Current HIV/AIDS Reports (2024) 21:140–151
https://doi.org/10.1007/s11904-024-00695-z

REVIEW

Synthetic Controls for Implementation Science: Opportunities for HIV Program Evaluation Using Routinely Collected Data

Sara Wallach[1,2] · Suzue Saito[2] · Harriet Nuwagaba-Biribonwoha[1,2] · Lenhle Dube[3] · Matthew R. Lamb[1,2]

Accepted: 28 February 2024 / Published online: 13 March 2024

Abstract

Purpose of Review HIV service delivery programs are some of the largest funded public health programs in the world. Timely, efficient evaluation of these programs can be enhanced with methodologies designed to estimate the effects of policy. We propose using the synthetic control method (SCM) as an implementation science tool to evaluate these HIV programs.

Recent Findings SCM, introduced in econometrics, shows increasing utility across fields. Key benefits of this methodology over traditional design-based approaches for evaluation stem from directly approximating pre-intervention trends by weighting of candidate non-intervention units. We demonstrate SCM to evaluate the effectiveness of a public health intervention targeting HIV health facilities with high numbers of recent infections on trends in pre-exposure prophylaxis (PrEP) enrollment.

Summary This test case demonstrates SCM's feasibility for effectiveness evaluations of site-level HIV interventions. HIV programs collecting longitudinal, routine service delivery data for many facilities, with only some receiving a time-specified intervention, are well-suited for evaluation using SCM.

Keywords Synthetic control method · Program evaluation · HIV PrEP · Design-based methods · Implementation science

Introduction

The synthetic control method (SCM) was first described by Abadie and Gardeazabal in 2003 and expanded on by Abadie, Diamond, and Hainmueller in 2010. It is a design-based approach to causal inference related to difference-in-difference (DiD) analyses, employed when randomized controlled trials are not feasible, where researchers evaluate the effects of an intervention across both space and time [1, 2]. In traditional DiD analyses, a non-intervention comparison unit is chosen via matching; parallel trends in outcome characteristics between intervention and matched non-intervention units in this pre-intervention period are used as evidence of the absence of time-varying confounding. DiD analyses are extremely useful when matched comparison groups are available.

However, in program implementation and evaluation, facilities selected to receive an intervention are often decidedly different from those not receiving the intervention on factors related to the outcome. There are often non-random reasons for why we intervene where we do. In these situations, finding a comparison unit appropriate for a valid DiD analysis is challenging.

SCM provides a way forward by manufacturing an appropriate comparison group from a weighted combination of possible non-intervention units. This method, transparent and data-driven in its comparison unit selection, meets many of the assumptions inherent in DiD analyses and has been labeled "arguably the most important innovation in the policy evaluation literature in the last 15 years" [3].

We pose that SCM should be used more frequently in routine program monitoring and evaluation, specifically for large-scale HIV service delivery programs. Although SCM is currently most widely used in policy evaluation, the types of questions often being asked, and the types of data routinely available for use in conducting program evaluations, are directly analogous to the situations where SCM has been most useful. For example, HIV and related program service

✉ Sara Wallach
sara.wallach@columbia.edu

[1] Department of Epidemiology, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York 10032, USA

[2] Mailman School of Public Health, ICAP at Columbia University, 722 West 168th Street, New York 10032, USA

[3] Government of the Kingdom of Eswatini, Ministry of Health, Mbabane, Eswatini

implementation, like those funded through the United States President's Emergency Plan for AIDS Relief (PEPFAR), are shifting from generalized support to more targeted service delivery, and evaluations of interventions focused on addressing gaps in optimal HIV treatment and prevention service delivery are vital [4].

SCM is a natural fit for HIV service delivery program evaluation for several reasons. First, this shift to targeted service delivery creates conditions where providers and partners launch interventions heterogeneously within their care portfolios, with, for example, some facilities receiving interventions and others not. Facilities not receiving the intervention can contribute towards a counterfactual comparison group for those receiving the intervention. Second, routinely reported information of key HIV and related indicators, like PEPFAR Monitoring, Evaluation, and Reporting (MER) Indicators [5], are available longitudinally across long time periods, allowing for sufficient pre-intervention data to model trends. Finally, the non-random nature of many service delivery evaluations can make other study designs for evaluation less rigorous.

In this review, we will begin with a brief overview of SCM as a tool useful for large-scale evaluation of public health interventions. Because we found no examples of using SCM for HIV program evaluation outside of own evaluations to date, we will extend this review to its application in evaluation of policies and evaluations in public health and the social sciences more broadly. Next, we will describe the SCM and present an example of its use for program monitoring and evaluation in an HIV service delivery setting, focusing on assessing changes in PrEP enrollment following a public health response intervention targeted at facilities with high numbers of recent HIV infections in Eswatini. Finally, we will discuss the strengths and limitations of SCM for evaluating HIV service delivery programs.

Review of Synthetic Control Method Applications in Public Health Evaluations

Early applications of SCM were focused on policy evaluation. For example, Abadie et al.'s 2010 paper estimated the causal effect of California's Tobacco Control Program on cigarette sales [2]. A 2018 literature review identified 38 studies using SCM in health research, with most focusing on state or national policy interventions [6]. However, a few studies included in this review focused on interventions below the state or national level, including one investigating the effectiveness of a pay-for-performance policy on 30-day hospital mortality, two assessing school food programs, and one focused on food labeling [7–9]. Public health-relevant studies published after this review have used SCM to evaluate the effects of Medicaid expansion on cardiovascular disease, the impact of the Convention on the Rights of the Child on child mortality and vaccination rates, and the effect of air-quality regulations in Seoul, Korea, on cardiovascular mortality, and more [10–12].

SCM has also been used in evaluations of targeted interventions of direct analog to those relevant in HIV implementation science. However, there are very few studies specifically focusing on HIV. One study in the 2018 literature review used SCM to compare life expectancy, mortality, and birth rates in countries heavily impacted by HIV (Mozambique, South Africa, and Zimbabwe) against a synthetic control of other countries in Sub-Saharan Africa less impacted by HIV [13]. To our knowledge, no other HIV-related papers using SCM have been peer-reviewed since this review; however, one IZA Discussion Series paper, examining the impact of the introduction of highly active anti-retroviral therapy on economic indicators, and a masters dissertation examining the effect of a needle exchange program on HIV- and hepatitis-related healthcare visits, have expanded the use of SCM into the field of HIV [14, 15].

There are several studies directly applicable to the types of situations encountered in HIV program evaluation. A 2024 study used SCM to evaluate the effectiveness of a targeted mosquito sterilization program for dengue control in Singapore, comparing Dengue rates in towns receiving interventions to a synthetic control built from 30 or non-intervention towns [16]. A 2015 report focused on evaluating a heterogeneously implemented policy intervention (free primary care in Zambia), a frequent intervention target in program evaluations [17]. SCM has also recently been used to evaluate the impact of COVID-19-related policies on a variety of outcomes, including on COVID-19 cases, deaths, vaccination rates, and air pollutants [18–21]. Regarding program evaluations, one study used SCM to investigate the impact of pneumococcal conjugate vaccine programs and another evaluated a firearm violence prevention program [22, 23]. The one evaluating the firearm violence prevention program investigated an effect at the site-level [23].

While the use of SCM is lacking in the field of HIV program evaluation, we feel that the applications listed above are directly relevant to questions often raised in this field. Below, we argue that SCM is uniquely suited to many types of program and policy evaluations routinely encountered by researchers focusing on HIV service implementation evaluation. Specifically, we propose the use of SCM to answer research questions involving the implementation of a specified intervention within part, but not all, of a service provider's portfolio, such as specific health facilities or regions within a country introducing a new standard of care, receiving enhanced clinical training, or targeting a new population of interest. In these settings, traditional approaches for evaluation might be difficult due to the myriad differences related to the outcome of interest between clinics or regions chosen versus those not chosen for such interventions, as well as a lack of high-quality data measuring these proposed differences. However, in these settings, there likely is substantial longitudinal data measuring trends in the outcome of interest and numerous facilities or districts, similar in many respects, but not receiving an intervention from which a synthetic control can be built.

The Method and an Application

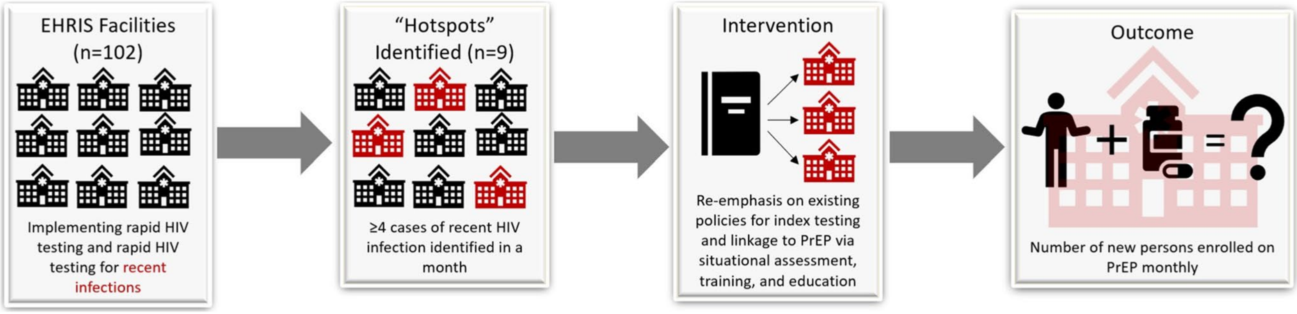
When and How to Use the Synthetic Control Method

SCM was developed as an extension of difference-in-difference approaches to policy evaluation, replacing the choice of a single non-intervention comparison unit with one synthesized from a weighted average of many potential non-intervention units [1, 2]. The reasoning is intuitive: in situations where a given location (such as a district or health facility) receives an intervention, but other areas do not, how best can we approximate outcome trends in the intervention unit had there actually been no intervention? In the case of HIV implementation science, we are often tasked to evaluate the effects of a targeted intervention focusing on, for example, low-performing health facilities or under-serviced geographic areas. While traditional DiD methods require identification and matching of this intervention unit with a non-intervention unit, the SCM leverages information from a "donor pool" of similar units to *statistically manufacture an approximation for pre-intervention trends observed in the intervention unit*. By using a weighted average of several non-intervention units, like districts or facilities, rather than relying on matching against a single unit, synthetic control methodologies begin with the assumption that "a combination of unaffected units often provides a more appropriate comparison than any single unit alone" [24••].

To illustrate this concept, consider an example using SCM for routine monitoring and program evaluation recently conducted in Eswatini. Beginning in 2019, all health facilities offering HIV testing services began classifying HIV-positive test results based on likely timing of infection (recent or long term) using the Recent Infection Testing Algorithm (RITA) based on a Rapid Test for Recent Infections (RTRI) and baseline viral load as part of routine HIV testing services. As part of the Eswatini HIV-1 Recent Infection Surveillance (EHRIS) program, potential "hotspots" of recent infection, defined as any facility identifying ≥ 4 RITA recent cases within a month, are flagged for potential public health responses [25, 26]. This facility-level public health response, triggered by identification of HIV infection "hotspots," is our intervention of interest. In Eswatini, this public health response includes re-emphasizing fidelity to existing national policies supporting providing index testing services for contacts of all newly identified people with HIV and linkage to PrEP services for contacts of index cases testing HIV-negative. This is achieved by (1) conducting in-person situational assessments at health facilities of gaps in implementation of these national policies and (2) increasing training and education to healthcare workers at these facilities on these policies. As part of routine program evaluation activities, we wanted to assess whether this facility-level public health response—this renewed emphasis on index testing and, ultimately, linkage to PrEP for those testing negative—affected the outcome of PrEP enrollment, defined as the new persons enrolled on PrEP monthly. This is illustrated in Fig. 1.

We used aggregated HIV service data available for routine reporting; no participants were enrolled, no identifying information on individuals was collected, and no additional data points were collected for this analysis. We used what is routinely available to HIV service delivery programs: aggregated, longitudinal data. In this case, these data were aggregated at the facility level, from October 2019 to December 2022.

The intervention time-point was August 2020, resulting in ten pre- and 28 post-intervention months. This is the point at which the public health responses were implemented, after facilities with "hotspots" of recent infection were identified. Nine intervention facilities (out of


EHRIS Facilities (n=102)

Implementing rapid HIV testing and rapid HIV testing for recent infections

"Hotspots" Identified (n=9)

≥4 cases of recent HIV infection identified in a month

Intervention

Re-emphasis on existing policies for index testing and linkage to PrEP via situational assessment, training, and education

Outcome

Number of new persons enrolled on PrEP monthly

Fig. 1 Illustration of the EHRIS program and subsequent public health response intervention triggered by HIV "hotspot" identification

102 facilities with data available that were implementing EHRIS) received the intervention in August 2020.

When there are multiple treated units, intervention effects can be estimated for each intervention unit, or the intervention units can be combined into an aggregated unit such as a geographic region or district [27–29]. Constructing separate synthetic controls for each intervention avoids interpolation biases but is burdensome. Construction of an aggregate intervention unit is acceptable if the pre-intervention outcome has a similar average and range in the intervention units and donor pool [27]. Investigators may also desire to use SCM to evaluate an intervention that has been rolled out in a staggered or stepwise fashion. When an intervention is staggered, there are three ways to perform SCM. Investigators may fit a separate synthetic control for each intervention unit and average; they may pool weights to construct a synthetic control that better approximates a counterfactual for all the intervention units; or they may use partially pooled weights—an intermediate between these two options [30••]. These extensions are described in detail by Ben-Michael et al., Abadie, and others [24••, 27–29, 30••].

For the purposes of this analysis, we averaged the nine intervention facilities to represent a single intervention unit; subsequent analyses can explore whether the effect of the intervention was greater or lesser at specific facilities, but here, we focused on the average effect.

We first examined our data visually. Figure 2 provides a summary of average PrEP enrollment trends for the intervention and non-intervention facilities. We saw increases in PrEP enrollment averages in both groups, with this increase starting in January 2021. The trend in PrEP enrollment appears generally higher in the intervention facilities after August 2020. This analysis is a good starting point, and we have enough information here to support performing a difference-in-difference analysis if we could find an appropriate comparison unit. However, because the intervention facilities were chosen deliberately based on identified increases in recent infections, which might be related to PrEP enrollment trends regardless of any intervention, we did not want to use traditional matching to perform this assessment. Further, we had substantial longitudinal measures of a few key indicators across many health facilities, as is typical of large-scale service delivery projects, but not the granular data often needed for proper matching. Because of these conditions—no clear matching prospects, a large donor pool of facilities, and adequate longitudinal data on key measures—we decided this situation was ideal for SCM.

Creating the Synthetic Control

We began by creating our synthetic control. Originally described by Abadie and Gardeazabal [1], synthetic controls are constructed as the weighted average of "candidate" non-intervention units that maximizes the fit between the pre-intervention trends in the outcome of interest in the intervention unit and the synthetic control using simple ordinary least squares regression:

$$\sum_{m=1}^{k} v_m \left(X_{1m} - X_{0m}W\right)^2$$

W is the weight assigned to each candidate unit, X_0 reflects a vector of pre-intervention characteristics of candidate unit m, and X_1 reflects a vector of pre-intervention characteristics of the intervention unit. The variable ν_m is

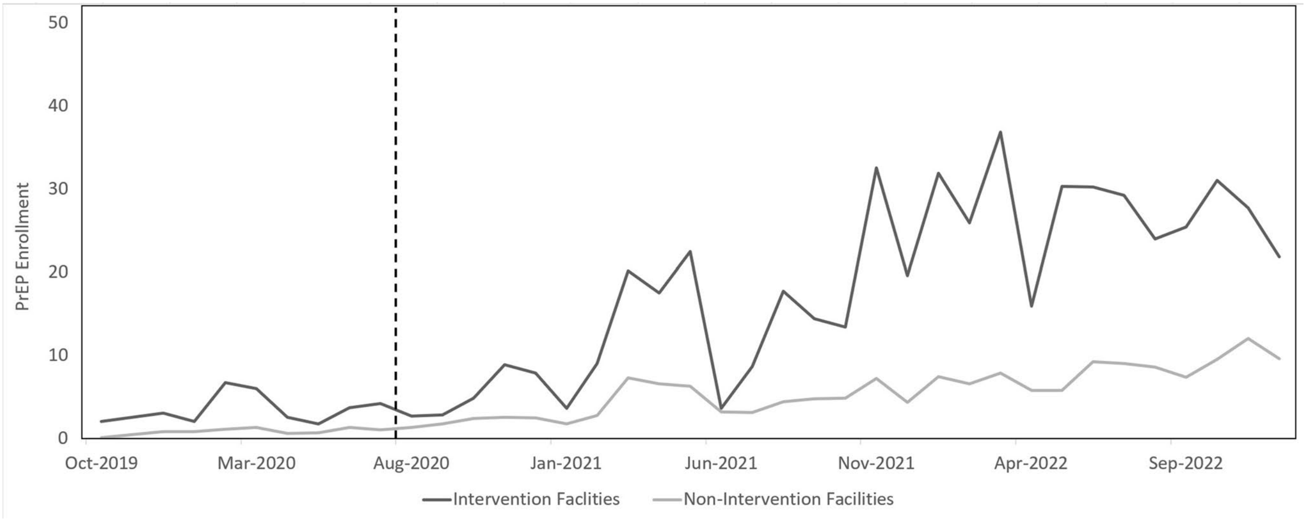


Fig. 2 Average PrEP enrollment for intervention vs. non-intervention facilities, October 2019 to December 2022

optional but can be used to change the relative importance of any predictors in the vector. In its simplest form, ν_m is set to 1 and the vectors X_1 and X_0 contain only repeated measures of the outcome of interest at time-points prior to intervention implementation. Other predictors thought to improve model fit can also be added. Each candidate non-intervention unit is assigned a weight W that is chosen to minimize the total value of the above equation; in traditional SCM, the candidate weights W are non-negative and sum to exactly 1.0. [27]. Therefore, to create a synthetic control is essentially a two-step process in which we (1) create predictor weights, including, minimally, pre-treatment repeated measures of the outcome of interest and, potentially, additional predictor variables of exogenous trends, and (2) create non-intervention unit weights.

In its simplest formulation, the synthetic control is generated exclusively from a weighted average of pre-intervention trends in the outcome of interest; no additional predictors are required. This makes this approach extremely well-suited to HIV implementation science, where longitudinal information is routinely available on key programmatic outcomes, but additional information is not routinely available on potential sources of confounding. The overlying goal of SCM is to create a comparison group that closely mimics pre-intervention trends in the outcome of interest as experienced by the intervention groups. Any predictor weighting that accomplishes this goal is considered valid. If this goal is accomplished without additional predictors, so much the easier. When the initial model fit is insufficient, or when the investigator wants to ensure that the intervention and synthetic control are similar on specific characteristics, predictors may be included in the model to the extent that these improve model fit. Whether to include predictors, and how many, is a decision made by the investigator, who is likely to factor in several considerations: the likelihood that other measured variables might influence trends, comparison of the overall model fit, how affected the predictors are by the intervention, and statistical considerations [27]. If investigators choose to use predictors, it is recommended that they consider various combinations of predictors and observe model fit. Additionally, if used, the predictors are weighted by importance, statistically, and sum to 1.

Consequently, to complete the first step in the process to construct our SCM, we considered two models: one including only pre-intervention PrEP enrollment (the outcome) in the vector of predictors and one including additional predictors of PrEP enrollment routinely available to HIV program implementers: monthly counts of (1) individuals accepting an index testing referral, (2) individuals tested for HIV, (3) individuals testing positive for HIV, (4) individuals starting ART, (5) the type of clinic (i.e., clinic, hospital, or public health unit), and (6) the outcome of PrEP enrollment. These predictors are correlated with each other but not perfectly. We decided to test both models, labeling the model with only PrEP enrollment as a predictor of the "primary" model and the model with additional predictors as the "secondary" model. For this analysis, we did not choose different combinations of the additional predictors for the simplicity of illustrating this practical SCM example.

SCM is available through several statistical software packages. We used the "synth" and "augsynth" R packages, for this analysis [31–33]. Additional packages are available in STATA ("synth," "allsynth," etc.) [34]. In brief, the "synth" R package creates predictor and donor pool candidate weights, manufactures a synthetic control, provides trend plots comparing the intervention unit and synthetic control, and produces the mean squared prediction error for the model [31]. For the first step in the process to create a synthetic control, we focused on the predictor weights for our primary and secondary models. As expected, the weight for PrEP enrollment in the primary model was 1.0. In the secondary model, predictor weights ranged from 0.7% for the predictor "individuals tested for HIV" to 53.8% for the predictor "individuals accepting an HIV index testing referral." This is illustrated in Table 1.

For the second step in the process, weighting the donor pool candidates, the pool of potential candidates should meet these minimally sufficient conditions: (1) they should have adequate longitudinal outcome measures to establish a stable assessment of pre-intervention trends in the outcome of interest; (2) their outcome trends should not be impacted by the intervention under investigation; (3) they should be of the same general "type" as the intervention unit (i.e., facilities should be in the donor pool for investigations of facility-level interventions, countries for country-level interventions); and (4) units that have received unique "shocks" related to the outcome should be excluded [2, 27]. Longitudinal pre-intervention outcome measures are important because SCM uses the trend in pre-intervention time-points to estimate the trend in the outcome in the post-intervention period. Abadie et al. assert that SCM should not be used when the SCM's pre-intervention outcome trend does not closely match that of the intervention unit [27]. However,

Table 1 Predictor weights for the SCM

Predictor	Weight (primary model)	Weight (secondary model)
Accepted HIV index testing referral	n/a	0.538
Tested for HIV	n/a	0.007
Tested positive for HIV	n/a	0.036
Started ART	n/a	0.294
Clinic type	n/a	0.014
Enrolled in PrEP	1	0.111

"closely match" is not defined and may be considered relative to the investigator's goals and the quality of the data used in the analysis.

Our analysis had a natural donor pool: 93 facilities that were implementing EHRIS (i.e., rapid HIV testing and rapid HIV testing for recent infections) but were not flagged as "hotspots" of recent infection and, therefore, did not receive the public health response intervention. Like the intervention units, these non-intervention units were health facilities with adequate measurement of the longitudinal outcome measures necessary to establish a stable assessment of pre-intervention trends in the outcome of interest. Unlike the intervention units, they should not have been impacted by the intervention. Additionally, none of these facilities received any unique "shocks" that would have impacted our specified outcome of PrEP enrollment.

Using the "synth" R package, we created non-intervention unit weights for donor pool facilities to manufacture the synthetic control for both our primary and secondary models. For the primary model, the synthetic control was a weighted combination of all 93 donor units, ranging from 0.5 to 6% each. For the secondary model, three units from the donor pool comprised the synthetic control, with weights ranging from 2.3 to 78.9%. This is illustrated in Table 2.

Using both these predictor and facilities weights, we created the synthetic control and statistically and visually assessed its performance.

Analyzing the Synthetic Control Model

The ability of the SCM to unbiasedly estimate the causal effect of the intervention on the intervention unit is premised on having a good pre-treatment model fit. Statistically, this is determined by comparing the synthetic control's pre-intervention fit with observed trends in the intervention unit using the mean squared prediction error (MSPE) [2]. Steps taken to reduce over-fitting follow conventional prediction model building approaches and include dividing the pre-intervention periods into training and validation periods and testing different groups of predictors [27]. To evaluate the effect of the intervention on the outcome, as with standard DiD analyses, investigators plot the trends in the outcome in the pre-intervention and post-intervention periods for the intervention unit and the synthetic control and compare visually. Investigators can construct 95% confidence intervals and estimate *p*-values using a *t*-statistic, plot the trend in the difference between the intervention unit and the synthetic control in the pre- and post-intervention time periods, and calculate the magnitude and significance of the average treatment effect on the treated (ATT) unit compared to the synthetic control.

Figure 3 presents the results of our SCM analysis using the "synth" R package using both the primary and secondary models. A visual inspection showed similar results, but with the secondary model (3b) having somewhat better pre-intervention fit. For example, in the primary model, notice the spike in the synthetic control around March 2020 that is not observed in the intervention unit. Tables 3 and 4 compare the predictor means for the intervention unit to the predictor means for the synthetic control and the entire donor pool. This allowed us to see how closely the synthetic control approximates the intervention. In Table 3 for the primary model, the intervention and synthetic control means for PrEP enrollment are identical (3.425 monthly enrollments) while

Table 2 Weighted facilities in the synthetic control (secondary model)

Facility	Weight
Facility A	0.789
Facility B	0.189
Facility C	0.023

Table 3 Mean values for the intervention, synthetic control, and sample for each predictor (primary model)

Predictor	Intervention mean	Synthetic control mean	Donor pool
Enrolled in PrEP	3.425	3.425	0.800

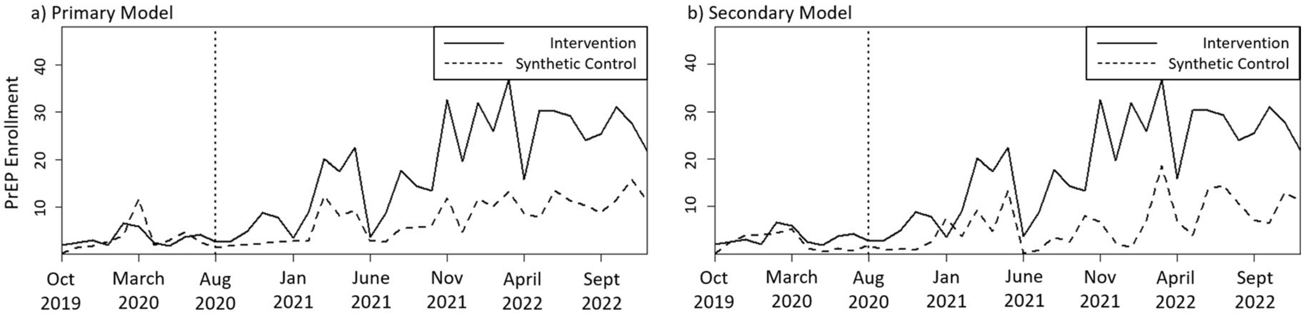


Fig. 3 Trends in PrEP enrollment for the intervention unit and the synthetic control, October 2019 to December 2022. **a** Primary model, only includes PrEP enrollment as a predictor. **b** Secondary model, additionally includes other predictors ("synth" package)

Table 4 Mean values for the intervention, synthetic control, and sample for each predictor (secondary model)

Predictor	Intervention mean	Synthetic control mean	Donor pool
Accepted HIV index testing referral	5.458	4.619	0.742
Tested for HIV	98.440	56.749	39.068
Tested positive for HIV	31.380	18.468	10.778
Started ART	25.520	13.785	6.262
Clinic type	1.000	1.000	1.172
Enrolled in PrEP	3.425	2.356	0.800

the donor pool mean is much lower (0.8 monthly enrollments). In Table 4, the intervention and synthetic control means for each predictor are still quite dissimilar. This result is very different from the intervention and synthetic control means outlined in the literature, which resemble the closeness of the synthetic control and intervention unit in Table 3 [1, 2, 27]. This suggests that the secondary model may not be a good fit for our analysis, that the predictor means across donor facilities are too different, and that these very different means could be an artifact of using programmatic data. More investigation was warranted.

The MSPE for the primary model was 5.1, while the MSPE for the secondary model was 3.7, indicating that the secondary model had better predictive accuracy. Consequently, the statistical and visual assessments of the synthetic control models agreed.

Using the "augsynth" package, we deepened our analysis. The "augsynth" package, more recently developed than the "synth" package, can be used in both traditional synthetic control and augmented synthetic control analyses (see "Extensions of the Synthetic Control Method" section*)*. This package provides graphs, tables, and gap plots estimating the average treatment effect among the treated and the difference in the outcome between synthetic control and intervention unit, with confidence intervals and *p*-values, at each timepoint. We created gap plots for the primary and secondary models to examine the difference in the outcome between the intervention unit and synthetic control (the solid line), with confidence intervals (the gray shading), at each timepoint (Fig. 4). In our visual inspection, in contrast to what we determined from the results of the "synth" package, we determined that the primary model appears to have a better pre-intervention fit for the outcome. Notice the smoother pre-intervention fit line in the primary model compared to the secondary model. The ATTs for the primary and secondary models were 9.53 ($p = 0.066$) and 10.19 ($p = 0.004$) more monthly PrEP enrollments, respectively, for the intervention unit than the synthetic control in the post-intervention period. Because of these conflicting results, to decide what model was best, we used an SCM extension with augmented synthetic controls (see "Extensions of the Synthetic Control Method" section).

Extensions of the Synthetic Control Method

Augmented Synthetic Controls

Approaches to statistically improve pre-treatment model fit are available with the augmented synthetic control method (ASCM). ACSM involves the use of advanced regression techniques and more flexibility in donor pool weights to avoid overfitting [35••]. The choice to use ASCM instead of SCM should be based on how much bias exists between the outcome model's fitted values for the intervention unit

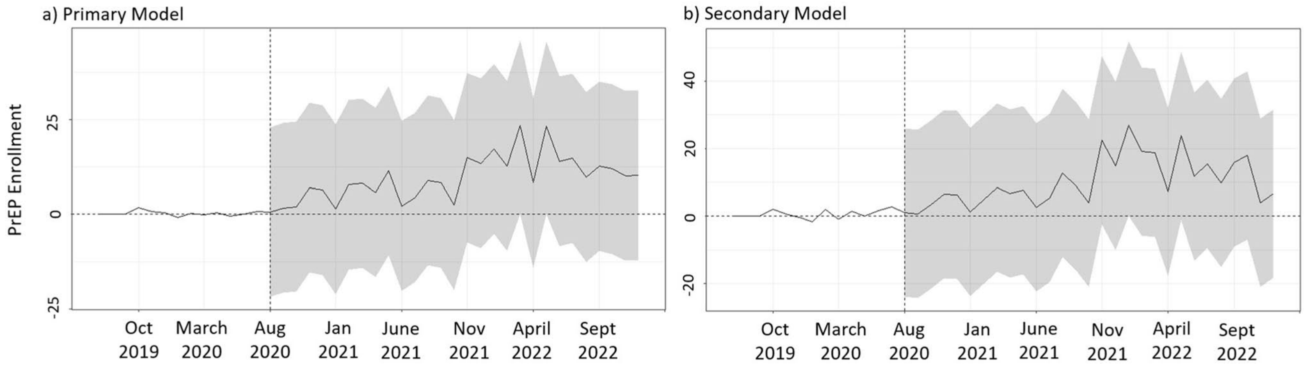


Fig. 4 Difference in PrEP enrollment between the intervention unit and synthetic control. **a** Primary model, only includes PrEP enrollment as a predictor. **b** Secondary model, additionally includes other predictors ("augsynth" package)

and synthetic control [35••]. Researchers must decide how much bias they can tolerate based on their data source, contextual factors, and research question. For more information, see Ben-Michael et al. [35••].

For our analysis, because of the noted discrepancies in the results, we ran the augmented synthetic control models both with and without predictors (i.e., the primary and secondary models), and we found that the bias estimation was 0.06 PrEP monthly enrollments for the primary model and 7.01 PrEP enrollments for the secondary model. While the bias estimation for the primary model was a very small percentage of its ATT, the bias estimation for the secondary model was more than two-thirds of its ATT. Therefore, in consideration with our visual examination of pre-intervention trends, MSPE comparison, and Tables 3 and 4 comparisons, we determined that the primary model was the right model for interpretation. We also determined that, because of the small bias estimation, conducting ASCM was not necessary.

For demonstration, Fig. 5 shows the augmented difference, with ASCM, in the outcome between the intervention unit and the synthetic control, with confidence intervals, at each time-point. The augmented primary model (Fig. 5a) appears to be very similar to the regular primary model, further supporting our use of the regular primary model. The ATTs for the augmented primary and secondary models were 9.48 ($p = 0.149$) and 5.58 ($p = 0.998$) more monthly PrEP enrollments for the intervention unit compared to the synthetic control, respectively.

Sensitivity Analyses—Robustness Checks

Sensitivity analyses examining the robustness of the study findings to different exposure or outcome operationalizations are recommended when performing SCM. Common sensitivity analyses include the following: "different outcome" placebo tests (replacing the outcome of interest with one not expected to be impacted by the intervention), "in time" placebo tests (replacing the actual timing of the intervention with a random time not related to it), and "in space" placebo tests (creating a synthetic control for every non-intervention unit) are commonly applied to assess the robustness of study assumptions [27, 36]. In our example, we performed a "different outcome" placebo test using ART initiation as the outcome instead of PrEP enrollment (Fig. 6a). As hypothesized, there was no difference between the performance of the intervention and synthetic control. For an "in time" placebo test, we replaced the real intervention time-point with a randomly selected time-point (March 2021) (Fig. 6b). As hypothesized, there was no difference between the performance of the intervention and synthetic control.

However, the results of our "in space" placebo test provided some evidence warranting caution in interpreting our findings. Using a gap plot illustrating the differences between every candidate-turned-intervention unit and its synthetic control, investigators can examine an array of effects and evaluate them all against the effect difference between the actual intervention unit and its synthetic control. We expect the magnitude of the effect should be at the outer range or, ideally, beyond the range of placebo effects. If it instead falls within the range of placebo effects, this undermines a claim for causation [27]. We see that the magnitude of the effect of the intervention unit is at the higher end of the plot but within the range of placebo effects in the post-intervention period (Fig. 6c). This suggests that chance alone cannot be ruled out as an explanation for our findings.

Discussion of Our Synthetic Control Method Results

The SCM analysis found that post-intervention trends in PrEP enrollment increased compared to the hypothetical situation of no intervention in the synthetic control. This difference was lagged (not apparent until about 6-month post-intervention) and largely sustained over time. Across all facilities, there was a decrease in PrEP enrollments in June 2021, corresponding to the beginning of a period of anti-monarchy protests, political unrest, and associated

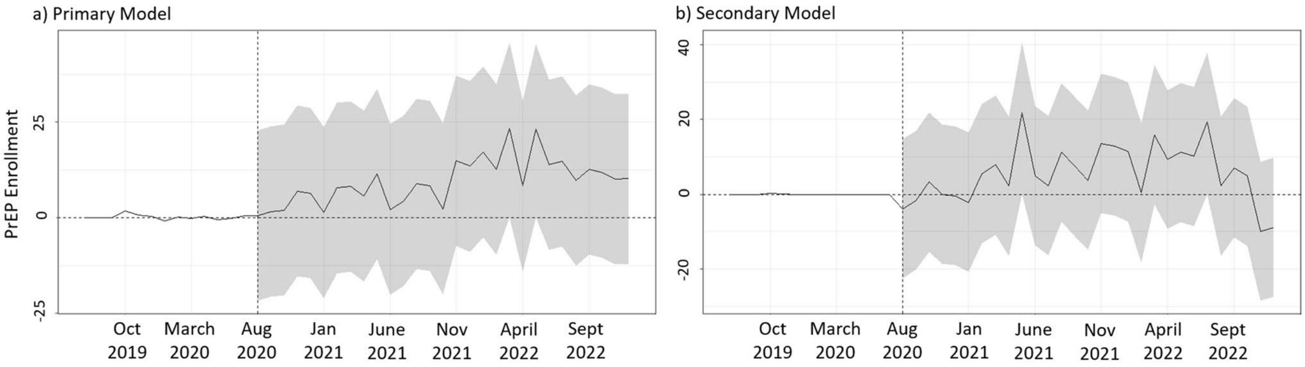


Fig. 5 Difference between the intervention unit and synthetic control using augmented synthetic control method. **a** Primary model, only includes PrEP enrollment as a predictor. **b** Secondary model, additionally includes other predictors ("augsynth" package)

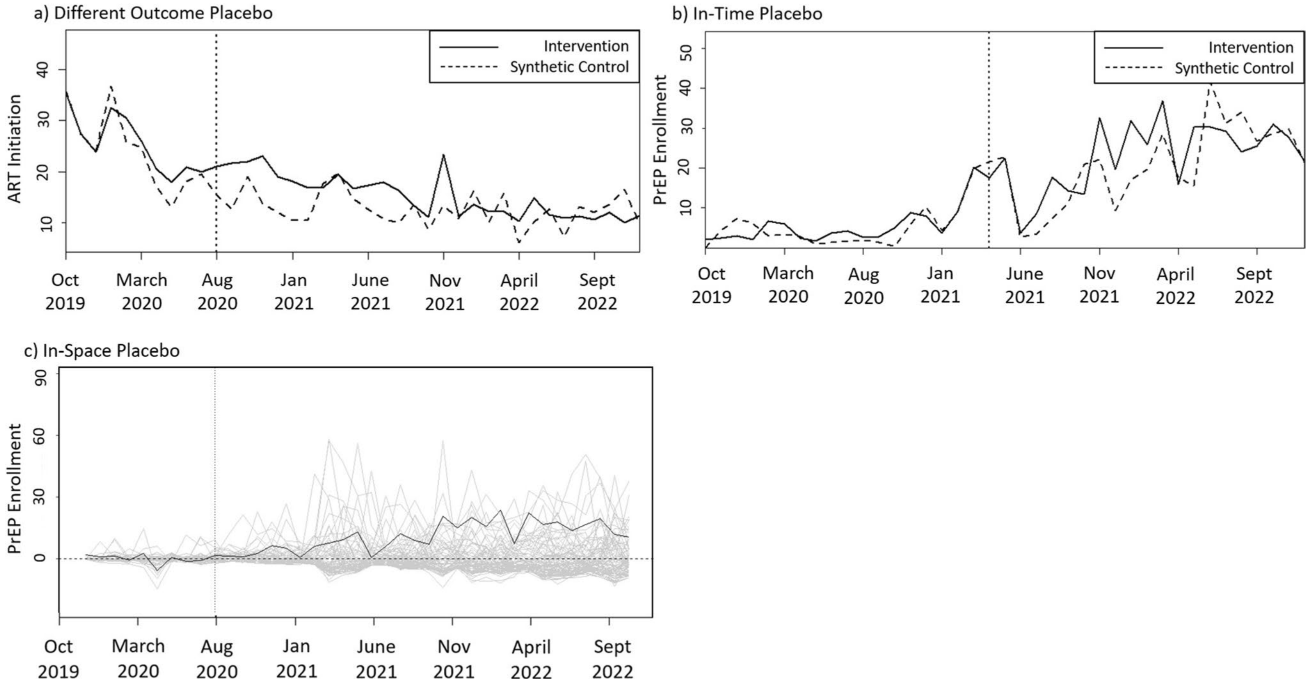


Fig. 6 Robustness checks ("synth" package). **a** and **b** Plots of trends in the outcome. **c** A gap plot

limitations in mobility in Eswatini [37–39]. This highlights the importance of contextual knowledge across evaluation and interpretation.

A potential limitation of this analysis is the limited availability of predictors for model building. In the future, we could consider using facility characteristics, like catchment area and staff size, or other predictors of PrEP enrollment. Moreover, it is possible that our chosen predictors are also affected by the intervention, if health facility staff move from facility to facility, making it an imperfect proxy for control of time-invariant confounding. Furthermore, rather than classifying our facilities as clinics, hospitals, or public health units, we could have used additional information to more systematically classify the facilities to ensure similarity between our intervention and donor pool units.

We conducted both SCM and ASCM, with and without predictors, to illustrate their use, their differences, and the implications of their results for this routine evaluation example in Eswatini. We also conducted robustness checks via placebo testing, with mixed results. Two placebo tests supported our conclusion, while the third did not, highlighting the need for multiple methods of placebo testing.

Based on the ATT analysis of the primary and secondary models for regular SCM, the difference in monthly PrEP enrollment between the intervention and synthetic control in the post-intervention period was 9.53 and 10.19, respectively. This is a less-than-one-PrEP enrollment difference, but only the secondary model had statistical significance.

Observed increases in PrEP enrollment resulting from a low-burden public health intervention would be meaningful if causal. Enrolling 9.53 more people on PrEP each month, per the 9 intervention facilities, translates to an additional 1029 persons on PrEP per year. If this intervention were rolled out to all facilities, in the event of "hotspots," this could mean an additional 10,635 persons on PrEP per year.

Assumptions – Difference-in-Difference and the Synthetic Control Method

In addition to the identifiability assumptions, the main assumptions necessary in difference-in-difference analysis are the parallel trends assumptions, or the assumption that the trends in the outcome in the pre-intervention period between the intervention and non-intervention units are parallel, and the assumption that important unmeasured variables are either time-invariant attributes of the study units or time-varying factors that are homogenous across study units [40]. SCM addresses the first assumption completely because it weights candidate non-intervention units to create parallel trends in the pre-intervention phase. The second assumption, however, is nearly unverifiable, but points to the importance of understanding the context of your intervention and data. This supports our argument that SCM is well-suited for evaluations of HIV and related program service implementation, as implementers are keenly aware of their implementation context.

Shi et al. describe assumptions required for the SCM estimator to validly measure the average causal effect of the intervention on the intervention units [41]. Briefly, under the identifiability assumptions of consistency (assuming that observed outcomes are a realization of potential outcomes under that treatment scenario), an interactive fixed effects model (assuming that the intervention is the only source of time-varying difference in causes of the outcome of interest), no interference between intervention and non-intervention units, and that effects of unmeasured confounding on the intervention units can be matched by a weighted average of unmeasured confounding on a set of the non-intervention units, which requires the assumption that a set of weights exists that can satisfy this assumption, a causal inference can made [41]. More informally, the identifiability assumptions unique to the SCM involve the assumption of no time-varying confounding between intervention and non-intervention units. Hollingsworth and Wing add an additional requirement of no period perfect multicollinearity of common factors, implying that imperfect multicollinearity is acceptable [42].

In our example, we assumed consistency. Additionally, because the intervention was targeted at specific facilities, so that only these facilities received the assessment, training, and education that enforced fidelity to national policies and guidelines, we were reasonably confident in no interference between the intervention and non-intervention units. We do acknowledge, however, that some facilities have high staff turnover resulting in potential interference that would underestimate the effect we observed. We also assumed that the intervention and synthetic control were exchangeable because we manufactured that exchangeability based on pre-intervention PrEP enrollment trends, which was successful for the primary model (Table 3).

Strengths and Limitations

SCM is a transparent, rigorous, data-driven, and efficient approach to approximating a counterfactual for non-experimental DiD analyses. Unlike traditional DiD analysis, it leverages information across an entire pool of candidate non-intervention units to synthetically construct equivalent pre-intervention trends (that are parallel and of similar magnitude). Additionally, SCM does not require an excessive amount of data for many different variables, but does require data measured over time, making it extremely useful for evaluations involving routinely collected data. Its results are also straightforward and easily interpreted.

Limitations include the requirement for sufficient pre-treatment information and a reasonably sized donor pool from which to construct the synthetic control. Units in the donor pool should have experienced the same “shocks,” or time-varying factors, as the intervention units and be of the same general type (i.e., facilities). As in traditional DiD analysis, SCM is still potentially subject to bias due to contamination or spillover from the intervention units to the non-intervention units and from anticipation bias, or when intervention units react ahead of actual implementation in anticipation [6]. Moreover, considerable contextual knowledge is necessary for both choosing adequate units for the donor pool and understanding the true nature of the effect observed.

Conclusions

SCM is a rigorous, data-driven approach to achieving a counterfactual comparison unit for DiD analyses. It is not limited by the assumptions inherent in traditional DiD, and it controls for observed and unobserved time-varying confounders. Its use in public health has increased steadily since its introduction, but SCM is rarely used to examine the effectiveness of programs or site-level interventions. We believe SCM has a place in routine HIV service delivery program evaluation. SCM is an efficient and rigorous analytic tool with which to carry out effectiveness evaluations.

We illustrated the use of SCM to evaluate the effect of EHRIS-associated interventions on monthly PrEP enrollment. We found that PrEP enrollment increased more in the aggregated intervention unit, and this remained true throughout the post-intervention period, but this increase was not fully robust to alternate explanations. It is important to understand the limitations of SCM, as well as its various novel elements, to appropriately use this method for a particular research question, context, and dataset. This example also demonstrates how our understanding of pre-intervention fit, suitable donor units, and predictor choice influences the results of our analysis. Additionally, we must strive to increase the quality of programmatic data to use SCM to evaluate HIV service delivery programs.

We believe SCM’s potential as an evaluation tool in implementation science, and its potential as the method grows and improves will be substantial. This is particularly true if we can routinize SCM for up-to-date and automatic DiD analysis generation.

Author Contribution S.W., M.R. L., and S.S. conceptualized the manuscript.

S.W. and M.R.L wrote the main text and prepared the tables and figures.

H.N.B and L.D. support the EHRIS program described in the text and provided key details on implementation and context.

All authors reviewed the manuscript.

Funding Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number T32AI114398. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declarations

Competing Interests The authors declare no competing interests.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as: •• Of major importance

1. Abadie A, Gardeazabal J. The economic costs of conflict: a case study of the Basque country. American Economic Review. 2003;93(1):113–32.
2. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc. 2010;105(490):493–505.
3. Athey S, Imbens GW. The state of applied econometrics: causality and policy evaluation. Journal of Economic Perspectives. 2017;31(2):3–32.
4. Reimagining PEPFAR's strategic direction fulfilling America's promise to end the HIV/AIDS pandemic by 2030. [Internet]. U.S. Department of State, Office of the Global AIDS Coordinator and Health Diplomacy; 2022 Sep. Available from: https://www.state.gov/wp-content/uploads/2022/09/PEPFAR-Strategic-Direction_FINAL.pdf
5. PEPFAR Fiscal Year 2023 Monitoring, evaluation, and reporting (MER) indicators. [Internet]. United States Department of State. [cited 2024 Feb 20]. Available from: https://www.state.gov/pepfar-fy-2023-mer-indicators/
6. Bouttell J, Craig P, Lewsey J, Robinson M, Popham F. Synthetic control methodology as a tool for evaluating population-level health interventions. J Epidemiol Community Health. 2018;72(8):673–8.
7. Kreif N, Grieve R, Hangartner D, Turner AJ, Nikolova S, Sutton M. Examination of the synthetic control method for evaluating health policies with multiple treated units. Health Econ. 2016;25(12):1514–28.
8. Bauhoff S. The effect of school district nutrition policies on dietary intake and overweight: a synthetic control approach. Econ Hum Biol. 2014;12:45–55.
9. Kiesel K, Villas-Boas SB. Can information costs affect consumer choice? Nutritional labels in a supermarket experiment. Int J Ind Organ. 2013;31(2):153–63.
10. Garber MD. Precision and weighting of effects estimated by the generalized synthetic control and related methods: the case of Medicaid expansion. Epidemiology. 2024;35(2):273–7.
11. Reinbold GW. Effects of the Convention on the Rights of the Child on child mortality and vaccination rates: a synthetic control analysis. BMC Int Health Hum Rights. 2019;19(1):24.
12. Kim SY, Kim H, Lee JT. Health Effects of air-quality regulations in Seoul metropolitan area: applying synthetic control method to controlled-interrupted time-series analysis. Atmosphere. 2020;11(8):868.
13. Karlsson M, Pichler S. Demographic consequences of HIV. J Popul Econ. 2015;28(4):1097–135.
14. Ejrnæs M, García-Miralles E, Gørtz M, Lundborg P. When death was postponed: the effect of HIV medication on work, savings and marriage. SSRN Journal [Internet]. 2023 [cited 2023 Oct 11]; Available from: https://www.ssrn.com/abstract=4527782
15. Eriksson A. The effect of needle exchange program on HIV and hepatitis: evidence from Sweden [Internet]. Uppsala University; 2022. Available from: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1678269&dswid=-6691
16. Lim JT, Bansal S, Chong CS, Dickens B, Ng Y, Deng L, et al. Efficacy of Wolbachia-mediated sterility to reduce the incidence of dengue: a synthetic control study in Singapore. The Lancet Microbe. 2024 Feb;S266652472300397X.
17. Lepine A, Lagarde M, Le Nestour A. Free primary care in Zambia: an impact evaluation using a pooled synthetic control method. SSRN Journal [Internet]. 2015 Jun [cited 2024 Feb 20]; Available from: http://www.ssrn.com/abstract=2520345
18 Yang X. Does city lockdown prevent the spread of COVID-19? New evidence from the synthetic control method. glob health res policy. 2021;6(1):20.
19. Mader S, Rüttenauer T. The effects of non-pharmaceutical interventions on COVID-19 mortality: a generalized synthetic control approach across 169 countries. Front Public Health. 2022;4(10): 820642.
20. Lang D, Esbenshade L, Willer R. Did Ohio's vaccine lottery increase vaccination rates? A pre-registered, synthetic control study. J Exp Polit Sci. 2023;10(2):242–60.
21. Cole MA, Elliott RJR, Liu B. The impact of the Wuhan Covid-19 lockdown on air pollution and health: a machine learning and augmented synthetic control approach. Environ Resource Econ. 2020;76(4):553–80.
22. Bruhn CAW, Hetterich S, Schuck-Paim C, Kürüm E, Taylor RJ, Lustig R, et al. Estimating the population-level impact of vaccines using synthetic controls. Proc Natl Acad Sci USA. 2017;114(7):1524–9.
23. Buggs SA, Webster DW, Crifasi CK. Using synthetic control methodology to estimate effects of a *Cure Violence* intervention in Baltimore. Maryland Inj Prev. 2022;28(1):61–7.
24.•• Abadie A. Using synthetic controls: feasibility, data requirements, and methodological aspects. Journal of Economic Literature. 2021;59(2):391–425. **This work provides an overview of the synthetic control method, its extensions, and related methods, as well as its strengths and limitations.**
25. HIV recency testing adopted as key component of HIV testing in Eswatini. ICAP News & Events [Internet]. 2022 Jan 21; Available from: https://icap.columbia.edu/news-events/hiv-recency-testing-adopted-as-key-component-of-hiv-testing-in-eswatini/
26. Kim AA, Behel S, Northbrook S, Parekh BS. Tracking with recency assays to control the epidemic: real-time HIV surveillance and public health response. AIDS. 2019;33(9):1527–9.
27. Abadie A, Diamond A, Hainmueller J. Comparative politics and the synthetic control method: comparative politics and the synthetic control method. American Journal of Political Science. 2015;59(2):495–510.
28. Hainmueller J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. Polit anal. 2012;20(1):25–46.
29. Robbins MW, Saunders J, Kilmer B. A Framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. J Am Stat Assoc. 2017;112(517):109–26.
30.•• Ben-Michael E, Feller A, Rothstein J. Synthetic controls with staggered adoption. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2022;84(2):351–81. **This article provides an overview of how to use the synthetic control method with staggered adoption units and provides an example of its use.**
31. Abadie A, Diamond A, Hainmueller J. Synth: An *R* package for synthetic control methods in comparative case studies. J Stat Soft [Internet]. 2011 [cited 2023 Oct 11];42(13). Available from: http://www.jstatsoft.org/v42/i13/

32. Bogard M. R code from an R package for synthetic control methods in comparative case studies. [Internet]. BioSciEconomist. 2019. Available from: https://gist.github.com/BioSciEconomist/de89cb79970cd13e0d6adb69396b4190

33. Ben-Michael E. augsynth: The augmented synthetic control method. [Internet]. 2021. Available from: https://github.com/ebenmichael/augsynth/blob/master/vignettes/singlesynth-vignette.md

34. Wiltshire JC. allsynth: (Stacked) synthetic control bias-correction utilities for Stata. [Internet]. 2022. Available from: https://justinwiltshire.com/allsynth-stacked-synthetic-control-biascorrection-utilities-for-stata

35.•• Ben-Michael E, Feller A, Rothstein J. The augmented synthetic control method. Journal of the American Statistical Association. 2021;116(536):1789–803. **This article provides an overview of the augmented synthetic control extension of the synthetic control method and illustrates an example of its use.**

36. Lipsitch M, TchetgenTchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383–8.

37. Agence France Presse. eSwatini youth stage rare rural protest against monarchy. News 24 [Internet]. 2021 Jun 20; Available from: https://www.news24.com/news24/africa/news/eswatini-youth-stage-rare-rural-protest-against-monarchy-20210620

38. Masuku L. Anti-monarchy protests in African kingdom eSwatini turn violent. Reuters [Internet]. 2021 Jun 29; Available from: https://www.reuters.com/world/africa/anti-monarchy-protests-african-kingdom-eswatini-turn-violent-2021-06-29/

39. Magome M. Eswatini imposes curfew to quell pro-democracy protests. The Associated Press [Internet]. 2021 Jun 29; Available from: https://apnews.com/article/africa-southern-africa-democracy-d0660fe7e44d0719d4f41ba136d79679

40. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. Annu Rev Public Health. 2018;39(1):453–69.

41. Shi C, Sridhar D, Misra V, Blei DM. On the assumptions of synthetic control methods. [Internet]. arXiv; 2021 [cited 2024 Feb 23]. Available from: http://arxiv.org/abs/2112.05671

42. Hollingsworth A, Wing C. Tactics for design and inference in synthetic control studies: an applied example using high-dimensional data. SSRN Journal [Internet]. 2020 [cited 2024 Feb 22]; Available from: https://www.ssrn.com/abstract=3592088

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.