Research Article

Layla Parast*, Priscillia Hunt, Beth Ann Griffin, and David Powell

When is a Match Sufficient? A Score-based Balance Metric for the Synthetic Control Method

https://doi.org/10.1515/jci-2020-0013
Received Dec 17, 2019; accepted Oct 01, 2020

Abstract: In some applications, researchers using the synthetic control method (SCM) to evaluate the effect of a policy may struggle to determine whether they have identified a "good match" between the control group and treated group. In this paper, we demonstrate the utility of the mean and maximum Absolute Standardized Mean Difference (ASMD) as a test of balance between a synthetic control unit and treated unit, and provide guidance on what constitutes a poor fit when using a synthetic control. We explore and compare other potential metrics using a simulation study. We provide an application of our proposed balance metric to the 2013 Los Angeles (LA) Firearm Study [9]. Using Uniform Crime Report data, we apply the SCM to obtain a counterfactual for the LA firearm-related crime rate based on a weighted combination of control units in a donor pool of cities. We use this counterfactual to estimate the effect of the LA Firearm Study intervention and explore the impact of changing the donor pool and pre-intervention duration period on resulting matches and estimated effects. We demonstrate how decision-making about the quality of a synthetic control can be improved by using ASMD. The mean and max ASMD clearly differentiate between poor matches and good matches. Researchers need better guidance on what is a meaningful imbalance between synthetic control and treated groups. In addition to the use of gap plots, the proposed balance metric can provide an objective way of determining fit.

Keywords: Synthetic control method; Non-experimental study; Matching methods; Gun violence

2020 Mathematics Subject Classification: 62D20, 62P25

1 Introduction

Researchers using the synthetic control method (SCM) to evaluate the effect of a strategy may struggle to determine whether they have identified a "good match" between the control group and treated group. Without an appropriate counterfactual, it is recommended that researchers not use the SCM for causal inference, but there is little guidance about how to determine whether the estimated synthetic control is satisfactory [1]. While the SCM has proven valuable in empirical crime research [2–6, *e.g.*,], we focus on an example in which it is unclear whether the SCM provides a useful counterfactual.

In 2005, an interagency working group of California law enforcement officials and crime researchers entered a partnership to design interventions to reduce gun violence in Los Angeles (LA). One intervention, the gun letter program, was implemented by the LA City Attorney's Office in which letters were sent to purchasers of handguns during the 10-day waiting period. Letters advised the purchasers that the dealer record of sale

***Corresponding Author: Layla Parast:** RAND Corporation, 1776 Main Street, Santa Monica, CA; Email: parast@rand.org
Priscillia Hunt: RAND Corporation, 1776 Main Street, Santa Monica, CA; Institute of Labor Economics (IZA), Schaumburg-Lippe-Strasse 5-9, 53113 Bonn, Germany
Beth Ann Griffin, David Powell: RAND Corporation, 1200 S Hayes St, Arlington, VA 22202

for the new weapon is in his or her name, and that failure to properly record any transfer of the weapon with California's Department of Justice is a crime. The letter further emphasized that whenever an improperly transferred gun is found at a crime scene, the LA City Attorney's Office would prosecute the original owner. The aim of the letter was to reduce rates of firearm violence by deterring transfers of legally purchased weapons to individuals prohibited from purchasing them, known as "straw purchases" [7]. After a pilot randomized trial in two neighborhoods of LA [8], the letter program was fully implemented citywide between January 1, 2013, and September 1, 2015. In 2017, Hunt *et al.* (2017) [9] applied the SCM to evaluate the impact of the letter strategy on firearm crime in LA. However, upon observation, authors determined the match between synthetic LA and LA did not appear good enough, so they did not continue with inference. Therein lies the problem for many applications of SCM.

When using SCM, the appropriateness of fit between the synthetic control group and treated group is typically tested visually and either accepted or discarded by the researcher. As Ben-Michael *et al.* 2018 [10] pp 6 note, "[t]here is little guidance about what constitutes poor fit…and common practice is fairly ad hoc." That is, it is often difficult to determine what constitutes a bad enough match that one should not proceed with inference. Specifically, there is not a test metric to objectively assess whether the synthetic control and treated groups are statistically equivalent. This has left many researchers trying to determine- what is a good enough match?

In this study, we present a new metric to help analysts determine whether they have a good enough match to conduct causal inference using SCM. The metric is borrowed from the propensity score literature to objectively examine whether a SCM match should be considered sufficient to proceed with inference. We additionally explore and compare other potential metrics using a simulation study. For our proposed metric, we assess different ways to optimize balance for cases like our case study, examining impacts of the firearm letter intervention (hereafter, Firearm Letter Study), where optimal balance is not obtained between the treated unit and synthetic control, and we explore how changing the donor pool and preintervention duration influences test statistics.

2 Methods

The SCM, introduced and developed in Abadie and Gardeazabal (2003) [11] and Abadie *et al.* (2010, 2015) [12, 13], allows for the analysis of data from studies with only one treated unit. While difference-in-difference and fixed effect models use an unweighted mean of the control units to develop a counterfactual for the treated unit, the synthetic control method uses a weighted mean of the control units, which is then referred to as the "synthetic control". The weights are constructed optimally in a manner that minimizes the pre-intervention differences in the outcomes between the treated unit and the synthetic control. This additional flexibility potentially permits construction of a synthetic control with similar pre-existing trends and levels as the treated unit, even when the unweighted outcomes of the control units are systematically different. Thus, in the LA Firearm Letter Study example [9], authors generate a "synthetic LA" for comparison to LA by using the annual weighted means of firearm crime in medium to large cities throughout the U.S., prior to the intervention in 2013. Notably, the SCM approach has been described as "arguably the most important innovation in the policy evaluation literature in the last 15 years" [14].

A key assumption of the SCM is that the treated unit is within the convex hull of the potential control units, an assumption equivalent to the parallel trends assumptions necessary for difference-in-differences. The convex hull assumption states that there exist non-negative weights, which sum to one, such that the weighted average of the outcomes of the control units is close to the outcomes of the treated unit in the pre-treatment period. This assumption is testable by examining the "fit" in the pre-period between the treated unit and its synthetic control. Typically, the appropriateness of this fit is tested visually using gap plots. We propose using a score-based balance metric.

2.1 Proposed Balance Metric for SCM

The balance metric we introduce to the SCM setting is pulled from the propensity score literature [15–19]. Specifically, we propose using the Absolute Standardized Mean Difference (ASMD) between the weighted synthetic control (SC) and the treated unit, examining each pre-intervention time period individually. By definition, the ASMD for a given factor equals the absolute difference between the weighted SC group and the treated unit (LA in our application), divided by the standard deviation of the given factor in the SC control group:

$$ASMD_{year} = \frac{|Y_{treated,year} - Y_{SC,year}|}{sd(Y_{SC,year})}$$

where $Y_{treated,year}$ is the outcome of the treated unit in the pre-intervention year, $Y_{SC,year}$ is the weighted mean outcome in the SC in the pre-intervention year where the weights are specified by the SCM, and $sd(Y_{SC,year})$ is the weighted standard deviation of the outcome in the SC group in the pre-intervention year. The estimated standard deviation in the denominator has to come from the control group in this case since there is no variability in the treated condition (single group) within a single year. The estimate provides us with a sense of how much variability there typically might be for the measure of interest (here, the preintervention outcome in a particular year prior to the intervention) among controls that are given a non-zero weight by the SCM, and accounting for the estimated weights. The ASMD provides a way to gauge how similar the treated unit and its synthetic control are on each preintervention year used to match in the SCM.

ASMDs can be used to help quantify the size of the imbalances shown in typical SCM gap plots by providing a simple numerical summary for researchers to compute and assess in a SCM application. We propose to examine two summaries of the ASMD values across years: the maximum ASMD denoted as max_{ASMD} and the mean of the ASMD values across years denoted as $mean_{ASMD}$ which can be expressed as:

$$mean_{ASMD} = {T_0}^{-1} \sum_{year}^{T_0} \frac{|Y_{treated,year} - {Y_{control,year}}'w|}{sd(Y_{SC,year})}$$

and

$$max_{ASMD} = \max_{year} \frac{|Y_{treated,year} - {Y_{control,year}}'w|}{sd(Y_{SC,year})}$$

where T_0 is the number of pre-intervention years and w is the vector of SCM weights.

2.2 Alternative Potential Metrics for SCM

While to our knowledge there have been no proposed metrics to assess balance in the SCM context, quantities other than our proposed ASMD metric may also prove useful. For example, because the SCM chooses weights, w, as a solution to the constrained optimization problem

$$\min_{w} \sum_{year}^{T_0} \left(Y_{treated,year} - {Y_{control,year}}'w\right)^2$$

subject to the weights summing to 1 and being non-negative, a logical metric to consider would be the root mean squared error (RMSE)

$$RMSE = \sqrt{{T_0}^{-1} \sum_{year}^{T_0} \left(Y_{treated,year} - {Y_{control,year}}'w\right)^2}$$

which directly targets the optimization function within SCM. A potential disadvantage of this approach is that unlike the ASMD, this metric is relative to the scale of the outcome rather than on a standardized scale.

Therefore, another logical metric would be the RMSE with standardization *i.e.*

$$SRMSE = \sqrt{T_0{}^{-1} \sum_{year}^{T_0} \left\{ \frac{Y_{treated,year} - Y_{control,year}{}'w}{sd(Y_{SC,year})} \right\}^2}$$

where again, $sd(Y_{SC,year})$ is the weighted standard deviation of the outcome in the SC group in the pre-intervention year. Lastly, although proposed within a framework to obtain an "augmented" SCM estimate, the "estimated bias" proposed by Ben-Michael (2018) [10] could also be considered as a metric to assess balance. Using their approach, the estimated bias due to imbalance is obtained by first fitting an outcome model *e.g.* a simple linear model to predict the post-intervention outcome using the lagged outcomes in previous years as the predictors (model fit only among controls). Then, the bias is estimated as the difference in the predicted outcome when the model is applied to the treated lagged outcomes, and the average of the weighted predicted outcomes when the model is applied to the control lagged outcomes. Unlike ASMD, this metric is on the scale of the outcome, rather than a standardized metric.

3 Simulation Study

3.1 Simulation Setup

We use a simulation study to examine and compare our proposed metric with the alternative metrics described above, and to explore appropriate cut-off values for the various metrics, *i.e.* a threshold at which point one should consider the match sufficient or not. To examine and compare the metrics, we consider an ideal metric to be such that higher values of the balance metric (indicating bad balance) correspond to higher bias in the estimated intervention effect with a monotone increasing correspondence. With respect to a cut-off value for a particular metric, an ideal cut-off would be such that the probability of identifying a match as having poor balance increases as the bias in the estimated intervention effect increases. For the ASMD metric used in the propensity score literature, values (mean or max) less than 0.1 are considered to be small imbalances, whereas ASMD values of 0.10 to 0.40 are considered moderate imbalance, and greater than 0.40 are large imbalances. However, though we borrow from propensity score work, the SCM is distinct from propensity score analysis and thus, application of the same thresholds without exploration may not be appropriate.

We examine 14 simulation settings. For all settings, simulated datasets were constructed utilizing the LA Firearm Letter Study (where there was one treated city – LA – and 102 control cities and 12 years of pre-intervention data), the intervention effect was set to be zero, and simulation results summarize over 1000 iterations. In settings 1-7, for each iteration, 10 control cities were randomly selected and the treated outcomes were generated as weighted combinations of these 10 cities. In setting 1, the weights used were: (0.25,0.20,0.15,0.10,0.05,0.05,0.05,0.05,0.05,0.05); that is, the treated outcomes were a perfectly weighted combination of a subset of the control cities. In setting 2, treated outcomes were generated using the same weights as setting 1 but with added error generated from a Normal(0,0.025) distribution. In settings 3 through 6, the error variance was increased to 0.05, 0.1, 0.2, and 0.3, respectively. In setting 7, the weights used were: (0.4, 0.2, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10) *i.e.*, outside the convex hull, and the variance was 0.50. In settings 8-14, for each iteration, 3 control cities were randomly selected and the treated outcomes were generated as weighted combinations of these 3 cities. In setting 8, the weights used were: (0.3, 0.4, 0.3); in setting 9, treated outcomes were generated using the same weights as setting 8 but with added error generated from a Normal(0,0.025) distribution. In settings 10 through 13, the error variance was increased to 0.05, 0.1, 0.2, and 0.3, respectively. In setting 14, the weights used were: (0.4, 0.4, 0.4) *i.e.*, outside the convex hull, and the variance was 0.5.

For each setting and each iteration we calculate the: (1) mean ASMD, (2) max ASMD, (3) RMSE, (4) SRMSE, (5) the estimated bias metric proposed by Ben-Michael (2018)[10], (6) the bias of the intervention effect estimate. Since the intervention effect is set to be zero, the intervention effect bias is calculated as the difference between the post-intervention treated outcome and the weighted combination of the post-intervention con-

trols using the estimated SCM weights. In addition, because the interpretation of the bias of the intervention effect estimate depends on the scale of the outcome, we standardize it by dividing by the simple unweighted standard deviation of all post-intervention outcomes. Thus, the bias of the intervention effect estimate is on a Cohen's *d* scale where 0.2 is generally considered a small effect size, 0.5 is considered a moderate effect size, and 0.8 is considered a large effect size [20].

3.2 Simulation Results

Figure 1a shows the average metric within settings 1-7 by plotting the average of the metrics against (the absolute value of) the intervention effect bias where each point represents one setting, *e.g.* the first point on the left is setting 1, the last point on the right is setting 7. While we wouldn't necessarily expect a strictly linear trend in these figures, we would expect, for a reasonable metric, that the magnitude of the metric increases monotonically as the intervention effect bias increases. A very steep increase may raise concerns about high sensitivity in practice, while a small slope may raise concerns about lack of sensitivity. As with the evaluation of most statistical metrics, we seek a balance between these two extremes. Examining settings 1-7 first, the magnitudes of the metrics all generally increase as the matches become more imperfect *i.e.* from setting 1 to setting 7, in order. This figure illustrates that all metrics have the desired property that they increase as the match becomes more imperfect and the intervention effect bias increases. In addition, these results show that max ASMD is very sensitive and increases rapidly as the intervention effect bias increases, while the Ben-Michael estimated bias increases much more slowly. The RMSE, SRMSE, and mean ASMD metrics lie between, with moderate increases as the bias increases. This general pattern is also observed for settings 8-14, shown in Figure 1b, which use weighted combinations of 3 cities instead of 10. Of course, given the construction of the metrics, the units of the metrics are not directly comparable – while the ASMD metrics are on an effect size scale, the Ben-Michael estimated bias and RMSE are relative to the scale of the outcome, and SRMSE is on the scale of a root mean squared effect size. Therefore, for RMSE and Ben-Michael estimated bias, examination of the metrics and thresholds would vary depending on the scale of the outcome. Such tailoring of the scaling and thresholds for different applications is generally not desirable. In addition, while SRMSE involves stan-

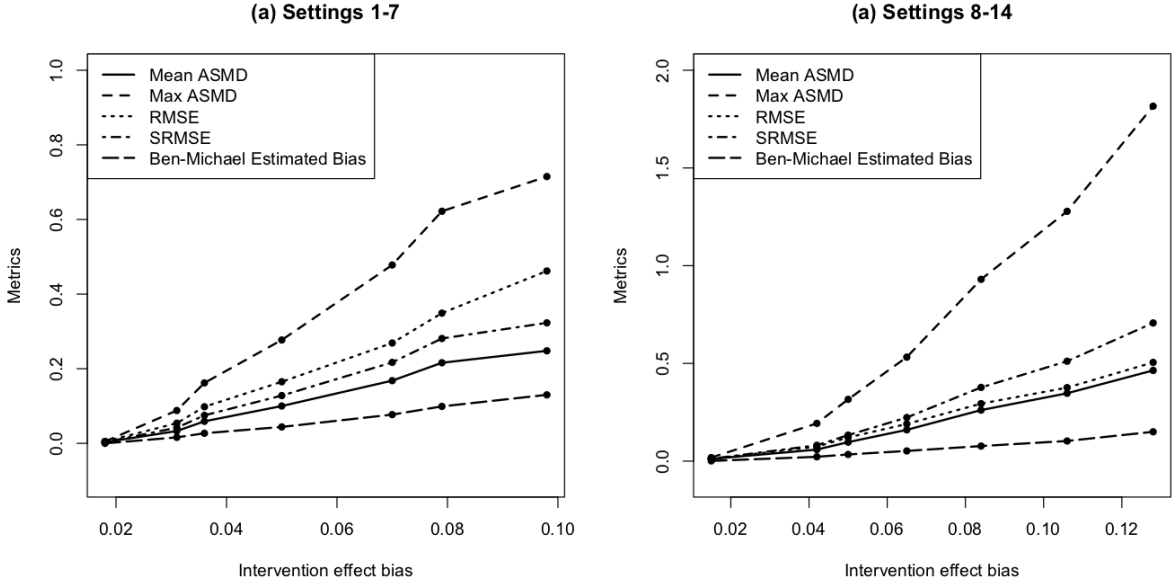


Figure 1: Simulation Study Results; metrics versus bias in intervention effect estimate averaged across 1000 replications; (a) each point reflects one simulation setting, for settings 1-7, in order from left to right, (b) each point reflects one simulation setting, for settings 8-14, in order from left to right

dardization, the metric scale may not be easily interpretable. For these reasons, we prefer the ASMD metrics as the effect size scale is broadly applicable, easy to interpret, and largely familiar to applied researchers.

With respect to selecting a specific threshold, we focus specifically on the mean and max ASMD; in the Supplementary Material, we examine various thresholds for RMSE, SRMSE, and the Ben-Michael estimated bias metrics. Figure 2 shows the proportion of simulation iterations with a calculated metric above each threshold versus (the absolute value of) the intervention effect bias for the mean ASMD and max ASMD. The property we wish to see is that a greater proportion of iterations are identified as above the threshold (indicating a bad match) as the intervention effect bias increases, and that a very small proportion of iterations are identified as above the threshold when the intervention effect bias is small (indicating a good match). This figure shows that for mean ASMD, 0.1 appears to reflect these desirable properties. For example, when the intervention effect bias is small, between 0.02 and 0.04, the proportion above the 0.1 threshold is substantially less than 20%; in contrast, the proportion above 0.05 in this bias region is quite a bit higher, reaching over 40%. When the intervention effect bias is larger, between 0.08 and 0.10, the proportion above the 0.1 threshold is over 80%; in contrast, the proportion above the 0.2 threshold in this bias region is around 50% or less. Thus, given our ideal expectations, the 0.10 threshold appears optimal. In contrast, for max ASMD, 0.1 appears to be a bit too sensitive, and 0.2 may be slightly preferred. Results are similar for settings 8-14 (not shown).

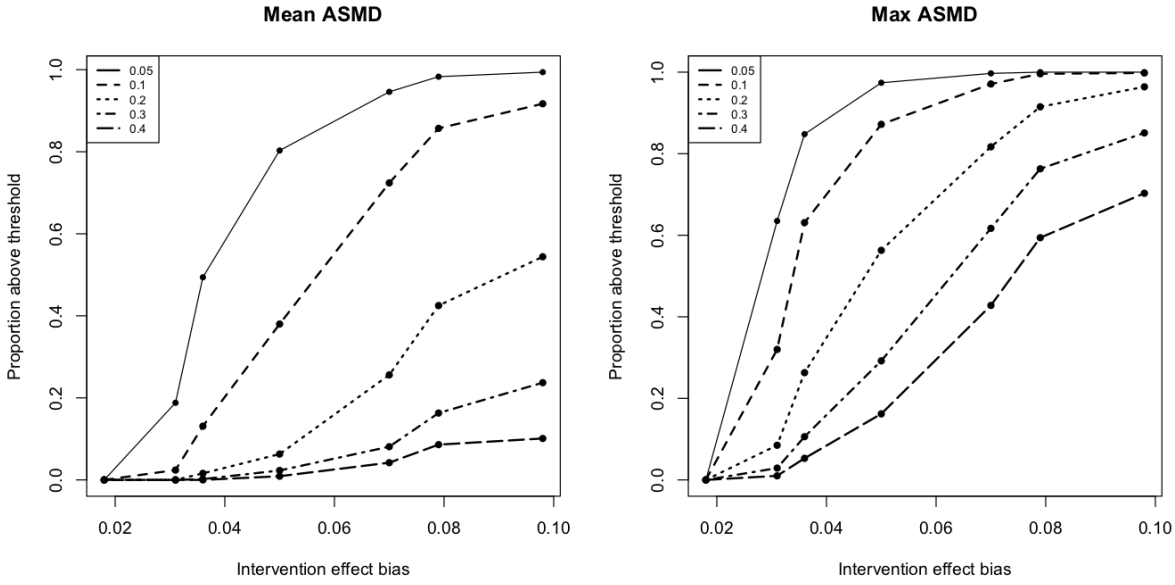


Figure 2: Simulation Study Results; proportion of simulation iterations above each threshold for Mean ASMD and Max ASMD summarized across 1000 replications; each point reflects one simulation setting, for settings 1-7, in order from left to right

Though we explore the Ben-Michael estimated bias metric as proposed in Ben-Michael (2018) [10], one could consider a standardized version wherein the bias is divided by the simple unweighted standard deviation of all post-intervention outcomes, thus making the metric on an effect size scale rather than on the scale of the outcome. Results for this standardized version are shown in Figure S6 in the Supplementary Materials.

Based on these results, we explore the use of the mean and max ASMD in the LA Firearm Letter Study Application below and utilize a threshold of 0.1, as this was identified as ideal for the mean ASMD, and is somewhat conservative for the max ASMD.

4 LA Firearm Letter Study

The focus of this study is to introduce a method for assessing the balance between a SC unit and a treated unit. To do this, we use an application to the LA Firearm Letter Study. Specifically, we use SCM to obtain a counterfactual for the LA firearm-related crime rate based on a weighted combination of control units in a donor pool of cities. We match on each year of the pre-intervention period, which has become common in the literature [10, 21–23, see], and then we calculate our proposed balance metrics for each matching factor (here, the preintervention outcomes).

Specifically, we first estimate SCM and calculate the balance metrics when using the full donor pool and all pre-intervention years for which we collect data. Given the data collection burden, we necessarily had to limit the data collection for the donor pool. The donor pool includes medium- to large-California CA cities and large, non-CA cities. We chose large agencies for two reasons: they are more likely to report data for the full 12 months of the year as does LA [24], and they may have a relatively similar urban density or confluence of people that would influence levels and rates of firearm violence. We also include medium-population cities in CA under the assumption they may have similar culture and policies as LA, thereby leading to similar rates of firearm violence absent the intervention. In sum, we selected cities that we anticipated would have more similar levels and rates of firearm violence. The full pre-intervention period is 2000 to 2012 inclusive, 13 years prior to the intervention. Then we explore the impact of changing the donor pool and number of pre-intervention years used in the SCM. The second donor pool is reduced to the CA cities only so that any state policy shock to gun crime would similarly affect the synthetic control and treated groups, thus isolating the impact of the gun letter policy. The second pre-intervention period is shorter, five years, to demonstrate the implication of pre-intervention durations on our balance metric results. Importantly, while we examine two different lengths for the pre-intervention time period, this is only done for the purpose of demonstrating the performance of the metrics when the time periods are different and does not imply that one should arbitrarily decrease the pre-intervention time period for analyses simply to get a better match.

Given the SCM is a comparative case study approach, we provide further details as to why the cities selected would be suitable for the donor pool to the comparison group. First, none of the cities adopted a similar intervention during the period of our study. The 10-day waiting period in CA makes it possible to send a letter in the period between purchasing and obtaining a firearm, and potentially dissuade 'straw purchasers' from obtaining a gun to transfer to someone else. Therefore, only states with a waiting period could feasibly implement a letter intervention. Ten other states have waiting periods of at least 3 days, allowing for time to send a letter that could prevent acquiring a firearm intended for 'straw purchase' [25]. According to our searches, no other cities had similar letter-writing interventions. Second, we restrict the donor pool to cities with characteristics that are similar to LA. As previously described, the cities included are relatively larger. As we will later show in the data subsection, other characteristics known to be correlated with firearm violence are also similar to LA. Third, it is important to exclude cities from the donor pool that may have experienced large shocks to firearm crime due to legislation during the intervention period. While government agencies around the country work to prevent gun violence through a number of local strategies, gun policy is an area where few policies have been adopted, and even fewer have been implemented that have large effects [32].

Finally, in an effort to understand the magnitude of the bias that may result in the intervention effect estimation given potential imbalance in the matching, we quantify the Ben-Michael estimated bias due to imbalance (also examined in the simulation study) by fitting an outcome model and producing an augmented SCM estimate, as suggested in Ben-Michael (2018) [10]. We use a simple linear model to predict the outcome in 2013 (first post-intervention year) using the lagged outcomes in previous years as the predictors (model fit only among controls). The bias is estimated as the difference in the predicted outcome when the model is applied to the treated lagged outcomes, and the average of the weighted predicted outcomes when the model is applied to the control lagged outcomes. The augmented estimate, as proposed in Ben-Michael (2018) [10], is then the weighted outcome in the treated group, plus this bias term.

The SCM is implemented in *R* 3.6.1 using synth [26] and the balance metric calculations are also written in *R*.

4.1 Data

This study uses Uniform Crime Report (UCR) data, available at Inter-university Consortium for Political and Social Research (ICSPR), because they have the feature of uniformly defining and collecting data across agencies, making the data particularly useful for exploiting cross-jurisdictional variation. UCR data is collected by the Federal Bureau of Investigation from law enforcement agencies submitting month-specific information on the number of incidents reported to law enforcement. Since we use reported crime and not actual crimes occurred, in the unlikely event that the intervention affected the propensity to report crimes, we might erroneously attribute the intervention effect to an actual change in crime rather than changes in reporting to police. It is reasonable to assume, however, the letter intervention did not affect individuals' willingness to report a murder, robbery, or aggravated assault with a firearm.

For this study, the relevant crime types are Part 1 violent offenses (homicide, rape and sexual assault, robbery, and aggravated assault); we exclude rape and sexual assault because the summarized data does not indicate whether a firearm was present. The police incident data include variables needed to study the gun letter program at the city-level, including crime type, whether a weapon was used, type of weapon, and where the incident occurred.

We use the UCR data at the police jurisdiction level, which is approximately a city. As previously described, we collected data on cities outside of California with a large population (more than 500,000 people) and California cities with medium to high population (more than 100,000 people). A key limitation of the data is that not all agencies report consistently over time [27]. As such, we only include cities without missing data [24]. Therefore, results are not driven by agencies stopping reporting firearm crime, *e.g.* missing data. Additionally, there can be errors in how data is reported by law enforcement. This may be due to pressure on some law enforcement agencies to make the numbers "look good", or because of changes in a department regarding who records crimes and a lack of training to record crimes properly. Indeed, we are aware of LA police department misclassifying aggravated assaults as simple assaults between 2005 and 2012; but that review, showed that LA police department did not misclassify assaults with a firearm [28]. While we cannot be certain about every city in the data set, we are not aware that the cities with the greatest weight in the synthetic control LA started misrecording and misreporting firearm crimes in 2013, when the intervention occurred.

The estimating sample comprises 103 jurisdictions, including Los Angeles, 32 large non-California cities with a population greater than 500,000 or more in at least one year since 1980, and 70 California cities with a population greater than 100,000 or more in at least one year since 1980. This provides a good mix of large jurisdictions from across the country and medium to large jurisdictions facing the same policies from the state. The final analytical sample consists of 1442 jurisdiction-year observations. This study was approved by RAND's Institutional Review Board.

4.2 Measurement

We utilize agency-level incident data from 103 agencies across the U.S. over a 14-year period (2001–2014) to derive the outcome measure of total crimes (murder, robbery, and aggravated assault) involving a firearm. The count of murders with a firearm is based on the total number of non-negligent murder and justifiable homicide events committed with a handgun, rifle, shotgun, or any other type of gun or firearm. The number of robberies with a firearm is the total number of known robbery offenses that were committed with any type of firearm. Unlike burglaries, robberies presume a victim that was hurt or threatened during the theft. And, the aggravated assault count is total number of known aggravated assault offenses that were committed with any type of firearm. Examples include attempted murder and threatening the victim. The classification of these offenses is based solely on police investigation, rather than the determination of a court, jury, or medical examiner, for example.

Given the differences in the sizes of cities and that SCM is not designed for count data, we generate total crimes involving a firearm as a rate per 1,000 population by adding the counts of murder, robbery, and

aggravated assault in each jurisdiction-year and use population data provided in the UCR dataset for each agency-year. Figure 3 shows the rates of total crimes with a gun during the pre-intervention period for LA and all donor pool cities.

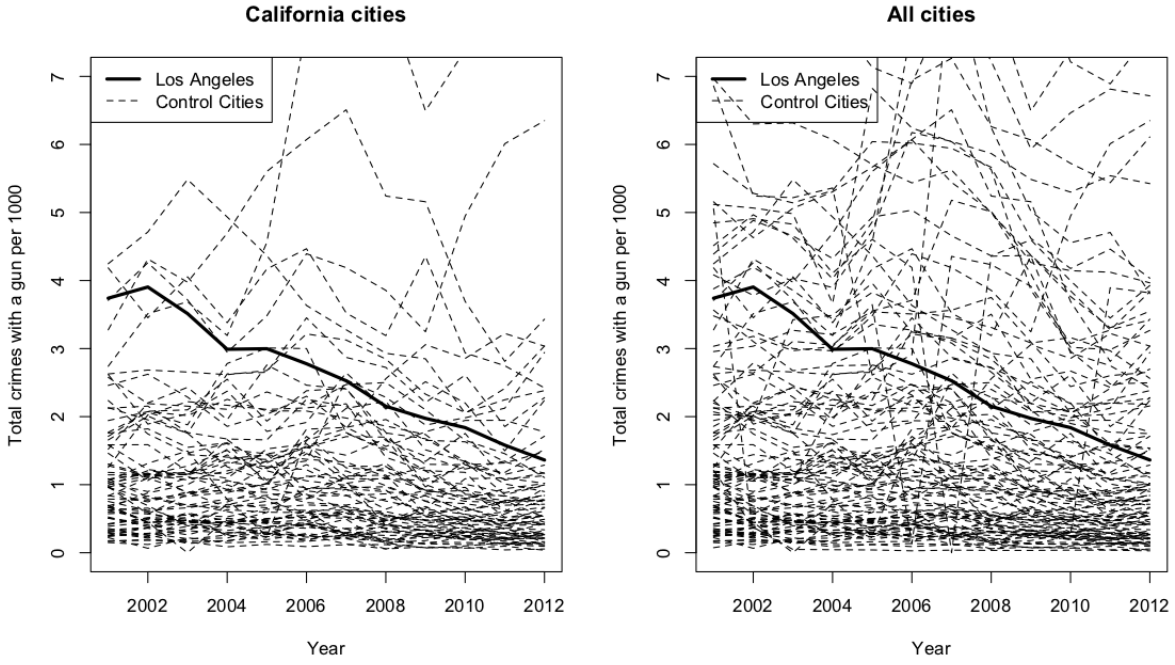


Figure 3: Total crimes with a gun per 1000 during the preintervention period among control cities

4.3 Results

Table 1 displays the calculated weights for each city (restricted to cities with weights greater than 0.005) using each combination and shows that the cities that get non-zero weights differ depending on the donor pool and pre-intervention duration; the full table of cities and weights is shown in Table S4 in the Supplementary Material. When using CA cities only, changing the pre-intervention duration changes the subset of cities that receive a non-zero weight with the exception of a single city, Inglewood (a city adjacent to the city of LA and within LA county), which receives a non-zero weight with either duration. When using all years as the duration, changing the donor pool moderately changes the cities that receive a non-zero weight - two cities, San Bernardino and Fontana, receive a non-zero weight with either duration but the remaining two cities that receive a non-zero weight differ. When using 5 years as the duration, changing the donor pool also changes the cities with non-zero weights with South Gate and Norwalk selected using either duration but the remaining cities differing. Using 5 years and all cities, in particular, results in several cities with small non-zero weights (see Table S4 in the Supplementary Material). Importantly, since letters were sent to all handgun purchasers (purchased in LA) residing in city of LA zip codes, including individuals living in zip codes that extended outside of the city limits, we investigated potential treatment contamination by examining whether any of the cities comprising synthetic LA were cities with at least one LA zip code. Only one city, Inglewood, with positive estimated weights had zip codes that overlap with Los Angeles. Specifically, two zip codes (of the 129 Los Angeles zip codes) overlap with two of the twelve zip codes of Inglewood. Since Inglewood has weights of 0.02 to 0.05 in three of the four estimates and 0.00 in a fourth estimate, it is unlikely that treatment contamination or spillover will drive our results described below.

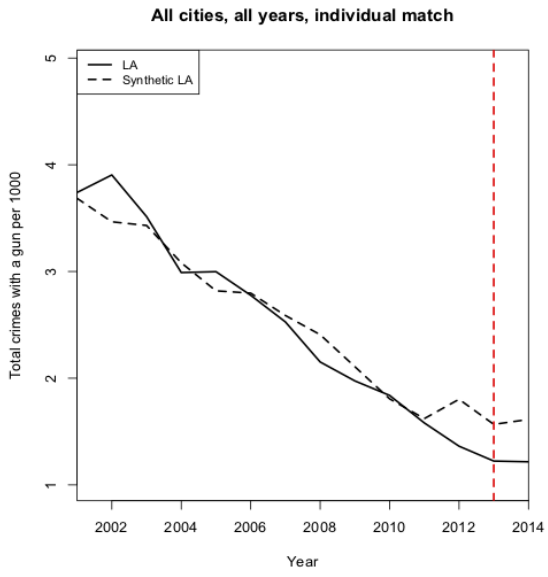
Table 1: Calculated weights for each city, by donor pool and pre-intervention duration (CA = California), ordered by magnitude of weight (restricted to cities with a weight greater than 0.005)

Donor pool: CA cities only				**Donor pool: All cities**			
All years		**5 years**		**All years**		**5 years**	
Weights	Cities	Weights	Cities	Weights	Cities	Weights	Cities
0.472	SAN BERNARDINO	0.534	SOUTH GATE	0.557	FONTANA	0.274	SOUTH GATE
0.251	FONTANA	0.239	NORWALK	0.28	DALLAS	0.258	EL MONTE
0.227	SANTA ANA	0.168	FRESNO	0.105	SAN BERNARDINO	0.177	COLUMBUS
0.05	INGLEWOOD	0.042	INGLEWOOD	0.058	VISALIA	0.077	NORWALK
		0.016	STOCKTON			0.057	NEW ORLEANS
						0.028	INGLEWOOD

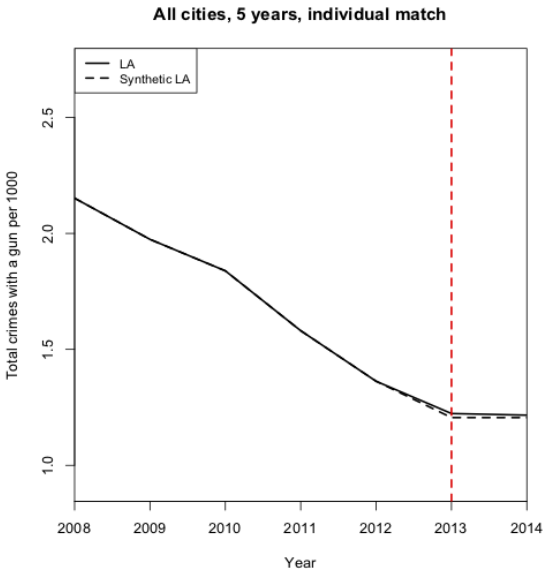
Figure 4 displays a typical SCM illustration in which synthetic LA is compared to real LA in the pre-intervention period (note that the gap plot is created by taking the difference between the two plotted lines). Panels A and B demonstrate the matches when using all cities in the donor pool and the pre-intervention duration is 12 years (panel A) versus 5 years (panel B). Panels C and D demonstrate the matches when using CA cities only in the donor pool and the pre-intervention duration is 12 years (panel C) versus 5 years (panel D). Figure 4 shows that using all years of the pre-intervention period (Panels A and C) results in matches that are not ideal, while using 5 years results in very good matching (Panels B and D). However, the question is – are the matches shown in Panel A and C bad enough that we should not proceed with inference? While it is obvious from Figure 4 that using only 5 years is better, arbitrarily choosing a shorter pre-intervention period would be concerning. In particular, a shorter time period may result in omitting potentially important information in the pre-intervention time period. As with many statistical problems, there is always a necessary trade-off between the desire to have an adequate match or fit and concerns about over-fitting. On the one hand, SCM is subject to the curse of dimensionality whereby the probability that exact balancing weights exist vanishes as the number of time periods grows [29]. On the other hand, improving the fit by reducing the pre-intervention period can lead to poor estimates of the LA counterfactual if it fails to capture important trends that are predictive of the outcome.

When we visually examine Panel B and D, using CA cities results in a relatively worse match than when using all cities. This is expected and is especially not surprising as LA is the largest city in California and unique in its composition and geographic dispersion, making it very different from other California cities but potentially similar to other large cities outside of California. It is also worth noting here that if we were to use CA cities only with a pre-intervention duration of five years, we observe a potentially "good match", and the impact of the intervention may be significant with LA experiencing *greater* gun violence after the intervention than the synthetic LA. In panel B, however, using all cities seems to result in a better match, and there appears to be relatively little difference between synthetic LA and treated LA (*i.e.*, no impact of the intervention).

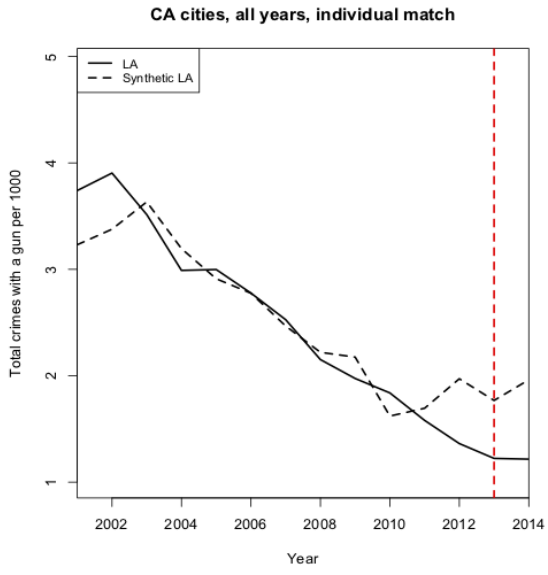
Notably, after visually inspecting Figure 4, it becomes more apparent why these figures are an insufficient way to determine whether the match is good enough. Table 2 illustrates the use of the two metrics we propose for assessing SCM matching quality, the maximum ASMD and the mean ASMD, where the ASMD is calculated for each pre-intervention year. These values indicate that matching is clearly not adequate when using CA cities only with a pre-intervention period of 12 years (max ASMD = 0.535, mean ASMD = 0.196). It is also clear that there are relatively small imbalances between synthetic LA and LA when using a donor pool of CA cities only or all cities with a pre-intervention duration of five years, since both the mean and max ASMD are well below the 0.10 threshold. The assessment is less clear when using a donor pool of all cities and a pre-intervention period of 12 years; the max ASMD indicates a moderate imbalance (0.225), while the mean ASMD suggests a small imbalance (0.077). Given the foundational SCM papers suggest only conducting inference if a good match is available [12], we would recommend not conducting inference if any one of the metrics is



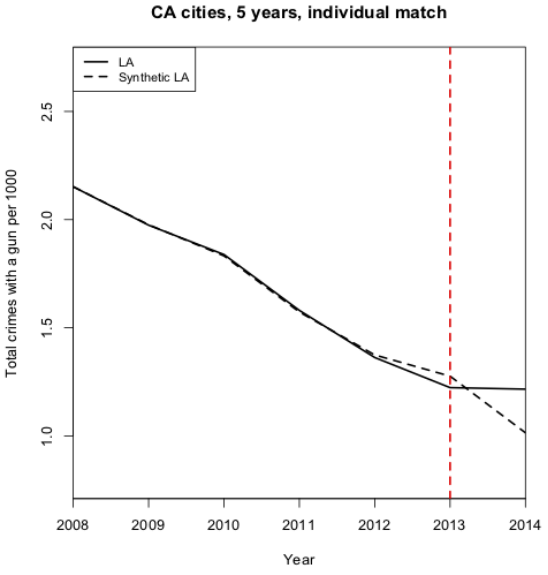
(a) Panel A



(b) Panel B



(c) Panel C



(d) Panel D

Figure 4: Outcome over years for treated group (Los Angeles [LA]) and the synthetic control (synthetic LA); Panel A uses all cities and all years for matching; Panel B uses all cities and 5 years for matching; Panel C uses California (CA) cities only and all years for matching; Panel D uses CA cities only and 5 years for matching; individual match indicates the matching was based on each individual pre-intervention year outcome measure

Table 2: Max and mean Absolute Standardized Mean Difference (ASMD), by donor pool and pre-intervention duration (CA = California)

	Donor pool: CA cities only		Donor pool: All cities	
	All years	5 years	All years	5 years
max ASMD	0.535	0.007	0.225	0.006
mean ASMD	0.196	0.004	0.077	0.004

greater than 0.1. Therefore, while Panel A may seem like a good enough match if examining the gap plots, these balance metrics would be one way to objectively determine the match is in fact insufficient.

Table 3 shows the Ben-Michael estimated bias due to imbalance using the augmented SCM approach of Ben-Michael *et al.* (2018) [10]. We show the value of total number of gun crimes per 1,000 in LA in 2013 (row A), for synthetic LA (row B), and the differential (row C). We then show the augmented SCM weighted synthetic LA (row D) and the differential with real LA (row E). The Ben-Michael estimated bias due to imbalance is the difference between the augmented synthetic LA and synthetic LA, or equally, the difference between the two estimated intervention effects (row F). The results show that the Ben-Michael estimated bias is large when using all years – large enough that the estimate of the intervention flips from a reduction in gun crime rate (−0.544 and −0.344) to an increase in gun crime rate (0.495 and 0.251). When using only five years for matching, the Ben-Michael estimated bias is much smaller (−0.018 and 0.0). Using the match with the least amount of Ben-Michael estimated bias, the intervention effect is close to zero (0.017).

Table 3: Estimated bias due to imbalance, by donor pool and pre-intervention duration (CA = California; LA = Los Angeles; SCM = synthetic control method)

		Donor pool: CA cities only		Donor pool: All cities	
		All years	5 years	All years	5 years
A	Outcome in LA in 2013 (total crimes per 1,000)	1.223	1.223	1.223	1.223
B	Outcome in synthetic control in 2013 using SCM weights (total crimes per 1,000)	1.768	1.276	1.567	1.206
C	Intervention effect in 2013 using SCM (A-B) (total crimes per 1,000)	−0.544	−0.053	−0.344	0.017
D	Outcome in synthetic control in 2013 using Augmented SCM weights (total crimes per 1,000)	0.728	1.259	0.973	1.206
E	Intervention effect in 2013 using Augmented SCM (A-D) (total crimes per 1,000)	0.495	−0.035	0.251	0.017
F	Ben-Michael estimated bias due to imbalance (D-B) or (E-C) (total crimes per 1,000)	−1.04	−0.018	−0.595	0

5 Conclusion

In this study, we present metrics to help analysts determine whether they have a good enough match to conduct causal inference using SCM. The challenge with identifying relevant balance metrics for an SCM application is that there is only one treated unit, and most test statistics commonly used to assess balance require a distribution in both groups (*e.g.* Kolmogorov Smirnov (KS) test). This paper suggests the field can learn from recent work in the statistical literature to improve how to assess balance; see for example Griffin *et al.* (2014) [30]. Researchers need better guidance on what is a meaningful difference; in addition to examining gap plots, we assess the implications of using the mean and max ASMD balance metrics traditionally used in propensity score methods. The ASMD tell us more clearly when there may be a problem of equivalence between control and treatment groups, with an established threshold (*e.g.* 0.10) for the mean of the covariate balance metrics (mean ASMD) or the maximum of the balance metrics (max ASMD). In addition, while recent work on SCM has explored methods to improve matches including relaxing the convex hull constraint, allowing the sum of the weights to exceed one, and allowing some weights to be negative [21], this metric could similarly be useful within these alternative methods. Rather than relying on the ASMD as a strict decision rule, we encourage the use of this metric in addition to the gap plot, to allow researchers to further understand the quality of the

match and to provide a quantifiable way of measuring and comparing match quality. Importantly, while we examined ASMD and various potential thresholds across a range of simulation scenarios, future use of this metric would benefit from further theoretical development to understand its properties in a general setting.

In an application to the LA Firearm Letter Study, we show that while the gap plots were not ideal, one could argue that they seemed "good enough", and we demonstrate how decision-making can be improved by using ASMD. The mean and max ASMD clearly show poor matches and good matches, with a borderline case in which the max ASMD value would lead to a rejection of the match and the mean ASMD would result in a conclusion of a small imbalance. In that case, we argue a researcher should carefully consider the implications of conducting inference. We recommend conducting causal inference only if the ASMD values are below the threshold of 0.10. Two lengths of the pre-intervention time period were examined in the LA Firearm Letter Study – 12 years and 5 years – for the purpose of illustrating the proposed metrics. However, it is important that one not arbitrarily choose a shorter time period simply to obtain a better match, *i.e.* to obtain a metric value less than 0.10, as this can lead to poor estimates of the treatment counterfactual if the utilized time period fails to capture important trends that are predictive of the outcome.

Our proposed metrics along with the alternative metrics examined in the simulation study highlight an important area of potential future research – while the standard synthetic control approach chooses weights as a solution to the constrained optimization problem shown in equation shown above, one might consider selecting weights according to some other optimization function. For example, if the aim is to minimize the mean AMSD, the metric itself could be used as the optimization function. More generally, there are a variety of methods that have been recently proposed within the propensity score literature in an effort to identify optimal weights to balance treated and control groups in an observational setting, and these methods may extend nicely to the SCM. For example, one could consider an approach similar to the stable balancing weights proposed in Zubizarreta (2015) [31] which would constrain the absolute difference in means of the weighted pre-intervention outcomes to be less than user-specified thresholds. Such alternative approaches may be attractive if one is not able to obtain a good match with the standard SCM weight selection.

Some limitations of our study are as follows. First, our data was limited to jurisdictions without missing data and thus, may not be representative of all jurisdictions that would have been eligible based on size. It is possible that different or better matches could be obtained if these data were available. That said, the focus of our study is to demonstrate the properties of a balance metric, and we would not expect that this lack of data would affect the intuition behind our proposed approach. Second, as mentioned earlier, it was important to carefully consider and possibly exclude cities from the donor pool that may have experienced large shocks to firearm crime due to legislation during the intervention period. Using the RAND (2020) [32] Gun Policy in America database [32], we reviewed the extent to which any of our non-CA cities contributing more than 0.01 weight to synthetic LA had key policy changes in the pre- or post-period of the treatment. For the analysis using a 5-year pre-period (2008-2013), there were two non-CA cities with policy changes including Columbus, OH (weight = 0.18) and New Orleans, LA (weight = 0.06). Ohio made one law effective in the pre-period, 2008, that further expanded Castle Doctrine (which designates a person's legally occupied place as a place in which that person has protections and immunities permitting one, in certain circumstances, to use force to defend oneself) to include vehicles (Ohio Rev. Code Ann. § 2901.09 [33]). Louisiana passed one law in the post-period, 2014, that extended a prohibited possessor law to individuals with domestic violence restraining order (LA Rev. Stat. § 14:95.10 [34]). While we cannot rule out that this may contribute to the findings, these two cities contributed less than a quarter of the weight to synthetic Los Angeles and passed two different laws, years apart. Therefore, we would argue it is unlikely these policy changes affected the matching. For the analysis including all years of data, one city outside of California had a positive weight, Dallas, TX (weight = 0.28), where in 2007, the state expanded its law to include anywhere a person has the legal right to be (*i.e.* they do not have the duty to retreat), commonly referred to as Stand Your Ground Law (Tex. Penal Code § 9.32(C) [35]). Again, while we cannot rule out the policy change may have affected the matching, it would be very limited since we match on 12 years of data, each year individually. In addition, we focused only on a single outcome in this study – total crimes which included murder, robbery, and aggravated assault. It is possible that results may be different if each crime type was examined individually. Lastly, to assess the estimated bias of imbalance, we used a simple linear model to predict the outcome in first post-intervention year as

suggested by Ben-Michael *et al.* (2018) [10]. Notably, this is not necessarily the true bias (only an estimate) and more flexible outcome models could be considered to accommodate potential model mis-specification.

References

[1] Bouttell, J., Craig, P., Lewsey, J., Robinson, M., and Popham, F. (2018). Synthetic control methodology as a tool for evaluating population-level health interventions, Journal of Epidemiology and Community Health, 72: 673–78.

[2] Saunders, J., Lundberg, R., Braga, A. A., Ridgeway, G., & Miles, J. (2015). A synthetic control approach to evaluating place-based crime interventions. Journal of Quantitative Criminology, 31(3), 413–434.

[3] Robbins, M. W., Saunders, J., & Kilmer, B. (2017). A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. Journal of the American Statistical Association, 112(517), 109–126.

[4] Donohue, J. J., Aneja, A., & Weber, K. D. (2019). Right-to-Carry Laws and Violent Crime: A Comprehensive Assessment Using Panel Data and a State-Level Synthetic Control Analysis. Journal of Empirical Legal Studies, 16(2), 198–247.

[5] Pinotti, P. (2015). The economic costs of organised crime: Evidence from Southern Italy. The Economic Journal, 125(586), F203–F232.

[6] Loeffler, C. E., & Chalfin, A. (2017). Estimating the crime effects of raising the age of majority: Evidence from Connecticut. Criminology & Public Policy, 16(1), 45–71.

[7] Ridgeway, G., Pierce, G.L., Tita, G., and Wintemute, G. (2008). Strategies for disrupting illegal firearm markets: A case study of Los Angeles. Rand Corporation Report: TR-512-NIJ.

[8] Ridgeway, G., Braga, A.A., Tita, G., and Pierce, G.L. (2011). Intervening in gun markets: an experiment to assess the impact of targeted gun-law messaging. Journal of Experimental Criminology, 7: 103–09.

[9] Hunt, P.E., Parast, L., and Weinberger, G. (2017). Can an Informative Letter Reduce Gun Crime and Be Cost-Effective: A Study of Los Angeles. Santa Monica, CA: RAND Corporation.

[10] Ben-Michael, E., Feller, A., and Rothstein, J. (2018). The augmented synthetic control method, arXiv preprint arXiv:1811.04170.

[11] Abadie, A., and Gardeazabal J. (2003). The economic costs of conflict: A case study of the Basque Country. American Economic Review, 93: 113–32.

[12] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105: 493–505.

[13] Abadie, A., Diamond, A. and Hainmueller, J. (2015). Comparative politics and the synthetic control method. American Journal of Political Science, 59(2): 495–510.

[14] Athey, S. and Imbens, G.W., (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives. 31(2): 3–32.

[15] Griffin, B. A., McCaffrey, D. F., Almirall, D., Burgette, L. F., & Setodji, C. M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. Journal of causal inference, 5(2).

[16] Austin , P. C. (2008). A critical appraisal of propensity score matching in the medical literature from 1996 to 2003. Statistics in Medicine. 27: 2037–2049.

[17] Austin, P. C., Mamdani, M. M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Statistics in Medicine. 25: 2084–2106.

[18] Austin, P. C., Grootendorst, P., Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Statistics in Medicine. 26: 734–753.

[19] Ho, D. E., Imai, K., King, G., Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 15: 199–236.

[20] Cohen, J. S. (1988). Statistical Power Analysis for the Behavioral Sciences. Erlbaum: Hillsdale, NJ.

[21] Doudchenko, N., & Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis (No. w22791). National Bureau of Economic Research.

[22] Powell, D. (2018). Imperfect Synthetic Controls: Did the Massachusetts Health Care Reform Save Lives?. Working Paper. Accessed September 26, 2019: https://www.rand.org/content/dam/rand/pubs/working_papers/WR1200/WR1246/RAND_WR1246.pdf

[23] Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W. and Wager, S., 2019. *Synthetic difference in differences* (No. w25532). National Bureau of Economic Research.

[24] Lynch, J. P., & Jarvis, J. P. (2008). Missing data and imputation in the uniform crime reports and the effects on national estimates. Journal of Contemporary Criminal Justice, 24(1), 69–85.

[25] Giffords Law Center to Prevent Gun Violence. Waiting Periods. https://lawcenter.giffords.org/gun-laws/policy-areas/gun-sales/waiting-periods/ Accessed February 21, 2020.

[26] Abadie, A., Diamond, A., & Hainmueller, J. (2011). Synth: AnRPackage for Synthetic Control Methods in Comparative Case Studies. Journal of Statistical Software, 42(13). doi:10.18637/jss.v042.i13
[27] Committee on National Statistics Council (CNSC): Division on Behavioral and Social Sciences and Education; National Research. (2014). Data from Law Enforcement Agencies. in Kruttschnitt C, Kalsbeek WD and House CC (eds.), Panel on Measuring Rape and Sexual Assault in Bureau of Justice Statistics Household Surveys. National Academies Press Washington DC.
[28] Poston, Ben. LAPD records reveal flaws in crime reporting. Los Angeles Times. https://documents.latimes.com/lapd-crime-data/ Accessed Feburary 21,2020.
[29] Ferman, B. and C. Pinto (2018). Synthetic controls with imperfect pre-treatment fit.
[30] Griffin, BA., Ramchand, R., Almirall, D., Slaughter, M.E., Burgette, L.F, and McCaffery, D.F. (2014). Estimating the causal effects of cumulative treatment episodes for adolescents using marginal structural models and inverse probability of treatment weighting. Drug and Alcohol Dependence, 136: 69–78.
[31] Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association, 110(511), 910-922.
[32] RAND Corporation. Gun Policy in America. https://www.rand.org/research/gun-policy.html Accessed February 21, 2020.
[33] Ohio Revised Code Annotated § 2901.09.
[34] Louisiana Revised Statute § 14:95.10.
[35] Texas Penal Code § 9.32(C).

Supplemental Material

Figure S5 shows the proportion of simulation iterations with a calculated metric above each threshold vs. the intervention effect bias for each of these metrics, using threshold values 0.05, 0.1, 0.2, 0,3, and 0.4. This figure shows that a threshold of 0.2 for RMSE and SRMSE has desirable properties, while for estimated bias, 0.3 appears more desirable due to the smaller scale for the estimated bias. Importantly, while these thresholds may be reasonable to examine in this simulation, because RMSE and Ben-Michael estimated bias are relative to the scale of the outcome, a different threshold would need to be considered depending on the scale of the outcome. As discussed in the main text, this is our motivation behind focusing on the ASMD metrics which are on the effect size scale. Figure S6 shows simulation results for a standardized version of the Ben-Michael estimated bias. Table S4 shows the full set of estimated weights from the LA Firearm Letter Study.

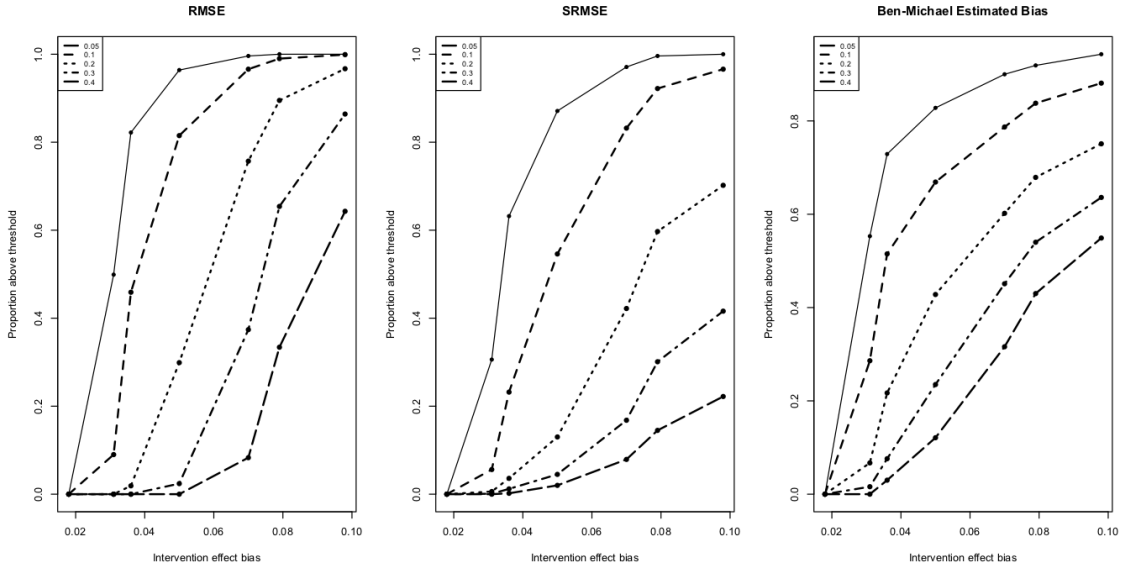


Figure S5: Simulation Study Results; proportion of simulation iterations above each threshold for RMSE, SRMSE, and estimated bias summarized across 1000 replications; each point reflects one simulation setting, for settings 1-7, in order from left to right

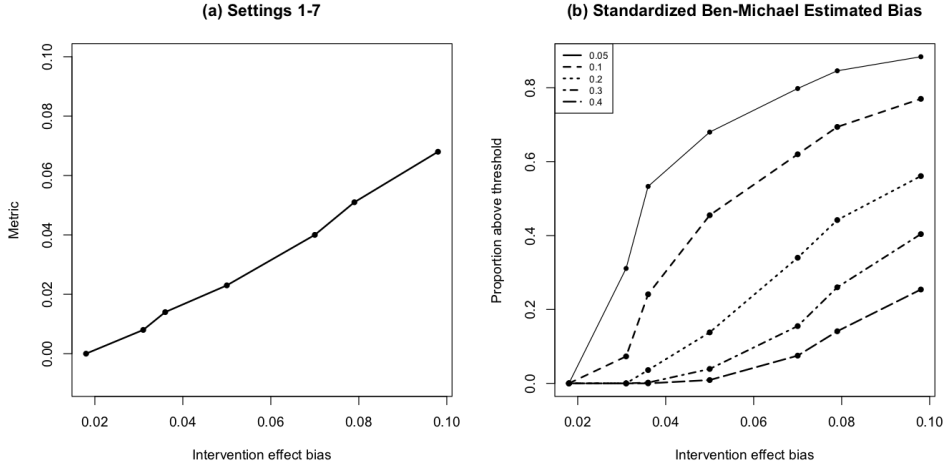


Figure S6: Simulation Study Results examining a standardized version of the Ben-Michael estimated bias metric; (a) metric versus bias in intervention effect estimate averaged across 1000 replications where each point reflects one simulation setting, for settings 1-7, in order from left to right, (b) proportion of simulation iterations above each threshold for the metric summarized across 1000 replications; each point reflects one simulation setting, for settings 1-7, in order from left to right

Table S4: Calculated weights for each city, by donor pool and pre-intervention duration (CA = California), ordered by magnitude of weight

Donor pool: CA cities only				**Donor pool: All cities**			
All years		**5 years**		**All years**		**5 years**	
Weights	Cities	Weights	Cities	Weights	Cities	Weights	Cities
0.472	SAN BERNARDINO	0.534	SOUTH GATE	0.557	FONTANA	0.274	SOUTH GATE
0.251	FONTANA	0.239	NORWALK	0.28	DALLAS	0.258	EL MONTE
0.227	SANTA ANA	0.168	FRESNO	0.105	SAN BERNARDINO	0.177	COLUMBUS
0.05	INGLEWOOD	0.042	INGLEWOOD	0.058	VISALIA	0.077	NORWALK
0	ANAHEIM	0.016	STOCKTON	0	ALBUQUERQUE	0.057	NEW ORLEANS
0	ANTIOCH	0	ANAHEIM	0	ANAHEIM	0.028	INGLEWOOD
0	BAKERSFIELD	0	ANTIOCH	0	ANTIOCH	0.005	CONCORD
0	BERKELEY	0	BAKERSFIELD	0	ATLANTA	0.005	FREMONT
0	BURBANK	0	BERKELEY	0	AUSTIN	0.005	HAYWARD
0	CARLSBAD	0	BURBANK	0	BAKERSFIELD	0.004	GARDEN GROVE
0	CHULA VISTA	0	CARLSBAD	0	BALTIMORE	0.004	HUNTINGTON BEACH
0	CLOVIS	0	CHULA VISTA	0	BERKELEY	0.004	SANTA ANA
0	CONCORD	0	CLOVIS	0	BOSTON	0.003	AUSTIN
0	CORONA	0	CONCORD	0	BURBANK	0.003	EL CAJON
0	COSTA MESA	0	CORONA	0	CARLSBAD	0.003	ORANGE
0	DALY CITY	0	COSTA MESA	0	CHARLOTTE-MECKLENBURG	0.003	VICTORVILLE
0	DOWNEY	0	DALY CITY	0	CHICAGO	0.002	BURBANK
0	EL CAJON	0	DOWNEY	0	CHULA VISTA	0.002	CARLSBAD
0	EL MONTE	0	EL CAJON	0	CLEVELAND	0.002	CHICAGO
0	ESCONDIDO	0	EL MONTE	0	CLOVIS	0.002	CLOVIS
0	FAIRFIELD	0	ESCONDIDO	0	COLUMBUS	0.002	COSTA MESA
0	FREMONT	0	FAIRFIELD	0	CONCORD	0.002	DALY CITY
0	FRESNO	0	FONTANA	0	CORONA	0.002	DOWNEY
0	FULLERTON	0	FREMONT	0	COSTA MESA	0.002	FRESNO
0	GARDEN GROVE	0	FULLERTON	0	DALY CITY	0.002	GLENDALE
0	GLENDALE	0	GARDEN GROVE	0	DENVER	0.002	HONOLULU
0	HAYWARD	0	GLENDALE	0	DETROIT	0.002	IRVINE
0	HUNTINGTON BEACH	0	HAYWARD	0	DOWNEY	0.002	MURRIETA

Table S4: ... continued

Donor pool: CA cities only				**Donor pool: All cities**			
All years		**5 years**		**All years**		**5 years**	
Weights	Cities	Weights	Cities	Weights	Cities	Weights	Cities
0	IRVINE	0	HUNTINGTON BEACH	0	EL CAJON	0.002	NEW YORK
0	LANCASTER	0	IRVINE	0	EL MONTE	0.002	RANCHO CUCAMONGA
0	LONG BEACH	0	LANCASTER	0	EL PASO	0.002	RIVERSIDE
0	MODESTO	0	LONG BEACH	0	ESCONDIDO	0.002	ROSEVILLE
0	MORENO VALLEY	0	MODESTO	0	FAIRFIELD	0.002	SAN MATEO
0	MURRIETA	0	MORENO VALLEY	0	FORT WORTH	0.002	SANTA CLARA
0	NORWALK	0	MURRIETA	0	FREMONT	0.002	SANTA CLARITA
0	OAKLAND	0	OAKLAND	0	FRESNO	0.002	SANTA MARIA
0	OCEANSIDE	0	OCEANSIDE	0	FULLERTON	0.002	SANTA ROSA
0	ONTARIO	0	ONTARIO	0	GARDEN GROVE	0.002	SIMI VALLEY
0	ORANGE	0	ORANGE	0	GLENDALE	0.002	SUNNYVALE
0	OXNARD	0	OXNARD	0	HAYWARD	0.002	TEMECULA
0	PALMDALE	0	PALMDALE	0	HONOLULU	0.002	THOUSAND OAKS
0	PASADENA	0	PASADENA	0	HOUSTON	0.001	ALBUQUERQUE
0	POMONA	0	POMONA	0	HUNTINGTON BEACH	0.001	ANAHEIM
0	RANCHO CUCAMONGA	0	RANCHO CUCAMONGA	0	INDIANAPOLIS	0.001	BAKERSFIELD
0	RIALTO	0	RIALTO	0	INGLEWOOD	0.001	BERKELEY
0	RICHMOND	0	RICHMOND	0	IRVINE	0.001	BOSTON
0	RIVERSIDE	0	RIVERSIDE	0	JACKSONVILLE	0.001	CHULA VISTA
0	ROSEVILLE	0	ROSEVILLE	0	LANCASTER	0.001	CORONA
0	SACRAMENTO	0	SACRAMENTO	0	LAS VEGAS METRO PD	0.001	DENVER
0	SALINAS	0	SALINAS	0	LONG BEACH	0.001	EL PASO
0	SAN DIEGO	0	SAN BERNARDINO	0	MEMPHIS	0.001	ESCONDIDO
0	SAN FRANCISCO	0	SAN DIEGO	0	MILWAUKEE	0.001	FAIRFIELD
0	SAN JOSE	0	SAN FRANCISCO	0	MODESTO	0.001	FONTANA
0	SAN MATEO	0	SAN JOSE	0	MORENO VALLEY	0.001	FORT WORTH
0	SANTA CLARA	0	SAN MATEO	0	MURRIETA	0.001	FULLERTON
0	SANTA CLARITA	0	SANTA ANA	0	NASHVILLE	0.001	LANCASTER
0	SANTA MARIA	0	SANTA CLARA	0	NEW ORLEANS	0.001	LAS VEGAS METRO PD

Table S4: ... continued

Donor pool: CA cities only				**Donor pool: All cities**			
All years		**5 years**		**All years**		**5 years**	
Weights	Cities	Weights	Cities	Weights	Cities	Weights	Cities
0	SANTA ROSA	0	SANTA CLARITA	0	NEW YORK	0.001	LONG BEACH
0	SIMI VALLEY	0	SANTA MARIA	0	NORWALK	0.001	MODESTO
0	SOUTH GATE	0	SANTA ROSA	0	OAKLAND	0.001	MORENO VALLEY
0	STOCKTON	0	SIMI VALLEY	0	OCEANSIDE	0.001	OCEANSIDE
0	SUNNYVALE	0	SUNNYVALE	0	OKLAHOMA CITY	0.001	ONTARIO
0	TEMECULA	0	TEMECULA	0	ONTARIO	0.001	OXNARD
0	THOUSAND OAKS	0	THOUSAND OAKS	0	ORANGE	0.001	PALMDALE
0	TORRANCE	0	TORRANCE	0	OXNARD	0.001	PASADENA
0	VALLEJO	0	VALLEJO	0	PALMDALE	0.001	PORTLAND
0	VENTURA	0	VENTURA	0	PASADENA	0.001	SAN ANTONIO
0	VICTORVILLE	0	VICTORVILLE	0	PHILADELPHIA	0.001	SAN DIEGO
0	VISALIA	0	VISALIA	0	PHOENIX	0.001	SAN FRANCISCO
0	WEST COVINA	0	WEST COVINA	0	POMONA	0.001	SAN JOSE
				0	PORTLAND	0.001	SEATTLE
				0	RANCHO CUCAMONGA	0.001	TORRANCE
				0	RIALTO	0.001	VALLEJO
				0	RICHMOND	0.001	VENTURA
				0	RIVERSIDE	0.001	VISALIA
				0	ROSEVILLE	0.001	WEST COVINA
				0	SACRAMENTO	0	ANTIOCH
				0	SALINAS	0	ATLANTA
				0	SAN ANTONIO	0	BALTIMORE
				0	SAN DIEGO	0	CHARLOTTE-MECKLENBURG
				0	SAN FRANCISCO	0	CLEVELAND
				0	SAN JOSE	0	DALLAS
				0	SAN MATEO	0	DETROIT
				0	SANTA ANA	0	HOUSTON
				0	SANTA CLARA	0	INDIANAPOLIS

Table S4: ... continued

Donor pool: CA cities only				Donor pool: All cities			
All years		**5 years**		**All years**		**5 years**	
Weights	Cities	Weights	Cities	Weights	Cities	Weights	Cities
				0	SANTA CLARITA	0	JACKSONVILLE
				0	SANTA MARIA	0	MEMPHIS
				0	SANTA ROSA	0	MILWAUKEE
				0	SEATTLE	0	NASHVILLE
				0	SIMI VALLEY	0	OAKLAND
				0	SOUTH GATE	0	OKLAHOMA CITY
				0	STOCKTON	0	PHILADELPHIA
				0	SUNNYVALE	0	PHOENIX
				0	TEMECULA	0	POMONA
				0	THOUSAND OAKS	0	RIALTO
				0	TORRANCE	0	RICHMOND
				0	TUCSON	0	SACRAMENTO
				0	VALLEJO	0	SALINAS
				0	VENTURA	0	SAN BERNARDINO
				0	VICTORVILLE	0	STOCKTON
				0	WASHINGTON	0	TUCSON
				0	WEST COVINA	0	WASHINGTON
				0	MILWAUKEE		