Evaluation of Section 1115 Demonstrations That Change Over Time

White Paper

July 2026

R. Vincent Pohl and Monica Farid

This white paper was prepared on behalf of the Centers for Medicare & Medicaid Services as part of the Medicaid 1115 Demonstration Support Contract (contract number: 75FCMC19D0091/75FCMC24F0198). Under the contract, Mathematica provides technical assistance focused on states' section 1115 demonstration evaluation designs and reports. This white paper is intended to support states and their evaluators by providing a nontechnical summary of best practices for supporting evaluations of section 1115 demonstrations that change over time.

Contents

Tables

Figures

I. Introduction

In states with section 1115 demonstrations, Medicaid policies vary over time. Policy changes occur when a state first implements a demonstration, but many states also add or remove demonstration policies during a demonstration period via amendments. States also commonly change policies between an initial demonstration and its renewal. These policy changes over time both present an opportunity for the evaluation of demonstrations and can introduce complications that states and their independent evaluators need to address. In addition, most demonstrations include several distinct policies. States may implement these policies under a single demonstration approval or through multiple Medicaid authorities. When multiple authorities are involved, they may differ in their goals, eligible populations, geographic scope, implementation timelines, or reporting requirements, making the policies difficult to evaluate as a single, coordinated intervention. Such complex and multi-component demonstrations require careful evaluation planning to account for the differential effects of changes in the composition of authorities within a demonstration.

Evaluations of section 1115 demonstrations typically use data from before the demonstration start date and during the demonstration period to assess how changes in outcomes over time relate to the introduction of demonstration policies.[1] Methods that evaluators can use to analyze these data and to estimate demonstration impacts include difference-in-differences, interrupted time series, and pre-test/post-test designs (see box).

When a state introduces a demonstration for the first time and implements all demonstration policies on the same date, it is straightforward to define the pre-period and post-period for evaluation purposes. In this case, the period before implementation (including the planning period, if applicable) is the baseline or pre-period, while the period starting at implementation is the demonstration or post-period. In practice, however, many demonstrations depart from this clear-cut pattern. This white paper seeks to provide guidance to states for evaluating section 1115 demonstrations in the context of the common but challenging scenarios described below.

Empirical time-series methods

Difference-in-differences design: In this design, evaluators measure the pre-post difference in an outcome for the demonstration group minus the pre-post difference for the comparison group.

Interrupted time-series design: In this design, which can be employed when an intervention occurs at a specific point in time, data are collected at several points before and after the intervention (a time series). Evaluators compare the level and slope of the post-intervention time series to the pre-intervention time series.

Pre-test/post-test design: This nonexperimental design lacks a comparison group but includes pre-intervention observations that enable evaluators to measure the change in outcomes between the periods before and after the intervention.◢

[1] An exception is a purely cross-sectional design that uses only data from the post-period ("posttest-only with nonequivalent groups design"). This research design does not enable evaluators to account for secular trends that may differ between the demonstration and comparison group and is only appropriate if there are no time invariant confounding factors or trend differences between the groups. See "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations" (Bradley et al. 2020), available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/comparison-grp-eval-dsgn.pdf.

Demonstrations that add or amend policies during the demonstration period. States do not always implement all demonstration policies at the same time (see example in panel A of Figure 1). States may add new policies via an amendment midway through a demonstration approval period. They may also implement some demonstration policies later. For example, after CMS issued guidance for demonstration initiatives to improve care transitions for incarcerated individuals in April 2023, several states amended existing demonstrations to include a reentry initiative, and several of these demonstrations anticipated a delayed implementation period before they would begin delivering services due to the novel systems coordination challenges this policy presents.[2] Aside from planned later-implementation start dates, states may also delay implementing demonstration policies due to state legislative approval or unanticipated resource constraints. This was particularly common during the COVID-19 public health emergency when states delayed or paused implementing demonstration policies that could have increased beneficiary cost-sharing or led to disenrollment. States may also amend policies that were part of the original demonstration approval, such as eligibility conditions or benefit structures, to reflect implementation challenges, stakeholder feedback, or changing state legislative priorities. There are several challenges in contexts like these, including which intervention start dates to use and how to evaluate the impact of both the demonstration as a whole and of individual policies.

Demonstrations that introduce multiple individual policies at the same time. Many section 1115 demonstrations fall into this category (see example in panel B of Figure 1). For example, during the 2022–2026 approval period, the Arkansas Health and Opportunity for Me (ARHOME) demonstration included premium assistance for coverage through Qualified Health Plans and supports for people with high-risk pregnancies, people in rural areas with substance use disorder (SUD) or serious mental illness, and young adults at risk for long-term poverty.[3] Some demonstrations that introduce multiple policies at the same time operate under a single authority while other such demonstrations operate under multiple authorities that are aligned in their goals and intended outcomes or under substantially different authorities. For example, the California Advancing and Innovating Medi-Cal (CalAIM) initiative was implemented through a coordinated set of section 1115, section 1915(b), and Medicaid state plan authorities.[4] Different authorities authorize different components of the initiative: in the original CalAIM approval package, 12 Community Supports, including housing navigation and medically tailored meals, were approved through the Section 1915(b) waiver, while recuperative care and short-term post-hospitalization housing were authorized through the section 1115 demonstration. Although these components have different legal and administrative bases, they were designed to advance the same broader goals of integrated, person-centered care, and improved health outcomes. When all demonstration policies are implemented at the same time and may impact the same beneficiaries, it is generally difficult to isolate the impacts of individual policies.

---

[2] CMS issued this guidance in state Medicaid director letter #23-003, available at https://www.medicaid.gov/federal-policy-guidance/downloads/smd23003.pdf. States that amended demonstrations to include a reentry initiative include Arizona, Kentucky, Montana, and Vermont, among others.

[3] Demonstration approval and other documents for the ARHOME demonstration are available at https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/126401.

[4] Demonstration approval and other documents for the CalAIM demonstration are available at https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/81046.

Demonstrations implemented in phases by region, population, or other features. Finally, some states do not implement demonstration policies at the same time for all eligible beneficiaries or in all parts of the state (see example in panel C of Figure 1). A staggered roll-out by county or other sub-state area may be necessary because it takes more time to develop the required resources and infrastructure in some parts of the state—for example, when there are not enough providers in rural areas or new data sharing protocols must be established with key implementation partners. For example, the matching plan policy for dually eligible beneficiaries, which is part of the CalAIM demonstration for the 2022–2026 approval period, was operational in 12 of the state's counties in 2023 and five additional counties in 2024. Other instances of staggered roll-out include reentry demonstrations where states often phase in benefits sequentially by carceral facility, as different facilities become ready for program implementation. While scenarios like this can seem challenging to evaluate, they also provide an opportunity for implementing rigorous empirical designs that exploit the variation in when different demonstration populations are exposed to demonstration policies.

Figure 1 illustrates the three cases above. In panels A and B, the initial demonstration approval date is in 2018, and the renewal approval date is in 2023. In panel A (demonstrations that add or change policies over time), the demonstration adds a new policy at the start of the demonstration renewal period (2023) and two other new policies during the renewal period (in 2025 and in 2027).[5] In panel B (demonstration renewals with multiple new policies), the demonstration introduces new policies or changes existing polices at the start of the renewal period (2023). In panel C (demonstrations with staggered rollout), a demonstration policy becomes effective in different counties at different times. For example, it becomes effective in County A in 2021, followed by county B in 2024. County C is not affected by the demonstration throughout the study period.

The remainder of this paper discusses mixed methods approaches for evaluating section 1115 demonstrations in the categories above. Rigorous and comprehensive evaluations of section 1115 demonstration encompass an in-depth assessment of demonstration implementation (implementation research) and quantitative analyses of demonstration outcomes. Supplementing quantitative methods with comprehensive contextual analyses is particularly important in settings when implementation did not proceed as planned.

Implementation research

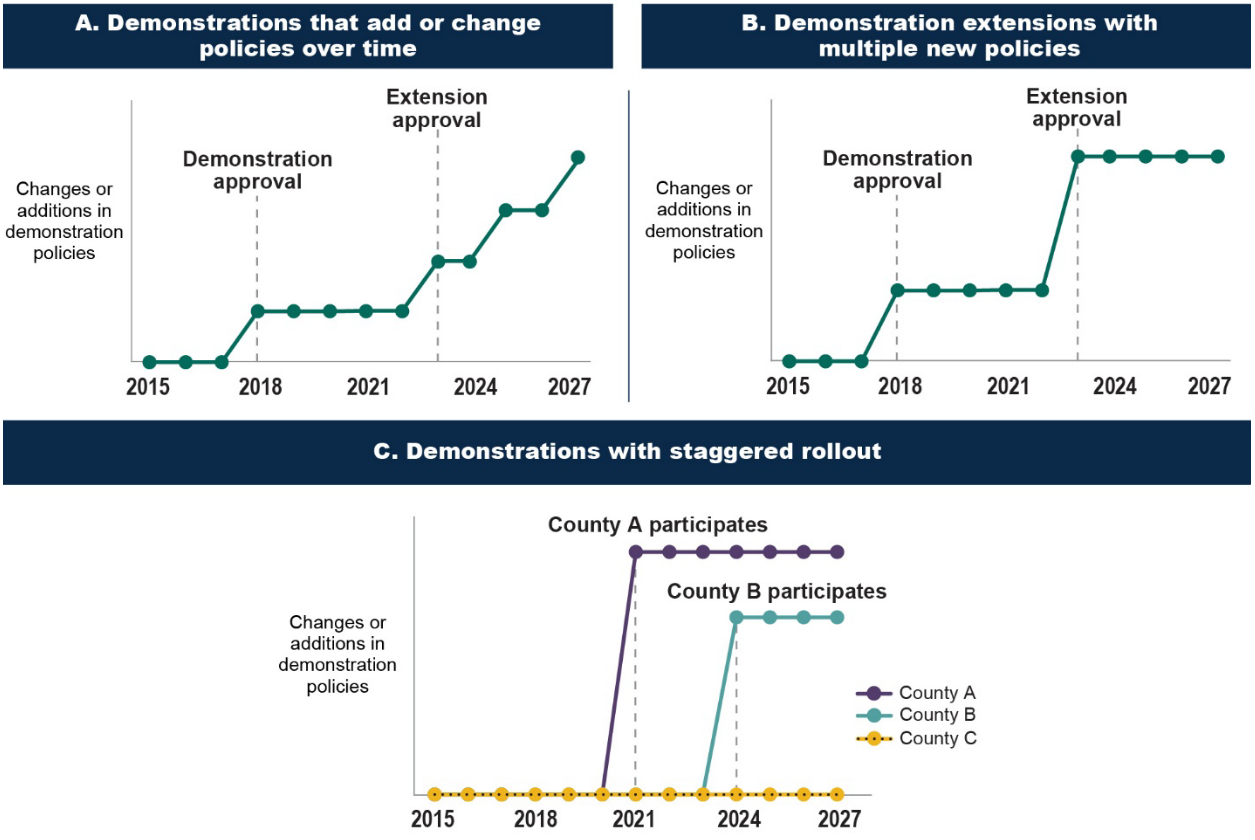
Implementation research is the study of how an intervention—such as a section 1115 demonstration—is put in place, how it proceeds over time, and the reasons why it does or does not proceed as expected (Felland and Bradley 2020). Although an intervention's design anticipates certain outcomes and an impact evaluation assesses in retrospect whether the intervention achieved those outcomes, implementation research intentionally tracks the intervention closer to real time to understand the reasons behind its progress (Felland and Bradley 2020; Werner 2004).

The paper discusses evaluation approaches for three categories of section 1115 demonstrations: demonstrations that add or change policies over time (Section II); demonstrations that introduce multiple individual policies at the same time (Section III); and demonstrations implemented in different parts of the state at different times (Section IV). We provide

[5] There are several reasons why all demonstration policies may not be introduced at the same time, including: (1) demonstration policies that were initially approved were rolled out in a staggered manner; (2) the state introduced an amendment during the demonstration period; or (3) a previously paused demonstration policy resumed.

concluding remarks in Section V. The appendix summarizes recent methodological advances that can help states evaluate demonstrations with a staggered rollout.

Figure 1. Illustration of different types of policy changes in section 1115 demonstrations



II. Demonstrations That Add or Change Policies During an Approval Period

In this section, we discuss demonstrations that introduce individual policies at different times during a demonstration period. Some states plan to add demonstration policies at different times when applying for demonstration approval. For example, during the 2022–2027 approval period of the Oregon Health Plan demonstration, housing supports for populations at risk of homelessness were scheduled for no sooner than November 2024 and benefit eligibility of youth with special healthcare needs started in January 2025.[6] Demonstrations also may experience unplanned delays in implementation or suspend programs due to unforeseen external factors. Other states may introduce a policy in stages, such as starting with a subset of services and later expanding them. For example, under the TennCare III demonstration during the 2021–2030 approval period, Tennessee initially implemented a limited subset of home and community-based services for targeted populations and later expanded the scope and availability of these services as capacity and infrastructure developed.[7] In either scenario, states should use a mixed methods approach to understand implementation and guide the evaluation, as well as estimate period-specific demonstration impacts. In cases where policies affect different populations, states can also estimate individual policy impacts. This applies to cases where policies are introduced at different times or at the same time, as long as the populations they affect are distinct. For example, if a state implements a SUD demonstration and a premium assistance policy at the same time, it can estimate impacts of these policies separately because they are likely to have different eligibility criteria.

A. Use mixed methods approaches to understand how demonstration policies were implemented and to guide the impact evaluation design

States should use qualitative and quantitative implementation data to guide the impact evaluation. Understanding the implementation context is important for evaluations as it informs the most suitable evaluation approach. For example, if ramp-up was slower than planned or occurred in several stages, states could consider estimating period-specific impacts for each phase of the post-period. Gathering information on the implementation process is especially important in demonstrations experiencing unplanned delays, because these may deviate more from the planned implementation.

States should carefully document when demonstration policies were actually implemented, whether this differed across regions, the reasons behind any implementation delays, how rollout occurred (for example, the speed of ramp-up), and the broader context in which the demonstration policies were implemented. The following questions may serve as a guide for understanding the implementation context:[8]

[6] Demonstration approval and other documents for the Oregon Health Plan demonstration are available at https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/82956.

[7] Demonstration approval and other documents for the TennCare III demonstration are available at https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/83206.

[8] See "Conducting Robust Implementation Research for Section 1115 Demonstration Evaluations" (Felland and Bradley 2020) available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/Implementation-rsch.pdf.

- Is the intervention being carried out as planned and proceeding as expected? Why? What factors are helping (facilitators), and what factors are posing challenges (barriers)?
- Does the intervention's progress vary depending on where it is taking place and who is affected? Are there differences in context or characteristics that might explain this variation?
- What else is changing at the same time that could be affecting the intervention or its outcomes?

States can gather this information from several sources:

Monitoring metrics. Monitoring metrics that encompass process measures can shed light on the extent to which the demonstration has made progress towards full implementation. For example, states with a SUD demonstration may use monitoring metrics that show the share of demonstration participants who accessed early intervention services.[9] This can help show how much demonstration processes have ramped up and whether they are reaching eligible populations.

Policy-specific process indicators. States may collect additional process indicators to track operational steps. For example, the Arizona Health Care Cost Containment System (AHCCCS) demonstration provides integrated physical and behavioral health services through Regional Behavioral Health Authorities (RBHA) to several populations including beneficiaries with serious mental illness during the 2022–2027 approval period.[10] To assess progress toward implementation, the state could track the number of physical and behavioral health providers contracted with RBHAs and how this number changes as implementation ramps up.

Surveys and interviews. Surveys and interviews with providers, beneficiaries, and other stakeholders in the demonstration are a key source of data for understanding how the demonstration was implemented and stakeholders' experiences with implementation. For example, such data collection with providers and health systems could solicit information on challenges and facilitators of service delivery and the speed of ramp-up. Likewise, data collection with beneficiaries could ask about beneficiary awareness and experience of demonstration services. Surveys and interviews can complement one another. Surveys enable the evaluator to gather comparable information about the implementation across potentially large groups of respondents and assess prevalence of the responses. Through open-ended questions and targeted probing, semi-structured interviews enable the evaluator to gather rich, nuanced descriptions of the implementation and context that could help guide what survey questions to ask or help explain patterns observed in survey responses.

States should also try to document as much information as possible about the implementation context in interim and summative evaluation reports. In addition to informing the evaluation design, such information also helps states and CMS interpret evaluation results. Combining and synthesizing implementation and outcomes research improves the rigor of the overall evaluation.

[9] Monitoring metrics for SUD demonstrations are available in the "SUD Metrics" tab at https://www.medicaid.gov/medicaid/section-1115-demonstrations/downloads/monitor-rpt.xlsx.

[10] Demonstration approval and other documents for the AHCCCS demonstration are available at https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/80976.

B. Estimate period-specific impacts to understand impacts of different policies

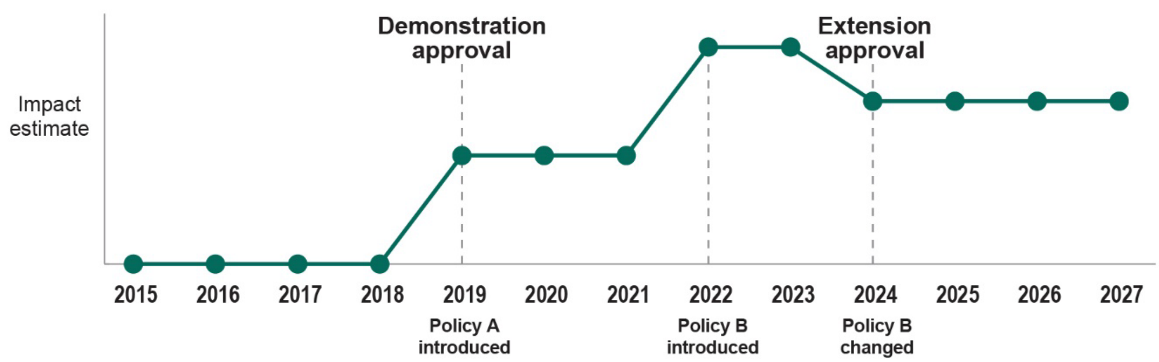
An advantage of settings in which policies are not all introduced at the same time is that states could potentially identify the impact of individual demonstration policies. Table 1 summarizes different types of these settings, and how evaluators can take advantage of them to estimate period-specific or policy-specific impacts.

Table 1. Options for estimating period-specific impacts

Demonstration policy setting	Evaluation option
Individual demonstration policies affect different populations and there is little or no overlap in these populations	Estimate policy-specific impacts by using each policy-specific population to estimate corresponding impacts, supplemented by mixed-methods approach (Section II.A).
Individual policies are targeted at a similar population and introduced at different times	Estimate the impacts of individual policies by using the start of the demonstration period and tracking year- or quarter-specific impacts and compare these impacts with the times when different policies were introduced. The demonstration impact during a specific period represents the combined impact of the demonstration policies in effect during that period. Supplement with mixed-methods approach.
Individual policies are targeted at a similar population and introduced at the same time	Use mixed-methods approach.

Figure 2 shows a visual example of yearly demonstration impacts for a section 1115 demonstration that introduced or changed policies in 2019, 2022, and 2024. Yearly impacts help evaluators understand how demonstration effects evolved as the new policies were introduced.

Figure 2. Illustration of period-specific impacts



An important caveat is that, in practice, policy impacts may change over time, making it difficult to isolate the impact of a specific demonstration policy during periods in which more than one policy affecting the same population has been introduced. For example, if one policy is introduced at the beginning of the demonstration period and a second is introduced in the third year of the demonstration period, an evaluator might not be able to recover the impact of the second policy independent of the first. This is because the impact of the first policy in year one may be different from its impact in year three.

However, if there is no or little overlap between populations affected by different policies, states can generally estimate policy-specific impacts. For example, for the evaluation of a demonstration with a postpartum coverage and a SUD policy, there may be overlap between the affected populations, but states can estimate the impact of the postpartum policy by focusing on recent mothers without a SUD and the impact of the SUD policy by excluding recent mothers from the estimation sample. Similarly, states can distinguish policy-specific impact when there is little temporal overlap between policies. For example, if one policy is implemented at the beginning of the approval period and another policy in year four of the approval period, the state can estimate the impact of the first policy using the first three years. Section III describes alternative strategies for understanding the impacts of individual policies.

III. Demonstrations Where Individual Policies Are Introduced at the Same Time

When multiple demonstration policies are implemented at the same time, it is difficult to ascertain the impacts of individual policies using time-series approaches.[11] However, evaluating the impact of individual demonstration policies remains an important evaluation priority. For future section 1115 demonstration innovations, it is vital for states and CMS to understand the specific policies and demonstration services that are main contributors to desired changes in the outcomes of interest. In some cases, specifically when demonstrations operate under substantially different authorities, evaluators can often estimate impacts of individual policies because the underlying populations do not fully overlap (for example, pregnant and postpartum women and beneficiaries with a SUD diagnosis) or because some policies may be testing novel approaches while others are longstanding. Below we outline two strategies for evaluating individual policies when demonstrations operate under a single authority or affect a homogeneous population.

A. Use mixed method approaches

As described in Section II.A, qualitative and quantitative implementation data are important sources of information that states and their evaluators can use to design an impact evaluation approach and interpret quantitative findings. This also applies to the evaluation of demonstrations with multiple policies implemented at the same time. Qualitative evidence, such as beneficiary and provider surveys and interviews with providers and other stakeholders, can provide rich data on how different policies affect beneficiary outcomes and help states disentangle impacts of different policies. For example, the Utah Medicaid Reform 1115 Demonstration includes several policies during the 2022–2027 approval period: integrated managed care plans providing state plan benefits to the adult expansion population, a dental benefit, a serious mental illness benefit, and a SUD program, among other policies.[12] To understand the impacts of each of these individual policies, an evaluator can ask providers or other key informants about the effects of each of these policies on their access to care and their health. Evaluators can conduct both quantitative analyses and a qualitative synthesis of provider responses to shed light on how each of the individual policies impacted the outcomes of interest.

B. If individual policies affect different eligibility groups, use a regression discontinuity design

Alternative methods that do not use time-series data, such as regression discontinuity designs, might be better suited for identifying the effect of individual policies if eligibility for these policies depends on a threshold. For example, in a section 1115 demonstration with many policies, including a policy to reduce copayments for beneficiaries under a certain income threshold, evaluators can use regression

[11] The issues and solutions in this section also apply to states with multiple separate demonstrations that are approved for the same or overlapping period.

[12] Demonstration approval and other documents for the Utah Medicaid Reform 1115 Demonstration are available at https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/83321.

discontinuity designs to specifically evaluate the impact of the copayment reduction.[13] Using a regression discontinuity design to identify the effects of individual policies is also helpful in instances where policies are introduced at different times, but their impacts are expected to increase or decrease over time. In this case, evaluators can use a regression discontinuity design to estimate separate impacts for each demonstration year among beneficiaries whose eligibility for a benefit depends on an income threshold. They can then contrast these impact estimates with demonstration year-specific effects of other demonstration policies that affect a broader beneficiary population.

[13] For guidelines on how best to implement regression discontinuity designs in demonstration evaluations, see the white paper "Regression Discontinuity Designs in the Evaluation of Section 1115 Demonstrations" (Farid et al., 2023), available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/regression-discontinuity-designs.pdf.

IV. Demonstrations That Are Implemented for Different Populations at Different Times (Staggered Implementation)

States may implement some demonstration policies at different times for different populations of Medicaid beneficiaries in a planned or unplanned way. Populations can be defined by areas of residence within the state (such as counties), beneficiary characteristics, or other demonstration features (for example, by carceral facility in reentry demonstrations). Implementing demonstration policies in some parts of a state is possible because CMS can waive the statewideness requirements of Medicaid policies under section 1115 demonstrations. For example, during the 2021–2030 approval period of the Texas Healthcare Transformation and Quality Improvement Program, the state divides its counties into Regional Healthcare Partnerships (RHPs), and each RHP develops its own programs and a different timeline.[14] In this context, the staggered nature of the implementation was known beforehand, though the timing of program implementation by different RHPs was not. In other section 1115 demonstrations, implementation may be staggered across geographic areas due to other reasons, such as unexpected and unplanned differences in infrastructure and ramp-up efforts across regions or financial constraints. Demonstration implementation could also be staggered by age group. For example, during the 2022–2027 approval period of the Oregon Health Plan demonstration, the demonstration covers early period screening, diagnosis, and treatment (EPSDT) for children from birth to age 21 effective January 1, 2023. For young adults through 25, the demonstration covers EPSDT effective January 1, 2025. The approach described in this section also applies when states have waitlists for benefits or services, and as a result, beneficiaries start participating in the demonstration at different times. For example, the Washington Medicaid Transformation Project 2.0 demonstration includes the Foundational Community Supports Program during the 2023–2028 approval period, which offers housing and employment services subject to provider capacity, leading to deferred service receipt for eligible beneficiaries.[15] Finally, some states pilot demonstration policies to a small group of people or within a small area, before expanding eligibility more broadly. Panel C in Figure 1 illustrates such cases of a staggered rollout.

Section 1115 demonstrations with staggered implementation across geographic areas may allow for more convincing causal impact identification, because they enable evaluators to better account for any secular state-specific time trends. Having multiple time periods in which groups of beneficiaries become exposed to demonstration policies alleviates concerns that contemporaneous trends in outcomes lead to misleading or incorrect impact estimates. Staggered rollout not only provides a source of causal identification, but it may also be more practical for states that do not have the resources for an immediate statewide rollout.

In settings with staggered implementation, evaluators typically use difference-in-differences methods to compare beneficiaries who have been exposed to the demonstration (for example, because they live in an

[14] Demonstration approval and other documents for the Texas Healthcare Transformation and Quality Improvement Program demonstration are available at https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/83231.

[15] Demonstration approval and other documents for the Washington Medicaid Transformation Project 2.0 demonstration are available at https://www.medicaid.gov/medicaid/section-1115-demo/demonstration-and-waiver-list/83531.

area where the demonstration was rolled out early) to those who have not yet received demonstration services (because they live in an area where the demonstration has not yet been rolled out). The latter group acts as the comparison group. If there is a group of beneficiaries who are never exposed to demonstration policies, evaluators may also include them as a comparison group. In choosing whether to include this group, evaluators should keep in mind the likely similarity of the "never-treated" group to the groups exposed to demonstration policies. For example, if the state only chose to roll out the demonstration in urban areas, beneficiaries in rural areas who were never exposed to the demonstration are likely not a good comparison group. On the other hand, comparable geographic regions that have been exposed to demonstration services and only differ in when exposure occurred are likely more comparable to one another.

Evaluators have traditionally implemented difference-in-differences using a two-way fixed effects regression model to estimate the causal impact of a demonstration policy when the state implements a demonstration for different regions at different times during an approval period.[16] This approach involves comparing outcomes pre- and post-implementation after controlling for time trends (time fixed effects) and time invariant region characteristics (region fixed effects).

However, a growing literature[17] has shown that impacts estimated with the standard two-way fixed effects model described above are biased if treatment effects change over time or across regions. In the context of section 1115 demonstrations, it is likely that demonstration impacts vary across groups or over time. For example, a demonstration policy such as increased access to SUD treatment services could have a different impact in rural counties compared to cities. Because of barriers to implementation or slow beneficiary outreach, demonstration impacts may become larger over time. Several methods have been developed that circumvent this bias. Broadly, the goal of these methods is to carefully select treatment and comparison groups and then compare each group (or set of units that are treated at the same time) to a clean comparison group—that is, a comparison group that has never been treated or that has not yet been treated. States can use several approaches that are further described in the Appendix.

[16] We use region as a running example in this section, but groups could be defined in other ways. For example, if staggered rollout is not based on geography but based on different eligibility groups, then the two-way fixed-effects model would include "eligibility-group fixed effects" rather than "region fixed effects."

[17] Examples include Goodman-Bacon (2021), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfoeille (2020) and Sun and Abraham (2021).

V. Conclusion

Section 1115 demonstrations that evolve over time or introduce multiple policies at the same time present evaluation challenges for states in terms of defining suitable baseline periods, forming appropriate hypotheses, and understanding the impacts of individual policies, among other challenges. When considering how to evaluate these types of demonstrations, states should prioritize rigorous approaches for newer policies, for which less evidence from other states exist. This includes mapping demonstration goals to logic models and hypotheses and empirical research designs following the suggestions in this white paper that yield reliable impact estimates. For longstanding policies with an exhaustive evidence base, states can consider a more pragmatic evaluation approach that can track whether states meet demonstration goals, without necessarily accounting for more complex timing issues. This may also apply to cases where the policy is not longstanding in a specific state, but other states have the same longstanding policy with exhaustive evaluation evidence.

The best empirical approach to evaluate demonstrations that change over time will depend on the timing of the introduction of various demonstration policies and on the data available. This white paper discussed strategies that states and their independent evaluators can use to evaluate such time-varying demonstrations. In addition to selecting a suitable quantitative approach, states should consider a comprehensive mixed methods evaluation to understand how policies are implemented over time.

Recommendations for section 1115 evaluation designs. Evaluators should define pre- and post-periods using guidelines outlined in this paper and should consider estimating period-specific impacts in settings where demonstration services change over the evaluation period, or when the evaluation period encompasses multiple demonstration periods. Evaluators should also form hypotheses with care, keeping in mind that the definition of the baseline period affects whether outcomes might be expected to change relative to the baseline. Finally, comprehensive implementation research to gain an understanding of how the demonstration rolled out will inform lessons learned and best practices.

Recommendations for section 1115 evaluation reports. In section 1115 demonstration evaluation reports, evaluators should provide sufficient detail such that readers can clearly identify the baseline and intervention periods and interpret period-specific impacts, if those are estimated. In addition, states should summarize information about how the demonstration was implemented as well as any state initiatives or external policies contemporaneously implemented that could affect the demonstration population in interim and summative evaluation reports. This information can help the states and CMS interpret evaluation results. Finally, states should use period-specific impact estimates and implementation evidence to inform changes or updates to the demonstration and broader Medicaid policy.

References

Baker, A., D.F. Larcker, and C.C.Y. Wang. “How Much Should We Trust Staggered Difference-In-Differences Estimates?” *Journal of Financial Economics*, vol. 144, no. 2, May 2022, pp. 370–395. https://doi.org/10.1016/j.jfineco.2022.01.004.

Borusyak, K., X. Jaravel, and J. Spiess. “Revisiting Event Study Designs: Robust and Efficient Estimation.” *The Review of Economic Studies*, vol. 91, no. 6, November 2024, pp. 3253–3285. https://doi.org/10.1093/restud/rdae007.

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. Available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/comparison-grp-eval-dsgn.pdf.

Callaway, B., and P.H.C. Sant’Anna. “Difference-in-Differences with Multiple Time Periods.” *Journal of Econometrics*, vol. 225, no. 2, December 2021, pp. 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer. “The Effect of Minimum Wages on Low-Wage Jobs.” *The Quarterly Journal of Economics*, vol. 134, no. 3, 2019, pp. 1405–1454. https://doi.org/10.1093/qje/qjz014.

de Chaisemartin, C., and X. D’Haultfoeuille. “Two-Way Fixed Effects and Difference-in-differences with Heterogeneous Treatment Effects: A Survey.” *The Econometrics Journal*, vol 26, no. 3, September 2023, pp. C1–C30. https://doi.org/10.1093/ectj/utac017.

de Chaisemartin, C., and X. D’Haultfœuille. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review*, vol. 110, no. 9, 2020, pp. 2964–2996. https://doi.org/10.1257/aer.20181169.

Deb, P., E.C. Norton, J.M. Wooldridge, and J.E. Zabel. “A Flexible, Heterogeneous Treatment Effects Difference-in-Differences Estimator for Repeated Cross-Sections.” NBER Working Paper 33026. Cambridge, MA: National Bureau of Economic Research, May 2025. Available at https://www.nber.org/system/files/working_papers/w33026/w33026.pdf.

Deshpande, M., and Y. Li. “Who Is Screened Out? Application Costs and the Targeting of Disability Programs.” *American Economic Journal: Economic Policy*, vol. 11, no. 4, 2019, pp. 213–248. https://doi.org/10.1257/pol.20180076.

Farid, M., W. Zhu, and A. Hill. “Regression Discontinuity Designs in the Evaluation of Section 1115 Demonstrations.” Washington, DC: Mathematica, February 2023. Available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/regression-discontinuity-designs.pdf.

Felland, L., and K. Bradley. “Conducting Robust Implementation Research for Section 1115 Demonstration Evaluations.” Washington, DC: Mathematica, October 2020. Available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/Implementation-rsch.pdf.

Gardner, J. “Two-Stage Differences in Differences.” ArXiv:2207.05943 [Econ], July 2022. https://doi.org/10.48550/arXiv.2207.05943.

Goodman-Bacon, A. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*, vol. 225, no. 2, December 1, 2021, pp. 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014.

Imai, K., and I. Song Kim. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis*, vol. 29, no. 3, July 2021, pp. 405–415. https://doi.org/10.1017/pan.2020.33.

Liu, L., Y. Wang, and Y. Xu. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science*, vol. 68, no. 1, January 2024, pp. 160–176. https://doi.org/10.1111/ajps.12723.

Pohl, R.V., L. Lei, and M. Niedzwiecki. "Matching Methods for the Evaluation of Section 1115 Demonstrations." Washington, DC: Mathematica, February 2023. Available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/matching-methods.pdf.

Roth, J., P.H.C. Sant'Anna, A. Bilinski, and J. Poe. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics*, vol. 235, no. 2, August 2023, pp. 2218–2244. https://doi.org/10.1016/j.jeconom.2023.03.008

Sun, L., and S. Abraham. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*, vol. 225, no. 2, December 2021, pp. 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006.

Werner, A. A Guide to Implementation Research. Washington, DC: The Urban Institute Press, 2004.

Wooldridge, J. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." *Empirical Economics*, vol. 69, August 2025, pp. 2545–2587. https://doi.org/10.1007/s00181-025-02807-z.

Appendix: Difference-in-Differences for Staggered Implementation

As we describe in Section V, some states implement section 1115 demonstrations for different groups (such as parts of the state) at different times (staggered rollout). This setting lends itself to an impact evaluation using difference-in-differences methods. Applying difference-in-differences requires data from the demonstration group (beneficiaries, providers, or other entities affected by the demonstration policy being evaluated) and a comparison group (those unaffected by the policy) both before and after implementation. When a state implements the demonstration in all parts of the state at the same time, it is straightforward to define demonstration and in-state or out-of-state comparison groups and pre- and post-periods (baseline and follow-up periods). In the case of staggered rollout, this is more complicated because there are multiple demonstration and comparison groups and pre- and post-periods. Until recently, researchers estimated impacts in staggered rollout settings using standard difference-in-differences methods, which typically meant including dummy or indicator variables for each demonstration and comparison group and for pre- and post-periods (the two-way fixed effects estimator). As Goodman-Bacon (2021) and others have shown, applying standard difference-in-differences methods can lead to biased impact estimates when there are multiple demonstration groups that start being affected by a policy at different times and if impacts vary over time or across groups.[18] In this appendix, we further explain the reason for this bias and then summarize methods that evaluators can use to estimate unbiased demonstration impacts when states implement policies in a staggered manner.[19] We conclude by providing a list of software packages that can help evaluators implement these methods.

A. Difference-in-differences with variation in treatment timing leads to biased impact estimates

Goodman-Bacon (2021) has shown that using standard difference-in-differences methods leads to biased impact estimates when groups start to be affected by a policy (that is, treated) at different times. In the two-way fixed effects method, researchers use a regression model with dummies (binary indicator variables) for each group and time period and a post-treatment indicator as independent variables. The post-treatment indicator equals one if a group is treated in a given time period and zero otherwise. In the context of section 1115 demonstrations used as an example in Section V (illustrated in Figure 1, panel D), the post-treatment indicator reflects, for example, in which counties and years a state has implemented a demonstration policy. In this regression model, researchers interpret the estimated coefficient on the post-treatment dummy as the average impact of the demonstration. As Goodman-Bacon (2021) showed, this coefficient is a biased estimate of the true impact. The bias arises because the two-way fixed effects estimator is a weighted average of all possible two-group/two-period difference-in-differences impacts with some weights possibly being negative. To provide more intuition, we first describe the idea of two-group/two-period impacts and then explain why weights may be negative and how that leads to bias.

A two-group/two-period impact refers to the impact that researchers would estimate when comparing outcomes between one treated and one comparison group with a single pre-period and a single post-

[18] Bias is the difference between the impact estimate and the true impact (which evaluators do not observe). When evaluating a section 1115 demonstration, it is important to use methods that yield unbiased impact estimates, so evaluators can learn the true demonstration impact.

[19] Roth et al. (2023), Baker et al. (2022), and de Chaisemartin and D’Haultfœille (2022) provide more technical summaries of these methods.

period. However, in the case of staggered implementation, there are several of these comparisons. For example, a state may implement a demonstration policy in county A in 2025, in county B in 2026, and never in county C. Then, with data available from 2024 through 2026, evaluators could make the two-by-two comparisons shown in Table A.1. As the table shows, there are four such comparisons. Note that county B can be the comparison group when considering the year 2024 and 2025, because it is not affected by the demonstration policy yet (row 3). Similarly, county A can serve as the comparison group for county B when considering the years 2025 and 2026, because county A's treatment status does not change: it is affected by the policy in both years (row 4). The latter comparison seems counterintuitive, because researchers usually think about comparison groups as not being affected by the policy in question. In this context, however, a comparison group is defined as group whose treatment status does not change across two years. (This group may be employed to "difference out" trends in outcomes common to all counties.)

Table A.1. Possible two-by-two comparisons with three groups and three periods

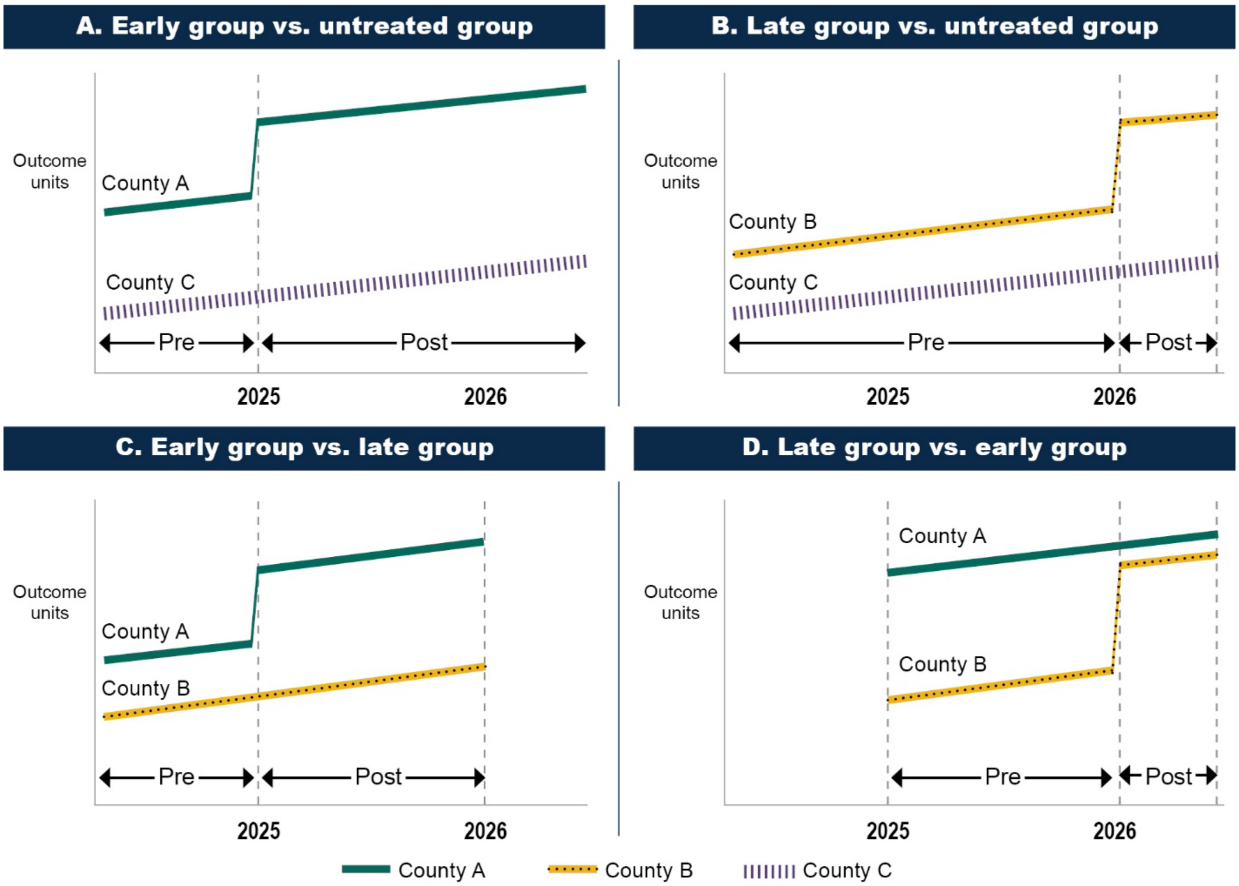
Two-by-two comparison	Demonstration/ treatment group	Comparison group	Pre-period	Post-period
#1	County A	County C	2024	2025 and 2026
#2	County B	County C	2024 and 2025	2026
#3	County A	County B	2024	2025
#4	County B	County A	2025	2026

Figure A.1 illustrates these comparisons graphically, using a hypothetical outcome path. The figure also shows that not all available years of data are used when comparing counties A and B (see panels C and D). In each two-by-two comparison, the difference-in-differences impact is defined as the difference between (1) the vertical difference between the two outcome lines in the post-period and (2) the vertical difference in the pre-period. The figure highlights that with staggered implementation, there are multiple difference-in-differences comparisons. Researchers implicitly aggregate these comparisons into a single impact estimate when using standard difference-in-differences methods.

When aggregating the two-by-two comparisons into a single impact estimate, each two-by-two estimate implicitly receives a weight—that is, the overall impact is a specific weighted average of the two-by-two impacts. These weights depend on the size of each group and the length of time during which a group was affected by the treatment. In the example in Table A.1, if there are more Medicaid beneficiaries in county A than in counties B and C, the two-by-two comparisons involving county A would receive more weight. In addition, comparison #1 would receive more weight than comparison #3 because the former involves two years in the post-period.

The problem with the average treatment effect estimated by a two-way fixed effects model is that the two-by-two comparisons receive weights that are incorrect. Attaching the wrong weight to two-by-two comparisons leads to biased impact estimates. In some cases, the weights can be negative. Negative weights arise when groups that are already treated act as comparisons, as in comparison #4 in Table A.1, where county A is affected by the demonstration policy in both years but acts as a comparison to county B, which is only treated in 2026.That implies that a two-by-two comparison can yield a positive impact estimate, but this positive impact estimate would reduce the overall difference-in-differences impact because it receives a negative weight in the weighted average. For example, if the demonstration impact in county A is larger in 2026 than in 2025, there is an increase in the outcome that will be subtracted from the outcome difference in county B due to the mechanics of difference-in-differences estimation. Subtracting this positive change leads to the negative weight and introduces bias.

Figure A.1. Illustration of two-by-two comparisons



Source: Adapted from Figure 2 in Goodman-Bacon (2021).

The main take-away from this discussion is that evaluators should not use the standard two-way fixed effects method involving group and time dummies and a single post-treatment indicator when estimating demonstration impacts in settings with staggered implementation, because there is always a possibility that impacts could change over time or vary across groups.[20] This also applies when impacts are not expected to change over time but are expected to differ according to when groups of beneficiaries first became subject to a demonstration policy. For example, states may expect different demonstration impacts for beneficiaries for whom a policy was implemented earlier rather than later because, for example, implementers learn from early implementation experiences and become more effective with later rollouts. In the following sections, we present alternative methods that evaluators can use instead.

[20] If the two-by-two comparisons happen to produce *identical* impact estimates, then it does not matter what weight is assigned to each comparison. However, there is no reason a priori to expect identical impact estimates from each two-by-two comparison.

B. Estimating period-specific impacts can avoid bias

Instead of using the standard difference-in-differences model with a single dummy variable representing which groups were affected by a demonstration policy, evaluators can estimate more flexible models that involve multiple post-treatment dummies for each time unit (for example, year or quarter) in the post-period and for each group. (Beneficiaries who are first exposed to a demonstration policy at the same time are also referred to as a treatment cohort.) In the example shown in Table A.1, evaluators would include separate post-treatment dummies for county A in 2025 and 2026 and for county B in 2026. Hence, instead of implicitly aggregating the two-by-two comparisons into a single impact estimate, they would explicitly estimate individual demonstration impacts for counties A and B in each year. In addition to avoiding the bias described above, this approach would also show whether impacts differ between counties or over time.

A growing literature proposes to use alternative methods to estimate impacts when implementation is staggered, and impacts vary over time or across groups with different implementation dates. These methods can be categorized into three groups:

1. **Estimating weighted group-time average treatment effects.** These methods involve estimating cohort-specific treatment effects (where a cohort is the set of units that is treated at the same time). Each cohort of treated units is compared to units not yet treated or never treated, then the average treatment effect is obtained by taking a weighted average across cohorts. Examples include Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfoeille (2020), Sun and Abraham (2021), Imai and Kim (2021), Deb et al. (2025), and Wooldridge (2025).
2. **Stacked regression.** These methods involve creating several small data sets where each treatment cohort is matched to clean (not-yet-treated controls). These data sets are then stacked, and evaluators compare outcomes before and after an intervention controlling for data set-specific group and time effects. Examples include Cengiz et al. (2019) and Deshpande and Li (2019). Gardner (2022) provides an overview.
3. **Imputation methods.** These methods involve a two-stage approach in which evaluators estimate the correct counterfactual outcome for each unit by regressing the outcome on time and unit fixed effects using observations only for units and time periods that are not yet treated. Using these methods, evaluators can then estimate unbiased treatment effects. Examples include Borusyak et al. (2024), Gardner (2022), and Liu et al. (2024).

The first category of methods is easy to implement using recently developed and widely available statistical packages. Here, we highlight articles by Sun and Abraham (2021) and Callaway and Sant'Anna (2021), which are in the first category of methods. Both articles propose flexible methods to estimate cohort- and time-specific impacts but do so slightly differently.

Sun and Abraham (2021) extend the two-way fixed effect regression model to include additional terms as independent variables: interactions between group indicators and time indicators as independent variables in difference-in-differences models. Specifically, they define treatment cohorts according to when an intervention was first implemented. For example, if a state implements a demonstration policy in multiple counties at the same time, these counties would all be in the same treatment cohort. If a demonstration period is from January 1, 2022, through December 31, 2026, there might be 2022, 2023, 2024, 2025, and 2026 cohorts; accordingly, evaluators would define five cohort indicators. These cohort indicators are interacted with time indicators reflecting how many years have elapsed relative to the start of the implementation for that group. For example, for a county where the policy was implemented in 2023, the indicator for the first relative year would be equal to one in 2023, the indicator for the second relative year would be equal to one in 2024, and so on. In this example, there are five cohorts and up to five relative time periods. However, not all relative time periods are defined for all cohorts. For example, for counties with an implementation start of 2025, there are three relative time periods (2025, 2026, and 2027). Overall, there are 15 cohort and time-specific demonstration impacts to be estimated. Evaluators can interpret these individual impacts and aggregate them to an overall demonstration impact. Sun and Abraham (2021) show that this approach avoids the bias described in Appendix Section A.

Alternatively, Callaway and Sant'Anna (2021) propose a method that also constructs cohort- and time-specific impact estimates, but they do so by using a weighted difference-in-differences approach. The weights are calculated using propensity scores—that is, estimated probabilities of being in a given treatment cohort. Conceptually, this is a double robust approach that has desirable statistical properties.[21] After estimating cohort- and time-specific impact estimates, Callaway and Sant'Anna (2021) show how to aggregate them to produce more meaningful estimates of average impacts. Evaluators may use different schemes for weighting the individual (cohort- and time-specific) impact estimates into an overall average impact estimate, depending on the research question. For example, evaluators may be interested in the average impact throughout the approval period in counties where the demonstration was implemented in 2023 or in the average impact across all counties in 2025. This method enables them to do so flexibly and avoid the issues described above.

C. Software packages for implementing difference-in-differences estimators

Table A.2 shows available packages for Stata and R that evaluators can use to implement the estimators mentioned in this white paper.

[21] See the white paper "Matching Methods for the Evaluation of Section 1115 Demonstrations" (Pohl et al., 2023), available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/matching-methods.pdf, for a discussion of propensity score weighting and double robust estimation in the context of section 1115 demonstrations.

Table A.2. Stata and R software packages

Paper	Stata package	R package
Borusyak et al. (2024)	did_imputation, event_plot[a]	didimputation[b]
Callaway and Sant'Anna (2021)	csdid[c]	did[d]
Cengiz et al. (2019)	stackedev[e]	
de Chaisemartin and D'Haultfœuille (2020)	twowayfeweights, fuzzydid, did_multiplegt[f]	DIDmultiplegt[g]
Goodman-Bacon (2021)	bacondecomp[h]	bacondecomp[i]
Imai and Kim (2021)		wfe[j]
Sun and Abraham (2021)	eventstudyweights, eventstudyinteract[k]	fixest[l]

Source: https://asjadnaqvi.github.io/DiD/.

[a] Available from https://github.com/borusyak/did_imputation.

[b] Available from https://github.com/kylebutts/didimputation.

[c] Available from https://github.com/friosavila/csdid_drdid.

[d] Available from https://github.com/bcallaway11/did.

[e] Available from https://github.com/joshbleiberg/stackedev.

[f] Available from https://faculty.crest.fr/xdhaultfoeuille/software-programs/.

[g] Available from https://cran.r-project.org/web/packages/DIDmultiplegt/index.html.

[h] Available from https://ideas.repec.org/c/boc/bocode/s458676.html.

[I] Available from https://github.com/evanjflack/bacondecomp.

[j] Available from https://github.com/insongkim/wfe.

[k] Available from https://github.com/lsun20/EventStudyWeights and https://github.com/lsun20/EventStudyInteract.

[l] Available from https://lrberge.github.io/fixest/.

Medicaid.gov
Keeping America Healthy
Centers for Medicare & Medicaid Services

CMS • 7500 Security Boulevard, Baltimore, MD 21244

CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES

Mathematica Inc.

Our employee-owners work nationwide and around the world.

Find us at **https://www.mathematica.org** and **edi global.com**.

Mathematica
Progress Together