MATHEMATICA Policy Research

White PAPER

BY JAMES D. RESCHOVSKY, JESSICA HEERINGA, AND MAGGIE COLBY

Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations

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 intended to support states and their evaluators in selecting the most appropriate comparison group and evaluation design.

CONTENTS

FIGURES

I. INTRODUCTION AND PURPOSE

Section 1115 demonstrations provide flexibility to states to design and test specific policy approaches to promote the objectives of Medicaid and better serve their Medicaid populations.[1] The Affordable Care Act strengthened federal requirements for evaluations of section 1115 demonstrations, requiring that states use a range of evaluation strategies with the approval of the Centers for Medicare & Medicaid Services (CMS) (Musumeci et al. 2018).[2] Thus, states' section 1115 demonstration evaluations can include a blend of quantitative impact evaluations and qualitative information about the implementation of the demonstration that provides useful context for the impact findings.

This guidance document focuses on quantitative impact evaluations, which assess the causal effects of an intervention by comparing outcomes under the demonstration's policies with an estimate of what would have happened under a counterfactual—that is, what would have happened in the absence of those policies or if the policies had been implemented differently. These evaluations typically address the following types of questions:

- How did the demonstration affect beneficiary coverage, cost, quality, or access to care?
- How did the demonstration affect providers and how they treat beneficiaries?
- Were there any unintended effects?
- Did the policy change have differential effects on different beneficiary populations or, for a given population, under different circumstances (for example, high versus low unemployment)?
- Do the demonstration's impacts increase over time? In a renewed demonstration, are past gains being maintained?

In designing evaluations to address these questions, evaluators face the challenge of determining how to isolate the impact of the intervention on an outcome from other factors that could influence that outcome. The validity of an evaluation—the extent to which we can appropriately attribute changes in outcomes to the policy intervention—is a central focus of evaluation design. Different evaluation designs are subject to different types of threats to validity (see the text box on threats to validity on page 17). Moreover, many of the steps in conducting an evaluation, from measurement of variables to specification of statistical models, can influence an evaluation's validity.

The gold standard for impact evaluations is experimental studies, often referred to as randomized control trials. Individuals are randomly selected to either receive or not receive the intervention, forming what are termed treatment and control groups, respectively. Random assignment seeks to ensure that these two groups are nearly identical with respect to factors that may influence the outcome being studied. As a result, any difference in mean outcomes between

[1] For more information about the role of section 1115 demonstrations in promoting Medicaid objectives, please see: https://www.medicaid.gov/medicaid/section-1115-demo/about-1115/index.html.

[2] For more information, please see 42 CFR 431.424.

the treatment and control groups after the policy has been implemented can be attributed to the effects of the intervention.[3] However, experimental evaluations are often impractical. By their very nature, they are *ex ante* evaluations; that is, they must be planned before the intervention being studied is implemented. Moreover, randomly assigning who gets and does not get the intervention can be controversial, and experimental evaluation designs can often be expensive.[4]

Evaluation timing vis-à-vis intervention design and implementation

Preferably, evaluations should be designed before demonstration implementation to permit the broadest set of evaluation options. These "*ex ante evaluations"* may involve random assignment, staged implementation of the intervention, or primary data collection of baseline values that would typically be infeasible if the demonstration has already been implemented.

Alternatively, *ex post evaluations* are planned after the design, and sometimes after the implementation, of the demonstration. *Ex post* evaluation designs can often be rigorous—particularly when administrative data are used to obtain pre- and post-implementation information on both the treatment group and a credible comparison group.

Although experimental designs should be used when practical, states and their evaluators most frequently use quasi-experimental designs to evaluate section 1115 demonstrations as these designs are more feasible to implement. Quasi-experimental designs are observational studies that identify a comparison group that did not receive the intervention and is as similar as possible to the treatment group in terms of baseline (pre-intervention) characteristics.[5] When data availability severely constrains options, states sometimes use nonexperimental designs. These designs do not include a unique comparison group and are therefore inferior to both experimental and quasi-experimental designs as they do not incorporate a credible counterfactual. They are also subject to a broader set of threats to validity.

This guidance document focuses on comparison group selection for quasi-experimental designs. It is intended to help states that are developing their evaluation designs identify the best evaluation designs and comparison groups, given the state context. In Section II, we describe key activities to perform before selecting comparison groups and evaluation designs; in Section III, we present comparison group options and discuss key considerations for selecting comparison groups and evaluation designs; and in Section IV, we provide a brief overview of statistical and other methods that are needed to support and draw appropriate inferences from the comparisons. To highlight various types of comparison groups, we draw on examples from approved section 1115 demonstration evaluation design plans and reports.

[3] While experimental designs are least likely to suffer from internal threats to validity, they are not totally immune to bias. For instance, differential attrition among members of the treatment and control groups caused by the intervention could threaten the validity of results.

[4] Oregon used an experimental design to evaluate the effects of an 1115 demonstration. In 2008, Oregon wanted to expand eligibility for the Oregon Health Plan, but lacked the funds to fully insure the targeted expansion population. Thus, the state created a lottery by which individuals who applied were randomly selected to receive coverage through the plan. Doing so allowed experimental studies that assessed the impacts of expanding Medicaid coverage to the target population, using those applicants who lost the lottery as the control group (Finkelstein et al. 2012).

[5] Typically, the term "control group" refers to untreated individuals in experimental studies, while "comparison group" describes untreated individuals in quasi-experimental and nonexperimental designs.

II. FOCUSING THE EVALUATION THROUGH LOGIC MODELS AND EVALUATION QUESTIONS

There are a variety of section 1115 demonstration types, including eligibility and coverage alternatives, healthy behavior incentives, and benefit changes for nondisabled adults; managed care expansions and mandatory enrollment; interventions targeted at populations with special needs, such as populations needing long-term services and supports or behavioral health care services; and delivery system and payment reform demonstrations (Hinton et al. 2017). Each demonstration type adopts policy interventions to influence targeted outcomes. They also raise unique policy and research questions, not only about whether the intervention achieved its objectives, but also about how best to target the intervention, create circumstances that foster intervention effectiveness, and avoid unintended consequences.

Section 1115 demonstrations frequently include multiple interventions, which, in turn, may affect different beneficiary populations or different provider groups. Each intervention may be hypothesized to affect different types of outcomes, which can often be measured using different data sources. As a result, state evaluation designs often specify multiple evaluation research methods and draw upon multiple data sources. For instance, if a given intervention is hypothesized to affect quality of care, some quality metrics might be obtained from claims or encounter data, while others are assessed through beneficiary surveys such as the Consumer Assessment of Healthcare Providers and Systems. These different data sources may involve different beneficiary samples and also require different evaluation designs and analytic methods. In this section, we discuss the key preparatory steps that states, as well as their evaluators, should take to gain a better understanding of how the intervention and other factors may affect key outcomes and to target the evaluation and selection of comparison groups on the most critical or high-priority policy or research questions.

A. Developing logic models

New Hampshire's Building Capacity for Transformation Demonstration Evaluation. In the evaluation design for its Delivery System Reform Incentive Payment (DSRIP) demonstration, New Hampshire outlined DSRIP activities and short-, intermediate, and long- term outcomes. Building from the logic model, the state planned data collection and analyses to assess the link between DSRIP activities and outcomes (New Hampshire Department of Health and Human Services 2017).

An important first step in designing an evaluation is to develop a logic model, which visually depicts the theory of change or mechanisms by which the demonstration intervention is thought to achieve its targeted outcomes. Although other terms such as "driver diagrams" are used, we use the generic term "logic models" here. To develop a logic model, the state or its evaluator should have a firm understanding—informed by past research or grounded theory—of how the intervention intends to achieve its targeted outcomes. However, evaluators should develop models that go beyond showing only the direct causal links between the intervention and key outcomes.[6] Logic models

[6] See the guidance document from the Centers for Medicare and Medicaid Innovation Learning and Diffusion Group (Centers for Medicare & Medicaid Services 2013) for a description of the process for developing a driver diagram and Weiss (1998) for a discussion of developing a program theory of change to support evaluation design. Renger and Titcomb (2002) provide a useful example of developing a logic model for program evaluation.

should be able to help the evaluator identify: (1) short-term, intermediate, and long-term outcomes that might be measured; (2) mediating factors that influence the ability of the strategies to impact the outcomes,[7] and (3) potential confounding variables that are correlated with both the intervention and outcome and which may bias evaluation results if not controlled for. By identifying potential confounding variables, logic models will help inform whether potential comparison groups are sufficiently similar to the treatment group to support a good, unbiased evaluation design and assist in selecting the statistical methods by which comparison groups can be made more similar to the population subject to the demonstration (that is, the treatment group). These additional factors might include beneficiary, provider, managed care organization (MCO), or community characteristics, as well as macroeconomic, policy, and regulatory changes. The mediating factors identified in the logic model should also include factors that that may be difficult to measure, such as patient motivation and engagement. Identifying these extraneous factors will aid evaluators in choosing the best design, guiding data collection, developing statistical controls, and understanding limitations of their evaluations.

B. Focusing the evaluation through research questions and hypotheses

Developing research or policy questions to guide the evaluation. Informed by the program logic model, the state or evaluator should focus the evaluation through the specification of research questions and hypotheses.[8] Because section 1115 demonstrations frequently involve multiple components that may affect various beneficiary (and at times provider) populations, and each component may affect various outcomes, the state or evaluator should articulate overarching research questions and then outline specific, targeted research questions that address specific components. Each of these component research questions can then form the basis for a part of the evaluation design, targeted data collection, and analysis. Section 1115 demonstrations frequently renew and amend ongoing demonstration efforts. Generally, CMS is interested in evaluations that focus on the new components of the demonstration, provided that the original demonstration (and earlier amendments) has been evaluated or is in the process of being evaluated.

Identifying the right counterfactual. Impact evaluations compare outcomes of the group receiving an intervention with what would have occurred absent the intervention or under a different intervention. This alternative state defines the counterfactual. Evaluations require a counterfactual in order to attribute observed outcomes to the intervention. The appropriate counterfactual should be informed by the key policy questions of the evaluation. For instance, if a section 1115 demonstration is testing how the introduction of premiums for certain beneficiary groups affects these groups' access to care, quality, and cost outcomes, then the most appropriate counterfactual may be a similar beneficiary group within the state that is not responsible for paying premiums for their Medicaid coverage. The state could also compare beneficiary groups with different premium responsibilities if the demonstration varies the amount or timing of

[7] In driver diagrams, these factors are "secondary drivers" that influence the primary drivers or strategies used to influence change. For more information about driver diagrams, see https://innovation.cms.gov/files/x/hciatwoaimsdrvrs.pdf.

[8] Rossi, Freeman, and Lipsey (2003) provide a more detailed discussion about formulating evaluation questions.

premium requirements across beneficiaries or geographic areas (perhaps in a staged rollout of the intervention).

Sometimes, it may also be helpful to compare how the state's demonstration affected outcomes compared with other states that implemented similar interventions. For example, a state that implements a new managed long-term services and supports (MLTSS) program for its disabled beneficiary population would logically choose the counterfactual of continued coverage of this beneficiary population using FFS (perhaps using a comparison group of disabled beneficiaries in the state not covered by the MLTSS). However, it might also compare outcomes for beneficiaries under its MLTSS program to outcomes for similar groups of disabled beneficiaries in other states that are using MLTSS. The first counterfactual implies an evaluation that assesses whether the move from FFS long-term services and supports to MLTSS affected outcomes for the impacted beneficiary population, whereas the second informs whether the implementation of the demonstration was as effective as compared to MLTSS programs in other states. In other situations, a state may wish to compare how their demonstration affected outcomes in comparison with other states that attempted to achieve the same policy goal, but using a different approach.

Special considerations for demonstrations likely to affect enrollment

If the demonstration seeks to expand eligibility to new populations or implement new policies that are likely to reduce enrollment, states should consider primary data collection strategies prior to demonstration implementation. For eligibility expansions, it may be helpful to conduct primary data collection for the population likely to be newly affected by the demonstration, for example, through a survey of uninsured individuals. For policies likely to result in some beneficiaries losing their Medicaid eligibility, baseline beneficiary surveys that can be repeated after the policy has been implemented may be the best approach. Alternatively, states should consider using beneficiary observations from other states that did not implement similar policies for a comparison group.

Beyond the choice of a counterfactual, several factors influence comparison group selection and evaluation design decisions. To a significant degree, the choice of comparison group and evaluation design rests on the availability of data. That said, there may be times when the evaluator has multiple options for constructing a comparison group to support a given design. Each option should be assessed in terms of whether the data would support a strong design. To the extent feasible, it is best to triangulate evaluation results by using multiple evaluation designs/comparison groups to address a research question, as discussed in greater depth in Section IV.

Most frequently, treatment and comparison groups are collections of Medicaid beneficiaries, although at times, the comparison group might include similar patients who are not Medicaid beneficiaries. Indeed, some section 1115 demonstration interventions target nonbeneficiaries who are likely to become beneficiaries (for example, low-income pregnant women). Additionally, there are section 1115 demonstrations that focus interventions on providers. Although provider-based interventions will most often be assessed on the basis of impacts on their patients, evaluations may include a comparison group of other providers who are not subject to the intervention (and whose patients would be members of a patient-level comparison group). For simplicity, we will hereafter refer to treatment and comparison groups of beneficiaries in this guidance document, but readers should recognize that occasionally there will be circumstances where other group definitions are appropriate.

III. KEY CONSIDERATIONS FOR SELECTING COMPARISON GROUPS AND EVALUATION DESIGNS

This section describes some of the most common quasi-experimental and nonexperimental designs that could be applied to different demonstration policies, illustrates how comparison group selection and evaluation designs go hand-in-hand, and presents key considerations that guide these choices.[9]

States and evaluators must balance multiple considerations as they design their evaluations. The state health system context within which section 1115 demonstrations are implemented may create challenges and opportunities for comparison group selection; for example, the intervention may affect all beneficiaries enrolled in managed care, leaving only FFS beneficiaries as a potential in-state comparison group. In many cases, both the selection of evaluation design and comparison group are constrained by available data sources. In addition, section 1115 demonstrations may introduce multiple policy interventions concurrently and these interventions may apply to varied beneficiary groups. Thus, evaluators may need to incorporate more than one design and comparison group to adequately address all the high-priority research questions relevant for a demonstration.

Appendix A contains a flowchart that focuses on quasi-experimental designs and poses a series of questions to help guide the identification of potential comparison groups and related evaluation designs. Our discussion in this section is framed around the same questions posed in this flowchart. The first set of questions focuses on identifying comparison groups among beneficiaries who are not eligible for the demonstration. The second set focuses on identifying a subset of beneficiaries who are subject to (or eligible for) the intervention to serve as the comparison group. Finally, we briefly describe the types of nonexperimental evaluation designs that may be used when a comparison group is not feasible. States should consider all of the questions to identify all feasible options and to guide selection of the strongest designs.

A. Comparison group options using beneficiaries not eligible for the intervention

1. Does a threshold value determine eligibility for treatment (for example, income or disability) with data available for individuals on both sides of the threshold?

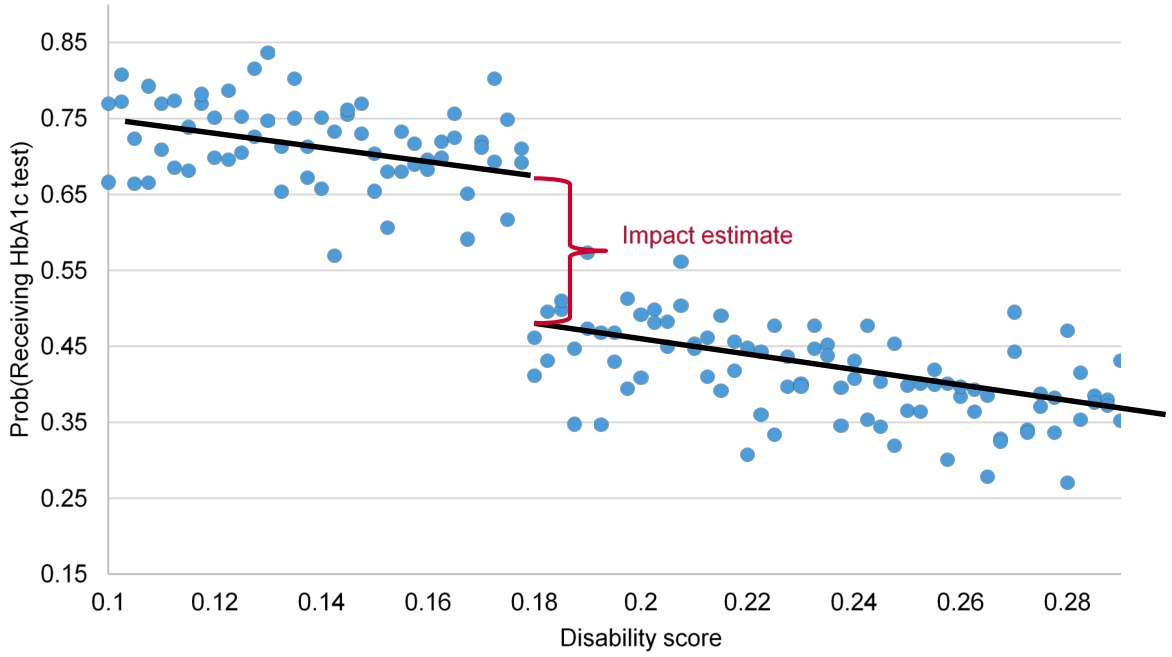
Section 1115 demonstrations often target an intervention to individuals on the basis of a scalar measure, that is, a measure for which eligibility for the intervention is determined by a cutoff or threshold, such as income or a disability severity score. The individuals just below the threshold are similar to the eligible individuals just above the threshold and therefore may constitute a viable comparison group.

[9] Readers should refer to one of the many resources on evaluation design for a more thorough discussion of design options and their relative merits and drawbacks. For instance, Campbell, Stanley, and Gage (1963) outline various experimental and quasi-experimental designs and make methodological recommendations that are broadly applicable to a number of social science applications. Other sources include Rossi, Freeman, and Lipsey (2003); Langbein (1980); and Weiss (1998).

In these cases, evaluators may use a **regression discontinuity (RD) design.** We illustrate this graphically in Figure 1 using the example of the RD design outlined in the Arkansas Works demonstration evaluation design (Arkansas Center for Health Improvement 2017). Because the intervention was applied to beneficiaries with disability scores below a certain threshold (0.18), the design essentially compares outcomes between those just above and just below this disability risk-score cutoff. Figure 1 presents the probability of receiving the hemoglobin HbA1c test for individuals along a range of disability scores. The heavy black lines are the regression lines, and the 0.2 discontinuity in the probability of receiving the test at the disability score threshold of 0.l8 represents the impact estimate.

Arkansas Works (formerly Health Care Independence Program) Demonstration Evaluation. Arkansas expanded Medicaid eligibility in 2014 to childless adults and parents with incomes up to 138 percent of the federal poverty level. Beneficiaries with a demonstrated level of frailty—as determined by a scale based on survey questions answered by beneficiaries—obtained coverage under traditional FFS Medicaid. Below the threshold, newly eligible beneficiaries obtained coverage through qualified health plans (QHPs) offered through the state's health insurance exchange. Given that a threshold score determined eligibility, the state's evaluation employs a regression discontinuity design to evaluate outcomes (including access to care, use of preventive services, and continuity of care) for beneficiaries enrolled in QHPs relative to a comparison group composed of newly eligible individuals with FFS coverage (the counterfactual) (Arkansas Center for Health Improvement 2017).

Figure 1. Example of regression discontinuity design



Source: Example based on Arkansas Center for Health Improvement (2017); data points are illustrative only and not based on actual data.

The RD design is generally thought to be a strong design. Of note, this design does not require pre-intervention observations, as most quasi-experimental designs do. The addition of a comparison group, composed of individuals who fall both above and below the eligibility

threshold, would make the evaluation stronger in what is called a **comparative regression discontinuity design**. One important limitation of the RD design is that it provides an impact estimate relevant only for those close to the threshold that determined eligibility for the intervention; it may not be appropriate to extrapolate this estimate to those who do not fall close to the threshold.

2. Are there beneficiaries who are not subject to the intervention but are in other eligibility groups with similar characteristics?

Demonstration provisions may apply only to some Medicaid eligibility groups. When this targeting is unrelated to characteristics likely to affect outcomes (such as health status), evaluators may want to consider beneficiaries in other eligibility groups as a comparison group.

O_1	X	O_2
O_1		O_2

Such a comparison group could support a **nonequivalent control group design** (also referred to as the **untreated control group design with pretest and posttest).** This design permits causal inferences, as it includes a treatment and comparison group and a pre-intervention or baseline observation (denoted as O_1 in the figure) and post-intervention observation (O_2) for both groups. Similar characteristics and levels of pre-intervention outcomes for the treatment and comparison groups can give the evaluator some confidence about the adequacy of the comparison group. Impact estimates are based on the change in outcomes pre- and post-implementation ($O_2 - O_1$) among members of the treatment group in comparison to the corresponding change among members of the comparison group. A key consideration is that the post-intervention observation, O_2, should be made only when the evaluator is confident that the intervention has had an effect. If made too early, the intervention may appear ineffective when eventually it does have an effect on beneficiaries. Although we illustrate this design with only one pre- and post-intervention observation in the accompanying figure, it is always preferable to have multiple observations over time, especially in the post-implementation period.[10]

When considering this approach, evaluators must assess how similar the comparison group will be to those subject to the intervention. Different eligibility groups can differ markedly with respect to income, disability, or other factors. A good comparison group should have substantial overlap with the treatment group in terms of the important characteristics likely to affect program outcomes. Informed by the logic model, the evaluator needs to exercise judgment regarding the key characteristics that should be used to identify similar individuals. In Section IV, we will describe how statistical matching techniques can help ensure some equivalency between the groups. Simply using regression analysis to control for covariates without matching may not be sufficient to avoid bias in impact estimates.

[10] As discussed below, if there are multiple observations both before and after the intervention took place for the treatment group and a comparison group, this similar—and very strong—design is called the comparative interrupted time series.

3. Was the intervention implemented only in certain parts of the state or otherwise less than statewide?

At times, section 1115 demonstrations are not implemented statewide to all members of the target population. Rather, states may implement new initiatives in certain geographic areas, involve only beneficiaries enrolled in only some MCOs, or in other ways limit which beneficiaries are subject to the demonstration. These decisions may reflect a phased implementation strategy or the state's judgement that certain interventions are more feasible in certain areas (for example, urban areas) or for certain beneficiary groups than others. Under these circumstances, beneficiaries who were not affected by the intervention may serve as a potential comparison group.

California's Medi-Cal 2020 (formerly the Bridge to Reform) Demonstration Evaluation. Under the Bridge to Reform demonstration, the state transitioned its seniors and persons with disabilities (SPD) population into the managed care delivery system in a subset of the state's counties operating specific plan models (Two-Plan and Geographic Managed Care) between 2011 and 2012**.** Under its Medi-Cal 2020 demonstration, renewed in 2015, the state seeks to evaluate the impact of mandatory managed care enrollment for the SPD population in these counties on beneficiary satisfaction, access to care, costs of care, and quality outcomes. To evaluate intervention effects, the state planned to draw potential comparison groups from counties in which SPD beneficiaries were not mandatorily enrolled in managed care or were enrolled in managed care through an alternative existing approach—the county-operated health system model (California Department of Health Care Services 2017).

There are some cautions about using a comparison group of this type, however. Evaluators will need to fully understand the reasons behind the decisions to implement the intervention in a targeted fashion. Of concern is whether these decisions were based on perceptions of beneficiary need or on factors that are related to how effective the intervention implementation might be. For instance, if a demonstration were implemented only in those parts of the state where it was thought that the performance of the local health care system was better than that of other local systems, and thus the capacity to implement the demonstration was greater, then treatment-comparison group differences could reflect these underlying health system characteristics and not exclusively the impact of the demonstration. In a similar vein, if the demonstration were limited to urban areas, then the evaluator would need to be cautious about assuming that rural and urban beneficiaries would respond to the intervention similarly. Evaluators may be able to mitigate the influence of confounding contextual differences to the extent that differences are measurable and can be statistically controlled for. However, the evaluator should be sensitive to unmeasurable factors. For this type of comparison group, evaluators may use a comparative time series design (described below) or nonequivalent control group design, depending on data availability.

4. Can beneficiaries with similar characteristics from other states or nationally be identified?

If an acceptable comparison group cannot be identified from within the state, or if it is preferable to use beneficiaries from other states because their circumstances represent the desired counterfactual condition not present in the demonstration state, a comparison group may be constructed from beneficiaries in other states or nationally. However, state Medicaid programs differ considerably in terms of eligibility requirements, benefits offered, delivery systems, and implementation contexts. Moreover, low-income populations may differ in important ways relevant to the evaluation across states. To ameliorate some of the challenges with using comparison groups drawn from other states, evaluators should use statistical approaches, such as

propensity score matching to ensure treatment and comparison groups are as similar as possible. External comparison groups can also be used to complement analyses with in-state comparison groups, recognizing that each approach has its limitations and strengths. Generally, cross-state data and comparisons would be used in nonequivalent control group designs.

Montana Health and Economic Livelihood Partnership (HELP) Demonstration Evaluation. The HELP demonstration extends Medicaid eligibility to parents and childless adults with incomes up to 133 percent of the federal poverty level who receive services through a third-party administrator (unless they meet certain exemptions). The evaluator intends to use three national surveys (American Community Survey, Behavioral Risk Factor Surveillance System, and Current Population Survey) to identify pre- and post-implementation outcomes (such as health insurance coverage and access to care) for the treatment group and comparison groups from communities and states with comparable Medicaid populations identified in the three data sources (Social & Scientific Services, Inc. 2017).

Data sources for information on beneficiaries from other states

Data sources that support individual-level matching. There are several sources of information on beneficiaries from other states. Most notably, national-level enrollment, claims, and encounter files for Medicaid beneficiaries from the Transformed Medicaid Statistical Information System (T-MSIS) are slated to become available for all states in 2018. These data enable detailed matching of beneficiaries, drawn from comparison states, to the treatment group. Second, states that collect data on patients using surveys can consider replicating questions that are found in regularly repeated national surveys such as the American Community Survey (ACS) and the Behavioral Risk Factor Surveillance System (BRFSS). The ASC and BRFSS are most likely to have samples of sufficient size to support state-to-state comparisons. Alternatively, other national surveys such as the Current Population Survey (CPS), the National Health and Nutrition Examination Survey (NHANES), and the Medical Expenditure Panel Survey (MEPS) have small state-specific subsamples. As such, these surveys are unlikely to support a state-specific comparison group but may support a national comparison group. A key advantage of using respondents of these various surveys is that they can provide a relatively low cost comparison group for a state's survey of demonstration participants. Additionally, because state Medicaid programs vary considerably in their approaches, state-specific survey samples can also provide a broader range of counterfactuals than may be available from any in-state comparison group alone. However, there are several limitations of survey data. The number of health-related questions likely to be pertinent to a state's demonstration evaluation is often limited, particularly in surveys that are not primarily focused on health, such as the ACS and CPS. Moreover, different survey procedures can affect how people respond to otherwise identical questions, so comparisons of responses across surveys could be subject to bias. Finally, national or statewide surveys often are unable to accurately identify respondents covered by Medicaid. Beneficiaries often respond incorrectly about their source of health insurance in surveys, as they may not be aware of their eligibility or may erroneously report they have private coverage if they are covered by a private insurer who contracts with a state Medicaid program. To effectively use survey data, evaluators should consult the survey's technical documentation. A final source of comparison group data, specifically for Medicaid 1115 demonstrations focused on pregnancy, childbirth, and neonatal outcomes, is vital statistics data available nationwide from the Centers for Disease Control and Prevention. States can also link vital statistics data from state sources to state Medicaid administrative data to examine outcomes for comparison and treatment group beneficiaries (Kranker et al. 2014).

Aggregate data to provide context for comparison. Various sources of aggregate cross-state data on Medicaid beneficiaries may be available for descriptive comparisons to put state-specific results in context. For instance, coalitions of states support coordinated data collection on specific populations. The NCI-AD™ (National Core Indicators–Aging and Disabilities) is an initiative designed to support states' interest in assessing the performance of their programs and delivery systems to improve services for older adults and individuals with physical disabilities. Other NCI-AD initiatives concern populations with developmental disabilities. The National Committee for Quality Assurance (NCQA) Quality Compass data for Medicaid permits comparisons of HEDIS (Healthcare Effectiveness Data and Information Set) measures for Medicaid beneficiaries enrolled in managed care organizations, against which states might benchmark their performance. Minnesota, for example, is planning to use such data in an evaluation of the state's Prepaid Medical Assistance Project Plus Demonstration, which is a long-standing demonstration that provides full medical assistance benefits for children 12 to 23 months of age and pregnant women during the period of presumptive eligibility. To assess the effects of the demonstration on the populations targeted by the demonstration, the state plans to measure quality outcomes at the state level and compare the state's performance to national benchmarks available in the NCQA Compass data (Minnesota Department of Human Services 2017).

5. Are there non-beneficiary individuals who are similar to those in the treatment group?

States may select individuals who are not Medicaid beneficiaries but have similar characteristics to serve as a comparison group, although this option is not frequently used. States and evaluators should typically consider this option when comparison groups among Medicaid beneficiaries cannot be found—for instance, if the demonstration covers all beneficiaries in the target population. Evaluators should assess whether these non-beneficiary individuals are similar enough to beneficiaries in the treatment group and whether there are enough common data for both beneficiaries and comparison group members to support a strong comparison group. Such a comparison group would most likely be used in a **nonequivalent comparison group design**.

Georgia's Healthy Babies Demonstration Evaluation. This demonstration enrolls Medicaid beneficiaries in Georgia who give birth to very low birth weight babies in an outreach program to improve birth outcomes for future pregnancies. The state's evaluator selected a comparison group of privately insured women with an educational attainment level of high school or less, leveraging a statewide surveillance database on maternal behaviors and outcomes. The evaluator indicated that focusing on low educational attainment was a strategy for identifying privately insured women with income levels that were likely to be comparable to the treatment group (Georgia Department of Community Health and Emory University 2016).

Nonbeneficiary individuals can include those who meet all enrollment requirements for the Medicaid demonstration but who have not enrolled in the state's Medicaid program, as well as otherwise similar individuals who do not meet enrollment requirements. Constructing a comparison group of this type faces two challenges. First, many states may find it challenging to identify a data source with which to identify comparison group members and obtain information on the characteristics and outcomes of these individuals. State or federal surveys are most commonly used to address this issue. Second, particularly in the case of eligible individuals who are not beneficiaries, the evaluator should be cognizant that the choice of whether to enroll in Medicaid may be associated with unobserved characteristics that are related to outcomes. If so, evaluation results may be affected by selection bias, which is discussed in the next section.

B. Comparison group options using beneficiaries who are members of the intervention's target population

1. Do beneficiaries have the choice to participate in the demonstration?

Some 1115 demonstrations give beneficiaries the choice of whether to participate (as well as the choice to withdraw from participation). A logical choice for a comparison group under these circumstances may appear to be eligible beneficiaries who choose not to participate. Although this option may be viable in some situations, the evaluator needs to be concerned about the possibility of **selection bias.** Selection bias arises when participants who choose to be subject to an intervention have baseline characteristics that are systematically different from nonparticipants along dimensions that will very likely affect program outcomes. Unless these characteristics can be measured and statistically controlled for—which is seldom the case—evaluation results are likely to be biased. Selection bias most likely skews evaluation results in the positive direction—that is, making the intervention seem more successful than it may actually be. For example, if an intervention targets high-cost, high-needs beneficiaries who agree to participate, there may be differences in outcomes between participants and nonparticipants because participants have greater motivation to improve their health than those who decline to participate.

Selection bias can be mitigated in several ways. If all of the potential confounding factors that affect whether or not beneficiaries elect to participate are measurable, then using nonenrollees as a comparison group is acceptable, with appropriate statistical controls for relevant differences between the two groups. However, if the logic model suggests that unobserved factors influence selection—which is most likely—then eliminating the risks of selection bias through design is challenging.[11] At a minimum, evaluators should strive to use their logic models to identify the threat of bias, make an assessment of how serious the bias is likely to be, and gauge the likely direction of the bias. Evaluations that may be affected by selection bias can at times provide useful information, although results should always be cast in light of the expected size and direction of the bias.

Options for addressing missing baseline data to support the selection of a comparison group

If baseline data are not available and evaluators have the opportunity to collect primary data on treatment and comparison group members before implementation of the demonstration, one option is to conduct a survey during the post-implementation period. This survey would collect time-invariant personal characteristics (for example, race, gender, education) and ask retrospective questions about respondents' characteristics and outcomes during the baseline period. The survey responses would then be used to match members of the comparison group to the treatment group. The survey could be used to gather post-implementation outcome information, or baseline characteristics could be used in conjunction with post-implementation outcome data from other sources (for example, healthcare use documented through claims/encounter data).

Responses to retrospective survey questions are generally subject to recall error. For instance, telescoping (making things more recent than they were) is common. However, if the survey is administered to both the intervention and comparison groups, then the biases would presumably affect both groups. The evaluator will need to determine what respondents can reasonably be expected to remember and how critical likely response errors might be in the context of the evaluation design.

Selection bias is not a concern, however, if the evaluation is structured in an **intent-to-treat (ITT)** framework. ITT evaluations ask about the effects of an intervention as it is implemented, including the effects on members of the target population who choose not to enroll, failed to fully comply with the requirements of the intervention, or withdrew. Section 1115 demonstration evaluations should most often be designed in an ITT framework, as policymakers are most interested in the effects of policies as they exist. To mitigate potential selection bias in an ITT evaluation, the evaluator should define the treatment group as including all eligible beneficiaries, regardless of whether they choose to participate, and identify a similar comparison group from outside of the population eligible to participate. The implicit assumption is that the distribution of unobserved factors affecting participation—such as motivation—would be identical between treatment and comparison groups, after matching or statistically controlling for measurable characteristics. Various types of comparison groups among those described in this guidance document could be used in an ITT evaluation. Evaluations that focus only on members who

[11] There are some statistical models that attempt to account for selection bias. However, typically, these models require finding a measurable factor that is meaningfully related to decisions to participate, but is not related to the outcomes that are the focus of the evaluation. Finding such "identifying variables" can be very challenging. If the source of selection bias is thought to be time invariant among beneficiaries, and panel data are available on a sample of program participants spanning the pre- and post-intervention periods, then statistical models can control for individual beneficiary differences, allowing for accurate impact estimates among program participants. Using this approach would not allow inferences about how the program would work relative to those selecting not to participate.

chose to join the treatment group are ***per protocol*** **(PP)** evaluations, which can be thought of as proof of concept tests. The results of PP evaluations can be informative for program administrators and CMS. In particular, if the evaluation finds no or only small impacts on those beneficiaries who chose to participate in the demonstration, then it is likely that expanding the demonstration to others (say, by making the intervention mandatory) would produce worse results as the treatment group expands to include both those motivated and unmotivated to participate. Conversely, the results of a PP evaluation in the presence of selection bias means that the results of the evaluation (even if they show positive impacts) cannot be extrapolated to other eligible beneficiaries who chose not to participate in the demonstration intervention.

2. Is implementation of the intervention staggered (and the timing is unrelated to outcomes)?

Some section 1115 demonstrations are intended to provide a proof of concept and are implemented as pilot interventions before being taken to scale. Alternatively, some states may choose to adopt a phased implementation wherein only certain areas or beneficiary groups are initially eligible for the intervention. When states employ small-scale testing or piloting, the beneficiaries who were not affected by the initial intervention rollout can serve as a comparison group. In a **delayed treatment control group (or pipeline) evaluation design**, those beneficiaries who might be enrolled at a later date serve as the comparison group for the treatment group of early enrollees.

Early	O_1	X	O_2		O_3
Delayed	O_1		O_2	X	O_3

There are three important cautions about this option. First, program administrators might prioritize the enrollment of certain types of treatment group members, which would make later enrollees less comparable for the purposes of evaluation. The second caution is that as administrators gain greater experience through implementation, the nature of the intervention might change over time. If the intervention evolves over time, early and late enrollees who are similar in their baseline characteristics may experience somewhat different interventions. Referring to the figure, the estimated impacts on the delayed group, measured by $O_3 - O_2$, may not mimic those found in the early group, measured by $O_2 - O_1$. Third, it may take time for the data to show changes in outcomes related to the intervention, or the effect on outcomes could also be cumulative over time (rather than a fully realized outcome at one point in time), which could influence the extent to which differences between the treatment and comparison group are detected. Assessment of the effects of the demonstration across varying periods of implementation can be accomplished with additional waves of subsequent participants and observations.

Statistical analysis using this design is similar to that used for a nonequivalent comparison group design. However, a somewhat more complex estimation specification would be required to control for secular changes and to assess the effects of greater time in the program.

3. Is eligibility for the intervention triggered by an exogenous event (for example, pregnancy)?

Targeted cohort	O_{2017}	X	O_{2018}
Earlier cohort	O_{2016}		O_{2017}

Some section 1115 demonstrations target specific populations that may become eligible for the intervention (and possibly for Medicaid) as a result of a well-defined event or trigger, such as pregnancy or attaining a certain age.[12] Under these circumstances, there may not be a distinct population from which to draw a contemporaneous comparison group. Evaluators may be able to employ a **cohort design** in these situations and use earlier cohorts as the comparison group. This approach is feasible because the triggering event is unrelated to the implementation of the demonstration. For example, a state may evaluate an intervention aimed at pregnant women that is designed to improve maternal and infant outcomes. If the demonstration began identifying women beneficiaries who became pregnant in 2017 and observed birth and postpartum outcomes through 2018, the evaluator may be able to use an earlier cohort of women who became pregnant in 2016 as the comparison group, with observations on maternal and infant outcomes taken through 2017. This design assumes adjacent cohorts are similar, although it is subject to a threat to validity if something unrelated to the demonstration changed between 2016–2017 and 2017–2018 that would affect birth outcomes (such as a clinical advance in maternal care).

4. Are observations on the treatment group available for at least several time periods before and after the intervention?

O_1	O_2	O_3	O_4	X	O_5	O_6	O_7	O_8

For an established beneficiary group, pre-intervention data on beneficiary characteristics and outcomes are often available from enrollment, claims, or encounter data and other administrative data sources. Once a demonstration has been implemented for several years, available data would support an **interrupted time series design**. In this design, the pre-intervention observations serve as the comparison group for the post-intervention treatment group.[13] This design does not entail a distinct comparison group *per se*; rather, the repeated observations before the intervention, if sufficient for observing the variation and secular patterns of the outcome in question, allow the evaluator to assess whether the level or trend shifted between the periods before and after the intervention. Evaluators should strengthen this design by using regression analysis to control for other potential confounding factors. We illustrate this design in the figure above with a total of eight observations over time, indicated by O_1 through O_8. The X indicates when the intervention began. The number of observations is arbitrary; generally speaking, the more observations over a

[12] In the case of pregnancy, some low-income women will be eligible for Medicaid beforehand, whereas others may become eligible as a result of their pregnancy. Thus, pre-pregnancy Medicaid administrative data may not be available for all members of the treatment group. In the case of pregnancy-related programs, each state maintains a vital statistics database that includes information on women giving birth (for example, maternal age, marital status), delivery outcomes (for example, preterm, cesarean deliveries), and newborn outcomes (for example, birth weight), information that can be used to create a comparison group and measure outcomes. Importantly, it also includes principal payment source for the delivery, allowing evaluators to focus the evaluation on births paid by Medicaid.

[13] The repeated observations could be on a panel of beneficiaries or on repeated cross-sections of beneficiaries.

longer time span the better, making this approach most appropriate for final rather than interim evaluations.

One threat to drawing conclusions from this design is the possibility that another occurrence (for example, an economic recession or policy change) coincided with the implementation of the demonstration, confounding comparisons of pre-post observations.[14] States and evaluators may mitigate this risk through identification of external events that could influence the outcomes achieved, and statistically controlling for these external factors when possible.

New York State Health and Recovery Plans Demonstration Evaluation. Approved in October 2015, the New York Health and Recovery Plans (HARP) demonstration enrolls Medicaid adult beneficiaries with serious mental illness or substance abuse disorder into HARPs, which are specialty lines of business operated by Medicaid MCOs. To evaluate the effects of HARPs on health, behavioral health, and social functioning outcomes, the state intends to conduct an interrupted time series analysis in which non-HARP enrollees in the pre-intervention period serve as a comparison for HARP enrollees in the post-intervention period. To strengthen the design, the state intends to use a regression specification (segmented regression) to test whether HARP implementation was associated with either an immediate change in outcomes or a change in the time trend of the outcome measures (New York State 2017).

Evaluators may further limit the risk of biased results due to concurrent external events by adding a comparison group that was not subject to the intervention but for which data are available for the same set of time periods. Observing whether the pre-post intervention change differed between the treatment and comparison group enhances confidence that the differential change observed in the treatment group was due to the intervention and not some external event that occurred concurrent with the implementation of the demonstration.[15] This approach is called the **comparative interrupted time series design**.

C. Options when no viable comparison group is available

In this section, we present options when an experimental or quasi-experimental design is infeasible because of data limitations or the inability to identify a viable comparison group. In these situations, evaluators should consider nonexperimental evaluation designs. Nonexperimental designs, characterized by the lack of either a comparison group or baseline observations, are vulnerable to most threats to internal validity. We briefly describe several types of nonexperimental designs below.

Case study or one-group posttest-only design. This design involves making observations on the treatment group in the post-intervention period only. Under this design, evaluators cannot assess whether the group experienced any change in outcome measures because no pre-intervention observations were made. Evaluators can strengthen this design by adding multiple post-intervention observations to assess trends in outcomes after the demonstration's implementation. However, the evaluator will

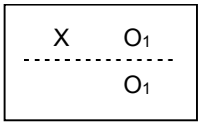
X	O_1

[14] This threat to validity is often called "history."

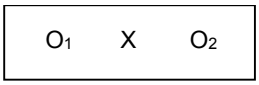
[15] Statistical models using a difference-in-differences specification would be appropriate in this situation. Angrist and Pischke (2009) provide a more detailed description of difference-in-differences models, which are considered a strong evaluation design.

not be able to know whether these trends were the result of, or independent of, the intervention. Evaluators should avoid this design if possible.

Posttest-only with nonequivalent groups design. A variation on the case study design includes a comparison group for which there was no pre-intervention observation. This design therefore offers no mechanism by which the evaluator can assess whether this comparison group is equivalent (or at least similar) to the group receiving the intervention in the pre-intervention period. The design is subject to various threats to validity. Evaluators should be especially alert to whether selection bias is likely to affect the comparison between intervention and comparison group outcomes.



Pretest-posttest design. Although the pre-intervention observation in this design allows measurement of the change in outcomes before and after the intervention, the possibility that other external factors, independent of the intervention, caused this change also makes it a weak design. Evaluators should be particularly sensitive to other factors that might explain changes between O_1 and O_2. Multiple pre- and post-implementation observations can help strengthen this design.



IV. STATISTICAL METHODS AND OTHER STRATEGIES TO SUPPORT THE USE OF COMPARISON GROUPS

A. Statistical methods to support the use of comparison groups

Although a full discussion of the statistical methods used by evaluators is beyond the scope of this document, we briefly describe in this section how statistical methods can help guide the selection of an evaluation design and enhance the confidence that can be placed in quasi-experimental evaluation results.

Power calculations. Beyond selection of a comparison group that is credible, the size and characteristics of the treatment and comparison groups must be sufficient to support the evaluation. Statisticians use well-established formulas to assess an evaluation's **statistical power**. Statistical power refers to the likelihood that a study will detect an effect when there is in fact an effect to be detected. When statistical power is high, the probability of concluding there is no effect when, in fact there is one (a type II error), declines. Fundamentally, statistical power in an evaluation refers to the reliability of the evaluation, that is, the extent to which the evaluation design would produce the same result if it were possible to repeat the evaluation multiple times on different samples of beneficiaries. Statistical power is affected primarily by the expected size of the demonstration's effect and the size of the samples used to detect it, although other aspects of the evaluation design can affect power. Larger effects are easier to detect than smaller effects, and large samples offer greater test sensitivity than small samples. Statistical power tests should be part of the evaluation design phase as they will inform whether efforts should be made to change the number of beneficiaries involved in the evaluation or whether an evaluation component should be abandoned (or replaced with qualitative analysis) because it is unlikely to reliably determine the effects of a demonstration intervention.[16]

[16] Murnane and Willett (2011) provide a non-technical discussion of statistical power and sample size.

Threats to validity: Internal and external validity

Evaluation designs should be assessed in terms of potential threats to internal and external validity. ***Internal validity*** refers to the extent to which a causal conclusion based on a study is warranted (for example, whether and by what magnitude the policy intervention affected outcomes of interest). Internal validity depends in part on the extent to which the evaluation design effectively controls for confounding factors that influence the program outcomes. Alternatively, ***external validity*** refers to the extent to which causal inferences in evaluation research can be generalized to other situations and to other people. There are various common types of threats to internal and external validity.[a]

Threats to Internal Validity

Instrumentation. Observed changes seen between observation points (for example, pre- and post-implementation) may be due to changes in the testing procedure (for example, changes to the content or the mode of data collection).

Regression. Measured changes in program effects may be due to the tendency of extreme pre-intervention scores to revert to the population mean once measured again. This threat affects evaluations of programs for which participants are selected on the basis of extreme pretest (baseline) results (e.g. high pre-implementation health care use), as their post-implementation scores will tend to shift toward the mean score, regardless of the efficacy of the program.

Maturation. Observed changes in program effects could be due to physical or mental changes that occur within the participants themselves. In general, the longer the time from the beginning to the end of a program, the greater the maturation threat.

Testing. Changes in program effects may be due in part to pre-implementation data collection such as a survey, which may convey knowledge to the participants.

History. Observed program results may be explained by events, experiences, or other policy changes that impact the participant between pre- and post- implementation measurements.

Selection. Differences in post-implementation outcome results between a treatment group and nonequivalent comparison group could be due to preexisting differences between the groups rather than the impact of the program itself. This threat is of particular concern when the treatment and comparison groups are significantly different from one another in terms of unobserved characteristics that may be associated with program outcomes.

Threats to External Validity

Interaction of selection and treatment: This threat occurs when the intervention's impact only applies to the particular group involved in the evaluation and may not be applicable to other individuals with different characteristics.

Interaction of testing and treatment: This threat occurs when the design involves a baseline measurement (for example, a survey of participants) that influences the treatment or how individuals respond to the treatment. Therefore, the treatment effects may not be generalizable if implemented without the baseline measurement.

Interaction of setting and treatment: When the results are affected by the setting of the program, evaluations are subject to the threat that the results may not apply if the intervention were implemented in a different setting.

Interaction of history and treatment: If the intervention is evaluated in a given time period, replicating the evaluation in a future time period may not produce similar results; in other words, an aspect of the timing of the intervention (perhaps a major event) may have influenced the treatment effects.

Multiple treatment threats: This threat occurs when the intervention exists in an ecosystem that includes other interventions. As a consequence, the treatment effects may not be generalizable to other contexts.

[a] Adapted from Ranker et al. (2015).

Ensuring the equivalence of treatment and comparison groups. To make valid causal inferences from quasi-experimental evaluations, evaluators must ensure that treatment and comparison groups are similar with respect to the characteristics of the groups' members. The degree to which the treatment and comparison groups are similar is often referred to as **covariate balance.** For example, if the treatment group primarily consists of older adults, then a comparison group consisting primarily of younger adults would not be balanced on age. Balance should be achieved across all covariates, especially those that the logic model suggests are particularly influential on outcomes.

Propensity score matching. Propensity score methods have been developed to facilitate covariate balancing by combining all matching variables in a single common metric—the propensity score (Rosenbaum and Rubin 1983).[17] The propensity score is an estimate of the likelihood of treatment after controlling for baseline characteristics. Under this approach, the evaluator can match, stratify, or weight observations on just the propensity score. Balance on the propensity score, however, does not guarantee that all individual covariates will be balanced, so evaluators should also examine the balance of individual covariates after propensity score methods have been applied and make adjustments accordingly. When the two groups achieve balance across their covariates, the likelihood that they are also similar with respect to unobserved covariates is assumed to be enhanced.

Using statistical models to generate impact estimates. Another important step to ensure that intervention and comparison groups are equivalent and reduce the threat of bias is that impact estimates should be calculated using statistical models that contain covariates thought to affect the outcome.[18] These statistical models, typically regression models, will control for differences in covariates that persist after propensity score methods are applied. The use of regression models alone to adjust for differences in the distributions of covariates is not a substitute for propensity score methods. The combination of matching and regression is preferred because regression can reduce treatment/comparison differences that persist after matching occurs, allow control of covariates not used in matching, and permit the evaluator to specify hypothesized interactions and nonlinear relationships. Matching has the advantage over regression in that it does not impose any parametric structure to control for differences between treatment and comparison groups. Finally, when differences between treatment and comparison groups are reduced through matching, regression models are not forced to inappropriately extrapolate beyond the range of observed values in the comparison group, which could bias results.[19]

---

[17] For reviews of propensity score methods, see Stuart (2010) and Austin (2011).

[18] These steps are not necessary in experimental studies because members of the treatment and control groups are randomly assigned and presumably identical with respect to both measured and unmeasured attributes. However, multivariate regression models are typically used in experimental evaluation studies so as to adjust for treatment-control differences that occur because of chance or differential attrition.

[19] Moffitt (1991) and Murnane and Willett (2011) provide accessible, general guidance for states or evaluators who are interested in learning about alternative ways to specify equations in order to generate desired impact estimates and how specific evaluation designs lend themselves to statistical model specifications. For more detailed technical discussion of these methods, states or evaluators should refer to Angrist and Pischke (2009) and Lance et al. (2014).

Evaluators must make many decisions regarding the evaluation's statistical analysis, including choices about which covariates to control for and which statistical models to use. While the choices leading to the "preferred" model may be well-reasoned, it is important to test how robust the findings are to the choices that led to the preferred model. The process of systematically testing assumptions that led to the preferred model against reasonable alternatives is called **sensitivity analysis** and should be routinely conducted.

B. Other strategies to support evaluation design and corroborate findings

Given the limitations of various evaluation designs and comparison group options, the best strategy to gain confidence in evaluation results can be to **triangulate**, or corroborate, them through multiple analyses. Triangulation might involve the use of different metrics focused on measuring the same general outcome (such as access to care or care quality). It could also involve applying different evaluation designs to the same or similar outcomes or metrics. If the direction and magnitude of impacts are generally consistent across alternative ways of addressing the same research question, then greater confidence can be placed on the overall evaluation conclusions. Furthermore, if more rigorous evaluation designs consistently find that confounding is not present or important, then the evaluator can attach greater confidence to related results that come from nonexperimental evaluation designs that are unable to adjust for variables originally thought to be potential confounders. Finally, quantitative evaluation results should be triangulated with results from qualitative analyses, which can validate and add depth to the interpretation of quantitative impact evaluation results, regardless of the level of rigor possible in comparison group selection and evaluation design.

V. CONCLUSIONS

Section 1115 demonstration evaluations can present many challenges for states and evaluators. Demonstrations are often multifaceted, involving multiple interventions and different beneficiary populations. State evaluations may therefore need to address a variety of research questions, each of which may require unique data and evaluation designs. Clear evaluation goals and detailed program logic models can help to guide the selection of outcomes, the counterfactual, comparison groups, and evaluation designs and can inform decisions about when new data collection may be necessary.

States and their evaluators must inevitably balance real-world data and budget constraints with the desire for rigor. Given this need, the selection of the most appropriate evaluation designs and comparison groups can help to improve both the rigor and efficiency of evaluations by focusing resources on evaluation approaches that are most likely to generate reliable evidence. Evaluators should use statistical techniques to help overcome limitations in the evaluation designs and comparison groups they select and employ a mix of quantitative and qualitative analyses to corroborate research findings.

REFERENCES

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

Arkansas Center for Health Improvement. "Arkansas Works Programs Proposed Evaluation for Section 1115 Demonstration Waiver, February 6, 2017." Little Rock, Arkansas: Arkansas Center for Health Improvement, 2017. Available at: https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/ar/Health-Care-Independence-Program-Private-Option/ar-works-draft-eval-dsgn-2017-2021.pdf. Accessed December 11, 2017.

Arkansas Center for Health Improvement. "Arkansas Health Care Independence Program ("Private Option"): Proposed Evaluation for Section 1115 Demonstration Waiver, February 20, 2014." Little Rock, Arkansas, 2014. Approved by the Centers for Medicare & Medicaid Services on March 24, 2014. Available at: https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/ar/Health-Care-Independence-Program-Private-Option/ar-private-option-eval-design-appvl-ltr-03242014.pdf. Accessed December 11, 2017.

Austin, P. C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research*, vol. 46, no. 3, May 2011, pp. 399–424.

California Department of Health Care Services. "Seniors and Persons with Disabilities: Final Evaluation Design: November 2017." Approved by the Centers for Medicare & Medicaid Services on November 3, 2017. Available at: https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/ca/medi-cal-2020/ca-medi-cal-2020-spds-appvd-eval-design-11032017.pdf. Accessed March 14, 2018.

Campbell, Donald T., Julian C. Stanley, and N. L. Gage. *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Houghton, 1963.

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 March 23, 2018.

Kranker, Keith, So O'Neil, Vanessa Oddo, Miriam Drapkin, and Margo Rosenbach. "Strategies for Using Vital Records to Measure Quality of Care in Medicaid and CHIP Programs." Cambridge, Massachusetts: Mathematica Policy Research, January 2014. Available at: https://www.medicaid.gov/medicaid/quality-of-care/downloads/using-vital-records.pdf. Accessed June 4, 2018.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J.P. Newhouse, H. Allen, K. Baicker, and Oregon Health Study Group. "The Oregon Health Insurance Experiment: Evidence from the First Year." *The Quarterly Journal of Economics*, vol. 127, no. 3 (August 2012), pp. 1057-1106.

Gaudette, E., G.C. Pauley, and J.M. Zissimopoulos. "Lifetime Consequences of Early-Life and Midlife Access to Health Insurance: A Review." *Medical Care Research and Review,* November 2017: 1077558717740444.

Georgia Department of Community Health and Emory University. "Annual Report: Planning for Healthy Babies Program 1115 Demonstration in Georgia, Year 5, December 21, 2016. Available at https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/ga/ga-planning-for-healthy-babies-annual-rpt-2015.pdf. Accessed December 11, 2017.

Hinton, E., M. Musumeci, R. Rudowitz, and L. Antonisse. "Section 1115 Medicaid Demonstration Waivers: A Look at the Current Landscape of Approved and Pending Waivers." Washington, DC: Kaiser Family Foundation, September 2017. Available at http://files.kff.org/attachment/Issue-Brief-Section-1115-Medicaid-Demonstration-Waivers-A-Look-at-the-Current-Landscape. Accessed December 11, 2017.

Lance, P., D. Guilkey, A. Hattori, and G. Angeles. *How do we know if a program made a difference? A guide to statistical methods for program impact evaluation*. Chapel Hill, North Carolina: MEASURE Evaluation, 2014. Available at https://www.measureevaluation.org/resources/publications/ms-14-87-en. Accessed March 23, 2018.

Langbein, Laura Irwin. *Discovering Whether Programs Work: A Guide to Statistical Methods for Program Evaluation*. Goodyear Publishing Company, 1980.

Minnesota Department of Human Services. "Minnesota Prepaid Medical Assistance Project Plus (PMAP+) (No. 11-W-0039/5), Attachment B: Evaluation Plan 2016 to 2020." Approved by the Centers for Medicare & Medicaid Services on August 9, 2017. Available at https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/mn/mn-pmap-ca.pdf.

Moffitt, Robert. "Program Evaluation with Nonexperimental Data." *Evaluation Review,* vol. 15, no. 3, June 1991, pp. 291–314.

Murnane, Richard J., and John B. Willett. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press, 2011.

Musumeci, MaryBeth, Robin Rudowitz, Elizabeth Hinton, Larisa Antonisse, and Cornelia Hall. "Section 1115 Medicaid Demonstration Waivers: The Current Landscape of Approved and Pending Waivers." Washington, D.C.: Kaiser Family Foundation, 2018. Available at: http://files.kff.org/attachment/Issue-Brief-Section-1115-Medicaid-Demonstration-Waivers-The-Current-Landscape-of-Approved-and-Pending-Waivers. Accessed March 22, 2018.

New Hampshire Department of Health and Human Services. "New Hampshire Building Capacity for Transformation – Delivery System Reform Incentive Payment (DSRIP) Demonstration Waiver Evaluation Design: August 2017." Approved by the Centers for Medicare & Medicaid Services on September 5, 2017. Available at: https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/nh/building-capacity/nh-building-capacity-transformation-appvd-eval-dsgn-09052017.pdf. Accessed December 11, 2017.

New York. "Evaluation Framework for the New York State Behavioral Health Partnership Plan Demonstration Amendment—NYS MMC Behavioral Health Carve-In and Health and Recovery Plans Demonstration Period: October 1, 2015 through December 31, 2020." Approved by the Centers for Medicare & Medicaid Services on May 10, 2017. Available at https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/ny/medicaid-redesign-team/ny-medicaid-rdsgn-team-harp-eval-dsgn-appvl-05102017.pdf. Accessed December 11, 2017.

Ranker, L., W. DeJong, and R. Schadt. "Program Evaluation." Boston, Massachusetts: Office of Teaching & Digital Learning, Boston University School of Public Health, 2015. Available at: http://sphweb.bumc.bu.edu/otlt/mph-modules/ProgramEvaluation/index.html. Accessed March 23, 2018.

Renger, R., and A. Titcomb "A three-step approach to teaching logic models." *American Journal of Evaluation*, vol. *23*, no. 4, 2002, pp. 493-504.

Rosenbaum, P. R., and D. B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, vol. 70, no. 1, 1983, pp. 41–55.

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman. *Evaluation: A systematic approach*. Sage publications, 2003.

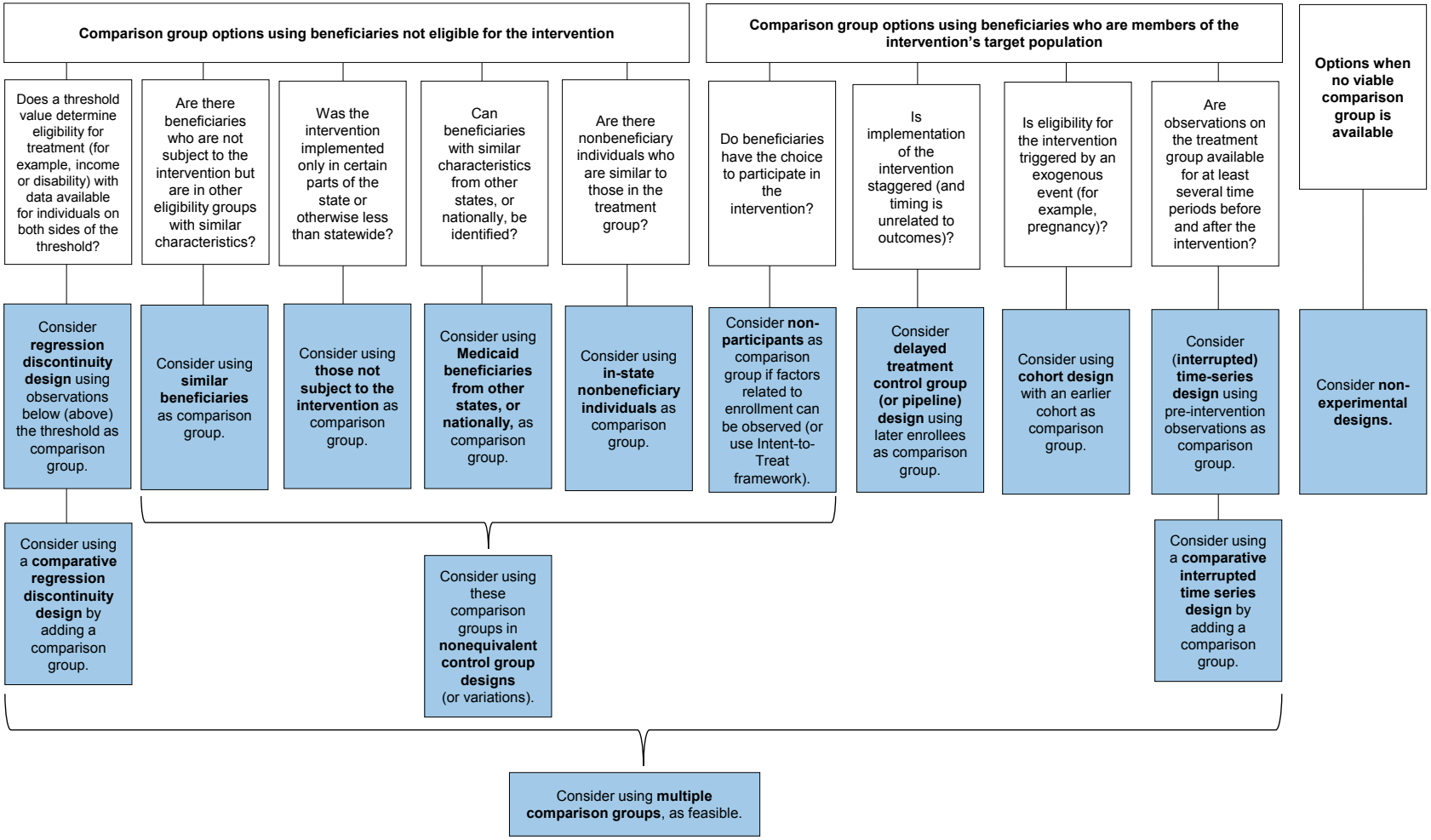
Social & Scientific Systems Inc. Evaluation Design Report for Montana HELP Federal Evaluation. Silver Spring, Maryland: Social & Scientific Systems, May 16, 2017. Approved by the Centers for Medicare & Medicaid Services May 31, 2017. Available at: https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/mt/HELP-program/mt-HELP-program-fed-state-eval-dsgn-051617.pdf. Accessed December 8, 2017.

Stuart, E. A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

Weiss, Carol H., *Evaluation: Methods for Studying Programs and Policies,* 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1998.

APPENDIX A: FLOWCHART TO GUIDE THE IDENTIFICATION OF COMPARISON GROUPS AND EVALUATION DESIGNS

Figure A.1. Questions to guide the choice of comparison group and evaluation design



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

APPENDIX B: GLOSSARY

Case study design: This nonexperimental design involves making observations on the treatment group in the post-intervention period only because no pre-intervention observations had been made. This type of design is also called "one group posttest-only design" (see below).

Cohort design: This design is relevant when there is no distinct population from which to draw a contemporaneous comparison group. When an event unrelated to the implementation of the demonstration, for example pregnancy, triggers eligibility for the intervention, earlier cohorts of individuals may serve as a comparison group.

Comparative interrupted time series design: This design employs an interrupted time series design (see below) but adds time series data from a nonequivalent comparison group over the same period. Adding the comparison group protects the evaluation from the threat to validity known as history.

Comparative interrupted time series design. This designs adds a contemporaneous comparison group to the interrupted time series design. Importantly, the addition of a comparison guards against making incorrect inferences on program impacts when another event coincident to the intervention can affect program outcomes.

Comparative regression discontinuity design: This rigorous design adds a comparison group—one with observations both above and below the eligibility threshold—to a regression discontinuity design

Comparison group: In quasi-experimental designs, the comparison group is composed of individuals who closely resemble the treatment group with respect to demographic or other variables but are not receiving the intervention. The comparison group represents the evaluation's counterfactual (that is, what would have happened absent participation in the intervention).

Confounding variables: A confounding variable is an outside influence that changes the effect of a dependent and independent variable or, in the context of demonstration evaluations, that influences both the treatment and the outcome. Confounding is the bias that arises when such variables are not controlled for, that is, when the treatment and comparison groups differ with respect to confounding variables. Selection of appropriate comparison groups, the use of statistical methods such as propensity score matching to ensure treatment and comparison groups are similar with respect to these variables, and inclusion of confounding variables in statistical models can help to reduce the threat of bias from confounding.

Control group: In experimental designs, the control group includes randomly selected individuals who do not receive the intervention. Therefore, observations on the control group serve as the evaluation's counterfactual.

Counterfactual: In an experimental or quasi-experimental evaluation, comparison groups represent the state that exists absent the intervention—that is, the counterfactual. In some cases,

counterfactuals may represent alternative interventions to achieve program goals rather than the absence of the intervention.

Delayed treatment control group (or pipeline) design: Appropriate for interventions that are implemented in a staged fashion, this design exploits variation in the timing of program implementation, using eligible participants who have not yet received the program as a comparison group.

Ex ante evaluation: These evaluations must be planned prior to the implementation of the intervention and may involve random assignment, staged implementation of the intervention, or primary data collection in service of the evaluation.

Experimental design: This design entails randomized assignment to treatment and control groups, controlling for systematic differences between individuals who are subject to the intervention and those who are not; therefore, among other designs, it is the least likely to suffer from threats to internal validity.

Ex post evaluation: These evaluations are planned after the design, and sometimes after the implementation, of the intervention. The nature of the intervention, the timing of its implementation, the assignment of people to the treatment group, and available data will influence the evaluation design.

External validity: This type of evaluation validity relates to the extent to which findings are generalizable to other contexts or populations.

Impact evaluation: An impact evaluation assesses the changes that can be attributed to a particular intervention, such as a project, program or policy. Ideally, an impact evaluation measures intended as well as unintended outcomes.

Intent-to-treat evaluation: An intent-to-treat (ITT) evaluation assesses outcomes of the initial population to whom the intervention was offered, including those who chose to receive the intervention and those who did not so choose or who withdrew from the intervention or failed to fully comply with the intervention's requirements.

Internal validity: This type of evaluation validity refers to the extent to which (1) potential confounding variables are adequately controlled for and (2) the design enables researchers to draw conclusions about the relationship between the intervention and outcome.

Interrupted time series design: In this design, which can be employed when an intervention occurred at a specific point in time, data are collected at several points before and after the intervention (a time series). If the intervention has a causal impact, the post-intervention time series will have a different level or slope.

Nonequivalent comparison group design: This quasi-experimental design compares post- versus pre-intervention outcomes among members of a treatment group with that of another group with similar characteristics. This type of design is also called "untreated control group design with pretest and posttest" (see below).

Nonexperimental design: This design is characterized by the lack of either a comparison group or baseline observations. Because it measures only post-intervention effects, it is vulnerable to most threats to internal validity.

One-group posttest-only design: This design, also referred to as case study design, involves making observations on the treatment group in the post-intervention period only; changes in outcome measures cannot be assessed because no pre-intervention observations were made.

Per protocol evaluation: A per protocol evaluation assesses the impact of an intervention on those who were fully exposed to the intervention, and thus does not account for those who refused the intervention, withdrew from it, or otherwise failed to follow intervention rules or expectations.

Posttest-only with nonequivalent groups design: This nonexperimental design includes a comparison group but lacks pre-intervention observations on that group. It therefore offers no mechanism by which the evaluator can assess whether this comparison group is equivalent (or at least similar) to the group receiving the intervention in the pre-intervention period.

Pretest-posttest design: This nonexperimental design lacks a comparison group but includes pre-intervention observations that allow evaluators to measure the change in outcomes between the periods before and after the intervention.

Quasi-experimental design: Quasi-experimental designs are observational studies that identify a comparison group that did not receive the intervention but is as similar as possible to the treatment group in terms of baseline (pre-intervention) characteristics.

Reliability: In the context of program evaluation, reliability is related to the statistical power of an evaluation design and is the degree to which the design would produce similar results if repeated on different samples.

Regression discontinuity design: This type of design is appropriate when an intervention is targeted to individuals who meet an eligibility threshold, such that individuals close to the eligibility threshold are similar to the treatment group and may serve as a comparison group.

Selection bias: A threat to internal validity, selection bias is introduced when individuals who choose to participate in an intervention have baseline characteristics (in particular unmeasured characteristics) that are systematically different from those of nonparticipants along dimensions that will very likely affect program outcomes.

Sensitivity analysis: An analytic approach to testing how estimation results change when assumptions regarding the relationship between the independent and dependent variables or other assumptions vary from the primary model used.

Treatment group: This group is composed of individuals who are subject to the intervention, either on the basis of randomization in experimental designs or through other circumstances such as meeting program eligibility criteria in quasi-experimental designs.

Triangulation: The process of validating evaluation results and increasing confidence in the effects of an intervention by comparing related evaluation results obtained using multiple data sources, different evaluation designs and comparison groups, and across related outcome measures.

Untreated control group design with pretest and posttest: This design involves a pretest/posttest design with a comparison group**.** It is also called "nonequivalent control group design."

Validity: In the context of evaluation design, the validity of an evaluation refers to the degree it is free from potential bias stemming from such things as measurement error or confounding.

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