UNIVERSITY OF CALIFORNIA

Los Angeles

Causal Inference with Longitudinal Data:
Moving Beyond Difference-in-Difference

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Health Policy and Management

by

Landon Manzano Gibson

2020

ABSTRACT OF THE DISSERTATION

Causal Inference with Longitudinal Data:
Moving Beyond Difference-in-Difference

by

Landon Manzano Gibson
Doctor of Philosophy in Health Policy and Management
University of California, Los Angeles, 2020
Professor Frederick Zimmerman, Chair

Difference-in-Difference is a widely used method in health policy and health services research for estimating a causal effect. Unfortunately, the validity of difference-in-difference is difficult to evaluate without a tool to directly assess the parallel trends assumption. For example, existing tools indirectly examine the parallel trends assumption using pre-treatment observations. Developments in the methodological literature have given rise to an alternative class of estimators – Synthetic Controls – that do not make the parallel trends assumption and to sensitivity analysis tools that provide a novel approach for directly evaluating the parallel trends assumption

The first chapter of this dissertation develops guidelines for the use of synthetic control methods alongside difference-in-difference. Synthetic control methods are a valuable tool because they don't assume parallel trends; however, they are not without assumptions of their own. This chapter provides guidance for the utilization of synthetic controls and difference-in-difference, and proposes several post-estimation validity analyses to further evaluate the assumptions made by each method.

The second chapter examines the effect of Medicaid Expansion on State Medicaid spending. The analysis is done using a subset of states among which the parallel trends assumptions is tenuous. Using a kernel-balanced synthetic control, and the post-estimation analyses introduced in the first chapter, this paper shows no evidence for Medicaid Expansion increasing or decreasing State Medicaid spending over a three-year period.

The third chapter extends a suite of sensitivity tools for estimating the sensitivity of difference-in-difference to unobserved time-varying confounders – parallel trends violations. The tools utilize the explanatory power of observed covariates to estimate how strong unobserved confounders must be to change the conclusions. They not only relax the strict binary nature of classic indirect parallel trends tests, but also utilize the post-period outcome data to directly examine the parallel trends assumption.

The dissertation of Landon Manzano Gibson is approved.

Warren S. Comulada

Chad Hazlett

Jack Needleman

Frederick Zimmerman, Committee Chair

University of California, Los Angeles

2020

Dedicated to Anita Araiza Manzano . . .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The past four years have been an incredible journey for which I owe a great amount of gratitude to friends, family, and colleagues. Here I would like to acknowledge some of those who were apart of this journey.

I would like to thank Professor Frederick Zimmerman, my advisor and the chair of my dissertation committee. Fred has been an excellent mentor who is caring, thoughtful, and supportive. I entered UCLA as a masters student and quickly began to wonder if I was capable of pursing a Ph.D., but I wasn't sure if I would be able to succeed. Fred encouraged me to pursue a Ph.D., and after I was accepted he continued to support my intellectual journey through different departments and topics that lead me to this thesis. It has been an absolute pleasure working with Fred.

I would also like to thank my fellow students in my MPH cohort because they made life in a new place very enjoyable. Thank you as well to my Ph.D. cohort - Kate McBride, Eryn Block, Lucia Felix, and Pragya Bhuwania - who provided a space where I could test new ideas, explore new topics, and practice sharing my knowledge. I'm very glad to have shared this unique experience with them.

Thank you to my parents, Jess and Beverly, who have always pushed me to achieve and do more. I would likely still be riding bikes in Marin had they not insisted I attend graduate school. Thank you to my colleagues at PPIC, especially Laurel Beck, who took the time to teach me fundamental research skills that would be critical to this journey. Finally, thank you to Hannah who had to tolerate my absent mindedness, the ever hazy finish line, and the working on weekends. Hannah, not only have you been incredibly supportive and encouraging but you also pushed me to grow as a human being these past four years for which I will be ever the better for. Thank you, Hannah.

VITA

2010–2013 Undergraduate Research Assistant, UC Irvine School of Medicine, University of California, Irvine.

2013 B.S. (Biological Science), University of California, Irvine.

2013–2014 Quality Assurance Data Analyst, Foothill Community Clinics.

2014–2016 Research Associate, Public Policy Institute of California.

2016–2020 Graduate Student Researcher, UCLA Center for Health Advancement, University of California, Los Angeles.

PUBLICATIONS

Health Training Pathways at California's Community Colleges (with Sarah Bohn and Shannon McConville), December 2016

Career Technical Education in Health at California's Community Colleges (with Sarah Bohn and Shannon McConville), December 2016

Anticipating Changes in Regional Demand for Nursing Homes (with Laurel Beck), November 2016

Modeling the Impact of Early Childhood Interventions on Health Equity for San Diego County (with Natalie Rhoads, Nathaniel W. Anderson, and Frederick J. Zimmerman), October 2019

CHAPTER 1

Difference-in-Difference and Synthetic Control Methods for Causal Inference: Implementation and Evaluation

1.1 Introduction

An important aspect of health policy research is the evaluation of the effect of a program, policy, or project on an outcome of interest. The gold standard for evaluation of programs and policies is the randomized control trial (RCT); however, an RCT is often infeasible due to either ethical or practical limitations. When a program, policy, or project — hereafter referred to as a treatment — is non-random, researchers utilize a variety of techniques, depending on the structure of the data and context of the treatment, to estimate the causal effect of treatment on an outcome of interest. In this paper we will focus on a specific data structure where: treatment occurs uniformly in time across one or many treated units and there is at least one observation of the outcome of interest in both the pre- and post-treatment periods. The data structure described gives rise to panel data, or time-series cross-sectional data, where the treatment occurs at an aggregate level such as the county or state and data on each state or county is observed for two or more time periods.

Motivated by recent developments in the methodological literature and the widespread utilization of panel data estimators in the health policy literature, we review three methods: difference-in-difference, mean-balanced synthetic control, and kernel-balanced synthetic control. Difference-in-Difference is widely utilized within health policy to estimate causal effects in longitudinal data structures (Ryan et al., 2015). Mean-balanced synthetic control – developed by Abadie et al. (2010) – is a quasi-experimental method with complimentary properties to difference-in-difference, but has seen limited adoption in the health policy literature (Bouttell et al., 2018). Kernel-balanced synthetic control, is a recent extension of the mean-balanced synthetic control developed by Hazlett and Xu (2018a). This paper 1) provides pre-estimation guidelines for when each method is most appropriate, 2) proposes strategies for deciding between methods when multiple methods may be appropriate, and 3) provides post-estimation guidelines for evaluating the validity of mean and kernel-balanced synthetic control models.

This paper is structured as follows. Section 1.2 provides notation, introduces the underlying causal framework, and presents the methods. Section 1.3 presents a decision tree to aid researchers in implementing and evaluating difference-in-difference alongside the synthetic control methods. Section 1.4 proposes strategies that assist with deciding between methods when multiple methods are appropriate. Section 1.5 details guidelines for evaluating the validity of results from each method. Section 1.6 and 1.7 discuss future research directions and concludes.

1.2 Background

1.2.1 Notation

Throughout this paper i represents units and t represents time periods. In typical policy studies, i often represents states or counties and t represents years or months. Y_{it} is the outcome for unit (state, county, city, country) i in time t, D_{it} is a binary time varying treatment indicator that is zero for all units in the pre-treatment period and zero for only the control units in the post-treatment period, X_{it} is a matrix of time varying covariates, and τ is the treatment effect.

This paper also utilizes potential outcomes notation, a causal framework which motivates many methods, including those discussed in this paper. Under the potential outcomes framework, every unit has two potential outcomes – an outcome with treatment and an outcome without treatment. We represent the potential outcomes of unit i in time t as $\{Y_{it}^1, Y_{it}^0\}$ where Y_{it}^1 is the outcome under treatment, such as California expanding Medicaid in 2014, and Y_{it}^0 is the outcome under the control scenario, California doesn't expand Medicaid in 2014. Once treatment occurs only one of the two potential outcomes is observed, which means we can observe Y_{it}^1 for California – because it expanded Medicaid in 2014 – but we are unable to observe Y_{it}^0 for California.

Hypothetically, the treatment effect, τ, for unit i is estimated as the difference in the potential outcomes, $\tau = Y_{it}^1 - Y_{it}^0$, but only one potential outcome is ever observed. Researchers use methods such as randomized control trials to impute the unobserved potential outcome for unit i by exchanging the unobserved counterfactual for the observed control outcome. Quasi-experimental methods such as difference-in-difference and synthetic controls also exchange the counterfactual for an observed outcome but make additional assumptions in order to do so.

1.2.2 Difference-in-Difference

Difference-in-Difference estimates the treatment effect, τ, by taking the difference in the outcome pre- and post-treatment by treatment status and a second difference between the treated and control units. For two time-periods or more, difference-in-difference is estimated as:

$$Y_{it} = b_0 + \tau D_{it} + b_1 \lambda_i + b_2 Z_t + \epsilon_{it} \tag{1.1}$$

where λ_i are unit fixed effects and Z_t are time fixed effects.

Note that Equation 1.1 does not include covariates, other than treatment status, because difference-in-difference is structured such that time-invariant covariates (sex, race, education) do not bias the estimates as long as their effect on the outcome is constant over time *or* the means are balanced between the treated and control units. Time-varying covariates, however, require that their effect on the outcome is constant over time *and* the means are balanced between the treated and control groups (Zeldow and Hatfield, 2019). When one of the aforementioned conditions is not met, for time-varying covariates, then the covariate is a confounder and can bias the results. Differences-in-difference assumes that the conditions for time-varying covariates are met, which is commonly known as the *Parallel Trends* assumption.

Put another way, parallel trends assumes that the evolution of the control group's outcome – over the post-treatment period(s) – is the same as the treated group's *unobserved* post-treatment outcomes under the control scenario, or counterfactual. For example, if state per-capita income is an unobserved time-varying confounder, parallel trends says that the *change* over time in the post-treatment period is the same for both the treated counterfactual and the control units. The assumption of parallel trends is a crucial assumption for the validity of causal inferences from difference-in-difference models. Yet, the published literature includes many examples of inadequate – or even non-existent – evaluation of this assumption. Section 1.4 below offers suggestions for appropriate evaluation of the parallel trends assumption.

1.2.3 Mean-Balanced Synthetic Control

An alternative method to difference-in-difference is the mean-balanced synthetic control method developed by Abadie et al. (2010). Synthetic control methods are motivated by the idea that a weighted combination of multiple control units can serve as a better comparison group than any individual control unit. For example, under the potential outcomes framework the researcher's goal is to estimate the counterfactual, Y_{it}^0, for the treated unit for which only Y_{it}^1 is observed. Difference-in-Difference estimates Y_{it}^0 using one or many control units without adjustment. The mean-balanced synthetic control method instead imputes the unobserved counterfactual of the non-treatment outcome for the treated group, Y_{it}^0, as a weighted average of observed control units where the weights are estimated to minimize the difference between the synthetic control and the treated unit's outcome in the pre-treatment period. Notably, synthetic control methods do not make the parallel trends assumption in order to estimate an unbiased effect.

The mean-balanced synthetic control does, however, assume that the potential outcomes are independent of the treatment assignment conditional on a linear combination of past outcomes - known as the linearity in prior outcomes (LPO) assumption. That is, by balancing on prior outcomes the treatment is no longer related to the potential outcomes and the synthetic control can be used as a substitute for the unobserved treatment counterfactual. The LPO assumption fails if the potential outcomes are a linear function of more than just the first moment, or their interactions, of the outcome data.

A visual comparison is provided below to help instill intuition. In Figure 1.1 the solid black line describes the treated unit's outcome over time, the dashed black line describes the control group's simple average outcome over time, the grey lines are the individual control units, and the red line is the weighted combination of control units, or Synthetic Control. Note, treatment begins after 2013 but due to there being one observation per year it seems as though treatment begins having an effect between 2013 and 2014.

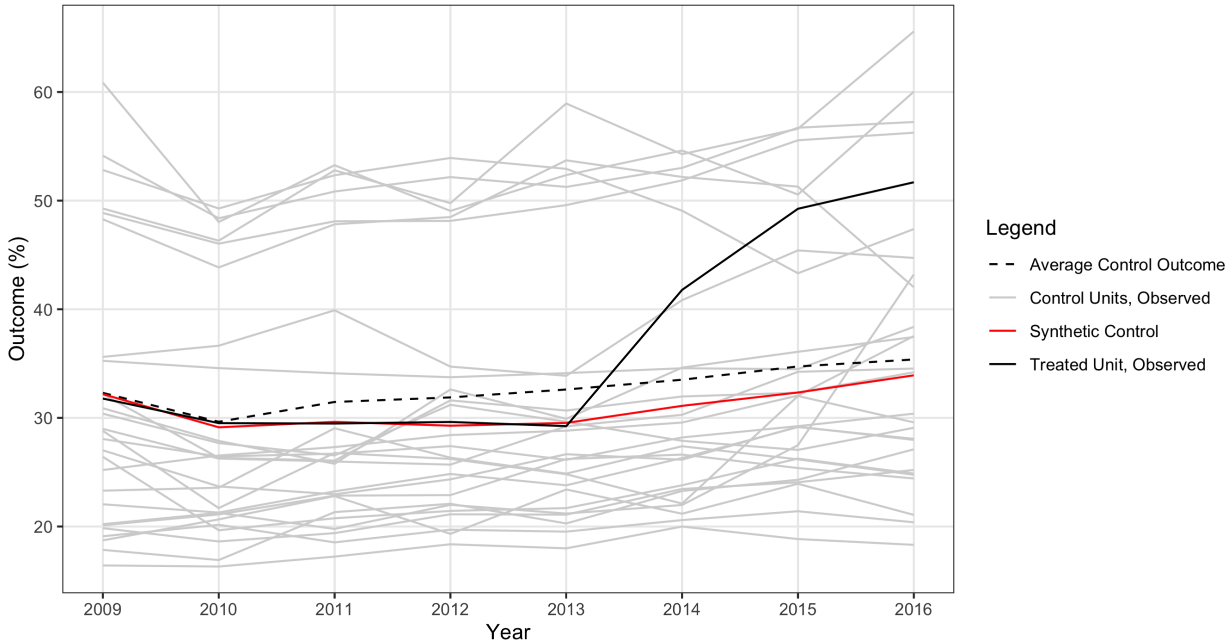


Figure 1.1: Synthetic Control and Observed Control

The control unit created by the Synthetic Control method, the red line, is nearly identical to the treated unit through the last pre-treatment period in 2013. In order to impute the counterfactual, the weights are applied to the control units in the post-treatment period. The treatment effect is then estimated by taking the difference between the treated unit's outcome and the synthetic control outcome in each post-treatment period – 2014 through 2017 – and averaging them to obtain the ATT for the treatment period.

The Synthetic Control method can be implemented in Stata, R, or MatLab using the *Synth* package, which was developed by Abadie et al. (2011).

1.2.4 Kernel-Balanced Synthetic Control

The kernel-balanced synthetic control is an extension of the mean-balanced synthetic control that modifies the balancing procedure used to estimate the weights (Hazlett and Xu, 2018a).

Conveniently, the intuition is the same as the mean-balanced synthetic control, but the key difference in the balancing procedure. The kernel-balanced synthetic control uses a procedure that tries to balance on more than just the mean structure such as the variance of the kurtosis of the outcome data.

As discussed in the previous section, the objective of synthetic control methods is to find a weighted combination of observed control units to generate a counterfactual that is nearly identical to the treated unit's *mean* outcome over time, in the pre-treatment period (Figure 1.1). Kernel-balanced synthetic control has a similar objective, but it approximately balances on not just the mean structure but also *higher moments* of the data in order to minimize the distance between the treated and control units, on those moments, given some constraints. By balancing on higher order moments of the data, kernel-balancing relaxes the LPO assumption because it models a more flexible linear combination of the prior outcomes (Hazlett and Xu, 2018b).

The kernel-balanced synthetic control relaxes the assumptions of the mean-balanced synthetic control because it is a more flexible model, but that flexibility comes at a cost. The increased flexibility means that the kernel-balanced synthetic control is less efficient than the mean-balanced synthetic control and, in some cases, may not have a feasible combination of weights that improve balance. The feasibility limits occur because the kernel-balanced synthetic control is asking much more from the available data by matching on higher order moments, which when combined with the weight constraints, can result in a synthetic control that marginally improves balance beyond the unweighted control units.

When weights that improve balance are infeasible, the kernel-balanced synthetic control can be modified using a demeaning procedure. The demeaning procedure first subtracts the average pre-treatment outcome, by treatment status, from the treated and control groups pre-period data and then estimates the kernel-balanced synthetic control. The benefit of first demeaning the data is that it reduces the chance that weights cannot be found due to feasibility constraints by making the treated and control outcomes more similar. The downside, however, is that the demeaning procedure induces a parallel trends assumption (Hazlett and Xu, 2018a). While both the demeaned kernel-balanced synthetic control and difference-in-difference assume parallel trends they differ in that the demeaned kernel-balanced synthetic control still tries to balance time-varying confounders by balancing pre-treatment outcomes. Difference-in-Difference, however, makes a stronger parallel trends assumption that both observed and unobserved time-varying covariates are not confounders.

The kernel-balanced synthetic control method can be implemented in R using the Trajectory Balancing package, *Tjbal*[1].

1.3 Decision Tree for Implementation and Evaluation

This section provides a decision tree to assist researchers in deciding between difference-in-difference and synthetic control methods, and in determining whether a mean-balanced or kernel-balanced synthetic control is most suited to their data. The decision tree begins by considering the number of pre-periods and asks the researcher to subset their data to include only the pre-period; the purpose of which is to reduce the chance a researcher simply selects a method based on the ATT. Second, it contains an estimation step that is followed by a set of specification analyses. If the specification analyses are positive, the tree reaches a terminal node but if they are negative the tree moves onto the method that makes more restrictive assumptions. The decision tree repeats until either the specification analyses return results in support of the model or it reaches difference-in-difference. In the decision tree, we shorthand the names of the methods to DiD for Difference-in-Difference, MB for mean-balanced synthetic control, and KB for kernel-balanced synthetic control.

The number of pre-treatment periods is important because the unbiasedness of both synthetic control methods is related to the number of pre-treatment periods. Abadie et al.

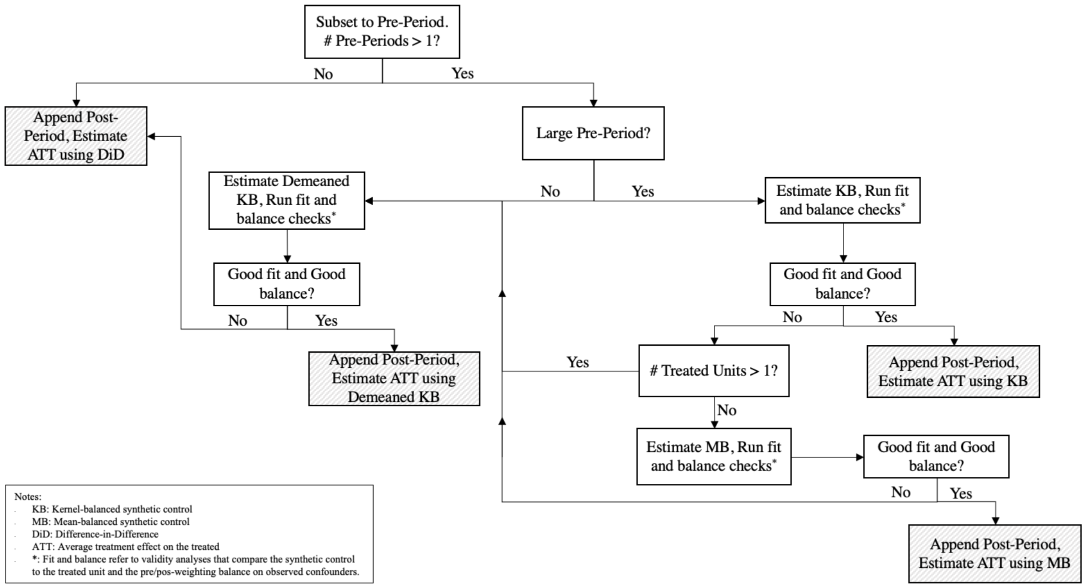
[1] For more information on how to install and use the package please visit the maintainer's website: http://yiqingxu.org/software.html

(2010) suggest that researchers utilize the mean-balanced synthetic control when the number of pre-treatment periods is large, but without further defining large. Work by Ferman (2020) reiterates the necessity of having a large number of pre-treatment periods and add that the ratio of the number of pre-periods to control units is important; however, he also avoids committing to a specific cut-off. For example, if there are few control units than more pre-treatment periods are required compared to when there are a wealth of control units. Keeping with previous work, we also refrain from suggesting specific pre-period thresholds for when to utilize synthetic control methods. Ultimately, specification analyses for both the fit of the synthetic control and pre/post balance on confounders will be critical for deciding whether or not there are enough pre-periods, which is reflected in the decision tree.

The specification analyses are a critical step in the decision tree because the results help to inform the researcher if they should estimate the ATT or utilize a different method that makes more stringent assumptions. The decision tree emphasizes two specification analyses: the fit of the synthetic control to the pre-treatment outcome for the average treated unit and the pre/post balance on confounders. The purpose of the specification analyses is to inspect a critical assumption that by finding weights that balance the pre-treatment outcome data also balance confounders (observed and unobserved) (Abadie et al., 2010).

Comparing the fit of the synthetic control with the average treated outcome is the first part of the aforementioned assumption. While exact balance is not often achieved in real world data the specification analysis still looks for a good match that is preferably nearly exact. In addition, checking pre/post-weighting balance on observed confounders is informative for evaluating if the weights balance observed confounders, and by extension, unobserved confounders. Note, this is an opportunity for the researcher to incorporate their contextual knowledge by making an argument that the included confounders are important. Later, we will introduce two validity analyses that can be implemented after a method is settled upon. The reason for not using the validity analyses until after the ATT is estimated is because there is a risk of overfitting to the pre-period data - discussed later in Section 1.5.2.

If the specification analyses return negative results, there is a decision point that evaluates the number of treated units. The number of treated units matters because the mean-balanced synthetic control is designed for data structures with only one treated unit; whereas, the kernel-balanced synthetic control and difference-in-difference can easily accommodate one or many treated units. There exist modifications of the mean-balanced synthetic control that allow for the use of multiple treated units, but there are technical issues that complicate the procedure by which the weights are estimated (Alberto Abadie and Jérémy L'Hour, 2019). One simple workaround is to estimate an individual synthetic control for each treated unit and then aggregate the effect estimates but it is not clear how to estimate uncertainty. These extensions are not the focus of this paper and we refer readers to a more detailed review in Abadie (2019).


Subset to Pre-Period.
# Pre-Periods > 1?
No
Yes
Append Post-Period, Estimate ATT using DiD
Large Pre-Period?
No
Yes
Estimate Demeaned KB, Run fit and balance checks*
Estimate KB, Run fit and balance checks*
Good fit and Good balance?
No
Yes
Good fit and Good balance?
No
Yes
Append Post-Period, Estimate ATT using Demeaned KB
Yes
# Treated Units > 1?
Append Post-Period, Estimate ATT using KB
No
Estimate MB, Run fit and balance checks*
Good fit and Good balance?
No
Yes
Append Post-Period, Estimate ATT using MB

Notes:
- KB: Kernel-balanced synthetic control
- MB: Mean-balanced synthetic control
- DiD: Difference-in-Difference
- ATT: Average treatment effect on the treated
- *: Fit and balance refer to validity analyses that compare the synthetic control to the treated unit and the pre/pos-weighting balance on observed confounders.

Figure 1.2: Estimation Decision Tree for Difference-in-Difference (DiD), Mean-balanced (MB) and Kernel-balanced (KB) Synthetic Control Methods.

When the number of pre-treatment periods is large then a kernel-balanced synthetic control is the recommended starting point. The primary reason is that the kernel-balanced synthetic control makes fewer assumptions than the mean-balanced synthetic control (Section 1.2). The specification analyses then follow to assert how well the synthetic control fits the outcome data and whether or not balance was improved for observed confounders. When the specification analyses return positive results - discussed further in Section 1.5.2 - the kernel-balanced synthetic control can be used to estimate the ATT.

When the specification checks for the kernel-balanced synthetic control return poor results, because the fit was poor and balance was not improved on observed confounders, then a mean-balanced synthetic control is recommended, if there is only one treated unit. If the specification analyses are positive for the mean-balanced synthetic control then it can be used to estimate the ATT. Alternatively, if there are multiple treated units then it is recommended that the demeaned kernel-balanced synthetic control is estimated instead of a mean-balanced synthetic control.

When the three models above have been exhausted - mean, kernel, and demeaned kernel-balanced synthetic controls - then difference-in-difference is the recommended model to implement. Difference-in-difference is the last model to be recommended because it makes the strongest assumptions out of all the methods on the decision tree. For example, while the demeaned kernel-balanced synthetic control does assume parallel trends, it also attempts to obtain balance on time-varying confounders - observed and unobserved - by constructing a weighted combination of control units to be similar to the treated units. Difference-in-Difference, however, simply assumes parallel trends without adjusting imbalances on time-varying confounders further; unless they are included as covariates in the model. Therefore, we strongly recommend the use of sensitivity analysis tools to evaluate the sensitivity of the estimates to hypothetical unobserved time-varying confounders, which can be implemented using the *sensemakR* package in R or Stata for more information see Cinelli and Hazlett (2020).

1.4 Post-Estimation Analyses to Decide Between Methods

The purpose of this section is to discuss ways to evaluate the assumptions of each method, post-estimation, and how to use that information to identify possible sources of bias for each of the methods.

Difference-in-Difference vs. Synthetic Control Methods

A key distinction between difference-in-difference and both mean-balanced and kernel-balanced synthetic controls is the parallel trends assumption. If the parallel trends assumption holds, difference-in-difference is an unbiased estimator of the counterfactual, but when the parallel trends assumption doesn't hold, difference-in-difference is biased, and either the mean or kernel-balanced synthetic controls should be used, assuming the data structure is appropriate.

The conclusions drawn from a biased difference-in-difference estimator, however, may or may not change depending on the sensitivity of the results to unobserved confounding with strength similar to observed confounders. For example, if the treatment effect estimated by difference-in-difference is large and positive the researcher would conclude that the treatment had a positive impact. One could then imagine an unobserved confounder, or group of unobserved confounders, that have a non-zero relationship with the outcome, which would bias the treatment effect estimate. However, if the hypothetical unobserved confounders have a weak relationship with the outcome, the conclusion that the impact of treatment was positive is unlikely to change despite the point estimate changing. Considering the sensitivity of the conclusions on a continuum based on hypothetical confounding is known as sensitivity analysis.

We propose the use of sensitivity analysis to assert how strong a violation of the parallel trends assumption is required to change the conclusions drawn from difference-in-difference. Sensitivity analysis considers hypothetical unobserved confounders one or many times stronger than an observed confounder, which was argued to be the most important confounder by the researcher, and calculates how the hypothetical confounder changes the results. For example, if the conclusions of the study only change when an unobserved confounder is ten times stronger than the observed confounder, argued as being the most important confounder by the researcher, the results are likely not very sensitive to unobserved confounding because it would take an unrealistically strong hypothetical confounder to change the conclusions. However, when the results are sensitive such that a plausibly strong unobserved confounder can change the conclusions of the study, either synthetic control method is a more appropriate estimator because they do not assume parallel trends.

Mean vs. Kernel-balanced Synthetic Control

A key difference between the mean and kernel-balanced synthetic control is the Linearity in Prior Outcomes (LPO) assumption. The mean-balanced synthetic control strictly assumes that the treated unit's counterfactual outcome is a linear function of the prior outcomes, whereas the kernel-balanced synthetic control relaxes the LPO assumption and thereby makes fewer and less restrictive assumptions compared to the mean-balanced synthetic control (Hazlett and Xu, 2018a). Unfortunately, it is not possible to examine empirically whether or not the LPO assumption holds.

An alternative way to investigate which method is most appropriate is to determine which model best fits the pre-treatment outcomes. A problem with this approach is that the kernel-balanced synthetic control is balancing on multiple moments of the outcome – eg. the mean, variance, kurtosis, etc – (Hazlett, 2020) while the mean-balanced synthetic control is balancing on only the first moment of the outcome. As a result, The kernel-balanced synthetic control may be less biased than the mean-balanced synthetic control but obtain worse balance on the outcome mean because the mean is not the only element in the objective function that it minimizes.

A complementary way to compare the two synthetic control methods is to estimate the amount by which balance – by treatment status – improves on observed covariates pre vs. post-weighting. Balance on covariates is crucial because synthetic control methods are unbiased if the weights that balance the outcome also balance observed and unobserved covariates (Abadie et al., 2010). Implementation of this analysis requires that covariates are not included in the balancing procedure, otherwise weights will be found that balance the covariates regardless. Other literature has suggested including covariates in the balancing procedure, but recent work has pointed to issues with that technique that can lead to researcher induced bias. Therefore, exclusion of the covariates has the advantage of reducing researcher bias and providing an excellent analysis with which to evaluate the validity of the synthetic control. Furthermore, balance checks are an excellent tool because it incorporates the researcher's content expertise by relying on the researcher's ability to identify, justify, and assess the relative importance of potential confounders.

In summary, the decision for whether to implement a mean or kernel-balanced synthetic control relies on the researcher conducting post-estimation analyses – pre-treatment outcome balance and pre-treatment covariate balance – and making a decision based on the entirety of the results. The mean-balanced synthetic control is preferred over the kernel-balanced synthetic control if the kernel-balanced synthetic control fails to improve balance on covariates, whereas the mean-balanced synthetic control does. In the scenario that both methods perform equally well on the above analyses we recommend the utilization of the kernel-balanced synthetic control because it makes fewer and less restrictive assumptions compared to the mean-balanced synthetic control.

1.5 Evaluating the Validity of the Results

The purpose of this section is to detail an analysis plan for evaluating the validity of causal effects estimated from the methods of interest. The section is broken into two parts: a brief section reviewing diagnostic analyses for difference-in-difference (Section 1.5.1), and a section presenting an analysis plan for evaluation of causal effects from the synthetic control methods (Section 1.5.2).

1.5.1 Difference-in-Difference

Indirect Analyses of Parallel Trends

Parallel trends analyses are commonly utilized to evaluate whether or not the assumptions holds in the pre-treatment period; however, the parallel trends assumption is specific to the post-treatment period. Therefore, parallel trends analyses are not directly examining the parallel trends assumption because such analyses lack data on the relevant period. Therefore, we will further refer to parallel trends analyses as *indirect* evaluations of the parallel trends assumption. Related to this limitation of parallel trends analyses is that they are not viable when the number of pre-treatment periods is equal to one because they compare the trends, again, only between two pre-treatment periods.

Indirect evaluation of the parallel trends assumption is often done informally via visualization and formally with a hypothesis test. Note, these analyses are limited to using pre-treatment data and therefore do not directly examine the parallel trends assumption. Instead, they examine whether or not trends are parallel in the *pre-treatment* period. A graphical examination of the trends of the outcome by treatment status is a common way for researchers to begin examining common trends in the pre-treatment period. It is often easy for researchers to identify non-parallel pre-treatment trends. Another analysis for parallel pre-treatment trends is to estimate a modified difference-in-difference model that allows for group-specific linear time trends (Bilinski and Hatfield, 2018).

$$Y_{it} = b_0 + b_1 G_i + b_2 \lambda_i + b_3 Z_t + \sum_{t=1}^{T} b_t G_i Z_t + \epsilon_{it} \tag{1.2}$$

where λ_i and Z_t are unit and time fixed effects, respectively, T represents the total number of time periods, and G_i is a time-invariant treatment indicator.

The above model estimates a coefficient for each time period (b_t) that describes the difference in trend between the treated and control group. Each coefficient also has a p-value based on the null hypothesis that the trends between the treated and control groups are similar ($b_t = 0$)[2]. One must be careful when interpreting the results of this analysis for two reasons. First, a failure to reject the null hypothesis is not a confirmation that parallel trends holds because one can only find evidence against the null hypothesis, not accept the null hypothesis. This appears to be a common problem in the literature. Second, the hypothesis test is only examining parallel trends in the pre-treatment period but the parallel trends assumption involves post-treatment trends. Therefore, the analysis does not directly examine the parallel trends assumption unless one makes strong, untestable, assumptions as to how the trends develop in the post-treatment period.

Sensitivity Analysis

An alternative procedure considers how strong unobserved time-varying confounding would have to be to change the estimated effect. Sensitivity metrics – developed by Cinelli and Hazlett (2020) and extended in the third chapter of this dissertation – may be anchored to observed time-varying confounders to understand how strong unobserved confounders would need to be, relative to observed confounders, to change the conclusions of the study. This approach has a number of advantages over the hypothesis tests described above: 1) Sensitivity analysis is a direct investigation of the parallel trends assumption rather an indirect examination of parallel trends in the pre-treatment period and 2) Sensitivity analysis can be conducted even when there is only one pre-treatment period. A full presentation of these techniques is beyond the scope of this paper. Interested readers are referred to Cinelli and Hazlett (2020) and the third chapter of this dissertation for more information.

[2] The null hypothesis is reversed in this scenario, it should be structured such that the alternative hypothesis contains the quantity we wish to prove ($b_t = 0$). The consequence of this reversal is that the chance we fail to reject the null despite the null being false (incorrectly conclude that parallel trends hold) is not fixed at 5%. For more information please refer to (Bilinski and Hatfield, 2018)

1.5.2 Mean and Kernel-Balanced Synthetic Controls

Below we discuss the specification analyses introduced in the decision tree and introduce validity analyses of both mean and kernel-balanced synthetic controls. The purpose of these analyses is to examine the assumptions underlying both methods. Some of these methods were covered in Section 1.4 but are discussed here with more detail.

Pre-Treatment Outcome Balance: An important, and simple, specification analysis is to determine how well the counterfactual constructed by either synthetic control method fit the pre-treatment means and variance of the treated unit(s) outcome (Abadie et al., 2010; Abadie, 2019). For the mean-balanced synthetic control, this amounts to a visual inspection of the estimated counterfactual outcome trend against the treated units outcome trend, as in Figure 1. Alternatively, one can generate a balance table which reports the pre-treatment outcomes over time for the average treated unit, the unweighted average of the control units, and the synthetic control. A positive result would be for the difference between the synthetic control and treated unit(s) outcome to be near zero in all pre-treatment periods, or that the synthetic control is a near exact match to the average treated unit. For the mean-balanced synthetic control a good fit is a synthetic control that is as close to the average treated unit in the pre-treatment period as possible. Comparison of the kernel-balanced synthetic control to the average treated unit, however, is more complicated.

The kernel-balanced synthetic control balances the outcome across many moments of the data and therefore imbalances on the raw outcome data are not dispositive. One may inspect the fit of the estimated synthetic control to the mean structure and the variance, but keeping in mind that the kernel-balanced synthetic control may be balancing on higher order moments as well. Therefore, any large gaps between the synthetic control and the average treated unit could be because the kernel-balancing procedure is matching on higher order moments of the data or because it has difficulty finding weights to match. A mean-balanced synthetic control can be utilized to investigate whether or not there is a feasibility issue -

weights don't exist to obtain balance - because the mean-balanced synthetic control only tries to balance the means and therefore eliminates the possibility that visible imbalances are due to priority on higher order balances. For example, if the mean-balanced synthetic control fits near exact with the average treated unit then it can be concluded that feasibility is not of concern and instead kernel-balancing is prioritizing balance on higher order moments.

Confounder Balance Checks: A complementary specification analysis to evaluate pre-treatment outcome balance - for either the mean or kernel-balanced synthetic control - is to examine balance on what the researcher deems to be important confounders. Balance checks are crucial because both methods assume that the weights balance the pre-treatment outcomes, and observed confounders, then unobserved confounders are also likely to be balanced (Abadie, 2019). However, the confidence that balance checks give towards the methods is related to the importance of the confounders on which balance is checked. For example, obtaining balance on confounders that are only weakly related with the outcome is much less important than obtaining balancing on confounders with a strong relationship with the outcome and treatment. Therefore, it is critical that the researcher incorporates their content knowledge to argue that the confounders they include in the balance check are important. Balance checks on covariates can be implemented by simply comparing the difference in means between the treated and control units pre- and post-weighting. It is expected that the difference in means will be near zero post-weighting.

Validity Check - Placebo Treatment Period: Placebo treatment periods are useful for evaluating the ability of either synthetic control method to estimate an accurate counterfactual given the number of pre-periods available. The placebo treatment period analysis works by sub-setting the data to just the pre-treatment periods, making the final period(s) a false treatment window, estimating the synthetic control method of choice, and reporting the ATT for the false treatment window. The analysis assumes that there is no treatment effect in the pre-treatment periods, and as a result we would expect the ATT returned by this analysis to be zero for all pre-treatment periods. When the ATT is non-zero than we must question other events that occurred during the pre-treatment period. In the absence of alternative explanations than the validity of the synthetic control to estimate the counterfactual is of concern.

Robustness - Pre-Treatment Period Length: Robustness of the synthetic control can be explored by iteratively reducing the pre-treatment period by one period and estimating the ATT each time. For example, if the pre-treatment period consists of ten pre-periods the check is implemented by removing the period furthest from treatment, estimating the ATT, and repeating removing the next period furthest from treatment. A robust synthetic control would report ATTs similar to the main effect even as the pre-treatment period decreases; whereas a less robust synthetic control will have varying results with shorter pre-treatment periods. Further work is needed to investigate analyses for evaluating whether or not the difference in the ATTs for fewer pre-periods is similar or not to the main effect.

1.6 Discussion

Mean and kernel-balanced synthetic control methods are valuable additions to the applied researcher's toolkit that complement existing methods. Both synthetic control methods are complementary to difference-in-difference because they don't make the parallel trends assumption, and as a result, are very useful for making causal inferences in the presence of unobserved time-varying confounding. To be clear, synthetic controls are not an end-all solution to unobserved confounding and still require post-estimation analyses to evaluate the validity of the results as detailed in Section 1.4. Furthermore, alternative post-estimation analyses relating to the weights or that utilize other characteristics of the data, such as a placebo analysis that uses units that experience treatment at a later date, are also viable and we expect there to be significant creativity with creation of additional analyses.

The literature on synthetic control methods is rapidly expanding as researchers develop extensions and furthering the theoretical underpinnings. Extensions have been proposed that extend synthetic control methods to allow for multiple treated units and reduce bias when the pre-treatment fit is poor. Furthermore, there are a number of related methods such as the Generalized Synthetic Control and Matrix Completion (Xu, 2017; Athey et al., 2018). An overview of extensions and related methods is available in Abadie (2019).

The literature for difference-in-difference is also rapidly growing with extensions being developed for more specific and complex data structures. For example, difference-in-difference has been extended to cases with: small numbers of treated groups, variation in treatment timing, multiple time periods, or sharp and fuzzy treatments. There have also been advancements in quantifying the validity of difference-in-difference estimates. Concerns have been raised with the current indirect parallel trends analysis paradigm and as such has lead to the growth of sensitivity analysis as an alternative (Bilinski and Hatfield, 2018; Roth, 2018).

Future research for causal inference with panel data is an exciting area with an abundance of innovation in the areas of estimation and validity analysis. Future work should continue to develop post-estimation tests to determine which method -– mean or kernel-balanced synthetic control -– is more appropriate. For example, one may develop balance checks that compare densities rather than moments. Other developments may focus on determining how much imbalance is acceptable by utilizing sensitivity techniques based on the strength of relationship between a covariate, treatment, and the outcome.

This paper builds upon prior work by Samartsidis et al. (2019) in a number of ways. First, this paper develops a decision tree to aid applied researchers with implementation of difference-in-difference, mean-balanced synthetic control, and kernel-balanced synthetic control. Second, this paper highlights the utility of the kernel-balanced synthetic control because it relaxes the assumptions of the mean-balanced synthetic control. Third, this paper provides detailed guidelines for evaluating the validity of synthetic control methods and references a novel sensitivity tool for evaluating the validity of difference-in-difference. Fourth, we generalize our implementation guidelines to scenarios with one or many treated units and one or many pre-treatment periods.

1.7 Conclusion

Quasi-experimental designs such as difference-in-difference and synthetic controls are useful for learning causal relationships when treatment is non-randomly assigned. Because data in health policy is often collected overtime quasi-experimental designs that exploit the nature of panel data are valuable tools for an applied researcher. However, these tools are not end-all solutions to the problem of unobserved confounding by way of non-randomized treatment assignment. Care must be taken to determine both pre and post-estimation which design is appropriate and whether or not the assumptions underlying a given design hold.

CHAPTER 2

The Effect of Medicaid Expansion on State Medicaid Spending

2.1 Introduction

In 2014 the landmark Affordable Care Act (ACA) began nationwide implementation. Included in the ACA was an important provision to expand eligibility for Medicaid to all adults with incomes up to 138% of the federal poverty line (FPL). This was a significant expansion in two ways. First, Medicaid had previously been limited to children and their custodial parents (Herz et al., 2016). Second, Medicaid eligibility was contingent on very low-income thresholds in many states; with a median income limit for working parents of only 68% of the federal poverty line (Heberlein et al., 2012). Doubling this income limit would therefore bring large numbers of people into eligibility (Mitchell, 2018).

To address concerns of cost increases, the federal government committed to paying 100% of the costs for newly eligible individuals in the first two-years with an increasing state share in the subsequent years. Many policy-makers, especially Republicans, voiced concern about the costs of such an expansion on their state budgets (Alonso-Zaldivar, 2014). They argued that expanding Medicaid eligibility would increase state costs because it would raise the profile of Medicaid; and therefore, induce more previously eligible people to sign up than had previously done so. This effect of raising the uptake rate of those previously eligible was termed "the woodwork effect" (Frean et al., 2016; Kenney et al., 2016). Understanding the fiscal effects of Medicaid expansion is important due to concern about the financial impacts on states as Medicaid enrollment grows; the current study seeks to explore this concern.

Unlike the ACA, which was implemented nationally, the decision whether or not to expand Medicaid eligibility was to be made at the state-level, which lead to only some states adopting the new eligibility limits. Because the expansion of Medicaid to low-income adults below 138% of the FPL did not occur uniformly across all states, a natural experiment was established with some states selecting into treatment and others selecting out of treatment. Medicaid expansion has been studied extensively since its implementation, using the aforementioned natural experiment, with researchers looking at a number of outcomes such as: access to care, health and quality of care, cost of care, and hospital financial performance (Mazurenko et al., 2018). Researchers have also studied the impacts of Medicaid expansion on state budgets, with a number of detailed studies on the fiscal impacts of Medicaid Expansion at the national, multi-state, and individual state level. A meta-review of studies on the effects of the Medicaid expansion under the ACA found that expansion states experienced budget savings, revenue gains, and economic growth (Antonisse and Garfield, 2018). To our knowledge, however, there has only been one study that has looked at the effect of Medicaid Expansion on state budgets across multiple states by utilizing a quasi-experimental methodology. The study utilized a difference-in-difference design to analyze the effect of Medicaid expansion on total state spending and state spending on Medicaid using data on all states through FY 2015 (Sommers and Gruber, 2017).

This paper makes four additions to earlier work by adding new data, improving treatment and control definitions of states following Medicaid expansion, and making two important methodological improvements. As a first contribution, this study expands prior work to include an extended post-treatment period from fiscal year 2015 to fiscal year 2017, which allows for an investigation into whether or not Medicaid expansion had an impact on state spending on Medicaid that persisted over time. A longer time horizon is motivated by descriptive evidence that enrollment continued to increase in the two years following Medicaid expansion (Rudowitz et al., 2016). Therefore, this paper analyzes the hypothesis that the woodwork effect was larger in later years than in the first fiscal year following expansion.

The second contribution is to revisit and improve the treatment and control definitions that have been utilized by prior work (Sommers and Gruber, 2017; Garrett and Kaestner, 2015; Hu et al., 2016). The improvements are made by considering the Medicaid eligibility criteria of states as of December 2009. December 2009 was an important date because it dictates, under the ACA, which Medicaid expenses qualify for the special Medicaid match rate and which do not. For example, states that had already expanded Medicaid to childless adults – and as a result would have no newly eligible enrollees – would receive a special match rate that would become equivalent to the enhanced match rate in 2020. The present study utilizes variation in the Medicaid eligibility criteria as of December 2009 to create groupings of states and assign them to three categories: treated, control, and excluded.

The third and fourth contributions are methodological advancements. The first methodological advancement is related to the concern, not fully addressed in prior work, of a violation of the parallel-trends assumption due to uncontrolled time-varying confounders. When the assumption of parallel trends is not warranted, difference-in-difference estimation provides biased estimates of the causal effect of the policy change. To address this issue, this study utilizes a kernel-balanced synthetic control which is apart of the trajectory balancing framework (Hazlett and Xu, 2018a). Kernel-balanced synthetic control imputes the counterfactual by weighting the control units, much like the mean-balanced synthetic control method (Abadie et al., 2010). A kernel-balanced synthetic control has two main advantages over the mean-balanced synthetic control and difference-in-difference: 1) The kernel-balanced synthetic control does not assume parallel trends; and 2) it relaxes the linearity in prior outcomes (LPO) assumption made by the mean-balanced synthetic control. Therefore, the kernel-balanced synthetic control requires the fewest assumptions in order to be unbiased, compared to the mean-balanced synthetic control and difference-in-difference.

The second methodological contribution is the introduction and application of several validity analyses for the kernel-balanced synthetic control. In this paper we utilize four different validity checks: visualizing the fit of the synthetic control to the average treated outcome in the pre-treatment period, analyzing the assumption that balance is found on important confounders, investigating the credibility of the synthetic control by backdating the treatment timing into the pre-treatment period, and a robustness analyses examining the sensitivity of the results to the number of pre-periods.

2.2 Methods

2.2.1 Data

The primary data source for this analysis came from the National Association of State Budget Officers (NASBO) annual reports. The data provides state-level information related to total spending by category which is then broken down by source: general fund, other state funds, bonds, and federal funds. The data is available from State Fiscal Year (FY) 1991 through 2017; fiscal years begin July 1st and end June 30th, and they are denoted by the year in which the period ends. Furthermore, all years of data were not utilized because there were six missing data points: four in Nevada, one in Mississippi, and one in New Mexico. The missing data points for Nevada, New Mexico, and Mississippi, were also missing from the NASBO reports for the associated year. Therefore, it was likely that the aforementioned states did not submit data for those years. There was a missing data point for Wyoming in FY 1999 but the data was imputed using data in the online report for FY 1999.

2.2.2 Measurement

The primary outcome of interest was Medicaid spending by source. The sources available in the NASBO data are: federal funds, general funds, other funds, and bond funds. General funds, Other funds, and bond funds are spending sources for the state government, but only general and other Funds are related to Medicaid spending. General funds, according to

NASBO's definition, consist of revenues from broad state taxes, and are the primary financing mechanism for Medicaid by the state. Other funds are specific taxes that are restricted for specific activities or governmental functions. For Medicaid, Other funds include provider taxes, fees, donations, assessments, and local funds. To capture total state spending on Medicaid, we combine the spending from the general and other funds.

After constructing the state spending on Medicaid outcome as the combination of general and other state funds, the total state spending on Medicaid by the state population within each year - using National Institutes of Health Surveillance, Epidemiology, and End Results (SEER) data - was normalized to obtain per-capita spending. Therefore, the primary outcome was the per-capita state spending on Medicaid.

In addition to the primary outcome of interest there were three time-varying confounders in the sensitivity analysis: unemployment rate, state per-capita revenue, and the labor force participation rate. Unemployment rate data are annualized unemployment rates at the state level reported by the Bureau for Labor and Statistics. The data is available for all years used in the analysis (FY 1998 – FY 2017). State per-capita revenue was generated by combining revenue data from NASBO with SEER population data. The revenue data is included in the NASBO dataset used for the primary outcome. The labor force participation rate is available for all years at the monthly level. Annual labor force participation rate is calculated by averaging seasonally adjusted monthly data from the St. Louis Federal Reserve Economic Data (FRED).

2.2.3 Treatment Assignment

Treatment was the adoption of the full Medicaid expansion by a state in 2014. Treatment assignments are based on prior studies that examined the effect of Medicaid expansion, but are slightly different. (Garrett and Kaestner, 2015; Hu et al., 2016; Sommers and Gruber, 2017). This analysis focused on the nine states that had no partial Medicaid expansions prior to 2014. The more narrow treatment group was selected because their pre-treatment spending on Medicaid was different from the other treated states. Prior work, however, included states with prior partial Medicaid expansions for parents and/or childless adults before they fully expanded Medicaid under the ACA in 2014 in the treatment group (Table 2.1).

Table 2.1: Treated States

Expansion in 2014 and No Prior Expansion for Parents or Childless Adults	
Arkansas	New Hampshire*
Kentucky	New Mexico
Michigan*	North Dakota
Nevada	Ohio
	West Virginia
*: Expanded in 2014 but later than Q1 2014	

Control states were states (Table 2.2) that did not expand Medicaid in 2014 or after 2014 under the ACA. States that did not expand Medicaid in or after 2014 exhibit two policy behaviors. First, the majority of control states (N = 16) did not expand Medicaid in or after 2014 and they had no prior partial expansions for parents and/or childless adults. Second, a smaller number of control states (N = 3) did not expand Medicaid in or after 2014 but they did have a partial expansion for parents and/or childless adults. For example, Wisconsin expanded Medicaid eligibility to all adults under 100% FPL, but was not included as a treated state because they did not expand Medicaid enough to receive the federal funding benefits the treated states did. The distinction between the types of expansions was because any expansion short of 138% FPL was not granted the federal funding benefits associated with Medicaid expansion under the ACA.

Table 2.2: Control States

No Expansion in 2014 and No Prior Expansion		**No Expansion in 2014 but Limited Prior Expansion for Parents and/or Childless Adults**
Alabama	North Carolina	Maine
Florida	Oklahoma	Tennessee
Georgia	South Carolina	Wisconsin
Idaho	South Dakota	
Kansas	Texas	
Mississippi	Utah*	
Missouri	Virginia*	
Nebraska	Wyoming	

*: Expanding Medicaid under the ACA in 2019

Excluded from the analysis were a total of 23 states that were divided into three categories (Table 2.3. The first category contains states that expanded Medicaid between 2015 and 2017 (N = 5). States expanding after 2014 were excluded because the treatment being examined was specific to Medicaid expansion in 2014 and the post-treatment period of interest was from 2015 through 2017. The second category were states that expanded Medicaid but had prior partial expansions for parents and/or childless adults after 2010 (N = 13). The decision to remove these states from the treated group was a divergence from previous literature. The states with prior partial expansions were excluded because they had large increases in spending between 2010 and 2014, which led to a large decrease in 2014 when the enhanced federal match rate began. The third category are states that had prior full expansions for parents and childless adults which then expanded under the ACA in 2014 (N = 5). The five states with full prior expansions of Medicaid have been used as control units in prior work (Garrett and Kaestner, 2015). Using states with full prior expansions as control units in this study, however, would be inappropriate as they were expected to have high per-capita spending that decreases, less dramatically than states with prior partial expansions, after 2014, when they accessed the special match rate. (Kaiser Family Foundation, 2013).

Table 2.3: Excluded States

Expansion Post-2014	**Expansion in 2014 Plus Prior Full Expansions for Parents and Childless Adults**	**Expansion in 2014 Plus Prior Partial Expansions for Parents and/or Childless Adults**	
Pennsylvania (1/1/2015)	Delaware	Arizona	Iowa
Indiana (2/1/2015)	Washington, D.C.	California	Maryland
Alaska (9/1/2015)	Massachusetts	Connecticut	Minnesota
Montana (1/1/2016)	New York	Colorado	New Jersey
Louisiana (7/1/2016)	Vermont	Hawaii	Oregon
		Illinois	Rhode Island
			Washington

2.3 Statistical Methods

The pre-period was from FY 1998 through FY 2013 and the post-period was from FY 2015 through FY 2017. The analysis did not include FY 2014 because FY 2014 contains six months of pre-treatment data and six months of post-treatment data that were non-separable. The analytical approach was a kernel-balanced synthetic control from the Trajectory Balancing framework (Hazlett and Xu, 2018a). Estimation of the kernel-balanced synthetic control was done using the *tjbal* package in R. A more thorough discussion of trajectory balancing is below.

2.3.1 Notation

G_i is a time-invariant indicator of treatment status for a state i with $N_{G=1}$ treated states and $N_{G=0}$ control states. Y_{it} is the outcome variable of interest which in this case is the per-capita state spending on Medicaid for a state i in year t. A total of T years of the outcome variable are observed for all states and the year prior to treatment is denoted as T_0.

2.3.2 Identification Assumption: Conditional Ignorability on Past Outcomes

Identification of the ATT_t relies on the conditional ignorability assumption where, in this instance, the non-treatment potential outcome (Y_{it}^0) is independent of treatment conditioning on a large feature expansion ($\phi()$) of the pre-treatment outcomes.

$$Y_{it}^0 \perp\!\!\!\perp G_i | \phi(\mathbf{Y}_{it}) \tag{2.1}$$

where $\mathbf{Y}$ is a $N \times T_0$ matrix of observed pre-treatment outcomes and ϕ is a large feature expansion of Y_{it} which we explain in more depth below.

The identification strategy also assumed that there exist weights such that the pre-treatment trends of the synthetic control are similar to the treated unit which then balances observed and unobserved confounders (Abadie et al., 2010).

2.3.3 The Kernel-balanced Synthetic Control

Synthetic control methods (mean or kernel-balanced) are motivated by the underlying idea that the data can be used to construct a control unit which is a better counterfactual than the average of the observed control units. For example, a difference-in-difference approach would utilize all control units equally to estimate the ATT. Synthetic control methods, however, construct a new control unit from the observed controls. If the above identifying assumptions hold, the ATT can then be estimated using the constructed, or synthetic, control unit.

The synthetic control unit is constructed as a weighted combination of the observed control units. For example, a study may be conducted with one treated unit and several control units. A synthetic control method will assign weights – that are non-negative and sum to one – to the control units such that one control unit may receive a weight of 0.5, another a weight of 0.1, and so on. The methodology for estimating the weights is described in more detail below. Once the weights have been estimated the counterfactual in the post-treatment period is imputed as the weighted combination of the observed control units outcome in the post-period. The ATT can be estimated as the difference between the average treated outcome and the synthetic control.

The ATT_t is given by:

$$\widehat{ATT_t} = \frac{1}{N_{G=1}} \sum_{G_i=1} Y_{it} - \sum_{G_i=0} w_i Y_{it}, \forall t > T_0 \tag{2.2}$$

where w_i are time-invariant weights estimated by an approximate balancing procedure and the ATT can be estimated in each post-period as the difference between the observed treated outcome and the kernel-balanced synthetic control.

The kernel-balanced synthetic control utilizes a kernel balancing procedure in order to estimate weights (Hazlett, 2020). The balancing procedure begins with the outcome data, $\mathbf{Y}$, which is a matrix with N rows – one for each unit – and T columns – one for each time period. Therefore, Y_i is a vector of length T for unit i. The outcome data is then transformed using a feature map that maps $Y_i \xrightarrow{\phi} \phi(Y_i)$ where $\phi(Y_i)$ is an expanded feature set that may included second-order polynomials and other higher order expansions of Y_i. Ideally, we would like to obtain balance on $\phi(Y_i)$, but it is difficult to know what feature map should be used.

The kernel-balancing procedure uses a Gaussian kernel to implicitly construct $\phi(Y_i)$ (Hazlett, 2020). The gaussian kernel maps $\mathbf{Y}$ to a kernel matrix $\mathbf{K}$:

$$\boldsymbol{K}_{ij} = k(Y_i, Y_j) = \exp(\frac{- \parallel Y_i - Y_j \parallel^2}{2\sigma^2}) \tag{2.3}$$

where $\mathbf{K}$ is an $(N \times N)$ matrix, Y_i is a $(T \times 1)$ vector of outcomes for unit i, Y_j is also a $(T \times 1)$ vector of outcomes for unit $j \neq i$.

Ideally the balancing procedure would try to find balance on the columns of the kernel matrix between the treated and control units. Unfortunately, finding exact balance on all N-dimensions of $\mathbf{K}$ is often infeasible, especially when N is large.

As a solution, the kernel balancing procedure finds *approximate balance* by minimizing a worst-case bias bound:

$$\min_{w_0} \parallel (w_1^T \mathbf{V}_1 - w_0^T \mathbf{V}_0)\mathbf{A}^{1/2} \parallel_2 \tag{2.4}$$

where w_0 are the weights for the control units that minimize the bias bound and w_1 are fixed weights for the treated units that are equal to one, $\mathbf{V}$ is an $N \times R$ matrix of eigenvectors which is subdivided by treatment status into $\mathbf{V}_0$, an $N_{G=0} \times R$ matrix, and $\mathbf{V}_1$, an $N_{G=1} \times R$ matrix, $\mathbf{A}$ is an $R \times R$ matrix with the eigenvalues on the diagonals.

The weights for the control units, w_0, are chosen using an iterative procedure. The procedure begins by ordering the columns of $\mathbf{A}$ in descending order such that the first column contains the largest eigenvalue; the associated columns in $\mathbf{V}$ are also reordered. The procedure then iteratively includes an increasing number of columns, estimates the weights to balance the subset $\mathbf{V}$, and calculates the bias bound for each iteration. The procedure ends when either the bias bound achieves a minimum or all R columns are utilized. For example, the procedure begins with the first column of $\mathbf{A}$ and $\mathbf{V}$, estimates weights to balance the first column of $\mathbf{V}$ between the control units and the average treated unit, and calculates the bias bound. The procedure repeats using the first two columns of $\mathbf{A}$ and $\mathbf{V}$, then the first three, and so on until either the bias bound achieves a minimum or all columns of $\mathbf{A}$ and $\mathbf{V}$ are utilized.

With each iteration of the aforementioned procedure the weights are estimated to achieve exact or near exact balance between the control units $\mathbf{V}_0$ and the average treated unit $\frac{1}{N_{G=1}} \sum \mathbf{V}_1$. The weights are found using entropy balancing which maximizes the entropy measure $\sum_i w_i log(w_i)$ (Hainmueller, 2012). Trajectory balancing has other options for estimating the weights, but the analysis in this paper utilizes the entropy balancing approach.

In this paper, a kernel-balanced synthetic control is used to estimate the average ATT from FY 2015 through FY 2017 as well as the ATT in each year. The ATT is interpreted as the average treatment effect on the treated of Medicaid expansion on the per-capita state spending on Medicaid. Inference is performed using a bootstrap procedure.

2.3.4 Validity Analyses

In addition the main analyses, five additional validity analyses were conducted to explore the underlying assumptions of the kernel-balanced synthetic control, the validity of the synthetic control, and the robustness of the results.

The first two validity analyses examined the assumption that by finding a near exact fit of the synthetic control with the treated outcome, in the pre-treatment period, confounders will be balanced (Abadie et al., 2010). The first validity analysis visualized the fit of the synthetic control to the observed treated outcome, which is the first part of the aforementioned assumption. Ideally, the synthetic control will be a near exact match with the observed outcome trajectory for the treated unit. A limitation of this analysis, however, is that it doesn't take into account the fact that the kernel-balancing procedure balances across different moments of the data, and not just the means. As a result, the mean balance examined in this analysis may be poor but that could be because the higher order moments take priority when minimizing the bias bound. In this paper we propose estimating a mean-balanced synthetic control and visualizing how well it fits the treated outcome data. The logic is such: if the mean-balanced synthetic control is able to achieve a good fit then it is likely that trajectory balancing is prioritizing other moments of the data and is therefore not a feasibility issue.

The second validity analysis examined the second piece of the assumption discussed in the prior paragraph – that balance is achieved on both observed and unobserved confounders. The balance check was implemented by comparing the absolute mean difference between the treated and control groups on observed confounders before and after weighting. The assumption would be supported if the absolute mean difference shifted towards zero, post-weighting. The analysis is limited, however, because balance on unobserved confounders cannot be investigated; but it can still be useful if balance is achieved on what are considered, by expert content knowledge, to be the most important confounders.

The third validity analysis is known as *backdating* and is similar to an analysis proposed in other work (Abadie, 2019). Backdating works by iteratively moving the treatment period back in time – into the pre-treatment period – and estimating the synthetic control with the reduced pre-period. In this paper the backdated treatment period is a three-year window and doesn't include the post-treatment period. Backdating analyzes the credibility of the synthetic control to accurately model the treated units outcome trajectory by comparing the synthetic control to a ground truth, the observed treated outcome. For example, if the backdated synthetic control estimates an ATT of zero for the three-year treatment window – in the pre-treatment period – then the synthetic control is able to accurately predict the treated units pre-period trajectory. However, if the synthetic control is unable to accurately predict the treated units pre-period trajectory then we would be concerned about the ability of the synthetic control to model the counterfactual

Fourth, the effect of using all pre-periods in the data was examined by reproducing the main analysis but with an incremental reduction in the number of pre-periods, in chronological order. The purpose for incrementally reducing the number of pre-periods was to provide transparency around the decision to utilize all of the pre-periods. The ATT was reported for each incremental reduction in the number of pre-periods to reduce any concern of researcher bias.

2.4 Results

2.4.1 Descriptive Analysis

The data in Figure 2.1 visualize the trends in annual per-capita spending from the state's general and other funds on the Medicaid program. The per-capita state spending on Medicaid was relatively similar starting in FY 1996 and slowly began to diverge through FY 2017. The blue colors highlight the treated states and the color changes from light to dark blue when they enter the treatment period in FY 2014. The treated states, aside from Ohio which was not included in the analysis, are similar in magnitude to the control states. Ohio was an outlier because they passed federal funding for Medicaid through their general fund (Potamianos et al., 2018). For some fiscal years the amount of federal funding passed through the general fund is reported to NASBO, however, for other years it was not reported. As a result, this analysis did not include Ohio because it was not possible to separate federal and state funding for Medicaid.

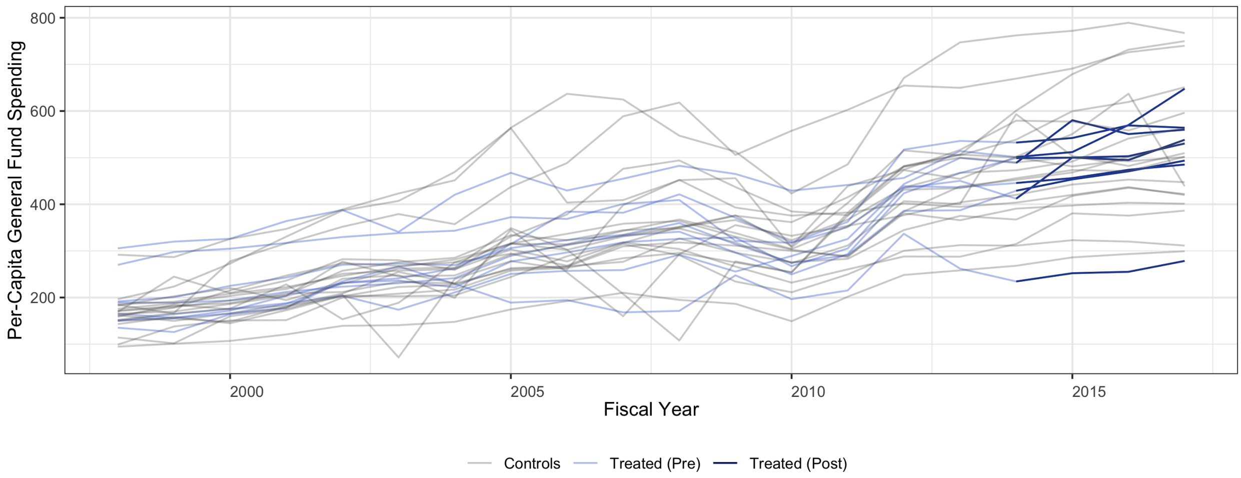


Figure 2.1: Trends in Per-Capita General & Other Fund Spending on Medicaid by State

The trends in per-capita state spending on Medicaid were mostly increasing across all states over time, but there was a noticeable dip that occurs between 2008 and 2012 across almost all states in the sample. A possible explanation for the decrease in per-capita state spending on Medicaid was the American Recovery and Reinvestment Act (ARRA). ARRA provided federal funding to state Medicaid programs to assist with increases in Medicaid expenditures following the recession. The ARRA relief lasted from October 2008 to June 2011 (FY 2009 through FY 2011) (Kaiser Family Foundation, 2011).

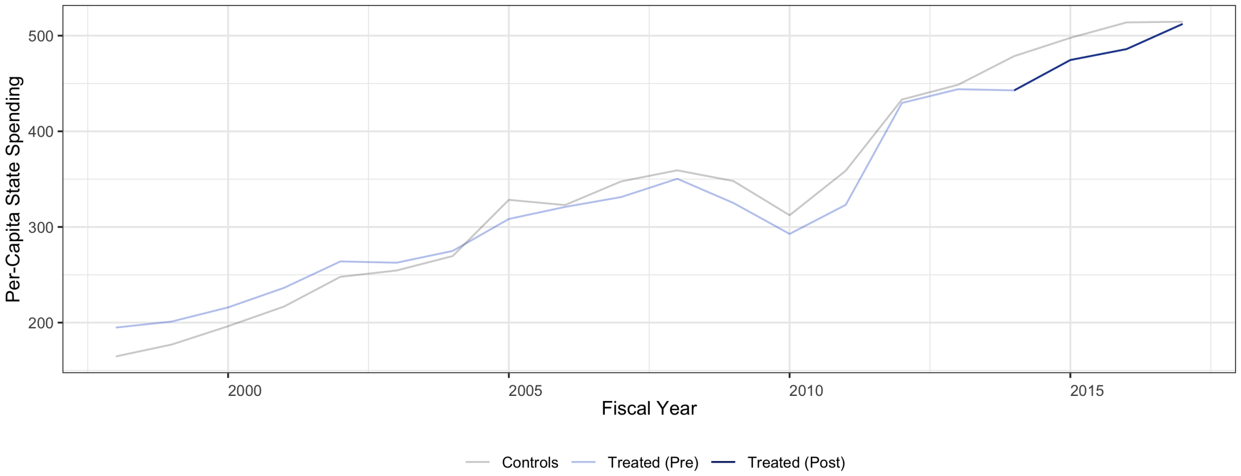


Figure 2.2: Trends in Unweighted Average State Per-Capita Spending from General & Other Funds on Medicaid by Expansion Status

In Figure 2.2, the control and treated groups average per-capita combined general and other fund spending on Medicaid were relatively similar in magnitude and trend. There were instances where the trends diverged such as just prior to treatment. Furthermore, the magnitude of per-capita state spending on Medicaid has nearly doubled, on average, since 1995 and continues to rise. The treated group's average per-capita state spending on Medicaid from FY 2015 to 2017 was $490.92 with a high of $512.25 in FY 2017. The control group's average per-capita state spending on Medicaid from FY 2015 to 2017 was $508.70 with a high of $514.55 in FY 2017.

2.4.2 Main Results: The Effect of Medicaid Expansion on State Medicaid Spending

Treatment effects were estimated using a kernel-balanced synthetic control and confidence intervals were estimated using the bootstrap method. The top five weighted states were: Nebraska, Utah, Kansas, Florida, and South Dakota. The ATT over the entire treatment period (FY 2015 to FY 2017) was a decrease of -$12.31, or -%2.50, decrease in per-capita spending from state funds on Medicaid. The year-by-year ATT ranged from -$20.42 in FY 2016 to $0.66 in FY 2017 (Table 2.4). Although the year-by-year results show a sudden increase in the treatment effect in FY 2017 (perhaps reflecting the increasing state share of newly eligible costs), they are also statistically insignificant. The results do not providence evidence in support of the hypothesis that Medicaid expansion increased per-capita State spending on Medicaid.

Table 2.4: Effect of Medicaid Expansion on Per-Capita State Spending on Medicaid

Fiscal Year	**ATT**	**Confidence Interval**
2015	-$17.78	-$82.60, $48.25
2016	-$20.42	-$74.67, $33.83
2017	$0.66	-$56.66, $58.98

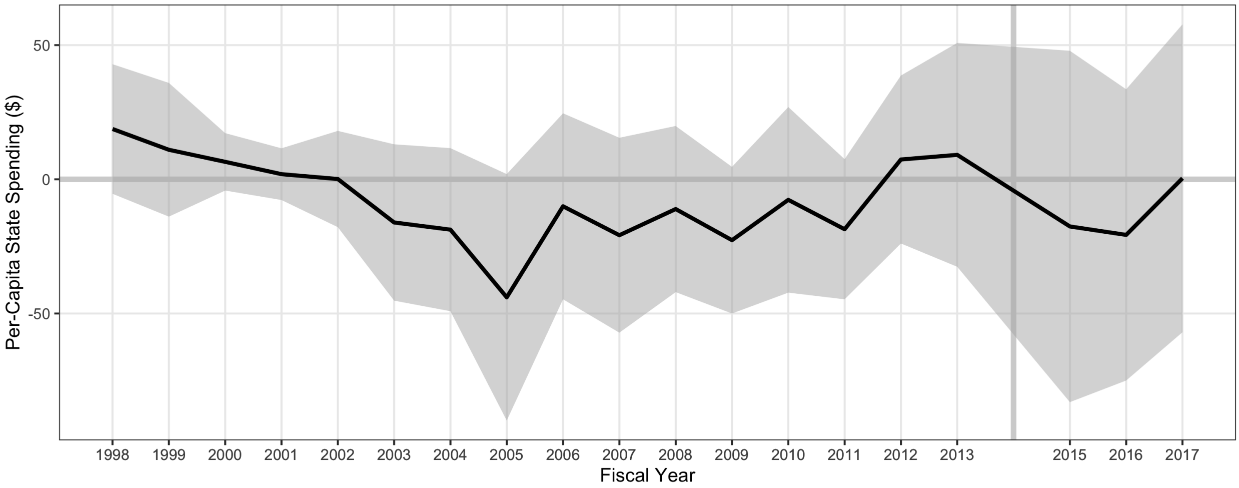


Figure 2.3: Combined General Fund & Other Fund Spending on Medicaid – Average Treatment Effect on the Treated by Year

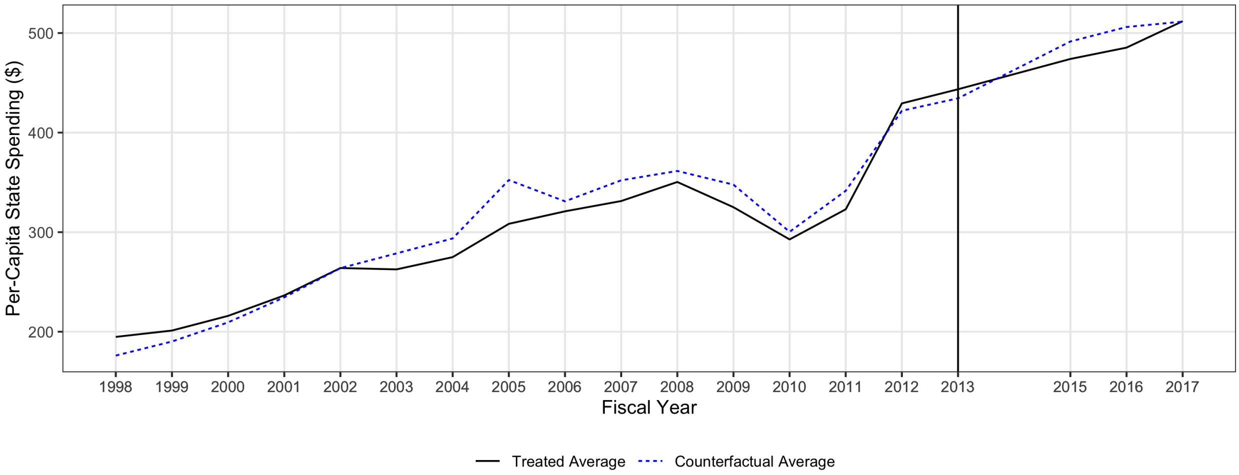


Figure 2.4: Kernel-Balanced Synthetic Control Fit Compared to Treated Group

2.4.3 Validity Analyses

Five distinct validity, robustness, and sensitivity checks were conducted to evaluate the validity and credibility of the synthetic control.

2.4.3.1 Validity of the Synthetic Control - Main Effect

Important to the validity of the kernel-balanced synthetic control is the quality of the estimated synthetic control. Figure 2.4 displays the mean per-capita state spending on Medicaid for the treated units compared to the synthetic control estimated by the kernel-balanced synthetic control. The beginning of the treatment period is denoted by a vertical line. Perfect balance — rarely attainable with real-world data — would align the synthetic control and the observed data exactly for each period prior to the grey line.

In these data, the kernel-balancing procedure was not able to find a set of weights such that the synthetic control exactly matches the treated group's outcome trajectory over the pre-period. Overall, The magnitude of the gaps between the synthetic control and treated groups mean is of some concern, especially when considered relative to the ATT. The average distance between the treated group and the synthetic control group's means, in the pre-period, was -\$7.22 and the maximum absolute distance was \$43.99, which occurred in FY 2005. The largest gaps between the synthetic control and the treated group means occur at the beginning and the mid-point of the pre-treatment period in FY 1998, 1999, and between 2003 and 2009. As the pre-period approaches the treatment period the difference between the two groups becomes small.

Further analysis compared the fit of the kernel-balanced synthetic control with a mean-balanced synthetic control to evaluate whether or not weights exist that balance the mean structure – a feasibility analysis. Because a mean-balanced synthetic control is solely focused on finding mean balance it provides a check on the feasibility of the weights for the mean structure. The fit of the mean-balanced synthetic control in Figure 2.5 supports the validity of the kernel-balanced synthetic control because the mean-balanced synthetic control fit better to the periods that were difficult. Therefore, it is likely the case that the kernel-balanced synthetic control is prioritizing other moments of the data and sacrificing mean balance as a result.

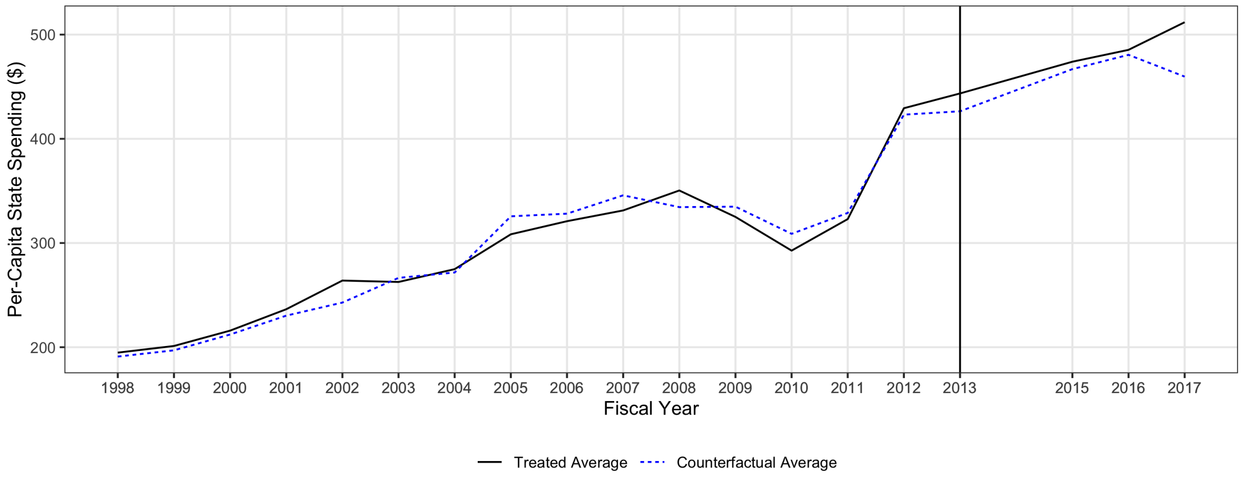


Figure 2.5: Mean-Balanced Synthetic Control Fit Compared to Treated Group

2.4.3.2 Post-Weighting Confounder Balance Checks

Important to synthetic control identification strategy is improved balance among time-varying confounders between the treated and weighted control units compared to the unweighted control units (Abadie et al., 2010). Therefore, a validity analysis is to compare balance of observed time-varying confounders before and after weighting. Figure 2.6 compares balance on time-varying confounders with and without the weights estimated by the kernel-balancing procedure; it also includes 95% confidence intervals. The confounders examined were the unemployment rate, state revenue per-capita, and the labor force participation rate (Sommers and Gruber, 2017).

In Figure 2.6, the results of the balance checks - for the main analysis - provide little support for the validity of the synthetic control. For example, prior to weighting the unemployment rate and state income per-capita had absolute mean differences, by treatment status, near zero. Labor force participation rate, however, was very different from zero with an absolute mean difference of 0.47. After weighting. balance improved slightly for both the unemployment rate but state income per-capita became slightly worse. Balance on labor force participation improved the most post-weighting such that the point estimate was closer to zero but it was still not inline with the absolute mean differences for unemployment rate and state income per capita.

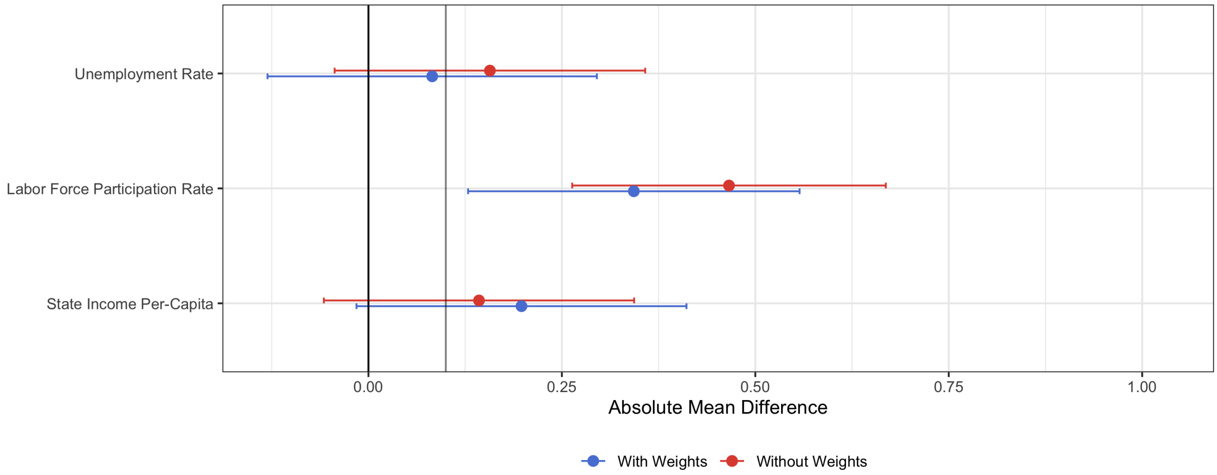


Figure 2.6: Time-Varying Confounder Balance Before and After Weighting with Confidence Intervals

2.4.3.3 Backdated Treatment Period

The backdated treatment period analysis reproduced the main analysis but changed the definition of the treatment period. The actual treatment period (FY 2015 - FY 2017) was removed from the data such that the data consisted only of the pre-treatment period. The treatment period was then reassigned to a three-year period within pre-period. For example, the full pre-period was from FY 1998 through FY 2013. The treatment period was first assigned to FY 2011 through FY 2013, a three-year window similar to the actual treatment period. The three-year post-treatment window was iteratively shifted back in time by one pre-period such that it runs from FY 2010 through FY 2012, FY 2009 through FY 2011, and so on. The three-year window was utilized because the actual post-treatment period was a three-year window. Therefore, we are interested in understanding how well the kernel-balanced synthetic control predicts the observed treated outcome over a placebo treatment period of length equal to the true treatment period.

The average of the ATTs over the backdated post-periods, reported in Table 2.5, were mixed because the periods near the true treatment period returned results that support the credibility of the synthetic control, but results for earlier periods generally did not. The results just prior to treatment returned small, near zero ATTs, which support the ability of the kernel-balanced synthetic control to correctly model the treated unit. However, as the placebo treatment window moved back in time the ATT became large and for placebo treatment windows between FY 2005 - 2007 and FY 2007 - 2009 the ATT is on average -$52.63, which is large especially when compared to the main ATT.

The results raise some concern over the ability of trajectory balancing to correctly model the average treated unit between FY 2005 and FY 2010. However, these concerns are somewhat assuaged by the near zero ATTs just prior to the true treatment period, which indicates that while the kernel-balanced synthetic control is having difficulties with FY 2005 through FY 2010 it is seemingly able to correctly model the treated unit for periods after FY 2010.

Table 2.5: Placebo Treatment Period Average Effects on the Treated

Placebo Treatment Years (FY)	**ATT**	**Confidence Interval**
2011 - 2013	$0.91	-$34.38, $36.2
2010 - 2012	-$4.24	-$61.24, $52.77
2009 - 2011	-$28.48	-$45.93, $8.96
2008 - 2010	-$0.28	-$67.05, $66.48
2007 - 2009	-$42.80	-$96.70, $11.10
2006 - 2008	-$66.08	-$170.51, $38.36
2005 - 2007	-$49.02	-$114.27, $16.23
2004 - 2006	-$21.35	-$76.87, $34.16
2003 - 2005	-$27.46	-$69.96, $15.03
2002 - 2004	-$16.74	-$57.11, $23.62
2001 - 2003	-$7.58	-$35.19, $20.03

Despite the null results, of some concern is the strong ATT for placebo treatment years FY 2005 through FY 2010. A possible explanation is the American Recovery and Reinvestment Act (ARRA), which is described in the exploratory analysis (Figure 2.1). There may have been a differential impact of ARRA between treated and control groups due to the way ARRA was implemented. ARRA was implemented as a flat percentage point decrease in every state's share of Medicaid expenses of 6.5 percentage points. For example, a state that pays half of their Medicaid costs would experience a 12% decrease while a state that pays 25% of their Medicaid costs would have experienced an 24% decrease in the amount they paid. However, the average share of Medicaid costs for the treated states in FY 2012 was 36.00% and for control states it was 37.36%.

2.4.3.4 Robustness to Different Pre-Period Lengths

The purpose of this validity analysis was to provide transparency as to the decision to utilize all available pre-periods, and examine the robustness of the synthetic control to different pre-period lengths. The ATT was used to evaluate the robustness of the results to different pre-period lengths.

The results in Table 2.6 support the conclusion that the synthetic control is robust to the number of pre-treatment periods. The first row is the same as the main results and the second row begins in FY 1999 instead of FY 1998. The average ATT across all post-periods varied as the number of pre-periods decreased. The average effect ranged from a low of -\$10.42 to a high of \$26.64 with an average of -\$0.63. Despite the variation in the average ATT, the confidence intervals all included zero and lead to the same conclusion as the main results.

Table 2.6: Average ATT of Medicaid Expansion on State Spending on Medicaid with Decreasing Number of Pre-Periods (Main analysis in Bold)

Pre-Period Start (FY)	Average ATT	Confidence Interval
1998	**-$4.21**	**-$57.96, $49.54**
1999	-$6.66	-$57.61, $44.29
2000	-$8.85	-$42.68, $24.98
2001	-$8.08	-$57.98, $41.82
2002	-$7.87	-$57.68, $41.93
2003	-$10.42	-$55.52, $34.68
2004	-$9.10	-$55.17, $36.98
2005	$1.97	-$24.86, $28.81
2006	$23.05	-$18.92, $65.01
2007	$1.69	-$34.45, $37.83
2008	-$4.04	-$64.46, $56.38
2009	-$0.26	-$61.21, $60.69
2010	-$2.69	-$53.86, $48.48
2011	$26.64	-$8.05, $61.33

2.5 Discussion

The main results do not provide evidence that Medicaid Expansion lead to an increase in state spending on Medicaid. States that expanded Medicaid under the ACA with no prior expansions had, on average, a non-significant decrease of -$12.31 or -2.50% in per-capita state spending on Medicaid between FY 2015 and 2017. The overall effect was estimated over the three fiscal years following the policy change, which builds upon prior work using only one fiscal year post-treatment. The ATT in FY 2017 suggests that the policy effect grew stronger over time – relative to the ATT in FY 2015 and FY 2016 – and was estimated here at \$0.66 or 0.13% increase in per-capita state spending on Medicaid in FY 2017, but it also overlaps with the decrease in the federal match rate for the newly eligible population from 100% to 95%.

Other quasi-experimental work on this topic was limited to one prior study which the findings in the current paper are consistent with. For example, prior work found that state spending increased by a small, non-significant amount of 2.4% in FY 2015 (Sommers and Gruber, 2017). Similarly, this paper found a small non-significant change in state-spending, but the effect was of the opposite sign as it was a decrease of -3.70% over the same time-period. However, direct comparison of this paper to prior work requires caution due to differences in treatment assignments. For example, prior work utilized all 50 states in their quasi-experimental analysis, whereas the current study utilized only 27 states.

While this paper was not the first quasi-experimental paper to study the effects of Medicaid expansion on state Medicaid spending, it was the first to utilize an alternative methodology - a kernel-balanced synthetic control - that does not assume parallel trends. By not assuming parallel trends the kernel-balanced synthetic control method provides the opportunity to make causal inferences in cases where the parallel trends assumption is unlikely to hold. The kernel-balanced synthetic control does, however, make the assumption that by finding good balance on the outcome confounders - observed and unobserved - are balanced. This paper presented a process for analyzing the assumptions of trajectory balancing through a series of validity checks on the fit of the synthetic control and the degree to which balance was achieved on observed confounders.

The additional analyses contributed some evidence in support of the model used for the main analysis. For example, there were small improvements in balance on some time-varying confounders (Figure 2.6). Balance on time-varying confounders post-weighting is important for identification with synthetic control methods and the improvements for the state income per-capita, and minimal change for the unemployment rate contributed to the validity of the synthetic control but the balance for the labor force participation rate was still poor.

What is more supportive is that the kernel-balanced synthetic control matched the treated outcome well in most of the pre-periods and for the periods that it matched poorly on a mean-balanced synthetic control was able to match well. Therefore, the kernel-balanced synthetic control was likely prioritizing higher order moments over mean balance for some pre-periods.

Limitations

There are three limitations of note to this study. First, the data was voluntarily reported by the states to the National Association of State Budget Officers (NASBO). The danger of using voluntarily reported data is that the labeling of aggregate funding and revenue streams may not be done consistently across states. States, however, had the option of including footnotes to explain any unusual state budget practices, exceptions, or when data was unavailable. A review of the footnotes for FY 1998 through FY 2017 was conducted and identified two states with unusual funding reports (Ohio and Connecticut) as well as other reporting errors that had already been corrected within the data. Despite these concerns, the NASBO data has been used for descriptive analysis by policymakers, government organizations, and commissions such as the CBO, GAO, and MACPAC (Congressional Budget Office, 2007; Government Accountability Office, 2012; MACPAC, 2017).

Second, this paper estimated a group average effect among a collection of states rather than a state specific effect. For example, there was variation across states in Medicaid programmatic structure related to not only eligibility rules, but enrollment and renewal procedures, and cost sharing practices as well (Heberlein et al., 2012). The effect estimated in this paper did not illuminate how the aforementioned programmatic designs directly influence state spending. Rather, the effect contributes to understanding whether or not there was, on average, an increase in state spending on Medicaid following a Medicaid expansion for adults with incomes up to 138% of FPL in 2014. Therefore, the results are limited to answer the posed question and do not provide further insight to whether or not specific programmatic decisions reduced state spending on Medicaid. Furthermore, the results do not lay claim that all states in the treated group experienced a small, non-significant change in per-capita state spending on Medicaid. Rather, the results speak to the average change in per-capita state spending on Medicaid among the treated group.

Third, there are differences in the time-cycle of state budgets. For example, in 2011 there were four states with biennial sessions of which 3 are included in this analysis (Nevada, North Dakota, and Texas). In addition, Arkansas and Oregon switched from a biennial session to an annual session in 2009 and 2011, respectively. Oregon, however, is not included in this analysis. Analysis without states with biennial budgets results in estimates closer to null and an analysis without Arkansas results in similar estimates to the main effects in this paper.

2.6 Conclusion

Medicaid expansion and its surrounding media attention led to concerns that there would be a sudden increase in previously eligible individuals newly enrolling in Medicaid, the woodwork effect. These concerns led to the hypothesis that Medicaid expansion will increase state spending on Medicaid, notwithstanding the generous federal match for expansion. Concerns of a woodwork effect were used to justify non-expansion in many states (Alonso-Zaldivar, 2014).

The current work contributes to the methodological literature by expanding the role of the kernel-balanced synthetic control (Hazlett and Xu, 2018a), a type of synthetic control method, in the health policy literature. The method is an alternative to difference-in-difference that does not assume parallel trends, which makes it a valuable estimator for the empirical health policy researcher. The current paper presented four ways to assert the validity of the estimator and we stress the importance of validity analysis with any synthetic control methodology. A kernel-balanced synthetic control can be easily implemented in R using the *tjbal* package.

This paper contributes to a subset of the expansive Medicaid expansion literature focused on estimating the financial impacts of Medicaid expansion to local, state, and federal governments. The results of this study suggest that Medicaid expansion did not greatly increase state spending on Medicaid from FY 2015 through FY 2017. States that expanded Medicaid eligibility had an average decrease of -$19.10 or 3.86% in per-capita spending on Medicaid from FY 2015 through FY 2016. The rationale for failing to expand Medicaid does not withstand empirical scrutiny in this analysis.

Appendix A

Additional Tables and Figures

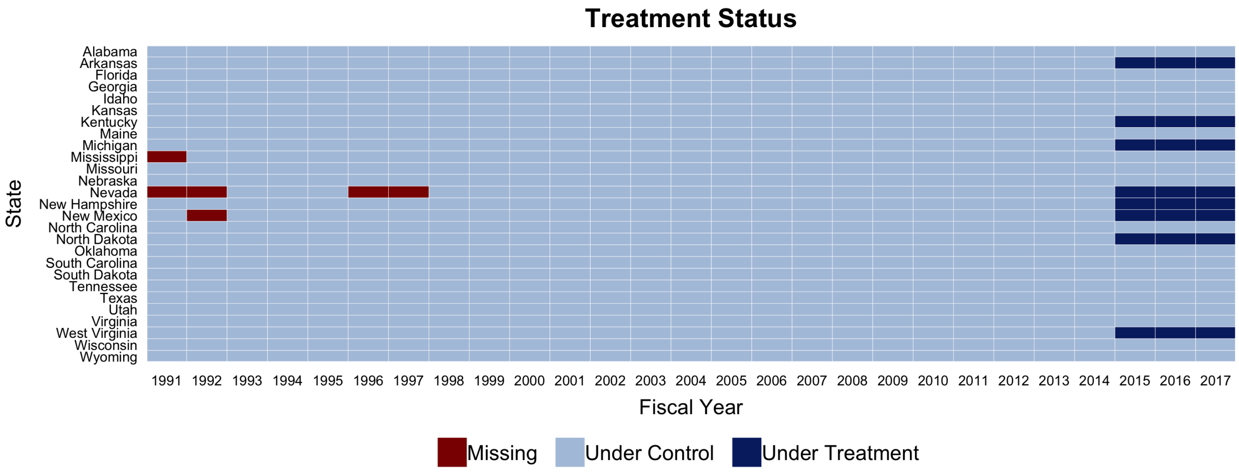


Figure A.1: Missingness by State and Treatment Status for States in Analytic Dataset

Figure A.2, below, details the weights assigned by trajectory balancing in the main analysis. The states in dark blue are treated states and the control states are on a gradient of light blue that becomes darker as the weight becomes greater. The five states with the largest weights are (in descending order): Tennessee, South Carolina, Nebraska, Florida, and Alabama.

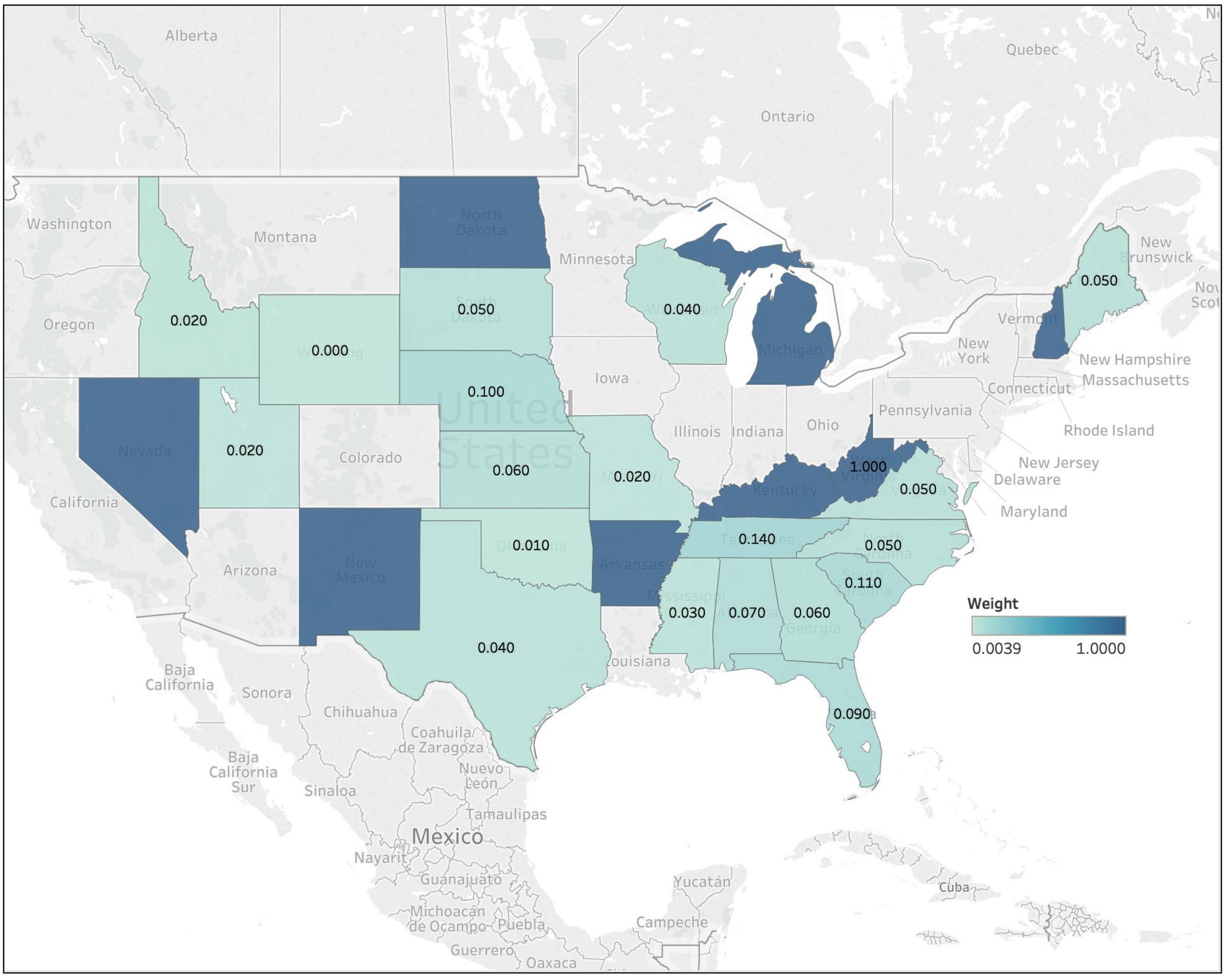


Figure A.2: Map of Weights Assigned by the Kernel-Balanced Synthetic Control

Table A.1: Percent Change in Enrollment (Average of July to Sep 2013) vs. Sep 2014 – Control States

State	**Percent Change: Pre-ACA to Sep 2014**
Alabama	9%
Florida	9%
Georgia	16%
Idaho	20%
Kansas	6%
Maine	N/A
Mississippi	10%
Missouri	-2%
Nebraska	-1%
North Carolina	9%
Oklahoma	2%
South Carolina	0%
South Dakota	0%
Tennessee	11%
Texas	6%
Utah	0%
Virginia	3%
Wisconsin	5%
Wyoming	5%

Appendix B

Visualizing Trends in State Medicaid Eligibility Rules

This section provides additional material that visualizes the Medicaid eligibility rules prior to Medicaid expansion and for a short period after, which has important implications to how the definition of treatment and control groups. Medicaid eligibility rules, based on income, can be broken into four groups: Pregnant women, Parents, Children, and Other Adults (Non-disabled and Non-seniors). The four graphics below visualize the eligibility rules based on data from Kaiser Family Foundation: Trends in Medicaid Expansion, for each of the aforementioned groups.

Figure D1 displays the trends in Medicaid eligibility limits for other adults from 2011 to 2019. Other adults were largely not covered prior to Medicaid expansion except for a handful of states: Delaware, Massachusetts, New York, and Vermont. Note, in the graphic below Massachusetts does not appear to cover childless adults, but this was because they covered childless adults through an alternative program known as Commonwealth Care. Furthermore, the majority of control states who did not provide coverage to other adults did not change their coverage policies following Medicaid expansion in 2014. States that provided coverage prior to Medicaid expansion continued to provide stable coverage following Medicaid expansion.

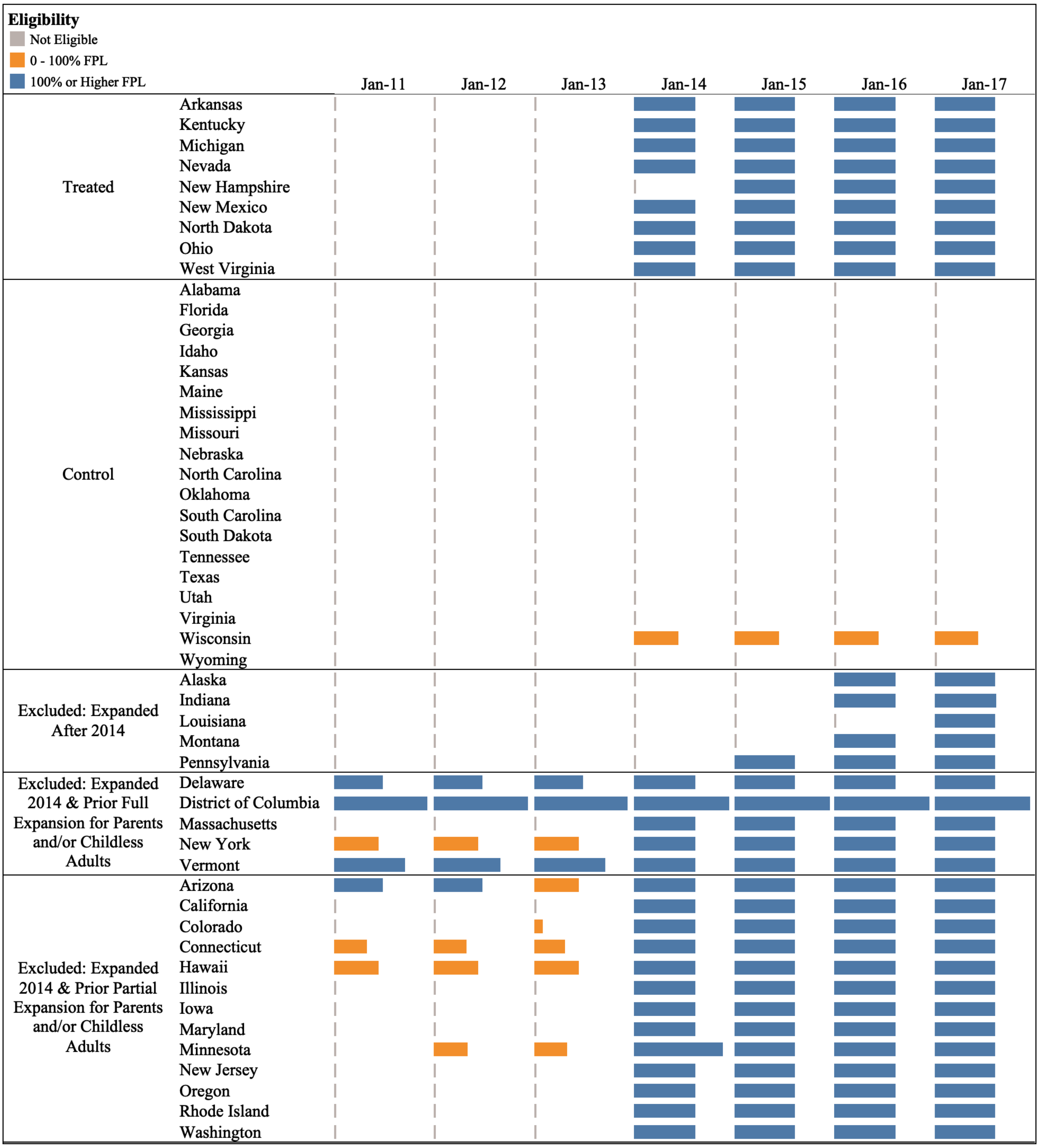


Figure B.1: Medicaid Income Eligibility Limits for Other Non-Disabled Adults, 2011 to 2017

Figure D2, below, visualizes trends in Medicaid eligibility for parents from 2009 to 2017. Notably, many states provided coverage for parents to varying degrees prior to Medicaid expansion, and following it. Among the control states there was significant variation in the eligibility limits for parents, but it remained largely stable overtime within states. The treated states exhibited minor variation prior to Medicaid expansion, but following Medicaid expansion covered parents up to 138% of the FPL.

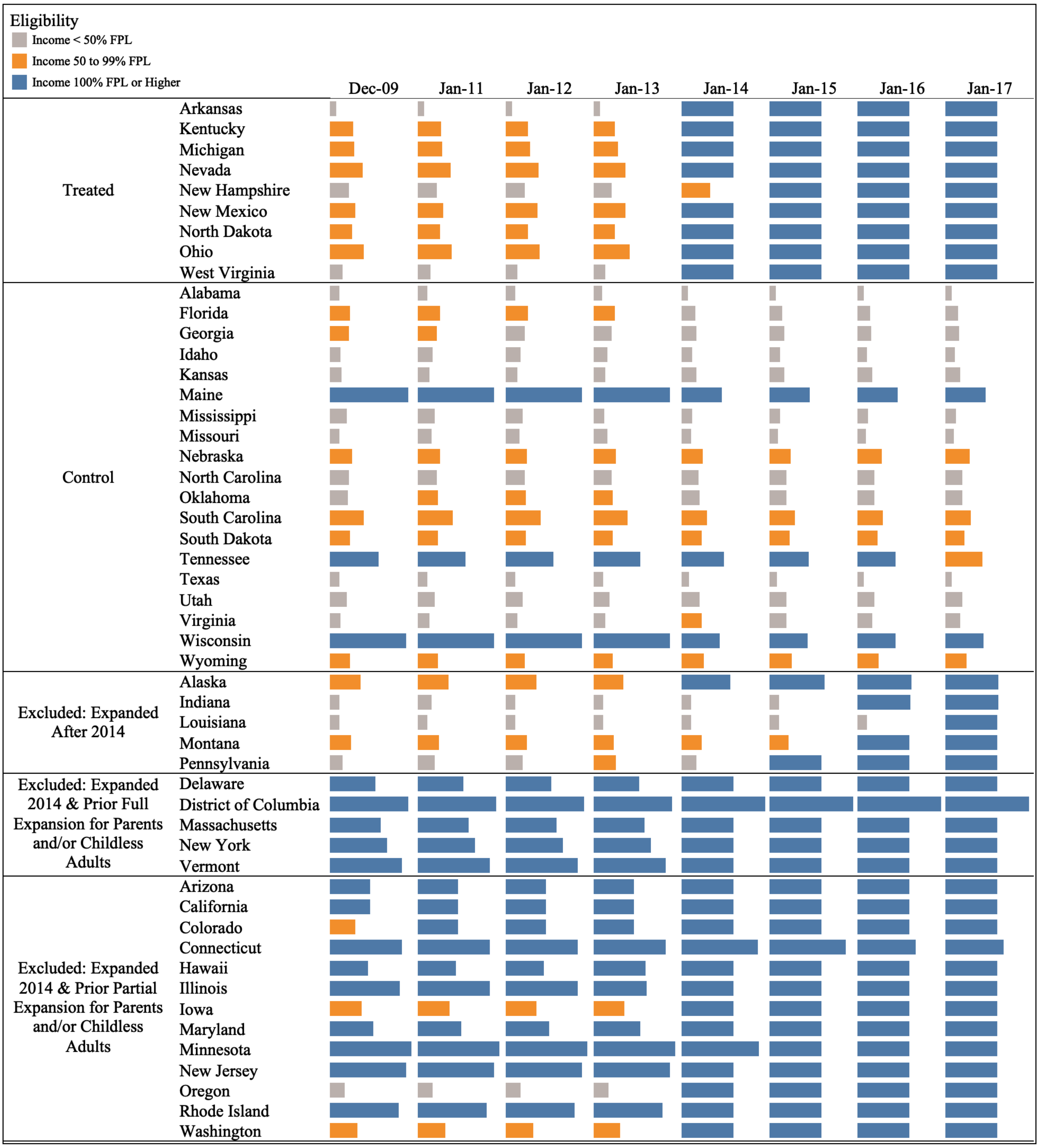


Figure B.2: Medicaid Income Eligibility Limits for Parents, 2009 to 2017

Figure D3, below, visualizes the trends in Medicaid eligibility limits for pregnant women from 2006 to 2017. Overall, pregnant women received much better coverage, prior to the

ACA, compared to other adults and parents. The range of eligibility limits, however, was much larger and ranges from 133% FPL up to 300% FPL, in some years. Interestingly, nearly all states increased the eligibility limit in 2014 such that the median eligibility limit between the treated and control groups were nearly identical (200% FPL and 201% FPL).

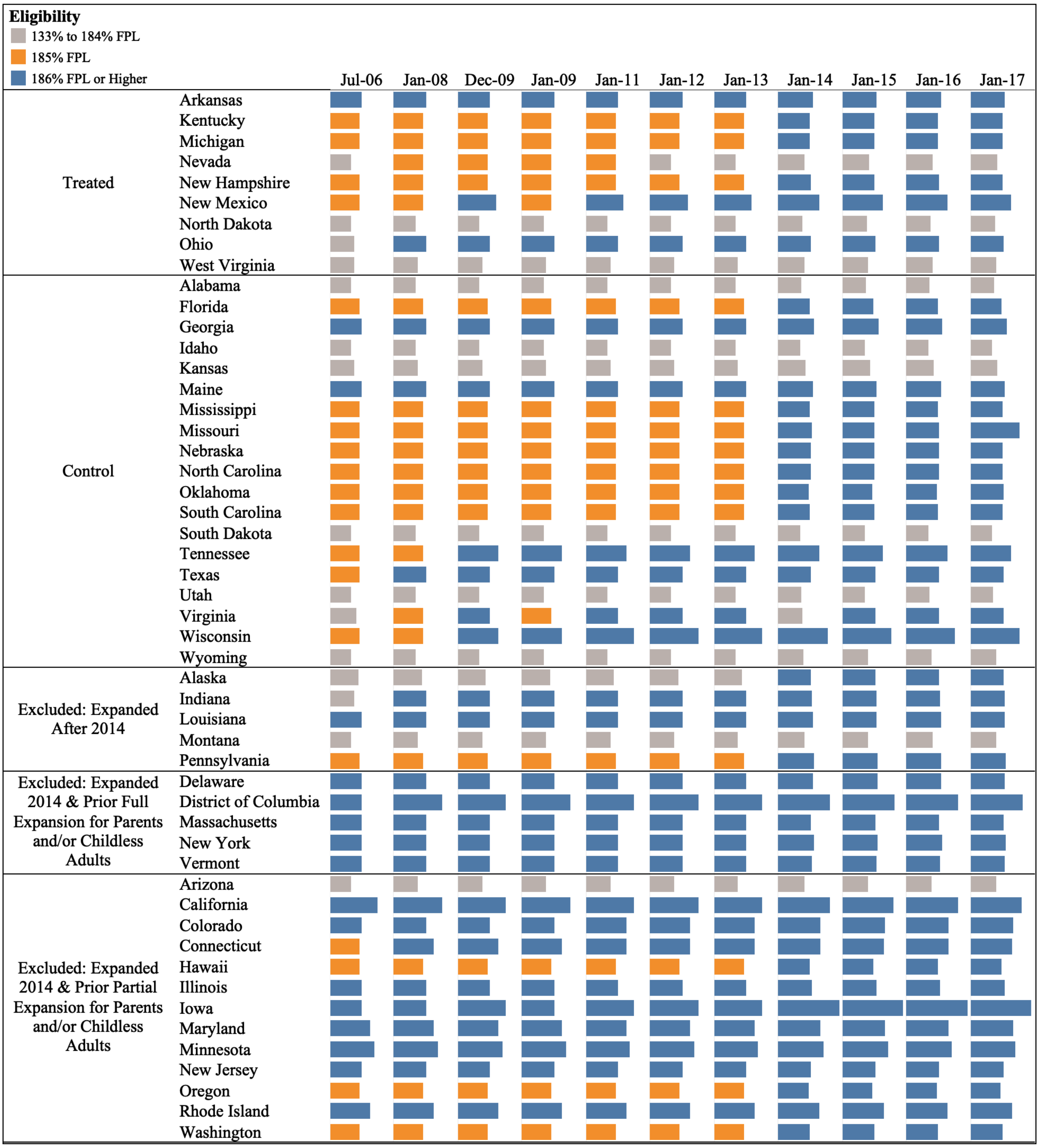


Figure B.3: Medicaid Income Eligibility Limits for Pregnant Women, 2006 to 2017

Children, who were primarily funding through CHIP, received the most generous eligibility limits across all states. Figure D4, below, visualizes trends in Medicaid and CHIP eligibility limits from 2009 to 2017. Overall, the eligibility limits vary from state to state within both the control and the treated groups.

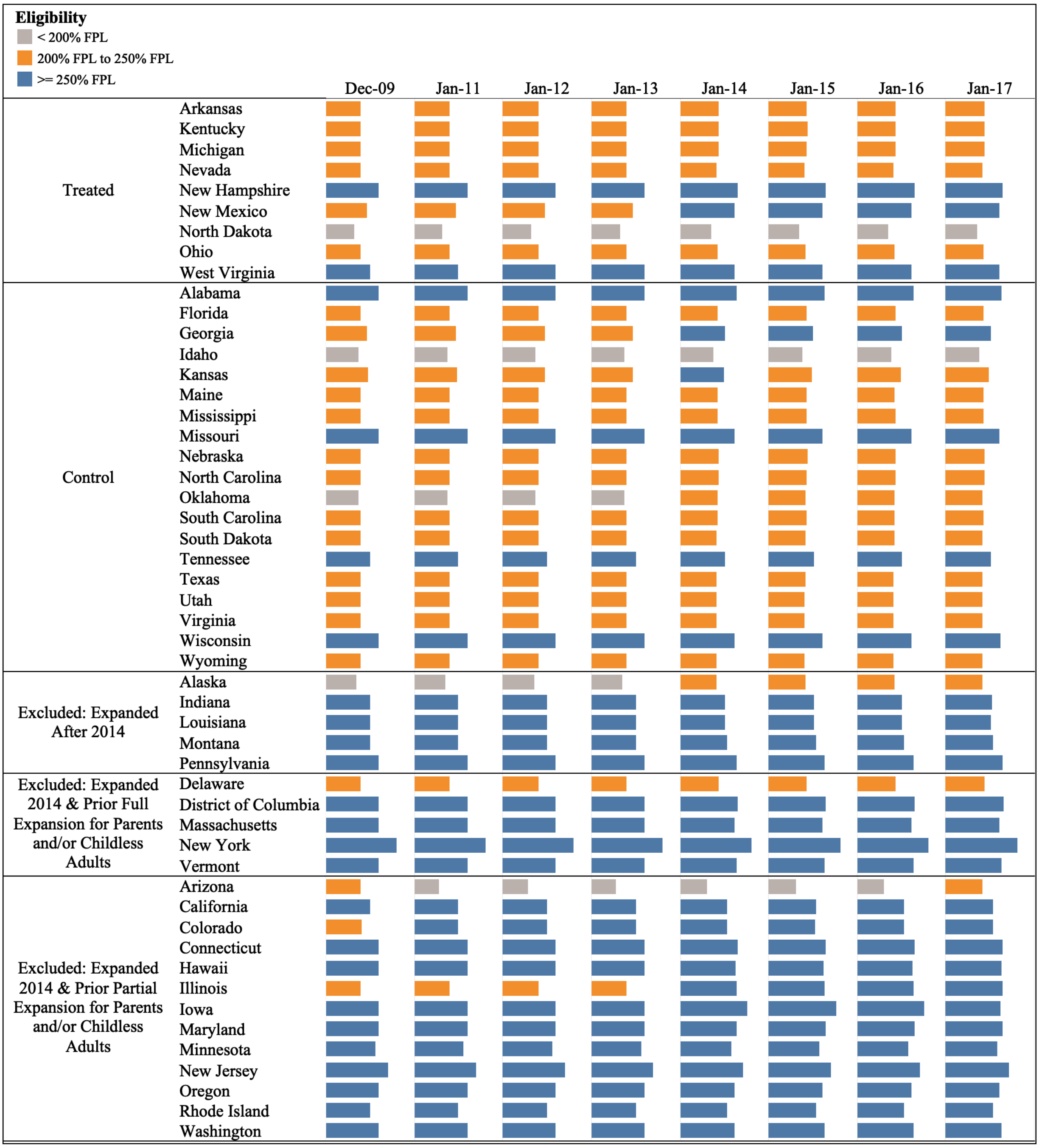


Figure B.4: Medicaid and CHIP Income Eligibility Limits for Children, 2009 to 2017

In addition to visualizing the trends we can summarize the average eligibility for pregnant women and parents as of December 1st, 2009. December 2009 was an important date because the eligibility limits, as of that date, were utilized for determining the thresholds for traditional matching under the FMAP and newly eligible enhanced federal matching. Table B.1, below, summarizes the average Medicaid income eligibility limits as of December 2009 for the states in the treated group.

The results in table B.1 describes large variation in Medicaid eligibility limits across states for both pregnant women, parents, and in the averages across both categories. For example, the average Medicaid eligibility limits among the treated states in December 2009 had a maximum of 151% in New Mexico and a minimum of 92% in West Virginia.

Table B.1: Medicaid Income Eligibility Limits as of December 2009

State	**Pregnant Women**	**Parents**	**Average**
Arkansas	200%	17%	109%
Kentucky	185%	62%	124%
Michigan	185%	64%	125%
Nevada	185%	88%	137%
New Hampshire	185%	49%	117%
New Mexico	235%	67%	151%
North Dakota	133%	59%	96%
West Virginia	150%	33%	92%

Appendix C

Supplemental Analysis 1

The supplemental analysis below replicates the main analysis but removes Florida from the pool of control states. Florida was removed from the pool of control states because in 2011 Florida began including Medicaid administrative expenses in the overall state funds reported to NASBO. As a result, the state funds spent on Medicaid increased from 2010 to 2011.

The results of the supplemental analysis are below in Table C.1. The overall ATT was slightly smaller than the ATT from the main analysis (-10.75 vs -12.31), and the year-by-year ATTs follow the same trend as the main results, but are slightly smaller.

Table C.1: Effect of Medicaid Expansion on Per-Capita State Spending on Medicaid - No Florida

Fiscal Year	**ATT**	**Confidence Interval**
2015	$-14.06	-$74.06, $45.94
2016	$-19.91	-$67.12, $27.80
2017	$1.72	-$48.92, $52.35

Appendix D

Supplemental Analysis 2

The supplemental analysis below replicates the main analysis but removes Michigan and New Hampshire from the pool of treated states. Michigan and New Hampshire are removed because they expanded much later in 2014. Thus, the below estimates are for states that expanded in January 2014.

The results of the supplemental analysis are below in Table D.1. The overall ATT was larger than the ATT from the main analysis (16.86 vs -12.31), and the year-by-year ATTs followed a different trend from the main analysis.

Table D.1: Effect of Medicaid Expansion on Per-Capita State Spending on Medicaid - No Michigan or New Hampshire

Fiscal Year	**ATT**	**Confidence Interval**
2015	$18.78	-$20.16, $57.71
2016	$7.86	-$30.96, $46.68
2017	$23.93	-$19.07, $66.94

Appendix E

Revisiting Sommers & Gruber

Despite being able to replicate Exhibit 2 from Sommers & Gruber, this paper was unable to replicate their model results. In their paper, Sommers & Gruber state that they used a two-way fixed effects difference-in-difference model that controlled for the annual unemployment rate and the per-capita income. The replication controls for both yet estimated a negative effect and statistically significant effect that is different from Sommers & Gruber who estimated a small, non-significant, positive effect.

Below are the results from a model with the same treated and control groups as Sommers & Gruber, a model without Ohio, a model without Connecticut, and then a model without either Ohio or Connecticut. The ATT was estimated for FY 2014 and FY 2015. The percent ATT was calculated by exponentiating the raw log ATT and then taking the difference from one. In addition, one of the models utilized a more general economic indicator - labor force participation rate - that was annualized in line with the fiscal year data. For example, the unemployment rate was annual running from Jan 1st through the end of the year, but the labor force participation rate was calculated to be annual July 1st through June 30th.

Table E.1: Replicating Sommers & Gruber - Effect Sensitivity to Ohio and Connecticut

Model	**Raw Log ATT**	**P-Value**	**Percent ATT**
Replicated Estimate	-0.079	0.041	-7.62%
No Ohio	-0.0859	0.028	-8.23%
No Connecticut	-0.053	0.073	-5.14%
Neither	-0.059	0.046	-5.72%
LFP - Neither	-0.069	0.018	-6.67%

CHAPTER 3

Measures for Quantifying the Sensitivity of Difference-in-Difference Estimates to Parallel Trends Violations

3.1 Introduction

One of the primary challenges of estimating a causal effect, when treatment assignment is not random, is controlling for all confounders. Confounders pose a threat to the unbiased estimation of a causal effect because of their relationship with both treatment assignment and the outcome, which could lead to a spurious result. A spurious result exists when the effect of an unobserved confounder acting on the treatment and the outcome was incorrectly determined to be the effect of treatment. For example, the decision of a state to expand Medicaid under the Affordable Care Act (ACA) in 2014, was likely driven by multiple variables such as economic conditions, political landscape, and state finances; which also may influence outcomes of interest such as state spending on Medicaid. In addition, longitudinal data is often used to estimate the causal effects of policies and programs. Causal analysis of longitudinal data is subject to confounding from both time-invariant and time-varying confounders. Therefore, to estimate a causal effect one needs to control for confounding from both sources.

Difference-in-difference is a commonly utilized methodology to estimate a causal effect in longitudinal data, especially in health policy (Ryan et al., 2015). To identify a causal effect difference-in-difference removes time-invariant confounding by differencing (estimating a difference between) pre- and post-treatment periods. A second difference is then taken between the treated difference and the control difference in order to remove time-varying confounding. The second difference, however, assumes that the control unit(s) experience all of the same time-varying confounders as the treated unit(s).

The assumption that the control unit experiences the same magnitude and type of time-varying confounding as the treated unit is known as the *parallel trends* assumption. If the underlying trends between treatment and control units are not in fact parallel, then the impact of unobserved confounders would be erroneously attributed to the treatment effect. Therefore, the identification of a causal effect estimated by difference-in-difference relies on the validity of the parallel trends assumption. For parallel trends to hold perfectly, the control group must experience the same magnitude and type of time-varying confounding as the treatment group.

Hypothesis tests are frequently used to indirectly evaluate the validity of the parallel trends assumption. Examples of indirect parallel trends tests include interrupted time-series, interactions of time dummies and a time-invariant treatment indicator, and comparison to a difference-in-difference-in-difference estimator (Courtemanche et al., 2017; Sommers and Gruber, 2017; Yue et al., 2018). A key concern with the indirect tests is that they only examine whether or not the parallel trends assumption is valid in the pre-period. For example, it is common for researchers to interact a time-invariant treatment indicator with time dummies and evaluate the significance of the coefficient for periods prior to treatment. This is problematic because there is no guarantee that if parallel trends holds in the periods just prior to treatment, it will hold through the post-treatment period.

An alternative approach to indirectly evaluating the parallel trends assumption is to estimate how strong the unobserved time-varying confounders need to be to send the estimated effects to zero, a type of sensitivity analysis. Sensitivity analysis estimates the magnitude of bias, due to hypothetical unobserved time-varying confounders, that would be required to change the conclusions of the study. For example, sensitivity analysis may find that the bias induced by a hypothetical unobserved confounder is minimal, and while it may change the effect estimate the conclusions drawn from estimate are not meaningfully changed. For example, if the effect estimate is a 10% increase then for the conclusions to change the effect estimate would need to be zero or negative. Conversely, an indirect hypothesis test of parallel trends may return a significant result and the authors would conclude that the parallel trends assumption, in the pre-period, is violated and the results are biased without further consideration as to how strong of a parallel trends violation is required to change the conclusion drawn from the results.

A suite of sensitivity analysis tools developed by Cinelli & Hazlett (2020), further referred to as CH, focuses on expanding the utility and implementation of sensitivity analysis for standard linear regression analyses(Cinelli and Hazlett, 2020). The authors sought to encourage a more nuanced discussion of how biased an estimated effect may be through the use of sensitivity analysis tools. Traditionally, discussions focus on whether or not unobserved confounders exist and place emphasis on the researcher controlling for all confounders. In much empirical work, however, it is challenging to control for all confounders, which is problematic when research is being evaluated on the presence of unobserved confounding. The tools developed by CH are intended to supplement the conversation of whether or not unobserved confounders exist by quantifying how strong hypothetical unobserved confounders need to be in order to change the conclusions drawn from a study. In other words, the sensitivity analysis tools are a way to judge the bias of an estimate in scenarios where it is difficult or impossible to control for all unobserved confounders.

In the motivating example used by CH, the authors were interested in the effect of violence on one's perspectives of peace and war. The study consisted of individuals who experienced violence at some point in the past and identified the causal effect under the strategy that allocation of violence was random, conditional on village and gender. A reviewer may critique the analysis under the hypothesis that an individual's distance to the center of the village was a critical unobserved confounder because individuals closer to the edge of the village had less time to react to an attack and therefore faced violence at a higher rate than those nearer the center. CH utilized their sensitivity tools to estimate the strength of a hypothetical unobserved confounder, or any combination of hypothetical confounders, required to send the estimated effect to zero and thereby change the conclusions of the study. The authors found that proximity to the center of the village would need to have a relationship with the treatment four times stronger than that of gender, an important observed confounder, to send the estimated effect to zero. Using the results from the sensitivity analysis, the discussion of the validity of the results focused on whether or not content experts think that the strength of location, or some other unobserved confounder, was four times that of gender. The approach encourages thoughtful contextual discussion that is assisted by sensitivity analysis.

The current paper contributes to the rapidly expanding difference-in-difference literature by extending the sensitivity tools developed by CH to the specific use case of difference-in-difference. In the context of difference-in-difference the sensitivity tools assert the validity of the parallel trends assumption by estimating how sensitive the results are to unobserved time-varying confounders. The tools have three valuable characteristics. First, the tools utilize the post-period data as well as pre-period data. For example, rather than only focusing on the pre-period, the sensitivity tools also use data in the transition from the pre-period to the post-period, which means the metrics can be estimated when there is only one pre-period. Second, the tools promote discussion, based on expert content knowledge, of the plausibility that a time-varying unobserved confounder(s) of a given strength, exist. Third, the method can be used to estimate how much stronger the unobserved time-varying confounder(s) need to be, relative to an observed time-varying confounder, to meaningfully change the conclusions of the study. Calculating the sensitivity of the results to an observed confounder of significant importance is valuable for contextualizing discussion of how hypothetical unobserved confounders may change the conclusions.

Section 3.2 reviews the Omitted Variable Bias (OVB) framework, discusses how CH utilize a reparamterization of OVB to develop their sensitivity analysis tools, and how we extend their tools to difference-in-difference. Section 3.3 discusses the methodology for simulating data with which to explore the characteristics and properties of the tools. Section 3.3.3 explores the properties and performance of the proposed sensitivity tools via simulation. Section 3.4 provides three applied examples of the sensitivity tools. Sections 3.5 and 3.6 provide discussion and conclude.

3.2 Background

The method of inquiry utilized by CH builds upon prior literature on omitted variable bias (OVB) as a tool for sensitivity analysis. The classical omitted variable bias framework begins with the following regression for the true data generating process:

$$Y = \hat{\tau}D + \hat{\beta}X + \hat{\gamma}Z + \hat{\epsilon}_{full} \tag{3.1}$$

where Y is the outcome, $\hat{\tau}$ is the ATT, D is a treatment indicator, X is an observed covariate, and Z is an unobserved covariate. However, because Z is unobserved the regression that is estimated is:

$$Y = \hat{\tau}_{res}D + \hat{\beta}_{res}X + \hat{\epsilon}_{res} \tag{3.2}$$

where $\hat{\tau}_{res}$, $\hat{\beta}_{res}$, and $\hat{\epsilon}_{res}$ are the coefficient estimates when Z is not included in the regression. The question is, how does $\hat{\tau}_{res}$ compare to $\hat{\tau}$? The classical OVB solution answers this as:

$$\hat{\tau}_{res} = \hat{\tau} + \hat{\gamma}\hat{\delta} \tag{3.3}$$

Where $\hat{\gamma}$ - broadly speaking - describes how the unobserved confounder Z *impacts* the the outcome, and $\hat{\delta}$ describes how *imbalanced* the treatment and control groups are on Z.

When $\hat{\gamma}$ is equal to zero - Z has no relationship with the outcome - then $\hat{\tau}_{res} = \hat{\tau}$; and the same holds when $\hat{\delta}$ is zero. However, when $\hat{\gamma}$ and $\hat{\delta}$ are non-zero then $\hat{\tau}_{res}$ is a biased estimator of $\hat{\tau}$ (Cinelli and Hazlett, 2020). Therefore, OVB estimates bias due to unobserved confounders as the product of the imbalance, with respect to treatment, and the impact on the outcome.

In econometrics the OVB framework is used to evaluate the bias that may be induced by unobserved confounders by using the hypothesized directionality of the partial correlation of the confounder with the outcome (impact or $\hat{\gamma}$) as well as with treatment (imbalance or $\hat{\delta}$). The classical OVB framework, however, has a number of limitations that reduce its functionality to simple scenarios with a single binary confounder. The main shortcoming is the difficulty of generalizing classical OVB to multiple confounders. For example, if there are confounders Z1, Z2, and Z3, combining their partial correlations with the outcome and with treatment becomes difficult because of the many ways in which these confounders may combine. Motivated by the issues with classical OVB, CH reformulate it using an R^2 parameterization to obtain (Cinelli and Hazlett, 2020):

$$|\hat{\text{bias}}| = se(\hat{\tau}_{res})\sqrt{\frac{R^2_{Y\sim Z|D,\boldsymbol{X}}R^2_{D\sim Z|\boldsymbol{X}}}{1-R^2_{D\sim Z|\boldsymbol{X}}}(\text{df})} \tag{3.4}$$

Where $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is the variance of Y that would be explained by a confounder Z after accounting for D, a binary indicator of treatment status, as well as all of the observed confounders, $\boldsymbol{X}$. $R^2_{D\sim Z|\boldsymbol{X}}$ is the variance of D that would be explained by a confounder Z after accounting for $\mathbf{X}$. $se(\hat{\tau}_{res})$ is the standard error of the estimated treatment effect from the observed – or restricted – model, and df is the degrees of freedom of the observed model.

The core idea in the R^2 parameteriziation remains the same: there exists some confounder – that has an effect on the outcome and on treatment take-up – and the interest is to understand how strong that confounder needs to be related with both treatment status and the outcome in order to make an observed treatment effect zero. The advantage of using an R^2 measure is that it is scale-free, easily interpretable, and doesn't require distributional assumptions on how to combine multiple confounders. Using Equation 3.4, CH proposed several values for sensitivity analysis with an emphasis on two: the robustness value (RV) and the extreme scenario value (EV).

3.2.1 The Robustness Value

The robustness value, which ranges from zero to one, tells the researcher the minimum association – as measured by variance explained – of a hypothetical unobserved confounder with the outcome and with treatment in order to make the estimated treatment effect zero. A robustness value near zero suggests that the estimated effect could be eliminated by confounders that explain a small percentage of the remaining variance in the treatment and the outcome. Conversely, a robustness value near one suggests that the estimated effect could be completely eliminated - relative to a threshold set my the researchers - by unobserved confounders that explain a large amount of the remaining variance in treatment and the outcome. As an example, consider the aforementioned bias = imbalance × importance formula; the robustness value is the scenario where imbalance = importance. The robustness value can be estimated by:

$$RV_q = \frac{1}{2}\sqrt{f_q^4 + 4f_q^2} - f_q^2 \tag{3.5}$$

Where $f_q = q|f_{Y\sim D|\boldsymbol{X}}| = \frac{q|\hat{\tau}_{res}|}{se(\hat{\tau}_{res})\sqrt{\text{df}}}$ is the partial Cohen's f and q is a proportion reduction in the treatment effect, of some value, which would be problematic to the conclusions drawn from the study. The value q that is most often of interest is when the estimated treatment effect is reduced by 100% or $q = 1$. Equation 3.5 is obtained by considering the scenario where unobserved confounders explain the same amount of the remaining variance in both the outcome and treatment status ($R^2_{Y\sim Z|D,\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}} = RV_q$) and then solving Equation 3.4 for RV_q (Cinelli and Hazlett, 2020).

3.2.2 The Extreme Scenario Value

The extreme scenario value considers the case where a hypothetical unobserved confounder(s) explain all of the remaining variance in the outcome ($R^2_{Y\sim Z|D,\boldsymbol{X}} = 1$), and estimates how much of the remaining variance in treatment status ($R^2_{D\sim Z|\boldsymbol{X}}$) they need to explain to send the estimated effect to zero. The extreme scenario value fixes the amount of remaining variance in the outcome explained by a confounder and seeks to determine what value of $R^2_{D\sim Z|\boldsymbol{X}}$ is required to send the estimated treatment effect to zero. The extreme scenario value can be obtained by first considering an alternative version of the bias formula derived by CH:

$$q|f_{Y\sim D|\boldsymbol{X}}| = |R_{Y\sim Z|D,\boldsymbol{X}} \times f_{D\sim Z|\boldsymbol{X}}| \tag{3.6}$$

The above formula is obtained by substituting $f^2_{D\sim Z|\boldsymbol{X}} = \frac{R^2_{D\sim Z|\boldsymbol{X}}}{1-R^2_{D\sim Z|\boldsymbol{X}}}$ into the bias formula and simplifying. When $R^2_{Y\sim Z|D,\boldsymbol{X}} = 1$ and $q = 1$ the above equation simplifies to $f^2_{Y\sim D|\boldsymbol{X}} = f^2_{D\sim Z|\boldsymbol{X}}$. A confounder that explains all of the remaining variance in the outcome would need to satisfy $f^2_{Y\sim D|\boldsymbol{X}} = f^2_{D\sim Z|\boldsymbol{X}}$ which implies that $R^2_{Y\sim D|\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}}$. Therefore, the amount of variance in the outcome explained by the treatment after accounting for the covariates, $R^2_{Y\sim D|\boldsymbol{X}}$, is the extreme scenario value.

3.2.3 The Relative Confounding Strength Value

Equation 3.4 also lends itself to a useful tool, the relative confounding strength (RCS) value. The RCS value considers how the estimated effect would change if there were unobserved confounders K times stronger than an observed covariate, of the researchers choosing. The RCS value begins by calculating how strongly related an observed confounder is with the outcome and with the treatment, and those values are then used as hypothetical stand-ins for an unobserved confounder. For example, a researcher may argue that they have controlled for one of the most important confounders, Z, and using the RCS value they estimate how the observed effect sized changes if a hypothetical unobserved confounder that is K times stronger than Z was included in the model. The researcher can report how the estimated effect changes for different K and may find that when $K = k$ the estimated effect is reduced to a point such that the conclusions drawn from the study change.

3.2.4 Extending Cinelli & Hazlett to Difference-in-Difference

The sensitivity tools developed by CH are a suite of general tools for regression. This paper is interested in utilizing them for the specific use case of assessing the presence of time-varying confounding in difference-in-difference. To calculate the robustness and extreme scenario values requires the effect estimate, the standard error, and the degrees of freedom. Utilization of the tools, however, is complicated by the common use of cluster-adjusted standard errors in difference-in-difference models because the bias formula (Equation 3.4) derived by CH utilizes the unadjusted standard error. The proposed solution is to utilize a first-difference model such that cluster-adjusted standard errors are no longer required.

The first-difference model is estimated as:

$$\Delta y_{it} = \alpha_0 + \tau \Delta D_{it} + \Delta \epsilon_{it}, t = 2, 3, ..., T \tag{3.7}$$

Where ΔD_i is the change in a time-varying treatment status and τ is the effect of interest. A first-difference estimator is used because if the error term ϵ_{it} follows a random walk then the usual standard errors are asymptotically valid (Wooldridge, 2003). With the usual OLS standard errors in hand, the robustness and extreme scenario values are easily estimated either through an online tool or the R or Stata package, *sensemakr*. Alternatively, if the cluster-robust standard errors are similar to the usual OLS standard errors then the sensitivity tools are also estimable without further modification to the regression.

3.3 Simulation

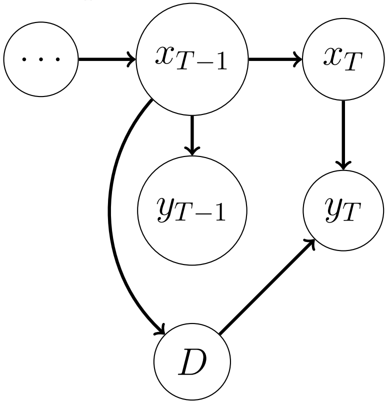
The purpose of the simulation analysis is to quantify the functionality of the sensitivity tools with respect to parallel trends, the violation of which will be referred to as time-varying confounding. Specifically, how do the sensitivity tools developed by CH, behave when time-varying confounding is present compared to when it is not present, and as the magnitude of the time-varying confounding, with respect to the outcome, varies.

The behavior of the sensitivity tools under varying parameters was explored using simulated data. The simulation consisted of two parts: an outcome model and a confounded treatment assignment model. Confounding was induced by using the same covariate in both the outcome and treatment assignment models.

3.3.1 Data Generating Process

Data was generated based on the following directed acyclic graph (DAG) with treatment occurring at time $t = T$:

Figure 3.1: Graphical Data Generating Process



Which was implemented using the following structural causal model:

$$D_i = Binom(\pi = \frac{1}{(1 + exp(-2x_{i,T-1}))}) \tag{3.8}$$

$$Y_{it} = \beta X_{it} + 10I(D_i = 1, t = T) + \epsilon_{it} \tag{3.9}$$

where i represents each unit, t is an indicator of the time period with T denoting the time-period just prior to treatment, D is a time-invariant indicator of treatment, and Y is the outcome. D_i is drawn from a binomial distribution where the probability for unit i is a log-link function of $x_{i,T-1}$ - the X values just prior to treatment - with X being drawn from a multivariate standard normal distribution with serial correlation of 0.6,0.4,0.2,0.1. Y_{it} is a function of the confounder X_{it} and its coefficient β, a treatment effect - fixed at 10 - and an indicator of whether or not a unit is treated and in the post-period, and random noise. For the Monte Carlo simulation β takes the values of 0,5,10,20,40.

3.3.2 Estimation

After simulating the data, a first-difference model was estimated following Equation 3.10, and the results were used to estimate the robustness value and the extreme scenario value.

$$\Delta Y_i = \tau \Delta D_i + \Delta \epsilon_i \tag{3.10}$$

With the model fit, the *sensemakr* R package – developed by Cinelli and Hazlett (2020) – was used to estimate the robustness value and the extreme scenario value. *Sensemakr* estimates the robustness value according to Equation 3.5 which requires calculation of the partial Cohen's f ($f = \frac{|t_{res}|}{\sqrt{df}}$) for the treatment effect. Similar to the robustness value, *sensemakr* estimates the extreme scenario value as the ratio of the squared t-statistic to the squared t-statistic plus the degrees of freedom. The results were recorded and then the models were re-estimated for the next set of parameters. The process was repeated 1000 times for each parameter set.

3.3.3 Simulation Results

Table 3.1 reports the robustness value and extreme scenario value with varying strength of an unobserved time-varying confounder and by two instances of treatment assignment: random (confounding strength equals zero) and time-varying confounded. The results below are only for illustrative purposes and not to create a rule comparing the strength of unobserved confounding and the RV and EV estimates.

The robustness value for the point estimate is interpreted as follows: an unobserved confounder or confounders explaining at least RV% of the residual variance of both the treatment and the outcome would be enough to decrease the effect size by Q percent. Here Q is a value, determined by the researcher, at which the estimated effect is negligible. For example, Q is often set to 100% but in some context, such as clinical measures, a non-zero value may be regarded as being a negligible effect.

The EV value is interpreted as follows: if a confounder or set of confounders explained 100% of the residual variance of the outcome, they would need to explain EV% of the residual variance of treatment to reduce the estimated effect by Q percent. For example, after estimating a regression model if there exists some confounder or set of confounders Z, that explains 100% of the remaining variance in the outcome, then Z would have to explain EV% of the remaining variance in treatment to send the estimated effect to the negligible value, Q.

Table 3.1: Sensitivity Analysis Tool Results for Increasing Values of Confounding

Confounding Strength	**Robustness Value (RV)**	**Extreme Scenario Value (EV)**
0	86%	84%
5	54%	40%
10	37%	19%
20	25%	8%
40	18%	5%

Note: Results based on a simple structural causal model with a treatment, D, an outcome, Y, and a single confounder, Z. The reported results are from 1000 simulations using the aforementioned model.

For increasing magnitudes of the parallel trends violation, RV and EV decreased rapidly but at differential rates. In this example, when the strength of the unobserved confounder with the outcome was 20, or double the treatment effect, the robustness value reported that there would need to be an unobserved confounder that explains at least 25% of the residual variance of both the treatment and the outcome to send the estimated effect to zero. The extreme scenario value reported that if there was an unobserved confounder or set of confounders that explained 100% of the remaining variance of the outcome it would need to explain only 8% of the remaining variance of treatment to reduce the estimated effect to zero. Interestingly, the RV and EV values did not achieve values of 100% even when the confounding strength was equal to zero. This is due to the inclusion of random error in the data generating process. As the random error goes to zero we expect the RV and EV values to increase - assuming no confounding - because the observed covariates would explain all of the variance in the outcome.

The difference between the robustness value and extreme scenario value for increasing confounding strength in Table 3.1 was expected and is explainable by considering a more general form of what the two values are estimating. Recall the general bias formula, bias = imbalance × impact, where if one were to fix the bias and increase impact, the imbalance would have to decrease. Keeping said relationship in mind, the robustness value is the scenario where the imbalance is equivalent to the impact. Then, holding the bias fixed, the impact is increased to some maximum and as a result the imbalance required to achieve the set amount of bias must decrease, which is the extreme scenario value. To put it into the same language utilized in this paper, if we fix the bias to be the amount required to send the estimated effect to zero then as we increase the impact (association between the confounder and the outcome) it is required for the imbalance (association between the confounder and the treatment) to decrease.

3.4 Sensitivity Tools in Practice

Further exploration of the proposed sensitivity tools was done using examples from academic papers. Papers by Buchmueller and Carey (2018) and by Sommers and Gruber (2017) are used as examples to demonstrate how to use the tools in practice and the reanalyses are not intended to contest or challenge the authors or their conclusions. The first two examples reanalyze prior work on the effects of Prescription Drug Management Databases on Opioid related outcomes and on the effects of Medicaid expansion on State Medicaid spending.

3.4.1 Prescription Drug Monitoring Programs

Recent work by Buchmueller and Carey (2018) examined the effect of Prescription Drug Monitoring Programs (PDMP) on several opioid and adverse health related outcomes stratified by whether or not healthcare providers are required to access the PDMP prior to prescribing opioids. To identify the effect, they used a difference-in-difference framework where states without a PDMP were controls. Therefore, the authors assumed that the outcome trends through the post-treatment period for the control group are parallel to the unobserved counterfactual for the treated group, or that the time-varying confounding experienced by the treated units is experienced equally by the control units.

The results of the authors analysis are sensitive to time-varying confounding or the degree to which the parallel trends assumption holds, which the authors evaluated using an event study paradigm. From the binary interpretation standpoint, the tests concluded that the parallel trends assumption, for the pre-period, could not be rejected. However, sensitivity analysis provides a more nuanced approach to understanding how strong of a parallel trends violation the estimates can handle before the conclusions change. Therefore, the analysis is an excellent example of a scenario where the sensitivity analyses proposed in this paper are useful.

The purpose of this example is to estimate the sensitivity of the estimates of the effect of PDMP with "Must Access" provisions on a number of opioid outcomes. The sensitivity estimates in this paper are not for the original effect sizes but for effect sizes using updated treatment timings. The updated treatment timings are from recent work by Horwitz et al. (2018) that developed new treatment timings based on a detailed comparison of legislated policy implementation dates with the reported date PDMPs were usable. After estimating the effect sizes with updated treatment timings the R package *sensemakr* was used to calculate the RV and EV metrics proposed in this paper using the estimated effect, standard error, and degrees of freedom.

The outcomes analyzed were measures of opioid use and misuse constructed by Buchmueller and Carey (2018) from Medicare Part D claims data. There were a total of eight outcomes grouped into three general categories: quantity-based outcomes, "shopping" outcomes, and medical service outcomes. The quantity-based outcomes included the share of individuals taking any opioids (P(Taking Opiods), the share of individuals who received more than a 211 day supply of opioids (211+ Daily Supply), the mean daily morphine equivalence dose (120+ MED), and the share of individuals with multiple concurrent prescriptions for the same opioid drug (overlapping claims). The "shopping" outcomes included the proportion of individuals who received an opioid prescription from five or more providers and the proportion of individuals who filled a prescription at five or more pharmacies. The medical service outcomes - Excess Out-of-State (OOS) Pharmacies and Excess OOS Providers - compared, for a given individual, the proportion of opioid prescriptions either filled or received out of state with their non-opioid prescriptions either filled or received out of state. Each outcome was modeled using a two-way fixed effects model with no covariates, similar to Buchmueller and Carey (2018):

$$Y_{it} = \mu_i + \lambda_t + \tau D_{it} + \epsilon_{it} \tag{3.11}$$

Table 3.2 summarizes the results from the additional sensitivity analysis of the effects of the must-issue PDMP laws (using the estimation technique of the authors but with updated timing) to unobserved time-varying confounders by reporting the RV and EV estimates.

Table 3.2: Sensitivity Results for the Must-Issue PDMP Laws

Outcome	**RV**	**EV**
Taking Opiods	6.76%	0.49%
211 Days Supply	5.76%	0.35%
120+ Daily Med	2.14%	0.05%
Overlapping Claims	12.30%	1.70%
5+ Prescribers	7.70%	0.64%
5+ Pharmacies	13.21%	1.97%
Excess OOS Prescribers	7.39%	0.59%
Excess OOS Pharmacies	1.55%	0.02%

The robustness values (RV) ranged between 1.55% and 13.21% and are interpreted as the minimum percent of remaining variance that needs to be explained for both the treatment and outcome to send the estimated effect to zero. For example, the RV for overlapping claims is 12.30%, which is interpreted as: unobserved confounders that explain at least 12.23% of the remaining variance of both treatment and the outcome reduce the effect size to zero. The extreme value estimates (EV) vary slightly between 0.02% and 1.97%. The EV is interpreted as follows: if an unobserved confounder(s) explained all of the remaining variance of the "120+ Daily Med" outcome, they would only need to explain 0.05% of the remaining variance in treatment. Therefore, the EV metric considers the extreme scenario where if a confounder or confounders explained all the remaining variance in the outcome how much variance in the treatment would they need to explain, after accounting for the observed covariates.

The EV and RV measures for the PDMP model, which had no control variables, reflect the values they take when there is a large amount of unobserved confounding. The PDMP model was likely confounded by time-varying characteristics such as state financial resources, state ideology, and the perceived severity of the problem. For example, a state with limited resources may be unwilling to take-up treatment and be unable to commit resources to alternative programs that tackle opioid related issues.

The RV and EV measures were very small which indicates the effect estimates were very sensitive to time-varying confounding. A more intuitive way of interpreting the RV and EV estimates is to anchor them relative to observed confounders. Contextualizing the results with regards to observed confounders, however, is difficult because the data used by Buchmueller and Carey (2018) did not include time-varying confounders. The next example does observe two time-varying confounders and provides an example of how to utilize observed confounders to better understand the RV and EV estimates and presents the RCS estimate.

3.4.2 Effect of Medicaid Expansion on State Spending on Medicaid

A 2017 paper assessed the impact of Medicaid expansion on state Medicaid spending among states who expanded between FY 2014 and 2015 (Sommers and Gruber, 2017). To identify the effect, they utilized difference-in-difference controlling for the unemployment rate and state per-capita revenue. They tested the parallel trends assumption, in the pre-period, and found generally non-significant results which were used to support the validity of difference-in-difference. The paper concluded that Medicaid expansion had no effect on state Medicaid spending when they did not find significant results.

In this example, the sensitivity tools – robustness value, extreme scenario value, and relative confounding strength – are applied to the analysis by the authors for the total state Medicaid spending outcome. The example reanalyzed the original data and then estimated the sensitivity analysis tools. Following reanalysis, the example made a number of adjustments to the model and estimated the sensitivity analysis tools for each adjustment.

The original model was a two-way fixed effects model controlling for the unemployment rate and per-capita income. Table 3.3 reanalyzes the original model and makes three additional adjustments. The first adjustment was to remove Ohio and Connecticut from the data because of reporting issues (Potamianos et al., 2018). Second, the annual unemployment rate was replaced with the labor force participation rate (LFP). The LFP is a more holistic view of the job market because it includes discouraged workers who have given up looking for a job, whereas the unemployment rate only considers those who are employed or actively searching for a job. Third, a measure of state ideology was added to the model because Medicaid expansion was a divisive political issue and the states political ideology likely influenced both the decision to expand Medicaid as well as how the state allocates funds to the Medicaid program. Following each modification, the sensitivity metrics proposed in this paper are estimated using the *sensemakr* R package.

Table 3.3: Modification & Sensitivity Analysis of Sommers & Gruber (2018)

Model	**Effect**	**P-Value**	**RV**	**EV**
Original	-7.33%	0.0384	12.52%	1.76%
No Ohio or Connecticut	-5.59%	0.0427	12.51%	1.76%
LFP + No Ohio or Connecticut	-6.56%	0.0169	14.60%	2.43%
LFP + Ideology + No OH or CT	-6.58%	0.0147	14.92%	2.55%

Table 3.3 reports the treatment effect (percent change), the p-value, and the sensitivity

tools. The sensitivity tools were calculated using the estimated effect, standard error, and degrees of freedom from the two-way fixed effects model because the standard errors with clustering adjustment were similar to the unadjusted standard errors. The robustness values increased from 12.52% to 14.92%, and the extreme scenario values increased from 1.76% to 2.55%. These results can be contextualized by comparing them to the partial-R^2 of an important observed confounder such as the LFP. The partial-R^2 of the LFP with the outcome is 0.059 or 5.9%, and the partial-R^2 of state ideology with the outcome is 0.021 or 2.1%. If a partial-R^2 of 5.9% is typical for the kinds of state-level confounders that cause concern about the validity of difference-in-difference analysis of policy effects, then the concern may be assuaged. Of course, other contextual confounders may have very different partial-R^2 values, but this exercise at least helps to anchor expectations.

When the model controls for confounders the relative confounding strength (RCS) value can be estimated, which calculates how much the effect changes if there were an unobserved confounder K times stronger than a chosen observed confounder. For example, say that we believe the LFP is the confounder with the strongest relationship with the outcome and with treatment. The sensitivity framework developed by CH can be used to estimate how much the effect would change if an unobserved confounder existed that was K times stronger than the LFP. The estimation can be repeated across multiple values of K to understand the rate at which the effect changes as the strength of hypothetical confounder(s) increases relative to an observed confounder.

Table 3.4: Strength of Unobserved Confounder(s) Relative to Labor Force Participation Rate and Effect on Estimates

Strength of Unobserved Time-Varying Confounder(s)	**Adjusted Estimate**	**Adjusted SE**	**Adjusted T**	**Adjusted P-Value**
2x Labor Force Participation Rate	-5.60%	0.026	-2.2188	0.014
4x Labor Force Participation Rate	-4.49%	0.024	-1.9075	0.029
6x Labor Force Participation Rate	-3.34%	0.0219	-1.5524	0.061
8x Labor Force Participation Rate	-2.19%	0.0196	-1.1325	0.130
10x Labor Force Participation Rate	-1.01%	0.0168	-0.6051	0.272
12x Labor Force Participation Rate	0.19%	0.0135	0.1395	0.445

Table 3.4 reports the RCS estimates on how the estimated treatment effect, standard error, and t-value change as the strength of a hypothetical unobserved time-varying confounder(s) increases, relative to the LFP. For example, when the strength of a hypothetical unobserved time-varying confounder(s) was four times stronger than the LFP the estimated effect was no longer significant at the 0.05 level and decreased from the original value of -6.70% to -4.49%. Furthermore, when the strength of unobserved time-varying confounder(s) was 12 times stronger than the LFP the estimated effect was near zero as it decreased to 0.19%.

With these estimates in hand a more nuanced discussion can unfold regarding the presence of time-varying confounders. A reviewer may critique the paper for leaving out confounder Z. In the past, the author is limited in their ability to respond to the reviewer and must rely on arguments for why the unobserved confounder is likely of little import. With the RV, EV, and RCS estimates, however, the researcher can allow for confounder Z to be a plausible confounder and instead focus on whether or not confounder Z is likely strong enough to change the conclusions of the study. For the above example say that a reviewer posits the state immigration rate as an unobserved time-varying confounder. With the RV, EV, and RCS estimates the researcher can respond by saying that state immigration rate would have to be ten times stronger than the LFP to change the conclusions of the study. Furthermore, the state immigration rate would need to have a partial-R^2 of at least 14.92% with both the outcome and the treatment to send the effect to zero, which seems unlikely given that LFP has the highest observed partial-R^2 of 5.69%.

3.4.3 Effect of Medicaid Expansion on Medicaid Enrollment

In 2014 the Affordable Care Act gave states the option of expanding Medicaid coverage to childless adults with incomes below 138% of the Federal Poverty Line (FPL). Following expansion of Medicaid, a number of studies were published investigating the effect of Medicaid expansion on Medicaid enrollment among low-income adults. Despite variation in the data used to answer the aforementioned inquiry, each of the studies utilized difference-in-difference as the identification strategy to estimate the effect of interest. In order to justify the unbiasedness of difference-in-difference the studies visualized and tested for the presence of parallel trends in the pre-period and found evidence in favor of the parallel trends in the pre-period.

For this example, we conducted an analysis similar to prior work but used American Community Survey data on adults with incomes below 150% of FPL (Kaestner et al., 2017; Wherry and Miller, 2016). A first-difference model was used to estimate the effect and the sensitivity of the results to a parallel trends violation, or time-varying confounding, were explored using the robustness value and the extreme scenario value.

Table 3.5: Robustness and Extreme Scenario Values for Effect of Medicaid Expansion on Medicaid Enrollment

Model	**Effect**	**P-Value**	**RV**	**EV**
Naive	7.64	< 0.001	55.34%	40.68%
Naive + LFP	7.69	< 0.001	55.12%	40.37%
Naive + LFP + Ideology	7.70	< 0.001	57.45%	43.69%

The analysis found an effect of 7.64 percentage points with no control variables, which is in line with prior work by Wherry and Miller (2016) and Decker et al. (2017), who found percentage point increases of 10.5 and 7.3. The robustness and extreme scenario values for the analysis were 55.34% and 40.68% (Table 3.5). The initial analysis, however, was naive in that it ignored confounders such as labor force participation (LFP) and state political ideology. Thus, it is plausible that the observed effect is partially due to changes in LFP and state ideology over time. Controlling for the labor force participation rate changed the effect estimate to 7.69 percentage points, slightly increased the robustness value to 55.12% and the extreme scenario value to 40.37% (Table 3.5). Controlling for state ideology slightly changed the effect estimate to 7.70 percentage points and increased the RV and EV values to 57.45% and 43.69%.

Table 3.6: Strength of Unobserved Confounder(s) Relative to Labor Force Participation Rate and Effect on Estimates

Strength of Unobserved Time-Varying Confounder(s)	**Adjusted Estimate**	**Adjusted SE**	**Adjusted T**
1x State Ideology	7.49	0.627	11.948
2x State Ideology	7.29	0.600	12.154
5x State Ideology	6.67	0.506	13.175
10x State Ideology	5.61	0.273	20.543

In table 3.6 we estimate the RCS for an hypothetical unobserved confounder that is 1, 2, 5, and 10 times stronger than state ideology and find that the effect is robust because it remains significant despite hypothetical unobserved confounders of increasing strength. Interestingly, the treatment effect becomes more significant as the strength of the hypothetical confounder increases. The increasing precision of the effect is because state ideology has a very large partial-R^2 with the treatment and a small partial-R^2 with the outcome. Therefore, controlling for a hypothetical unobserved confounder that is strongly related to the treatment and negligibly related to the outcome would simultaneously increase the precision of the estimate and reduce the strength of the estimate.

A researcher might consider both state ideology and the LFP to be very important confounders that should be analyzed using the RCS values. The *sensemakr* package provides plots that are useful for summarizing the RCS values of multiple confounders. An example is provided below that visualizes the RCS values for both state ideology and the LFP 3.2

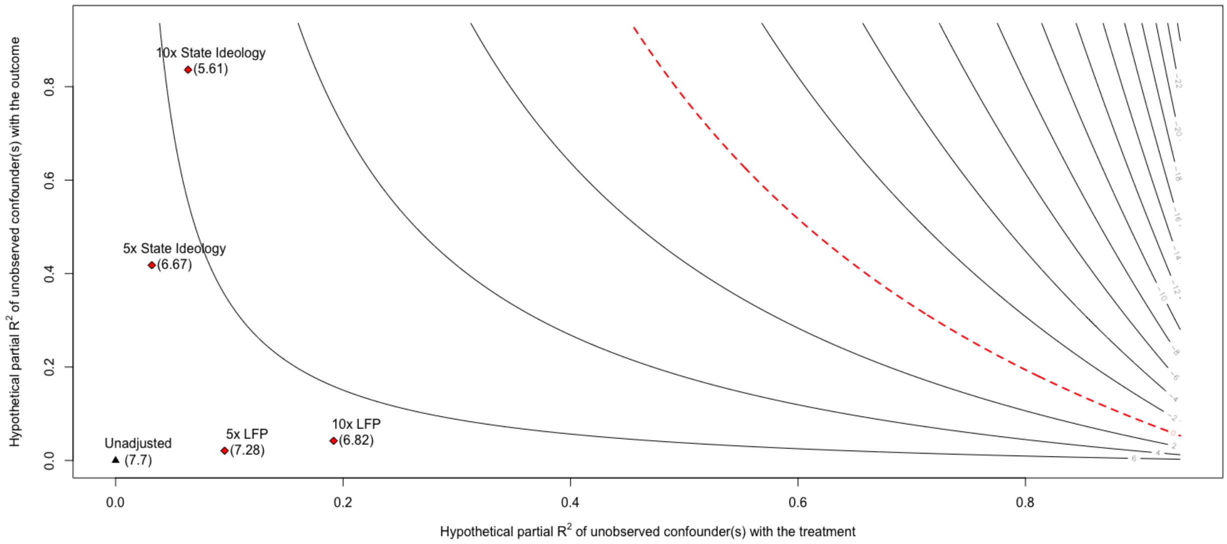


Figure 3.2: Visualizing Relative Confounding Strength

In the above contour plot each line represents the estimated effect for different values of hypothetical confounding, as measured by partial-R^2. The critical value, when the effect is zero, is denoted by the dashed red line. The observed effect size is denoted by the black triangle in the bottom left corner when the hypothetical confounder(s) has zero relationship with either the outcome or treatment. The key elements of the plot are the red diamonds which represent the effect size given a hypothetical unobserved confounder K times stronger than the listed observed confounder. For example, when considering a confounder five times stronger than the LFP the estimated effect is 7.28, and when the hypothetical confounder is ten times stronger than the LFP the estimated effect is 6.82.

We argue that this application may provide an approximation for the real world upper limit of EV and RV because confounding is minimal. Plausible confounders are state finances, the labor force participation rate, and state ideology. State finances, however, likely have little influence on treatment uptake because the federal government covered 100% of the new costs for the first two years and 90% of the costs after. Labor force participation does influence enrollment and states may consider labor force participation prior to expanding Medicaid. State ideology is likely strongly related to the decision to expand Medicaid due to the political decisiveness surrounding the topic, but is questionably related to the outcome.

3.5 Discussion

The range of examples discussed above provide a starting point for evaluating the continuum of values the EV and RV measures may take. The Medicaid expansion example (RV = 57%, EV = 44%) likely provides us with a feasible upper limit for the robustness and extreme scenario values in real-world applications. The PDMP example (RV 5%, EV 1%) captures the range of values for which we would be seriously concerned about in a real-world application. The second example represents results from a study where interpretation of the EV and RV is difficult, but presents the RCS values and partial R^2 for observed confounders as ways to contextualize the results.

For the examples with control variables, the RCS values were a very useful tool because they were readily interpreted within the context of the study by anchoring the sensitivity measures against a known confounder. In the state Medicaid spending example it was estimated that a hypothetical unobserved confounder(s) 12 times stronger than the labor force participation rate would send the estimated effect to near zero. In the Medicaid enrollment example a hypothetical unobserved confounder(s) would need to be more than ten times stronger than either state ideology or LFP to change the conclusions of the study.

We recommend that the sensitivity metrics are used to evaluate the strength of the evidence, on a continuum, for the conclusions drawn from a difference-in-difference model. While one may be tempted to create a threshold for the EV, RV, or RCS measures, doing so would defeat the purpose and intent of the sensitivity tools. The sensitivity tools should be used to understand how robust the effect estimate is to unobserved confounding and to promote discussion as to whether or not confounder(s) exist with enough strength to change the conclusions of the study.

The best utilization of the tools will come from reporting all three and thoroughly discussing whether or not unobserved confounders are strong enough to change the conclusions of the study. To this end, it may be helpful to report the partial-R^2 for each observed confounder to contextualize the results. For example the labor force participation rate in example two had the highest partial-R^2 at 5.89%. The other observed confounders in the model had much smaller partial-R^2s compared to the labor force participation rate. Therefore, it would need to be argued that the unobserved confounders would not only have to be more important than the majority of observed confounders, but also be 12 times more important than the confounder with the highest partial-R^2.

We also strongly recommend against using the tools as a measure on which to select different models. Using the proposed tools for variable selection would likely result in a biased effect because the tools are not identifying confounders and controlling for them; expert content knowledge is required to properly identify confounders. Models should be selected based on an identification strategy that seeks to make treatment assignment independent of the outcome. In the second example, we did not select our model based on the EV and RV, but instead sought to highlight how the EV and RV change as we make purposeful modifications to the model based on a causal structure.

When the results from the EV, RV, and RCS measures are ambiguous and the partial-R^2s from observed covariates are similar, interpretation can be difficult. An interesting companion to the proposed sensitivity tools is to estimate a synthetic control counterfactual and compare it with the difference-in-difference counterfactual. Then, a visual comparison of a synthetic control with difference-in-difference may be insightful for further evaluation of the validity of difference-in-difference. For example, the synthetic control may align with the difference-in-difference counterfactual, and the validity of difference-in-difference is supported. This approach requires further research but may be another useful sensitivity tool.

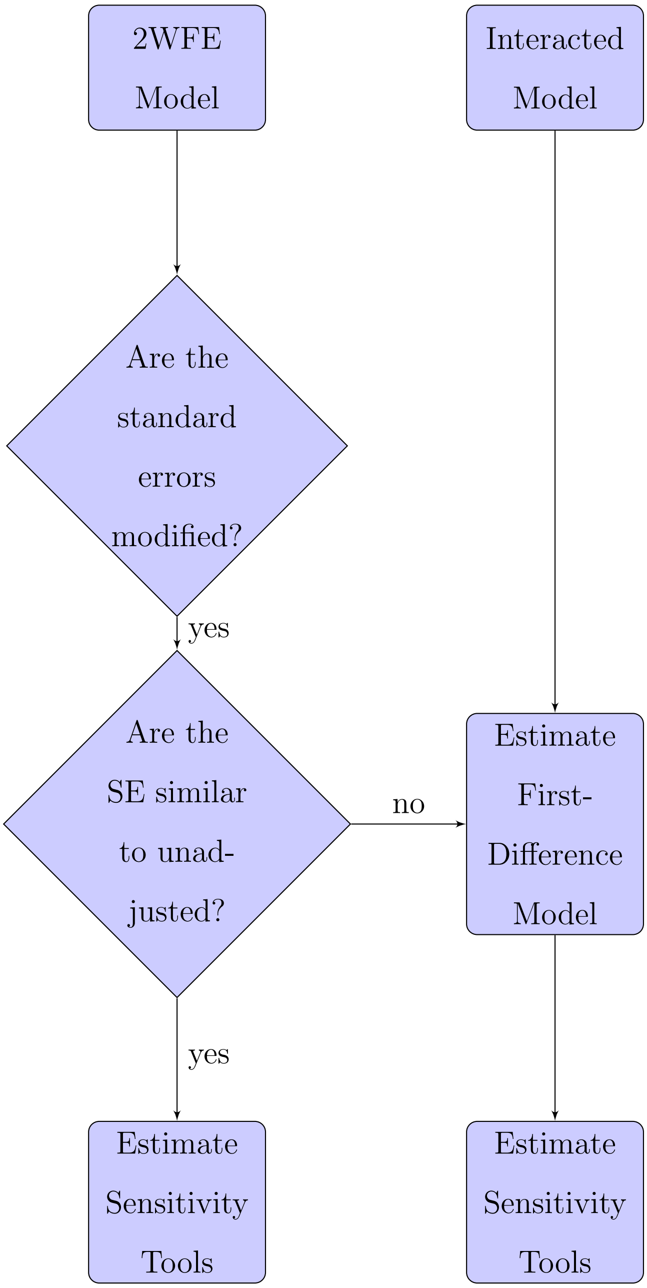
3.6 Conclusion

Difference-in-difference relies on the assumption that no time-varying confounding remains in order for the causal effect to be identified, known as the parallel trends assumption. Unfortunately, in health policy it is often infeasible to measure every confounder. Traditional methods for evaluating the feasibility of the parallel trends assumption are limited because they only evaluate whether or not parallel trends holds in the periods prior to treatment and make an additional assumption that the pre-period trends continue into the post-treatment period. In addition, indirect parallel trends tests result in binary conclusions that ignore the relationship between the magnitude of a parallel trends violation and the estimated effect. This paper contributes to the difference-in-difference literature by proposing a sensitivity analysis toolkit that analyzes how sensitive a causal estimate from difference-in-difference is to unobserved time-varying confounding, or parallel trends violations, on a continuum.

The sensitivity tools are a valuable addition to the social scientist's toolbox when making causal inferences with difference-in-difference. The toolkit can be easily implemented in R or Stata using the *sensemakR* package. The robustness, extreme scenario, and relative confounding values can be easily estimated by using the estimated effect, standard error, and degrees of freedom from the model. The tools quantify how much hypothetical unobserved confounding the estimated effect can handle before the conclusions drawn from the estimate change, and contextualize the sensitivity values based on observed confounders. Combined with expert content knowledge, we hope the sensitivity toolkit can create a more nuanced discussion between researchers about how unobserved confounding impacts causal results by providing quantitative guidelines for how sensitive causal effects from difference-in-difference are to unobserved confounding.

Appendix F

Utilization of the Tools - A Flow Chart


2WFE Model
Interacted Model
Are the standard errors modified?
yes
Are the SE similar to unadjusted?
no
Estimate First-Difference Model
yes
Estimate Sensitivity Tools
Estimate Sensitivity Tools

Appendix G

Example: Sensitivity Tools and Two-Way Fixed Effects Model Using R Software

```
# Model
fit1 = lm(log(mcaid_gfof) ~ Post + lfp + pc_inc + inst6017_nom + state +
    as.factor(year), data = d[d$year <= 2015 & !(d$state %in%
    c("Ohio","Connecticut")),])

# Record Standard Error for Treatment Effect
se = summary(fit1)[[4]][2,2]

# Estimate & Record Cluster Robust Standard Errors
fit1a = coeftest(fit1, vcov = vcovHC(fit1))
se.robust = fit1a[2,2]

# Compare, Find that they are similar. Estimate RV, EV, and RCS.
  se; se.robust

# Estimate Sensitivity Tools
  sens.fit1 = sensemakr(fit1 ,treatment = "Post",
                        benchmark_covariates = "lfp",
                        kd = c(2,4,6,8,10,12),)
  summary(sens.fit1)
  # Changing Bound Labels for Plot
  sens.fit1$bounds$bound_label = c("2x LFP","4x LFP","6x LFP","8x LFP","10x
      LFP","12x LFP")

# Contour Plot of Relative Confounding Values
  plot(sens.fit1)
```

Appendix H

Example: Sensitivity Tools and First Differenced Model Using R Software

```r
# Manually Create First-Differenced Dataset using plm Package
# d is a long dataset with State-Year rows

  # Defining P-Series
  d = pdata.frame(d,index = c("State","Year"))
  # Lagging Variables
  d$lag.outcome = plm::lag(d$outcome,k = 1,shift="time")
  d$lag.Post = plm::lag(d$Post,k = 1,shift="time")
  d$lag.LFP = plm::lag(d$LFP,k = 1,shift="time")

  # Calculating First-Difference
  d = d[complete.cases(d),]
  d$fd.outcome = d$outcome - d$lag.outcome
  d$fd.Post = d$Post - d$lag.Post
  d$fd.LFP = d$LFP - d$lag.LFP

  # Modeling
  fit1 = lm(fd.outcome ~ fd.Post + fd.LFP, data = d)

  # Sensitivity Analysis
```

```
fit1.s = sensemakr(model = fit1, treatment = "fd.Post", benchmark_covariates
   = "fd.LFP",kd = c(1,2,5,10))

# Contour Plot
fit1.s$bounds$bounds_label = c("1x LFP","2x LFP","5x LFP","10x LFP")
plot(fit1.s
```

Appendix I

Indirect Parallel Trends Tests & Statistical Power

Table I.1 presents the results of the Monte Carlo simulation that explored the ability of the indirect parallel trends test to correctly identify a parallel trends violation in the pre-period. It reports the proportion of times the indirect parallel trends test identified non-parallel trends in the pre-treatment period for a given sample size and pre-period year. For example, the columns indicate whether or not the parallel trends assumption held in the data generating process and the pre-period year, prior to treatment. The rows indicate the sample size being used for the given simulation run. The elements within the table were calculated as the percent of simulations where the classical indirect parallel trends test rejected the null hypothesis. Therefore, a value of 40% would be interpreted as follows: For a given sample, when parallel trends doesn't hold in a given pre-period, the test rejects the null hypothesis 40 percent of the time.

Table I.1: Probability of Rejecting Null Hypothesis - Classic Parallel Trends Testing

Sample Size	**Not Confounded**			**Confounded**		
	Year: T-3	Year: T-2	Year: T-1	Year: T-3	Year: T-2	Year: T-1
20	4%	6%	14%	2%	3%	7%
50	2%	10%	15%	6%	23%	53%
100	2%	9%	15%	9%	41%	47%
200	2%	8%	12%	43%	84%	80%

Table I.1 shows that as the sample size increases the probability of correctly rejecting the null hypothesis when parallel trends doesn't hold in the pre-period increases. This result was expected given the structure of the hypothesis test. The hypothesis test is structured such that the power of the test to reject the null hypothesis governs the probability with which we incorrectly conclude that parallel trends holds in the pre-period, rather than the alpha value. As a result, the probability of correctly rejecting the null hypothesis, parallel trends holds in the pre-period, increased with the sample size.

For comparison, the results of the sensitivity analysis tools were estimated and presented alongside the classical indirect parallel trends test in Table I.2. The sensitivity analysis tools were generally invariant to the sample size because sample size is not directly included in the metrics. Instead, the sensitivity tools are more responsive to changes in the precision of the estimate as well as the magnitude of the estimated treatment effect.

Table I.2: Sensitivity Analysis Tool Results for Increasing Sample Size

Sample Size	**Not Confounded**		**Confounded**	
	RV	EV	RV	EV
20	45%	28%	46%	29%
50	40%	22%	41%	23%
100	43%	25%	46%	28%
200	42%	24%	45%	26%

Bibliography

Alberto Abadie. Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, page 44, August 2019.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505, June 2010. ISSN 0162-1459. doi: 10.1198/jasa.2009.ap08746. URL https://amstat.tandfonline.com/doi/abs/10.1198/jasa.2009.ap08746.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synth: An R Package for Synthetic Control Methods in Comparative Case Studies. SSRN Scholarly Paper ID 1958891, Social Science Research Network, Rochester, NY, June 2011. URL https://papers.ssrn.com/abstract=1958891.

Alberto Abadie and Jérémy L'Hour. PenSynthPaper.pdf, 2019. URL https://drive.google.com/file/d/1RnzO_L32jYXxXdFu1IBoSTjIrTfVYcZt/view?usp=sharing&usp=embed_facebook. Library Catalog: drive.google.com.

Ricardo Alonso-Zaldivar. Medicaid surge triggers cost concerns for states - The Boston Globe, 2014. URL https://www.bostonglobe.com/news/nation/2014/05/26/medicaid-surge-triggers-cost-concerns-for-states/IPOJoELIXnZWbUZVXTMWjI/story.html.

Larisa Antonisse and Rachel Garfield. The Effects of Medicaid Expansion under the ACA: Updated Findings from a Literature Review, March 2018. URL https://www.kff.org/medicaid/issue-brief/the-effects-of-medicaid-expansion-under-the-aca-updated-findings-from-a-literature-review-march-2018/.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix Completion Methods for Causal Panel Data Models. Technical Report w25132, National Bureau of Economic Research, Cambridge, MA, October 2018. URL http://www.nber.org/papers/w25132.pdf.

Alyssa Bilinski and Laura A. Hatfield. Seeking evidence of absence: Reconsidering tests of model assumptions. *arXiv:1805.03273 [stat]*, May 2018. URL http://arxiv.org/abs/1805.03273. arXiv: 1805.03273.

Janet Bouttell, Peter Craig, James Lewsey, Mark Robinson, and Frank Popham. Synthetic control methodology as a tool for evaluating population-level health interventions. *Journal of Epidemiology and Community Health*, 72(8):673–678, August 2018. ISSN 0143-005X. doi: 10.1136/jech-2017-210106. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6204967/.

Thomas C. Buchmueller and Colleen Carey. The Effect of Prescription Drug Monitoring Programs on Opioid Utilization in Medicare. *American Economic Journal: Economic Policy*, 10(1):77–112, February 2018. ISSN 1945-7731. doi: 10.1257/pol.20160094. URL https://www.aeaweb.org/articles?id=10.1257/pol.20160094.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, February 2020. ISSN 13697412. doi: 10.1111/rssb.12348. URL http://doi.wiley.com/10.1111/rssb.12348.

Congressional Budget Office. The Impact of Unauthorized Immigrants on the Budgets of State and Local Governments. *Congressional Budget Office*, page 24, 2007.

Charles Courtemanche, James Marton, Benjamin Ukert, Aaron Yelowitz, and Daniela Zapata. Early Impacts of the Affordable Care Act on Health Insurance Coverage in Medicaid Expansion and Non-Expansion States. *Journal of Policy Analysis and Management*, 36(1):178–210, 2017. ISSN 1520-6688. doi: 10.1002/pam.21961. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pam.21961.

Sandra L. Decker, Brandy J. Lipton, and Benjamin D. Sommers. Medicaid Expansion Coverage Effects Grew In 2015 With Continued Improvements In Coverage Quality. *Health Affairs*, 36(5):819–825, May 2017. ISSN 0278-2715. doi: 10.1377/hlthaff.2016.1462. URL https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2016.1462.

Bruno Ferman. On the Properties of the Synthetic Control Estimator with Many Periods and Many Controls. *arXiv:1906.06665 [econ]*, April 2020. URL http://arxiv.org/abs/1906.06665. arXiv: 1906.06665.

Molly Frean, Jonathan Gruber, and Benjamin D Sommers. Premium Subsidies, the Mandate, and Medicaid Expansion: Coverage Effects of the Affordable Care Act. Working Paper 22213, National Bureau of Economic Research, April 2016. URL http://www.nber.org/papers/w22213.

Bowen Garrett and Robert Kaestner. Has the ACA Been a Job Killer? Technical report, Urban Institute, 2015.

Government Accountability Office. State and Local Government Pension Plans: Economic Downturn Spurs Efforts to Address Costs and Sustainability, 2012. URL https://www.gao.gov/assets/590/589043.pdf.

Jens Hainmueller. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46, 2012. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpr025. URL https://www.cambridge.org/core/product/identifier/S1047198700012997/type/journal_article.

Chad Hazlett. Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica*, 2020.

Chad Hazlett and Yiqing Xu. Trajectory Balancing: A General Reweighting Approach to Causal Inference With Time-Series Cross-Sectional Data. SSRN Scholarly Paper ID

3214231, Social Science Research Network, Rochester, NY, August 2018a. URL `https://papers.ssrn.com/abstract=3214231`.

Chad Hazlett and Yiqing Xu. Trajectory Balancing: A General Reweighting Approach to Causal Inference With Time-Series Cross-Sectional Data. SSRN Scholarly Paper ID 3214231, Social Science Research Network, Rochester, NY, August 2018b. URL `https://papers.ssrn.com/abstract=3214231`.

Martha Heberlein, Tricia Brooks, and Jocelyn Guyer. Performing Under Pressure: Annual Findings of a 50-State Survey of Eligibility, Enrollment, Renewal, and Cost-Sharing Policies in Medicaid and CHIP, 2011-2012. Technical report, Kaiser Family Foundation, 2012.

Elicia J Herz, Julie Stone, and Evelyne P Baumrucker. Medicaid Checklist: Considerations in Adding a Mandatory Eligibility Group. Technical report, Congressional Research Service, 2016.

Jill Horwitz, Corey S Davis, Lynn S McClelland, Rebecca S Fordon, and Ellen Meara. The Problem of Data Quality in Analyses of Opioid Regulation: The Case of Prescription Drug Monitoring Programs. Working Paper 24947, National Bureau of Economic Research, August 2018. URL `http://www.nber.org/papers/w24947`.

Luojia Hu, Robert Kaestner, Bhashkar Mazumder, Sarah Miller, and Ashley Wong. The Effect of the Patient Protection and Affordable Care Act Medicaid Expansions on Financial Wellbeing. Technical Report w22170, National Bureau of Economic Research, Cambridge, MA, April 2016. URL `http://www.nber.org/papers/w22170.pdf`.

Robert Kaestner, Bowen Garrett, Jiajia Chen, Anuj Gangopadhyaya, and Caitlyn Fleming. Effects of ACA Medicaid Expansions on Health Insurance Coverage and Labor Supply. *Journal of Policy Analysis and Management*, 36(3):608–642, 2017. ISSN 1520-6688.

doi: 10.1002/pam.21993. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pam.21993.

Kaiser Family Foundation. Impact of the Medicaid Fiscal Relief Provisions in the American Recovery and Reinvestment Act (ARRA), October 2011. URL https://www.kff.org/wp-content/uploads/2013/01/8252.pdf.

Kaiser Family Foundation. Summary of the Affordable Care Act. *Kaiser Family Foundation*, page 13, 2013.

Genevieve M. Kenney, Jennifer M. Haley, Clare Wang Pan, Victoria Lynch, and Matthew Buettgens. Children's Coverage Climb Continues: Uninsurance and Medicaid/CHIP Eligibility and Participation Under the ACA, June 2016. URL https://www.urban.org/research/publication/childrens-coverage-climb-continues-uninsurance-and-medicaidchip-eligibility-and-participation-under-aca.

MACPAC. Medicaid's share of state budgets : MACPAC, 2017. URL https://www.macpac.gov/subtopic/medicaids-share-of-state-budgets/.

Olena Mazurenko, Casey P. Balio, Rajender Agarwal, Aaron E. Carroll, and Nir Menachemi. The Effects Of Medicaid Expansion Under The ACA: A Systematic Review. *Health Affairs*, 37(6):944–950, June 2018. ISSN 0278-2715. doi: 10.1377/hlthaff.2017.1491. URL https://www.healthaffairs.org/doi/10.1377/hlthaff.2017.1491.

Alison Mitchell. Medicaid's Federal Medical Assistance Percentage (FMAP). Technical report, Congressional Research Service, 2018.

Paul Potamianos, Michael Cohen, Margaret Kelly, Zac Jackson, Dan Timberlake, Sandra Beattie, Liza Clark, Phil Dean, Duncan Baird, Teresa MacCartney, David Thurman, Jill Geiger, Marc Nicole, and John Hicks. 2018 State Expenditure Report - Fiscal Years 2016 to 2018. Technical report, National Association of State Budget Officers, 2018.

Jonathan Roth. Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends. *Harvard*, 2018.

Robin Rudowitz, Allison Valentine, and Vernon Smith. Medicaid Enrollment & Spending Growth: FY 2016 & 2017, October 2016. URL `https://www.kff.org/medicaid/issue-brief/medicaid-enrollment-spending-growth-fy-2016-2017/`.

Andrew M. Ryan, James F. Burgess, and Justin B. Dimick. Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences. *Health Services Research*, 50(4):1211–1235, August 2015. ISSN 1475-6773. doi: 10.1111/1475-6773.12270. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.12270`.

Pantelis Samartsidis, Shaun R. Seaman, Anne M. Presanis, Matthew Hickman, and Daniela De Angelis. Assessing the Causal Effect of Binary Interventions from Observational Panel Data with Few Treated Units. *Statistical Science*, 34(3):486–503, August 2019. ISSN 0883-4237. doi: 10.1214/19-STS713. URL `https://projecteuclid.org/euclid.ss/1570780981`.

Benjamin D. Sommers and Jonathan Gruber. Federal Funding Insulated State Budgets From Increased Spending Related To Medicaid Expansion. *Health Affairs*, 36(5): 938–944, May 2017. ISSN 0278-2715. doi: 10.1377/hlthaff.2016.1666. URL `https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2016.1666`.

Laura R. Wherry and Sarah Miller. Early Coverage, Access, Utilization, and Health Effects Associated With the Affordable Care Act Medicaid Expansions: A Quasi-experimental Study. *Annals of Internal Medicine*, 164(12):795, June 2016. ISSN 0003-4819. doi: 10.7326/M15-2234. URL `http://annals.org/article.aspx?doi=10.7326/M15-2234`.

Jeffrey M. Wooldridge. *Introductory econometrics: a modern approach*. South-Western College Pub, Australia ; Cincinnati, Ohio, 2nd ed edition, 2003. ISBN 978-0-324-11364-8.

Yiqing Xu. Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis*, 25(1):57–76, January 2017. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2016.2. URL https://www.cambridge.org/core/journals/political-analysis/article/generalized-synthetic-control-method-causal-inference-with-interactive-fixed-effects-models/B63A8BD7C239DD4141C67DA10CD0E4F3.

Dahai Yue, Petra W. Rasmussen, and Ninez A. Ponce. Racial/Ethnic Differential Effects of Medicaid Expansion on Health Care Access. *Health Services Research*, 53(5):3640–3656, October 2018. ISSN 1475-6773. doi: 10.1111/1475-6773.12834. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.12834.

Bret Zeldow and Laura A. Hatfield. Confounding and Regression Adjustment in Difference-in-Differences. *arXiv:1911.12185 [stat]*, November 2019. URL http://arxiv.org/abs/1911.12185.