for section 1115 demonstrations in general.)

Table III.1. Drawbacks of some comparison groups and potential approaches to overcoming them

Strengths	Common drawbacks	Suggested approach to overcome drawbacks
1. People in the same state who are not eligible for demonstration		
• Exposed to a similar environment (the state)	Potential spillover from demonstration	• Check for continuity of trends from pre-period to post-period—a departure from trend around the time of implementation may signal spillovers
	May be unlike demonstration group in terms of features like income, family structure, length of health insurance coverage	• Control for differences in observable characteristics • Use statistical matching strategies • If characteristics are non-overlapping, exercise caution in drawing any conclusions about the effect of the demonstration
2. Beneficiaries in other states		
• Similar on observable characteristics • Unlikely to be subject to spillovers	Are in a different policy environment	• Choose comparison states carefully—for example, by choosing states whose populations have similar demographic and economic characteristics and whose Medicaid policies before the demonstration were similar • Control for observable characteristics • Use statistical matching strategies
	Limited data availability	• Limit use of this comparison group to analyses based on data reliably collected from the same or similar sources, such as administrative data or national survey data
3. Non-Medicaid beneficiaries in the same state		
• Similar on observable characteristics (if using Medicaid-eligible individuals) • Exposed to a similar environment (the state)	Unmeasured characteristics associated with both decision to enroll and outcomes (selection bias)	• Limited remedies; can use matching techniques to achieve balance on observable characteristics, but unobservable characteristics still a concern
	Limited data availability	• Restrict use to analyses using national survey data

IV. CHECK FEASIBILITY OF CAUSAL INFERENCE

In addition to selecting a strong counterfactual, evaluators should verify that causal inference is possible and check for common threats to it.

A. Conditions for causal inference

Causal inference requires the following three conditions: (1) temporal precedence, (2) the association, and (3) elimination of confounding factors. We discuss each of these conditions in turn.

Temporal precedence requires the cause to come before the effect. In the context of section 1115 demonstrations, temporal precedence issues are most likely to occur when demonstration policies fail to take meaningful effect, calling into question whether a treatment could have affected the outcome of interest. If the policy is only partially implemented or the demonstration's requirements are not enforced, there can be a lack of meaningful effect. For example, if a demonstration includes a graduated co-payment structure for ED visits, it is important to establish that EDs are actually collecting co-payments. If not, then any changes in ED utilization are probably not related to the co-payment policy. As another example, if beneficiaries do not understand that financial incentives are involved in their taking or avoiding certain actions, it is difficult to argue that the beneficiaries considered the likely outcomes of their choices under the demonstration and then made a deliberate decision in response. Surveying beneficiaries to find out how well they understand program requirements and incentives is one way to check the plausibility of crediting the demonstration policies with changes in behavior.

Association requires that the treatment and outcome variables move reliably in the same or opposite directions. In other words, the policy change must be associated with the outcome.

Elimination of confounding factors ensures it is possible to see the true relationship between the treatment and outcome. As noted, matching methods and statistical controls may improve balance on observable characteristics. It is advisable to statistically control for confounding variables even if the demonstration and comparison group members were matched to ensure that any differences in outcomes would not be attributable to differences in those characteristics. Even for descriptive analyses, controlling for confounding factors such as age, gender, race, income, education, and health status will increase the usefulness of the analyses in understanding the effects of the intervention. In general, unadjusted rates are not valid measures of intermediate or long-term utilization outcomes, although they may be useful for monitoring changes in access to care over time and may be appropriate for answering subsidiary research questions. If the demonstration and comparison groups have non-overlapping characteristics, the analysis will not support causal inference. Often, unmeasured or unmeasurable characteristics such as motivation or concern about one's health are hypothesized to affect program participation or effectiveness. Controlling for observable differences cannot eliminate the threat that these variables will confound results, but it is normally assumed that doing so will reduce the potential for bias.

B. Common threats to causal inference

The conditions described above are necessary but not sufficient to establish causal inference. A number of other conditions can prevent evaluators from making valid determinations about whether an intervention caused an observed change in outcomes. Three of the most common threats to establishing causal inference are history, selection bias, and survivorship bias. States and their evaluators should be aware of these threats and develop plans to mitigate them. A detailed logic model may help an evaluator identify potential threats to validity.

History (also called confounding events) threatens causal inference when an event such as an unrelated policy change coincides with the timing of the demonstration policy and is expected to affect the outcome of interest. For example, a national public awareness campaign encouraging people to avoid using the ED for non-emergency care would be a confounding event in an evaluation of a graduated ED co-payment. In this case, it might be difficult to distinguish between the effects of the public awareness campaign and the effects of the co-payment. Evaluators may be able to reduce the history threat by using a comparison group exposed to the confounding event but not to the demonstration policy.

Selection bias occurs when the method for drawing individuals into the demonstration group is non-random. Evaluators should consider how individuals are being sorted into demonstration and comparison groups, and whether the selection mechanism might introduce confounding. If beneficiaries can choose to enroll in (or “self-select” into) a demonstration program, there will almost always be selection bias. For example, a program might allow beneficiaries to choose between two coverage options: (1) monthly payments and no point-of-service co-payments, and (2) point-of-service co-payments, but no monthly payments. Beneficiaries who believe they will use many health care services will be more likely to opt for the monthly payments, whereas those who expect to use few or none will be more likely to opt for the co-payments. This ability to self-select might make it difficult to assess the effects of either monthly payments or co-payments on beneficiaries’ behavior. External assignment can also introduce bias, especially if the state designates beneficiary populations who are expected to respond to the intervention in a particular way. An example of external assignment would be a state with a community engagement policy exempting certain individuals who are expected to have more difficulty complying with the requirement. Similarly, if decisions about where to implement a non-statewide policy are based on subjective judgments about the attributes of local areas (such as the availability of health providers), bias can result. Several methods exist to mitigate selection bias, although random assignment is the only one that completely removes it.

Survivorship bias (also called attrition bias) is closely related to selection bias. It happens when people drop out of an intervention in a non-random way. For example, healthy people may be more likely than sick people to drop insurance coverage if a monthly payment is required, because they derive fewer short-term benefits from the coverage. If the outcomes of people who drop out of the sample cannot be observed (because claims data are unavailable for individuals who disenroll from Medicaid), it is not possible to estimate the average treatment effect on all those exposed to the intervention. It may be possible to estimate the average treatment effect on the treated, or the average treatment effect among those who remain in the sample. The evaluator must decide if this estimate helps to address the research questions.

V. INTERPRET EVALUATION RESULTS

In this section, we discuss specific evaluation designs and assess the conditions under which they permit causal inference. We summarize several approaches used in evaluations of eligibility and coverage demonstrations that focus on non-disabled adults, and we highlight the limitations and common implementation pitfalls of each. These limitations should inform how states can work with evaluators to develop stronger evaluation designs, interpret the results generated through each method, and draw conclusions about demonstration policies.

Because most methods have limitations, it is prudent to use more than one method and outcome measure to examine each research question. If evaluators address a research question with a variety of methods and find consistent results, the conclusions will be more credible. Similarly, if a research question can be studied using more than one independent data source that produce similar findings, states and evaluators can have more confidence in the interpretation.

A. Experimental methods

If executed correctly, randomized controlled trials permit causal inference about the effects of demonstration policies. However, correctly executing a randomized controlled trial can be challenging (Nichol et al. 2010). Common pitfalls include the following:

- **Inappropriate application**. Not all outcomes lend themselves to a randomized controlled trial. Rare outcomes (such as hospitalization for specific conditions) or those that take a long time to develop (such as some cancers) are often better studied using other methods.
- **Inadequate randomization**. RCTs are only as reliable as their randomization methods. Simple randomization may be adequate with large sample sizes, but with smaller samples, it is often a good idea to use “block randomization,” which is designed to randomize beneficiaries into groups of roughly equal size. If a particular covariate (such as sex) is highly correlated with the outcome, or if the design involves subgroup analyses, it may be necessary to use stratified randomization to ensure balance. Various methods of randomization are available to evaluators, but in all cases it is crucial that randomization be correctly performed and documented.
- **Insufficient sample size**. Because RCTs are more expensive than other types of evaluations, there may be pressure to use the smallest possible sample. However, adequate statistical power is necessary to detect impacts, and it is advisable to be conservative when conducting power analyses (see Section VI). Insufficient sample size can also be a consequence of low recruitment levels for the sample, especially if beneficiaries must give their consent to be randomized and are reluctant to participate.
- **Failure to conduct intention-to-treat analysis**. In this type of analysis, all beneficiaries who are recruited for the intervention or enrolled in the trial are included in the analysis in the groups they were randomized into. This approach is intended to avoid bias that could result from non-random withdrawal (survivorship bias) or non-random failure to implement the intervention as intended. For example, demonstration group members who benefit from the treatment may be more likely to stay enrolled than demonstration group members who do not benefit. Failing to include the dropouts in the evaluation would systematically

exclude the beneficiaries who are least likely to benefit and would produce an overly optimistic assessment of the policy's impacts.

B. Non-experimental methods

In cases where randomization is not possible or desirable, non-experimental methods may be used, although they vary in quality and in their ability to generate causal impact estimates. First, we consider non-experimental methods when no pre-demonstration (baseline) data are available on participant characteristics and outcomes, as is often the case with demonstrations covering large groups of newly eligible beneficiaries. We then discuss how inference is affected if pre-demonstration data are available.

1. No pre-intervention data

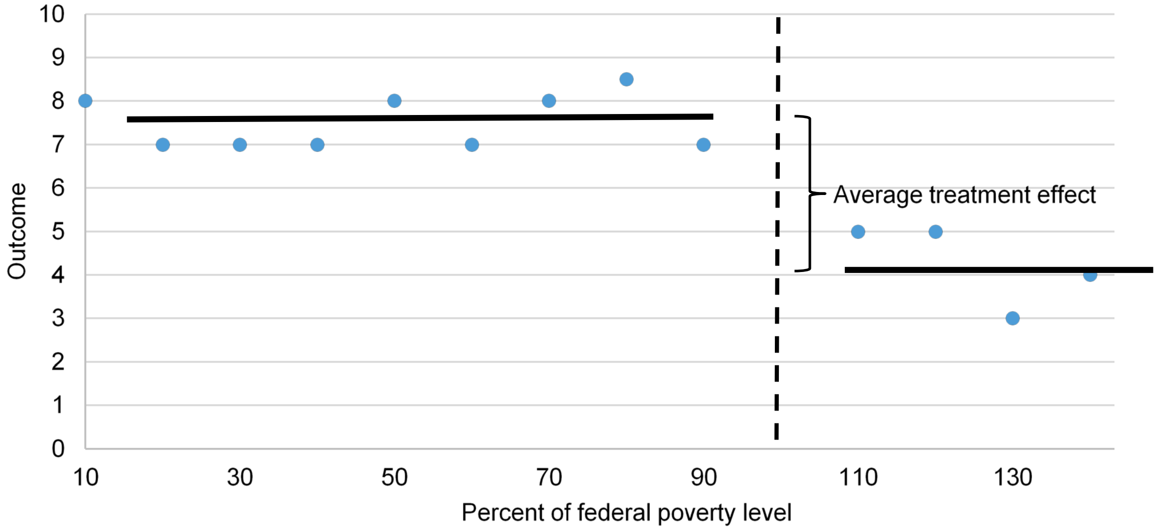
If no pre-intervention data are available, and no acceptable comparison group can be identified, the only feasible analysis is a **case study** (also called a one-group post-test–only design). Case studies are useful only for describing the characteristics of the treated group, and do not permit causal inference. Indeed, the lack of pre-intervention data means that the evaluator cannot even know whether outcomes changed before and after the demonstration program's implementation. However, descriptive approaches can be useful for demonstration monitoring and may support exploratory or subsidiary research questions.

The addition of a comparison group enables a **comparison of means** (also called post-test–only with non-equivalent comparison group). This design does not permit causal inference either, because it is impossible to tell whether any differences between outcomes in the demonstration and comparison groups are due to the intervention or due to differences between the two groups that existed before the intervention. The design is improved if observable characteristics are controlled for, but the presence of unobservable characteristics and the lack of information on pre-intervention outcomes limits the strength of the study.

Although an evaluation is seriously limited if there are no pre-intervention data, there is one special case in which a strong design can be used. If eligibility for a demonstration policy is determined by a threshold (such as income or age), then a **regression discontinuity** design is possible. For example, several section 1115 demonstrations use a threshold of 100 percent of the federal poverty level to distinguish beneficiaries who are subject to monthly payments or other requirements. The comparison group comprises individuals just below the threshold, and the treatment group comprises beneficiaries just above the threshold—who are by definition subject to the payment policy. Under the assumption that beneficiaries just above and below the threshold are similar in their unobserved characteristics, this design allows for a causal interpretation of the impact estimate.

To illustrate, in Figure V.1 the individuals above the threshold have, on average, a noticeably lower value on the outcome measure compared with individuals below the threshold. We can interpret the difference as the average treatment effect. If a second comparison group of individuals is available, and it includes both people above the threshold and people below it (all of whom are exempt from the policy) evaluators can use an even stronger design: **comparative regression discontinuity**.

Figure V.1. Regression discontinuity at eligibility threshold of 100% FPL



Common pitfalls and limitations of regression discontinuity are:

- **Poor external validity**. Regression discontinuity designs permit causal inference. Because the resulting impact estimate applies only to a small subset of the overall population (those just above and below the eligibility threshold), the estimates often have poor external validity for people whose values on the eligibility variable (for example, income) are far from the threshold. States should avoid attributing findings to beneficiaries whose values are far from those on the threshold.

- **Inadequate sample size**. The strength of the regression discontinuity approach derives from the similarity of beneficiaries just above and below the policy threshold. However, the number of beneficiaries close to the threshold could be limited. Comparing beneficiaries with incomes from 90 to 100 percent FPL to beneficiaries with incomes from 100 to 110 percent FPL allows evaluators to assess how the policy affected people who were similar in other respects, but the sample size would probably be small. Increasing the range to compare beneficiaries with incomes from 60 to 100 percent FPL to those with incomes from 100 to 140 percent FPL increases the sample size (resulting in more precise estimates), but calls the similarity of the two groups into question. This is a common issue in section 1115 demonstration evaluations that use an income threshold, because Medicaid beneficiaries with different income levels might be markedly different from each other. States must carefully define treatment and comparison groups to balance sample size and comparability.

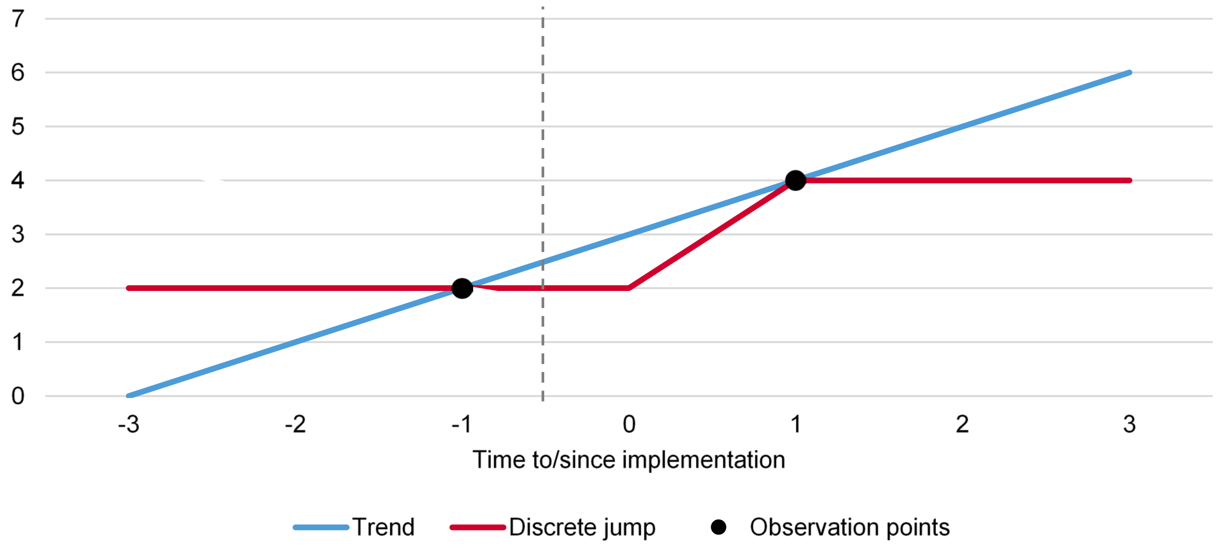
- **Potential for manipulation of the threshold**. In some cases, beneficiaries might be able to influence their exposure to a policy by manipulating their score on the eligibility variable—for example, by misreporting their income. Statistical checks for such manipulation (such as the McCrary test [2008]) are available and should be used, even if it seems unlikely that people would be able to "game" their treatment assignment.

2. Pre-intervention data are available

With pre-intervention data, evaluators can examine differences in outcomes before and after program implementation. As noted, pre-intervention data are rare in evaluations of eligibility and coverage demonstrations that apply to large groups of newly enrolled beneficiaries. National surveys may be a good source of pre-intervention data, but evaluators should consider carefully whether the survey population and available measures meet the needs of the evaluation.

If pre-demonstration data are available, but a suitable comparison group is not, the possibility for casual inference depends on how long the study population was observed before the start of the demonstration. If only one or just a few observations are available, the only feasible approach is a **one-group pre-test–post-test** design, which compares the outcomes for the study population before implementation and after implementation to see if they differ. This approach does not permit a causal interpretation. To see why, consider the two series in Figure V.2. If we only observed outcomes in time −1 and time 1, it would be impossible to distinguish whether the increase was part of a long-term trend (blue) or a discrete change at the time of implementation (red). Too many alternative explanations exist for any given observed pattern of outcomes, and without a comparison group, it is not possible to determine which explanations are valid.

Figure V.2. Pre-test–post-test designs do not permit causal inference

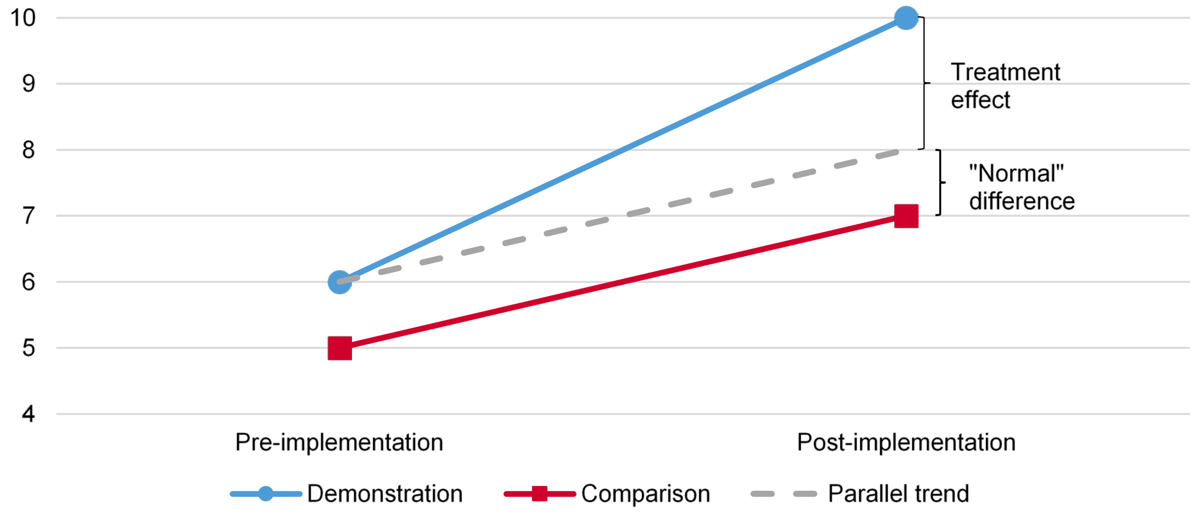


With enough observations before implementation, however, it may be possible to use an **interrupted time series** design, which may support causal interpretation. The extended pre-period allows the evaluator to check if there are indeed long-term trends that may explain a change in outcome from pre- to post-period. A common pitfall of this design is inappropriate application, such as when the treatment is introduced gradually, when strong seasonal effects exist, when pre-implementation trends are highly non-linear, or when the population being studied changes over time. In some cases, these issues can be addressed with the right modeling choices.

With both pre-implementation data and a comparison group, evaluators can both observe whether outcomes changed with the intervention and control for baseline differences between the demonstration and comparison group. They can therefore establish causality via a **non-equivalent comparison group** design. States can have more confidence in the evaluation findings if pre-intervention outcomes are similar for the demonstration and comparison groups (although they need not be identical), and if other characteristics of the two groups are similar.

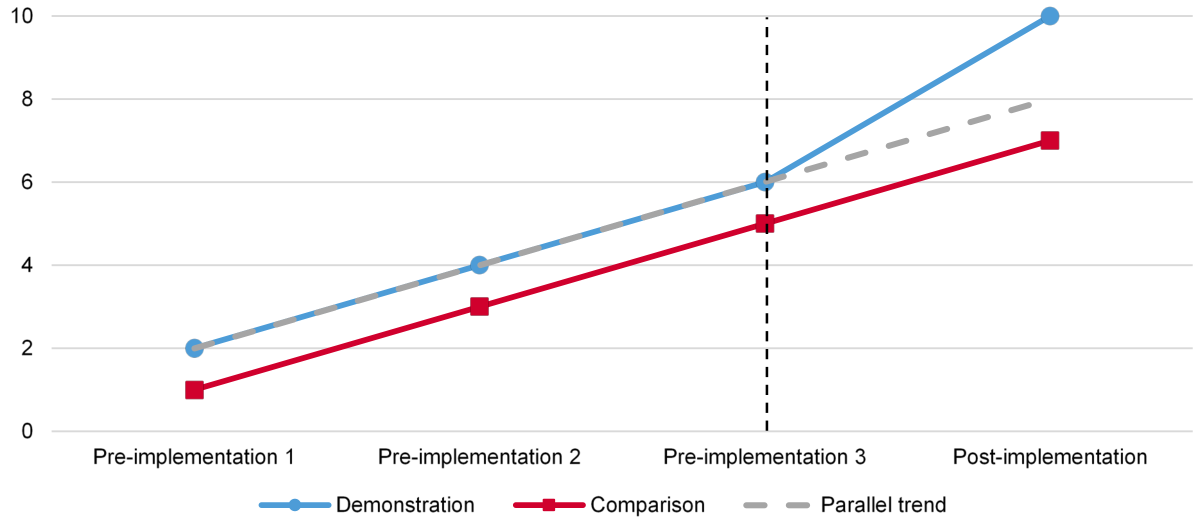
In program evaluation, the statistical technique that is used most often when both a comparison group and pre-intervention data are available is **difference-in-differences (DID)**. Under DID, the program impact is measured as the pre-post difference in an outcome for the demonstration group minus the pre-post difference for the comparison group. Figure V.3 illustrates DID methodology. The blue dots represent the observed pre- and post-implementation outcomes in the demonstration group, and the red squares represent the observed outcomes for the comparison group. Assuming *parallel trends*, the amount by which outcomes changed in the comparison group over time (from 5 to 7, or a difference of 2) is the amount by which outcomes in the demonstration group would have changed over time were it not for the demonstration (the gray dotted line). Given the differences in observed outcomes in the pre-demonstration period, a similar pre-post difference in the post-demonstration period would be considered normal. The additional difference between the demonstration and comparison groups, which is calculated as (10−6) − (7−5) = 2, and labeled as treatment effect, is attributable to the demonstration.

Figure V.3. Illustration of difference-in-differences



The parallel trends assumption of the difference-in-differences method can be tested if more than one pre-demonstration observation is available. Figure V.4 shows an example of data that exhibit parallel trends in the pre-intervention period.

Figure V.4. Parallel trends



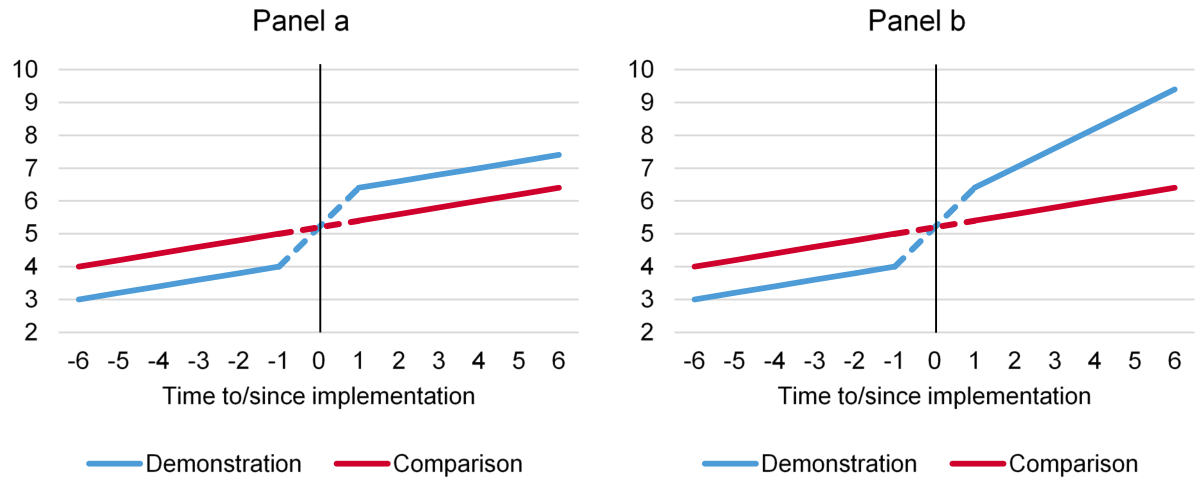
Short-, intermediate-, and longer-term impacts can be assessed using DID with additional post-implementation observations over time.

Common pitfalls in difference-in-differences designs include:

- **Correlation of demonstration exposure with the outcome at baseline**. The difference-in-differences approach relies on a person's exposure to the demonstration being "as good as random." If individuals are selected into the demonstration group because of their baseline values on the outcome measure, the DID approach is inappropriate and will yield biased results. Similarly, if the demonstration and comparison groups differ on some dimensions, this could result in bias from confounding.
- **Violations of the parallel trends assumption**. If multiple observations in the pre-implementation period are not available to verify the parallel trends assumption, evaluators should consider what factors might lead the demonstration and comparison groups to be on different trajectories even without the demonstration, such as simultaneous implementation of other policies or programs.
- **Shifting group composition**. If the DID approach is used with certain cross-sections of a population and repeated with the same cross-sections over time, evaluators should check that the characteristics of the demonstration and comparison groups remain stable. Otherwise, differences attributed to the demonstration may in fact be due to the different characteristics of the groups before and after the intervention.
- **Spillover effects**. Evaluators should consider whether people in the comparison group might be exposed to the demonstration in some form, which would result in a downward bias of any estimated impacts. This concern arises more often in designs using within-state comparison groups than in those using comparison groups from out of state.

If observations on both the demonstration and comparison groups are available for several time periods before and after implementation, a **comparative interrupted time series** design may be possible. Such a design is related to DID, but has the added benefit of testing whether the intervention changed just the level of the outcome (Figure V.5, Panel a), or if it also changed the long-term trend (Figure V.5, Panel b).

Figure V.5. Comparative interrupted time series



VI. MIND THE DETAILS

No matter what evaluation design is used, evaluators must make numerous decisions over the course of an evaluation. Some of them can seem small, but they can all affect the reliability and credibility of the findings. In this section, we suggest ways to increase the strength of an evaluation's evidence at various stages of analysis. Evaluators who address these issues can have greater confidence in their findings.

A. Before analysis: ensure sufficient statistical power, and consider Bayesian approaches

Power calculations should be done before collecting data to determine the likelihood that an evaluation will detect an effect when one is there to be detected. On the one hand, if a study is underpowered, statistical tests run a higher risk of false negatives, meaning they will fail to detect a real effect of a demonstration policy.[7] On the other hand, overpowered studies may produce statistically significant results for miniscule effects that are not meaningful for policy decisions. Overpowered studies can also waste an evaluation's resources by collecting data from a sample that is larger than it needs to be. Power calculations are especially important when data collection might be expensive and when it requires careful planning; beneficiary surveys are an example. States using beneficiary surveys or other primary data collection techniques should

[7] It is not, however, acceptable practice to run power calculations after completing analyses and then to use an underpowered study (that is, one with too few observations) as an explanation for failure to find a statistically significant result.

report power calculations in the evaluation plans they submit to CMS for approval, along with the minimum detectible effect sizes for key subgroups of interest.

In cases where lack of statistical power is a concern, states might consider **Bayesian estimation**. This approach incorporates information from other states' demonstrations or other relevant contexts to "add strength" to the estimates from the state's own demonstration. Demonstrations for which the interventions are selectively implemented (by geography, for example) may be good candidates for Bayesian approaches, as it is possible for the analysis to draw strength by including outcomes in untreated locations. In addition, states may find that Bayesian approaches generate more intuitive information about the impacts of the demonstration. Whereas the traditional (frequentist) approach produces estimates of the exact magnitude of the impact with an accompanying probability that the estimate was actually zero (the *p*-value), Bayesian approaches can tell the state what the probability was that the demonstration increased or decreased a particular outcome by more than a given amount (for example, a 50 percent probability that ED utilization decreased by at least 10 percent). However, this approach requires access to data on other demonstrations, it is complex to compute, and the evaluators must be experienced in Bayesian approaches (many are not).

B. During analysis: multiple comparisons and standard errors

Simultaneously conducting a number of statistical tests with different outcomes can create **multiple comparison problems** that lead evaluators to infer causality between demonstration policies and outcomes when it doesn't exist. Multiple comparisons using the same data source for the same population can result in statistically significant findings purely by chance. Take, for example, the case of a single hypothesis test intended to determine whether two groups are significantly different at the 5 percent significance level. This test has a 5 percent chance of finding a statistically significant result due to chance. With 20 comparisons (20 outcomes), the probability of finding a statistically significant result by chance increases to 64 percent. With 50 outcomes, the probability of finding a statistically result increases to 92 percent. To avoid multiple comparison problems, evaluators should pre-specify a minimum set of outcomes that are most useful for testing a single hypothesis. Evaluators can also use correction methods to decrease the likelihood of a false positive (Box 3).

Box 3. Statistical corrections to mitigate multiple comparison problems when using the same data source to generate multiple outcomes

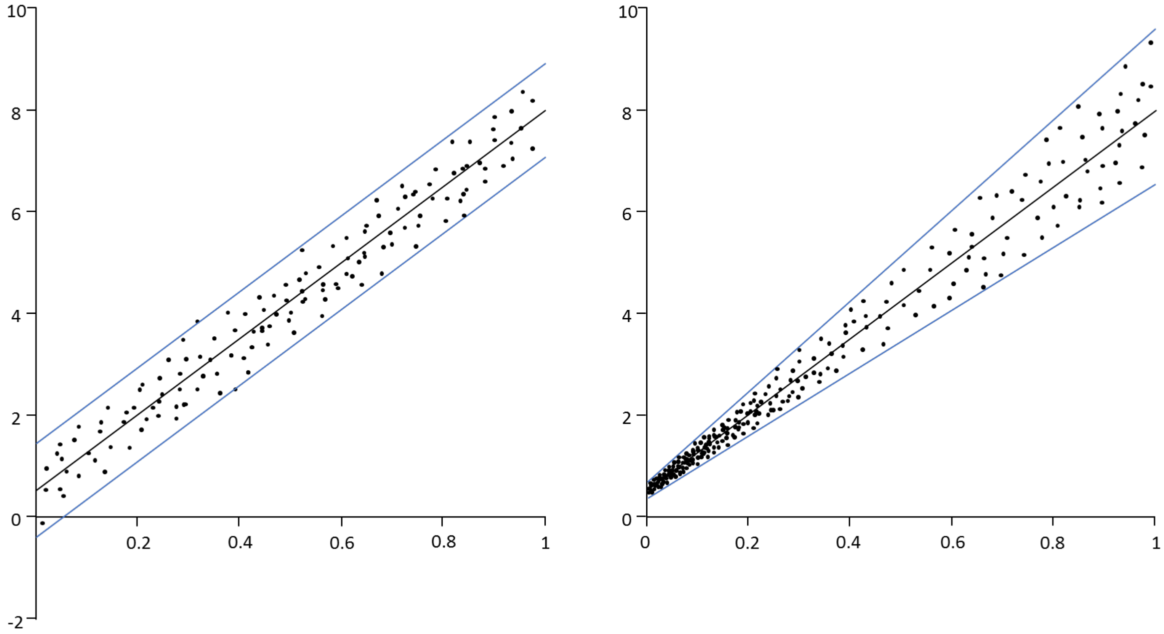
The *Bonferroni-Holm method* is a "step-down" method that controls for the familywise error rate, or the probability of seeing at least one false positive.

The *Benjamini-Hochberg method* is a "step-up" method that controls for the proportion of false positives among a set of significant results.

Standard errors measure the statistical accuracy, or precision, of an estimate. Standard errors are important in statistical analyses because common measures used in hypothesis testing—including confidence intervals, t-statistics, and *p*-values—all depend on them. However, because standard errors are themselves estimated from data, they can be incorrect if evaluators make unsound assumptions or use inappropriate methods to calculate them. Propensity score matching, for example, requires special estimation techniques to calculate unbiased standard errors. If the data being analyzed exhibit certain properties that are not accounted for in estimating standard errors, the estimates will be biased. Heteroscedasticity and clustering are common properties of data that can result in standard errors biased towards zero. These generate misleadingly narrow confidence intervals, large t-statistics, and low *p*-values, which in turn can lead evaluators to incorrectly conclude that the demonstration policies have an effect.

- **Heteroscedasticity** means that the amount of variation in a variable changes along the range of a second variable that predicts it (Figure VI.1). For example, variability in spending on medical services tends to increase with income. In this case, a regression estimate of the effect of income on medical services would be unbiased, but the standard errors would be too small—which could lead evaluators to make incorrect conclusions. Using "heteroscedasticity-robust" standard errors by default is good practice, because they do not cause bias in the absence of heteroscedasticity and improve inference in the presence of heteroscedasticity.

Figure VI.1. Bivariate models demonstrating homoscedasticity (left) and heteroscedasticity (right)



- **Clustering** in the data can also bias standard errors toward zero. Clustered observations are not independent of each other, or are grouped in some way. For example, the data are clustered when there are multiple observations on the same person over time or there are observations on individuals served by the same managed care organization. Correcting for clustering is important to avoid false positives. A rule of thumb is to adjust standard errors based on the lowest level of clustering. For instance, the standard error should be adjusted at the individual level if data are clustered by state, by managed care organization, and by individual.

C. Checking initial results: robustness checks, subgroup analyses, and placebo tests

Robustness checks are an important step in assessing the causality of demonstration policies because they give insight into whether results should be trusted and how generalizable they are. An analysis can be considered robust if evaluators are able to derive the same or similar results after changing their modeling approach by, for example, adding or removing covariates or outlier observations, changing inclusion or exclusion criteria for the study population, and testing alternative definitions of key variables. For instance, if the goal is to test a demonstration's effects on access to care, the state's evaluation might include a robustness test using an alternative definition of access to care. If the demonstration truly has an effect on access to care, then the effect should be present in models that use different measures of access to care (such as completion of a wellness visit and completion of any physician office visit). If estimates change significantly or in unpredictable ways, evaluators should be cautious about drawing conclusions about the effect of demonstration policies.

Subgroup analyses are advisable when the demonstration population comprises groups of people who might be expected to respond to the demonstration differently. These analyses can help evaluators assess the consistency of results for different groups of beneficiaries, different geographic areas in a state, or other constituent parts of a demonstration. However, subgroup analyses are also subject to problems arising from multiple comparisons and inadequate power, and, as noted in Section II of this guide, the measures must be valid and reliable within subgroups. Evaluators should keep these issues in mind to avoid reporting misleading subgroup effects.

A **placebo test**, or falsification test, is a replication of the analysis using an outcome variable that is not believed to be affected by the demonstration. A placebo test is another way of checking whether a model is reliably capturing the causal impact of the demonstration and nothing else. A result suggesting that a demonstration "affected" an outcome known to be unaffected is a signal that a study's estimated impact on the outcomes of interest should be viewed with suspicion. For example, an incentive to receive preventive cancer screenings would not be expected to change how often beneficiaries see their primary care physicians for flu symptoms. If an evaluation reveals that cancer screenings and visits for flu treatment change in similar fashion after the incentive is implemented, evaluators should not conclude that the incentive caused the increase in preventive screenings.

VII. CONCLUSIONS

Section 1115 demonstrations give states many opportunities to experiment with Medicaid coverage and eligibility reforms. Evaluations are a critical component of these policy experiments, enabling demonstration states, CMS, and other states considering demonstrations to understand how different program components influence outcomes such as access, utilization, and cost. It is therefore important for evaluations to employ sound research methodology that permits causal inference and enhances knowledge of the likely effects of future policies. As more and more states seek section 1115 demonstration authority for these reforms, the information in this guide can support states' searches for evaluation contractors, their collaboration on research design with evaluators and with CMS, and their interpretation of findings to inform demonstration improvements. Overall, generating high quality evidence supports best practices in Medicaid policy and a deeper understanding of this complex program.

REFERENCES

Centers for Medicare & Medicaid Services, Center for Medicare and Medicaid Innovation Learning and Diffusion Group. “Defining and Using Aims and Drivers for Improvement, A How-to Guide.” 2013. Available at https://innovation.cms.gov/files/x/hciatwoaimsdrvrs.pdf. Accessed April 24, 2018.

Columbia Center for New Media Teaching and Learning. “Measurement – Validity and Reliability.” Available at http://ccnmtl.columbia.edu/projects/qmss/measurement/validity_and_reliability.html. Accessed April 2, 2018.

Holland, P.W. “Statistics and Causal Inference.” *Journal of the American Statistical Association*, vol. 81, no. 396, 1986, pp. 945–960.

McCrary, J. “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test.” *Journal of Econometrics*, vol. 142, no. 2, February 2008, pp. 698–714. Available at https://www.sciencedirect.com/science/article/pii/S0304407607001133?via%3Dihub. Accessed April 24, 2018.

Nichol, A.D., M. Bailey, and D.J. Cooper. “Challenging Issues in Randomized Controlled Trials.” *Injury: International Journal of the Care of the Injured*, July 2010, Supplement 1: S20–S23. Available at http://www.injuryjournal.com/article/S0020-1383(10)00233-0/fulltext. Accessed April 24, 2018.

Reschovsky, J.D., J. Heeringa, and M. Colby. “Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations.” Report submitted to the Centers for Medicare & Medicaid Services, June 5, 2018.

Stuart, E.A. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, vol. 25, no. 1, 2010, pp. 1–21. doi:10.1214/09-STS313.

This page has been left blank for double-sided copying.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■
SEATTLE, WA ■ TUCSON, AZ ■ WASHINGTON, DC ■ WOODLAWN, MD