
EVALUATION DESIGN GUIDANCE FOR SECTION 1115 DEMONSTRATIONS FOR BENEFICIARIES WITH SERIOUS MENTAL ILLNESS/SERIOUS EMOTIONAL DISTURBANCE AND SUBSTANCE USE DISORDERS

Introduction

The purpose of this document is to support states in developing evaluation designs for Section 1115(a) Medicaid demonstrations that aim to provide and improve the continuum of care for (1) beneficiaries with serious mental illness (SMI) or serious emotional disturbance (SED) and (2) beneficiaries with substance use disorders (SUDs). The document communicates the expectations of the Centers for Medicare & Medicaid Services (CMS) for the rigor of demonstration evaluations, including assessments of cost implications.

Evaluation design guidance specific to each demonstration type is available in separate appendices to this document. States pursuing both SMI/SED and SUD demonstrations should consult Appendices A and B, respectively, to ensure that the state's evaluation design adequately addresses the guidance for each type of demonstration. CMS previously released stand-alone evaluation design guidance for SUD, which has not been substantively altered. It now exists as Appendix B. Appendix C provides a recommended method for assessing changes in total costs under either demonstration and for examining cost drivers within the Medicaid program as states expand the continuum of care for beneficiaries with SMI/SED or SUD.

This document and its appendices should be used as a basis for states' discussions with their independent evaluators. The document may also support evaluator procurement. It is not intended to replace the evaluation design process or to resolve all design-related issues that may arise. CMS encourages states to procure an independent evaluator to support the development of a robust draft evaluation design for CMS review. Ideally, states should identify an evaluation team before implementation to consult on implementation plans that support robust research designs and plan early data collection. CMS is available to provide further technical assistance to states if requested.

SMI/SED and SUD 1115 demonstrations and CMS evaluation policy

On November 13, 2018, CMS published a State Medicaid Director (SMD) letter that provided guidance on Section 1115 demonstrations for improving access to and quality of treatment for Medicaid beneficiaries with SMI or SED. Through this new policy, states have the flexibility to develop innovative solutions to address state-specific concerns about the SMI/SED care continuum and to phase in a range of strategies to address those concerns.

CMS previously published an SMD letter on November 1, 2017, with guidance on Section 1115 demonstrations for improving access to high quality, clinically appropriate treatment for opioid use disorder (OUD) and other SUDs. Through this policy, states have the opportunity to demonstrate how to implement best practices for improving OUD and other SUD treatment that take into account the particular challenges raised by the opioid epidemic in each state.

Consistent with the Administration's goal to hold states accountable for outcomes associated with their Medicaid programs, states with approved Section 1115 demonstrations are required to conduct regular monitoring and rigorous evaluation activities. CMS recommends an integrated

approach to monitoring and evaluation: monitoring data provide essential information on the demonstration's implementation, which creates the context for evaluations and informs the interpretation of results. Separately, CMS will provide guidance for monitoring approved demonstrations. CMS encourages states to share their monitoring protocols, data, and reports with their independent evaluators to support the evaluation.

To streamline the processes for monitoring and evaluation, CMS suggests metrics that can be used for both. However, CMS recommends separate sets of metrics for SMI/SED and SUD demonstrations. Monitoring data alone are not sufficient for evaluation purposes. Although some metrics may be identical for monitoring and evaluation, states are expected to develop evaluation designs that use the most robust methods feasible within the context of the demonstration. This may include collecting pre- and post-demonstration time series data and, when feasible, identifying comparison groups and applying statistical techniques to control for observable beneficiary characteristics.

As part of the requirements described in the special terms and conditions (STC) for the demonstrations, states are required to submit to CMS an evaluation design for each demonstration type within 180 days after approval. CMS must approve each evaluation design, and every effort should be made to adhere to the approved evaluation design when conducting analyses and developing the evaluation reports. However, states may request and CMS may agree to changes in the methodology in appropriate circumstances. States may contact their CMS project officer or the evaluation and monitoring technical assistance mailbox (1115MonitoringandEvaluation@cms.hhs.gov) for methodological guidance.

Section 1115 Medicaid Demonstration Evaluations

Under Section 1115 of the Social Security Act, states may apply for federal permission to implement and test new approaches for administering Medicaid programs that (1) depart from existing federal rules yet are consistent with the overall goals of the program, (2) are likely to meet the objectives of Medicaid, and (3) are budget neutral to the federal government. Section 1115 also requires that federal regulations specify requirements for demonstration evaluations. 42 CFR 431.424 outlines general evaluation requirements and evaluation design components such as hypotheses, data sources, and comparison strategies when feasible. 42 CFR 431.412 outlines requirements for including initial evaluation designs and evaluation reports in demonstration applications and renewal applications. STCs for each state further specify CMS's expectations for the contents and timing of evaluation designs and interim and summative evaluation reports.

How to use this technical assistance guide

This guidance is intended to support states in developing evaluation designs that will meet CMS's expectations for rigorous demonstration evaluations in compliance with requirements set forth in the STCs. Appendices A (SMI/SED) and B (SUD) include hypotheses, research questions, and evaluation approaches that will generate strong evaluation designs. The approaches in Appendices A and B differ in several ways. Appendix A was developed based on the goals and milestones of SMI/SED demonstrations that were outlined in the November 13, 2018, SMD letter. Appendix B was based on the SUD demonstrations that were outlined in the November 1, 2017, SMD letter. Recommended evaluation approaches are based on best practices in policy evaluation.

The content included in an evaluation design submitted to CMS by each state will need to be customized to each state's specific demonstration. For example, states may wish to modify or add hypotheses or research questions to ensure alignment with the state's Section 1115 demonstration policies. CMS recognizes that each state's Section 1115 demonstration will be unique. Therefore, the evaluation design submitted to CMS should capture those unique elements. This document discusses the common elements of the demonstration types as identified by each SMD letter. States should prioritize the information that will best provide evidence for the policies the state is testing under Section 1115 authority and that will best inform decisions regarding future policy.

This guidance also provides a general framework that states and their evaluators can use to organize the writing of their evaluation designs. The format for the evaluation design should be as follows:

- A. General background information
- B. Evaluation questions and hypotheses
- C. Methodology
- D. Methodological limitations
- E. Attachments

These sections correspond to the general evaluation design guidance that CMS has frequently incorporated as part of the STCs.¹ The remainder of this document provides guidance on how to complete each of these sections.

A. General background information

States should include basic information about the demonstration, as discussed below.

1. Demonstration name and approval date and evaluation time period

If the demonstration approval period and the evaluation time period are not identical, states should indicate the difference and provide a brief explanation (for example, the state was approved for the demonstration in September 2019 but did not start implementation until December 2019).

2. Goals of the demonstration's policies

This should include state-specific goals as articulated in the demonstration application or implementation plan and also the goals articulated in the respective SMD letter. The goals for

¹ This document and related appendices focus on evaluation of SMI/SED and SUD demonstrations and therefore build on CMS's general guidance, "Section 1115 Demonstrations: Developing the Evaluation Design," which is available at <https://www.medicare.gov/medicaid/section-1115-demo/downloads/evaluation-reports/developing-the-evaluation-design.pdf>. States with multiple components within their 1115 demonstrations (such as both SMI/SED and SUD components) are encouraged to clearly identify SMI/SED and SUD evaluation design components within their broader 1115 evaluation design. For example, states should include separate logic models for each of the SMI/SED and SUD demonstration components, as well as separate design tables describing the hypotheses and research questions for each component, following the guidance in Appendices A and B respectively.

SMI/SED and SUD demonstration types, as described in the corresponding SMD letters, are listed below.

The purpose of SMI/SED demonstrations is to test whether a new paradigm for delivering SMI/SED services for Medicaid enrollees achieves the following goals:

- a. Reduced utilization and length of stay in emergency departments among Medicaid beneficiaries with SMI or SED while awaiting mental health treatment in specialized settings
- b. Reduced preventable readmissions to acute care hospitals and residential settings
- c. Improved availability of crisis stabilization services, including services made available through call centers and mobile crisis units, intensive outpatient services, as well as services provided during acute short-term stays in residential crisis stabilization programs, psychiatric hospitals, and residential treatment settings throughout the state
- d. Improved access to community-based services to address the chronic mental health care needs of beneficiaries with SMI or SED including through increased integration of primary and behavioral health care
- e. Improved care coordination, especially continuity of care in the community following episodes of acute care in hospitals and residential treatment facilities

The purpose of SUD demonstrations is to test whether states that pursue delivery system transformation efforts for beneficiaries with SUDs are able to achieve the following goals:

- a. Increased rates of identification, initiation, and engagement in treatment
- b. Increased adherence to and retention in treatment
- c. Reductions in overdose deaths, particularly those due to opioids
- d. Reduced utilization of emergency departments and inpatient hospital settings for treatment when the utilization is preventable or medically inappropriate, through improved access to other continuum of care services
- e. Fewer readmissions to the same or higher level of care when the readmission is preventable or medically inappropriate
- f. Improved access to care for physical health conditions among beneficiaries

3. Description of the demonstration and implementation plan

This can be a summary of information provided in the implementation plan. The summary should also include a description of the mental health service delivery system and service availability at demonstration baseline. (SMI/SED demonstrations can draw on descriptions in their initial assessment of the availability of mental health services.) The description should also include information relevant to the evaluation design, such as implementation timing for the different activities planned under the demonstration and any plans for a staged rollout of relevant policies or programmatic changes that may create opportunities for within-state comparison groups.

4. Description of the population groups impacted by the demonstration

States should include a description of Medicaid eligibility groups that are subject to relevant policies or any other criteria used to define the group of beneficiaries impacted by the demonstration.

5. Other relevant contextual factors

States should note other state or local health care delivery or payment reform efforts that may affect either the demonstration or the evaluation or both. States should work with evaluators to identify the contextual factors most likely to affect demonstration outcomes.

B. Evaluation questions and hypotheses

1. Logic model

A primary purpose of Section 1115 demonstration evaluations is to determine whether the demonstrations are achieving their stated goals for each demonstration policy. Logic models help states and their evaluators harmonize the goals of the evaluation with the goals of the policy. Logic models should depict expected short-term and long-term outcomes for each policy. Relatedly, because hypotheses for some policies are logically dependent (that is, long-term outcomes are dependent upon realizing short-term outcomes), states and their independent evaluators should use evaluation results for short-term outcomes to help interpret results for long-term outcomes.

States should also include moderating factors and confounding (or contextual) variables in their logic models. Moderating factors are important preliminary outcomes that states should consider because they affect the relationship between the demonstration activities and one or more hypothesized outcomes; however, they are not themselves the policy goals. For example, client acceptance of any treatment offered moderates the degree to which demonstration activities for improving follow-up after discharge can influence preventable readmissions. Confounding or contextual variables may influence demonstration implementation or outcomes and can bias evaluation results, if the evaluation approach does not control for them. Beneficiaries' underlying health status—including, the extent of their mental and physical treatment needs—is an example of a confounding variable that the evaluation should control for in any regression model of the effect of demonstration policies on health outcomes. Finally, states may also wish to include potential unintended or adverse consequences. Appendix A includes an example logic model for SMI/SED demonstration activities that are intended to reduce preventable readmissions to acute care hospitals and residential settings.²

For SUD demonstrations, prior guidance required states to include a driver diagram in the evaluation design. States now have the option of submitting a driver diagram or logic model as part of their SUD evaluation designs. 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. CMS emphasizes logic models here to focus on relationships and logical dependencies among measurable policy outcomes. In practice, both logic models and driver diagrams can serve similar purposes in evaluation planning.

If a state opts to include a driver diagram rather than a logic model for its SUD demonstration evaluation design, the driver diagram should depict the relationship among the

² For more information on logic models, refer to “Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations,” which is available at <https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/causal-inference.pdf>.

demonstration's purpose (the goal of the demonstration), the primary drivers (specific actions or interventions) that contribute directly to realizing that purpose, and the secondary drivers that are necessary to achieve the primary drivers. A driver diagram is a "living tool" that should be updated regularly to reflect program refinements. The purpose stated in the example driver diagram in Appendix B for SUD demonstration policies is to reduce opioid-related overdose deaths. The primary drivers are delivery system components or factors that contribute directly to reductions in opioid-related overdose deaths. Secondary drivers are actions, interventions, or lower-level components that are necessary to achieve the primary drivers. As shown in Appendix B, Figure B.1, secondary drivers may relate to one or multiple primary drivers. The example also illustrates that primary drivers may relate to other primary drivers, but all primary drivers have a direct relational impact on the demonstration's overall purpose.³

2. Hypotheses and research questions

Evaluation documents must include a discussion of the hypotheses that a state intends to test, along with research questions that will address each hypothesis. Hypotheses should correspond to the policy goals for the demonstration and should clearly indicate the direction of the expected change—although, evaluators should use two-sided hypothesis tests to allow them to observe unexpected effects. Research questions should be listed under each hypothesis. Primary research questions address high-level effects of the changes in service delivery. Subsidiary research questions help states and CMS understand the observed effects on outcomes of interest in greater depth and detail.

Recommended hypotheses, research questions, measures, data sources, and analytic approaches for SMI/SED demonstrations are provided in Appendix A. A separate set of recommended hypotheses, research questions, measures, data sources, and analytic approaches for SUD demonstrations is provided in Appendix B. States may, however, modify the recommended hypotheses and research questions to more directly align with their demonstration activities. For each demonstration evaluation design, states may also add (1) hypotheses and research questions designed to evaluate unique or state-specific aspects of their policies and (2) research questions about demonstration processes and implementation that draw on monitoring or qualitative data and that are not already included in the appendix tables.

States are also asked to assess patterns and trends in Medicaid costs separately for each Section 1115 demonstration type. Appendix C recommends specific approaches for analyzing costs associated with SMI/SED and SUD demonstrations.

C. Methodology

1. Evaluation design summary

In this section, states should provide a brief narrative summary of the proposed evaluation design. Details on the proposed methodology should be described in subsequent sections and reflected in one or more design tables submitted as part of the evaluation design. Appendices A and B to this guidance contain suggested approaches to address key research questions for

³ If more detailed technical assistance related to driver diagrams is needed, additional tools and support can be found at <https://innovation.cms.gov/files/x/hciatwoaimsdrrvs.pdf>.

SMI/SED or SUD demonstrations. CMS anticipates that states will select at least one of the recommended approaches associated with each of the research questions included in a state's evaluation design. The use of multiple approaches can serve to strengthen confidence in findings. A state may also propose alternative approaches consistent with best practices in policy evaluation. The design tables submitted to CMS as part of a state's evaluation design should include only the approaches that the state intends to use after discussing potential approaches with an independent evaluator. Design tables must separately specify outcome measures, data sources, comparison strategies, and analytic approaches for each research question. The components for each research question must be presented in the same table so that CMS can assess how well they will work together for a given analysis.⁴

2. Target and comparison populations

States should define the target populations covered by their demonstrations in their monitoring protocols.⁵ CMS expects all states with approved SMI/SED and SUD demonstrations to make their best effort to incorporate comparison populations in their evaluation designs. Using comparison groups and quasi-experimental designs increases the validity of evaluation evidence and allows evaluators to attribute changes in outcomes to the policy intervention.⁶ Comparison group selection should be informed by the best opportunity to establish a counterfactual to test the effects of the demonstration. This would entail identifying a group of individuals who are similar to the demonstration group in their observable characteristics who do not have access to the same demonstration benefits (specific to SMI/SED or SUD). Evaluators would then compare access to and utilization of treatment services between the demonstration and comparison groups.

States should work with their independent evaluators to explore the feasibility of within-state or other-state comparison strategies. Involving evaluators in the demonstration design may facilitate this process. For example, states may consider staging implementation by geographic area, which would allow for comparisons of beneficiaries who have and have not received the intervention at each interval.⁷ In addition, other-state comparison groups may be possible.

⁴ The format of the design tables recommended for SMI/SED and SUD demonstrations was adapted from the format in the general CMS guidance provided at <https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/developing-the-evaluation-design.pdf>.

⁵ For SMI/SED demonstrations, states will describe the populations covered by the demonstration and the definitions for SMI and SED that they will use to calculate the monitoring metrics, including a list of diagnosis codes and service requirements.

⁶ For more guidance on the selection of comparison groups, please refer to the white paper, "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations," which is available at <https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/comparison-grp-eval-dsgn.pdf>.

⁷ Comparison group options are discussed in more detail in separate guidance on designing implementation to support rigorous evaluations: <https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/index.html>.

Comparison states should have a Medicaid program that is similar to the demonstration state in terms of trends in SMI/SED- or SUD-related outcomes.

In the evaluation design, a state's description of comparison groups should specify inclusion and exclusion criteria and explain how the state will overcome drawbacks in specific comparison group strategies. In some cases, it may be advisable to use propensity score matching or other statistical techniques to reduce dissimilarities in observable characteristics.⁸ For nationally recognized metrics, the state should compare performance to the national averages where possible (for example, National Quality Forum [NQF], Healthcare Effectiveness Data and Information Set [HEDIS], or Medicaid Adult Core set metrics).

Additional considerations when choosing a comparison group. CMS recognizes the challenges in identifying valid comparison groups for SMI/SED and SUD demonstrations. If a credible comparison group is difficult to identify, the state should engage with CMS for guidance. CMS will consider on a case-by-case basis the analytical approaches that do not require comparison groups. For example, information on treatment group outcomes prior to demonstration implementation will allow for interrupted time series analysis, which is an acceptable approach when credible comparison groups are unavailable. Multiple pre-period observations are necessary for this approach. In addition, using only the first year of implementation is not sufficient.⁹

3. Evaluation period

The evaluation design should clearly specify the evaluation time period, as well as which hypotheses, research questions, and data will be included in both the interim and summative (final) evaluation reports. If the state does not anticipate being able to address certain research questions in the interim evaluation report—either because of a lag in data availability or because outcomes are expected to occur after the interim evaluation period—the evaluation design should denote those research questions and explain the expected timing.

4. Data sources

In this section, states should provide details on the data sources for the evaluation, including all data sources listed in the design tables that a state submits to CMS. CMS recommends the use of both qualitative and quantitative data, depending upon the best way to address specific research questions. The evaluation design should describe all data sources to ensure that they can generate the planned outcome measures. States should also discuss the frequency of data collection, the process for demonstrating the accuracy and completeness of the data, and the limitations of the data. For all quantitative measures used to identify causal effects, states should

⁸ In difference-in-differences analyses, the similarity of pre-period trends between demonstration and comparison groups is more important than the similarity of observable characteristics. In addition, propensity score matching can be counterproductive. For a discussion of the bias introduced by matching in difference-in-differences analysis, see Daw, J. R. and Hatfield, L. A. (2018) "Matching and Regression to the Mean in Difference-in-Differences Analysis." *Health Services Research*, 53: 4138-4156. [doi: 10.1111/1475-6773.12993](https://doi.org/10.1111/1475-6773.12993)

⁹ In contrast, having pre-period data on both treatment and comparison group members permits more rigorous methodologies (for example, difference-in-differences), which are possible with only a single pre-period observation (for example, at demonstration baseline).

include statistical power calculations in the evaluation design to show that samples will be of sufficient size to support the proposed analyses. If power calculations suggest that analyses will be underpowered and it is not possible to generate larger samples, states should engage with CMS for guidance on the acceptable use of descriptive analyses.

When primary data (that is, data collected specifically for the evaluation) will be collected, states should discuss the method (for example, surveys, group interviews, or individual interviews) in detail, as well as the frequency and timing of the data collection. States that collect primary data should also consider the need for institutional review board approval.

Potential data sources for SMI/SED or SUD demonstration evaluations are listed below. CMS recognizes that not all data sources will be available for all states.

- Medicaid administrative data, including claims, encounters, enrollment, and demonstration monitoring data
- Medicare claims data for people dually eligible for Medicaid and Medicare
- Electronic health records and/or health information exchange clinical data repositories
- Beneficiary and provider surveys
- Qualitative data, including individual and group interviews with beneficiaries and/or key informants
- State data warehouses
- National survey data

a. Further considerations for Medicaid administrative data

Medicaid administrative data, including claims, encounters, and enrollment data, are necessary to answer research questions about the use of SMI/SED or SUD treatment services. Medicaid administrative data can support rigorous quasi-experimental evaluation designs because large numbers of observations are typically available both before and after demonstration implementation. Medicaid administrative data may also include data on demonstration operations (including monitoring data), which may be helpful for answering research questions about demonstration processes.

Medicaid administrative data may also be used to construct demonstration monitoring metrics or nationally recognized measures that are employed in the evaluation. Where possible, Medicaid-specific measures are preferable. SMI/SED or SUD monitoring metrics should be leveraged in the evaluation as appropriate. Recommended sources for Medicaid-specific measures that can be constructed from administrative data include the following:

- Medicaid Child Core Set, <https://www.medicare.gov/medicaid/quality-of-care/performance-measurement/child-core-set/index.html>
- Medicaid Adult Core Set, <https://www.medicare.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html>

- Medicaid Health Home Core Set, <https://www.medicaid.gov/state-resource-center/medicaid-state-technical-assistance/health-homes-technical-assistance/health-home-quality-reporting.html>

Other metric sets, not specific to Medicaid, can be identified from the following sources:

- National Behavioral Health Quality Framework (NBHQF),¹⁰ <https://www.nasmhpd.org/sites/default/files/SAMHSA%20Quality%20Improvement%20Initiative.pdf>
- NQF, https://www.qualityforum.org/Measures_Reports_Tools.aspx
- HEDIS measures, <http://www.ncqa.org/hedis-quality-measurement/hedis-measures>
- CMS Measure Inventory Tool, https://cmit.cms.gov/CMIT_public/ListMeasures
- Certified Community Behavioral Health Clinics measures, <https://www.samhsa.gov/section-223/quality-measures>

b. Further considerations for the use of beneficiary surveys

Beneficiary surveys developed and fielded by states can yield rich information on topics such as patient experiences, unmet care needs, barriers to care, service integration (including gaps in care coordination that are apparent to beneficiaries), and understanding patterns of SMI/SED or SUD treatment services. States that choose to field beneficiary surveys may wish to use existing survey items that reflect beneficiary experiences with behavioral health services, such as the Experience of Care and Health Outcomes survey, which is part of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) suite of surveys.¹¹

States that conduct surveys of current and/or former beneficiaries typically use single cross-sectional surveys, which gather data at a point in time. However, understanding certain outcomes of treatment services requires that states administer longitudinal or panel surveys that involve multiple survey waves that follow the same people over time.¹² For example, a state might survey the same beneficiaries repeatedly to assess changes over time for individual beneficiaries related to severity of symptoms, self-rated physical health, medication adherence, or frequency of office visits.

The description of a planned beneficiary survey in the evaluation design submitted to CMS should describe the sampling strategy and estimated number of completed surveys, power calculations, stratification by subgroups of particular interest, minimum detectable effect sizes for key subgroups of interest, frequency and timing of data collection, and the mode of data collection. The evaluation design should also describe how the state will meet challenges, such

¹⁰ The NBHQF was developed by the Substance Abuse and Mental Health Services Administration. It is available on the National Association of State Mental Health Program Directors' website.

¹¹ More information about the CAHPS survey can be found at <https://www.ahrq.gov/cahps/surveys-guidance/echo/index.html> and <https://www.ahrq.gov/cahps/about-cahps/index.html>.

¹² CMS guidance on designing beneficiary surveys to support rigorous evaluations is available at <https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/index.html>.

as reaching hard-to-reach populations, achieving sufficient response rates, weighting the sample for survey nonresponse, and avoiding survey attrition among beneficiaries.

c. Further considerations for the use of provider surveys

States may be able to obtain actionable information by conducting provider surveys on care coordination and approaches to integrated care. For example, states can use questions from the Behavioral Health Integration Capacity Assessment (BHICA), which allows behavioral health organizations to evaluate their care integration processes.¹³ The evaluation design submitted to CMS should describe the sampling strategy and the estimated number of completed surveys, stratification by subgroups of particular interest, frequency and timing of data collection, and the mode of data collection. The evaluation design should also describe how the state will meet challenges, such as achieving sufficient response rates, weighting the sample for survey nonresponse, and avoiding survey attrition for longitudinal surveys. Because provider surveys often focus on implementation factors, they are commonly used in descriptive analyses or as moderating variables that affect beneficiary outcomes (that is, as independent variables rather than outcome measures or dependent variables in regressions). To the extent that data from provider surveys will be used as outcome (dependent) variables in statistical analyses of demonstration effects, the evaluation design submitted to CMS must also include power calculations and minimum detectible effect sizes for key subgroups of interest.

d. Further considerations for the use of qualitative data

Rigorous collection of qualitative data and application of qualitative analytic methods may be appropriate when the goal is to learn about the understanding, experience, and perspective of beneficiaries or other stakeholders; when quantitative approaches are not feasible or sample sizes are very small; or to complement and contextualize quantitative outcomes. For example, qualitative data on demonstration implementation gathered through key informant interviews with field operations staff, providers, and consumer groups may suggest refinements to quantitative analyses and can provide context for observed outcomes. States can also use qualitative data to better understand the implementation structures and processes that lead to such outcomes. In addition, states can use qualitative data gathered through individual or group interviews with beneficiaries to inform the development of beneficiary survey instruments.

States planning to collect qualitative data should describe their planned methods in detail, including the process for identifying and selecting the respondent sample, whether respondents will be stratified by characteristics of interest, the targeted numbers of interviewees, and any incentives to be used in recruitment. States should also describe how they plan to process and analyze qualitative data, such as the application of a coding structure and analytic software, and the approaches they will use to identify key themes or taxonomies. Finally, states should indicate how they will use qualitative data in the evaluation, with a focus on how they will integrate qualitative and quantitative data.

e. Further considerations for national survey data

Existing national household surveys have several advantages as a data source. They are low cost, include items relevant to SMI/SED and SUD demonstrations, allow for comparisons with

¹³ More information about the BHICA can be found at <https://www.resourcesforintegratedcare.com/node/143>.

other states, and provide data for years both before and after demonstration implementation. Drawbacks of national surveys include that they do not allow states to isolate demonstration populations, are known to undercount Medicaid beneficiaries, and may not have sufficient sample sizes for the population of interest. States should determine whether samples for their population of interest contain enough observations to adequately power their planned analyses. In addition, most national surveys do not provide data until the year after they are fielded. Depending upon a state's evaluation period, it may be more feasible to include analyses of national survey data in summative reports than in interim reports.

One example of a national survey relevant for evaluations of SMI/SED and SUD demonstrations is the National Survey on Drug Use and Health (NSDUH). Sponsored by the Substance Abuse and Mental Health Services Administration, the NSDUH is a nationwide study of substance use, mental health, and other health-related issues in the United States. Each year, the NSDUH interviews approximately 70,000 people ages 12 and older who reside in households or noninstitutionalized group quarters. Other potentially relevant surveys include the Behavioral Risk Factor Surveillance System, which surveys adults ages 18 and older about health behaviors such as alcohol and tobacco use and generates state-level estimates, and the National Survey of Substance Abuse Treatment Services and National Mental Health Services Survey, which are annual surveys of service providers.

5. Analytic methods

In this section, states should provide details on all analytic methods planned for the evaluation, including each analysis listed in the design table submitted to CMS. CMS suggests a mixed-methods evaluation design that uses both qualitative and quantitative data to conduct descriptive and impact analyses. The objective of qualitative analyses is to understand demonstration operations and beneficiary experience and to support the design and interpretation of quantitative descriptive and impact analyses. The objective of quantitative analyses is to assess measured changes in demonstration outcomes, some of which may be attributable to the demonstration. States should explain how the evaluation will integrate findings from both types of analyses.

a. Further considerations for quantitative analyses

Suggested quantitative analyses include quasi-experimental and descriptive approaches. States can consult “Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations”¹⁴ and “Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations”¹⁵ from CMS for detailed discussions of analytic strategies, how they depend upon available comparison groups, and the extent to which they may support causal attribution. Although the first document applies these concepts through illustrations drawn from Section 1115 demonstrations with

¹⁴ “Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations” is available at <https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/causal-inference.pdf>.

¹⁵ “Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations” is available at <https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/comparison-grp-eval-dsgn.pdf>.

eligibility and coverage provisions, such as monthly payments, many of its recommendations can be generalized to evaluations of SMI/SED or SUD demonstrations.

- Quasi-experimental approaches such as regression discontinuity designs and difference-in-differences analysis may support causal inference, depending upon the strengths and limitations of the specific analysis. States may choose these approaches based on data and comparison group availability, but should also consider the trade-offs between them. For example, regression discontinuity designs offer strong internal validity, but the results do not necessarily generalize to the entire population subject to a demonstration.
- If it is not possible to identify a valid comparison group, states should consider using an interrupted time series design, which allows evaluators to (1) assess the impact of the demonstration on an outcome by using repeated pre- and post-demonstration observations to control for pre-existing trends and (2) assess changes in the strength and direction of the trend associated with implementation of the demonstration. Evaluators can strengthen this design by using regression analysis to control for other potential confounding factors. Evaluators should use an interrupted time series design only when several observations are available over longer time spans, making it most appropriate for summative rather than interim evaluations. Although the interrupted time series design is stronger than the pre-post design (which only compares one observation in each time period), neither accounts for unobserved external factors that may impact the outcome.
- Descriptive analyses of demonstration processes and outcomes may be necessary if no comparison group is available to address a particular research question or if power calculations suggest that a sample will be too small to observe a statistically significant effect. Descriptive analyses, including descriptive regressions, do not support causal inference about the effects of demonstration policies; therefore, states must interpret the results with caution. CMS will consider evaluation designs that are not experimental or quasi-experimental when a more rigorous design is not feasible.

b. Further requirements for the evaluation design

For each planned analytic approach included in a state's design tables (as submitted to CMS), the evaluation design should describe the target population and time points for data collection and outcome measures. For planned regressions, the evaluation design should specify the statistical model and list control variables. Depending upon the policy being evaluated and the expected outcome, important control variables may include demographic characteristics (such as age, race, and gender); household income; clinical characteristics (behavioral or physical health comorbidities); dual eligible status; and delivery system (managed care plan or fee-for-service).

c. Subgroup analyses

The suggested research questions for SMI/SED (Appendix A) and for SUD (Appendix B) do not include questions about subgroup effects. States should work with CMS to define important subgroups for their demonstration design and context. For example, it may be important to examine whether observed effects on overdose deaths differ by region, by presence and type of substance use disorder, by disability status, by racial or ethnic group, by age group, and/or by Medicaid eligibility group. States should plan to gather enough data to allow observation of

differences in relevant subgroup outcomes of at least 5 percent within 95 percent confidence intervals.

D. Methodological limitations

In this section, states should provide detailed information on the limitations of the evaluation. These could involve the availability of comparison groups, the sample sizes, the data sources, the collection process, or the analytic methods. States should also identify any efforts to minimize limitations, and they should acknowledge where limitations will preclude causal inferences about the effects of demonstration policies. In addition, this section should include any information about features of the demonstration that present methodological constraints that the state would like CMS to consider in its review.

E. Attachments

1. Independent evaluator

Describe the process the state will use to work with its independent evaluator to conduct the analysis and write the evaluation report. Explain how the state has involved the independent evaluator in the development of the evaluation design. Explain how the state will ensure that the independent evaluator will conduct a fair and impartial evaluation, prepare an objective evaluation report, and have no conflict of interest. The evaluation design should include “No Conflict of Interest” signed confirmation statements from the independent evaluator.

2. Evaluation budget

A budget for the evaluation shall be provided with the draft evaluation design. It should include the total estimated cost, as well as a breakdown of the estimated staff and the administrative and other costs for all aspects of the evaluation. Examples include but are not limited to the development and fielding of all survey and measurement instruments, quantitative and qualitative data collection, data cleaning and analyses, and report generation. A justification of the costs may be required by CMS if the estimates provided do not appear to sufficiently cover the costs of the draft evaluation design or if CMS finds that the draft evaluation design was not sufficiently developed. CMS may also request a justification if the costs appear excessively high or low relative to the proposed effort.

3. Timeline and major milestones

Describe the timeline for conducting the various evaluation activities, including dates for evaluation-related milestones (including those related to procurement of an outside contractor, if applicable) and deliverables. The final evaluation design shall include information regarding both interim and summative evaluations. Pursuant to 42 CFR 431.424(c)(v), this timeline should also include the due date for the final summative evaluation report.