Synthetic Control Methods for Estimating Effects of Super-Spreader Events

Walker Hughes, Kenton Young

December 2020

1 Introduction

Each August, hundreds of thousands of bikers and motorcycle enthusiasts converge on Sturgis, South Dakota to attend the annual Sturgis Motorcycle Rally. Despite occurring in the midst of the COVID-19 pandemic, the 2020 Sturgis Rally looked very similar to previous years' rallies with people tightly packed together and in close quarters. More than 400,000 [2] individuals attended this rally in the midst of the COVID-19 pandemic with minimal social distancing– the only requirement for entry to the rally was having a mask in one's possession [3]. By the end of August, cumulative numbers of COVID-19 cases in Meade had risen by 257%. But what if there had been no rally? Would we still have seen this sharp rise in cases? Synthetic control offer the answer. These methods provide meaningful insights into what would have happened to a treated unit in the counterfactual state of no treatment. This is an especially useful research method in comparative case studies of ”super-spreader” events, such as the Sturgis Motorcycle Rally.

Recent comparative case studies, like *The Contagion Externality of a Superspreading Event: The Sturgis Motorcycle Rally and COVID-19* by Dhaval, Friedson, McNichols, and Sabia (*Dhaval et al.*) have utilized synthetic control methods to estimate the effect this rally had on the rise of COVID-19 cases in South Dakota. However, *Dhaval et al.* were extremely conservative with the donor pool they considered when constructing these controls [3]. Though this was intended to not violate the Stable Unit Treatment Value Assumptions (SUTVA) of no interference or contamination of the donor pool, the researchers inadvertently omitted from their donor pool many US counties that best resemble Meade, South Dakota in terms of demographics, mask wearing sentiment, and total number of COVID-19 cases before the onset of the Rally. Many of these counties did not contribute large numbers of Sturgis Rally attendees [8], and thus do not pose a substantial contamination risk to the donor pool. Further, this study did not implement any placebo tests to test the robustness of their synthetic controls.

Our paper seeks to replicate the analysis of *Dhaval et al.* with a slightly relaxed SUTVA assumption, which allows for some of the counties they omitted to be included in the donor pool. We also utilize a new data set from the New York Times that these researchers did not utilize, which describes different counties' propensities to use masks that was conducted in July, 2020, the month before the Sturgis Rally began. This new information will ensure that our donor pool of US counties best matches Meade, South Dakota prior to the onset of the Rally. We also implement placebo tests for each synthetic control estimated in order to test their robustness and comment on their significance. This allows us to gain additional insights into the effect of super-spreader events on the rise of COVID-19.

2 Background and Literature Review

South Dakota is one of the least densely populated states in the United States. With a population of 900,000 people over 77,000 square miles, many health experts credit this sparsity as inducing a "natural social distancing" in the state even in the absence of the pandemic [2]. *Dhaval et al.* describe how as of July 31, 2020 (the week before the Sturgis Rally), only 8764 COVID-19 cases had been confirmed in South Dakota as a whole. Meade County, where the Sturgis Rally occurred, accounted for a mere 84 of these cases, less than 1 percent of South Dakota's total cases up to that date. On August 31, 2020, two weeks after the Sturgis events terminated for the year, total cases in Meade County totaled 300 [3], a 257% increase in cumulative cases since the start of the rally.

The Sturgis Rally is attended by hundreds of thousands of bikers from across the United States each year, though the distribution of locales from which these attendees come is not uniform. Though many Rally attendees are indeed South Dakota residents, most come from just eight of South Dakota's 66 counties: Lawrence, Pennington, Meade, Butte, Minnehaha, Lincoln, Brown and Davison counties [8]. Aside from these eight counties, the counties that contribute the most attendees to the rally are from Minnesota, Colorado, Texas and California [8].

Knowing which counties contribute high numbers of Sturgis Rally attendees is paramount, since the choice of donor pool is key to an effective and insightful synthetic control study. Ideally, our donor pool counties ought to match our treated county well on pre-treatment COVID-19 cases, demographic characteristics and social distancing standards, while not posing an immediate contamination risk through spillover effects. *Dhaval et al.* omitted from their donor pool all counties from South Dakota and its border states (North Dakota, Montana, Wyoming, Minnesota, Nebraska and Iowa), as well as any county in the United States that had at least a single Rally attendee [3]. Though done as a means to avoid contamination of their donor pool and thus maintain the Single Unit Treatment Value Assumption, the assumption that *all* of these counties posed a contamination risk was extreme since only a small subset of the counties omitted by *Dhaval et al.* contributed large numbers of Rally attendees. Further, since these counties are likely to match Meade County on a large number of observable characteristics like demographic composition and mask usage, which affect the spread of COVID-19, these counties should not be omitted from the donor pool. Our analysis relaxes the assumptions of *Dhaval et al.* and allows these counties to be considered in the donor pool. To allow for this to happen, however, we must establish our own identifying assumptions.

3 Identification Strategy

As mentioned, this paper aims to first recreate the synthetic control results of *Dhaval et al.* as a basis to compare our new estimates. Though these researchers matched on pre-Rally cases and urbanicity rates (a term they did not define in great detail) [3], they did not include in their analysis any summary statistics for how well they matched on these characteristics. We avoid this confusion by matching on pre-Rally case counts, county demographics and mask survey data. We then estimate a synthetic Meade County from a donor pool of counties from South Dakota and its border states, and another final synthetic Meade County from a donor pool of counties from across the US, including those from South Dakota and its border states. In each case, we match on pre-Rally cases 28 days prior to the Rally, mask-wearing survey data from the New York Times, and county demographic characteristics (These characteristics include county racial composition and portion of citizens ages 18-29, and age group shown to be very likely to disobey social distancing and mask mandates [4]). For computational purposes, we consider only the 21 counties from each donor pool that best match Meade on the basis of these characteristics. In each estimate we also omit counties that were ranked in the top 50 for 2020 Sturgis Rally attendance [8]. For each synthetic control we estimate cases for two weeks after the Rally's end.

In order to estimate these synthetic controls, we must ensure that we avoid contamination and interference in our donor pool. As such, we assume that the Sturgis Rally did not serve as justification for other counties to hold extremely large, socially-proximate gatherings in the period of four weeks prior to the rally, and two weeks after its end. If this did not hold, spillover effects would invalidate our results since there would be a higher likelihood of our treated unit intermingling with our control units. We also assume that the drastic rise in COVID-19 cases in Meade County following the rally is entirely due to the lack of social distancing at the Sturgis Rally, and that no other exogenous shocks contributed to the rise in cases.

Further, in order to use our COVID-19 cases data from the New York Times, we must assume that COVID-19 tests are reliable and give true positives and true negatives. While this may be our most lofty assumption given that COVID-19 tests are extremely new and can be unreliable [9], this ensures that the drastic rise in cases in Meade County cannot simply be attributed to a large batch of faulty tests, an outlier in the instrument used to calculate total new cases.

4 Data

Before diving into our results, we briefly introduce our data used. Our COVID-19 cases data comes from the New York Times, which has been tracking and recording total cumulative cases at the county level daily since the onset of the pandemic [5]. This data is highly reliable and has been used in several recent studies on the pandemic. We also utilize county-level demographic data obtained from the United States Census Bureau, specifically on the racial and age composition of each United States county. We then considered counties for which we had data available in each data set, leaving us with 2971 of the United States' 3007 total counties.

Finally, one major difference between our data and that of *Dhaval et al.* is that through new data, we avoid using mere *estimates* of social distancing sentiments. While *Dhaval et al.* created a proxy for social distancing based on foot traffic data obtained from SafeGraph.com, we avoid this complication by utilizing new mask survey data from the New York Times. This recently published data comes from a nation-wide survey the New York Times conducted in July, 2020, the month prior to the Sturgis Rally [5]. In this survey, each participant from a pool of 250,000 was asked, "How often do you wear a mask in public when you expect to be within six feet of another person?" and responded with either, "Never," "Sometimes," "Frequently," or "Always." These responses were aggregated at the county level, weighting by age and gender demographics for each county. This data provides a reliable social distancing metric to match on and is likely to be more reflective of the true, underlying distribution of mask wearing and social distancing than mere estimates would be. Further, we integrate this data when implementing our synthetic controls as a matching characteristic.

5 Results

5.1 Replication

We replicated the results of *Dhaval et al.* by estimating a synthetic control for Meade County by excluding all counties from South Dakota and its border states. With this pool of counties, we matched on slightly different county characteristics as described previously. While these authors also omitted any county from which a cell phone ping was detected at any Rally event [3], we omit this analysis do to computational constraints, and only omit counties that ranked in the top 50 for 2020 Sturgis attendance. Estimating this synthetic control revealed that our synthetic Meade County matched very well on the basis of pre-treatment number of cases. Our donor pool matched well on the basis of mask survey data, though the mean number of donor pool cases pre-treatment was about 3 cases lower than the truth [Table 1a, Appendix]. Our donor pool also matched fairly well on the basis of demographic characteristics.

Our synthetic estimate for this donor pool matches the trend very closely to that of *Dhaval et al.*, especially considering the tangency of our synthetic to the true number of cases just before the end of the rally, which the researchers also attain. Figure 1 in the Appendix shows this estimate with its associated placebo tests. Our estimate does not diverge too far from the truth until after the rally ends on August 16. Given that the Rally spanned 10 days and that COVID-19 symptoms can take up to two weeks to manifest themselves after exposure, this is consistent with what one might expect from such an event–the true effects were not felt until *after* the contagion period.

On the day prior to the Sturgis Rally, there were 82 confirmed cases in Meade County. On August 31, 2020, three weeks after the Sturgis events terminated for the year, total cases in Meade County were recorded at precisely 300 cases. Our synthetic control estimates that had no rally occurred, total cases on August 31 would have been approximately 131.7 in Meade County, a mere 43.9% of the true number of cumulative cases recorded. We also found that through placebo tests on our donor counties, Meade county did indeed experience the highest increase in cases after the treatment period compared to our donor counties. This is discussed in greater detail further on.

5.2 Donor Pool from South Dakota and Border States

Since *Dhaval et al.* omitted all counties from South Dakota and its border states, we next estimated another synthetic control for Meade County by only considering donor counties from these states, again excluding counties that were ranked in the top 50 for Sturgis attendance in 2020. With this donor pool, we were able to again match well on the basis of pre-Rally average cases, mask-wearing survey responses and demographic characteristics [Table 2a, Appendix]. It should be noted that this donor pool matches even more closely to Meade County on average pre-period cumulative cases and the portion of the population ages 18-29 than found previously. We also note that our matching on mask-wearing sentiment was surprisingly not as close as in our previous donor pool, though our synthetic Meade again matches extremely well on the basis on pre-Rally average cumulative cases.

This estimate was intriguing as we found that the synthetic Meade County constructed from only counties in South Dakota and its border states did not differ much from the previous estimation where we omitted these counties altogether, though this estimate did give slightly higher post-Rally numbers of cumulative cases [Figure 2, Appendix]. Like the previous estimate, our synthetic Meade attains near-tangency to the true number of cases on August 18 near the end of the rally. Also noteworthy is that of the counties in this donor pool, the counties with non-zero weights in the estimate were all from Iowa. While one may be skeptical of this result, it suggests that counties previously omitted from the donor pool match Meade fairly well on the basis of demographic characteristics and pre-period case counts.

With this donor pool, we found that on August 31, 2020, synthetic total cases in Meade County were estimated to be 134.3, only slightly higher than previously estimated. This amounts to 44.8% of the true cumulative number of cases recorded as of August 31. Further, in our placebo tests for this estimate, we found that our synthetic donor counties differed only slightly from the truth, suggesting that those counties experienced only mild effects from the Rally. Meade County by far experienced the greatest difference between its true cumulative cases and synthetic estimates, indicating it experienced the greatest effects of the Rally.

5.3 New Estimation Excluding High-Attendance Counties

We estimated our final synthetic control for Meade County with a donor pool that only excluded the top 50 counties that contributed Rally attendees in 2020. This relaxes the assumptions of *Dhaval et al.* by allowing counties from South Dakota and its border states to be considered in the donor pool. In this synthetic control, our synthetic Meade County and donor pool matched even closer to the true average number of total cases in Meade prior to the Rally than in any of our other donor pools [Table 3a, Appendix]. In our previous two estimates, the average number of pre-rally cumulative cases from our donor pools was off by a few cases of the true average of 63.928, however, this final donor pool's average comes within *fractions* of cases of the truth [Table 3a, Appendix].

Also of importance is that two counties from South Dakota's border states–Worth, Iowa and Barnes, North Dakota–had non-zero weights assigned to them in constructing this synthetic Meade. These counties were omitted from the donor pool *Dhaval et al.*, and their non-zero weights indicate their relative usefulness against the other counties in our donor pool when constructing our synthetic Meade. What is also noteworthy about this synthetic Meade County is that, unlike our previous two estimates, it never intersects Meade's true number of cumulative cases after the rally begins. Indeed, this estimate's slope remains relatively constant over time and thus avoids intersecting the truth, whereas our previous estimates both intersected the truth around August 15-18 near the end of the Rally.

This synthetic control estimates that on August 31, 2020, total cases in Meade County would have been 108.9 had the Sturgis Rally not occurred, only 36.3% of the true total of 300 cases. This synthetic Meade indicates that the Sturgis Rally had a greater effect on COVID-19 spread than our other two estimates, and many key demographic characteristics than our other estimates do. Several key insights from this analysis are described in the following section.

6 Inference

In order to address the topic of inference, we follow the approach of *Abadie et al.* in their paper *Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program* by asking the question of whether our results could have been driven merely by chance. Specifically, given a randomly chosen county from our donor pool, how often would we obtain as extreme a rise in cases after the onset of the Sturgis Rally?

(Note that this is similar to how a p-value gives us a probability of seeing data more extreme than we already observed given a null hypothesis). This question is answered through the use of placebo tests, where for each county in our donor pool, we estimate a synthetic control as we did for Meade County, and compare the difference between each county's true numbers of COVID-19 cases and their synthetic estimates. The estimated effect of the Sturgis Rally on a given county's change in cumulative COVID-19 cases is this difference. Since these donor pool counties did not receive treatment (ie., they did not host motorcycle rallies where 400,000 people attended), we would expect the differences between their true cumulative cases and synthetic estimates to be lower in magnitude than they were in Meade.

Figures 1, 2 and 3 in the Appendix depict these placebo tests. We found that in each of these tests, the effect of the Rally was far larger for Meade County than for any other county in our donor pools. In each case, our synthetics match very well in the pre-period before August 7, 2020, resulting in near-zero differences in the placebo graphs before the Rally. After the beginning of the Rally, we note that the placebo tests in Figures 1 and 3 resemble each other quite well, with only one or two other donor counties experiencing somewhat large rises in cases after August 7. The placebo test in Figure 2, however, has a far tighter spread in how far each donor county's synthetic case counts differ from their true number of cases. Remembering that this placebo test corresponds to the synthetic control estimated with donor counties from South Dakota and its border states alone, this suggests that the counties from this donor pool–which are on average geographically closer to Meade than those of our other donor pools–experienced far milder effects post-treatment.

These placebo tests are extremely compelling for two main reasons. First, they imply that Meade was indeed the county that experienced the greatest rise in cases after the onset of the rally. Second, Figure 2 suggests that the counties geographically closest to Meade experienced only mild rises in cases after the Rally. This supports our assumption that few counties from South Dakota and its border states truly posed contamination and interference threats to our donor pool. Also noteworthy is that, as described in the previous section, when counties from South Dakota and its border states were included in the donor pool for Figure 3, two of these counties had non-zero coefficients. Our findings not only suggest that including counties from South Dakota and its border states in our donor pools does not pose a serious threat to the validity of our experiment, but also provide compelling evidence that the Sturgis Motorcycle Rally truly had an enormous effect on the rise of COVID-19 cases in Meade relative to counties that best match Meade on various demographic and mask-wearing characteristics.

7 Conclusion

There is no question whether the 2020 Sturgis Motorcycle Rally lead to an increase in COVID-19 infections in Meade, South Dakota; rather, the question is by *how much*. Having estimated several synthetic controls for Meade with various donor pools of US counties, we estimate that on August 31, 2020, total cumulative cases in Meade County would have been a mere 36.3% - 44.8% of the true total of 300 cases had the Rally not occurred. We confirm that Meade County indeed experienced the greatest effects of this Rally compared to our donor pools through placebo tests. We also find that counties from both Iowa and North Dakota–counties omitted by *Dhaval et al.*–received non-zero weights in our estimates, indicating their relative usefulness in constructing our synthetic control. Considering our placebo tests, especially in Figure 2, our analysis suggests that including counties from South Dakota and its border states in our donor pool likely does not pose extreme contamination or interference risks.

Our findings are pertinent for policy makers and citizens alike. Not only do our analyses suggest that ”super-spreader” events can be extremely risky and have long-term effects on the spread of COVID-19, they also indicate that policy measures regarding public health should be taken extremely seriously, especially during a pandemic. Further research on the topic ought to explore how ”super-spreaders” affect inter-day rates of infection, and how long it may take for infection rates after a ”super-spreader” to converge to pre-spike levels. In all, this analysis suggests that the effects of super-spreader events can be extreme, difficult to control, and pose immediate public health risks to us all.

8 References

[1] Abadie et al. ”Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” Journal of the American Statistical Association, June 2010

[2] Cherney, Andrew. “The 2020 Sturgis Rally Stat Sheet.” Motorcycle Cruiser, Bonnier Corporation, 19 Aug. 2020

[3] Dave, Dhaval M., et al. “The Contagion Externality of a Superspreading Event: The Sturgis Motorcycle Rally and COVID-19.” IZA Institute of Labor Economics Publications, IZA Discussion Papers Series, Sept. 2020

[4] Morin, Rebecca. ”Young Americans less likely to social distance as coronavirus cases continue to rise, survey says.” USA Today, June 2020.

[5] New York Times. COVID-19 Data (2020). [US Counties and Mask Use data files]

[6] Porterfield, Carlie. “Sturgis Motorcycle Rally Attracts Thousands, With Few Masks And Little Social Distancing (Photos).” Forbes, Forbes Magazine, 8 Aug. 2020

[7] United States Census Bureau. County Population Characteristics (2019). [US County Population Characteristics data files]

[8] ”2020 Sturgis Motorcycle Rally Analysis.” Washington Post, Covid Alliance. Sept, 2020.

[9] ”Questions About Accuracy of Coronavirus Tests Sow Worry.” Wall Street Journal. April 2020.

9 Appendix

Table 1a: Figure 1 Average Values For Pre-August 7, 2020

Characteristic	Meade County	Donor Average	Synthetic
COVID-19 Cases			
Average Pre-Rally Cases	63.928	60.544	63.930
Wears Mask			
Never	0.123	0.099	-
Rarely	0.078	0.087	-
Sometimes	0.161	0.146	-
Frequently	0.214	0.216	-
Always	0.425	0.452	-
Demographics			
% Pop. Age 18-29	0.152	0.147	-
% White	0.905	0.862	-
% Black	0.020	0.087	-
% American Indian	0.032	0.020	-
% Asian	0.010	0.011	-
% Pacific Island	0.001	0.001	-

Average values for county characteristics matched on as percentages. Average pre-period cases of COVID-19 reported as mean of total cases before August 7.

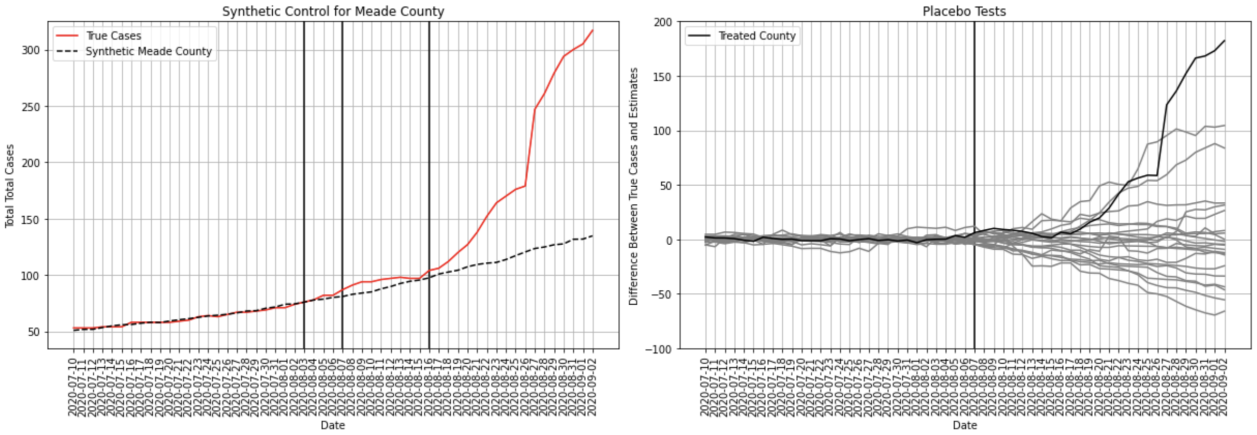

Figure 1: Replicated Synthetic Control for Meade County Excluding Counties From South Dakota and Border States. In the left figure, the vertical lines from left to right are the date of the first pre-rally event, the official start date, and the official end date. The vertical line on the right figure is the start date.

Table 1b: Counties and Weights			
County	Weight	County	Weight
Adams Ohio	0.224	Gates North Carolina	0.000
Dickinson Michigan	0.000	Carroll Illinois	0.325
Parke Indiana	0.000	Vernon Wisconsin	0.000
Wayne Kentucky	0.000	Shoshone Idaho	0.000
Ste. Genevieve Missouri	0.076	Juab Utah	0.000
Grand Utah	0.000	Tioga Pennsylvania	0.026
Cumberland Illinois	0.000	Aroostook Maine	0.000
Lincoln West Virginia	0.000	Taylor Georgia	0.000
Hansford Texas	0.000	McCreary Kentucky	0.000
Donley Texas	0.128	Vermillion Indiana	0.221
Randolph Missouri	0.000		

Table 2a: Figure 2 Average Values For Pre-August 7, 2020			
Characteristic	**Meade County**	**Donor Average**	**Synthetic**
COVID-19 Cases			
Pre-period Cases	63.928	64.811	63.985
Wears Mask			
Never	0.123	0.121	-
Rarely	0.078	0.121	-
Sometimes	0.161	0.151	-
Frequently	0.214	0.237	-
Always	0.425	0.370	-
Demographics			
% Pop. Age 18-29	0.152	0.150	-
% White	0.905	0.903	-
% Black	0.020	0.014	-
% American Indian	0.032	0.052	-
% Asian	0.010	0.011	-
% Pacific Island	0.001	0.001	-

Average values for county characteristics matched on as percentages. Average pre-period cases of COVID-19 reported as mean of total cases before August 7.

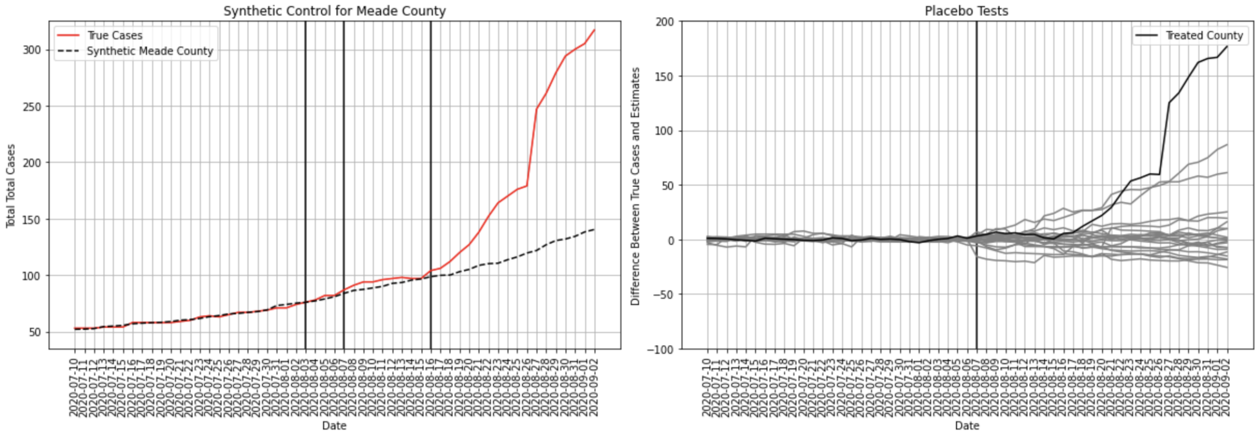

Figure 2: Synthetic Control for Meade County with Donor Pool Only From South Dakota and its Border States. In the left figure, the vertical lines from left to right are the date of the first pre-rally event, the official start date, and the official end date. The vertical line on the right figure is the start date.

Table 2b: Counties and Weights

County	Weight	County	Weight
Morrison Minnesota	0.000	York Nebraska	0.000
Meeker Minnesota	0.000	Monroe Iowa	0.000
Faribault Minnesota	0.000	Madison Iowa	0.322
Hughes South Dakota	0.000	Chippewa Minnesota	0.186
Fayette Iowa	0.076	Grundy Iowa	0.223
Delaware Iowa	0.000	Winnebago Iowa	0.000
Lyman South Dakota	0.000	Seward Nebraska	0.000
Albany Wyoming	0.000	Wabasha Minnesota	0.000
Roberts South Dakota	0.000	Union Iowa	0.167
Jackson Minnesota	0.000	Buchanan Iowa	0.000
Kossuth Iowa	0.027		

Table 3a: Figure 3Average Values For Pre-August 7, 2020			
Characteristic	**Meade County**	**Donor Average**	**Synthetic**
COVID-19 Cases			
Pre-period Cases	63.928	63.246	63.925
Wears Mask			
Never	0.123	0.104	-
Rarely	0.078	0.086	-
Sometimes	0.161	0.143	-
Frequently	0.214	0.213	-
Always	0.425	0.454	-
Demographics			
% Pop. Age 18-29	0.152	0.148	-
% White	0.905	0.860	-
% Black	0.020	0.092	-
% American Indian	0.032	0.018	-
% Asian	0.010	0.010	-
% Pacific Island	0.001	0.001	-

Average values for county characteristics matched on as percentages. Average pre-period cases of COVID-19 reported as mean of total cases before August 7.

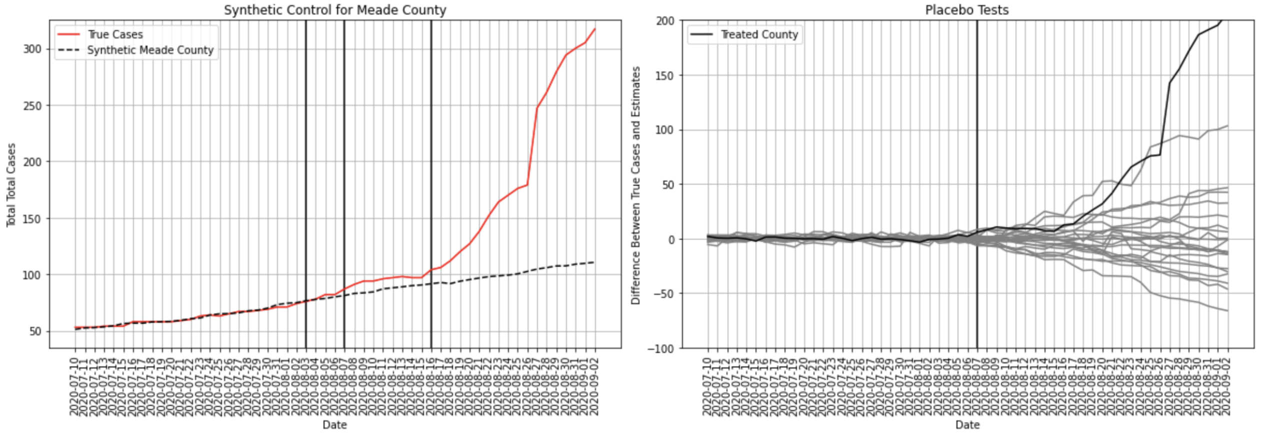

Figure 3: Synthetic Control for Meade County with Donor Pool Excluding only top 50 Counties for 2020 Attendance. In the left figure, the vertical lines from left to right are the date of the first pre-rally event, the official start date, and the official end date. The vertical line on the right figure is the start date.

Table 3b: Counties and Weights

County	Weight	County	Weight
Parke Indiana	0.000	Carroll Illinois	0.000
Grant Minnesota	0.000	Juab Utah	0.154
Worth Iowa	0.015	Tioga Pennsylvania	0.000
Cumberland Illinois	0.000	Barnes North Dakota	0.083
Wayne Kentucky	0.285	McCreary Kentucky	0.000
Dickinson Michigan	0.000	Vermillion Indiana	0.095
Donley Texas	0.000	Taylor Georgia	0.031
Hansford Texas	0.000	Cloud Kansas	0.000
Silver Bow Montana	0.000	Vilas Wisconsin	0.000
Vernon Wisconsin	0.295	Lynn Texas	0.000
Randolph Missouri	0.043		