Selection of Out-of-State Comparison Groups and the Synthetic Control Method

White Paper

October 2020

R. Vincent Pohl and Katharine Bradley

This white paper was prepared on behalf of the Centers for Medicare & Medicaid Services (CMS) as part of the Medicaid 1115 Demonstration Support Contract (contract number: HHSM-500-2014-00034I/75FCMC19F0008). Under the contract, Mathematica provides technical assistance focused on states' section 1115 demonstration evaluation designs and reports. This paper is intended to support states and their evaluators by describing how states can select comparison groups from other states and use the synthetic control method.

Contents

I. Introduction

Randomized controlled trials are the most rigorous approach to program evaluation, but they are often not feasible for evaluations of Medicaid section 1115 demonstrations due to political, ethical, or practical reasons. When randomizing beneficiaries is infeasible, the most rigorous choice for a non-experimental evaluation design involves a purposefully selected comparison group. As discussed in other evaluation resources released by the Centers for Medicare & Medicaid Services (CMS), selecting a comparison group that is similar to the demonstration group but not subject to demonstration policies can support causal inferences about demonstration effects.[1] In contrast, evaluation designs that track outcomes over time but do not use a comparison group make it more difficult to distinguish demonstration effects from the effects of confounding events, such as economic recessions or policy changes not related to section 1115 demonstrations.

States have two options when selecting a comparison group: (1) Medicaid beneficiaries or providers in the same state who are not subject to the demonstration policy being evaluated or (2) Medicaid beneficiaries or providers in other states that do not have the same demonstration policies. In-state comparison groups can be defined using characteristics that exempt the groups from the policy, such as Medicaid eligibility category, geographic location, age, or income. Out-of-state comparison groups are useful when a credible in-state comparison group is not available—for example, because the policy affects all Medicaid beneficiaries or providers, because beneficiaries not subject to section 1115 policies are very different from the demonstration population, or when states want to use both in-state and out-of-state comparison groups to increase confidence in their findings. This white paper focuses on selecting out-of-state comparison groups.

Several states with section 1115 demonstrations have included out-of-state comparison groups in their evaluation designs.[2] For example, West Virginia is using out-of-state comparisons to evaluate its section 1115 substance use disorder demonstration.[3] The evaluation of Montana's 2016–2020 demonstration included comparisons to other states that both did and did not expand Medicaid. Other states have explored the possibility of including out-of-state comparisons but have struggled to identify a suitable comparison state, either because other states have very different Medicaid programs or health system characteristics or because similar states use the same demonstration policies. This challenge may intensify as more states apply for section 1115 authority to test the same policies. Data availability for comparison

[1] See "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations" (Reschovsky et al. 2018) and "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations" (Contreary et al. 2018) at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

[2] The CMS-funded national evaluation of section 1115 demonstrations also used states without demonstrations as comparisons. Demonstration types in the national evaluation included (1) alternative Medicaid expansions with premium assistance focused on qualified health plans, premiums and other monthly payments, and beneficiary engagement/healthy behavior incentives; (2) managed long-term services and supports (MLTSS); and (3) delivery system reform incentive payment (DSRIP) programs. Evaluation designs are available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-federal-evaluation-meta-analysis/index.html.

[3] Approved evaluation designs are posted to the administrative record for each section 1115 demonstration: https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/index.html.

states must also be considered, given that evaluators must assess whether national surveys contain needed variables or whether they can obtain timely administrative data for other states.

To support states' efforts to include comparison groups for section 1115 demonstration evaluations in accordance with CMS guidance,[4] this white paper describes best practices for identifying suitable comparison states, including suggested selection criteria and data sources (Section II). We also describe a relatively new method of constructing a comparison group when there is no ideal comparison state, an approach called the synthetic control method (Section III).[5] The appendix provides an overview of recent extensions to the synthetic control method that states and their evaluators can use to strengthen synthetic control evaluation designs.

Section 1115 Medicaid demonstrations

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

[4] CMS guidance on developing evaluation designs and reports for section 1115 demonstrations is available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

[5] In an experiment, the group intentionally withheld from the intervention is typically called the control group, whereas in non-experimental evaluation designs, the group not subject to the intervention is referred to as the comparison group. Although we use the term "comparison group" in reference to non-experimental research designs, the literature on synthetic comparison groups uses the term "synthetic control method." We follow the literature and use the same term in our discussion of this method.

II. Selecting an Out-of-State Comparison Group

Randomized controlled trials are the best way to support causal inference because treatment and control groups are expected to be identical, except for the intervention being evaluated and any variation due to chance. If randomizing assignment to a demonstration or control group is not feasible, states should ensure that the comparison group comes as close as possible to representing the counterfactual (what would have happened if the intervention had not been implemented). In particular, the comparison group should have similar observed characteristics, be unexposed to the intervention, and be exposed to the reference policy environment (Contreary et al. 2018). A different white paper describes several comparison group options for section 1115 demonstration evaluations (Bradley et al. 2020), but this white paper focuses on data sources and methods, such as the synthetic control method, that are particularly helpful for out-of-state comparison groups.

Relevant observed characteristics might include demographic variables, beneficiaries' health status, or health care use. If providers are the target of the demonstration, evaluators should take into account both the similarity of providers and the beneficiaries they serve.[6] Variables that reflect the reference policy environment are usually state-level characteristics such as Medicaid eligibility levels, Medicaid managed care penetration, overall provider supply, and labor market features.

Choosing relevant variables for specific demonstration types. When considering which variables are most useful for assessing the similarity of comparison groups, states and their evaluators should prioritize the inclusion of confounding variables, which are variables that affect the outcome of interest and are related to the policy being evaluated. Doing so is important because it would otherwise be difficult or impossible to disentangle the effects of the demonstration policy from the effects of the confounding variable.

As a starting point, states should select variables that reflect the confounders they identify in their logic model for a demonstration policy (Contreary et al. 2018). For example, when evaluating a substance use disorder (SUD) demonstration designed to reduce overdose deaths, a good comparison state should have similar overdose trends as the state whose policy is being evaluated. Otherwise, it would not be clear whether differences in demonstration beneficiaries' employment outcomes are due to the policy or to differences in the availability of jobs. States should also include relevant variables measured at baseline, before the demonstration begins. For example, in a substance use disorder demonstration, it would be important to include substance use disorder rates before implementation of the demonstration policies as one of the variables to assess similarity between demonstration and comparison states.[7] States can disregard variables that are not related to the outcome and do not interact with the policy being evaluated. For example, the percentage of beneficiaries with diabetes might not be an essential characteristic for a comparison group used to evaluate a substance use disorder demonstration.

[6] For simplicity, we focus on beneficiaries in the remainder of this paper.

[7] There are limits to using preintervention beneficiary-level characteristics. When using formal matching methods, matching on preintervention characteristics can lead to biased impact estimates because of what is known as regression to the mean (Daw and Hatfield 2018). Regression to the mean refers to the fact that groups with very high (low) outcomes in the preintervention period often have lower (higher) outcomes in the postintervention period.

Matching methods

When using matching methods, researchers specify a range of matching variables relevant to the policy being evaluated with the goal of achieving balance—that is, equality of the distribution of these variables across intervention and comparison groups. At a high level, formal matching methods involve one of the following approaches:

- Identifying members of the comparison group who are similar to members of the intervention group based on a prespecified metric. Intervention and comparison group members who have no close match are often discarded from the analysis.
- Calculating a weight for each member of the comparison group that reflects how closely they match the demonstration group. Comparison group members with higher weights receive more importance in the analysis. An example is inverse propensity-score weighting. (The propensity score is the estimated probability that a beneficiary is a member of the demonstration group given his or her observed characteristics. Researchers estimate this probability for the demonstration and comparison group and use its inverse to reweight observations, which makes members of both groups balanced on observed covariates.)

Details may vary across matching methods. For example, comparison group members may be matched to intervention group members at a ratio of one to one or many to one, depending on the size of the pool of potential comparison group members. That is, researcher may be able to find multiple members of the comparison group who are similar to a single intervention group member based on the pre-determined matching criteria. Many-to-one matching can increase the precision of impact estimates. Comparison group members may be drawn with or without replacement, meaning that the same person in the comparison group may be determined to be the best match for multiple people in the intervention group. See Stuart (2010) for an overview, along with Imbens and Wooldridge (2009) and Imbens and Rubin (2015). The development of matching methods for policy evaluation is an active area of research (for example, see Iacus et al. [2019]).

Role of matching in comparison group selection. Use of statistical matching procedures is a common evaluation practice that helps to ensure similarity of demonstration and comparison groups (see box). After identifying a state or states that are similar to the demonstration state based on aggregate beneficiary and state characteristics (such as the overall percentage of beneficiaries with a substance use disorder), individual beneficiary characteristics may still not be well-balanced between demonstration and comparison states. To improve this balance, people in the comparison state can be matched to those in the demonstration state using individual-level characteristics. It is also possible to use statistical matching methods to select states, but evaluators more commonly apply matching to individuals within the comparison states after choosing states in a more ad-hoc way based on aggregate characteristics.[8]

Data availability considerations. States and their evaluators should consider available data sources when choosing variables on which to base the selection of comparison states. Many relevant data sources are publicly available national surveys, which enable comparisons of states along several dimensions.[9] For instance, the Behavioral Risk Factor Surveillance System (BRFSS) allows evaluators to estimate average state-level health status for Medicaid beneficiaries. The BRFSS and the American Community Survey are most likely to have samples of sufficient size to support state-to-state comparisons. Other national surveys, such as the Current Population Survey, the National Health and Nutrition Examination Survey,

[8] Evaluators can also use multilevel matching or matching on both state- and beneficiary-level (or provider-level) characteristics (see Arpino and Mealli [2011], Li et al. [2013], and Zubizaretta and Keele [2017]).

[9] Only restricted versions of some surveys, such as the National Health Interview Survey and the Medical Expenditure Panel Survey, contain state identifiers. Researchers can access these versions through special agreements with the agencies that collect these data.

and the Medical Expenditure Panel Survey, have smaller state-specific subsamples that may make them unsuitable, at least for states with smaller populations.

Some variables related to health service use, such as the number of visits to certain provider types, may not be available in surveys. In these cases, states can use administrative Medicaid data from the Transformed Medicaid Statistical Information System (T-MSIS), which provides enrollment, claims, and encounter files for Medicaid beneficiaries in all states. National survey and T-MSIS data might also be suitable sources for outcome measures. When selecting data sources to identify comparison states or to measure outcomes, states should consider the available release schedule for these data sets and whether it lines up with the evaluation timeline.[10]

Examples of variables to consider. Table II.1 provides a non-exhaustive list of variables that states and their evaluators can consider when selecting a comparison state. Evaluators are unlikely to use the entire list but can instead select the most important variables, given the demonstration design and intended outcomes. Table II.1 also lists suggested data sources for all variables, linked to web sites with more information.

Table II.1. Examples of variables for assessing the similarity of treatment and comparison states

Domain	Examples	Potential data sources
Beneficiary characteristics		
Demographic characteristics	• Percent non-White or Hispanic • Percent male • Age distribution • Percent living in poverty • Percent urban/rural • Percent without health insurance • Percent enrolled in employer-sponsored insurance • Percent without a high school degree • Percent with a college degree	• U.S. Census • American Community Survey • Current Population Survey
Health status of Medicaid beneficiaries	• Average self-assessed health status • Percent with diabetes • Percent with hypertension • Percent with substance use disorder • Percent with serious mental illness or serious emotional disturbance	• Behavioral Risk Factor Surveillance System • Youth Risk Behavior Surveillance System • National Health Interview Survey[a] • Medical Expenditure Panel Survey[a] • National Health and Nutrition Examination Survey[a] • National Survey on Drug Use and Health[a]

[10] States differ in the timeliness and quality of their data submissions to T-MSIS (see https://www.medicaid.gov/medicaid/data-systems/macbis/transformed-medicaid-statistical-information-system-t-msis/index.html). States can purchase T-MSIS research identifiable files for other states through ResDAC (see https://www.resdac.org/). They may also be able to acquire data through distributed data networks that facilitate cross-state sharing of aggregate information, such as the Medicaid Outcomes Distributed Research Network, a data-coordinating effort sponsored by a group of state universities (see https://egems.academyhealth.org/article/10.5334/egems.311/).

Domain	Examples	Potential data sources
Health care use among Medicaid beneficiaries	• Frequency of outpatient visits • Frequency of primary care visits • Frequency of specialist visits • Frequency of inpatient stays • Frequency of emergency department visits	• National Health Interview Survey[a] • Medical Expenditure Panel Survey[a] • Transformed Medicaid Statistical Information System Research Identifiable Files[b]
State-level characteristics		
Medicaid program characteristics	• Section 1115 demonstration policies • State-based or county-based program administration • Medicaid expansion to cover adult VIII group • Timing of Medicaid expansion • Income thresholds for different Medicaid eligibility groups • Percentage of population covered by Medicaid (use only from pre-implementation period if Medicaid coverage is an expected demonstration outcome) • Percentage of Medicaid beneficiaries in managed care, overall or by eligibility group • Medicaid value-based purchasing (yes/no indicators for different types of managed care quality initiatives, focus areas of performance measures)	• Centers for Medicare & Medicaid Services: section 1115 demonstrations, Medicaid managed care • Medicaid and CHIP Payment and Access Commission • Kaiser Family Foundation: state-specific information on Medicaid programs, Medicaid managed care
State health care market characteristics	• Hospital beds per 100,000 people • Physicians per 100,000 people • Provider characteristics (such as the fraction of solo practitioners or for-profit status of hospitals) • Level of market consolidation • Number of Federally Qualified Health Centers • Proportion of the state designated as health professional shortage areas • Number and characteristics of mental health and substance use treatment facilities	• Health Resources and Services Administration's Area Health Resource File • National Survey of Substance Abuse Treatment Services • National Mental Health Services Survey
State labor market characteristics	• Employment to population ratio • Unemployment rate • Number of jobs • Average wages	• American Community Survey • Current Population Survey • Quarterly Census of Employment and Wages • Local Area Unemployment Statistics

[a] This data source may only support evaluations involving the most populous demonstration and comparison states.

[b] States and evaluators may also be able to access administrative claims data through direct data sharing arrangements with other states, all-payer claims databases, and distributed data networks that arrange for aggregate-level data sharing among groups of states.

CHIP = Children's Health Insurance Program.

III. Synthetic Control Methods

A. Overview

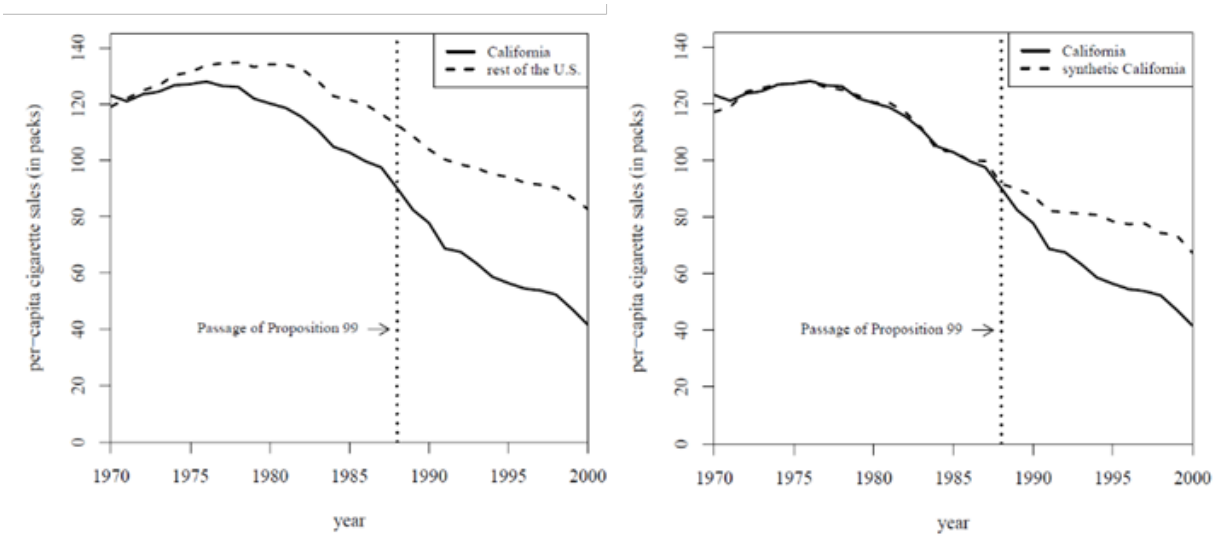
When states and their evaluators cannot find a suitable single state or set of states for a comparison group based on relevant beneficiary- and state-level variables, they can use the synthetic control method as an alternative approach to estimate the impact of a policy.[11] This method involves constructing a single comparison group from a pool of potential comparison states (the "donor pool") by combining them so that the newly constructed (synthetic) comparison group resembles the treatment group as closely as possible on levels and trends in preintervention outcomes. For example, a state may want to evaluate its section 1115 substance use disorder demonstration by comparing the outcomes with those in other states, but some potential comparison states may have higher rates of SUD among their Medicaid beneficiaries and others may have lower rates. In this case, the evaluator could compute the average SUD rate for the states with higher and lower rates to construct a single comparison group that had a similar rate of SUD before the intervention began.

With the synthetic control method, evaluators should use a large donor pool of potential comparison states, such as all states that have expanded Medicaid to cover the adult VIII group, and then assign different weights to each potential comparison state to form a weighted average. The method may assign a weight of zero to some or many of the potential comparison states, which implies that these states will not be part of the weighted average. One advantage of the synthetic control method is the intuitive appeal of these weights; a reader without a statistical background can easily see which states enter the synthetic comparison group and which do not.

Figure III.1 shows a graphical example of how the synthetic control method can be applied. This example comes from Abadie et al. (2010), who used the synthetic control method to estimate the impact of a cigarette tax increase on cigarette sales in California in 1988. The left panel shows that California has lower cigarette sales than the rest of the U.S., so it is not possible to evaluate the impact of the policy on cigarette sales by comparing the change in the two time series. Instead of using individual states as the comparison group, the synthetic control group method assigned weights of 16 percent to Colorado, 7 percent to Connecticut, 20 percent to Montana, 23 percent to Nevada, 33 percent to Utah, and 0 to all other states. As shown in the right panel, California and the synthetic comparison group had almost identical cigarette sales trajectories before the policy took effect, thereby allowing the authors to estimate the impact of the policy by comparing the difference in cigarette sales between California and "synthetic California."

[11] The synthetic control method was first proposed by Abadie and Gardeazabal (2003).

Figure III.1. Graphical example of the synthetic control method



Source: Abadie, A., A. Diamond, and J. Hainmueller. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association*, vol. 105, no. 490, June 2010, pp. 493–505. doi: 10.1198/jasa.2009.ap08746. Version: Authors' final manuscript. Reproduced under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported license (see https://creativecommons.org/licenses/by-nc-sa/3.0/legalcode).

To date, there are no published studies using the synthetic control method to study section 1115 demonstrations, but several studies have drawn on this method to evaluate other Medicaid policies. For example, Ghosh and Simon (2015) estimated the impact of a 2005 contraction in Medicaid coverage in Tennessee on hospitalization rates. To complement their difference-in-differences analysis using all other southern states as the comparison group, the authors used the synthetic control method to construct an additional comparison group. Other authors have assessed impacts of Medicaid expansions on different outcomes using the synthetic control method. For example, Freedman et al. (2017) estimated impacts on payer mix and hospitalizations, Kaestner et al. (2017) on health insurance coverage and labor supply, Peng (2017) on health insurance marketplace premiums, and Hu et al. (2018) on financial well-being. These authors used the synthetic control method because non-expanding states are different from expanding states in many ways, making them inappropriate comparison groups. Constructing synthetic comparison groups consisting of the weighted average of non-expanding states mitigated this issue.

B. Description of the method

This section provides a nontechnical description of the synthetic control method to give states a sense of how their evaluators would select a synthetic comparison group.[12] The standard synthetic control method assumes that there is one treated unit (for example, the state implementing the section 1115 demonstration) and a "donor pool" of potential comparison units (other states).[13] For all units, researchers observe the preintervention characteristics (for example, some of the variables listed in Table II.1) for years before the demonstration is deployed and the expected demonstration outcomes (for example, rates

[12] Abadie et al. (2015) provide a more technical description of the method.

[13] There are extensions to the synthetic control method that allow for multiple treated units. See the appendix for details.

of preventive care use if studying a healthy behavior incentive). The synthetic control method requires that data on all relevant measures are available for all states in the donor pool.

Therefore, to use this approach, it is necessary to collect aggregate (state-level) data from before the demonstration policy took effect, often including data on the expected demonstration outcomes. Because evaluation designs that use out-of-state comparison groups typically rely on national survey data or administrative Medicaid data, preintervention data are likely to be available.

Role of preintervention outcomes. In the first applications of the synthetic control method, researchers used only preintervention outcomes to obtain synthetic comparison groups, as in the example shown in Figure III.1. More recent applications also include "covariates"—or variables that are not outcomes—instead of or in addition to preintervention outcomes. Adding covariates can help making the synthetic comparison group more similar to the demonstration group based on some of the variables in Table II.1 instead of pre-intervention outcome trends alone.

When including preintervention outcomes to match the outcome trajectories of the treatment and synthetic control groups more closely, it helps to use as many periods of data as possible. In Figure III.1, for instance, the authors of the California study used almost 20 years of cigarette sales data before passage of the policy. Such a long time series is typically not feasible for evaluations of section 1115 demonstrations, but states and their evaluators could use several years of outcome data from some of the national surveys listed in Table II.1. For example, the BRFSS goes back several decades and can be used to estimate health and access to health care of Medicaid beneficiaries by state, which are relevant outcomes for several types of demonstrations. States can use their logic models for demonstration policies to identify relevant outcomes to include.

Tip: Narrow down the donor pool

It may be tempting to include all 50 states and the District of Columbia in the donor pool, but it may help to narrow down the pool as a first step.

- States that have deployed the same demonstration policies cannot be part of the donor pool because they would also be considered treated states.
- States in the donor pool should have a similar overall policy environment as the treated state. For example, it may be appropriate to only include states that have expanded Medicaid to cover the adult VIII group. This would avoid a situation in which a comparison state receives a large weight because of similarities in some covariates but does not share basic features of the Medicaid program with the treated state.
- States that experienced a large pre-implementation shock that sets them apart from the demonstration state should not be included in the donor pool. For example, states that have experienced a spike in opioid overdose rates may not be suitable for the donor pool in the evaluation of SUD demonstrations.

Synthetic control weights. When applying the synthetic control method to a list of preintervention characteristics and outcomes for the treated and potential comparison units, an algorithm will assign a single, optimal weight to each potential comparison unit such that the treated and potential comparison units are as similar as possible along the covariates and outcomes entering the model. If this weight is positive, the corresponding unit enters the synthetic comparison group; if the weight is zero, the unit does not enter the comparison group. The weights for all comparison units in the donor pool sum to one.

Relative importance of covariates. The synthetic control method is flexible in that it allows states and their evaluators to specify the relative importance of each preintervention covariate and outcome. For example, states may decide to construct a synthetic comparison group based on several covariates described in Section II, but they would like to emphasize those that are particularly relevant for the demonstration policy. In general, evaluators should place higher importance on variables that are more predictive of the outcome of interest. For example, in an evaluation of a substance use disorder demonstration in which an expected outcome is reduced admissions to hospitals and inpatient psychiatric facilities, important covariates could be the rate of substance use disorder among demonstration beneficiaries and the number of available treatment providers. Although evaluators may also want to build a synthetic comparison group that has similar demographic characteristics, they could assign lower relative importance to demographic covariates except when these covariates describe the prevalence of substance use disorder and the provider infrastructure.

Tip: Plot outcomes to visually inspect the synthetic comparison

A convenient feature of the synthetic control method is the ease with which its results can be presented graphically. As shown in Exhibit III.1, it is immediately clear how constructing a synthetic comparison group yields a more convincing counterfactual than using all untreated states with equal weights. Evaluators should plot outcomes over time for the demonstration group and the synthetic comparison group—not only to check whether the synthetic control method yielded a sufficiently similar comparison group, but because such a graph is a simple way to convey impact estimates to stakeholders.

Different comparison groups for different outcomes. The synthetic control method can be applied to one outcome measure at a time and can therefore produce a different comparison group for each outcome. That is, researchers can create a different synthetic comparison state for each outcome of interest. For example, in an evaluation of a retroactive eligibility policy, a different set of weights could be used to create a synthetic comparison state in an analysis of insurance coverage than the set of weights used to create a synthetic state in an analysis of health outcomes. This advantage of the synthetic control method allows evaluators to construct a more credible counterfactual for each outcome of interest. Weights used for the analysis of employment levels might place high importance on similarity in labor markets, whereas the weights used for the analysis of health outcomes might place high importance on beneficiaries' underlying health status. In addition, for each outcome of interest, evaluators typically include the corresponding preintervention outcome when estimating synthetic control weights; synthetic comparison groups also vary for this reason.

Estimation of policy impacts. Once the synthetic control method has determined the weights, estimation of impacts is straightforward. To estimate the impact of a demonstration policy, evaluations compare the mean outcome in the treatment group (Medicaid beneficiaries subject to the policy in the demonstration state) to the weighted mean outcome among beneficiaries in the comparison states, where the weights are given by the synthetic control method algorithm. Specifically, the impact estimate equals the difference between the treatment group mean and the weighted comparison group mean. There are publicly available software scripts that implement these estimates for the standard statistical packages.[14]

Statistical inference. To understand whether an estimated impact derived via the synthetic control method represents a "true" impact and not just noise, standard errors or confidence intervals are needed. These are typically obtained using falsification or placebo exercises. The details are beyond the scope of

[14] See https://web.stanford.edu/~jhain/synthpage.html for Stata and R packages.

this white paper, but briefly, these tests involve assigning a hypothetical demonstration policy to states in the donor pool that, in reality, did not implement a similar demonstration and then estimating the impact as if the donor states had actually implemented the policy. If this calculation yields a sizable impact estimate, it would cast doubt on the reliability of the initial impact estimate for the demonstration state.

C. The synthetic control method versus difference-in-differences designs

Existing section 1115 demonstration evaluations that involve an out-of-state comparison group typically rely on a difference-in-differences design. That is, they compare the outcomes of interest in the demonstration and comparison groups before and after the demonstration policy takes effect.[15]

The synthetic control and difference-in-differences methods have some features in common, but there are also some important differences. In the context of section 1115 demonstrations and out-of-state comparisons, both methods use other states to construct a counterfactual for the state that deployed the policy of interest. Difference-in-differences models simply take outcomes in comparison states and compare them to outcomes in the demonstration state, both before and after the demonstration was implemented, whereas the synthetic control method first derives weights using pre-intervention outcomes to construct a single synthetic comparison group before comparing post-intervention outcomes to derive an impact estimate. The data requirements for both approaches are thus similar, as both use data on preintervention outcomes but in different ways. Both methods are essentially longitudinal, and as a result, both are more robust when using multiple preintervention periods. However, it is possible to use difference-in-differences with only one pre-intervention period, whereas the synthetic control method always requires multiple pre-intervention periods.[16]

Another conceptual distinction is the way in which the methods assess similarities in preintervention outcomes. The key assumption that enables researchers to estimate causal effects using difference-in-differences is the parallel trends assumption—meaning that outcomes would have evolved in parallel in demonstration and comparison states in the absence of the demonstration. Whereas parallel trends can only be tested in difference-in-differences designs, applying the synthetic control method imposes this similarity, as shown in Figure III.1. The left panel in the figure shows nonparallel preintervention trends, which would preclude a difference-in-differences design. In contrast, the synthetic control method ensures that the intervention and synthetic comparison groups have the same preintervention outcomes because these outcomes entered the algorithm that was used to calculate the weights.

In summary, the synthetic control method and difference-in-differences models are both suitable methods for evaluations that involve out-of-state comparison groups. States can use difference-in-differences when they are able to identify groups of beneficiaries from one or more other states that have similar pre-intervention trends as the demonstration group. This approach does not necessarily require multiple years of pre-intervention data. In contrast, the synthetic control method can be a strong research design when no state has similar pre-intervention trends, but multiple years of pre-intervention data are available for other states, enabling construction of a synthetic comparison group.

---

[15] See "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations" (Contreary et al. 2018) for more details on difference-in-differences methods. Available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

[16] Computationally, although it is possible to implement a difference-in-differences design with one preintervention observation, the parallel trends assumption can be assessed only with multiple preintervention observations.

IV. Conclusions

Evaluation designs involving comparison groups make evaluations of section 1115 demonstrations more robust because they allow states and evaluators to disentangle policy effects from other concurrent changes that affect outcomes. Evaluators can use out-of-state comparison groups when no in-state comparison group is available or to corroborate findings from other evaluation strategies. This white paper provides practical suggestions for selecting or constructing out-of-state comparison groups to support the use of this strategy.

When choosing out-of-state comparison groups, states and their evaluators should select other states with observed characteristics and pre-intervention trends in the outcomes of interest that are similar to those of the demonstration state. This paper describes several types of characteristics to consider. If an out-of-state comparison group is difficult to find because no other states are similar to the demonstration state, the synthetic control method can help with constructing a comparison group. This method has been used in published research on several Medicaid-related outcomes.

References

Abadie, A., A. Diamond, and J. Hainmueller. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science*, vol. 59, no. 2, April 2015, pp. 495–510. doi: 10.1111/ajps.12116

Abadie, A., A. Diamond, and J. Haimueller. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association*, vol. 105, no. 490, June 2010, pp. 493–505. doi: 10.1198/jasa.2009.ap08746

Abadie, A., and J. Gardeazabal. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review*, vol. 93, no. 1, March 2003, pp. 113–132. doi: 10.1257/000282803321455188

Arpino, B., and F. Mealli. "The Specification of the Propensity Score in Multilevel Observational Studies." *Computational Statistics and Data Analytics*, vol. 55, no. 4, April 2011, pp. 1770–1780. doi: 10.1016/j.csda.2010.11.008

Ben-Michael, E., A. Feller, and J. Rothstein. "The Augmented Synthetic Control Method." 2018. Available at https://eml.berkeley.edu/~jrothst/workingpapers/BMFR_Synth_Nov_2018.pdf. Accessed June 30, 2020.

Bradley, K., J. Heeringa, R.V. Pohl, J.D. Reschovsky, and M. Samra. "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations." Washington, DC: Mathematica, revised October 2020.

Contreary, K., K. Bradley, and S. Chao. "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations." Oakland, CA: Mathematica Policy Research, June 2018.

Daw, J.R., and L.A. Hatfield. "Matching and Regression to the Mean in Difference-in-Differences Analysis." *Health Services Research*, vol. 53, no. 6, December 2018, pp. 4138–4156. doi: 10.1111/1475-6773.12993

Dube, A., and B. Zipperer. "Pooling Multiple Case Studies Using Synthetic Controls: An Application to Minimum Wage Policies." IZA Discussion Paper No. 8944. March 2015. Available at http://hdl.handle.net/10419/110107. Accessed June 25, 2020.

Ferman, B., C. Pinto, and V. Possebom. "Cherry Picking with Synthetic Controls." *Journal of Policy Analysis and Management*, vol. 39, no. 2, 2020, pp. 510–532. doi: 10.1002/pam.22206

Freedman, S., S. Nikpay, A. Carroll, and K. Simon. "Changes in Inpatient Payer-Mix and Hospitalizations Following Medicaid Expansion: Evidence from All-Capture Hospital Discharge Data." *PLoS ONE*, vol. 12, no. 9, September 2017. doi: 10.1371/journal.pone.0183616

Ghosh, A., and K. Simon. "The Effect of Medicaid on Adult Hospitalizations: Evidence from Tennessee's Medicaid Contraction." NBER Working Paper Series 21580. Cambridge, MA: National Bureau of Economic Research, September 2015. Available at http://www.nber.org/papers/w21580. Accessed June 25, 2020.

Hollingsworth, A., and C. Wing. "Tactics for Design and Inference in Synthetic Control Studies: An Applied Example Using High-Dimensional Data." May 2020. Available at https://osf.io/preprints/socarxiv/fc9xt/. Accessed June 25, 2020.

Hu, L., R. Kaestner, B. Mazumder, S. Miller, and A. Wong. "The Effect of the Affordable Care Act Medicaid Expansions on Financial Well-Being." *Journal of Public Economics*, vol. 163, July 2018, pp. 99–112. doi: 10.1016/j.jpubeco.2018.04.009

Iacus, S.M., G. King, and G. Porro. "A Theory of Statistical Inference for Matching Methods in Causal Research." *Political Analysis*, vol. 27, no. 1, January 2019, pp. 46–68. doi: 10.1017/pan.2018.29

Imbens, G.W., and D.B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction.* New York: Cambridge University Press, 2015.

Imbens, G.W., and J.M. Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, vol. 7, no. 1, March 2009, pp. 5–86. doi: 10.1257/jel.47.1.5

Kaestner, R., B. Garrett, J. Chen, A. Gangopadhyaya, and C. Fleming. "Effects of ACA Medicaid Expansions on Health Insurance Coverage and Labor Supply." *Journal of Policy Analysis and Management*, vol. 36, no. 3, 2017, pp. 608–642. doi: 10.1002/pam.21993

Li, F., A. Zaslavsky, and M. Landrum. "Propensity Score Weighting with Multilevel Data." *Statistics in Medicine*, vol. 32, no. 19, August 2013, pp. 3373–3387. doi: 10.1002/sim.5786

Peng, L. "How Does Medicaid Expansion Affect Premiums in the Health Insurance Marketplaces? New Evidence from Late Adoption in Pennsylvania and Indiana." *American Journal of Health Economics*, vol. 3, no. 4, 2017, pp. 550–576. doi: 10.1162/ajhe_a_00088

Powell, D. "Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?" Working Paper WR-1246. Santa Monica, CA: RAND Corporation, May 2018. doi: 10.7249/WR1246

Stuart, E.A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1–21. doi: 10.1214/09-STS313

Zubizarreta, J., and L. Keele. "Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System." *Journal of the American Statistical Association*, vol 112, no. 518, 2017, pp. 547–560. doi: 10.1080/01621459.2016.1240683.

Appendix: Recent Methodological Advances in the Synthetic Control Method

Several recent papers have proposed extensions to the synthetic control method that may be helpful for evaluation of section 1115 demonstrations who use an out-of-state comparison group. We briefly describe a few of these extensions below.

A. Pooling multiple case studies

The synthetic control method was developed to estimate the impact of a policy for one specific unit, such as a state that is implementing a section 1115 demonstration. Although this is the typical case for the evaluation of a section 1115 demonstration, in some instances, more than one state might adopt a policy. For example, researchers might want to evaluate the impacts of substance use disorder demonstrations across several states that have the same policy. In this case, Dube and Zipperer (2015) proposed estimating separate impact estimates for each policy change and then aggregating these estimates by calculating their mean or median. They also showed how to conduct statistical inference for this pooled estimate.

B. Imperfect synthetic control

A recent paper by Powell (2018) relaxed two assumptions underlying the synthetic control method: (1) the weighted average of potential comparison units always exists, and (2) the synthetic comparison group matches the preintervention outcome of the intervention group perfectly. Both assumptions are often not met in practice.

Powell developed an alternative approach that treats each state in the donor pool the same way that the intervention state is treated in the standard synthetic control method—that is, it also estimates synthetic control weights for each donor pool state. The next step is to estimate policy impacts by aggregating these weights over all possible comparison units. Finally, instead of using observed preintervention outcomes (which could be measured with noise), Powell's approach involves use of predicted preintervention outcomes when estimating synthetic control weights. Predicted outcomes are based on state-specific outcome trends.

C. Augmented synthetic control method

Using the synthetic control method, researchers can use multiple covariates and preintervention outcomes to estimate the optimal weights used to construct the synthetic comparison group. However, the more variables there are, the more difficult it becomes to construct a synthetic comparison state that is an exact match for the demonstration state (that is, a comparison state that has the same observed characteristics along several dimensions), which can bias the estimated impacts for a single outcome. This problem is known as the curse of dimensionality.

To solve this issue, Ben-Michael et al. (2018) proposed the augmented synthetic control method. With this method, researchers first estimate the bias by specifying a model for the outcome of interest and then adjust for this bias in the final impact estimate. Ben-Michael et al. showed in a simulation study that this method can improve the balance between the intervention and synthetic control groups.

D. Synthetic control using LASSO

As described above, the synthetic control method involves a number of decisions. For example, researchers must decide which states to include in the donor pool, which covariates and preintervention outcomes to use when calculating optimal weights, and how much relative importance to place on each covariate or outcome. The resulting impact estimate depends on these choices, but there are no general rules on how to make these decisions.

Hollingsworth and Wing (2020) applied machine-learning methods to this problem. They proposed a method for choosing among several sets of potential synthetic control weights, with a preference for less-complex models unless the data demand a model with higher complexity.[17] Intuitively, weights are shrunk toward or all the way to zero if the corresponding unit from the donor pool does not contribute sufficiently to improving the balance between the intervention and synthetic control groups.

E. Cherry picking with synthetic controls

With many different possible specifications—for example, regarding the number of pre- and postintervention periods and the number of covariates used in the synthetic control methods—it is possible to obtain different impact estimates for the same policy change. This may tempt "cherry picking" the specification that yields the most favorable finding. To avoid this problem, Ferman et al. (2020) recommend strategies to limit the possibility for specification searching. For example, they suggest always presenting results for many different specifications. They also show that using fewer than all available preintervention outcomes is preferred when other covariates enter the calculation of synthetic control weights.

[17] The method relates to LASSO regression models. LASSO (which stands for least absolute shrinkage and selection operator) regressions add an additional term to a standard regression model that emphasizes some desired quality of the resulting estimate. In this context, a LASSO regression prefers a synthetic comparison group consisting of fewer states if that is sufficient to obtain a balanced comparison group.

Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

mathematica.org