CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES

Mathematica
Progress Together

Matching Methods for the Evaluation of Section 1115 Demonstrations

White Paper

February 2023

R. Vincent Pohl, Lianlian Lei, and Matthew Niedzwiecki

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 use matching methods in their evaluations of section 1115 demonstrations.

Contents

I. Introduction

In evaluations of section 1115 demonstrations, it is often not feasible or desirable to randomly assign Medicaid beneficiaries to treatment and control groups. In the rare case of a randomized controlled trial (RCT), Medicaid beneficiaries are randomly included in a demonstration component (the treatment group) or randomly excluded (the control group), and evaluators compare the outcomes of these two groups that have identical characteristics, on average. An RCT therefore allows evaluators to interpret differences in average outcomes as the causal impact of the demonstration.

In practice, when an RCT is not feasible or desirable, the most rigorous evaluations of section 1115 demonstrations rely on identifying a suitable comparison group.[1, 2] Ideally, the comparison group would have observed and unobserved characteristics that are identical, on average, to the beneficiaries subject to the demonstration component being evaluated, just as in an RCT. In practice, however, comparison groups are often different from the demonstration group in at least some characteristics. For example, beneficiaries in the comparison group may be younger or older or have different health care needs than beneficiaries in the demonstration group, on average. If these differences are large, adjusting for them by including covariates in outcome regression models is often insufficient. For example, controlling for age in an outcome regression may be insufficient when the age distributions in demonstration and comparison groups are very different. In such a situation, matching methods allow evaluators to identify a comparison group most similar to the demonstration group, thereby approximating the ideal of an RCT, in which the comparison group has close to identical average characteristics.

Matching refers to any statistical method that can equate or "balance" covariates in demonstration and comparison groups (Stuart 2010). There are many different methods, each with its own advantages and limitations, that evaluators can use in different scenarios (see Section III). When evaluating demonstrations using matching methods, evaluators proceed in two stages: the evaluation design stage and the analysis stage (Imbens 2015). All matching methods are implemented at the design stage, when evaluators do not analyze outcomes but only consider the relationship between beneficiary characteristics and treatment status, that is, participation in a section 1115 demonstration. In contrast, the analysis stage focuses on estimating the relationship between treatment status and outcomes of interest. Typically, evaluators use matching methods when they are concerned that the comparison group differs from the demonstration group along important covariates that affect key outcomes. However, if beneficiaries in demonstration and comparison groups differ widely, matching methods cannot resolve the lack of overlap. In other words, matching is not a panacea for observed differences between the two groups, and evaluators should be careful to define a potential comparison group before applying matching methods (see Section II).

To support states' efforts to conduct rigorous evaluations of their section 1115 demonstrations, this white paper describes matching methods that evaluators can use to construct a comparison group that constitutes

[1] See "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations" (Bradley et al. 2020) 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 terms comparison group and control group both refer to individuals who are not subject to the intervention being evaluated, but researchers commonly use control group in the context of RCTs and comparison group for observational research designs, such as those involving matching.

a valid counterfactual for the Medicaid beneficiaries or health care providers that are affected by demonstration policies. We first briefly address comparison group selection (Section II) before providing an overview of the matching methods (Section III). We conclude by summarizing key decision points and the limitations of matching methods (Section IV). In particular, in contrast to an RCT, which equates both observed and unobserved characteristics of treatment and comparison groups, on average, matching can balance only observed covariates. By definition, evaluators cannot measure unobserved characteristics, such as beneficiaries' unmeasured health risks, and therefore cannot apply matching methods to this type of covariate. Although matching methods are a powerful tool for evaluations of section 1115 demonstrations, evaluators and states need to acknowledge that using them can yield valid impact estimates only if demonstration and comparison groups do not differ systematically along some unobserved characteristics.

Medicaid Section 1115 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.

II. Selecting a Potential Comparison Group

Before evaluators can use matching methods to balance the covariates of demonstration and comparison groups, they must define a potential comparison group (also referred to as a comparison beneficiary "pool").[3] The potential comparison group consists of individuals (such as Medicaid beneficiaries, people not covered by Medicaid, or health care providers) who are not affected by the component of the section 1115 demonstration being evaluated. Once evaluators have identified a potential comparison group, they can use the matching methods described in Section III to select individuals who will constitute the final comparison group used in the impact analysis stage. Evaluators can apply the matching methods described in this white paper whether the unit of analysis is a beneficiary, a health care provider, or any other entity affected by demonstration policies. For readability, we only refer to beneficiaries when describing the methods in general terms.

Selecting the potential comparison group is an important part of the evaluation design because it ensures that the final comparison group is similar to the demonstration group along key dimensions. Depending on the type of demonstration policy being evaluated, these important characteristics may include income or age. There must be some overlap between characteristics of the potential comparison group and the demonstration group before matching, because matching methods cannot resolve vast differences between the two groups. For example, if a demonstration component applies to beneficiaries aged 19 to 49, a potential comparison group should not consist only of beneficiaries aged 50 to 64 because there would be no overlap in age, an important explanatory variable for health-related outcomes.[4]

There are two types of comparison groups in the context of section 1115 demonstrations. In-state comparison groups consist of Medicaid beneficiaries who reside in the same state and are exempt from the demonstration component being evaluated or in some cases individuals residing in the same state but not covered by Medicaid. Out-of-state comparison groups are comprised of Medicaid beneficiaries in one or more other states that have not implemented the same type of demonstration. Bradley et al. (2020) discuss these comparison strategies in more detail.

To define an in-state comparison group, evaluators typically select Medicaid beneficiaries who are not eligible for the demonstration policy being evaluated. These comparison beneficiaries may be exempt because they have different Medicaid eligibility categories, geographic location within the state, income, or ages, among other reasons; or they may be in groups for whom demonstration policies will be rolled out later, as part of a phased implementation. For example, California's Drug Medi-Cal Organized Delivery System (for the demonstration period November 1, 2010 through October 31, 2015) rolled out substance use disorder services county by county, allowing beneficiaries in counties that would be phased in later to serve as comparisons for those in early implementation counties.[5] In some cases, evaluators can include an in-state comparison group consisting of non-Medicaid beneficiaries, such as low-income

[3] The methods described in this white paper apply regardless of the unit of analysis (Medicaid beneficiary, health care provider, or managed care entity), but for readability we refer to beneficiaries when describing matching methods generically.

[4] If there is no overlap but eligibility for a demonstration depends on a cut-off in a continuous covariate, such as income or age, evaluators can use a regression discontinuity design. For details, see "The Regression Discontinuity Design in the Evaluation of Section 1115 Demonstrations" (Farid et al. 2023) at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/regression-discontinuity-designs.pdf.

[5] The approved evaluation design is available at https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/ca/medi-cal-2020/ca-medi-cal-2020-organized-delivery-system-eval-design-06202016.pdf.

individuals who are commercially insured, but this comparison strategy typically requires access to an all-payer claims database. For example, the evaluation of the "Oregon Health Plan Substance Use Disorder 1115 Demonstration" (for the demonstration period April 8, 2021 through March 31, 2026) plans to use the state's broader population as a comparison group.[6] For all types of in-state comparison groups, characteristics are unlikely to perfectly align with those of the demonstration group, so evaluators should use matching methods to balance covariates.

For out-of-state comparison groups, evaluators select one or more states that have a similar Medicaid program as the demonstration state but have not implemented the policy being evaluated. For example, the evaluation of the substance use disorder component of the "Building and Transforming Coverage, Services, and Supports for a Healthier Virginia" section 1115 demonstration (for the demonstration period January 12, 2015 through December 31, 2024) plans to use Medicaid beneficiaries from neighboring states as a comparison group.[7] When using an out-of-state comparison strategy, selecting the final comparison pool consists of two steps. First, evaluators need to identify one or more states from which they will draw comparison beneficiaries. Evaluators should select comparison states that have similar state-level characteristics, including average demographics characteristics, the health status and health care use of Medicaid beneficiaries, Medicaid program characteristics, state health care market characteristics, and state labor market conditions.[8] Depending on the demonstration type, some characteristics may be more important than others. Pohl and Bradley (2020) provide details on how to assess the similarity of the demonstration and comparison states, and they list suitable data sources.[9] Second, depending on the demonstration type, evaluators may need to further restrict the beneficiary pool in comparison states. For example, when evaluating a demonstration aimed at group VIII adults, the beneficiary pool in comparison states should similarly be restricted to group VIII adults. On the other hand, all Medicaid beneficiaries are typically eligible to participate in substance use disorder demonstrations, so the comparison pool would not have to be restricted in that case.

Depending on the demonstration and its components, evaluators may need to define different comparison groups for evaluating different components. This may be the case when affected eligibility groups vary across demonstration components. For example, the "Arizona Health Care Cost Containment System" demonstration (for the demonstration period October 1, 2016 through September 30, 2021) includes components for Medicaid beneficiaries at risk of institutionalization, children in the custody of the Department of Child Safety, and beneficiaries with a serious mental illness, among others. Once the potential in-state or out-of-state comparison group is identified, evaluators can use the formal matching methods described in Section III to draw individual beneficiaries from the potential comparison group to construct the final comparison groups that align with each evaluation component.

[6] The approved evaluation design is available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/downloads/or-health-plan-sud-demo-appvd-sud-eval-des-09292022.pdf.
[7] The approved evaluation design is available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/downloads/va-gov-access-plan-gap-aprvd-eval-design-05062021.pdf.
[8] Examples of Medicaid program characteristics include Medicaid expansion to cover the adult VIII group, income thresholds for different Medicaid eligibility groups, and Medicaid managed care; examples of state health care market characteristics include number of hospital beds, physicians, and mental health facilities; and examples of state labor market conditions include unemployment rate and average earnings.
[9] For details on out-of-state comparison groups, see "Selection of Out-of-State Comparison Groups and the Synthetic Control Method" (Pohl and Bradley 2020) at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

For both in-state and out-of-state comparison groups, before using formal matching methods, evaluators can trim the comparison group pool manually by removing beneficiaries who differ substantially from the demonstration group along some important covariates. This is important because some of the matching methods described below work better when extreme outliers are removed beforehand. In practice, evaluators specify key covariates, such as age and income, and a threshold beyond which they will drop potential comparison beneficiaries. For example, beneficiaries could be dropped from the comparison pool if their income is more than 2.5 standard deviations above the maximum income among the demonstration group or more than 2.5 standard deviations below its minimum. Evaluators should determine rules for trimming the comparison pool beforehand instead of dropping beneficiaries post hoc. When defining the comparison group, evaluators should also attempt to approximate the characteristics of the demonstration group. For example, the "Healthy Indiana Plan" demonstration (for the demonstration period January 1, 2021 through December 31, 2030) includes beneficiaries between 19 and 64 years of age with incomes below 138 percent of the federal poverty level.[10] To ensure overlap in important covariates, such as age and income, the potential comparison group should also be restricted to beneficiaries within the same age and income ranges.

Both types of comparison groups have advantages and limitations. One benefit of using in-state comparison groups is that, when comparing demonstration beneficiaries who are subject to the same health care system and economic conditions, outcomes are not confounded by those state-specific factors. A limitation is the possibility that Medicaid beneficiaries in other eligibility categories or having other dissimilar characteristics are inherently different from the demonstration group. An advantage of an out-of-state comparison strategy is the ability to construct a comparison group consisting of beneficiaries in the same eligibility category who otherwise have similar characteristics. However, it can be challenging to identify states that have similar economic conditions and a similar Medicaid program. Whether an in-state or out-of-state comparison group is more suitable for evaluating a section 1115 demonstration or individual component depends on the policy context and available data sources. If feasible, evaluators can use both an in-state and an out-of-state comparison group, which allows them to triangulate demonstration impacts.

[10] The demonstration approval is available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/downloads/in-healthy-indiana-plan-support-20-ca-20210722.pdf.

III. Matching Methods

A. Overview

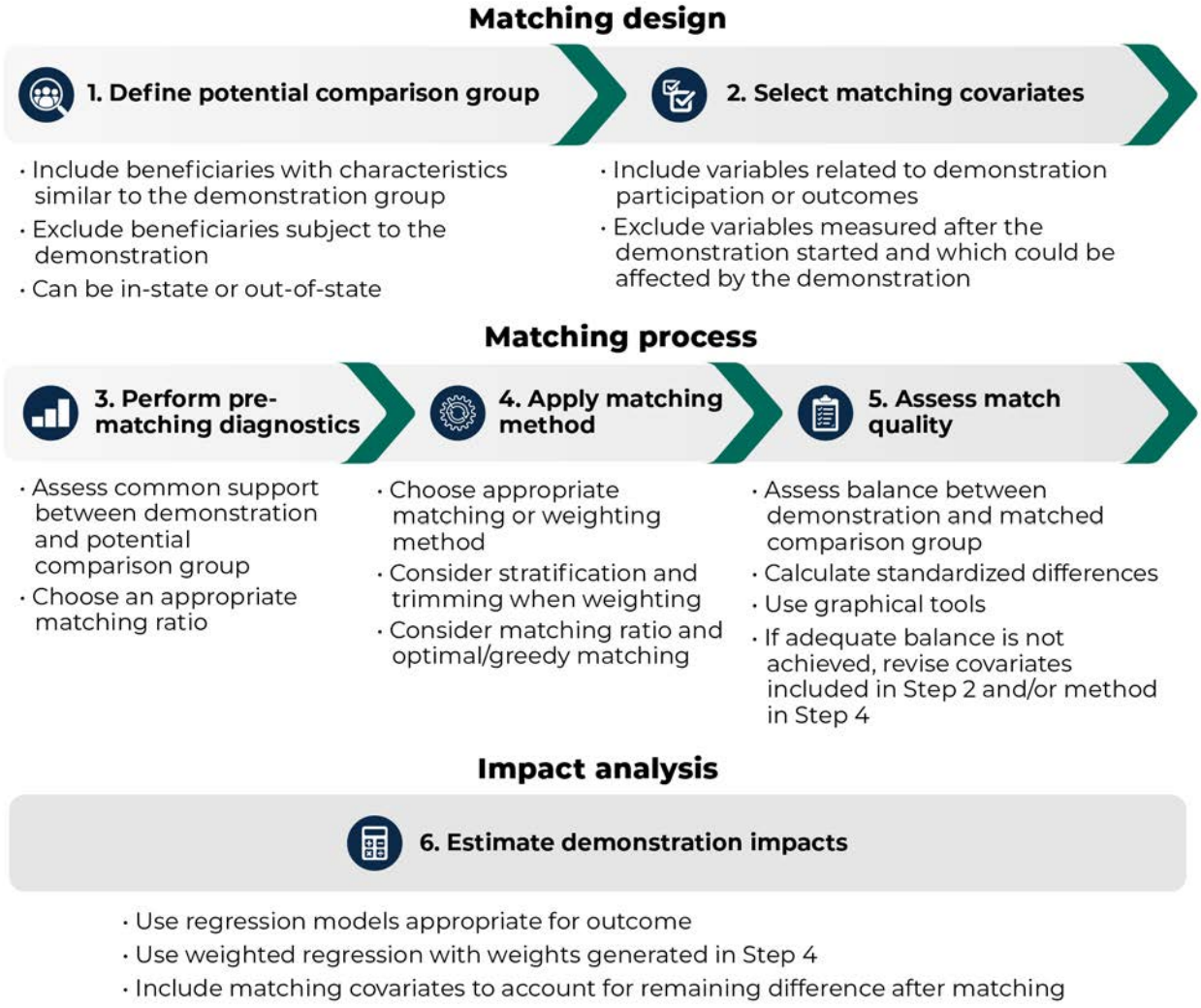
There are two main approaches that evaluators can use to balance the covariates in demonstration and comparison groups: weighting and matching.[11] When using weighting methods, evaluators estimate weights that reflect the propensity of each demonstration beneficiary and potential comparison beneficiary to participate in the demonstration, given their covariates.[12] Evaluators assign a weight to each demonstration and comparison group beneficiary and estimate impacts using weighted regression methods. This approach achieves balance between the covariates of demonstration and comparison groups. Rather than discarding unmatched beneficiaries as when using a matching method, weighting methods generally allow evaluators to use the full sample.[13]

When using a matching approach, a demonstration beneficiary will be matched to one or multiple comparison beneficiaries. That is, evaluators identify one or multiple beneficiaries in the potential comparison group (comparison pool) that resemble a demonstration beneficiary as closely as possible. The demonstration and matched comparison beneficiaries form a matched pair or matched set. Excluded from the analysis will be beneficiaries in the demonstration group for whom no similar comparison beneficiaries can be found and potential comparison beneficiaries who do not serve as a match for any demonstration beneficiary.

Figure III.1 summarizes the typical process that evaluators follow when using matching methods in the evaluation of section 1115 demonstrations. Overall, matching consists of three broad phases: (1) the matching design, (2) the matching process, and (3) the impact analysis. These phases are divided into six steps. We discuss Step 1 (defining a potential comparison group) above in Section II. The remainder of this section introduces the other steps of the weighting and matching processes in detail, including covariates to be included, pre-matching diagnostics, common matching methods, matching quality assessment, and analysis of outcomes. Matching is usually not a linear, but rather an iterative process, in which evaluators refine methods and the included covariates after assessing the quality of initial matches.

[11] Matching methods in a general sense encompass (1) procedures that involve generating weights that are used in outcome regressions (see Section III.D.1) and (2) matching procedures in a narrower sense, in which evaluators form matched pairs or sets of beneficiaries (see Section III.D.2).
[12] Evaluators can use a different type of weighting, the synthetic control method, when using an out-of-state comparison group. See Pohl and Bradley 2020 for details.
[13] When using a weighting method, it may be beneficial to trim extreme outliers from the sample before estimating demonstration impacts, see Section D.1.

Figure III.1. Summary of matching approach



B. Matching covariates

All matching methods rely on baseline characteristics of beneficiaries, health care providers, or other units of analysis. We refer to these characteristics as matching covariates. Matching methods can be successful in balancing demonstration and comparison groups only if they include the appropriate covariates. Therefore, evaluators must carefully consider which variables to include in and exclude from the matching models.

1. Variables to include

Evaluators should include any baseline covariates that could potentially confound impact estimates. Confounders are "variables that are thought to influence both the treatment and the outcome and can bias evaluation results if they are not controlled for" (Contreary et al. 2018, p. 3). Baseline covariates are determined before demonstration implementation and do not change due to the demonstration. The demonstration's logic model can guide evaluators in deciding which variables should be considered

confounders.[14] For example, one of the fundamental goals of Medicaid 1115 demonstrations is to improve beneficiaries' health outcomes, so evaluators should consider covariates that affect participation in a demonstration and beneficiaries' health outcomes. These covariates include baseline demographics, socioeconomic status, diagnoses, and health care utilization. Specifically, income determines Medicaid eligibility and is typically correlated with health. Existing health conditions may be a consideration when individuals decide to apply for Medicaid coverage and will affect subsequent health outcomes. Therefore, these and other variables should be included in the matching model that evaluators will use to assess the demonstration impact.

Evaluators should always include variables that are related to both demonstration participation and outcomes in matching models. However, selection of covariates related to either the outcome or demonstration participation is guided by a tradeoff between bias (the distance between estimated and true impact) and variance of the estimated impact. Specifically, if a covariate is related to the outcome but it is uncertain whether the covariate is related to demonstration participation, including it in the matching process can reduce bias even if it is not directly related to demonstration participation, and evaluators will generally want to include the covariate if the sample sizes are large. If not accounted for in the matching, the covariate might be an unmeasured confounder that will bias the estimated impact. For example, when estimating the health impact of a demonstration, evaluators should include neighborhood characteristics in matching models even though these covariates may not be directly related to demonstration participation, because they affect health outcomes. However, with a smaller sample size, including covariates that are possibly unrelated to the outcome may increase variance (that is, introduce too much "noise") into impact estimates and obscure any reduction in bias that is achieved by their inclusion.

Table III.1 provides a non-exhaustive list of variables that evaluators can consider when selecting a comparison group using matching methods. Evaluators are unlikely to use all possible variables from such a list but can instead select the most important confounding variables, given the demonstration design and intended outcomes. Table III.1 also lists suggested data sources for community-level variables that are publicly available; these sources are linked to websites with more information.

Table III.1. Examples of variables for matching model

Domain	Examples	Potential data sources
Beneficiary characteristics		
Demographic characteristics	• Age • Gender • Race/ethnicity • Language • Sexual orientation, gender identity	• Medicaid enrollment and claims data (for in-state comparison groups)

[14] Including confounders is related to the key concept of strong ignorability (exogeneity), which assumes that there are no unobserved differences between the demonstration and comparison groups, conditional on the observed covariates. To satisfy the assumption of ignorable treatment assignment, it is imperative to include in the matching process all variables that are believed to be related to both treatment assignment and the outcome (see, for example, Stuart 2010).

Domain	Examples	Potential data sources
Health characteristics in pre-implementation period[a]	• Heart disease • Cancer • Stroke • Lung disease • Diabetes • Hypertension • Substance use disorder • Serious mental illness/serious emotional disturbance	• T-MSIS Analytic Files (for out-of-state comparison groups) • American Community Survey • Behavioral Risk Factor Surveillance System • Consumer Assessment of Healthcare Providers and Systems
Health care use in pre-implementation period[a]	• Number of outpatient visits • Number of primary care visits • Number of specialist visits • Number of inpatient stays • Number of emergency department visits	
Medicaid eligibility and enrollment	• Continuous enrollment in Medicaid prior to the demonstration • Eligibility category • Participation in other demonstrations • Medicare-Medicaid dual enrollment • Income	• T-MSIS Analytic Files (for out-of-state comparison groups)
Provider characteristics		
Provider characteristics	• Training and experience of health care professionals • Practice or hospital size • Staffing • Ownership • Tax status (for profit or not for profit) • Hospital teaching status • Tenure of health care organization • Demographic and health characteristics of beneficiaries attributed to providers	• American Hospital Association • Medicare provider data
Community-level characteristics		
Neighborhood characteristics	ZIP-code level characteristics: • Urbanicity • Median income • Education • Housing characteristics • Deprivation index	• American Community Survey • Area Deprivation Index • See U.S. Community Indicators Project for additional data sources
Health care resources	County-level characteristics: • Number of hospital beds • Number of physicians • Number of primary care providers • Number of mental health providers • Number of nursing home beds	• Area Health Resources Files

[a] For diagnoses and health care use, states and evaluators need to make sure the observed time window between the demonstration and comparison individuals in the pre-implementation period is comparable. It is challenging if the demonstration or comparison group has not been enrolled in Medicaid before implementation of the demonstration; in that case, diagnoses and health care use derived from claims data would not be available.

2. Variables to exclude

Any variables that could be affected by the demonstration should be excluded from the matching process. One common example is health diagnoses and health care utilization that occurred after the demonstration began. Variables from post-implementation periods might be impacted by the implementation of the demonstration and may bias the estimated impact. Evaluators should instead use baseline characteristics (for example, baseline health diagnoses and health care utilization) in the matching process. If it is deemed critical to account for a variable that captures activity after demonstration implementation, it is better to exclude that variable in the matching procedure and include it as a covariate in the outcome analysis model. For example, evaluators could control local labor market conditions that may affect Medicaid enrollment by including contemporaneous unemployment rates as regression covariate but not as a matching variable.

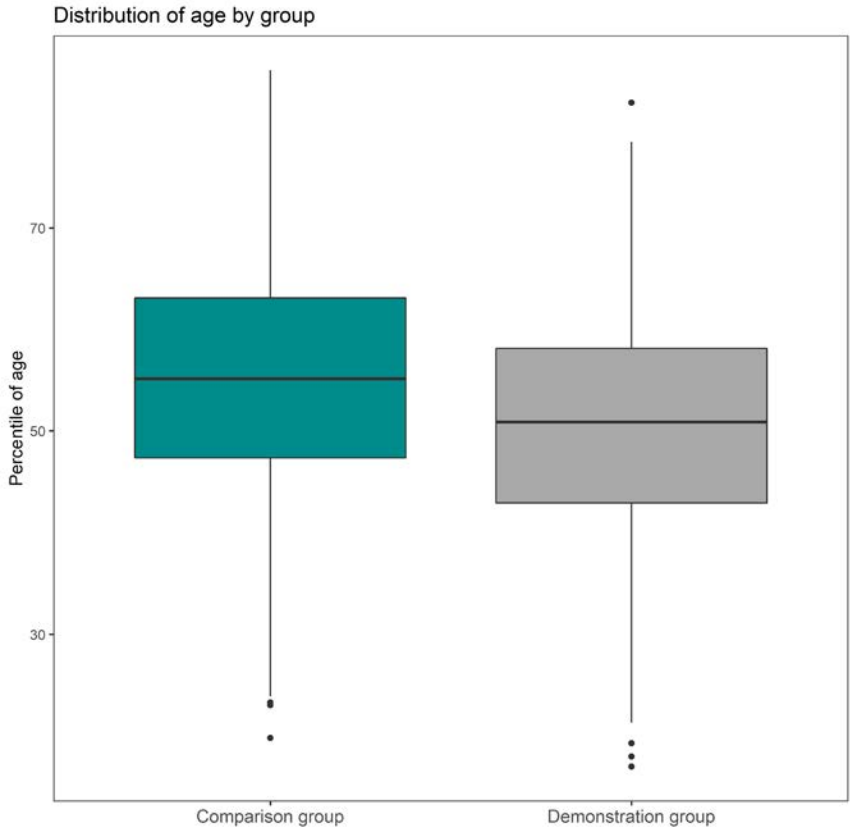
One challenge that can arise is that variables are not accurately and consistently measured in practice. For example, data from different sources (for example, Medicaid data from different states) might present different patterns of missingness. It is important to check the definitions, missingness, and distribution of the variables before matching, especially for key variables like race/ethnicity and income. Variables that are theoretically important but are differentially missing between the demonstration and comparison groups, may have to be excluded.

C. Pre-matching diagnostics

Pre-matching diagnostics constitutes an important step in the matching process that can help evaluators determine the feasibility of various matching methods. At this step, evaluators consider the entire comparison pool (see Section II), first assessing the overlap in covariate distributions (or "common support") between the demonstration and potential comparison groups. In practice, this is done by calculating the standardized difference between the demonstration and potential comparison group for all covariates.[15] If the standardized difference for many covariates is large, the comparison group may not constitute a suitable counterfactual, even after applying matching methods. A common threshold for a "large" standardized difference is above 0.25 (Imbens and Wooldridge 2009). In addition, evaluators can examine box plots of covariates in demonstration groups and potential comparison groups to assess overlap of the entire covariate distribution. Box plots (Figure III.2) show percentiles of a covariate distribution (25th percentile, median, and 75th percentile in the "box" and 5th and 95th percentiles on the end of the "whiskers" or vertical lines). Although there is no general rule for how to assess box plots, if the boxes depicting the range from the 25th to 75th percentile have little or no overlap, evaluators would conclude that demonstration and comparison groups are too different to apply matching methods. The pre-matching overlap in Figure II.2 would be considered adequate to proceed with applying matching methods.

[15] For each covariate, the standardized difference is defined as the difference between demonstration and comparison group means, divided by the square root of the average of demonstration and comparison group variances (Austin 2009).

Figure III.2. Example of box plot



As a second pre-matching diagnostic, evaluators should check the matching ratio for potential comparison to demonstration beneficiaries. The matching ratio is the number of beneficiaries in the potential comparison group divided by the number of beneficiaries in the demonstration group. If the matching ratio is low, that is, the number of potential comparison beneficiaries is relatively small compared to the demonstration group, there may not be enough comparison beneficiaries to provide a suitable match for each beneficiary in the demonstration group. See more details in the discussion of the number of matches in section III.D.2.d.

D. Common matching methods

In this section, we describe weighting and matching methods that evaluators can apply to the evaluation of a section 1115 demonstration. For each method, we first introduce commonly used approaches and then discuss specific considerations that can help evaluators make decisions about which methods to use and how to implement them. We include a hypothetical data set as a running example showing how these methods can be applied in practice. Evaluators can implement the weighting and matching methods using the statistical software packages Stata or R. The appendix contains list of commands for the methods introduced in this section.

1. Weighting methods

When using weighting methods, evaluators first derive weights that summarize the relationship between covariates and beneficiaries' treatment status (demonstration participation) in the design stage. Evaluators then apply these weights to the outcome regression by estimating weighted regressions in the analysis stage. The goal is to reweight the sample to balance or equate the covariate distributions of demonstration and comparison groups.

a. Propensity score weighting

Propensity score weighting (also known as inverse probability of treatment weighting) has become an increasingly popular matching method used in observational studies (Robins et al. 2000, Hirano and Imbens 2001). The propensity score is the probability of receiving a treatment (in this case, participating in a demonstration), given observed characteristics. For example, in a section 1115 demonstration that only covers beneficiaries with incomes below the federal poverty level, the propensity score would be higher for beneficiaries without a high school degree because educational attainment is typically correlated with income. A weight is calculated for each member of the demonstration and comparison group based on the probability of receiving treatment. Demonstration and comparison group members with higher weights receive more importance in the outcome analysis; that is, estimates of demonstration impacts rely more on beneficiaries with large propensity score weights than on those with smaller weights.

Evaluators usually estimate a logistic regression model to examine the characteristics associated with demonstration participation (for example, demographics, socioeconomic status, baseline diagnoses and health care utilization, and community characteristics). Evaluators should err on the side of including more variables when possible. Additional variables are useful for improving the propensity score models as long as they are not influenced by the demonstration itself. Even variables that are highly correlated with one another can improve the propensity score estimation, given a large sample size. The propensity score is the predicted probability of participating in the demonstration based on the logistic regression. Then, evaluators calculate weights based on the inverse probability of receiving and not receiving treatment and assign them to each member of the demonstration and comparison group, respectively.[16] Table III.2 shows a stylized example with three covariates for how to calculate the inverse propensity score weights based on a logistic model for demonstration participation. Propensity score weighting is an iterative process in which evaluators check balance between demonstration and comparison groups (Section III.E) and change the propensity score model if balance is not sufficient (for example, by adding matching covariates, interaction terms, or polynomials or categories of continuous covariates). For example, the evaluation of the substance use disorder component of the "New Jersey FamilyCare Comprehensive Demonstration" (for the demonstration period October 31, 2017 through June 30, 2022) plans to use a propensity score weighting approach that includes beneficiary characteristics such as sex, chronic disability payment score, race/ethnicity, and enrollment history.[17]

[16] For demonstration group beneficiaries, the weight is calculated as "1 / propensity score," and for members of the comparison group, it is calculated as "1 / (1 – propensity score)."

[17] The approved evaluation design is available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/downloads/nj-1115-request-cms-sud-eval-design-appvl-01302020.pdf.

Table III.2. Example for propensity score weighting

	Age	Diabetes	Zip code median income	Estimated propensity score	Inverse propensity score weight
D1	35	1	$18,000	0.8	1 / 0.8 = 1.25
D2	40	0	$30,000	0.6	1 / 0.6 = 1.67
D3	46	1	$35,000	0.4	1 / 0.4 = 2.50
C1	35	1	$18,000	0.8	1 / (1 – 0.8) = 5.00
C2	31	1	$25,000	0.5	1 / (1 – 0.5) = 2.00
C3	44	0	$50,000	0.1	1 / (1 – 0.1) = 1.11
C4	48	1	$36,000	0.3	1 / (1 – 0.3) = 1.43

C = comparison; D = demonstration.

There are extensions to propensity score weighting that evaluators can use. For example, the covariate balancing propensity score method proposed by Imai and Ratkovic (2014) estimates the propensity score similarly to what we described above but also imposes covariate balance. That is, in addition assigning a larger weight to comparison group members with a higher propensity score, this method also ensures that individual covariates are similar (balanced) between demonstration and comparison group. A generalization of this method, called the penalized covariate balancing propensity score, limits the variability of the propensity score weights to increase statistical power (Kranker et al. 2020).

b. Entropy balancing

Entropy balancing is a method for achieving covariate balance by involving a reweighting scheme that directly incorporates covariate balance into the weight function that is applied to the sample (Hainmueller 2011). The most important feature of entropy balancing is that it allows evaluators to obtain a high degree of covariate balance. Good covariate balance is a desirable outcome of a matching or weighting method because it means that demonstration and comparison group are very similar based on observed characteristics. Entropy balancing enables this high covariate balance by imposing a potentially large set of balance constraints that serve to equate the moments of the covariate distributions (for example, the means, variances, and covariances) between the reweighted demonstration and comparison group. It could (at least weakly) improve upon the balance that can be obtained by other conventional matching and weighting methods with respect to the specified balance constraints. Entropy balancing is unlike the propensity score weighting approach, in which evaluators first estimate the propensity score weights with a logistic regression and then compute balance checks to assess whether applying the estimated weights equates the covariate distributions for demonstration and comparison groups. Entropy balancing, in contrast, directly adjusts the weights to the sample, thereby obviating the need to continually check balances and search for propensity score models that may balance the prespecified covariates.

Evaluators begin by selecting a potentially large set of balance constraints, which ensures that the covariate distributions of the demonstration and comparison group match exactly on all prespecified variables (including different functions of the variables, such as interactions and squared terms).[18] Evaluators use entropy balancing to search for the set of weights that satisfies the balance constraints but

[18] For example, instead of only including age and race/ethnicity covariates, evaluators could add a squared term of age and interactions between age and race/ethnicity.

remains as close as possible to a set of base weights. This recalibration of the sample weights effectively adjusts for systematic and random differences in the demonstration and comparison groups.[19]

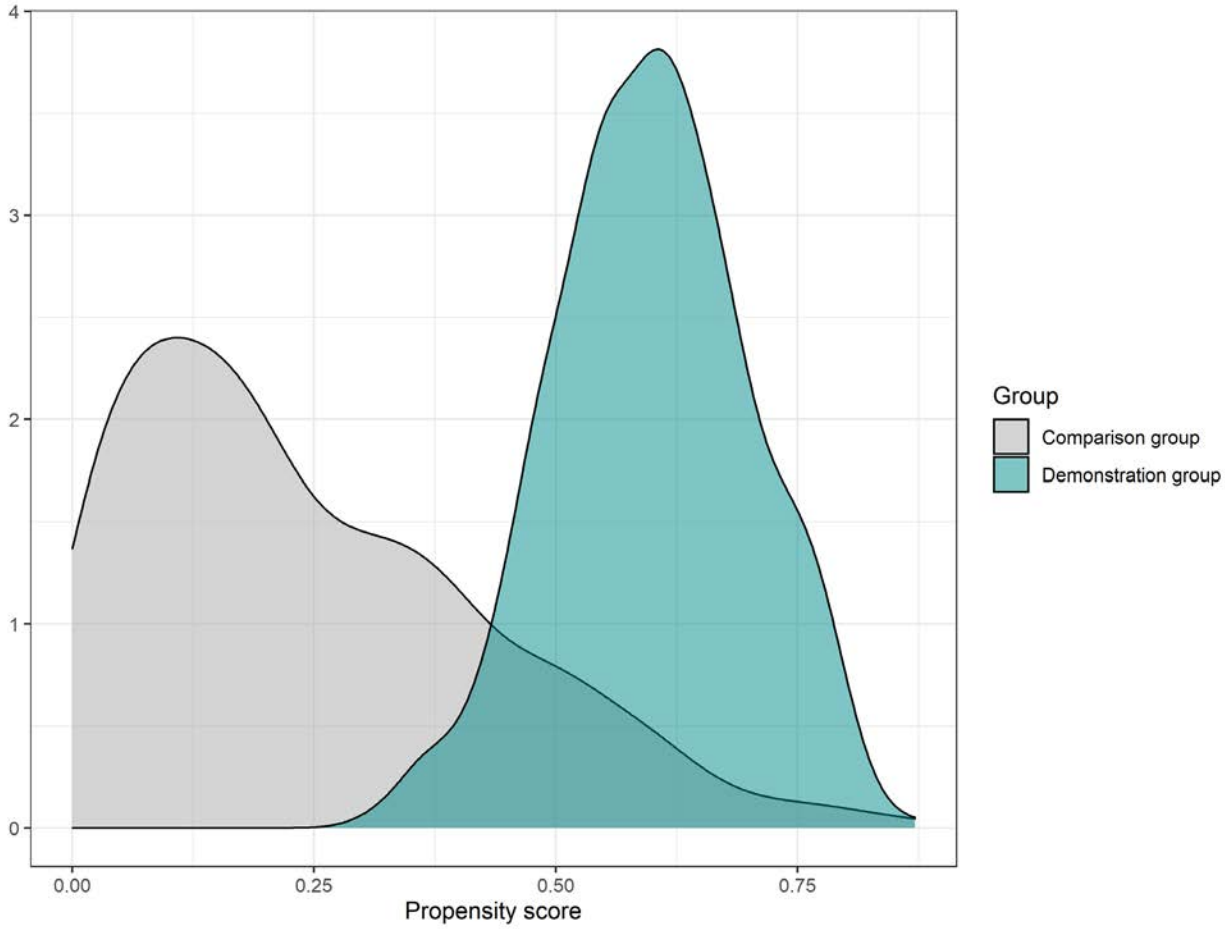
c. Considerations when using weighting methods

Stratification. Evaluators can stratify beneficiaries into groups with similar estimated propensity scores (for example, as defined by quintiles of the propensity score distribution) to improve the balance between demonstration and comparison beneficiaries within each stratum (Lunceford and Davidian 2004). This approach can lead to better balance overall than propensity score weighting without stratification. Overall demonstration impacts can then be estimated by combining stratum-specific impact estimates into a single value. Rosenbaum and Rubin (1984) show that creating five propensity score strata can remove 90 percent of the bias in the estimated impact that is due to the covariates that enter the propensity score model. Based on those results, the current convention is to use 5 to 10 strata. However, with larger sample sizes, more strata (for example, 10 to 20) may be feasible and appropriate.

Assessing common support and need to trim. One potential issue of the propensity score weighting approach is that of insufficient common support—or poor overlap in the distribution of estimated propensity scores between demonstration and comparison group members. For example, some of the comparison beneficiaries may be very different from all of the demonstration group members. In this case, it may be beneficial to explicitly restrict the analysis to those beneficiaries in the region of common support, for example, where the propensity score distributions of the demonstration and comparison groups overlap. The common support in Figure III.3 would not be considered sufficient because the demonstration and comparison group distributions do not overlap in the lower range of the propensity score.

[19] Stable balancing weights is a related weighting approach developed by Zubizarreta (2015). When using this method, evaluators select weights with the smallest possible variance across beneficiaries (stable) while also ensuring that covariates are similar, on average, between the demonstration and comparison groups (balancing). Entropy weighting and stable balancing weights fall under the class of weighting methods called "minimal dispersion approximately balancing weights." Both these weighting approaches minimize the dispersion or variance of the beneficiary-specific weights while imposing constraints on how different the covariates for the demonstration and comparison groups can be (Wang and Zubizarreta 2019).

Figure III.3. Example for assessing overlap between the demonstration and comparison groups



Assessing for extreme weights and need to trim. A potential drawback of the propensity score weighting approach is that the variance can be very large if the weights are extreme, that is, if the estimated propensity scores are close to 0 or 1. If the model is correctly specified and thus the weights are correct, then the large variance is appropriate. However, it might be problematic if some of the extreme weights are related more to the estimation procedure than to the true underlying probabilities. Weight trimming, which sets weights above some maximum to that maximum, has been proposed as one solution to this problem. The trimming level will depend on the size of weights and potential model misspecification (Stuart 2010).

2. Matching methods

When using matching methods (as opposed to weighting methods), evaluators identify comparison beneficiaries that are "close" to demonstration beneficiaries based on a distance measure. We describe a frequently used matching method involving a distance measure here: matching based on the propensity score.[20] Alternatively, evaluators can find a match with identical characteristics for each demonstration group beneficiary; we describe coarsened exact matching as an example.

[20] There are other matching methods based on distance measures, such as the Mahalanobis metric or the Euclidean metric. These methods are conceptually similar to matching methods that use the propensity score as a distance measure, and we do not describe them here.

a. Propensity score matching

Propensity score matching is a quasi-experimental method in which evaluators construct a comparison group by matching each beneficiary in the demonstration group with one or multiple beneficiaries in the comparison pool who have a similar estimated propensity score, that is, the probability of participating in the demonstration on the basis of prespecified covariates (Stuart 2010; see Section III.D.1.a on estimating the propensity score). Intuitively, beneficiaries with similar propensity scores also have similar characteristics as far as participation in the demonstration is concerned. Grouping beneficiaries with similar propensity scores replicates a randomized experiment, at least with respect to the observed covariates. Propensity score matching is one of the most commonly used matching approaches since Rosenbaum and Rubin (1983) introduced the technique. For example, the evaluation of the Illinois Behavioral Health Transformation demonstration (for the demonstration period July 1, 2018 through June 30, 2023) proposes to use propensity score matching with matching covariates including county of residence, age group, sex, income, Medicaid plan type, presence of children in the household, health, and health care use.[21] The evaluation for the Vermont Global Commitment to Health (for the demonstration period January 1, 2017 through December 31, 2021) proposes to use a propensity score matching approach to control for differences in demographic and delivery system characteristics between Medicaid beneficiaries who are aligned and not aligned with the state's accountable care organization.[22] Using the same hypothetical example as in Table II.2, Table III.3 shows how propensity score matching selects for each demonstration beneficiary (D1–D3) the comparison individual with the smallest distance between propensity score one by one. That is, the algorithm selects for D1 beneficiary C1, for D2 the comparison beneficiary from the remaining comparison beneficiaries with the smallest distance (C2), and so on.

Table III.3. Example for propensity score matching

	Age	Diabetes	Zip code median income	Estimated propensity score	Matched comparison beneficiary
D1	35	1	$18,000	0.8	C1
D2	40	0	$30,000	0.6	C2
D3	46	1	$35,000	0.4	C4
C1	35	1	$18,000	0.8	n/a
C2	31	1	$25,000	0.5	n/a
C3	44	0	$50,000	0.1	n/a
C4	48	1	$36,000	0.3	n/a

C = comparison; D = demonstration; n/a = not applicable.

b. Coarsened exact matching

Exact matching based on observed covariates between the demonstration and comparison groups is ideal because it ensures that both groups have the same characteristics, on average. However, requiring exact matches (for example, on the income level) often leads to many beneficiaries not being matched. Using exact matching can result in larger bias than if the matches are inexact, but more beneficiaries remain in the analysis. Neither approach works well when there are many covariates. Coarsened exact matching can

[21] The approved evaluation design is available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/downloads/il-behave-health-transform-appvd-eval-des-08132021.

[22] The approved evaluation design is available at https://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Waivers/1115/downloads/vt/Global-Commitment-to-Health/vt-global-commitment-to-health-eval-dsgn-appvl-20190609.pdf.

be used when there are not too many covariates to match on (five to eight, for example, depending on the number of values these covariates can take on) and the covariates may take on many values, such as income or age. The basic idea of coarsened exact matching is to coarsen covariates into a smaller number of values (Iacus et al. 2012). Then the exact matching algorithm is applied to the coarsened data to determine the matches. For example, instead of matching on the continuous income level, evaluators could recode the income into categories (such as 0–25, 25–50, or 50–75 percent of the federal poverty level, and so on) and then match the comparison beneficiaries based on these income categories. Continuing the example from above, age and median income are continuous covariates and can be coarsened into categories. Table III.4 shows an exact matching of demonstration and comparison beneficiaries based on coarsened covariates and diabetes diagnosis. Note that this algorithm does not find a match for one of the demonstration beneficiaries, who would be excluded from the analysis.

Table III.4. Example for coarsened exact matching

	Age	Diabetes	Zip code median income	Coarsened age category	Coarsened income category	Matched comparison beneficiary
D1	35	1	$18,000	35 – 39	$10K – $20K	C1
D2	40	0	$30,000	40 – 44	$20K – $30K	none
D3	46	1	$35,000	45 – 49	$30K – $40K	C4
C1	35	1	$18,000	35 – 49	$10K – $20K	n/a
C2	31	1	$25,000	30 – 34	$20K – $30K	n/a
C3	44	0	$50,000	40 – 44	$40K – $50K	n/a
C4	48	1	$36,000	45 – 49	$30K – $40K	n/a

C = comparison; D = demonstration; n/a = not applicable.

c. Considerations when using matching methods

Number of matches. When matching demonstration and comparison beneficiaries, evaluators need to choose how many matches to include, that is, the matching ratio. For each demonstration beneficiary, evaluators could include a single matched comparison beneficiary (1:1 matching), more than one matched comparison beneficiary (1:n), or fewer than one matched comparison beneficiary (n:1). The number of matches will depend on the relative sample size of the demonstration and potential comparison groups, how different their characteristics are before matching, and the likely quality of matches (Stuart 2010). When match quality is high, the evaluator could add more matches for each treated individual to increase the sample size. However, selecting multiple comparison beneficiaries for each demonstration beneficiary may increase bias, because the second and third closest matches, for example, are less similar to demonstration beneficiaries than the closest match. This bias may not be fully accounted for with regression adjustment. If the matching ratio is low, that is, the number of potential comparison beneficiaries is relatively small compared to the demonstration group, it would not be feasible to apply 1:n matching. In settings where the outcome data has yet to be collected, evaluators must also consider the cost arising from using multiple matches. In some cases, matching with a fixed ratio is not optimal because some demonstration beneficiaries may have many close matches while others have few. In such cases, evaluators can use variable ratio matching, which allows the matching ratio to vary, with different demonstration beneficiaries receiving different numbers of matches (Pimentel et al. 2015).

Matching with or without replacement. Evaluators need to decide whether potential comparison beneficiaries can be used as matches for more than one demonstration beneficiary, that is, whether the matching should be done with or without replacement (Stuart 2010). Matching with replacement can

often decrease bias, because comparison beneficiaries with characteristics similar to multiple demonstration beneficiaries can be used in the matched comparison group multiple times. This is particularly helpful in scenarios in which few comparison beneficiaries are comparable to the demonstration beneficiaries. However, statistical inference becomes more complex when matching with replacement, because the matched comparison beneficiaries are no longer independent. If some comparison beneficiaries can enter the matched sample multiple times, evaluators need to account for this, for example, by using frequency weights. However, it is also possible that the impact estimate is biased if only a small number of comparison beneficiaries enter the comparison group. When matching with replacement, evaluators should assess the number of times each comparison beneficiary is matched to a demonstration beneficiary (Stuart 2010). For example, it would be concerning if all demonstration beneficiaries were matched to only a small number of distinct comparison beneficiaries.

Optimal matching vs. greedy matching. One of the most common and easiest methods for matching without replacement is "nearest neighbor" matching (or "greedy" matching), which selects for each demonstration beneficiary the comparison individual(s) with the smallest propensity scores distance from that beneficiary. However, the order in which the demonstration members are matched may change the quality of the matches. Specifically, once the pair with the smallest distance is matched, the remaining demonstration and comparison beneficiaries might have a relatively large distance. Evaluators can avoid this issue by using optimal matching to consider the overall set of matches when choosing comparison beneficiaries, minimizing a global distance measure (Rosenbaum 1989). Table III.5 provides examples of greedy matching and optimal matching. Optimal matching achieves better balance by minimizing the total distance within matched pairs rather than minimizing the distance for each matched pair in one specific order.

Table III.5. Example of greedy vs. optimal matching

Greedy matching

	C1	C2	C3	C4	C5
D1	0.09	**0.06**	0.15	0.39	0.56
D2	0.34	0.12	**0.27**	0.63	0.49
D3	0.45	0.36	0.24	**0.11**	0.17
Total distance: 0.06 + 0.27 + 0.11 = **0.44**					

Optimal matching

	C1	C2	C3	C4	C5
D1	**0.09**	0.06	0.15	0.39	0.56
D2	0.34	**0.12**	0.27	0.63	0.49
D3	0.45	0.36	0.24	**0.11**	0.17
Total distance: 0.09 + 0.12 + 0.11 = **0.32**					

Notes: In the left panel, the greedy matching algorithm minimizes the distance within each matched set sequentially, that is, it selects for each demonstration beneficiary (D1–D3) the comparison individual with the smallest distance one by one. The algorithm selects for D1 the comparison individual with the smallest distance (C2), for D2 the comparison individual from the remaining comparison beneficiaries with the smallest distance (C3), and for D3 the comparison individual from the remaining comparison beneficiaries with the smallest distance (C4). In the right panel, where the order of matching is not relevant, the optimal matching algorithm minimizes the total distance across all matched sets, that is, it selects for each demonstration beneficiary a comparison beneficiary to minimize the total distance. The algorithm selects for D1 the comparison C1 with the second smallest distance and selects for D2 the comparison C2 and for D3 the comparison C4 with the smallest distances, such that the total distance is the smallest. Bolded numbers indicate best match.

C = comparison; D = demonstration.

Optimal matching with rolling enrollment. For many section 1115 demonstrations, beneficiaries enroll on a rolling basis, for example, when they become eligible. The enrollment dates and baseline periods for demonstration group beneficiaries are known, but the baseline period for potential comparison beneficiaries is not well defined. The same comparison beneficiary could be a good match for different demonstration beneficiaries at different times. In such cases, evaluators can use a promising new

approach called GroupMatch (an extension of optimal matching), which allows each comparison beneficiary to enter the comparison pool as several copies, each of which corresponds to a different enrollment date. Evaluators do not want to select more than one copy of the same comparison beneficiary from the comparison pool, because the matched comparison beneficiaries would no longer be independent, and GroupMatch ensures that only one copy of each comparison beneficiary can be selected for the comparison group (Pimentel et al. 2020). Table III.6 provides an example of GroupMatch when there are three potential comparison beneficiaries with two or three possible enrollment dates.

Table III.6. Example for optimal matching with rolling enrollment

Optimal matching with rolling enrollment								
		C1			C2		C3	
	E1.1	E1.2	E1.3	E2.1	E2.2	E3.1	E3.2	E3.3
D1	0.21	**0.09**	0.12	0.13	0.06	0.82	0.39	0.96
D2	0.41	0.34	0.29	**0.12**	0.21	0.73	0.63	0.54
D3	0.36	0.45	0.64	0.36	0.30	0.21	**0.11**	0.26

Notes: The matching algorithm selects at most one enrollment date (E#) from each unique comparison beneficiary (C#) by minimizing the total distance between demonstration and comparison observations. The algorithm selects for D1 the comparison E1.2 with the second smallest distance and selects for D2 the comparison E2.1 and for D3 the comparison E.3.2 with the smallest distances, such that the total distance is the smallest. Bolded numbers indicate best match.

C = comparison; D = demonstration.

Caliper matching. With caliper matching, evaluators form pairs of demonstration and comparison beneficiaries such that the difference in the measured distance (for example, propensity scores) between matched beneficiaries is no greater than a pre-fixed distance (the caliper width). This approach can also be used for individual covariates. For example, caliper matching could ensure that demonstration beneficiaries are matched only to comparison beneficiaries whose income differs at most by some prespecified amount. Evaluators can use calipers with a variety of matching methods, including optimal matching and matching with and without replacement. Using narrower calipers will result in the matching of more similar demonstration and comparison beneficiaries, based on observed covariates, reducing bias in the estimated impact of the demonstration.[23] However, using narrower calipers may result in a reduction in the number of matched beneficiaries and may worsen balance for other variables in the matching model. Smaller caliper width is necessary if the variance of the propensity score in the demonstration group is much larger, for example, more than twice the size, of the variance for the comparison group. Table III.7 shows an example for propensity score matching that imposes a caliper of five years for age. In contrast to propensity score matching without calipers, demonstration beneficiary D2 now gets matched to comparison beneficiary C3 because the age difference between D2 and C2 exceed five years.

Table III.7. Example for propensity score matching with caliper

	Age	Diabetes	Zip code median income	Estimated propensity score	Matched comparison beneficiary	
					Without caliper	With caliper
D1	35	1	$18,000	0.8	C1	C1
D2	40	0	$30,000	0.6	C2	C3
D3	46	1	$35,000	0.4	C4	C4

[23] For example, Austin (2011) shows that a caliper of 0.2 standard deviations of the logit of the propensity score can remove 98 percent of the bias in the estimated impact. The logit of the propensity score refers to the logistic function of the probability of participating the demonstration (propensity score).

	Age	Diabetes	Zip code median income	Estimated propensity score	Matched comparison beneficiary	
					Without caliper	With caliper
C1	35	1	$18,000	0.8	n/a	n/a
C2	31	1	$25,000	0.5	n/a	n/a
C3	44	0	$50,000	0.1	n/a	n/a
C4	48	1	$36,000	0.3	n/a	n/a

C = comparison; D = demonstration; n/a = not applicable.

E. Assessing match quality

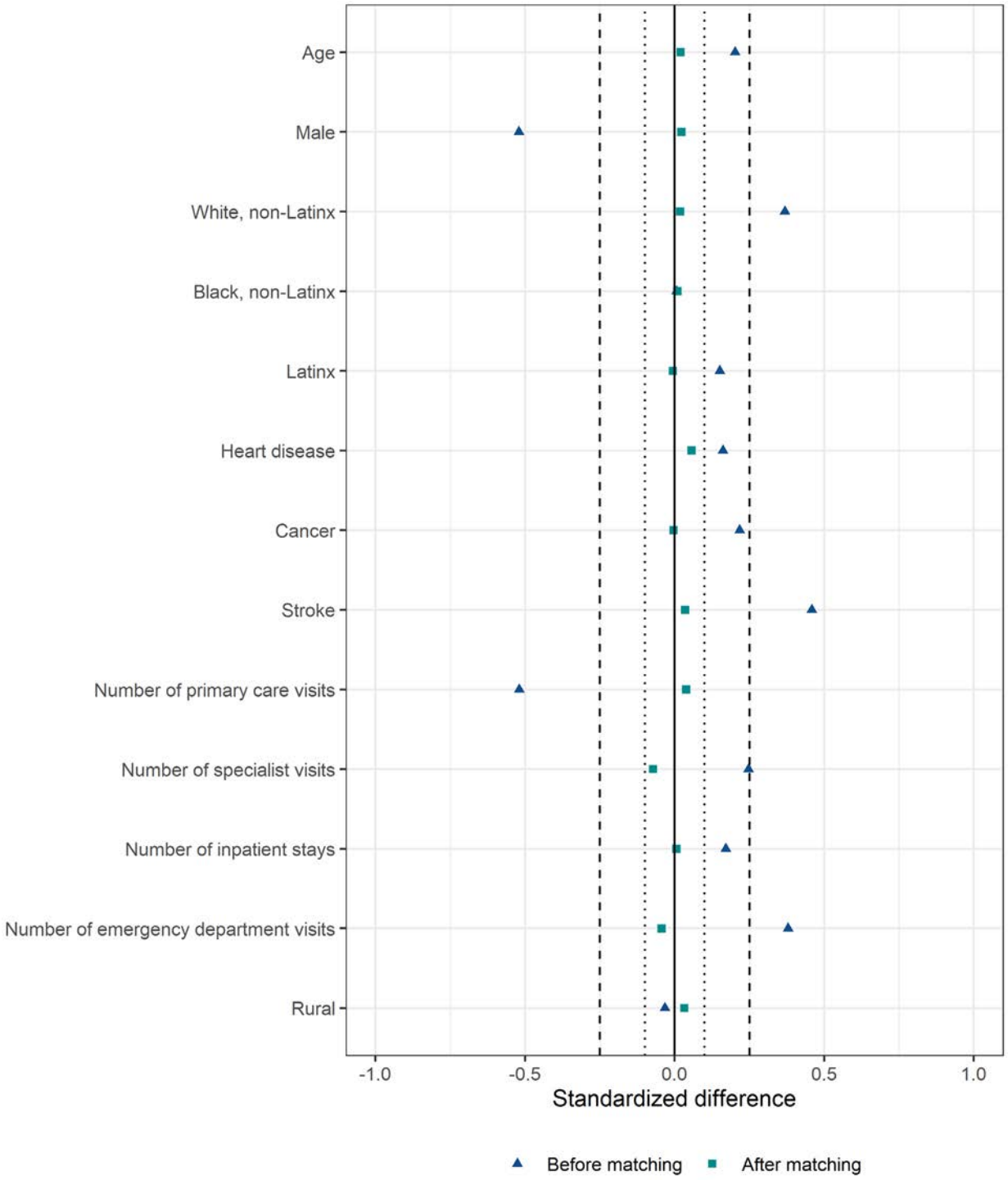
Once evaluators have identified a comparison group using one or more of the methods described in Section III.D, they must assess the resulting balance between demonstration and comparison groups. Evaluators using matching methods to select a comparison group often assess multiple methods and perform several iterations before selecting the final comparison group to use for the outcome analysis. If the evaluation includes subgroup analyses, evaluators should similarly assess balance for each subgroup separately. Achieving balance in every subgroup can be challenging, given the smaller sample sizes.

The main tools for assessing balance are (1) standardized differences between the demonstration and comparison group and (2) distributional plots such as density, box, or quantile-quantile (QQ) plots that compare the demonstration group to both the unmatched/unweighted and matched/weighted comparison groups.[24] Evaluators can also use box plots to assess pre-matching overlap between the demonstration and comparison group (Section III.C).

Standardized differences are an important measure for assessing whether demonstration and comparison groups have similar means. Evaluators should calculate standardized differences to assess all key covariates, regardless of whether they were included in the matching process. With beneficiary-level data, standardized differences after matching or weighting of less than 0.25 for all variables and less than 0.10 for key variables is generally considered to be well balanced (Institute of Education Sciences 2014). With provider-level data, evaluators typically have a smaller sample and therefore may not achieve balance as close as with beneficiary-level data. Standard practice is to report these results in a Love plot (Figure III.4), which shows the standardized differences between demonstration and comparison groups before and after the matching or weighting. In a Love plot, variables that move closer to a standardized difference of zero (the solid vertical midline) after matching or weighting show improved balance.

[24] For more details on how to assess match quality with the techniques recommended in this paper, see Austin (2009) and Stuart (2010).

Figure III.4. Example of Love plot



Note: Dashed lines represent 0.25 standardized differences, and dotted lines represent 0.1 standardized differences.

In addition to assessing the mean difference between key variables, evaluators should assess variable distributions before and after matching using density plots, box plots, and QQ plots. Such diagnostics can reveal cases in which treatment and comparison groups may have similar means (standardized differences close to zero) but the distributions remain different. In a density plot (Figure III.5), evaluators look for overlap between the distributions of key covariates. Ideally, distributions in demonstration and matched comparison groups look identical, but in practice it is often only feasible that matching improves overlap somewhat. Finally, QQ plots (Figure III.6) plot the quantiles (for example, percentiles) of covariates in the demonstration group against the quantiles in the comparison group, both before and after matching. Evaluators can use QQ plots to check that matching moves the plot closer to the 45-degree line. A perfect match on the particular covariate would be achieved if the QQ plot lies on the 45-degree line.

If good balance is not achieved in terms of standardized differences or distributions, especially for key covariates, evaluators can try several different approaches to improve balance. First, they can try adding calipers to key variables (see Section III.D.2.d). A caliper will improve balance for that key variable, but it will often worsen balance for other variables in the model. Second, evaluators using propensity score models can try changing the model by adding interaction terms, polynomials of continuous variables (such as squared or cubed terms), or categorical variables (such as an indicator variable for "ages 65 and older" in addition to a continuous measure of age). Finally, evaluators could consider including additional covariates or including other matching or weighting methods, described in Section III.D, if initial selections yield poor results (see Section III.B).

Figure III.5. Example of density plot

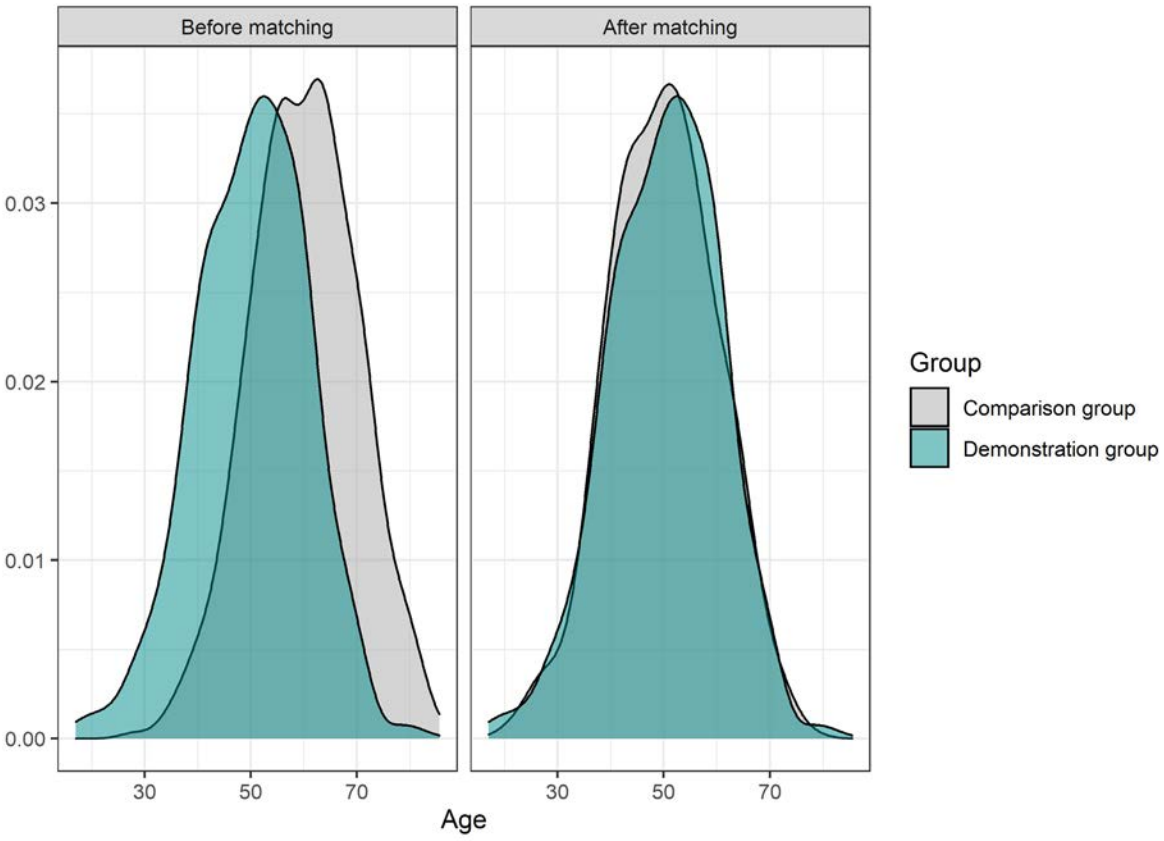
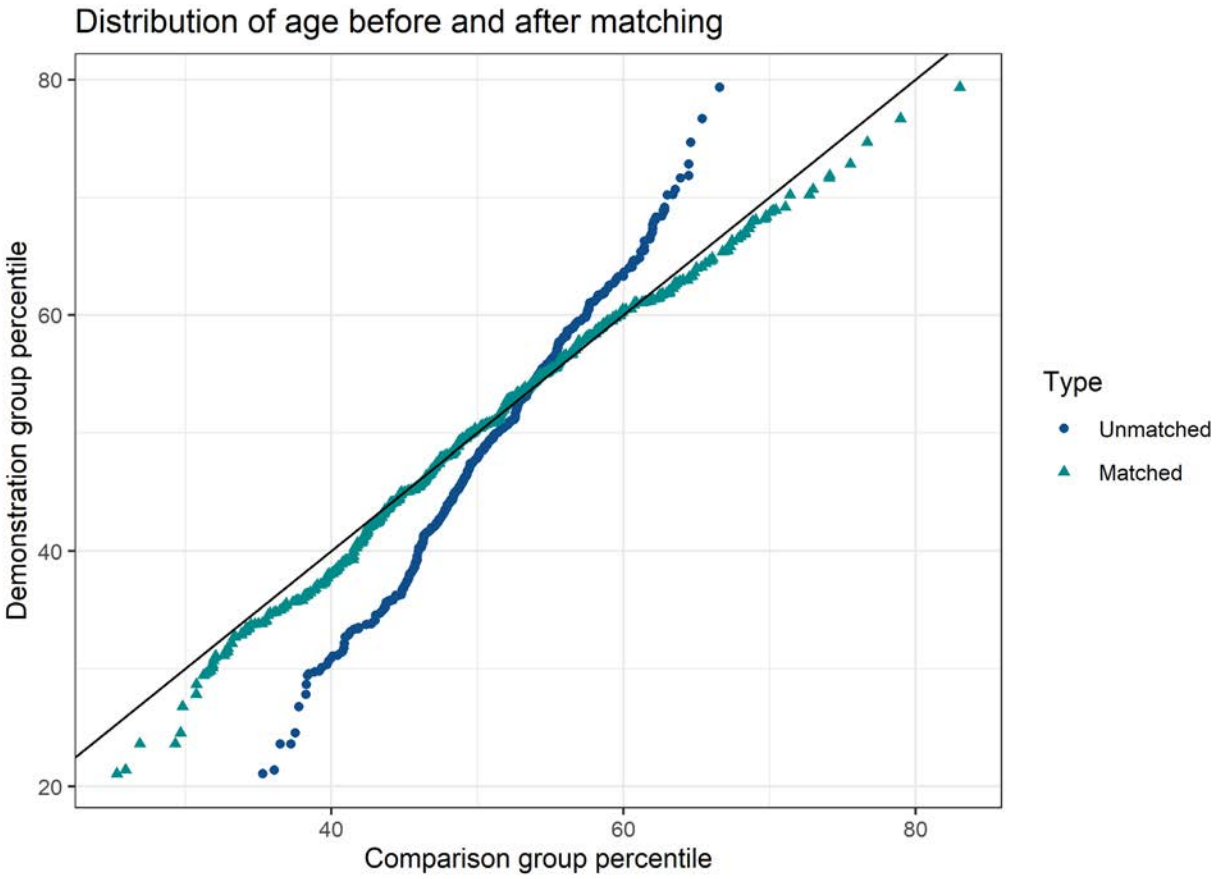


Figure III.6. Example of QQ plot



F. Outcome analysis with matched or weighted samples

In general, combining matching or weighting with regression adjustment, which is referred to as a "double robust" approach, leads to more reliable estimates of demonstration impacts because it creates doubly robust estimates. Regression adjustment after matching or weighting can adjust for residual differences that inevitably remain after matching or weighting, and it can be particularly important when matching or weighting does not achieve good balance. Regression adjustment without matching may not be a sufficient solution for avoiding biased impact estimates when covariates differ substantially among demonstration and comparison groups or when the relationships between the outcome and covariates cannot be accurately specified (Imbens 2015).

When using weighting methods, it is common to weight all observations in the outcome regression by the inverse of the propensity score, such that treatment assignment in the sample is independent of observed covariates. In this approach, evaluators may also want to trim comparison beneficiaries with propensity scores close to zero because they are very different from demonstration beneficiaries (see Section D.1.c). Where propensity score weights are large, evaluators can use stabilized inverse weights to account for observations with propensity scores close to zero (Austin and Stuart 2015).

When using matching methods, evaluators should also weight comparison observations to avoid the possibility that sets with more comparison beneficiaries have more importance in the outcome analysis.

Here, evaluators should weight each comparison observation by 1/n, where n is the number of comparison beneficiaries matched to a given demonstration beneficiary within a matched set. For example, if a demonstration beneficiary has three matched comparisons, each comparison would receive a weight of 1/3 while the demonstration beneficiary would receive a weight of 1.

When conducting subgroup analyses with matched comparisons (for example, by sex, or region), treatment beneficiaries could be matched to comparison beneficiaries who are not in the same subgroup if the evaluators did not use exact matching for subgroup identifiers. If the evaluators used exact matching for the subgroup identifier, all treatment and matched comparison beneficiaries will be in the same subgroup by construction. In cases where evaluators do not use exact matching and demonstration and matched comparison beneficiaries do not belong to the same subgroup, evaluators can drop discordant matches and reweight appropriately.

IV. Discussion

A. How to choose the best matching or weighting approach

When evaluators are choosing from the matching or weighting methods described in Section III.D to construct comparison groups, it is not always obvious which approach is best for the design of an evaluation. In general, it is usually best to try multiple approaches and choose the method that creates the best balance between demonstration and comparison groups. Good balance is easier to achieve when there is significant overlap between the treatment group and the comparison pool along key variables; this point highlights the importance of selecting an appropriate comparison beneficiary pool (see Section II). Which method to choose might also depend on which research questions the evaluators are trying to answer. The following factors should also be considered:

Statistical power. When analyses are underpowered, that is, when the sample is not large enough to detect meaningful demonstration impacts, evaluators should select a matching or weighting approach that preserves as many beneficiaries in the comparison group as possible. Weighting methods, which keep almost all possible comparison beneficiaries in the sample, are often best for improving power when there is concern about whether the sample is large enough to detect meaningful demonstration impacts. If the evaluators use matching, approaches that include higher ratios of comparison beneficiaries (1:n matching) can generate more statistical power than those that use lower matching ratios, assuming good balance can be achieved and the comparison pool is large enough.

Relative size of the treatment group and comparison beneficiary pool. When the comparison beneficiary pool is much larger than the demonstration group, finding high quality matches becomes easier. When there are relatively few comparison beneficiaries to choose from, matching with lower ratios of comparison to demonstration beneficiaries, matching with replacement, or weighting approaches could be better choices.

Number of variables included in the matching process. When evaluators wish to include a large number of variables in the matching or weighting approach, they should consider propensity score matching or weighting, because these models can easily handle large numbers of variables. When there are fewer variables, evaluators could use coarsened exact matching to assign one or more matched comparison beneficiaries to each demonstration beneficiary.

Whether subgroup analyses will be included. If the evaluator is conducting subgroup analyses (for example, by sex or age categories), matching methods preferably would use exact matching on subgroup identifiers to ensure that each matched set belongs to the same subgroup. In cases where exact matching on subgroup identifiers is not feasible, matching should also be done without replacement. Matching without replacement would avoid the possibility of a single comparison beneficiary being matched to more than one treatment beneficiary, each in different subgroups.

Cost for data processing or surveys for the comparison group. If evaluators plan to survey the comparison group to obtain additional information, and surveys are costly, evaluators may prefer to have fewer individuals in the comparison group, and therefore choose matching over weighting. Matching, especially with a lower matching ratio, can create a small but well-balanced comparison group. Weighting, on the other hand, removes relatively few individuals from the comparison group and leaves a larger comparison group.

B. Limitations of matching methods

Matching and weighting methods can greatly improve the quality of the comparison group by minimizing the differences between demonstration and comparison groups. However, matching and weighting approaches can account only for differences that are observed in the data. Matching and weighting cannot address unobserved confounders and cannot fully address selection bias when eligible beneficiaries can opt into the treatment group. Evaluators must continue to use best practices when designing their analysis to ensure a valid comparison group is selected.

Evaluators should use caution when combining matching methods with difference-in-differences regression models or models that use lagged outcome variables. For example, when evaluating the impact of a demonstration policy on emergency department visits, by matching on the frequency of emergency department visits during the baseline year, evaluators may erroneously select beneficiaries for the comparison group whose values are likely to revert to a long-run mean value that is different from the baseline year, which could lead them to conclude incorrectly that the demonstration reduced emergency department visits. This outcome is particularly likely when demonstration and unmatched comparison beneficiaries have very different average outcomes at baseline. Whether to include a baseline outcome variable may depend on the correlation of the outcome variables over time. If the baseline outcome variable is believed to be correlated with the outcome in the post-implementation period, including the baseline outcome variable in the matching will be appropriate and reduce the bias; otherwise, evaluators should not include baseline outcomes as a matching variable (see Daw and Hatfield 2018 for details).

C. Specifying and documenting matching or weighting approaches in section 1115 demonstrations

When specifying a matching or weighting approach in the evaluation design for a section 1115 demonstration, evaluators should specify sufficient details of the approaches they will consider using, to give a clear idea of how the comparison group will be constructed. Evaluation designs should specify which matching or weighting method evaluators will use or lay out decision rules according to which evaluators will select a specific method. In addition, the evaluation design should describe which variables will be included in the matching process, which variables will be used to assess balance after matching or weighting methods have been applied, and which variables will enter the outcome analysis as controls, including those not included in the matching process. The evaluation design should provide statistical power calculations for the analysis and should also specify the level of matching (beneficiary, provider, or other level), the matching ratio, and whether trimming or calipers will be used. Finally, the design plan should describe how the results from the matching approach will be incorporated into regression models or other parts of the outcome analysis plan.

When completing interim and summative evaluation reports, evaluators should describe which matching method they used. If they deviated from the plans described in the evaluation design, they should discuss the reasons for this deviation. Evaluators should also list the covariates used for matching and provide evidence for the match quality, using the tools described in Section III.E. A table listing covariates, along with standardized differences before and after matching and any diagnostic plots, can appear in an appendix to the evaluation report. Finally, evaluators should clearly indicate whether demonstration impacts were estimated on the basis of a matched sample or using outcome regressions that incorporate matching weights. Importantly, evaluators should discuss any limitations that arise from using matching

methods, such as the inability to control for unobserved confounders, which may lead to biased impact estimates.

D. Concluding remarks

Matching is often an important step in identifying a comparison group, and important for rigorous evaluations of section 1115 demonstration. An ideal comparison group has characteristics that are identical on average to the beneficiaries in the demonstration group, but in practice, these groups often differ for at least some characteristics. If these differences are large, adjusting for them by including covariates in outcome regression models is often insufficient. Matching can greatly reduce those observable differences. In general, combining matching or weighting with regression adjustment leads to more reliable, doubly robust estimates of demonstration impacts that can better inform future policy decisions. This white paper should act as a starting point to help states and their independent evaluators to select appropriate matching methods for their evaluations of section 1115 demonstrations.

References

Austin, P. C. "Balance Diagnostics for Comparing the Distribution of Baseline Covariates Between Treatment Groups in Propensity-Score Matched Samples." *Statistics in Medicine,* vol. 28, no. 25, November 2009, pp. 3083–3107. doi: 10.1002/sim.3697

Austin, P. C. "Optimal Caliper Widths for Propensity-Score Matching when Estimating Differences in Means and Differences in Proportions in Observational Studies." *Pharmaceutical Statistics,* vol. 10, no. 2, March 2011, pp. 150–161. doi: 10.1002/pst.433

Austin, P. C., and E. A. Stuart. "Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies." *Statistics in Medicine,* vol. 34, no. 28, December 2015, pp. 3661–3679. doi: 10.1002/sim.6607

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

Farid, M., W. Zhu, and A. Hill. "Regression Discontinuity Designs in the Evaluation of Section 1115 Demonstrations." Washington, DC: Mathematica, February 2023.

Hainmueller, J. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis,* vol. 20, no. 1, Winter 2012, pp. 25–46. doi: 10.1093/pan/mpr025

Hansen, B. B. "Optmatch: Flexible, optimal matching for observational studies." *New Functions for Multivariate Analysis,* vol. 7, no. 2, October 2007, pp. 18–24.

Hirano, K., and G. W. Imbens. "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services and Outcomes Research Methodology,* vol. 2, no. 3, December 2001, pp. 259–278. doi: 10.1023/A:1020371312283

Iacus, S. M., G. King, and G. Porro. "Causal Inference Without Balance Checking: Coarsened Exact Matching." *Political Analysis,* vol. 20, no. 1, Winter 2012, pp. 1–24. doi: 10.1093/pan/mpr013

Imai, K., and M. Ratkovic. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 76, no. 1, January 2014, pp. 243–63. doi: 10.1111/rssb.12027

Imbens, G. W. "Matching Methods in Practice: Three Examples." *Journal of Human Resources,* vol. 50, no. 2, March 2015, pp. 373–419. doi: 10.3368/jhr.50.2.373

Imbens, G. W. and J. M. Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature,* vol. 47, no. 1, March 2009, pp. 5–86. doi: 10.1257/jel.47.1.5

Institute of Education Sciences. "WWC Procedures and Standards Handbook (Version 3.0)." Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse, 2014.

Kranker, K., L. Blue, and L. Vollmer Forrow. "Improving Effect Estimates by Limiting the Variability in Inverse Propensity Score Weights." *The American Statistician,* vol. 75, no. 3, July 2021, pp. 276–87. doi: 10.1080/00031305.2020.1737229

Lenis, D., T. Q. Nguyen, N. Dong, and E. A. Stuart. "It's All About Balance: Propensity Score Matching in the Context of Complex Survey Data." *Biostatistics,* vol. 20, no. 1, January 2019, pp. 147–163. doi: 10.1093/biostatistics/kxx063

Lunceford, J. K., and M. Davidian. "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study." *Statistics in Medicine,* vol. 23, no. 19, October 2004, pp. 2937–2960. doi: 10.1002/sim.1903

Pimentel, S. D., L. V. Forrow, J. Gellar, and J. Li. "Optimal Matching Approaches in Health Policy Evaluations Under Rolling Enrolment." *Journal of the Royal Statistical Society: Series A (Statistics in Society),* vol. 183, no. 4, October 2020, pp. 1411–1435. doi: 10.1111/rssa.12521

Pimentel, S. D., F. Yoon, and L. Keele. "Variable-Ratio Matching with Fine Balance in a Study of the Peer Health Exchange." *Statistics in Medicine*, vol. 34, no. 30, December 2015, pp. 4070–82. doi: 10.1002/sim.6593.

Pohl, R. V., and K. Bradley. "Selection of Out-of-State Comparison Groups and the Synthetic Control Method." Washington, DC: Mathematica, October 2020.

Robins, J. M., M. A. Hernan, and B. Brumback. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology,* vol. 11, no. 5, September 2000, pp. 550–560.

Rosenbaum, P. R. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association,* vol. 84, no. 408, December 1989, pp. 1024–32. doi: 10.1080/01621459.1989.10478868.

Rosenbaum, P. R., and D. B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika,* vol. 70, no. 1, April 1983, pp. 41–55. doi: 10.1093/biomet/70.1.41

Rosenbaum, P. R., and D. B. Rubin. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association,* vol. 79, 1984, pp. 516–524. doi: 10.1080/01621459.1984.10478078

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

Wang, Y., and J. R. Zubizarreta. "Minimal Dispersion Approximately Balancing Weights: Asymptotic Properties and Practical Considerations." *Biometrika,* vol. 107, no. 1, March 2020, pp. 93–105. doi: 10.1093/biomet/asz050

Zubizarreta, J. R. "Stable Weights That Balance Covariates for Estimation with Incomplete Outcome Data." *Journal of the American Statistical Association,* vol. 110, no. 511, July 2015, pp. 910–22. doi: 10.1080/01621459.2015.1023805

Appendix: Examples for Stata and R Commands for Implementing Weighting and Matching Methods

Table A.1. Examples of Stata and R commands

Method	Stata	R
Propensity score weighting	The suite of commands `teffects` is built into Stata.[a] The subcommands `ipw`, `aipw`, and `ipwra` estimate impacts using propensity score weighting	The package `PSweight` estimates propensity score weights.[b] To install the package from CRAN: `install.packages("PSweight")`
	Alternatively, evaluators can manually calculate inverse propensity score weights by (1) estimating a logit model for the treatment status conditional on covariates, (2) predicting the propensity score (probability of being treated) for each beneficiary, and (3) calculating the inverse propensity score weights. Then, evaluators would estimate weighted outcome regressions using the weight from step (3).	
Entropy balancing[c]	User-written command `ebalance`. To install: `ssc install ebalance`	The package `ebal` can be installed from CRAN via `install.packages("ebal")`
Propensity score matching	Subcommand `psmatch` in the `teffects` suite	The package `MatchIt` can be installed from CRAN via `install.packages("matchit")` and has several options for propensity score matching[d]
Coarsened exact matching[e]	User-written command `cem`. To install: `ssc install cem`	The package `cem` can be installed from CRAN via `install.packages("cem")`
Optimal matching		The package `optmatch` can be installed from CRAN via `install.packages("optmatch")` and has options for using the propensity score[f]

[a] A detailed description is available at https://www.stata.com/manuals16/te.pdf.

[b] A detailed description is available at https://cran.r-project.org/web/packages/PSweight/PSweight.pdf.

[c] See https://web.stanford.edu/~jhain/ebalancepage.html for details.

[d] A detailed description is available at https://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf.

[e] See https://gking.harvard.edu/cem for details.

[f] A detailed description is available at https://cran.r-project.org/web/packages/optmatch/optmatch.pdf. Also see Hansen (2007).

CRAN = Comprehensive R Archive Network.

www.mathematica.org

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

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

Centers for Medicare & Medicaid Services
7500 Security Boulevard Baltimore, MD 21244