Title
Evaluation of Sutter Health Programs Utilizing the Synthetic Control Method: A Secondary Analysis of Electronic Health Record Data

Author
Qi, Wendy

Publication Date
2021

Peer reviewed|Thesis/dissertation

Evaluation of Sutter Health Programs Utilizing the Synthetic Control Method: A Secondary Analysis of Electronic Health Record Data

by

Wendy M. Qi

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:
Professor John M. Colford, Jr., Co-Chair
Professor Arthur L. Reingold, Co-Chair
Professor Alan E. Hubbard
Professor Alice R. Pressman

Spring 2021

Evaluation of Sutter Health Programs Utilizing the Synthetic Control Method: A Secondary Analysis of Electronic Health Record Data

Abstract

Evaluation of Sutter Health Programs Utilizing the Synthetic Control Method: A Secondary Analysis of Electronic Health Record Data

by

Wendy M. Qi

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor John M. Colford, Jr., Co-Chair
Professor Arthur L. Reingold, Co-Chair

Background: Health programs and interventions designed to improve patient outcomes within health care systems are implemented widely; however, many are implemented without a plan for evaluating the impact of the program on relevant and meaningful outcomes. Program evaluations are important for providing context, assessing whether the intervention is having its desired effect, deciding whether the program should be continued, identifying any unintended consequences, and highlighting areas for improvement. Many health programs fall under the category of observational studies, and thus methods such as matching, difference-in-differences (DiD), regression discontinuity, and pre-test/post-test are often used for evaluation. Recently, the synthetic control method (SCM) has surfaced as an important tool for program evaluation and has been described as "arguably the most important innovation in the policy evaluation literature in the last 15 years" [1]. SCM is motivated by the common difficulty in identifying a single control unit that approximates the most relevant characteristics of the treated unit. The central idea of SCM is that a combination of control units may provide a better "counterfactual" for the treated unit than any one single control unit alone. A data-driven approach is used to assign weights to potential control units to create a "synthetic" version of the treated unit that closely approximates the time series for the actual treated unit in the pre-intervention period. With this, predictions about what counterfactual trends would look like in the post-intervention period, had the intervention never been implemented, can be made.

Methods: In this dissertation, I apply SCM to evaluate hospital-level effects of three programs recently implemented within the Sutter Health system: (1) the Advancing Health Equity (AHE) asthma program at Alta Bates Summit Medical Center that brings culturally appropriate community-based care to African American/Black patients, and provides high-touch and high-tech counseling services to educate patients about

disease and medication self-management; (2) the ETOH-P program at Eden Medical Center that implements two protocols for treating patients who are at risk for developing alcohol withdrawal syndrome; and 3) a group-based lifestyle change program implemented within several Sutter clinics for diabetes management. For each program, we compare the hospital-level results from the SCM analysis to individual-level results from a propensity score matched analysis.

Significance: This dissertation illustrates the application of SCM to evaluate the impact of health care programs implemented within an open health care system such as Sutter Health. SCM has previously been applied to study a wide range of topics including political and economic effects following terrorist conflict[2], state-level policy changes[3]. health systems reforms[4,5], nutritional interventions[6], climate events such as drought[7], and most recently COVID-19 mitigation strategies and mandates[8–14]. To our knowledge, SCM has never been applied in such a setting, in which individuals choose to participate in the program or intervention. Lessons learned from this exercise provide valuable insight into the utility of this evaluation tool for health care systems research and offer both data and methodological considerations for future applications.

Dedication

For Mom, Dad and Stephanie. I am forever grateful for your support and encouragement.

And for Kyle. Your optimism is contagious.

Acknowledgements

I would first like to acknowledge my co-advisor, Jack Colford, for his consistent mentorship and unwavering support throughout my graduate training. Jack has always and continues to provide me with endless opportunities to learn and connect with others in the field. Jack is truly an unbelievable mentor and teacher.

I am also grateful to my co-advisor, Art Reingold, for providing ideas, inspiration, funding, and constant support throughout my entire PhD.

Thank you to Alan Hubbard for his valuable biostatistical input and constant guidance on this dissertation. I never thought that statistics would make sense to me, but I think some of it finally stuck thanks to you!

To Alice Pressman, I will never forget your generosity and kindness. Thank you for being my mentor, my friend, my confidant, my biostatistics resource, my second mom. Thank you for providing me an avenue to pursue my dissertation interests, for housing me and feeding me. I hope to continue learning from you throughout the rest of my career.

Thank you to my colleagues at Sutter Health RDD, Lydia Xu, Zijun Shen and Rob Romanelli, whose assistance and guidance in obtaining data for this dissertation was critical. I could not have done this without you all.

Lastly, thank you to my PhD cohort who provided incredible professional and personal support during our time at Berkeley together: Ruvani Jayaweera, Isaac Ergas, Sabrina Boyce, Eli Michaels, Mary Horton, Moon Choi, Chris Rowe and Cameron Adams.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

1.1 Motivation

Interventions and programs aimed at improving patient outcomes are implemented widely within the health care system setting. These health interventions can be educational programs, health policy changes, or health promotion campaigns that seek to improve or modify health behaviors and health outcomes among the population served. Assessing the impact of these health interventions and programs enables health care systems to learn and improve their procedures of care, as well as make decisions about scaling up programs. In many cases, these interventions and programs are developed and implemented without a subsequent plan for how to evaluate the success of the program. While patient testimonials and clinician stories on how these programs positively affect health outcomes are encouraging, it is often the case that health care systems wish to quantify the success of these programs in terms of both relevant health outcomes but also from a cost-effectiveness perspective. An impact evaluation is an empirical assessment of the program's effects; it measures the extent to which outcomes experienced by participating individuals were caused by the intervention, and which can be attributed to other unrelated factors[15]. In particular, rigorous impact evaluations are crucial to presenting data-driven evidence for program effectiveness in the health-care setting.

Two main evaluation study designs are frequently used within health care systems research. Experimental designs in the form of Randomized Controlled Trials (RCTs) are the gold standard for identifying causal impacts and involve random assignment of individuals to the intervention group and the control group. The main advantage of RCTs is their high internal validity, but RCTs can be limiting in terms of generalizability to other populations, and it is not always practical nor ethical to randomize interventions that are known to be beneficial. For these reasons, observational studies that do not utilize a non-random selection process are widely used. Observational studies aim to mimic the experimental design by creating equivalent intervention and control groups. Methods such as matching, difference-in-differences (DiD), regression discontinuity, and pre-test/post-test are often used in observational study designs [15].

Recently, the synthetic control method (SCM) has surfaced as an important tool for program evaluation[2,3]. Athey and Imbens describe the synthetic control approach as "arguably the most important innovation in the policy evaluation literature in the last 15 years" [1]. SCM originates from political science and economics, which uses comparative case studies to examine events or policy interventions that take place at an aggregate level and affect group-level entities such as schools, countries, regions, cities, etc. In the last several years, synthetic controls have been used to study a wide range of topics including political and economic effects following terrorist conflict[2], state-level policy

changes[3]. health systems reforms[4,5], nutritional interventions[6], climate events such as drought[7], and most recently COVID-19 mitigation strategies and mandates[8–14].

SCM is motivated by the common difficulty in identifying a single control unit that approximates the most relevant characteristics of the treated unit; thus, the idea behind the synthetic control approach is that a combination of control units may provide a better "counterfactual" for the treated unit than any one single control unit alone. In SCM, A data-driven approach is used to assign weights to potential control units to create a "synthetic" version of the treated unit. This eliminates the need for researchers to arbitrarily pick one control unit. The goal of the "synthetic" version of the treated unit is to closely approximate the time series for the actual treated unit in the pre-intervention period, such that it can be used to make predictions about what counterfactual trends would look like in the post-intervention period, had the intervention never been implemented[3,16].

Figure 1.1 below visually represents this concept in a mock example of the ETOH-P intervention. The red line represents the outcome trend for ETOH-P intervention patients. The blue line represents the outcome trend for the "synthetic" ETOH-P intervention patients and is constructed using weights applied to a pool of potential control units. As demonstrated, the blue "synthetic" line closely approximates the outcome trend of the actual red line during the pre-intervention weeks. After the intervention is implemented at week 10, the two lines diverge, with the blue "synthetic" line representing what would have happened in the red line, had the intervention never occurred (the counterfactual).

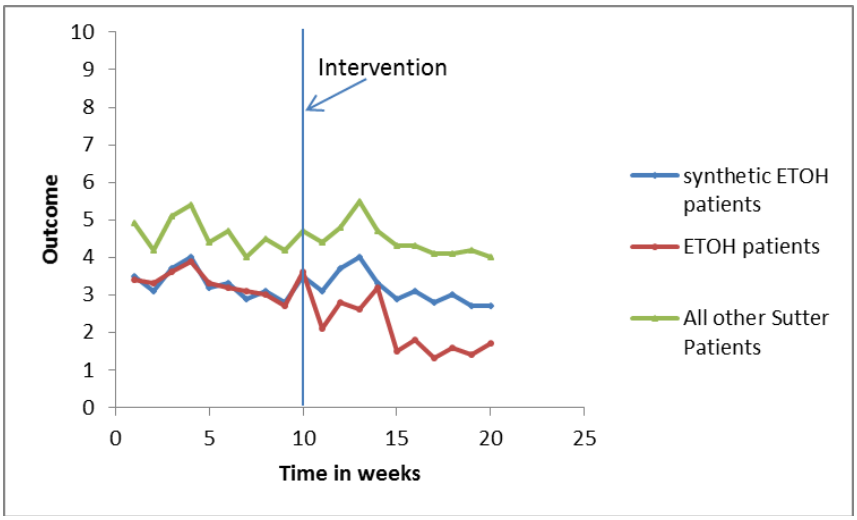


Figure 1.1 example of synthetic control method using mock data

We present a motivating model to more formally present details of the method, following previously published guidance[2,3,16,17]. We will proceed as if only one unit was exposed to the intervention of interest. In cases in which there are multiple units that received the intervention, it is advised to aggregate those exposed units together.

Suppose we have $j = 1, \dots, J+1$ hospitals for time periods $t = 1, \dots, T$, where the first hospital is the treated hospital and the remaining J control hospitals can possibly contribute to the synthetic control. This set of control clinics is termed the "donor pool". We define two potential outcomes. Y_{it}^N is the outcome that would be observed for hospital i at time t if hospital i is not exposed to the intervention for units $i = 1, \dots, J+1$ and time periods $t = 1, \dots, T$. Y_{it}^I is the outcome that would be observed for hospital i at time t if hospital i is exposed to the intervention in periods $T_0 + 1$ to T .

Let T_0 be the number of preintervention periods, with $1 \leq T_0 < T$. We observe Y_{it}^I in the post-intervention period for the treated hospital, but Y_{it}^N is unobserved for the treated hospital in the post-intervention period. The goal of the synthetic control method is to construct a synthetic control group that yields a reasonable estimate of this missing potential outcome. In doing so, we will be able to estimate the effect of the intervention on our outcome of interest for the treated hospital in the post-intervention period. Formally, this effect is the difference between the two potential outcomes for the intervention period:

$$\alpha_{it} = Y_{it}^I - Y_{it}^N$$

The synthetic control should resemble the treated hospital in terms of relevant pre-program characteristics and pre-intervention outcomes. We define a (Jx1) vector of positive weights$W = (w_2, \dots, w_{j+1})'$ such that $w_j \geq 0$ for $j = 2, \dots, J+1$ and $w_2 + \dots w_{j+1} = 1$ (sum to 1). Each particular value of the vector **W** represents a specific weight averaged of the control hospitals, and thus a separate synthetic control. Thus, the estimator of α_{it} can be expressed as

$$\widehat{\alpha_{1t}} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \text{ for } t \in \{T_0 + 1, \dots, T\}$$

Weights W* are chosen such that the synthetic control best approximates the treated hospital with respect to outcome predictors and linear combinations of pre-intervention outcomes. In some cases, it may not be possible to obtain a weighted combination of control units such that it matches exactly the treated hospital; in such scenarios where the fit is poor, it is not recommended to use a synthetic control.

Abadie et al.[16] caution that even if the synthetic control is a good approximation for the treated unit, interpolation biases can still be large if the simple linear model presented above does not hold over all the potential control units. This bias may be avoided by restricting the pool of donor control units to those with similar characteristics within a certain magnitude to the treated unit. The outcome variable is observed for T total periods, $t = 1, \dots, T$ for the intervention hospital Y_{1t} and for the control hospitals Y_{jt} where $j = 2, \dots, J+1$. $\mathbf{X_1}$ will be a (k x 1) vector containing the values of the pre-program characteristics of the treated clinic that we aim to match as closely as possible. $\mathbf{X_0}$ is a k x J matrix with values of the same variables for the control clinics. The preintervention characteristics in X0 and X1 may include pre-intervention values of the outcome(s). We will utilize the *synth* package (cite R synth package) to choose the vector W* that

minimizes the distance $||X_1 - X_0 W||$ between $\mathbf{X_1}$ and $\mathbf{X_0 W}$. The synth() function solves for W* by minimizing

$$||X_1 - X_0 W||v = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

Where V is a (k x k) matrix allowing for different weights to be applied to variables in $\mathbf{X_0}$ and $\mathbf{X_1}$ based on their predictive power on the outcome. We will utilize a data-driven procedure to choose V*, such that the mean-squared prediction error (MSPE) of the outcome is minimized over the pre-program years[2,16,17]

$$argmin_{v \in \mathcal{V}} (Z_1 - Z_0 W^*(V))' (Z_1 - Z_0 W^*(V))$$

Since large sample inferential techniques are not appropriate in this setting due to the small number of units, Abadie et al. (2010) propose a method of inference for synthetic controls based on exact inferential techniques; placebo/permutation test may be used to assess how unusual an effect would be if it were due to chance and thus provide context for the effect size. In short, the effect of the intervention is estimated separately for each unit in the donor pool through a series of placebo (permutation) tests; we iteratively apply the synthetic control method for each control hospital, reassigning the treatment to each control hospital as though it were the one that had implemented the intervention, and obtaining a distribution of "placebo effects". The estimated effect of the actual treated hospital can then be compared to the size of these other effect estimates. The effect of the treatment on the actual affected unit is considered significant when its magnitude is large relative to the distribution of "placebo effects". Typically, the permutation test results are compared for hospitals in which pre-program trends are well predicted by the synthetic control, because hospitals with a poorly matched synthetic control might appear to have more extreme differences resulting from an artifact of poor prediction.

It is possible that even if a synthetic control is able to closely approximate the outcome trend of the treated unit during the preintervention period, it may not be able to do so for all the units in the donor pool during permutation tests. Thus, Abadie et al. (2010) define the root mean squared prediction error (RMSPE) of the synthetic control estimator as:

For $0 \leq t_1 \leq t_2 \leq T$ and $j = \{1, \ldots, J+1\}$ let

$$R_j(t_1, t_2) = \left(\frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} \left(Y_{jt} - \widehat{Y_{jt}^N} \right)^2 \right)^{1/2}$$

where $\widehat{Y_{jt}^N}$ is the outcome on period t produced by a synthetic control for any unit j that is the treated unit and all others are part of the donor pool. The ratio between the post-intervention RMSPE and the pre-intervention RMSPE for unit j is

$$r_j = \frac{R_j(T_0 + 1, T)}{R_j(1, T_0)}$$

and measures the quality of the fit of a synthetic control for unit j during the postintervention period, relative to the quality of the fit in the pre-intervention period and is a useful test statistic for inference based on the permutation distribution. A p-value can be calculated to summarize the results of the permutation tests:

$$p = \frac{1}{J+1}\sum_{j=1}^{J+1} I_{+}(r_1 - r_j)$$

where I_{+} is an indicator function that returns one for non-negative arguments and zero otherwise[18].

SCM offers a suitable alternative evaluation approach when there is a small number of treated and control units. Unlike traditional DiD methods that rely on assumptions that the trend in the pre-implementation period would have continued into the post-implementation period had the intervention not occurred, SCM does not rely on such a parallel trends assumption to estimate the treatment effect. This can be advantageous in situations where it is difficult to establish whether the parallel trends assumption holds. Despite the practical advantages of SCM, successful application of the method depends heavily on data requirements as well as important contextual decisions during the data preparation process. While SCM has been applied widely to study the impact of laws, policies, reforms and other large-scale events, to our knowledge, it has not been used as an evaluation method for programs offered to patients within a health-care system setting. Programs delivered within health systems often have complex causal pathways that require understanding of the specific pathways to seeking care within the health care system, recognize higher level influences on individual behavior, consider multiple outcome measures both proximal and distal to the intervention, and recognize that strict adherence to the protocol is unlikely given geographical and resource variations within different sites of the health care system. This dissertation will apply the SCM to three programs implemented in recent years within Sutter Health, a large health care system in Northern California. In doing so, we hope to determine any impacts of the program, illuminate any methodological limitations, and provide guidance on when and how SCM should be applied within similar settings.

1.2 Specific Aims

In this dissertation, I conduct a SCM analysis of three programs implemented with the Sutter Health system. My specific aims are as follows:

1. To evaluate the impact of the Advancing Health Equity (AHE) Asthma program on 30-day and 90-day return to the emergency department for any reason and for breathing-related difficulties (Chapter 2).
2. To assess the impact of a new protocol for treating individuals at risk for developing Alcohol Withdrawal Syndrome on leaving against medical advice

(AMA), risk of intensive care unit (ICU) admission and hospital length of stay (Chapter 3).

3. To evaluate the impact of the Group Lifestyle Balance (GLB) program on mean weight and mean body mass index (BMI) (Chapter 4).

All three evaluations will utilize electronic health record (EHR) data from Sutter Health's Epic system and be compared against results of an individual-level propensity score matched analysis. The focus of Chapter 5 will be to summarize results from the three applications of SCM and discuss lessons learned from applying this method within a health-care system setting.

Chapter 2 - Evaluation of the Sutter Health Advancing Health Equity Asthma Program Utilizing Propensity Score Matching and the Synthetic Control Method

Abstract

Purpose

To evaluate the impact of the Advancing Health Equity (AHE) Asthma program on 30-day and 90-day all-cause return to the emergency department (ED).

Methods

We conducted an EHR-based retrospective cohort study of African American adults with asthma who enrolled in Sutter Health's AHE Asthma program. We utilized propensity score matching and synthetic controls to examine the program effect at the individual-level and at the hospital-level, respectively. To account for confounding based on comorbid COPD, we considered both the complete cohort and the cohort stratified by COPD status in the individual-level propensity score analysis. For the synthetic control analysis, we used the mean squared prediction error as a measure of fit between the treated unit and its synthetic control during the pre-program period. To obtain inference for the synthetic control method, we performed a series of placebo (permutation) tests in which we implemented the synthetic control method for each control hospital, as though it were the one that had implemented the asthma program. We then compared the treatment effect to the distribution of effects from these placebo tests.

Results

A total of 373 patients were enrolled into the asthma program from January 1, 2019 to February 29, 2020. After matching on the propensity score, the final analytic cohort consisted of 372 program participants and 1383 control patients. The odds of returning to the ED for breathing difficulty related reasons among program participants was 1.28 (0.99, 1.64) times the odds among non-participants. Stratified by COPD status, a similar effect was observed. For 30-day returns to the ED, program participants had 0.85 times the odds of returning compared to non-participants (0.62, 1.16). Again, this effect is mirrored when stratified by COPD status, with odds ratios of 0.63 (0.34, 1.20) and 0.89 (0.61, 1.29) for those with COPD and those without COPD, respectively. Lastly, the odds of returning within 90-days to the ED for program participants was 1.06 (0.83, 1.36) times that of non-participants. This was consistent among those without a history of COPD. Among those with COPD, program participants had 0.91 (0.54, 1.54) times the odds of 90-day return compared to non-participants.

In the synthetic control analysis, our results did not suggest a significant effect of the asthma program on 30-day or 90-day return to the ED, at either ABSMC location. The differences in outcomes between the actual ABSMC Ashby and its synthetic control indicate that there was an average increase during the program of approximately 0.52 returns and 1.2 returns, for 30-day and 90-day returns respectively. For ABSMC Summit, the average increase during the program was approximately 0.72 and 2.29 returns for 30-day and 90-day returns respectively. However, for both 30-day and 90-day returns to the ED, the synthetic ABSMC Ashby did not closely reproduce the outcome trend for the real ABSMC Ashby during the study period. The same trend is reflected in ABSMC Summit, where the synthetic ABSMC Summit also did not approximate closely the real ABSMC Summit.

Conclusions

We did not find evidence that the AHE Asthma program had an effect on our outcomes of interest at the individual level. Program participants appear to have fewer returns to the ED within 30-days, although this result was not significant. We also did not find any significant effects of the program at the hospital-level. Future explorations of synthetic controls in this setting should include additional hospital-level covariates to obtain better pre-program fit.

Abbreviations

ABSMC – Alta Bates Summit Medical Center
AHE – Advancing Health Equity
ED - Emergency department
ER – Emergency room
COPD – chronic obstructive pulmonary disorder
HEI – health equity index
AA – African American

2.1. Background

In the United States, the burden of asthma disproportionately affects low-income and minority populations; in particular, African American (AA) and Latino populations exhibit higher rates of asthma and poorer asthma outcomes, including hospitalizations and deaths[19–23]. Because asthma is considered an ambulatory care sensitive condition[24], ER visits and hospitalizations for asthma are a marker for poor self-management of asthma. Recent trends indicate that the black to white racial disparity in asthma hospitalization is widening among the U.S adult population[25]. Much of this disparity can be attributed to unequal access to preventative asthma care for Blacks/AA compared to Whites. One study found that significantly fewer Blacks/AAs report care that is consistent with asthma recommendations, including use of inhaled corticosteroids (ICS), self-management education, education to avoid triggers and use of specialist care[26].

Black and Hispanic individuals tend to underuse long-term control medications, due to a variety of reasons, including under-prescribing by physicians to patients who should receive them, as well as socioeconomic challenges or language and literacy barriers[23]. All these factors result in a pattern of health behavior characterized by under-use of routine health care services and a reliance on emergency care services for asthma.

Sutter Health is a large health-care system in Northern California that provides care to >3 million people per year across diverse urban and rural communities, and thus offers a unique opportunity to enhance our understanding of health disparities and further solutions for achieving health equity. In 2017, Sutter Health designed and implemented a novel Health Equity Index (HEI) to identify and quantify health inequities for ambulatory care sensitive condition management in health-care systems[27]. The HEI represents the average ratio of observed-to-expected hospital encounters for a given disease or diagnosis in a set period of time and can be stratified by racial and ethnic subgroups[28]. A ratio of 1.0 or less indicates that the outcomes for the group are at least as good as or better than expected under conditions of equity. A ratio greater than 1.0 indicates that the outcomes are not as good as expected for that particular group. Sutter Health's HEI is the first implemented health equity metric to combine real-time, health care system data with external demographic, prevalence and utilization statistics to produce a value that can be attributed to specific racial and ethnic groups.

In 2016, Alta Bates Summit Medical Center (ABSMC) had 649 patients who utilized the ED a total of 877 times for asthma-related reasons, resulting in an HEI value of 1.5. This was largely driven by Black/AA patients, in particular Black/AA women 60+ years of age and Black/AA men 45-64 years, who were disproportionately utilizing emergency services for asthma compared to what would have been expected under conditions of equity[27]. Based on this information, Sutter Health leaders at ABSMC began to design and implement a pilot program to address disparate outcomes for Black/AA patients with asthma who presented at ABSMC.

To thoroughly evaluate the impact of Sutter Health's AHE asthma program, we utilize two different approaches. First, we conducted a propensity score matched analysis at the patient level in which we compared AHE asthma program participants to control patients who did not receive the program. And second, we conducted a synthetic control analysis at the hospital level, in which ABSMC was compared to other Sutter Health EDs that did not implement the program. Specifically, we examined 30-day and 90-day return to the ED, as well as return to the ED for breathing related difficulties.

2.2. Methods

Study Design and Setting

This is a retrospective EHR-based observational study conducted at Sutter Health, a large, not-for-profit community-based health care delivery system in Northern California that provides medical services across 130 ambulatory clinics and 24 acute care hospitals, including 22 ED sites. All Sutter Health clinics and hospitals are linked by a

single electronic health record system (Epic, Verona, WI). Sutter Health has approximately 11 million ambulatory visits, 870,000 ED visits and 200,000 hospital discharges annually. This study was approved by the UC Berkeley Committee for the Protection of Human Subjects (CPHS) and the Sutter Health Institutional Review Board (IRB) with a Health Insurance Portability and Accountability Act (HIPAA) waiver of authorization and informed consent.

Asthma Program

In January 2019, ABSMC partnered with a community-based federally qualified health clinic, Lifelong Medical Center in East Oakland and Berkeley, to launch the AHE asthma program designed to address the specific needs of Black/AA patients who utilized the ED for asthma-related reasons. Implemented at the two separate ABSMC locations, Summit campus and Ashby campus, the program aims to bring culturally appropriate community based respiratory and primary care to patients, to provide education about the disease and medication self-management, and to use high-touch and high-tech real-time counseling services[27]. Although eligible patients are triggered by going to ABSMC's ED, the asthma program is considered an outpatient program. Patients are invited to participate in the program by trained program coordinators, and receive an intensive, personalized experience. The program consists of:

1. One-to-one educational sessions with a licensed respiratory therapist to review and teach proper medication protocols and adherence.
2. Home visits to assess for allergens and triggers and identify ways to ameliorate them.
3. Group classes to foster community and engage open discussions for long-term management.

Cohort identification

AHE program patients were identified via a monthly patient list sent from Lifelong Medical Clinic. In order to be eligible for the program, patients had to be aged 18 years or older, self-identify as African American, and have a discharge diagnosis of asthma from one of the two ABSMC ED campuses. Patients enrolled in the program from January 1, 2019 to November 11, 2020 were included in this study. To identify potential controls, we queried the Sutter Health Clarity EHR system using the same criteria for the same time period across the 20 non-intervention Sutter Health hospitals. Basic demographic information was also extracted from the EHR including age, sex, health insurance information, as well as medical history of COPD or any respiratory diseases within the last four years (**Appendix Table A1**).

Covariates

We selected covariates to include in the propensity score model based on available data, as well as a priori specification of variables that are believed to influence the probability of participating in the asthma program. These variables included age, sex, history of COPD, history of respiratory diseases, and health insurance type.

For the synthetic control analysis, we gathered hospital-level aggregate data on age, sex, race, ethnicity, and health insurance type from January 1, 2015 to February 29, 2020.

Outcome measures

We considered two primary outcomes of interest: return to the ED within 30 days and 90-days for any reason (both dichotomous outcomes). We assigned the anchor visit as the first initial ED visit during the study timeframe. We looked for any return to the ED within 30 days or 90 days after the initial anchor visit and created a binary indicator of having a return during the respective period. Because 30-day and 90-day all-cause return was an unspecific outcome, we also included return to the ED for breathing difficulty-related reasons as a secondary outcome for the propensity score matched analysis. We defined this for each patient as having at least one encounter visit for breathing difficulty reasons. In a subgroup analysis, we estimated the effect of the program on these outcomes, stratified by patient history of COPD.

For the hospital-level synthetic control analysis, we aggregated counts of 30-day and 90-day returns to the hospital level.

Statistical Analyses

All analyses were conducted in R version 4.0.3.

Propensity score matching

Observational studies are increasingly being used to estimate the causal effects of interventions. As it is not possible to randomly assign the asthma program intervention to patients, treatment selection will be influenced by subject characteristics. Thus, baseline characteristics between treated and untreated subjects may differ systematically and needs to be taken into account when estimating the effect of treatment on outcomes. Thus, we will utilize propensity score methods to correct for the treatment-selection bias imposed when estimating effects using observational data.

The propensity score[29] is the probability of receiving treatment conditional on observed baseline covariates. It is essentially a balancing score such that, conditional on the propensity score, the distribution of baseline covariates is similar between treated and untreated subjects[30,31]. In randomized controlled trials, the true propensity score is known. However, in observational studies, it is not known, but can be estimated using data. Propensity score matching, in which treated and untreated subjects are matched together based on a similar propensity score value, allows for estimation of the average treatment effect among the treated (ATT):

$$E[Y(1) - Y(0)|Z = 1]$$

where Y(1) is the potential outcome under active treatment and Y(0) is the potential outcome under control treatment. The ATT is defined as the average effect of the

treatment on the sub-population who ultimately received the treatment. Once a matched sample has been created, the treatment effect can be estimated by comparing the outcomes between treated and untreated subjects in the matched sets. However, since observations within a propensity score matched sample are no longer independent, as treated and untreated subjects have similar values of the propensity score, this lack of independent should be taken into account either by treating the matching variable as a fixed or random effect[30–32].

We estimated the propensity scores as the probability of receiving the asthma program conditional on baseline characteristics. We used the *MatchIt* package[33] to match asthma program patients to control patients 1:4, with replacement, using nearest neighbor matching with calipers width equal to 0.2 of the standard deviation of the logit of the estimated propensity score. We choose to match more than one control for each treated unit in order to increase the power of the procedure[34]. We checked covariate balance in the sample before and after matching by calculating the standardized mean difference (SMD)[32] and via plotting procedures. The SMD is not confounded by sample size and thus allows for appropriate comparison between the balance in the original sample with the matched sample[32]. We considered an SMD greater than 0.1 as the threshold for imbalance[35]. Variables that created imbalance were included in the matching selection model[36].

In a sensitivity analysis, we stratified the original cohort by COPD status, and then matched control patients to program participants within each stratification, again using the nearest neighbor method and a caliper width of 0.2.

We created a propensity score matched sample by matching program patients with controls that had a similar propensity score, thus treated and control patients within the same matched pair have a similar propensity score. We estimate the effect of the asthma program on our pre-specified outcomes as the difference in the probability of 30-day return and 90-day return between treatment groups by directly estimating the difference in proportions between treatment groups in the propensity score matched sample. Because these patients within the same matched set have baseline covariates that are, on average, more similar than two randomly selected treated and control patients, they can no longer be considered independent[32]. Therefore, we calculated Mantel Haenszel adjusted odds ratios, stratifying by the matched set, to account for the matched nature of the propensity score matched sample. We used McNemar's test for correlated binary proportions to assess statistical significance[30–32]. We chose this method, as opposed conditional logistic regression that is often used for matched-pairs data, because conditional logistic regression has been shown to result in biased estimation of odds ratios[32].

Synthetic control method

While the program is administered at the individual patient level, the health equity leadership committee at Sutter Health are also interested in the clinic-level effects of the program, and thus we will also perform a synthetic control analysis to assess any

hospital-level changes resulting from the program with the goal of evaluating the program for scale-up and roll-out to other Sutter Health clinics.

The synthetic control method (SCM) is an approach to program evaluation in which one or a small number of units are subject to intervention, and a comparative control unit is constructed such that the outcomes of the control units are weighted to construct the counterfactual outcome of the treated unit(s) in the absence of the treatment. This method has previously been used to analyze political and economic effects following large-scale events, state-level policy changes, health systems reforms, nutritional interventions, climate changes, and even the current COVID-19 pandemic[2,4,8,17,37–40] where it is difficult to find a single comparison unit that best approximates the relevant characteristics of the treated unit; indeed, a combination of units often provides a better comparison unit than any single unit alone.

SCM offers another tool for program evaluation by using the pre-treatment interval to construct a predicted course among the treated group (after initiation of the intervention) if, contrary to fact, they had continued without intervention. No extrapolation is required as weights are restricted to be non-negative and sum to one, and the weights are calculated and chosen without seeing the post intervention data, reducing the risk of cherry picking or p-hacking, since no outcome data during the intervention is used to create the synthetic control[18]. The contribution of each control unit to the synthesized control is made explicit and offers transparency in the selection of the best counterfactual[18]. Additionally, SCM provides a visual representation of the impact of the intervention and how it varies over time, as well as a clear visualization of the actual discrepancy between the treated unit and the convex combination of untreated units, something that propensity score methods do not provide. Plots are produced that display what the observed outcome looks like compared to what would be expected in the absence of the intervention. Lastly, predictions from SCM may serve as input to the propensity score model, by acting as a baseline counterfactual.

As the asthma program began in January of 2019, we define the years prior to 2019 as the pre-intervention period, and from January 1, 2019 to February 29, 2020 as the post-intervention period. We use a data-driven approach that takes a weighted combination of hospitals to create a synthetic “control” ABSMC hospital, which provides an estimate of the expected outcome in ABSMC if the asthma program had not been implemented. The hospitals that comprise ABSMC’s synthetic control are selected by the method based on their pre-2019 trends in covariate and outcome values. Those that are best able to predict the pre-2019 outcome trends for ABSMC are chosen to be included in the synthetic control. The expected outcome trend for ABSMC from January 2019 to February 29, 2020 in the absence of the asthma program are then compared against the observed outcome trend. The difference between the observed and expected values is the treatment effect of interest, or the impact of the asthma program.

We will utilize the *Synth* package[41] to choose the vector of weights by minimizing

$$\|X_1 - X_0 W\| v = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

Where V is a matrix allowing for different weights to be applied to variables in $\mathbf{X_0}$ and $\mathbf{X_1}$ based on their predictive power on the outcome. The *synth()* function also allows for a data-driven procedure to choosing V*, such that the mean-squared prediction error (MSPE) of the outcome is minimized over the pre-program years. Thus, the MSPE measures the quality of the fit of a synthetic control for the treated unit in the post-program period[2,17,37].

Traditional large sample inferential techniques are not appropriate in this setting due to the small number of units; however, exact inferential techniques, such as a permutation test may be used to assess how unusual an effect is[42]. As advised in prior literature, we applied the synthetic control method to each control hospital, as though it were the one that had implemented the asthma program beginning January 2019. The estimated effect for ABSMC can then be compared to the size of these other effect estimates from the placebo tests. We used the *SCtools*[43] package in R to generate and plot placebos as well as calculate and plot MSPE. We present the results of the test for all hospitals, in addition to the results for those hospitals with two times the mean squared prediction error (MSPE) observed for ABSMC.

2.3. Results

ABSMC consists of two separate sites, the Ashby campus and the Summit campus. While Abadie et al., 2010 suggests combining multiple treated units, we did not want to obscure differences in outcome trends and thus created two separate synthetic controls for each ABSMC hospital site that implemented the asthma program (**Figure 2.1**)

We show in **Figures 2.2 – 2.3** that plotting the outcomes of ABSMC compared to the 19 other control hospitals demonstrates that a simple average of the control units does not closely approximate the outcome trend in ABSMC. One hospital, California Pacific Medical Center (CPMC) Van Ness, did not have sufficient covariate data during the pre-period to be included. Thus, this motivates our decision to construct a synthetic control for each ABSMC site, with the goal of approximating a better counterfactual.

Study Cohort Description

We identified 373 patients who enrolled in the asthma program from January 1, 2019 to February 29, 2020, and 2,093 control patients that met the eligibility criteria for participating in the asthma program during the same time period. All but one (372 of 373) asthma program patients were matched to at least one control patient, although not all had four matches. Thus, the final matched data set consisted of 1755 total patients. After matching on the propensity score, we found that all covariates were balanced, as shown in **Table 2.1**. 76 asthma program patients had a history of COPD, compared to 308 control patients without any history of COPD. The back-to-back histograms in **Figures 2.4 and 2.5** show the distributions in propensity scores before

and after the match. This match suggests that the two groups are more similar in terms of their propensity scores after the match.

30-day and 90-d all-cause return

Among the program participants, 58 (15.59%) patients had an ED return within 30 days and 126 (33.87%) patients had an ED return within 90 days. We found that the odds of a 30-day return to the ED for any reason among program participants was 0.85 (0.62, 1.16) times the odds of return among control patients (**Figure 2.6).** When stratified by COPD status, the effect was similar. The odds ratio associated with 30-day return was 0.63 (0.34, 1.20) among those with a history of COPD, and 0.89 (0.61, 1.29) among those without a history of COPD. We saw a change in direction for the effect on 90-day returns. Program participants had 1.06 (0.83, 1.36) times higher odds of 90-day return compared to control patients (**Figure 2.7**). Although not significant, among those with COPD history, the odds of 90-day return among program participants was 0.91 (0.54, 1.54) times the odds among control patients. Among those without COPD history, the odds of 90-day return was 1.09 (0.82, 1.45) times higher among program participants.

Breathing Difficulty

Matched program participants returned to the ED for breathing difficulty related reasons in similar proportions to the control patients: 33.1% and 28.4 %, respectively. **Figure 2.8** shows that the odds of returning to the ED for breathing difficulty reasons was 1.28 (0.99, 1.64) times higher for patients in the asthma program compared to control patients. When stratified by COPD status, the adjusted odds ratio for returning to the ED for breathing difficulty was 1.56 (0.89, 2.75) and 1.25(0.95, 1.64), among those with history of COPD and no history of COPD, respectively.

Synthetic control results

A total of 19 other Sutter EDs comprised the potential control hospitals. Appendix figures A1 and A2 show a panel view of the available outcome data for each hospital at each month since January 2015. There was insufficient covariate data from one location, CPMC Van Ness Campus, and thus was dropped from the analysis.

Tables 2.2-2.5 compares the pre-program characteristics of ABSMC to those of the synthetic ABSMC and also to those of a population weighted average of all 19 hospitals in the donor pool. Overall, this suggests that the synthetic ABSMC provides a much better comparison for ABSMC than the average of our sample of other Sutter hospitals.

Tables 2.6-2.9 displays weights of each control hospital in the synthetic ABSMC. The weights indicate that 30-day return trends in ABSMC Ashby prior to the asthma program is best reproduced by a combination of CPMC Davies Campus, CPMC Mission Bernal Hospital, and Delta Medical Center. All other states in the donor pool are assigned zero weights. 90-day trends at ABSMC Ashby are best approximated by CMPC Davies Campus, Delta Medical Center, Eden Medical Center, Sutter Medical Center Sacramento and Sutter Solano Medical Center. The synthetic ABSMC Summit 30-day

return trend is comprised of only two hospitals. Most notably, the synthetic control for ABSMC Summit's 90-day return trend is comprised of a single unit, Delta Medical Center.

We used the mean squared prediction errors (MSPE) to measure fit between the treated unit and its synthetic control during the pre-program period. The MSPEs in the pre-program period for 30-day and 90-day returns are listed in **Table 2.10**.

Effect of the AHE Asthma program

Figure 2.9-2.12 displays the 30-day and 90-day returns to the ED for ABSMC Ashby and Summit and their respective synthetic controls for the period of January 1, 2015 to February 30, 2020. The number of 30-day and 90-day returns for the actual ABSMC Ashby were unsteady in the post-program period, dropping to zero in August/September of 2019 and spiking in November 2019. For both outcomes, the synthetic ABSMC Ashby does not closely reproduce the outcome for the real ABSMC Ashby during the study period; in fact, the two lines cross at multiple points during the post-program period. The same trend is reflected in ABSMC Summit, with the synthetic ABSMC Summit unable to approximate well the real ABSMC Summit.

Our estimate of the effect of the AHE asthma program on 30-day and 90-day return to the ED is given by the difference between the actual ABSMC and its synthetic version, or the "gap" between the two curves (**figures 2.13-2.16**). If our hypothesis is correct, we would hope to see a reduction in the number of 30-day and 90-day returns, and thus the gap (treated – synthetic) should be a negative number. However, as the figure indicates, the difference between the two series oscillates back and forth from positive to negative for the entire study period. The differences in outcomes between the actual ABSMC Ashby and its synthetic control indicate that there was an average *increase* during the study period of approximately 0.52 returns and 1.2 returns, for 30-day and 90-day returns respectively. For ABSMC Summit, the average *increase* during the program was approximately 0.72 and 2.29 returns for 30-day and 90-day returns respectively. Thus, our results do not suggest a pronounced effect of the asthma program on reducing 30-day or 90-day returns to the ED, at either ABSMC location.

Placebo tests

To evaluate the significance of our results, we use placebo tests to determine if our results could be driven entirely by chance (how often we would obtain results of this magnitude if we had chosen a hospital at random for the study instead of ABSMC). **Figures 2.17-2.22** displays the results of the placebo test. The gray lines represent the gap associated with each iteration of the test or the difference in 30-day returns between each hospital in the donor pool and its respective synthetic control. The superimposed black line denotes the gap estimated for ABSMC Ashby and Summit, respectively. As specified earlier, the placebo tests allow us to obtain synthetic control estimates for hospitals that did not actually implement the asthma program, and thus we can compare the estimated effect of the asthma program on ABSMC to the distribution

of placebo effects obtained for other hospitals. We consider the effect significant if the estimated effect for ABSMC is unusually large relative to the distribution of placebo effects. As **figures 2.17-2.22** illustrates, the estimated gap for ABSMC does not appear to be unusually large relatively to the distribution of the gaps for the other hospitals in the donor pool.

The pre-program mean squared prediction error (MSPE) for ABSMC Ashby (the average of the squared discrepancies between 30-day returns and 90-day returns in ABSMC Ashby and its synthetic counterpart during the pre-program period) is 2.9 and 5.6, respectively. These are fairly small MSPEs. The lack of fit is more pronounced for ABSMC Summit, with a MSPE of 11.2 and 22.2 for 30-day and 90-day returns respectively. The control hospital with the worst fit compared to its synthetic in the pre-program period is Sutter Medical Center Sacramento, with a MSPE of 11.6 for 30-day returns. This indicates that the outcome trend at Sutter Medical Center Sacramento is not well reproduced by a combination of all other hospitals. For 90-day returns, Delta Medical Center has the highest MSPE (26.7) compared to its synthetic version.

As Abadie et al. (2010) suggests, we repeat the placebo tests again, this time excluding hospitals with poor fit in the pre-program period (pre-program MSPE more than two times the MSPE of ABSMC). For ABSMC Ashby this meant discarding two hospitals with large pre-program MSPE values. Evaluating against only hospitals that have a low MSPE (good fit), the gap for ABSMC Ashby still does not appear unusual. We did not restrict the MSPE for ABSMC Summit as the MSPE for the actual treated unit was higher than all the other placebo test MSPEs.

Finally, we evaluate the ABSMC gap relative to the gaps obtained from the placebo runs by looking at the distribution of the ratios of post/pre-program MSPE. A ratio of the post-program MSPE to the pre-program MSPE measures the quality of the fit of a synthetic control for its treated unit in the post-program period, relative to the quality of the fit in the pre-program period. **Figure 2.23-2.26** displays the distribution of post/pre-program ratios of the MSPE for ABSMC and all 19 control hospitals. The ratio for ABSMC Ashby 30-day returns is about 1.25 and does not clearly stand out in the figure; two control hospitals achieved a larger ratio than ABSMC Ashby. Indeed, we obtained a p-value of 0.15 and 0.1 for 30-day and 90-day returns at ABSMC Ashby respectively. In other words, if we were to assign the program at random, the probability of obtaining a post/pre MSPE ratio as extreme or more extreme than that of the treated unit is 0.15 and 0.1, respectively. Likewise, the p-values for 30-day and 90-day returns at ABSMC Summit are 0.7 and 0.25 respectively.

This further confirms the trends shown in **figure 2.17** – **2.22**. The trend for the actual ABSMC Ashby and ABSMC Summit is not unusually large or small compared to the placebo trends, thus indicating that the effect is not significant.

2.4. Discussion

Selecting appropriate comparison groups is critical to assessing the impact of programs implemented within health care systems. We explore the construction of two different comparison groups applied to an asthma education program within Sutter Health.

Results from the propensity score matched analysis indicate that there was no significant effect of the AHE asthma program on return visits to the ED. While not significant, the magnitude of the odds ratio in the opposite direction we would expect if the program had its intended effect. The results for 30-day returns to the ED indicate that the effect of the program is in the direction we would hope, however the results are not significant. In addition, ED returns for breathing difficulty reasons were also not lower in the treated group.

Based on the synthetic control analysis, at the hospital level, we also did not find evidence that the AHE asthma program at ABSMC was associated with a decrease in the number of 30-day and 90-day returns to the ED, relative to the synthetic control. However, the credibility of a synthetic control estimator depends heavily on its ability to track the trajectory of the treated unit's outcome during the pre-period. As tables 6-9 indicate, the synthetic control may be underfitted, as weights are assigned to only a few hospitals or a single hospital, which suggests that the construction of the synthetic control is not really borrowing much information across the control units. This may suggest that SCM is not the best method for making a prediction model of the treated unit from the control units. Thus, due to the fact that the synthetic ABSMC failed to fit 30-day and 90-day returns for the real ABSMC during the pre-program, we cannot definitely say that there was no effect of the program in the post-program period. Similarly, placebo runs with poor fit prior to the program also do not provide information to measure the relative rarity of estimating a large post-asthma program gap for any given hospital.

Interestingly, for the outcome 90-day returns, Delta Medical Center was given a relatively large weight in forming the synthetic control for both ABSMC Ashby and Summit. However, in placebo tests in which we constructed a synthetic Delta Medical Center, the MSPE comparing the actual Delta Medical Center to its synthetic resulted in a higher value (11.6). One possible explanation of this is that ABSMC is more extreme in terms of the outcome compared to the control hospitals, so the method selects other extreme units as good comparators. However, when they rotate into the role of the "treated" unit in placebo tests, they are also difficult to fit well by a combination of all other hospitals.

Actual counts of the outcome were very low, both prior to program implementation and after program implementation, with many months in which the outcome was zero. As the nature of synthetic controls focuses on a single treated unit (or a small number of treated units), it is possible that small effects are indistinguishable from other influencers of the outcome in the treated unit, especially if the outcome is highly volatile[18] . As a

result of this, the impact of “small” programs with effects of a magnitude similar to the volatility of the outcome may be difficult to detect.

We acknowledge several limitations in our study. First, we had limited ability to replicate all of the relevant characteristics of the treatment population in the comparison group given the available data. It is possible that there are other confounding factors that we did not have data on, and thus were not able to adjust for in our analysis. Additionally, while the synthetic control method allowed us to capture the desired scope of impacts of the asthma program, it is non-specific and does not allow us to test the specific mechanism that led to the change. It is possible that unrelated factors we did not control for changed during the study period, and though unrelated to the asthma program, affected rates of returning to the ED. As such, a major limitation of the synthetic control method is that it does not allow us to distinguish between co-occurring events. To explore this further, future analyses should incorporate techniques such as a negative control, which estimates the effect of an exposure on an outcome that it should be plausibly impact, but which may be affected by a confounding factor[44]. If an effect is still observed, it can be assumed that there is confounding or bias present. Another valid concern in the context of this study is the potential for spillover effects. It is possible that other hospitals may have had their own asthma education programs, and thus we would expect the synthetic control estimator to be attenuated, thus underestimating the effect of the program.

This study also benefits from several strengths. We utilized two evaluation methodologies to study our outcomes at both the patient-level and the hospital-level. We chose to use the synthetic control method in order to improve our control group selection. One of the central motivations for using the synthetic control method was that it provides researchers with a quantitative tool to select appropriate comparison groups. In our analysis, a handful of control hospitals emerged as potential comparisons to ABSMC, which we displayed in **Table 2.3**, making explicit the contribution of each comparison unit to the counterfactual of interest. Traditional regression analysis fails to provide such a transparent list, as typically all units contribute to the regression fit and the contribution of units with positive regression weights may be counteracted by negative weights. In addition, because synthetic control weights are non-negative and sum to one, they avoid extrapolation outside the area of support, something that regression does due to the fact that weight may fall outside of the (0,1) bounds.

We believe that future explorations of the use of synthetic controls to evaluate a health care system program should incorporate additional hospital-level variables in order to obtain a better pre-program fit. Many of the predictor variables used in our synthetic control analysis were patient-level characteristics, aggregated to the hospital level. Information related to hospital resources, utilization costs, socioeconomic factors, etc. may provide added value.

2.5. Conclusions

Health care systems research needs additional program evaluation tools to quickly evaluate the success of their programs. In this paper, we utilize propensity score techniques to compare outcomes between asthma program participants and their matched control patients, as well as apply the data-driven synthetic control method to select the most appropriate comparison unit for ABSMC. We find that the AHE asthma program does not appear to have an effect at the individual level, or at the hospital level. Using placebo tests, we show that the effect of the asthma program is not unusually large compared to placebo estimates of the program effects among the control units.

2.6. Figures

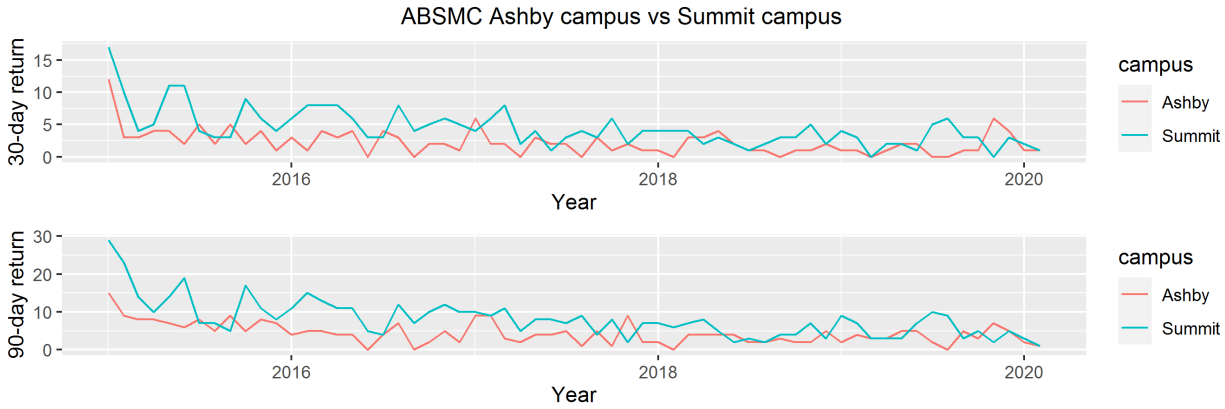


Figure 2.1. Trends in 30-day and 90-day return to the ED: ABSMC Ashby vs ABSMC Summit

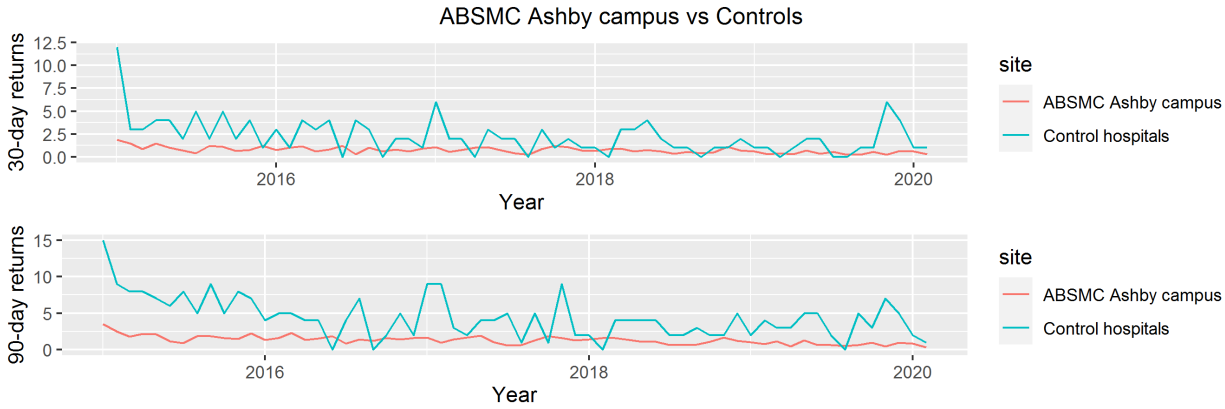


Figure 2.2. Trends in 30-day and 90-day return to the ED, ABSMC Ashby vs. other Sutter Health hospitals.

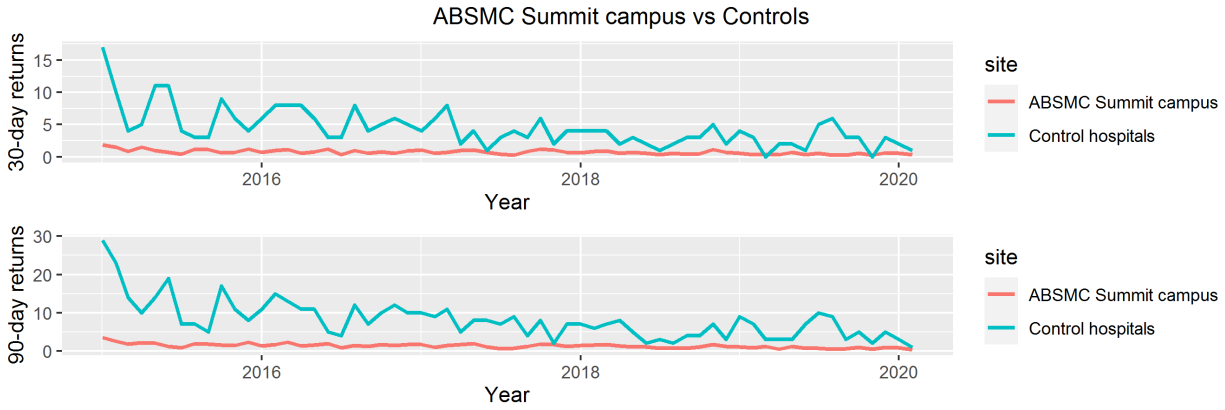


Figure 2.3. Trends in 30-day and 90-day return to the ED, ABSMC Summit vs. other Sutter Health hospitals.

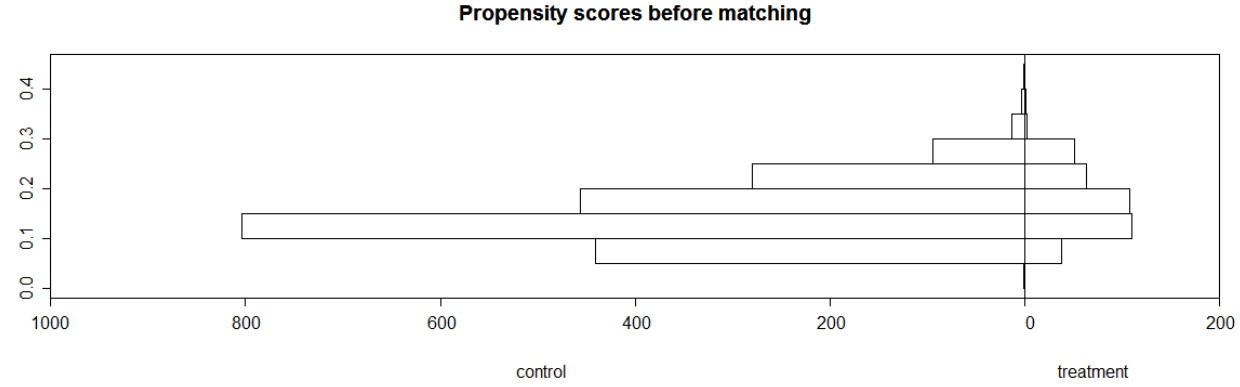


Figure 2.4. Distribution of propensity scores before matching.

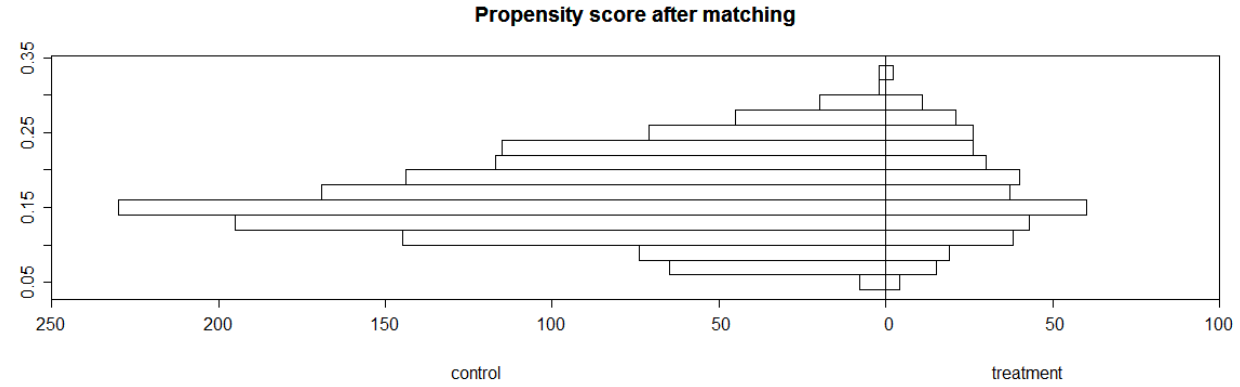


Figure 2.5. Distribution of propensity scores after matching.

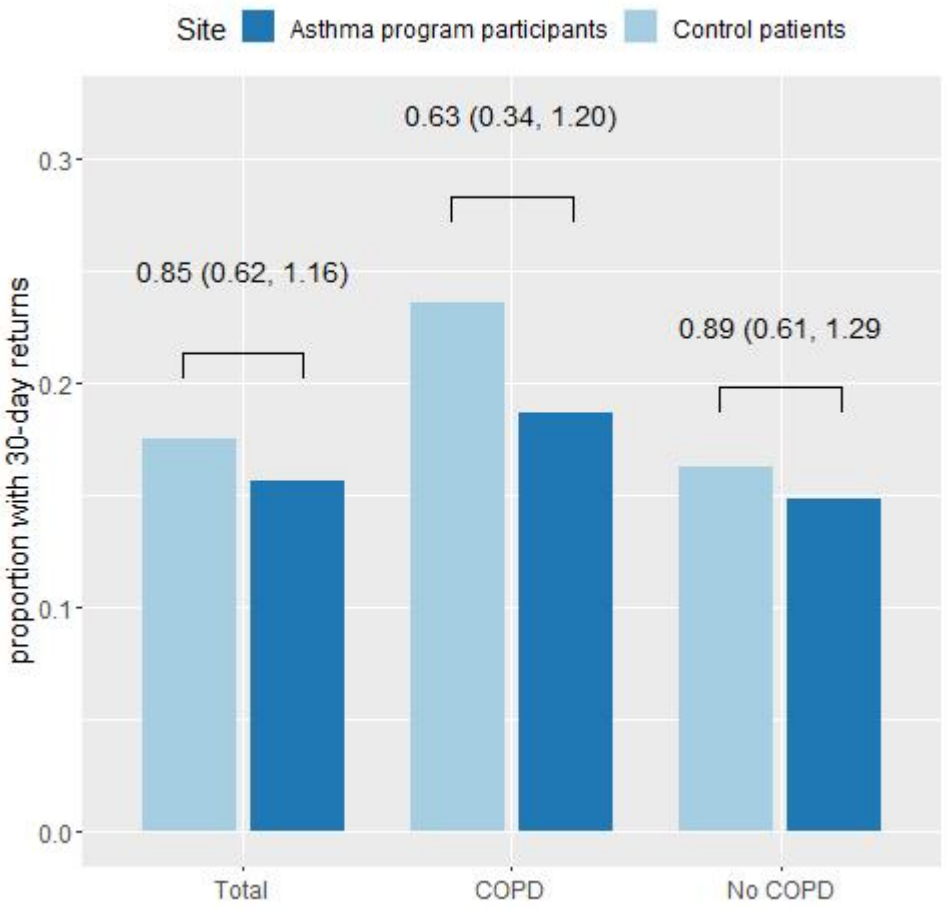


Figure 2.6. 30-day returns for asthma program participants and control patients.

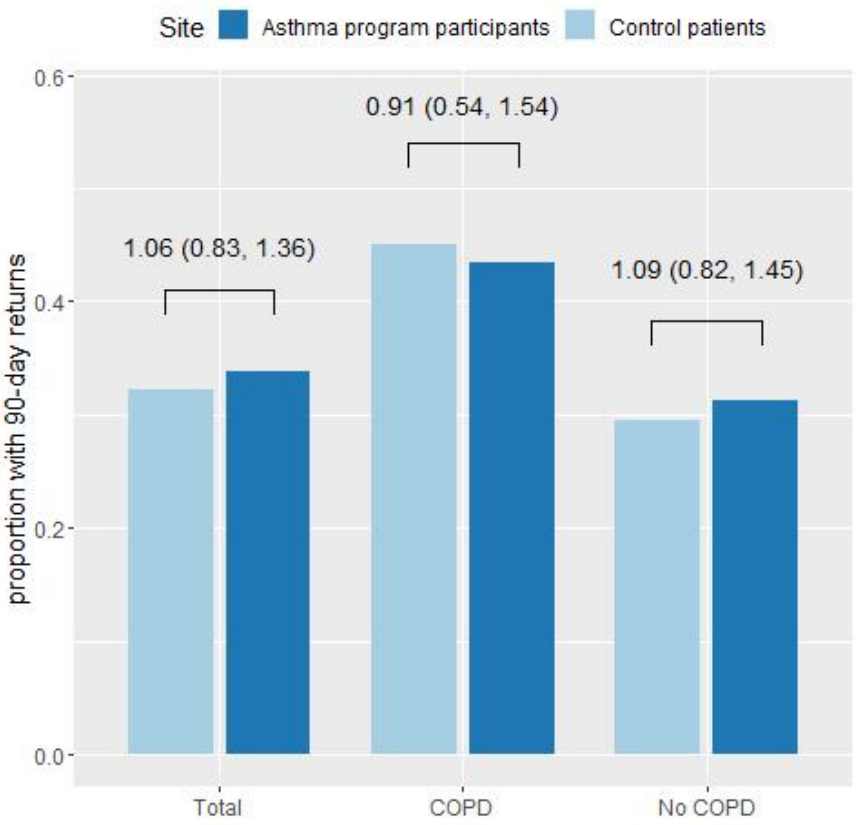


Figure 2.7. 90-day returns for asthma program participants and control patients.

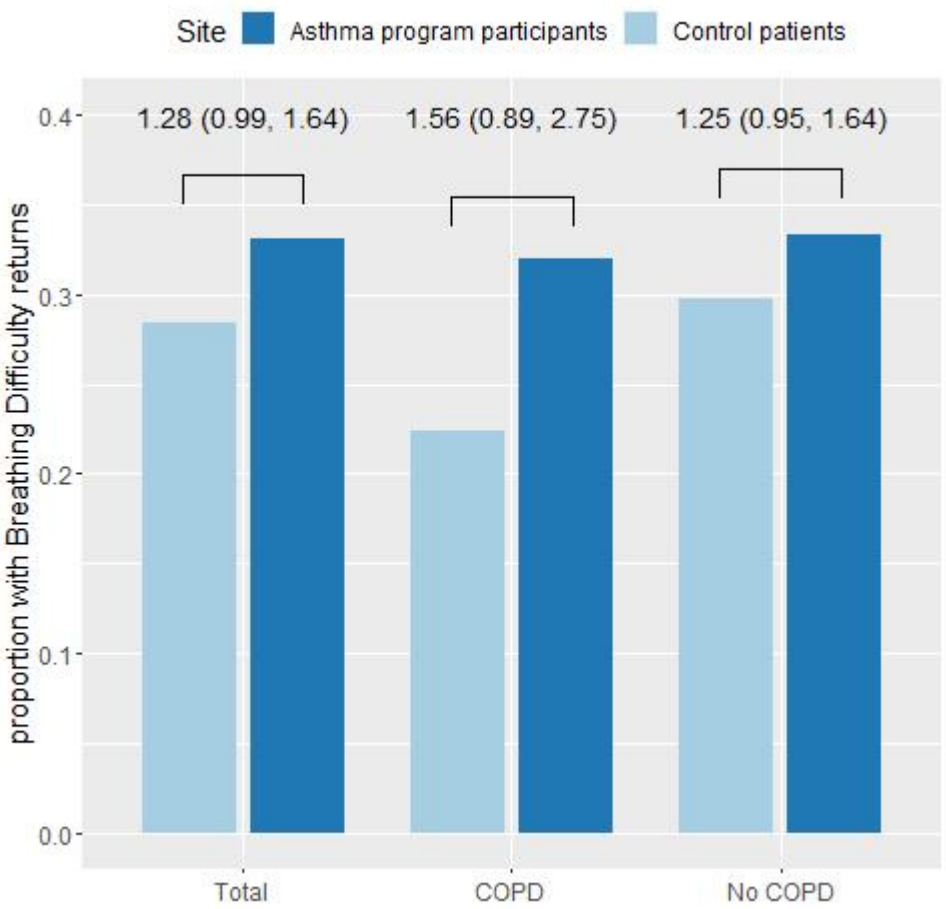


Figure 2.8. Breathing Difficulty returns for matching asthma program participants and control patients.

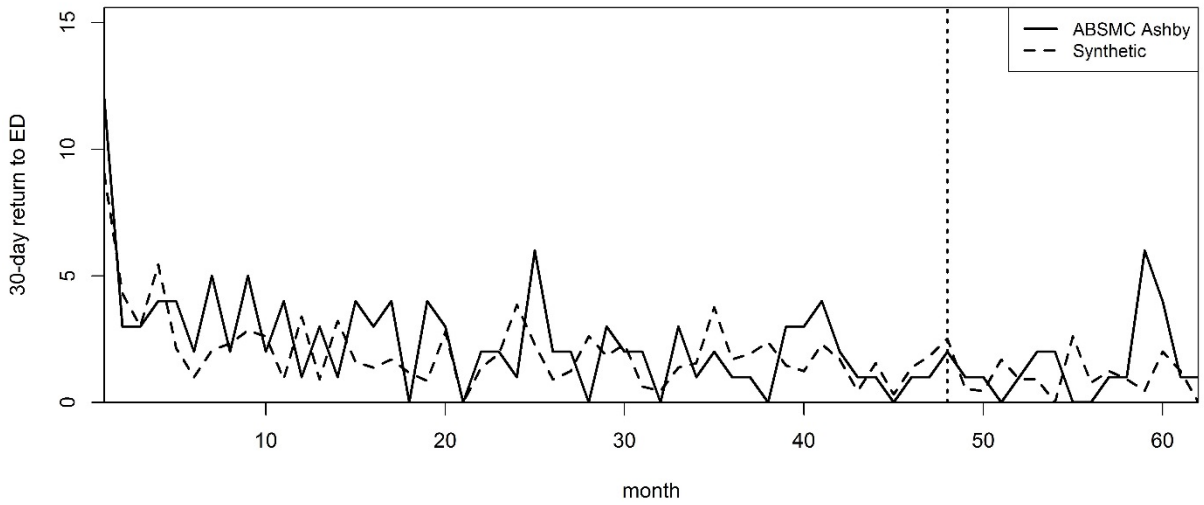


Figure 2.9. Trends in 30-day return to the ED, ABSMC Ashby vs Synthetic ABSMC Ashby.

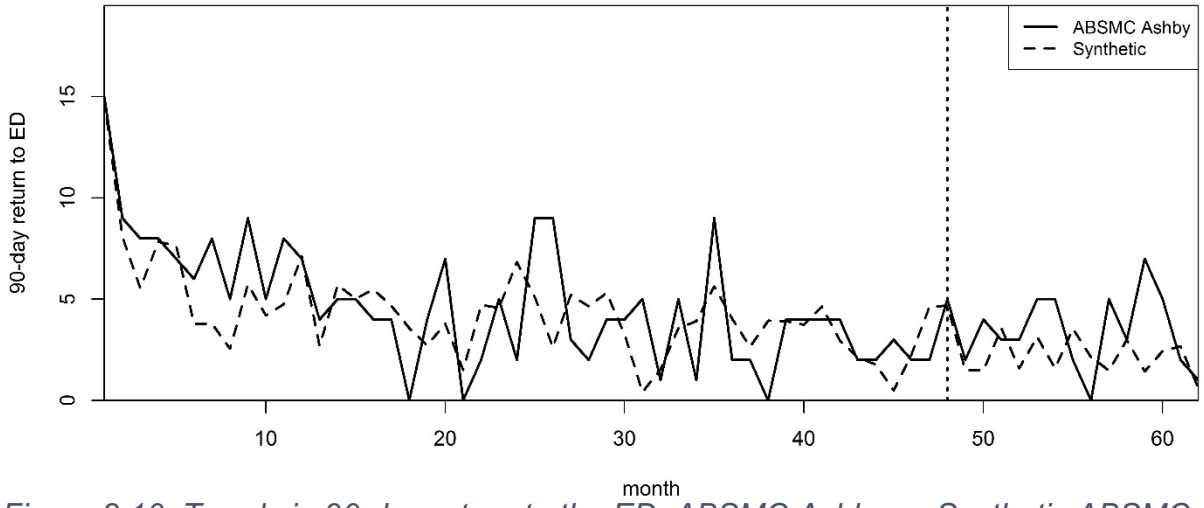


Figure 2.10. Trends in 90-day return to the ED, ABSMC Ashby vs Synthetic ABSMC Ashby.

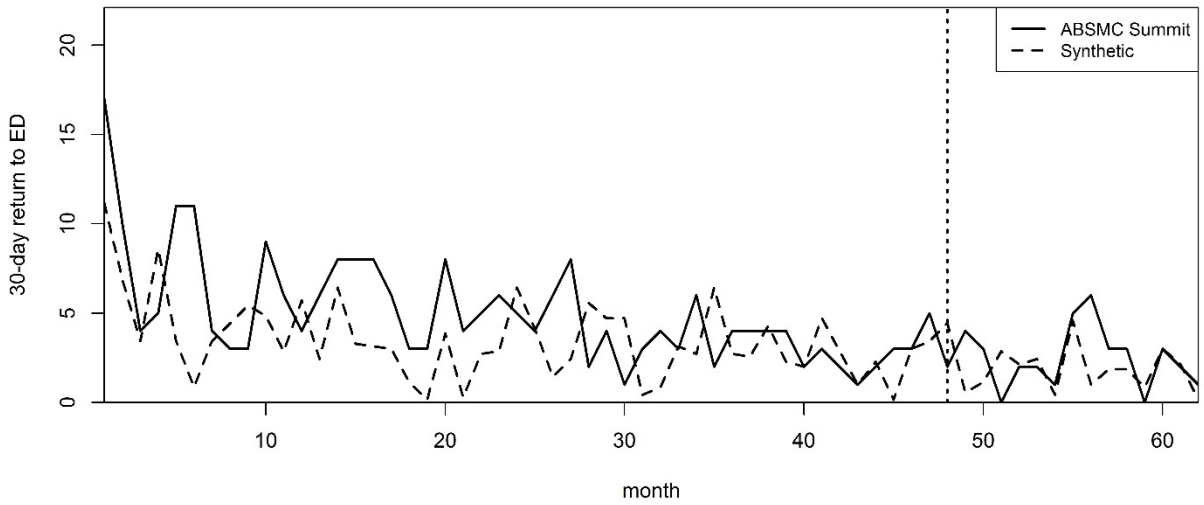


Figure 2.11. Trends in 30-day return to the ED, ABSMC Summit vs. Synthetic ABSMC Summit.

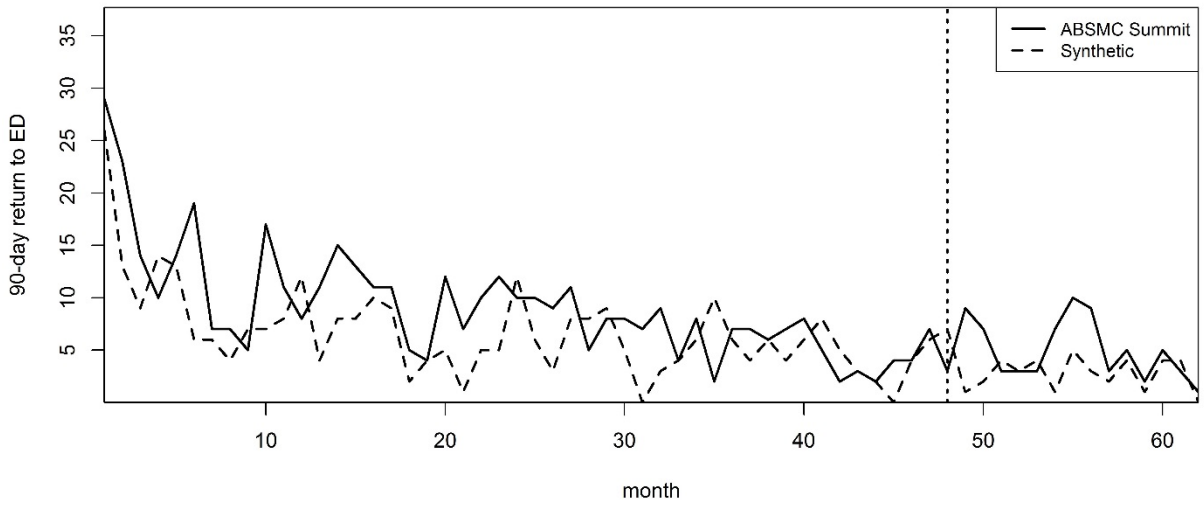


Figure 2.12. Trends in 90-day return to the ED, ABSMC Summit vs. Synthetic ABSMC Summit.

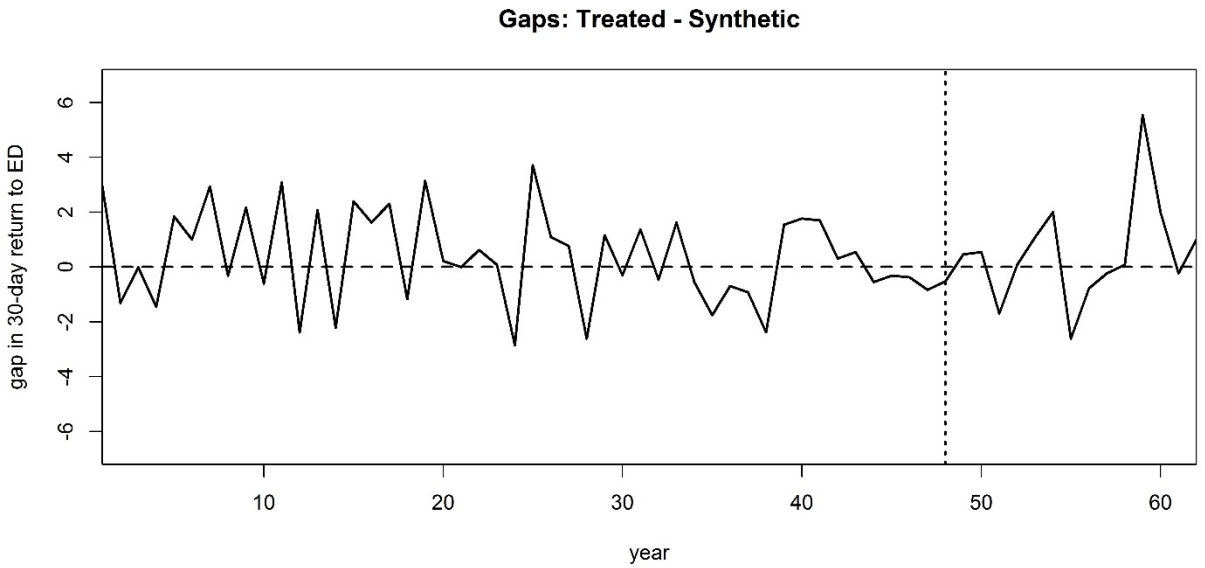


Figure 2.13. Gap in 30-day returns between ABSMC Ashby and synthetic ABSMC Ashby.

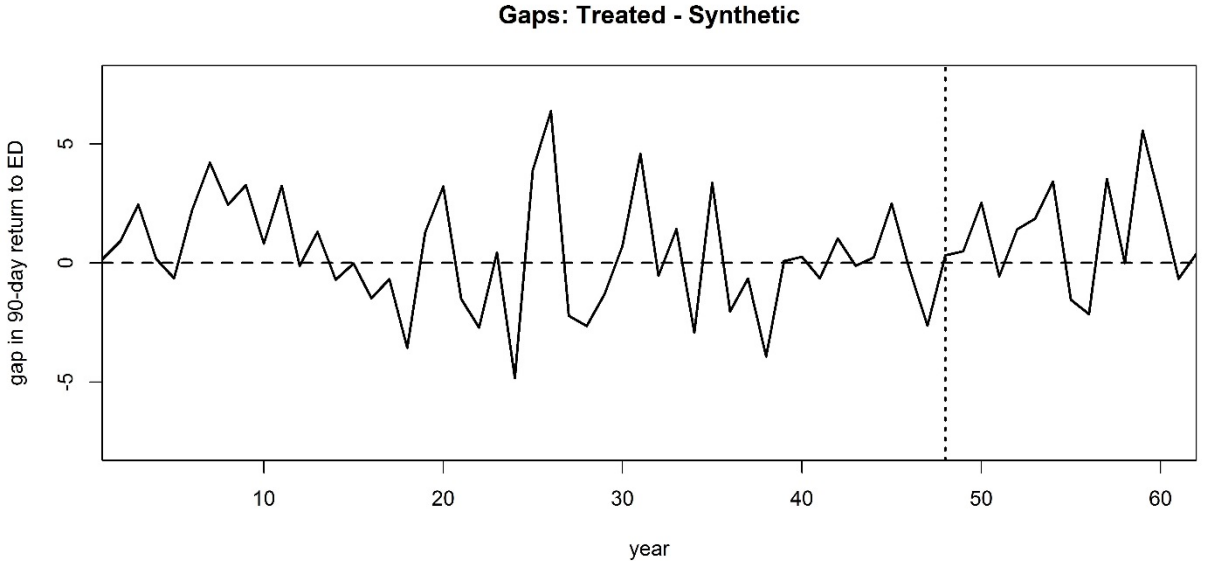


Figure 2.14. Gap in 90-day returns between ABSMC Ashby and synthetic ABSMC Ashby.

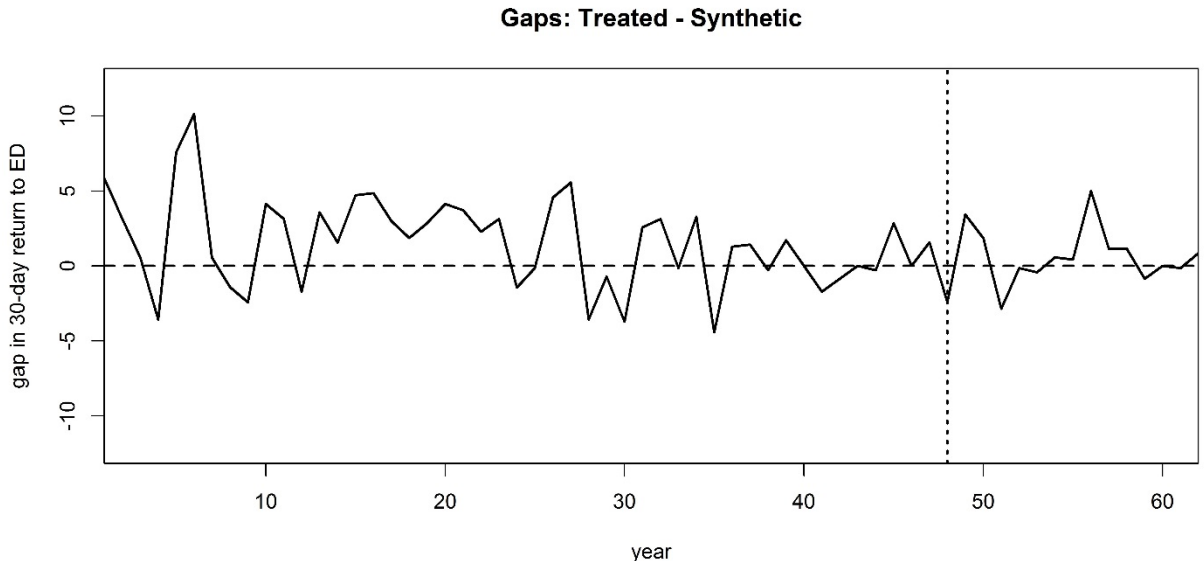


Figure 2.15. Gap in 30-day returns between ABSMC Summit and synthetic ABSMC Summit.

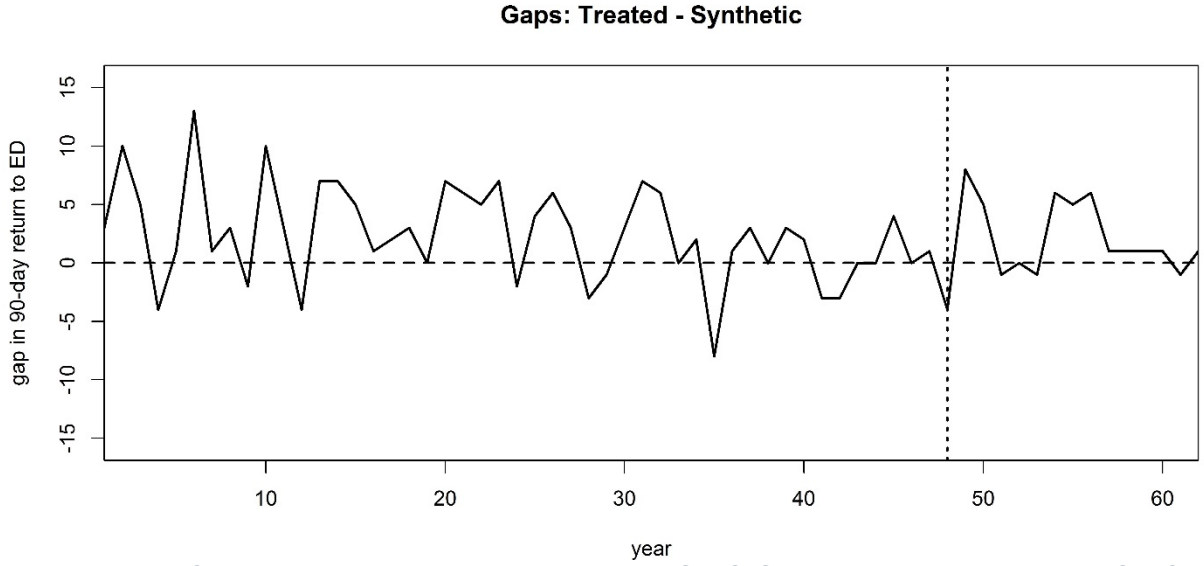


Figure 2.16. Gap in 90-day returns between ABSMC Summit and synthetic ABSMC Summit.

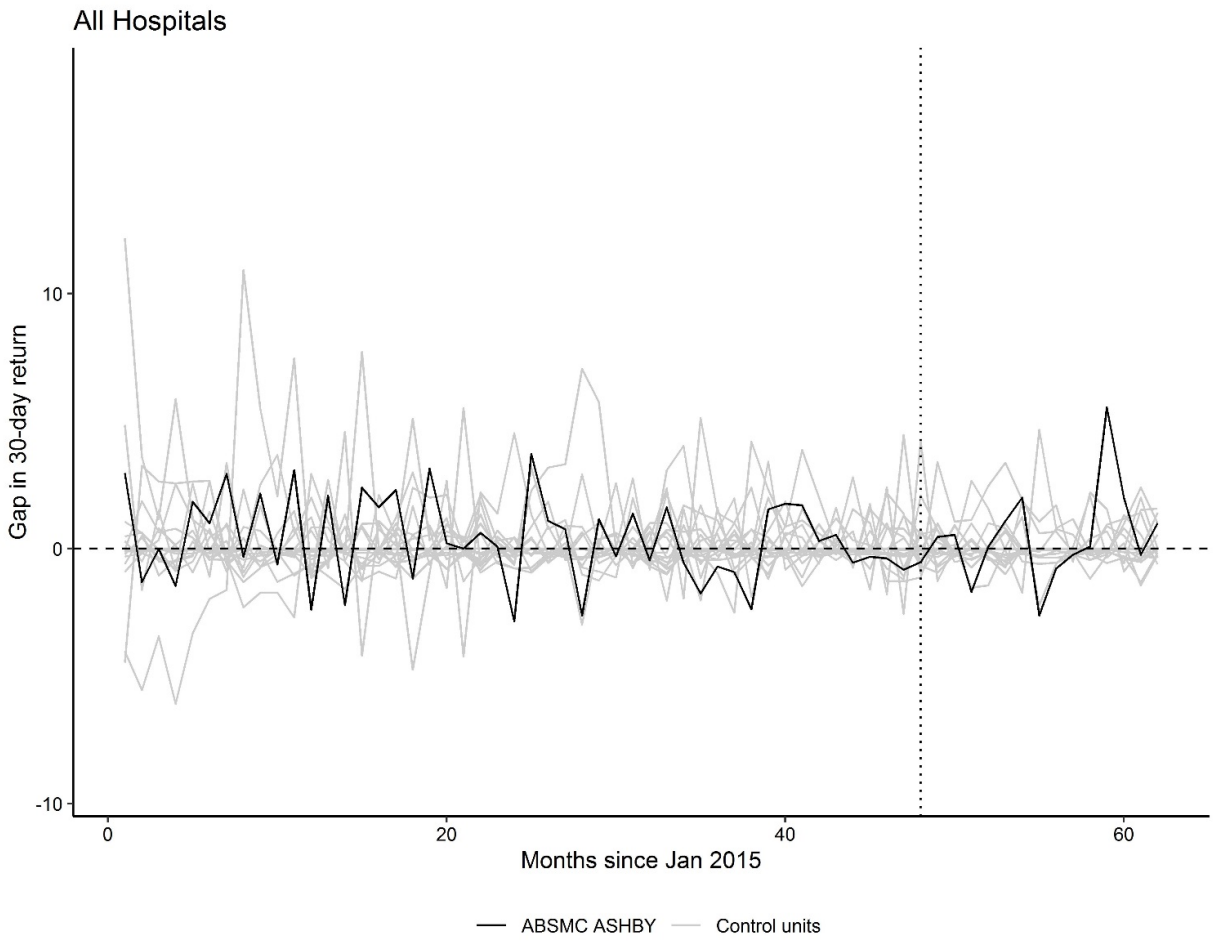


Figure 2.17. Gap in 30-day returns between control hospitals and ABSMC Ashby and their respective synthetic controls, all hospitals.

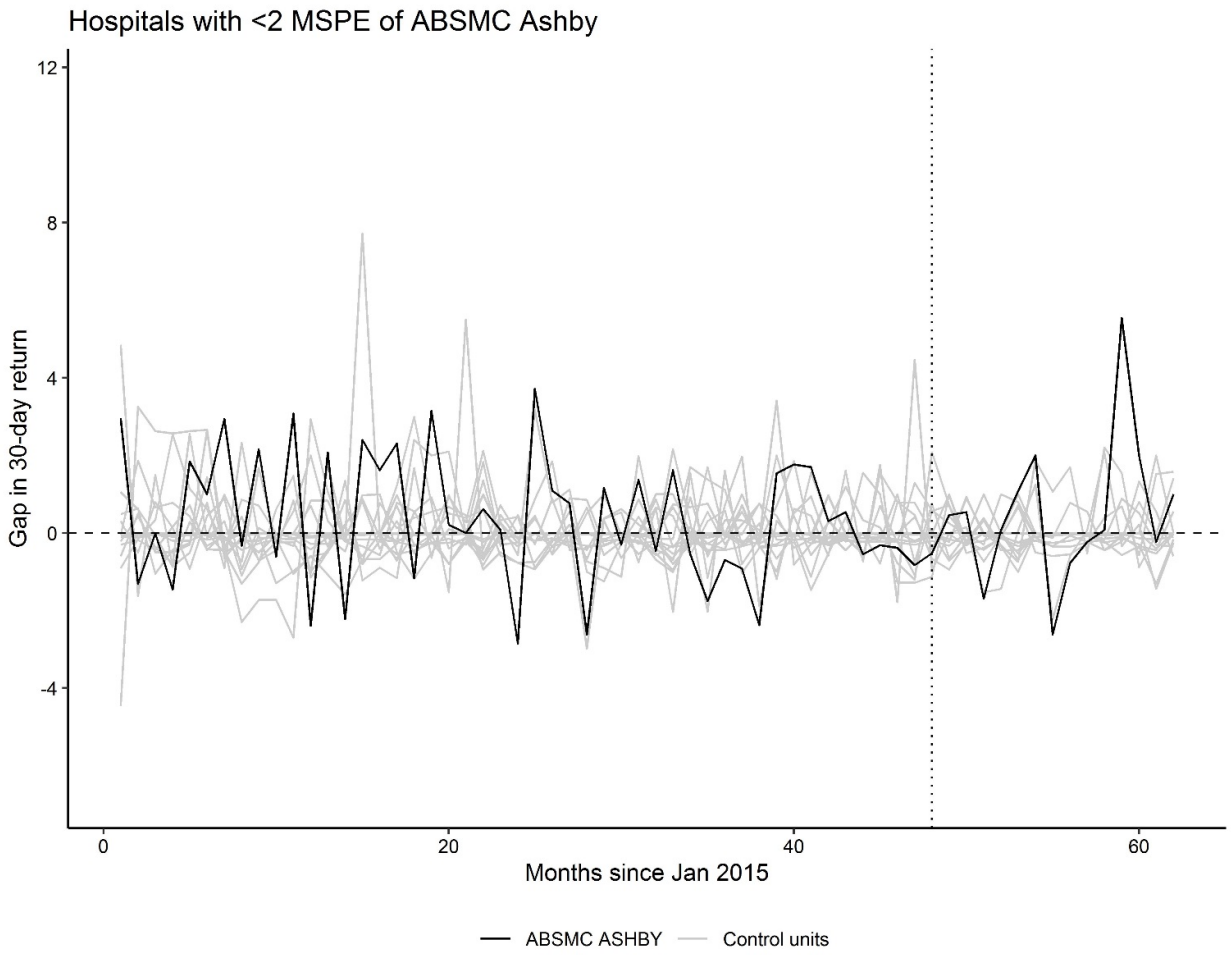


Figure 2.18. Gap in 30-day returns between control hospitals and ABSMC Ashby and their respective synthetic controls, hospitals with <2 MSPE of ABSMC Ashby.

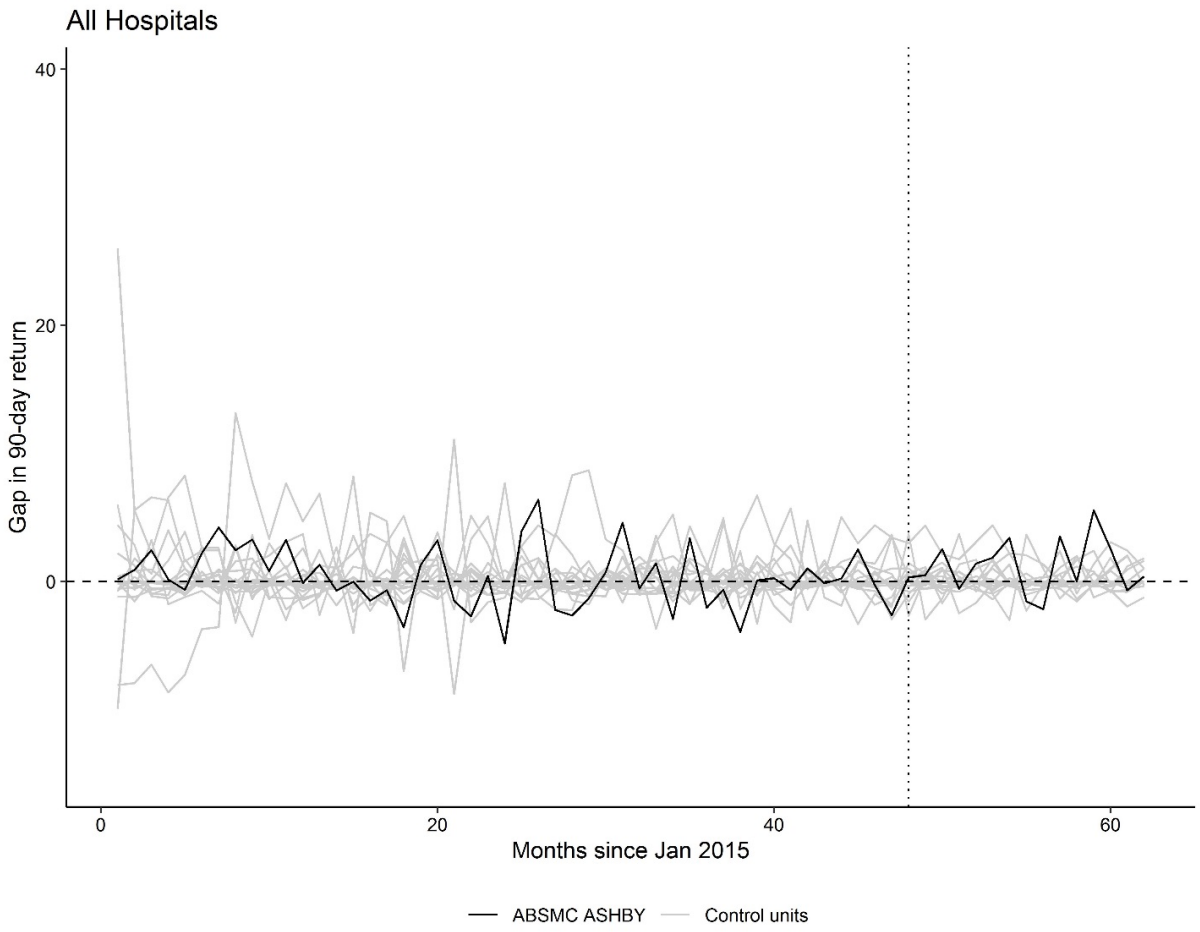


Figure 2.19. Gap in 90-day returns between control hospitals and ABSMC Ashby and their respective synthetic controls, all hospitals.

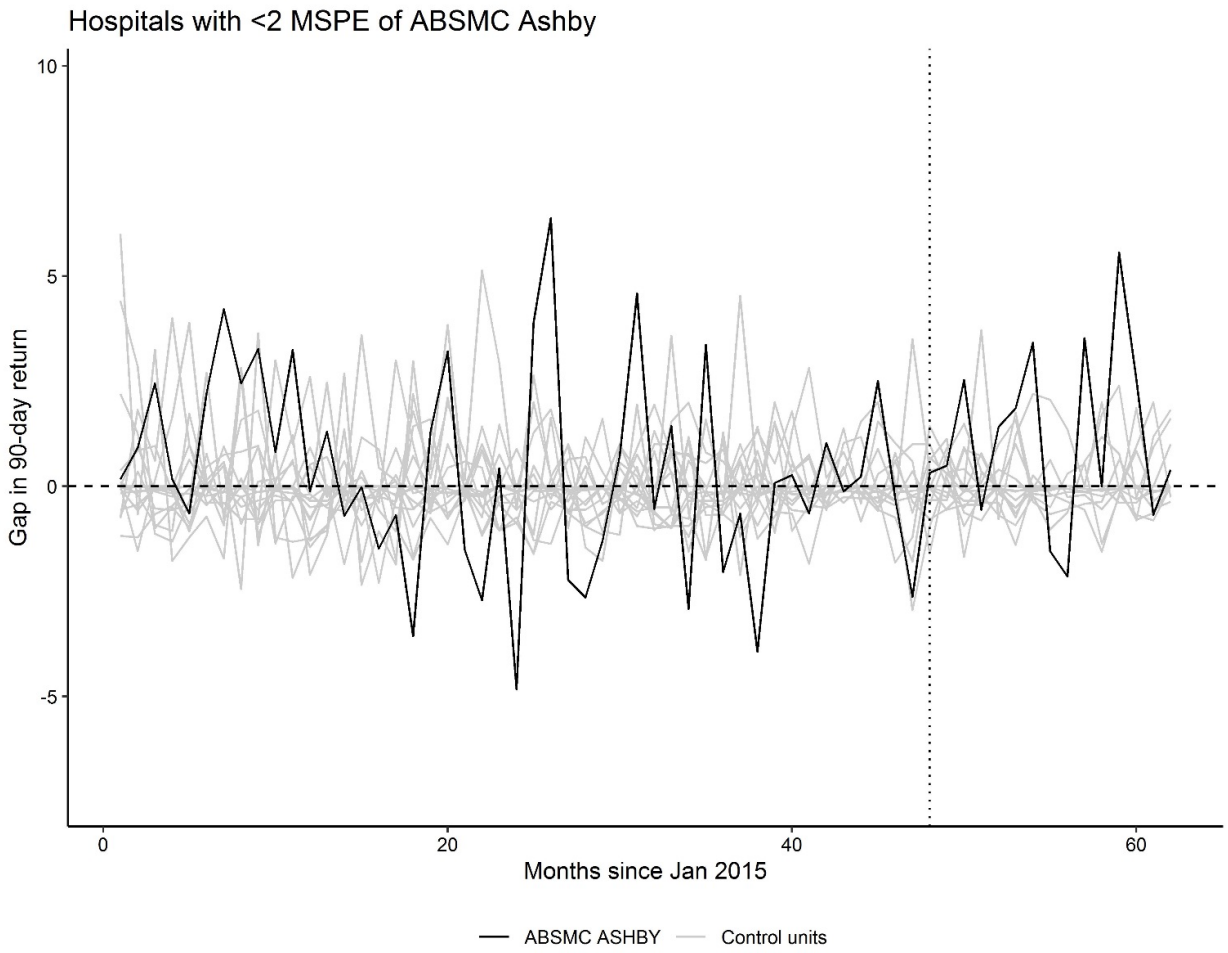


Figure 2.20. Figure 18. Gap in 90-day returns between control hospitals and ABSMC Ashby and their respective synthetic controls, hospitals with <2 MSPE of ABSMC Ashby

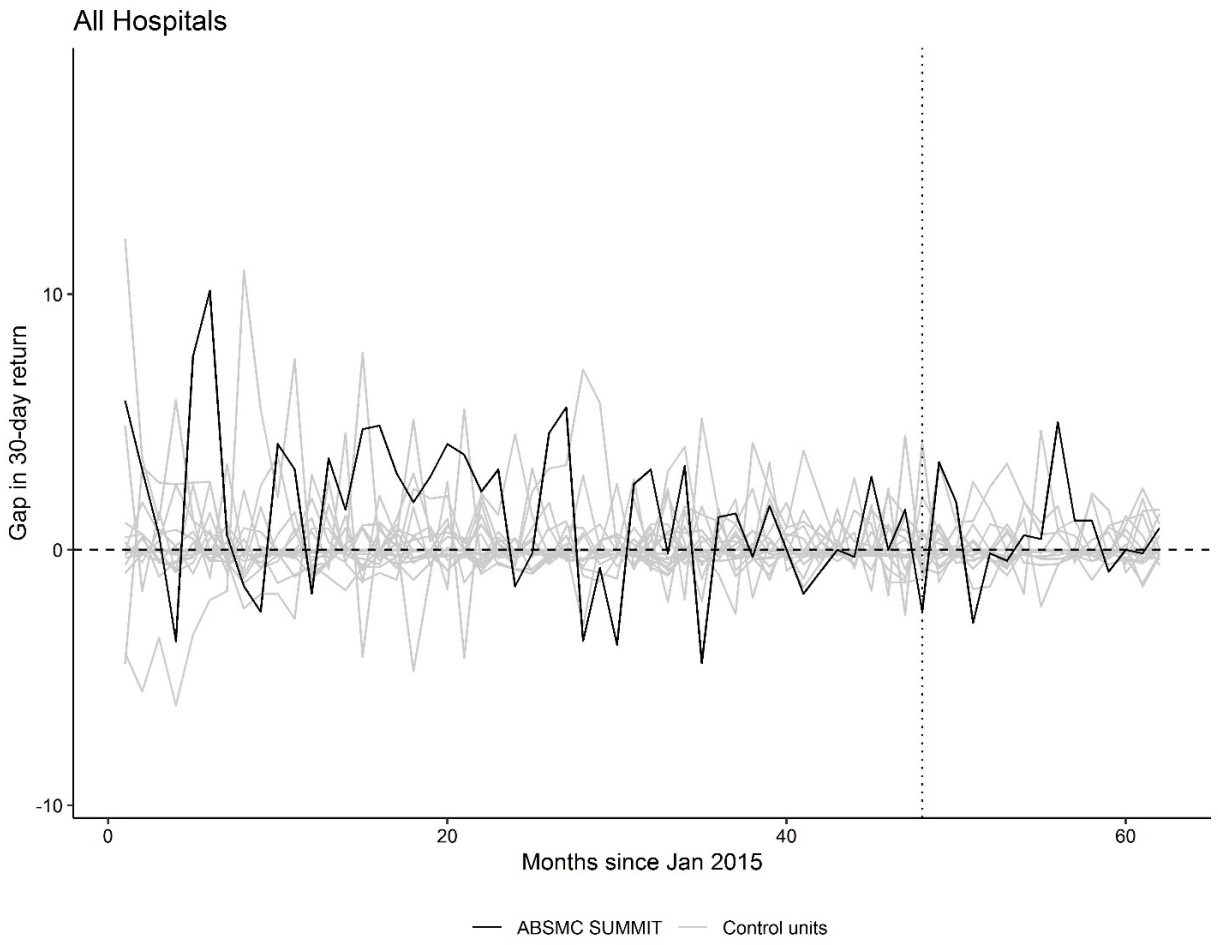


Figure 2.21. Gap in 30-day returns between control hospitals and ABSMC Summit and their respective synthetic controls, all hospitals.

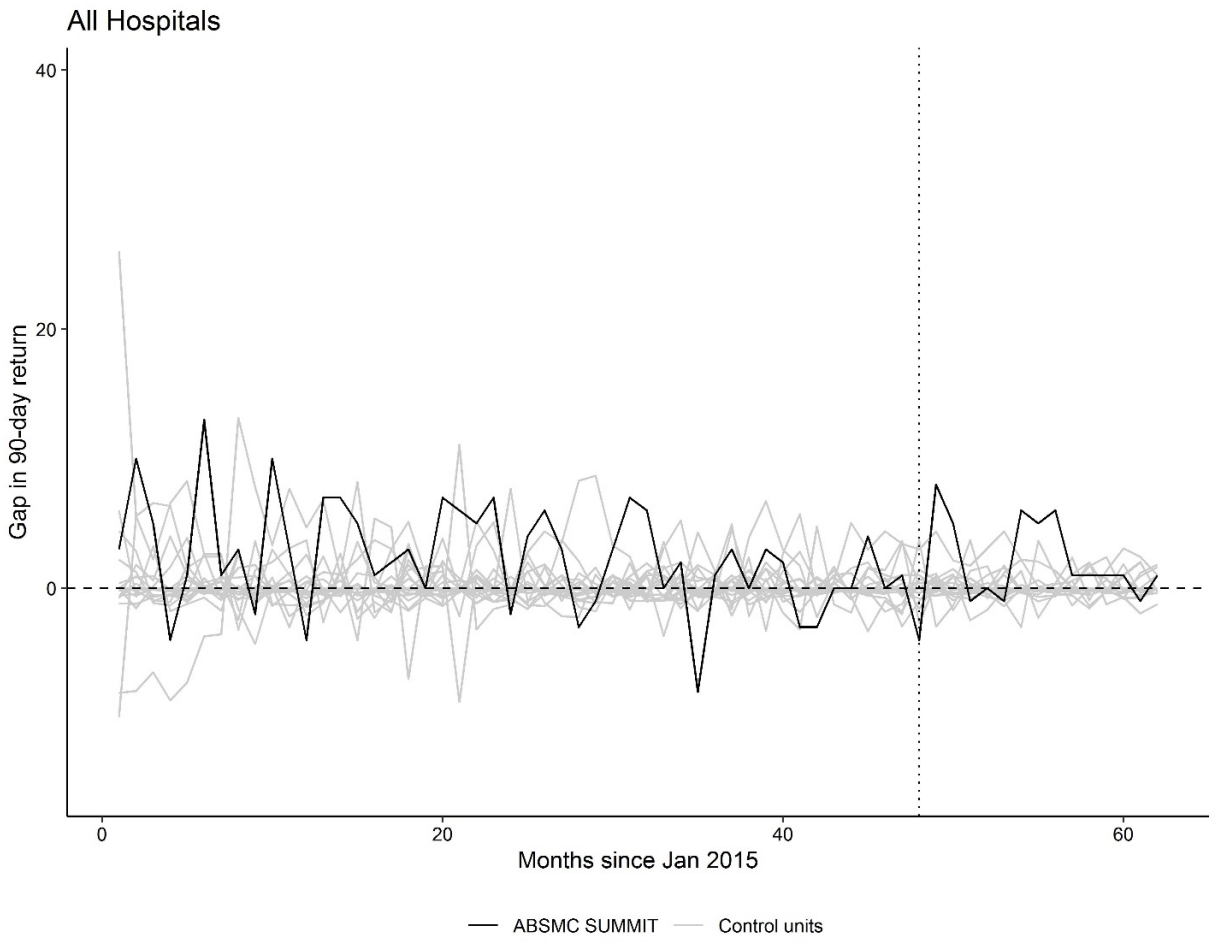


Figure 2.22. Gap in 90-day returns between control hospitals and ABSMC Summit and their respective synthetic controls, all hospitals

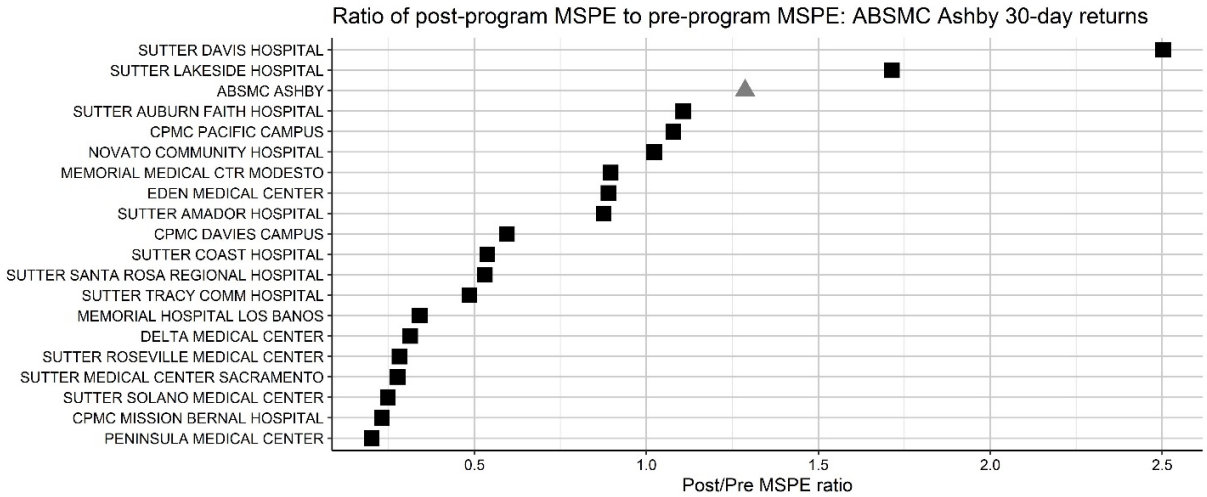


Figure 2.23. Ratios of post-program MSPE and pre-program MSPE, ABSMC Ashby vs control hospitals, 30-day returns.

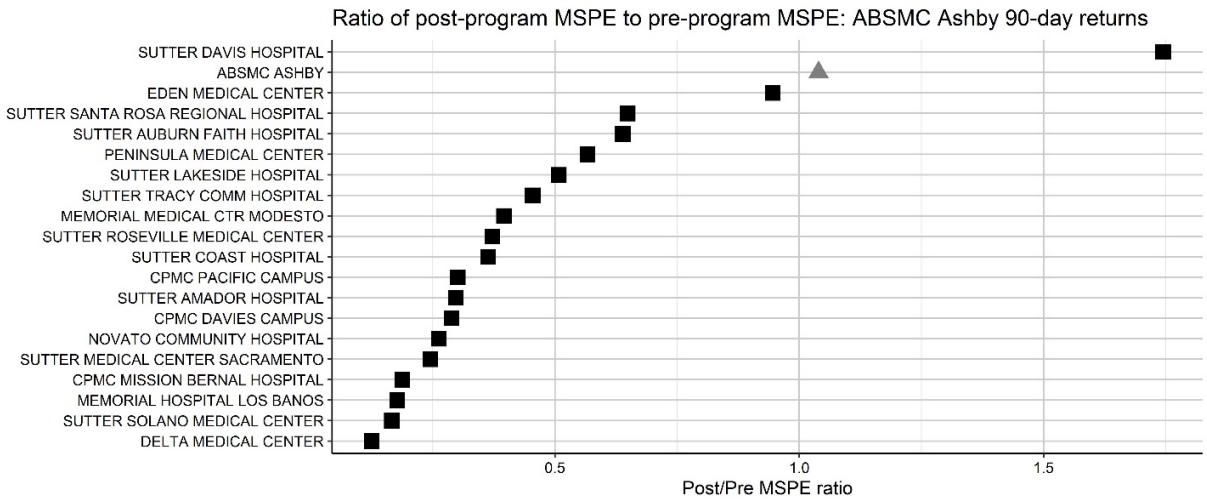


Figure 2.24. Ratios of post-program MSPE and pre-program MSPE, ABSMC Ashby vs control hospitals, 90-day returns.

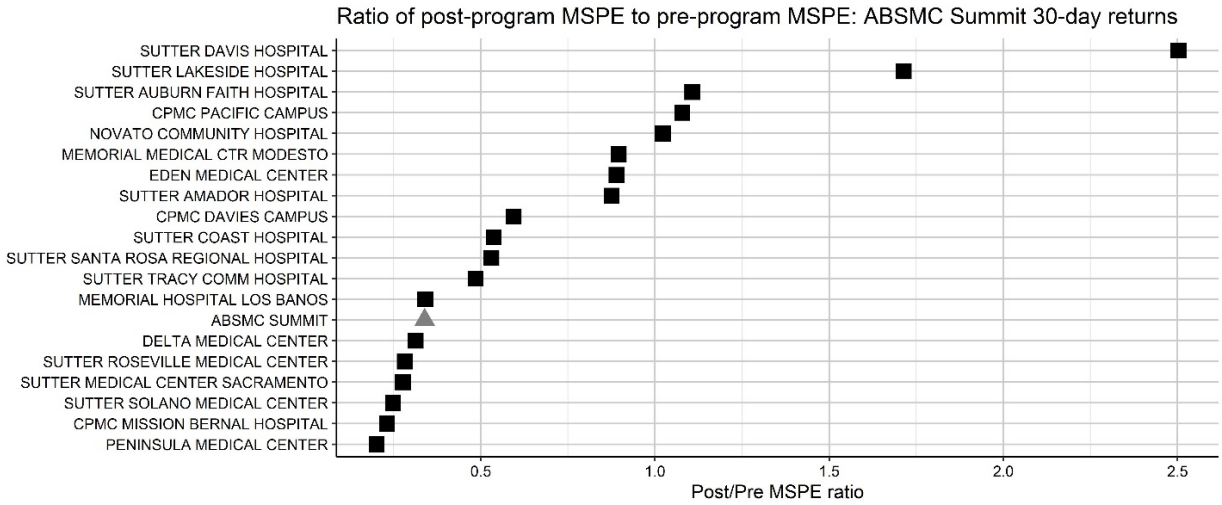


Figure 2.25. Ratios of post-program MSPE and pre-program MSPE, ABSMC Summit vs control hospitals, 30-day returns.

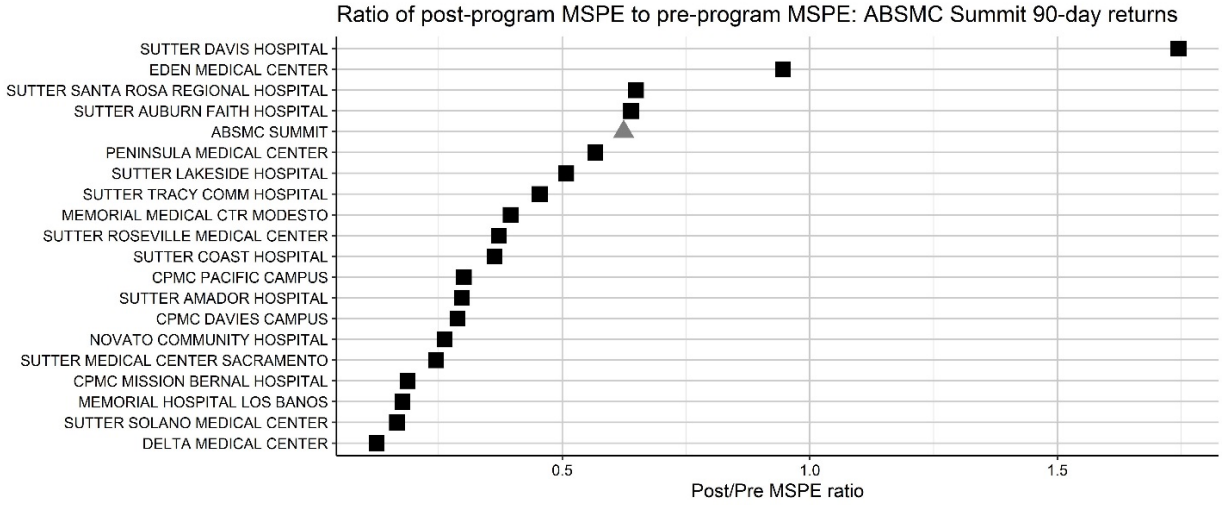


Figure 2.26. Ratios of post-program MSPE and pre-program MSPE, ABSMC Summit vs. control hospitals, 90-day returns.

2.7. Tables

Table 2.1. Baseline covariates for Asthma program participants and control patients.

	Unmatched Cohort			Matched Cohort		
	Program Participants	**Control Patients**		**Program Participants**	**Control Patients**	
	N = 373	**N = 2093**	**Standardized Mean Difference (SMD)**	**N = 372**	**N = 1383**	**Standardized Mean Difference (SMD)**
Demographics						
Mean Age, Years ± SD	46.08 ± 16.31	40.54 ± 15.46	0.349*	45.97 ± 16.19	44.11 (15.23)	0.118*
Female (%)	239 (64.1%)	1235 (59%)	0.104	239 (64.2%)	859 (62.1%)	0.044
History of COPD (%)	76 (20.4)	308 (14.7)	0.149*	75 (20.2)	246 (17.8)	0.061
History of respiratory disease (%)	270 (72.4)	1379 (65.9)	0.141*	269 (72.3)	953 (68.9)	0.075
Insurance type			0.226*			0.075
HMO, n (%)	45 (12.1)	369 (17.6)		45 (12.1)	151 (10.9)	
Medicaid/Medi-Cal	219 (58.7)	1055 (50.4)		219 (58.7)	820 (59.3)	
Medicare FFS	76 (20.4)	402 (19.2)		76 (20.4)	274 (19.8)	
Other, PPO/FFS, Self, Unknown	33 (8.8)	267 (12.8)		33 (8.8)	138 (10.0)	
*P<0.05 (derived from t tests or Chi-square tests of independence)						

Table 2.2. Predictors prior to the implementation of the asthma program, 30-day returns, ABSCM Ashby.

Predictors	Treated	Synthetic ABSMC Ashby	Sample Mean
Age (mean)	52.324	52.535	54.519
Hispanic	0.062	0.120	0.074
Non-hispanic	0.300	0.277	0.239
Unknown ethnicity	0.013	0.013	0.013
Asian	0.053	0.055	0.032
Black	0.074	0.064	0.026
Multirace	0.004	0.005	0.004
Unknown race	0.073	0.092	0.064
White	0.168	0.190	0.197
Male	0.132	0.185	0.144
Female	0.243	0.224	0.182
Other insurance	0.014	0.016	0.015
Public insurance	0.093	0.114	0.084
Medicare	0.079	0.096	0.09
Private insurance	0.170	0.136	0.108
Uninsured	0.001	0.001	0.001
COPD	0.001	0.001	0.001
Respiratory history	0.005	0.008	0.028
30-day returns	2.438	2.076	0.827

Table 2.3. Predictors prior to the implementation of the asthma program, 90-day returns, ABSMC Ashby.

Predictors	Treated	Synthetic ABSMC Ashby	Sample Mean
Age (mean)	52.324	53.128	54.519
Hispanic	0.062	0.086	0.074
Non-hispanic	0.300	0.253	0.239
Unknown ethnicity	0.013	0.012	0.013
Asian	0.053	0.039	0.032
Black	0.074	0.060	0.026
Multirace	0.004	0.004	0.004
Unknown race	0.073	0.068	0.064
White	0.168	0.177	0.197
Male	0.132	0.158	0.144
Female	0.243	0.193	0.182
Other insurance	0.014	0.013	0.015
Public insurance	0.093	0.103	0.084
Medicare	0.079	0.086	0.090
Private insurance	0.170	0.103	0.108
Uninsured	0.001	0.001	0.001
COPD	0.001	0.001	0.001
Respiratory history	0.001	0.001	0.028
90-day returns	4.688	4.381	1.498

Table 2.4. Predictors prior to the implementation of the asthma program, 30-day returns, ABSMC Summit

Predictors	Treated	Synthetic ABSMC Summit	Sample Mean
Age (mean)	56.94	51.404	54.519
Hispanic	0.057	0.096	0.074
Non-hispanic	0.373	0.240	0.239
Unknown ethnicity	0.023	0.013	0.013
Asian	0.074	0.025	0.032
Black	0.134	0.067	0.026
Multirace	0.004	0.005	0.004
Unknown race	0.078	0.077	0.064
White	0.160	0.171	0.197
Male	0.201	0.151	0.144
Female	0.253	0.198	0.182
Other insurance	0.017	0.009	0.015
Public insurance	0.121	0.124	0.084
Medicare	0.124	0.075	0.090
Private insurance	0.131	0.085	0.108
Uninsured	0.001	0.001	0.001
COPD	0.003	0.003	0.001
Respiratory history	0.015	0.076	0.028
30-day returns	4.979	3.521	0.827

Table 2.5. Predictors prior to the implementation of the asthma program, 90-day returns, ABSMC Summit.

Predictors	Treated	Synthetic ABSMC Summit	Sample Mean
Age (mean)	56.940	50.579	54.519
Hispanic	0.057	0.101	0.074
Non-hispanic	0.373	0.230	0.239
Unknown ethnicity	0.023	0.013	0.013
Asian	0.074	0.025	0.032
Black	0.134	0.071	0.026
Multirace	0.004	0.004	0.004
Unknown race	0.078	0.078	0.064
White	0.160	0.162	0.197
Male	0.201	0.149	0.144
Female	0.253	0.195	0.182
Other insurance	0.017	0.009	0.015
Public insurance	0.121	0.127	0.084
Medicare	0.124	0.064	0.090
Private insurance	0.131	0.084	0.108
Uninsured	0.001	0.001	0.001
COPD	0.003	0.001	0.001
Respiratory history	0.015	0.010	0.028
90-day returns	8.979	6.687	1.498

Table 2.6. Hospital weights for synthetic ABSMC Ashby in approximating the number of 30-day returns to the ED.

Hospital name	Weights
CCPMC DAVIES CAMPUS	0.225
CPMC MISSION BERNAL HOSPITAL	0.316
CMPC PACIFIC CAMPUS	0.000
DELTA MEDICAL CENTER	0.460
EDEN MEDICAL CENTER	0.000
MEMORIAL HOSPITAL LOS BANOS	0.000
MEMORIAL MEDICAL CTR MODESTO	0.000
NOVATO COMMUNITY HOSPITAL	0.000
PENINSULA MEDICAL CENTER	0.000
SUTTER AMADOR HOSPITAL	0.000
SUTTER AUBURN FAITH HOSPITAL	0.000
SUTTER COAST HOSPITAL	0.000
SUTTER DAVIS HOSPITAL	0.000
SUTTER LAKESIDE HOSPITAL	0.000
SUTTER MEDICAL CENTER SACRAMENTO	0.000
SUTTER ROSEVILLE MEDICAL CENTER	0.000
SUTTER SANTA ROSA REGIONAL HOSPITAL	0.000
SUTTER SOLANO MEDICAL CENTER	0.000
SUTTER TRACY COMM HOSPITAL	0.000

Table 2.7. Hospital weights for synthetic ABSMC Ashby in approximating the number of 90-day returns to the ED.

Hospital name	Weights
CCPMC DAVIES CAMPUS	0.168
CPMC MISSION BERNAL HOSPITAL	0.000
CMPC PACIFIC CAMPUS	0.000
DELTA MEDICAL CENTER	0.485
EDEN MEDICAL CENTER	0.287
MEMORIAL HOSPITAL LOS BANOS	0.00
MEMORIAL MEDICAL CTR MODESTO	0.000
NOVATO COMMUNITY HOSPITAL	0.000
PENINSULA MEDICAL CENTER	0.000
SUTTER AMADOR HOSPITAL	0.000
SUTTER AUBURN FAITH HOSPITAL	0.000
SUTTER COAST HOSPITAL	0.000
SUTTER DAVIS HOSPITAL	0.000
SUTTER LAKESIDE HOSPITAL	0.000
SUTTER MEDICAL CENTER SACRAMENTO	0.033
SUTTER ROSEVILLE MEDICAL CENTER	0.000
SUTTER SANTA ROSA REGIONAL HOSPITAL	0.000
SUTTER SOLANO MEDICAL CENTER	0.027
SUTTER TRACY COMM HOSPITAL	0.000

Table 2.8. Hospital weights for synthetic ABSMC Summit in approximating the number of 30-day returns to the ED.

Hospital name	Weights
CCPMC DAVIES CAMPUS	0.000
CPMC MISSION BERNAL HOSPITAL	0.000
CMPC PACIFIC CAMPUS	0.000
DELTA MEDICAL CENTER	0.858
EDEN MEDICAL CENTER	0.000
MEMORIAL HOSPITAL LOS BANOS	0.000
MEMORIAL MEDICAL CTR MODESTO	0.000
NOVATO COMMUNITY HOSPITAL	0.000
PENINSULA MEDICAL CENTER	0.000
SUTTER AMADOR HOSPITAL	0.000
SUTTER AUBURN FAITH HOSPITAL	0.000
SUTTER COAST HOSPITAL	0.000
SUTTER DAVIS HOSPITAL	0.000
SUTTER LAKESIDE HOSPITAL	0.000
SUTTER MEDICAL CENTER SACRAMENTO	0.142
SUTTER ROSEVILLE MEDICAL CENTER	0.000
SUTTER SANTA ROSA REGIONAL HOSPITAL	0.000
SUTTER SOLANO MEDICAL CENTER	0.000
SUTTER TRACY COMM HOSPITAL	0.000

Table 2.9. Hospital weights for synthetic ABSMC Summit in approximating the number of 90-day returns to the ED.

Hospital name	Weights
CCPMC DAVIES CAMPUS	0.000
CPMC MISSION BERNAL HOSPITAL	0.000
CMPC PACIFIC CAMPUS	0.000
DELTA MEDICAL CENTER	1.000
EDEN MEDICAL CENTER	0.000
MEMORIAL HOSPITAL LOS BANOS	0.000
MEMORIAL MEDICAL CTR MODESTO	0.000
NOVATO COMMUNITY HOSPITAL	0.000
PENINSULA MEDICAL CENTER	0.000
SUTTER AMADOR HOSPITAL	0.000
SUTTER AUBURN FAITH HOSPITAL	0.000
SUTTER COAST HOSPITAL	0.000
SUTTER DAVIS HOSPITAL	0.000
SUTTER LAKESIDE HOSPITAL	0.000
SUTTER MEDICAL CENTER SACRAMENTO	0.000
SUTTER ROSEVILLE MEDICAL CENTER	0.000
SUTTER SANTA ROSA REGIONAL HOSPITAL	0.000
SUTTER SOLANO MEDICAL CENTER	0.000
SUTTER TRACY COMM HOSPITAL	0.000

Table 2.10. Mean squared prediction error (MSPE) in the pre-program period.

Outcome	ABSMC Ashby	ABSMC Summit
30-day returns	2.9	11.2
90-day returns	5.6	22.2

Chapter 3 - A Propensity Score and Synthetic Control Approach to Evaluating a New Protocol for Patients at Risk for Alcohol Withdrawal Syndrome

Abstract

Purpose: To evaluate the impact of Eden Medical Center's (EMC) new alcohol (ETOH-P) withdrawal syndrome treatment protocol on risk of patient's leaving the hospital against medical advice (AMA), intensive care unit (ICU) admissions, and hospital length-of-stay (LOS).

Methods: We utilized EHR data from Sutter Health's Epic system to conduct a retrospective secondary analysis of the impact of the ETOH-P treatment protocol implemented at EMC beginning in April 2019. Using propensity score matched methods and synthetic controls, we assess the impact of the program on risk of patient's leaving the hospital AMA, ICU admissions, and hospital LOS at both the patient level and hospital level, respectively. We matched patients at EMC to patients who presented to other Sutter Health EDs based on relevant characteristics captured in a propensity score. We constructed a synthetic EMC using a combination of other Sutter Health EDs with similar characteristics during the years prior to April 2019 and performed a series of placebo tests to compare the treatment effect to the distribution of placebo effects.

Results
A total of 310 patients from EMC received the ETOH-P protocol from April 1, 2019 to March 2, 2020. Of the 310, 76 patients received only the active withdrawal order set, 159 patients received only the prophylaxis order set, and 67 patients received both order sets. 1229 control patients were matched to the ETOH-P protocol patients. Patients who received the ETOH-P protocol had 1.47 (0.96, 2.27) times the odds of leaving AMA and 1.00 (0.77, 1.30) times the odds of ICU admissions compared to those who did not receive the protocol. Neither of these estimates were statistically significant. Patients on the ETOH-P protocol had, on average, 0.19 (0.11, 0.26) more days in the hospital, compared to those not on the ETOH-P protocol, and this difference was statistically significant. We were not able to construct a synthetic EMC that closely mimicked the outcome trends of the actual EMC during the pre-intervention period. During the post-intervention period, the trends at EMC for all three outcomes were generally higher compared to the synthetic version, showing the opposite effect than we had originally hypothesized.

Conclusions

For both the propensity score matched analysis and the synthetic control analysis, we did not find evidence that the ETOH-P program led to fewer patients leaving AMA, fewer ICU admissions, or shortened length of stay. Future assessment of the ETOH-P program should focus on additional outcomes such as severity of AWS, and factor in additional information such as patient comorbidities and the number of diagnoses.

Abbreviations

EMC - Eden Medical Center
ED - Emergency department
AMA - Against Medical Advice
LOS - Length of stay
ICU - Intensive Care Unit
ETOH-P – Ethanol: in the context of this study, it refers to a new alcohol withdrawal syndrome prophylaxis protocol implemented at Eden Medical Center
AWS – Alcohol Withdrawal Syndrome
CIWA - Clinical Institute Withdrawal Assessment for Alcohol
SCM – Synthetic Control Method
PSM – Propensity Score Matching

3.1. Background

Alcohol abuse is a widespread societal and economic burden in the United States, resulting in 85,000 deaths and 200 billion dollars in spending annually[45]. On a larger scale, alcohol contributes substantially to the global burden of disease with 4-6% being attributable to alcohol and has been identified as a major risk factor for chronic disease and injury. In 2004, 3.8% of all global deaths were attributable to alcohol, 6.3% for men and 1.1% for women[46]. Unsurprisingly, alcohol-use disorders follow the same pattern as alcohol-attributable harm, with men having more disorders than women. Alcohol use disorders offer a major burden on American society, with one in four children younger than 18 years having some exposure to family alcohol problems[47].

In 2006, the estimated economic cost of excessive drinking was $223.5 billion, approximately $746 per person[48]. Excessive alcohol consumption also causes direct consequences such as premature death, increased health care costs, as well as indirect costs associated with property damage from fire and motor vehicle accidents, increased crime and criminal justice system costs, and lost worker productivity in the form of missed work, diminished output and reduced earnings potential[46,48].

The burden of alcohol-related problems on hospitals and health care systems is tremendous, both in terms of ED and out-patient services. In a prospective study

conducted at an inner-city hospital in north-west England, they found that overall 6.2% of all hospital admissions were due to alcohol-related problems, and that over 2,800 new out-patient visits were likely to have been generated over an 18-month period due to an initial alcohol-related visit[49]. Between 2006 and 2014, the number of ED visits involving alcohol consumption increased 61.6% in the United States[50]. The largest increase in rates for acute alcohol-related visits to the ED occurred in age groups 45 to 54 and 55 to 64, which is consistent with the age group that experienced the highest increase in alcohol-induced liver cirrhosis deaths (37%)[50]. EDs present a pivotal opportunity to detect and refer patients who misuse alcohol, as it is an appropriate place to offer patients initial health and education on their drinking habits. In a cost-benefit analysis assessing the impact of alcohol intervention programs in trauma centers on health care costs, the authors found that 27% of adult patients presenting to the ED were candidates for a brief alcohol intervention, and that the net cost of savings due to the intervention was $89 per patient screened, and $330 for each patient offered an intervention[51].

It has been well established that abrupt reduction, or total cessation of chronic alcohol use leads to serious complications. Diagnosis of alcohol withdrawal is defined by the Diagnostic and Statistical Manual of mental Disorders (DSM-5). In most cases, the symptoms of alcohol withdrawal are mild and do not require medical intervention, disappearing within 2-7 days after the last drink, but in more severe cases, medical intervention may be needed. Alcohol withdrawal syndrome (AWS) is a set of signs and symptoms that vary between mild to severe, typically developing in alcohol-dependent people within 6-24 hours of their last drink[52]. The symptomology of AWS can be divided into three categories of symptoms. The first consists of autonomic hyperactivity in the form of tremulousness, sweating, nausea, vomiting, anxiety and agitation, which appears within hours of the last drink and peaks within 24-48 hours. The second consists of neuronal excitation symptoms, which includes epileptiform seizures occurring within 12-48 hours of abstinence. The final set of symptoms consists of delirium tremens, which develops in a few patients (3-5%), characterized by auditory and visual hallucinations, confusion, disorientation, clouding of consciousness, impaired attention, and pronounced autonomic hyperactivity. In the most severe cases, death may occur from respiratory and cardiovascular collapse.

The pathophysiology of AWS is complex as chronic alcohol use affects various neurotransmitter system responses within the brain. Ɣ-aminobutyric acid (GABA), the major inhibitory neurotransmitter in the CNS and its receptor are downregulated during chronic alcohol use such that upon abstaining, this downregulation continues, leading to many of the AWS symptoms. Prolonged alcohol use also inhibits glutamate neurotransmitter activity, the major excitatory neurotransmitter in the CNS, and alters the conformational form of N-methyl-aspartate (NMDA) receptors[53]. Abrupt cessation of alcohol exposure results in brain hyper excitability, as receptors previously being inhibited by alcohol are no longer inhibited.

Medical history and physical examination are key in diagnosing the severity of alcohol withdrawal. Standardized clinical rating scales have been developed to help monitor the

severity of AWS. The Clinical Institute Withdrawal Assessment for Alcohol (revised) (CIWA-Ar) is a validated and reliable ten-item rating scale of patients' symptoms that may be used to guide clinical decision making[54]. CIWA-Ar scores of 8 points or lower correspond to mild withdrawal, while scores of 9 to 15 indicate moderate withdrawal, and scores greater than 15 points indicate severe withdrawal. The standardized treatment management for patients with suspected alcohol withdrawal is initial resuscitation and rehydration, and administration of a GABA agonist[53]. Benzodiazepines offer one treatment option by increasing the frequency of GABA-receptor channel opening and preventing progression to more serious withdrawal symptoms. There is no clear consensus on the best benzodiazepine treatment, but several are available for treatment including diazepam, lorazepam, midazolam, oxazepam, and chlordiazepoxide.

Current practice varies greatly both within the Sutter Health system and the overall medical community. While some hospitals utilize the CIWA assessment method and strictly Benzodiazepines for treatment, others have implemented protocols where providers prescribe both Benzodiazepines with additional medications to address hyper-adrenergic and delirium symptoms. Historically, the protocol at EMC has been to assess patients using the CIWA score and treat symptoms of AWS with benzodiazepines. While the CIWA tool has been shown to have high validity and inter-rater reliability, it requires the clinical team to wait for the patient to show signs and symptoms of withdrawal prior to starting treatment. However, as recent studies have shown that delirium increases hospital acquired conditions, length of stay, cost of care, death and long-term cognitive impairments[55], the question of why treatment can't be initiated sooner to prevent AWS symptoms before they begin is pertinent. Thus, in April of 2019, EMC implemented a new treatment protocol (ETOH-P) for patients at risk of developing AWS.

To thoroughly evaluate the impact Eden Medical Center's (EMC) ETOH-P protocol program, we utilized two different approaches. First, we conducted an analysis to determine the effect of the new protocol at the patient level in which we compared EMC participants to control patients who did not receive the ETOH-P protocol using a propensity score matched analysis. Second, we considered its effect at the hospital system level using a synthetic control method in which we compared EMC to other Sutter Health Eds that did not implement the ETOH-P protocol. Specifically, we examined the risk of leaving AMA, average hospital length of stay, and risk of ICU admission. The present analysis represents an application of the SCM method to an inpatient hospitalization protocol, where a patient receives the protocol if they present at the ED with indications of heading into alcohol withdrawal. We hypothesize that the implementation of the ETOH-P protocol within EMC will show an overall increase in the use of this multimodal approach to AWS, resulting in improved overall care for this patient population via fewer patients leaving the hospital AMA, fewer patients admitted to the ICU and shorter hospital stays.

3.2. Methods

Study Design and Setting

This is a retrospective EHR-based observational study conducted at Sutter Health, a large, private, and not-for-profit community-based health care delivery system in Northern California that provides medical services across 130 ambulatory clinics and 24 acute care hospitals, including 22 ED sites (**Appendix A3**). All Sutter Health clinics and hospitals are linked by a single electronic health record system (Epic, Verona, WI). Sutter Health has approximately 11 million ambulatory visits, 870,000 ED visits and 200,000 hospital discharges annually. This study was approved by the UC Berkeley Committee for the Protection of Human Subjects (CPHS) and the Sutter Health Institutional Review Board (IRB) with a Health Insurance Portability and Accountability Act (HIPAA) waiver of authorization and informed consent.

ETOH-P Program

The ETOH-P program, designed to reduce the number of patients who progressed to AWS, was implemented at EMC beginning April 2019. The program consists of two protocols, one aimed at preventing AWS (prophylaxis protocol) and the other at treating AWS earlier in its course with targeted pharmaceuticals (active withdrawal protocol). Upon admission to EMC, patients are identified as being at mild, moderate or severe risk of developing AWS based on their history of developing AWS. Patients are placed on either the Mild Risk Prophylaxis Protocol, the Moderate Risk Prophylaxis Protocol, or the Severe Risk Prophylaxis Protocol based on their history of withdrawals. If no withdrawal history is obtainable and the patient is known to drink more than three drinks per day, they are placed on the Mild Risk Prophylaxis Protocol. For each protocol, the ordering provider has the option of choosing between two benzodiazepine tapered dose regimens. Each patient will also be started on a course of supportive medications that include thiamine, folic acid and multivitamins. Vital signs and the Richmond Agitation Sedation Scale (RASS) are monitored continuously.

If active withdrawal develops (either present at the time of admission or otherwise), the patient is treated with a symptom-based Active Withdrawal Protocol that treats each symptom with an appropriate corresponding pharmacological medication. Patients are also classified into Mild Active Withdrawal, Moderate Active Withdrawal or Severe Active Withdrawal based on their symptoms. Benzodiazepines, in combination with medications for hyperadrenergic activity and delirium are used if those symptoms are also present. Patients are also started on a course of supportive medications that includes thiamine, folic acid and multivitamins. Vital signs and the RASS are monitored routinely.

Cohort identification

EMC patients who received the ETOH-P protocol were identified from the EHR if they received either the prophylaxis order set, or the active withdrawal order set. From the EHR, we retrospectively identified “control” patients who did not receive the protocol from a pool of eligible patients from all other Sutter Health hospitals. Patients were included in the study if they were

1. at least 18 years of age upon admission to the hospital
2. a Sutter in-patient during the time period of January 1, 2018 to March 4, 2020
3. had relevant ICD 9 or 10 codes

We excluded patients who had expired, visited historical/merged hospitals, had an acute neurological diagnosis, and those with a past medical history of dementia. For all patients, we extracted demographic information and information on primary insurance and hospitalization (admission date, discharge date, discharge diagnosis).

Covariates

Covariates included in the propensity score model were based on the available data, as well as a priori specification of variables that are thought to influence the probability of receiving the ETOH-P protocol. These variables included age, sex, ethnicity, race, median household income (MHI), and medical insurance type.

We obtained MHI by zip code for each California county from the U.S Census Bureau's publicly available data and assigned MHI based on the patient home address zip code. For the synthetic control analysis, we extracted and aggregated at the hospital level age, sex, race, ethnicity, and health insurance type from January 1, 2018 to March 4, 2020.

Outcome measures

We had three primary outcomes of interest: (1) the proportion of patients who left the hospital against medical advice (AMA), (2) average hospital length of stay (LOS), and (3) the proportion of patients who had an intensive care unit (ICU) admission. We created a dichotomous indicator of leaving AMA if the patient had a discharge status of "Left Against Medical Advice or Discontinued Care". For hospital length of stay, we took the maximum number of days in the hospital for each patient. Lastly, we created an indicator of ICU visit, if the patient ever had an ICU visit in any of their visits.

For the hospital-level synthetic control analysis, we aggregated outcomes to the hospital level.

Statistical Analyses

Propensity score matching

Observational studies are increasingly being used to estimate the causal effects of intervention. As it is not possible to randomly assign the ETOH-P program intervention to patients, treatment selection will be influenced by subject characteristics. Thus, baseline characteristics between treated and untreated subjects may differ systematically and needs to be taken into account when estimating the effect of treatment on outcomes. Thus, we will utilize propensity score methods to correct for the treatment-selection bias imposed when estimating effects using observational data.

The propensity score[29] is the probability of receiving treatment conditional on observed baseline covariates. It is essentially a balancing score such that, conditional on the propensity score, the distribution of baseline covariates is similar between treated and untreated subjects[30,31]. In randomized controlled trials, the true propensity score is known. However, in observational studies, it is not known, but can be estimated using data. Propensity score matching, in which treated and untreated subjects are matched together based on a similar propensity score value, allows for estimation of the average treatment effect among the treated (ATT):

$$E[Y(1) - Y(0)|Z = 1]$$

where Y(1) is the potential outcome under active treatment and Y(0) is the potential outcome under control treatment. The ATT is defined as the average effect of the treatment on those who ultimately received the treatment.

Once a matched sample has been created, the treatment effect can be estimated by comparing the outcomes between treated and untreated subjects in the matched sample. However, since observations within a propensity score matched sample are no longer independent, as treated and untreated subjects have similar values of the propensity score, this lack of independent should be taken into account[30–32].

A propensity score model was used to predict receipt of the ETOH-P program at intake, conditional on baseline characteristics shown in **Table 3.1**. We used the *MatchIt* package[33,56] to match EMC patients to control patients 1 to 4, with replacement, using nearest neighbor matching with calipers of width equal to 0.2 of the standard deviation of the logit of the estimated propensity score. We chose to match more than one control for each treated unit in order to increase the power of the procedure[34].

We checked covariate balance in the sample before and after matching by calculating the standardized mean difference (SMD)[32,35] and via plotting procedures (figures 3.1-3.4). The standardized mean difference is not confounded by sample size and thus allows for appropriate comparison between the balance in the original sample with the matched sample[32]. We considered an SMD greater than 0.1 as the threshold for imbalance[35,57]. The propensity score model for this study included all the variables that were imbalanced[36]. To avoid issues with multicollinearity, we created broader insurance group types of Medi-Cal, Medicare and Other/Unknown insurance (which includes self-pay and worker's compensation).

We created a propensity score matched sample by matching AWS protocol patients with controls that had a similar propensity score, such that the treated and control patients within the same matched pair have a similar propensity score. We estimated the effect of the ETOH-P protocol delivery method by directly estimating the difference in outcomes between the treatment group and controls in the propensity score matched sample.

Because these patients within the same matched set had baseline covariates that are, on average, more similar than two randomly selected treated and control patients, we could no longer consider them independent units[32]. Thus, we calculated Mantel Haenszel adjusted odds ratios for our binary outcomes, stratifying by the matched set, to account for the matched nature of the propensity score matched sample. We used McNemar's test for correlated binary proportions to assess statistical significance[30–32,58]. We chose this method, as opposed conditional logistic regression that is often used for matched-pairs data, as conditional logistic regression has been shown to result in biased estimation of odds ratios[30–32,58]. Accounting for the matched nature of the sample results in tests with appropriate type I error rates and confidence intervals because estimates of the standard error of the treatment effect are closer to the standard deviation of the sampling distribution of the treatment effect[58]. For count outcomes such as average hospital LOS, we calculated a difference in counts using conditional permutation tests within matched sets, each time calculating a paired t-test statistic.

Data extraction and statistical analyses were performed using R statistical software package v3.6.3 and Microsoft SQL Server Management Studio 13.0.15700.28.

Synthetic control method

While the program is administered at the individual patient level, the health equity leadership committee at Sutter Health are also interested in the hospital-level effects of the program, and thus we will also perform a synthetic control analysis to assess any hospital-level changes resulting from the program.

SCM is an approach to program evaluation in which one or a small number of units are subject to intervention, and a comparative control unit is constructed such that the outcomes of the control units are weighted to construct the counterfactual outcome of the treated unit(s) in the absence of the treatment. Statistical details on the method can be found in the appendix. This method has previously been used to analyze political and economic effects following large-scale events, state-level policy changes, health systems reforms, nutritional interventions, climate changes, and even the current COVID-19 pandemic[2,3,6,8,17,40,59,60] where it is difficult to find a single comparison unit that best approximates the relevant characteristics of the treated unit; indeed, a combination of units often provides a better comparison unit than any single unit alone.

SCM offers another tool for program evaluation, in which time series for the unit of interest in the period before the intervention are used to make predictions about what future trends would look like without the intervention. No extrapolation is required as weights are restricted to be non-negative and sum to 1, and the weights are calculated and chose without seeing the post intervention data, reducing the risk of cherry picking or p-hacking. The contribution of each control unit to the counterfactual is made explicit and offers transparency in the selection of the best counterfactual[18]. Additionally, SCM provides a visual representation of the impact of the intervention and how it varies over time, as well as a clear visualization of the actual discrepancy between the treated unit and the convex combination of untreated units, something that propensity score methods do not provide. Plots are produced that display what the observed outcome looks like compared to what would be expected in the absence of the intervention.

Lastly, predictions from SCM may serve as input to the propensity score model, by acting as a baseline counterfactual.

If the program proves to be successful at EMC, we are interested in spreading this program to other Sutter Health hospitals, therefore, we considered the effect of the new ETOH-P protocol at the hospital level. As the ETOH-P protocol was implemented beginning in April of 2019, we define the period from January 1, 2018 to April 1, 2019 as the pre-intervention period. We define April 1, 2019 to March 4, 2020 as the post-intervention period. We utilize the synthetic control method, a data-driven approach that creates a weight combination of control hospitals in order to create a "synthetic" EMC, providing an estimate of the expected outcome at EMC, had the ETOH-P protocol not been implemented. The hospitals that comprise the "synthetic" EMC are selected by the method based on their pre-program trends in both covariate and outcome values. The combination of control hospitals that are able to best match the pre-intervention outcome trends for EMC are chosen to be included in the synthetic control. The expected outcome trend for EMC in the post-intervention period in the absence of the ETOH-P protocol are then compared against the observed outcome trends. We estimate treatment effect of interest as the difference between the observed and expected values.

We will utilize the *Synth* package[41] to choose the vector of weights by minimizing

$$\|X_1 - X_0 W\| v = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

Where V is a matrix allowing for different weights to be applied to variables in $\mathbf{X_0}$ and $\mathbf{X_1}$ based on their predictive power on the outcome. The *synth()* function[61] also allows for a data-driven procedure to choosing V*, such that the mean-squared prediction error (MSPE) of the outcome is minimized over the pre-program years[2,17,37]. Thus, the mean squared prediction errors (MSPE) is used as a measure of fit between the treated unit and it's synthetic control during the pre-intervention period.

Traditional large sample inferential techniques are not appropriate in this setting due to the small number of units; however, exact inferential techniques, such as a permutation test may be used to assess how unusual an effect would be if it were due to change and thus provide context for the effect size. As advised in prior literature, we obtained inference for these estimates by using a series of placebo (permutation tests), in which we applied the synthetic control method iteratively across all the control hospitals in our sample, each time treating each hospital, as if it we the one that had implemented the ETOH-P protocol and calculating an estimate of the effect of the protocol[42]. We then compared the estimate of effect for EMC to the distribution of effects obtained from the placebo tests.

3.3. Results

Study Cohort Description

A total of 310 patients at EMC received the ETOH-P program from April 1, 2019 to March 2, 2020. Of the 310, 76 patients received only the active withdrawal order set, and 159 patients received only the prophylaxis order set. 67 patients received both order sets. 1229 controls patients were matched to the ETOH-P program patients.

Propensity score matching results

Matching on the propensity score improved covariate balance between the two treatment groups (**Table 3.1**) and reduced the final matched analysis cohort to 1550 patients. After the match, the two groups are more similar in terms of their propensity scores and share a common region of support (**Figure 3.1-3.4**).

The average hospital length of stay among the whole matched cohort, ETOH-P patients and control patients was 7.16 days, 7.32 days and 7.13 days respectively. 10.3% of ETOH-P patients left the hospital AMA compared to 7.3% of control patients who left AMA. 109 ETOH-P patients (35.2%) had an ICU admission, compared to 435 control patients (35.1%).

The results indicate that patients who received the ETOH-P protocol had 1.47 (0.96, 2.27) times the odds of leaving AMA compared to those who did not receive the protocol, however, this was not statistically significant (**Figure 3.8-3.10**). For ICU admissions, patients who received the ETOH-P protocol had 1.00 (0.77, 1.31) times the odds of ICU admissions compared to those who did not receive the protocol, though also not statistically significant. Patients on the ETOH protocol had, on average, 0.19 (0.11, 0.26) more days in the hospital, compared to those not on the ETOH protocol, and this difference was statistically significant.

Synthetic Control results

We show in **Figures 3.5-3.7** that plotting the outcomes of EMC compared to the 19 other control hospitals illustrates that a simple average of the control units does not closely approximate the outcomes at EMC. Thus, we sought to explore if the construction of a synthetic control would provide a better “counterfactual” to the EMC trend. Two hospitals, CPMC Mission Bernal Hospital and CPMC Van Ness Campus did not have any outcome data in the pre-program period, and thus were dropped from the potential control pool.

A total of 19 other Sutter Health hospitals comprised the potential pool of donor hospital. There was insufficient covariate data from two locations, CPMC Mission Bernal Hospital and CPMC Van Ness Campus, and thus were dropped from the analysis.

Table 3.2 compares the pre-intervention characteristics of EMC to those of the synthetic EMC and to those of the population weighted average of all 19 hospitals in the donor pool. Overall, this suggests that the synthetic ABSMC provides a much better comparison for EMC than the sample average of all other Sutter hospitals.

Tables 3.3-3.5 displays weights of each control hospital in the synthetic EMC. The weights indicate that leaving AMA trends in EMC prior to the ETOH-P protocol is best reproduced by a combination of ABSMC Ashby, ABSMC Summit, CPMC Davis Campus, Delta Medical Center, Sutter Davis Hospital, Sutter Medical Center Sacramento, and Sutter Tracy Community Hospital. All other states in the donor pool are assigned zero or near zero weights. Trends at EMC for ICU admissions is best represented by a combination of Delta Medical Center, Sutter Coast Hospital, Sutter Davis Hospital, and Sutter Santa Rosa Regional Hospital. Lastly, average hospital LOS at EMC is best approximated by ABSMC Ashby, Delta Medical Center, Peninsula Medical Center, Sutter Coast Hospital, Sutter Lakeside Hospital, Sutter Medical Center Sacramento, Sutter Santa Rosa Regional Hospital, Sutter Solano Medical Center, and Sutter Tracy Community Hospital.

The MSPEs in the pre-intervention period for leaving AMA, average hospital LOS and ICU admissions are listed in **Table 3.6.**

Effect of the ETOH-P protocol

For each outcome, the trends for EMC and its respective synthetic controls for the period of Jan 1, 2018 to March 4, 2020 are visualized in **Figures 3.11 – 3.13**. In general, the number of patients who left AMA and had an ICU admission was very low during the study period. As indicated in **Figures 3.11 – 3.13**, the pre-intervention trends for leaving AMA and ICU admission did not appear to be well approximated by their respective synthetic controls; in fact, the EMC curve crosses at multiple points with the synthetic control curve during the pre-intervention period. It does appear that the synthetic control for ICU admissions provides a better approximation of actual trends (**Figure 3.12**). After the implementation of the ETOH-P program, in general, the outcome trends at EMC remained higher compared to the synthetic EMC for all three outcomes.

Our estimate of the effect of the ETOH-P program on our outcomes of interest was determined by the difference between the actual EMC and its synthetic control version, or the “gap” between the two curves (**Figures 3.15-3.16**). If our hypothesis is correct, we would expect to see a reduction in the proportion of patients who left AMA, the average LOS and the proportion of patients who had an ICU admission, and thus the gap (treated – synthetic) would be a negative number. However, as **Figures 3.15-3.16** indicates, in general the difference between the two series remained positive during the entire study period. The differences between the actual EMC and its synthetic control indicate that there was an average increase in the proportion of patients who leave AMA during the program period of approximately 0.0003. There was an average increase in the proportion of patients admitted to the ICU of 0.0012 and an increase in the average

hospital length of stay of 2.03 days. Thus, our results did not suggest that the ETOH-P protocol led to a reduction in our outcomes of interest at EMC.

Placebo tests

To evaluate the significance of our results, we used placebo tests to determine of our results were driven entirely by chance (how often we would have obtained the results of this magnitude if we had chosen a hospital at random for the study, instead of EMC). The results of the placebo test are displayed in **Figures 3.16 - 3.18**. The gray lines represent the gap associated with each iteration of the test or the difference in outcomes between each hospital in the donor pool and its respective synthetic control. The superimposed black line denotes the gap estimated for EMC, the hospital that actually implemented the ETOH-P program. The placebo tests allowed us to obtain synthetic control estimates for hospitals that did not actually implement the ETOH-P program, and thus we can compare the estimated effect of the ETOH-P program on EMC to the distribution of placebo effects obtained for other hospitals. We considered the effect to be significant if the estimated effect for EMC is unusually large relative to the distribution of placebo effects. As **Figures 3.16 - 3.18** indicate, the estimated gap for EMC does not appear to be unusually large compared to the distribution of gaps for the other hospitals in the donor pool, for any of the outcomes.

The pre-intervention mean squared prediction error (MSPE) for EMC (the average of the squared discrepancies between the outcome of interest in EMC and its synthetic counterpart during the pre-intervention period) was 0.000000045, 0.00000018, and 1.34 for the proportion of patients who left AMA, the proportion of patients who had an ICU admission, and average LOS, respectively.

The control hospital with the worst fit compared to its synthetic in the pre-intervention period is Delta Medical Center for AMA, with an MSPE of 0.00000013. A higher MSPE indicates that the outcome trend at Delta Medical Center was not well reproduced by a combination of all other hospitals. For ICU admissions and average LOS, Sutter Santa Rosa Regional Hospital (0.00000057) and CPMC Davies Campus (23.4) had the highest MSPE, respectively, compared to its synthetic version.

Finally, we evaluated the EMC gap relative to the gaps obtained from the placebo runs by looking at the distribution of the ratios of post to pre-program MSPE. A ratio of the post-program MSPE to the pre-program MSPE measures the quality of the fit of a synthetic control for its treated unit in the post-program period, relative to the quality of the fit in the pre-program period. **Figures 3.19 – 3.21** displays the distribution of post/pre-program ratios of the MSPE for EMC and all 19 control hospitals. The ratio for EMC leaving AMA is about 2.7 and does not clearly stand out from the other ratios; in fact, five control hospitals achieved a larger ratio than EMC. We obtained a p-value of 0.32 and thus, if we were to have assigned the program at random, the probability of obtaining a post/pre MSPE ratio as extreme or more extreme than that of the treated unit is 0.32. The results from the ratio of post to pre-program MSPE and p-value were consistent with the placebo plots, which indicated that the effect of the ETOH-P program on leaving AMA was not significant.

The ratio for EMC average LOS was about 6.5. However, two control hospital achieved a larger ratio than EMC. We obtained a p-value of 0.2 from the placebo tests and thus, the effect of the program on average LOS was not significant. Lastly, the ratio for EMC ICU admissions was about 12.4. This was the largest ratio compared to all other control hospitals. Indeed, the p-value obtained from placebo tests was 0.05.

3.4. Discussion

To assess the impact of the ETOH-P program implemented at EMC to treat patients who present to the ED with AWS, we utilized a propensity score matched group of controls and a synthetic control. The propensity score matched analysis indicated that the ETOH-P program was actually associated with an increase in the proportion of patients who left AMA, the proportion of patients who had ICU admission and longer hospital LOS; while this effect was not statistically significant, it was in the opposite direction than we had originally hypothesized. Based on the synthetic control analysis, we did not find evidence that the ETOH-P program at EMC was associated with a decrease in the proportion of patients who left AMA, the proportion of patients who had an ICU admission, or the average LOS. In fact, we found that outcome trend at EMC for all three outcomes was generally higher than the synthetic control during the post-intervention period. For LOS, this effect was significant (p=0.05).

It is important to note that the credibility of the synthetic control estimator depends heavily on its ability to track the trajectory of the treated unit's outcome during the pre-intervention period. Thus, due to the fact that the synthetic EMC did not appear to adequately match the outcome trend of the real EMC during the pre-intervention period, we cannot make any definitive conclusions from this analysis. Similarly, placebo runs with poor fit prior to the program also do not provide information to measure the relative rarity of estimating a large post-ETOH-P program gap for any given hospital.

The results of our study contradict what has been found in previous studies; when a revised AWS protocol with focused use of benzodiazepines was implemented at a large tertiary academic medical center in adult patients, those who received the focused treatment had a significant decrease in hospital LOS and ICU length of stay[62,63]. However, in a recent study that also utilized routinely prospectively collected EHR data to investigate inpatients undergoing alcohol withdrawal, the authors found that higher maximum AWS scores are associated with increased LOS and in-hospital mortality[64]. In our study, categorization of whether a patient was "treated" or not was based solely on whether or not they visited EMC. It is possible that patients at EMC had higher AWS scores, leading to increased LOS and more severe progression of AWS resulting in ICU admission. Future work should take the severity of symptoms into consideration when evaluating the ETOH-P program.

Our study has several strengths. We utilized a large sample of patients and hospital stays from a diverse pool of patients that Sutter Health services. We used two different evaluation methodologies to study our outcomes of interest at the patient-level and hospital-level. We chose to use the synthetic control method in order to improve our control group selection. As **Figures 3.1-3.3** showed, a simple average of all 19 control hospitals did not closely mimic the outcome trends at EMC, thus we attempted to construct a synthetic control group that might better approximates these trends. Another central motivation for using the synthetic control method was that it provides researchers with a quantitative tool to select appropriate comparison groups. In our analysis, a handful of control hospitals emerged as potential comparisons to EMC, which we displayed in **Tables 3.5-3.7**, making explicit the contribution of each comparison unit to the counterfactual of interest. Traditional regression analysis fails to provide such a transparent list, as typically all units contribute to the regression fit.

We also recognize several limitations to our study. First, we had limited ability to replicate all of the relevant characteristics of the treatment population in the comparison group given the available data. It is possible that there are other confounding factors that we did not have data on, and thus were not able to adjust for in our analysis. Additionally, while the synthetic control method allowed us to capture the desired scope of impacts of the ETOH-P program, it is non-specific and does not allow us to test the specific mechanism that led to the change. It is possible that unrelated factors we did not control for changed during the study period, and though unrelated to the ETOH-P program, affected rates of our outcome. As such, a major limitation of the synthetic control method is that it does not allow us to distinguish between co-occurring events. To explore this further, future analyses should incorporate techniques such as a negative control, which estimates the effect of an exposure on an outcome that it should be plausibly impact, but which may be affected by a confounding factor[44]. If an effect is still observed, it can be assumed that there is confounding or bias present.

Another major limitation of our study is that other Sutter Hospitals have their own protocols for treated patients at risk for developing AWS, and thus the effect of the ETOH-P program within EMC on our three outcomes of interest may have been attenuated (underestimation of the effect). In fact, actual counts of all three outcomes were very low throughout the entire study period; there were several months in which there were zero patients who left AMA. As the nature of synthetic controls focuses on a single treated unit (or a small number of treated unit), it is possible that small effects are indistinguishable from other influencers of the outcome in the treated unit, especially if the outcome is highly volatile[18]. As a result of this, the impact of “small” programs with effects of a magnitude similar to the volatility of the outcome may be difficult to detect.

Lastly, we recognize that the time period of our study (April 1, 2019 to March 4, 2020) was relatively short, and that additional years of data may be beneficial to our ability to appropriately match the outcome trends at EMC during the pre-implementation period. As with other time series analyses, having sufficient data points prior to implementation and post-implementation is key. Since we only had 16 months of data prior to the

implementation of the ETOH-P protocol, and 11 months post-implementation, additional time points may help to reveal trends in the outcome.

Despite these limitations, we believe that our synthetic control analysis of the ETOH-P program demonstrates an application of the method to a program implemented within a health care system. Whether or not this method should be applied in such a setting depends on the research question of interest and availability of data. Future assessment of the ETOH-P program should also factor in additional information such as severity of AWS symptoms, patient comorbidities and the number of prior diagnoses.

3.5. Conclusions

The application of propensity score matching and the synthetic control method to evaluate implementation of an inpatient hospitalization protocol for AWS did not find evidence that the ETOH-P protocol was effective at reducing the number of patients leaving AMA, the number of patients with an ICU admission, or the average length of stay in the hospital. Future work in this area should incorporate additional years of pre-implementation data, as well as additional information on the severity of AWS symptoms, patient comorbidities and the number of prior diagnoses.

3.6. Figures

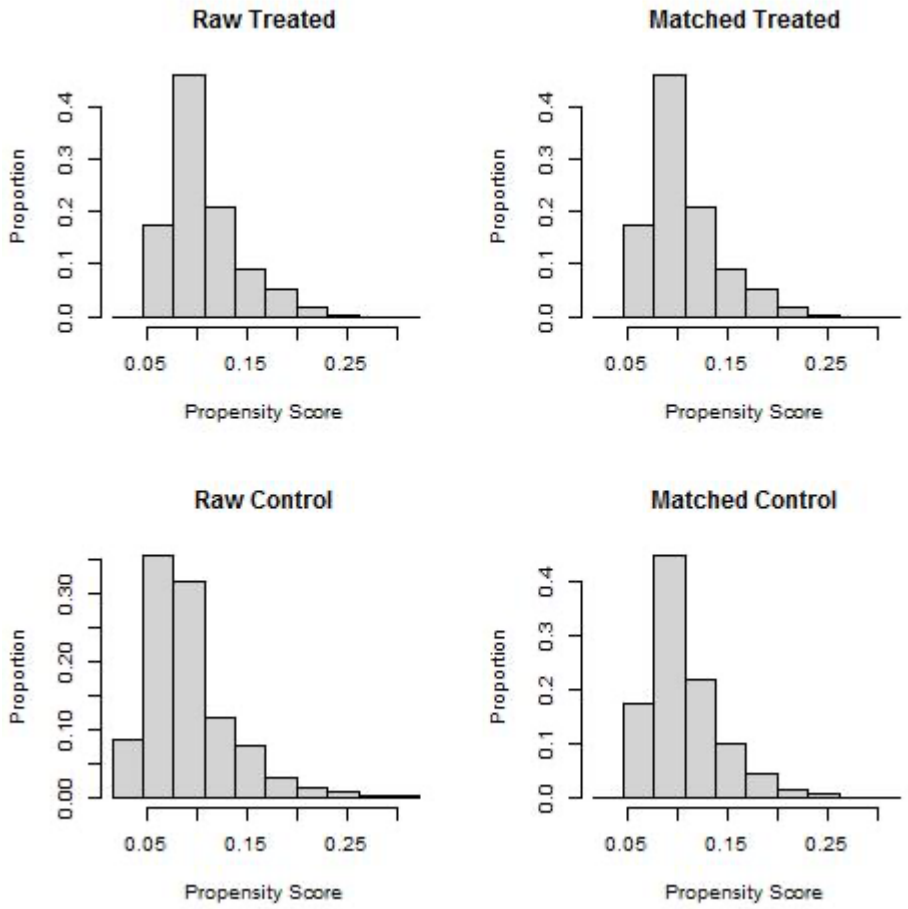


Figure 3.1. Histograms of the propensity score for the original cohort and the matched cohort, ETOH-P patients compared to control patients

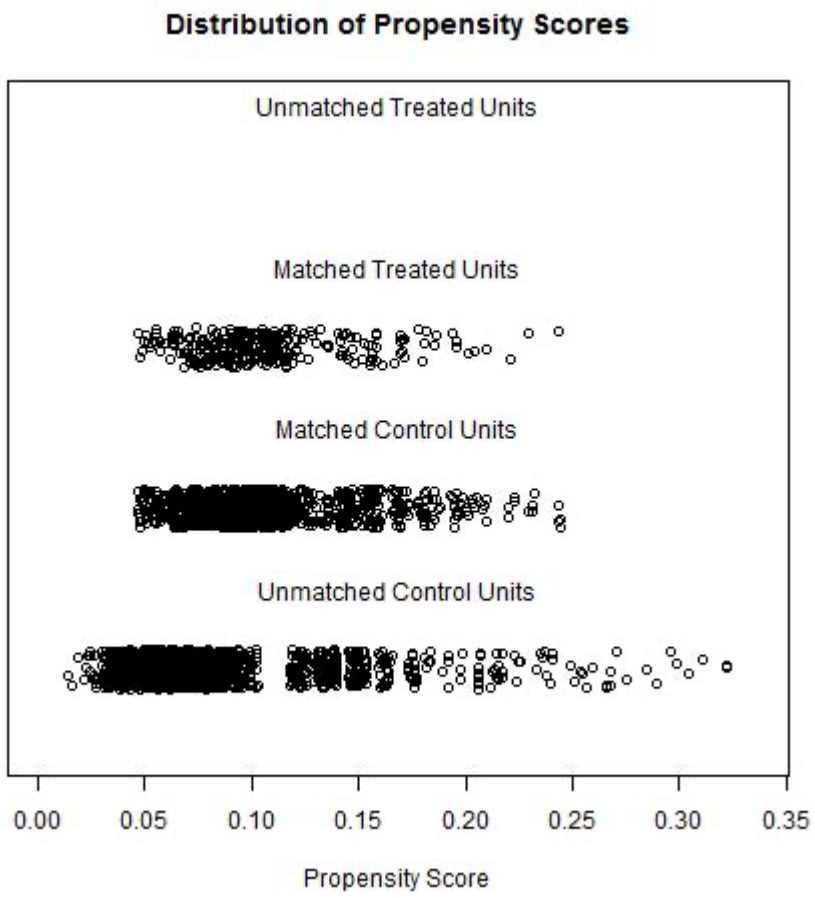


Figure 3.2. Jitter plots of the propensity score for the original cohort and the matched cohort, by ETOH-P patients compared to control patients.

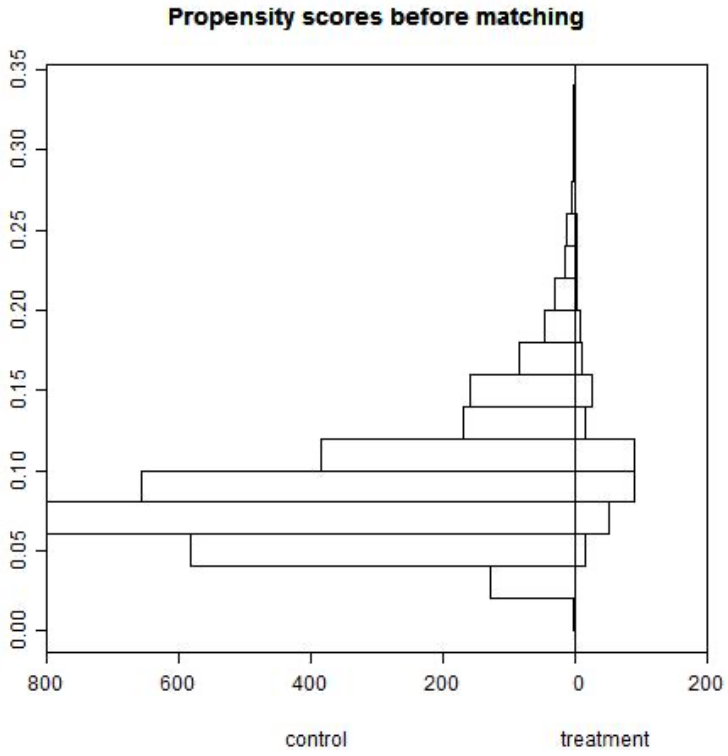


Figure 3.3. Distribution of propensity scores before matching.

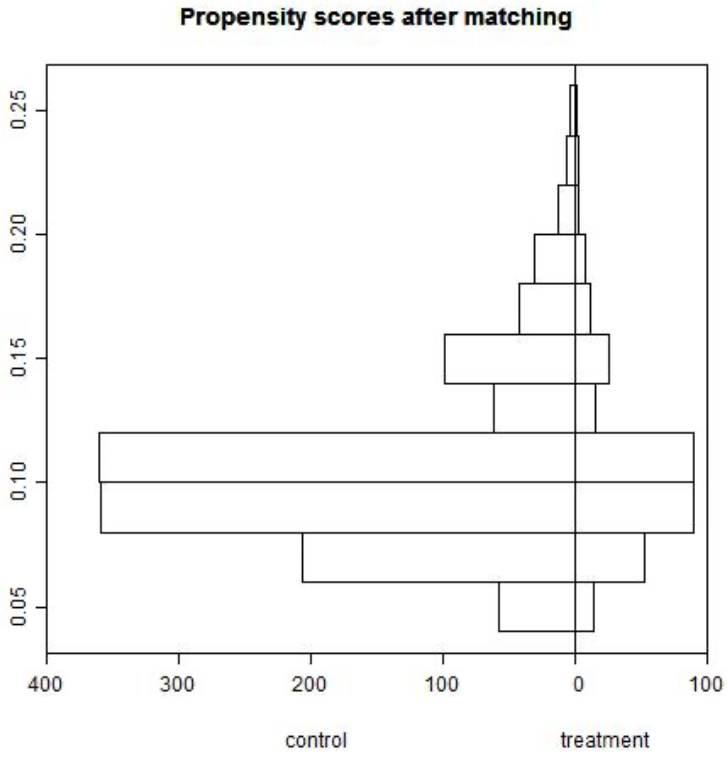


Figure 3.4. Distribution of propensity scores after matching.

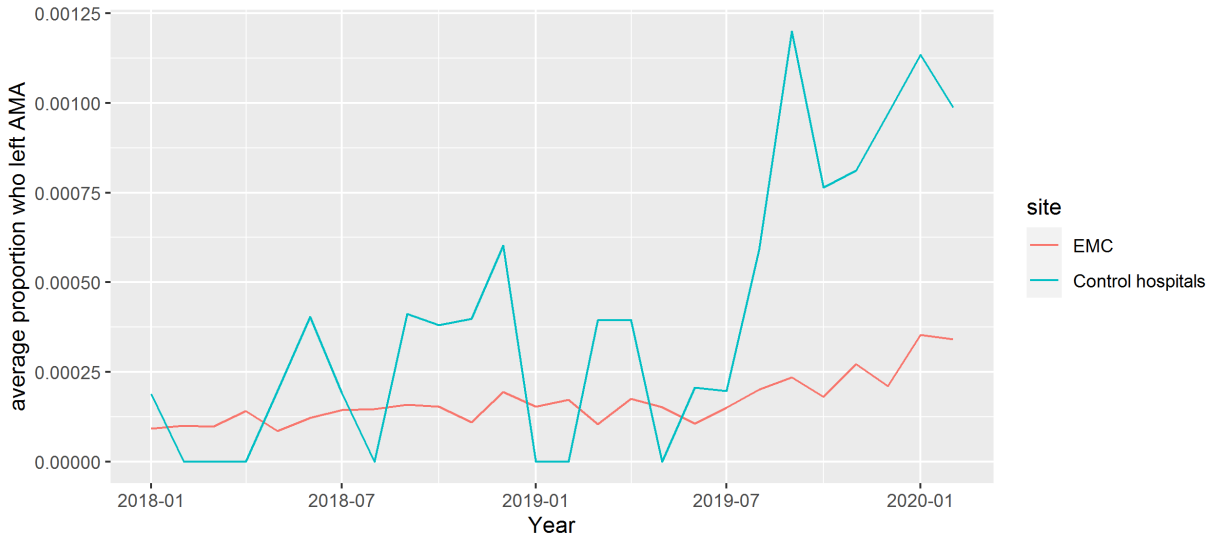


Figure 3.5. Trends in the proportion of patients who leave AMA at Eden Medical Center compared to control hospitals, April 1, 2019 to March 4, 2020.

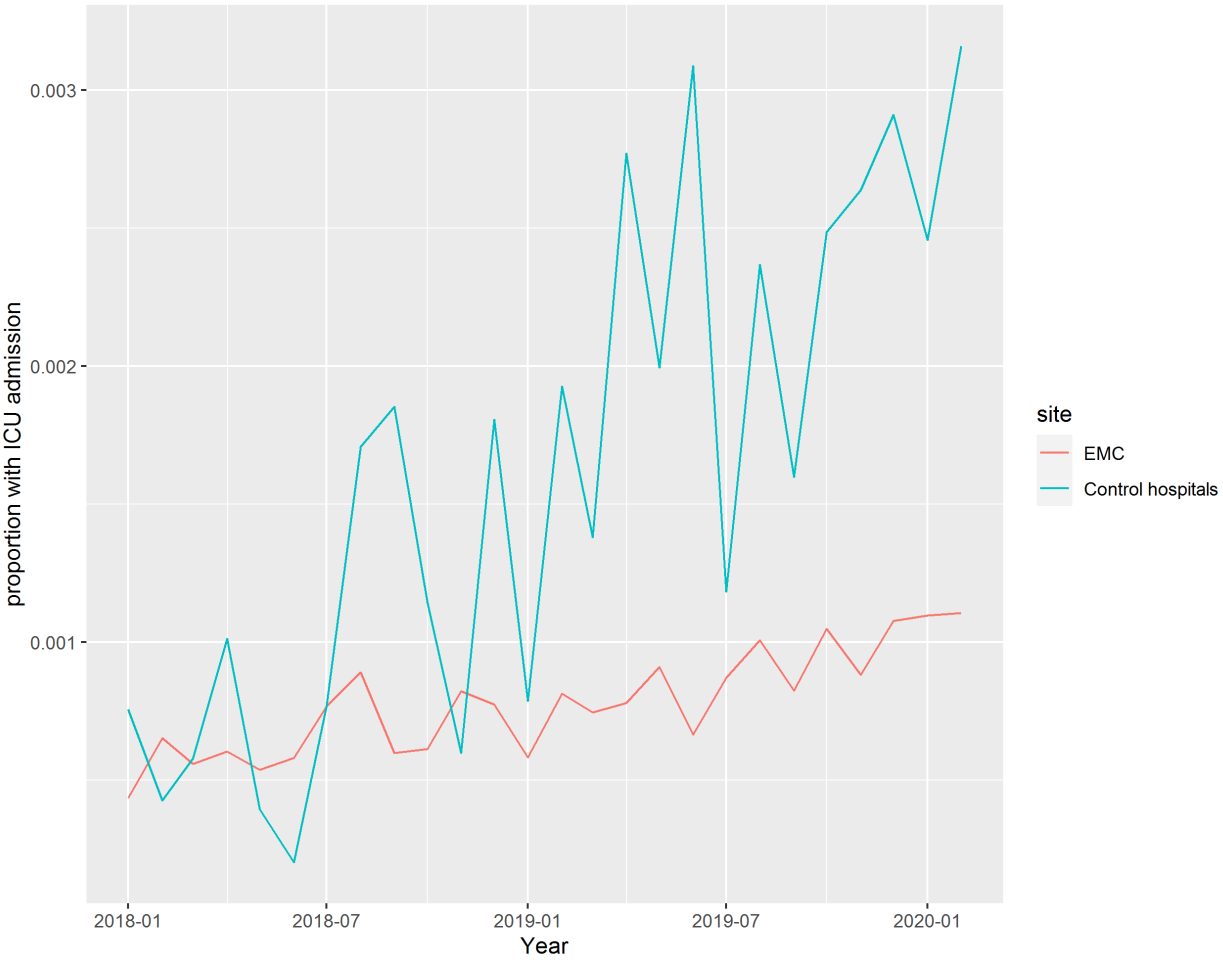


Figure 3.6. Trends in the proportion of patients who have an ICU admission at Eden Medical Center compared to control hospitals, April 1, 2019 to March 4, 2020.

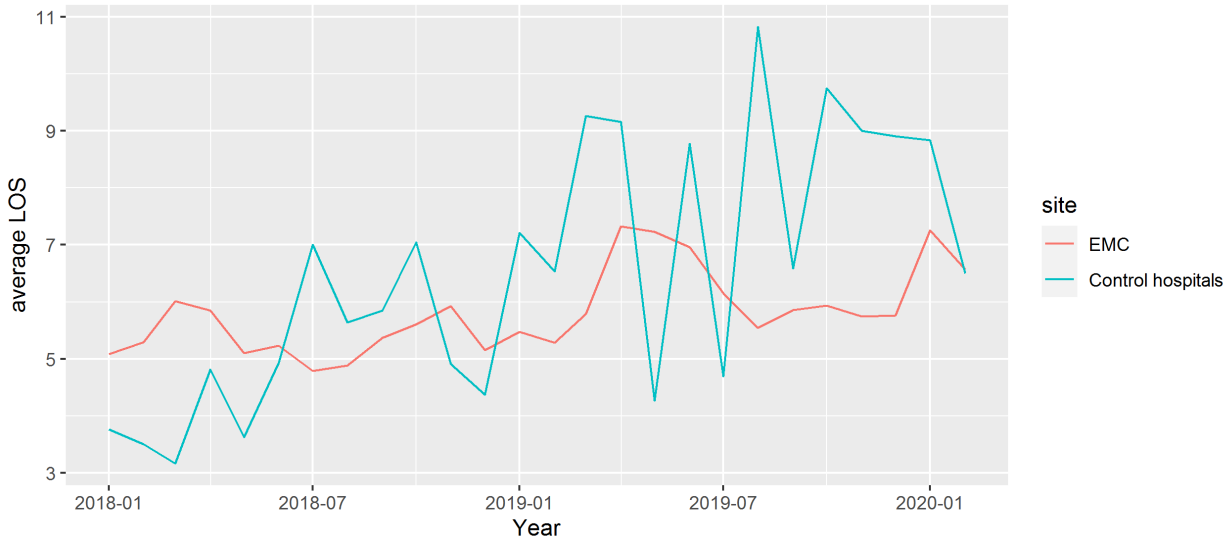


Figure 3.7. Trends in the average LOS at Eden Medical Center compared to control hospitals, April 1, 2019 to March 4, 2020.

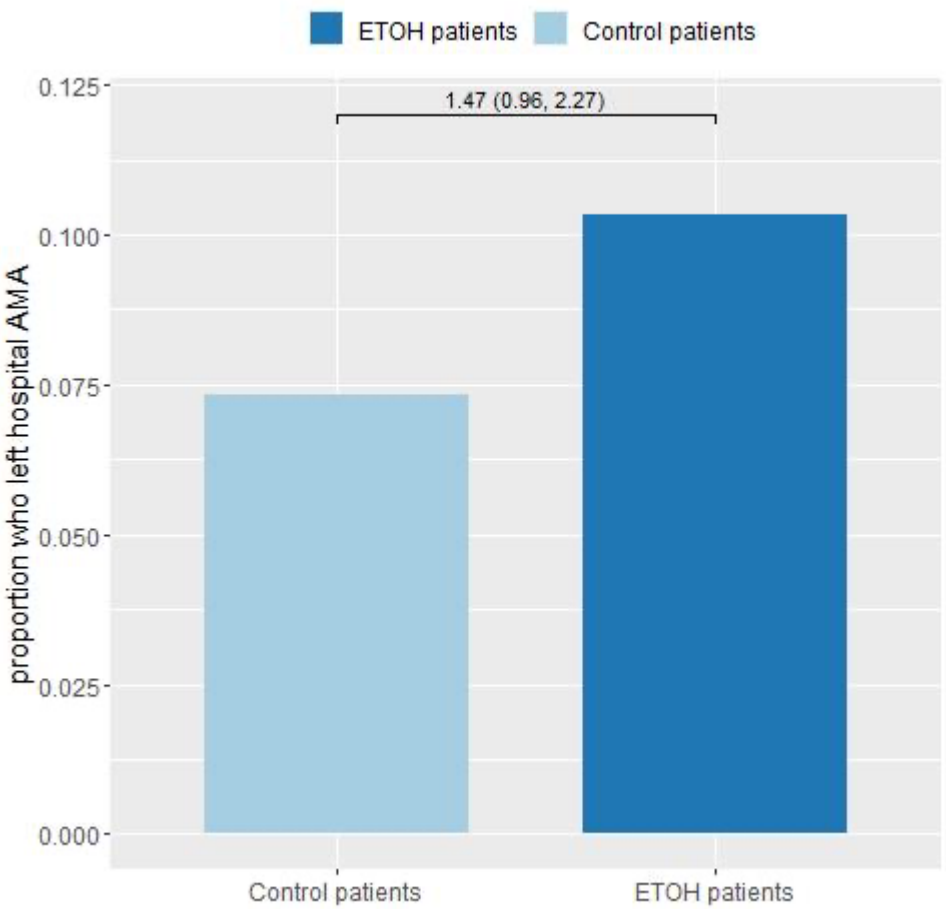


Figure 3.8. Proportion of patients who left AMA for ETOH-P protocol patients vs. control patients, April 1, 2019 – March 4, 2020. Patients who received the ETOH-P protocol had 1.47 (0.96, 2.27) times the odds of leaving AMA compared to control patients.

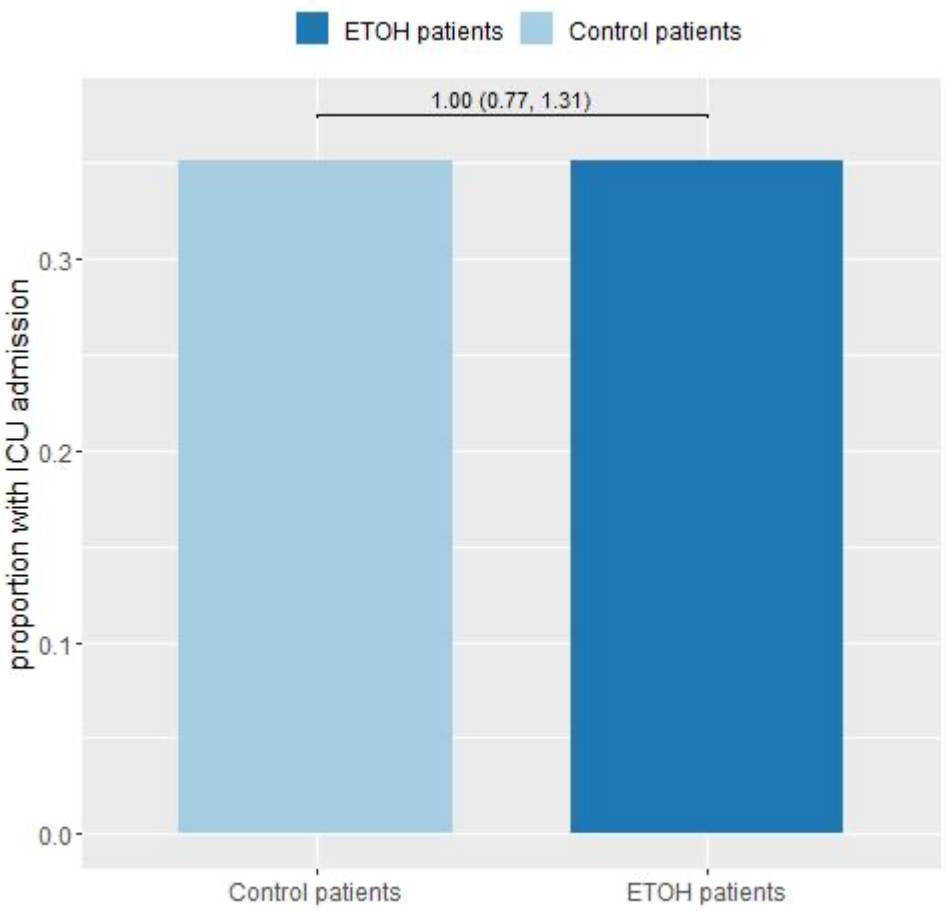


Figure 3.9. Proportion of patients who had an ICU admission for ETOH-P protocol patients vs. control patients, April 1, 2019 – March 4, 2020. Patients who received the ETOH-P protocol had 1.00 (0.77, 1.30) times the odds of an ICU admission compared to control patient

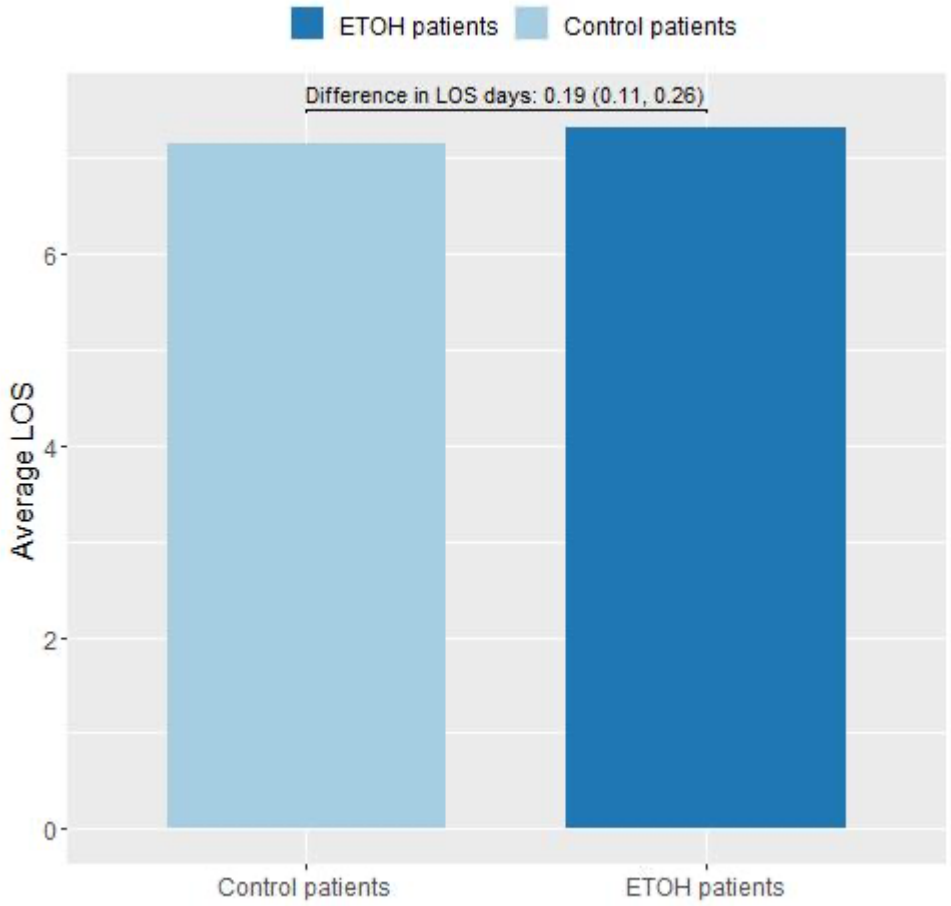


Figure 3.10. Average LOS for ETOH-P protocol patients vs. control patients, April 1, 2019 – March 4, 2020. Patients who received the ETOH-P protocol had on average, 0.19 (0.11, 0.26) more days in the hospital compared to control patients.

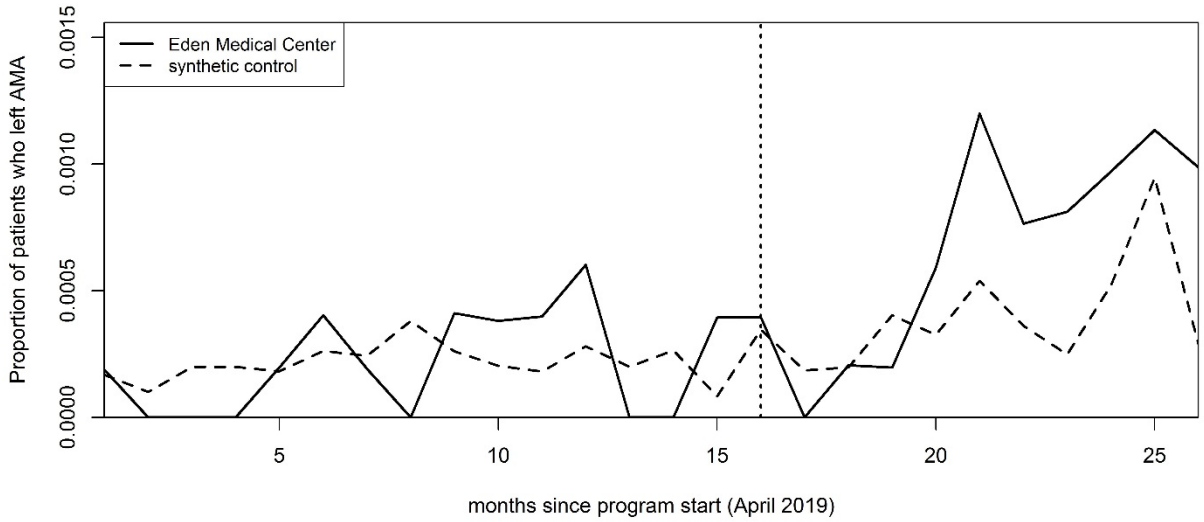


Figure 3.11. Trends in the proportion of patients who left AMA, EMC vs. Synthetic EMC, April 1, 2019 – March 4, 2020.

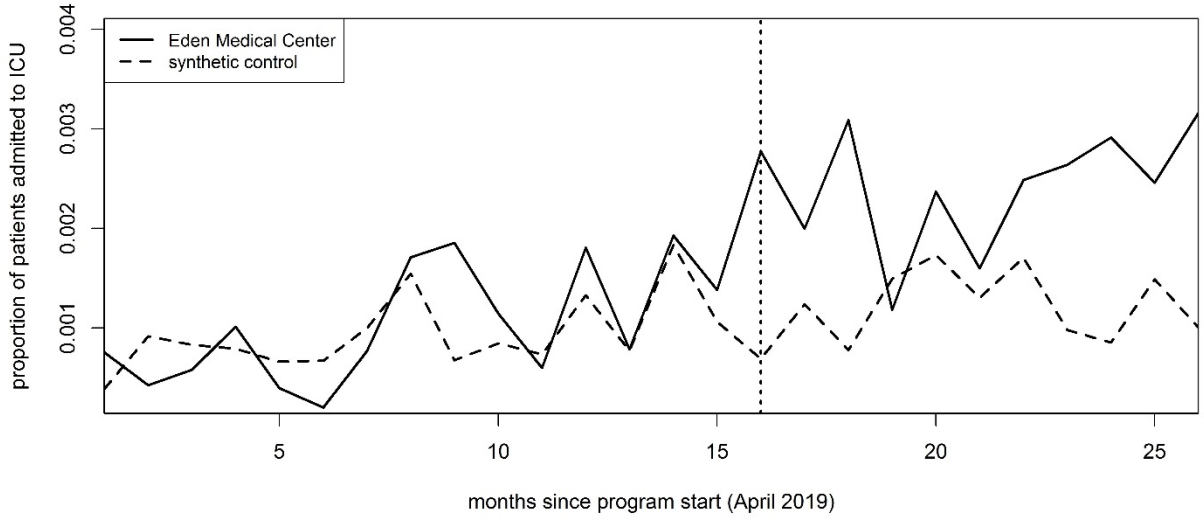


Figure 3.12. Trends in the proportion of patients who had an ICU admission, EMC vs. Synthetic EMC, April 1, 2019 – March 4, 2020.

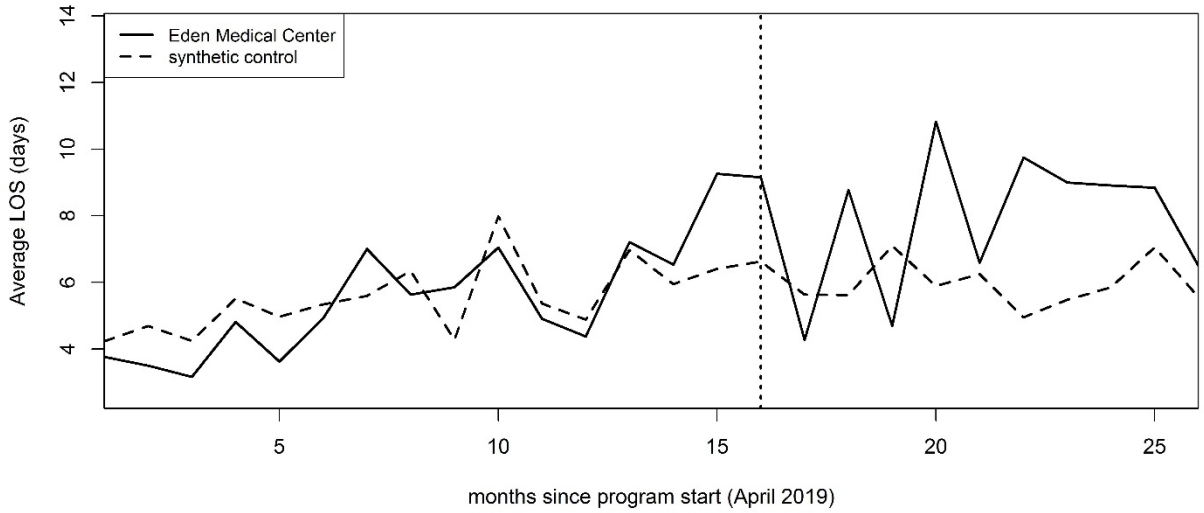


Figure 3.13. Trends in average LOS, EMC vs. Synthetic EMC, April 1, 2019 – March 4, 2020.

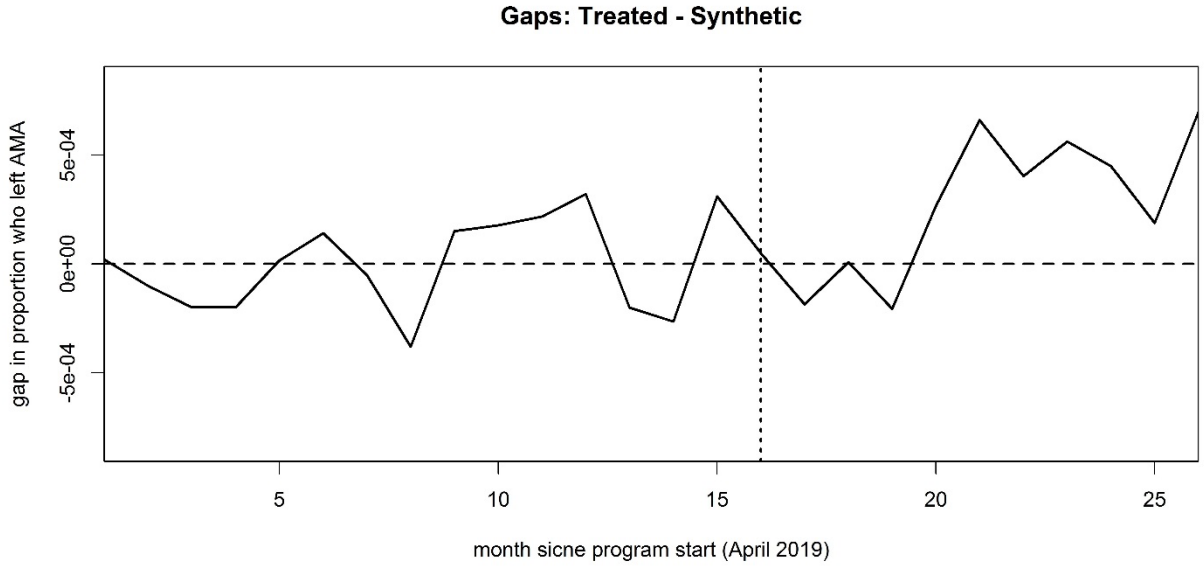


Figure 3.14. Gap in the proportion of patients who left AMA between EMC and Synthetic EMC, April 1, 2019 – March 4, 2020.

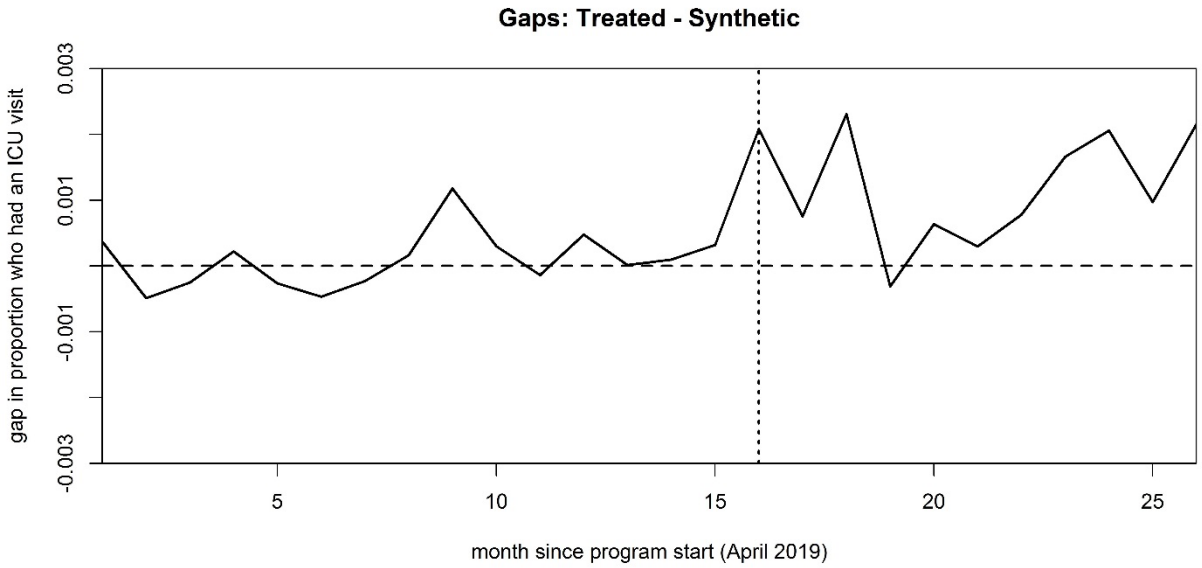


Figure 3.15. Gap in the proportion of patients who had an ICU admission between EMC and Synthetic EMC, April 1, 2019 – March 4, 2020.

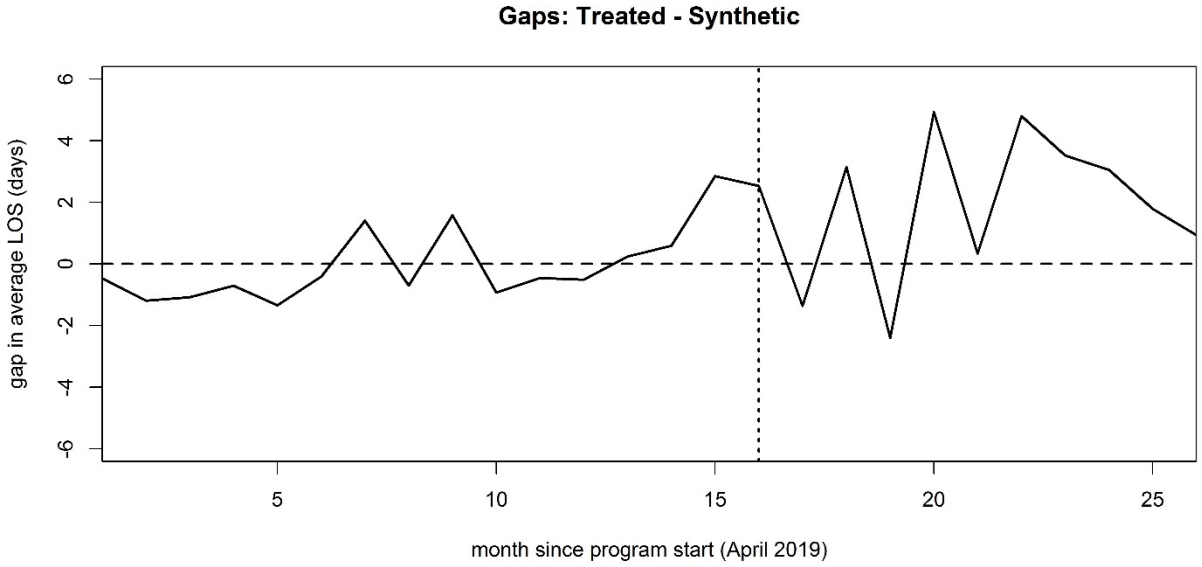


Figure 3.16. Gap in average LOS between EMC and Synthetic EMC, April 1, 2019 – March 4, 2020.

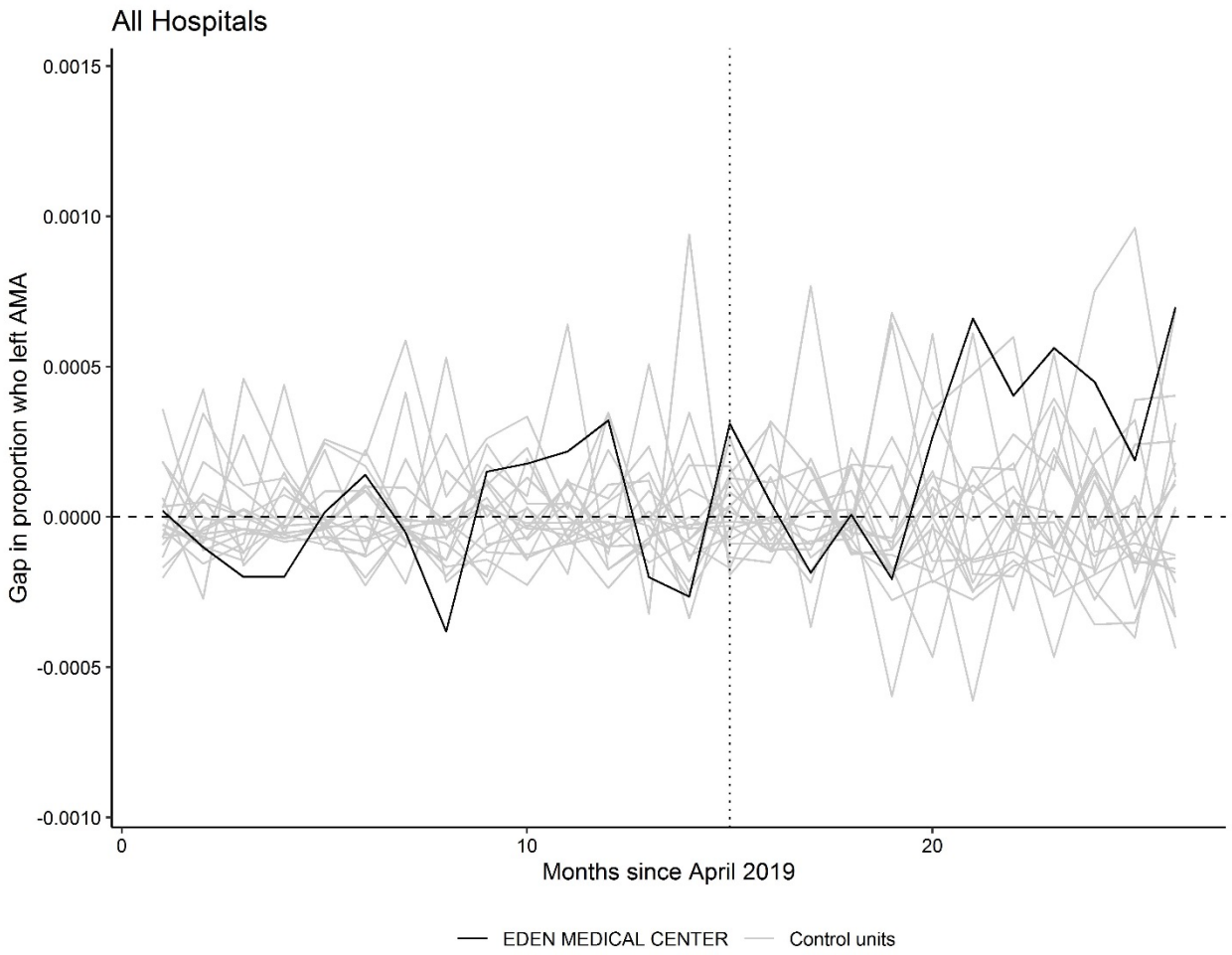


Figure 3.17. Difference in proportion who left AMA between control hospitals and EMC and their respective synthetic controls, all hospitals from April 1, 2019 – March 4, 2020. The superimposed black line denotes the gap estimated for EMC, the hospital that actually implemented the ETOH-P program.

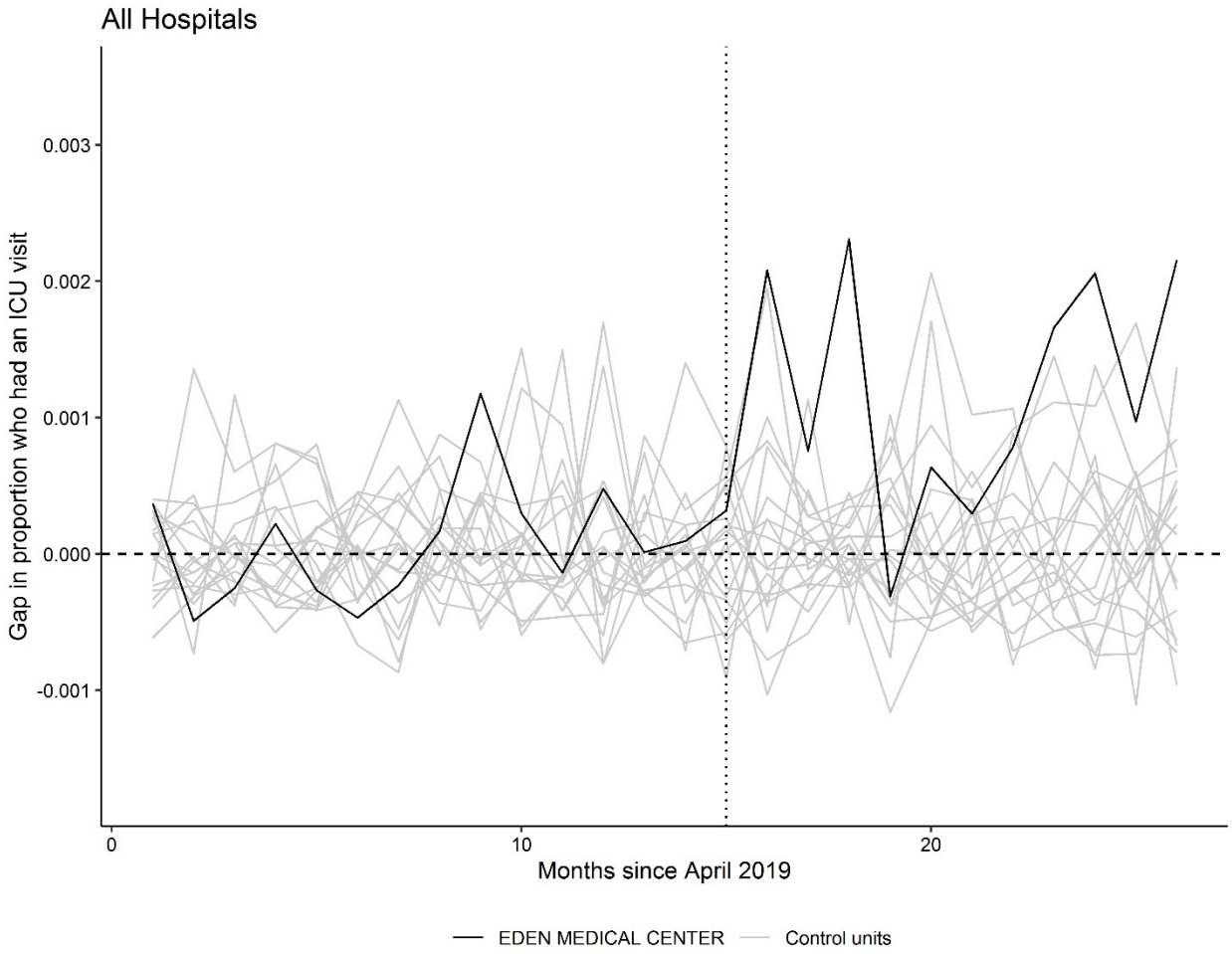


Figure 3.18. Difference in proportion who had an ICU visit between control hospitals and EMC and their respective synthetic controls, all hospitals from April 1, 2019 - March 4, 2020. The superimposed black line denotes the gap estimated for EMC, the hospital that actually implemented the ETOH-P program.

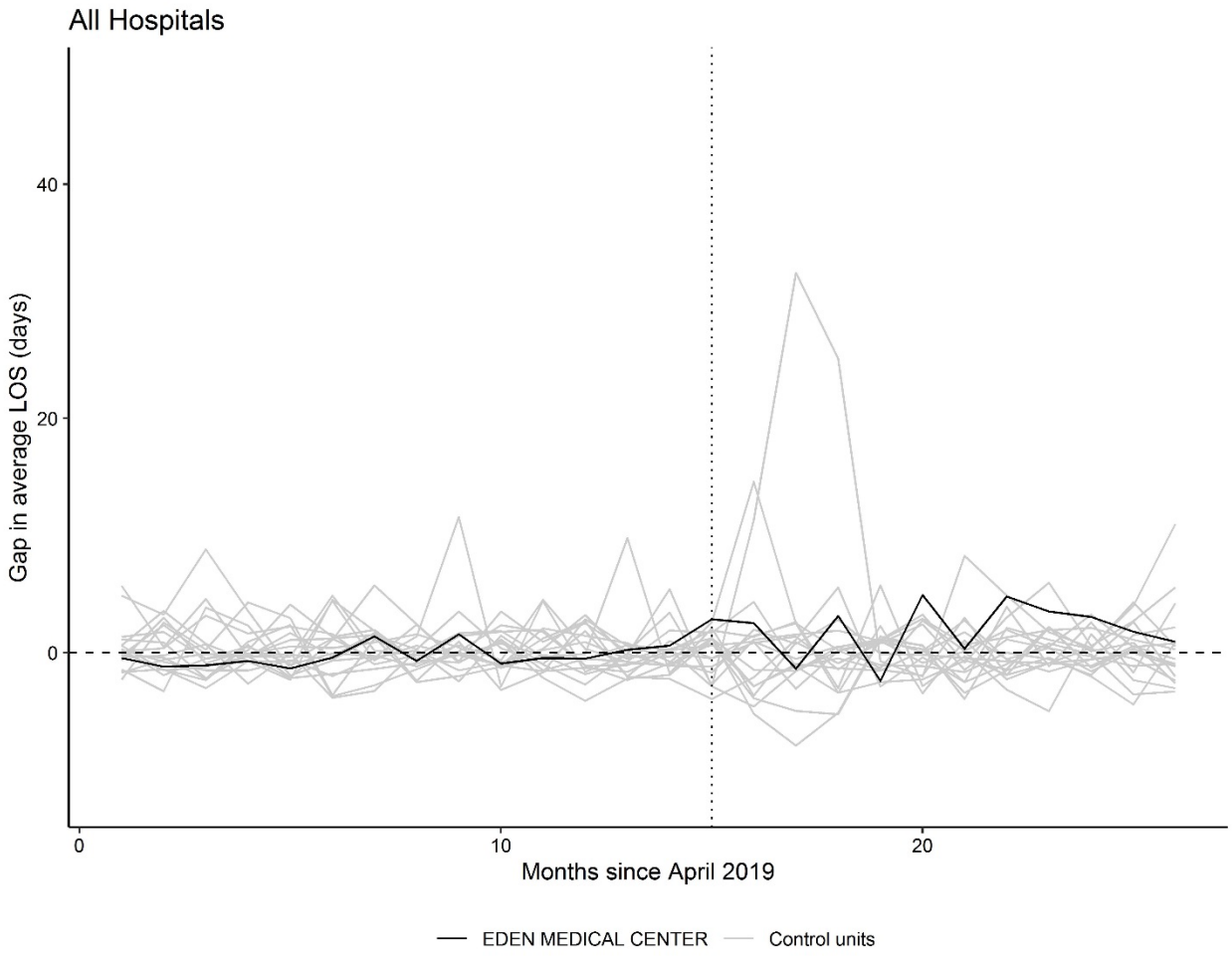


Figure 3.19. Difference in average LOS between control hospitals and EMC and their respective synthetic controls, all hospitals from April 1, 2019 – March 4, 2020. The superimposed black line denotes the gap estimated for EMC, the hospital that actually implemented the ETOH-P program.

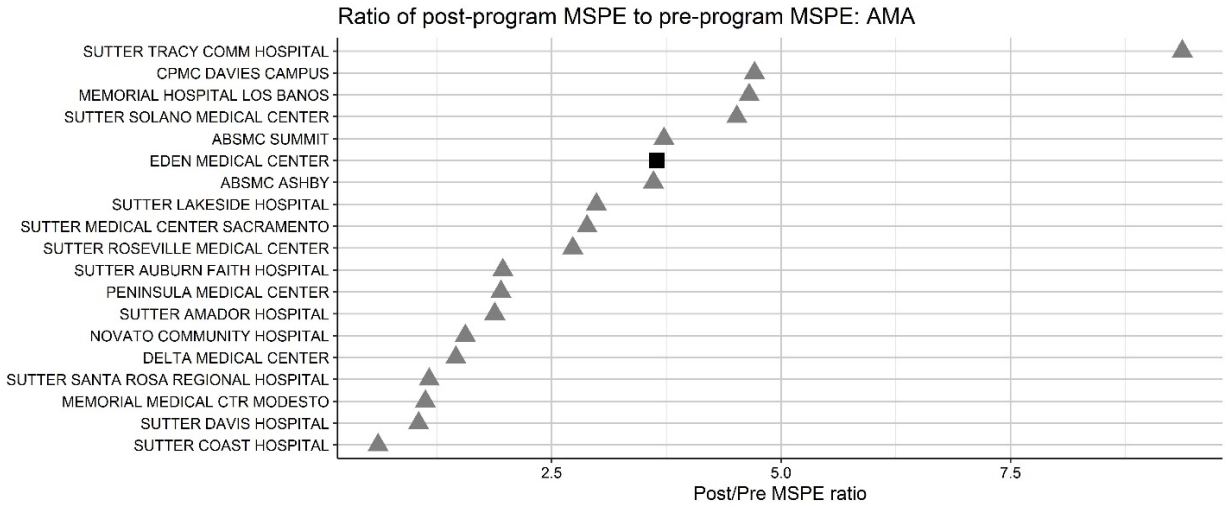


Figure 3.20. Ratios of post-intervention MSPE and pre-intervention MSPE for AMA at EMC and all hospitals in the donor pool.

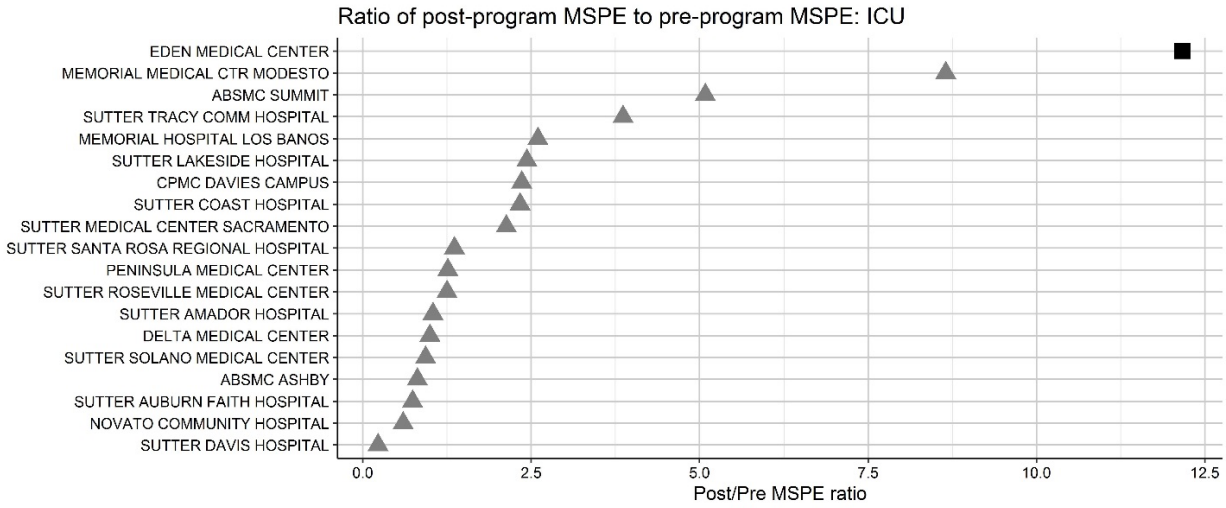


Figure 3.21. Ratios of post-intervention MSPE and pre-intervention MSPE for ICU admissions at EMC and all hospitals in the donor pool.

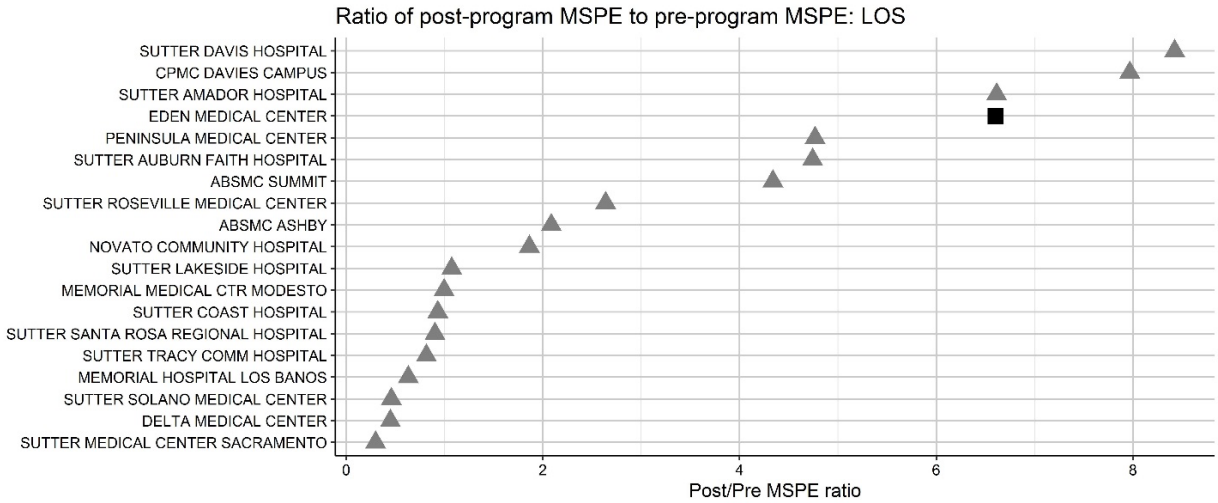


Figure 3.22. Ratios of post-intervention MSPE and pre-intervention MSPE for LOS at EMC and all hospitals in the donor pool.

3.7. Tables

Table 3.1. Baseline covariates for ETOH-P patients and control patients.

	Unmatched Cohort			Matched Cohort		
	ETOH-P patients	**Control Patients**		**ETOH-P patients**	**Control Patients**	
	N = 310	**N = 3,086**	**|SMD|**	**N = 310**	**N = 1229**	**|SMD|**
DEMOGRAPHICS						
Female, n (%)	67 (21.6)	960 (31.1)	0.217	67 (21.6)	263 (21.4)	0.005
Mean Age, Years ± SD	55.06 ± 15.37	53.88 (14.61)	0.078	55.06 ± 15.37	55.09 (14.47)	0.002
Ethnicity						
Hispanic	0.24(0.43)	0.18 (0.39)	0.144	0.24 (0.43)	0.23 (0.42)	0.037
Non-Hispanic	0.75 (0.43)	0.80 (0.40)	0.082	0.75 (0.43)	0.77 (0.42)	0.076
Unknown	0.01 (0.45)	0.01 (0.12)	0.121	0.01 (0.08)	0.00 (0.04)	0.048
Race category						
Black	0.15 (0.35)	0.14 (0.34)	0.022	0.15 (0.35)	0.14 (0.34)	0.027
Asian	0.06 (0.24)	0.03 (0.17)	0.148	0.06 (0.24)	0.05 (0.22)	0.043
Other	0.17 (0.38)	0.17 (0.38)	0.004	0.17 (0.38)	0.15 (0.36)	0.066
White	0.64 (0.48)	0.67 (0.47)	0.075	0.64 (0.48)	0.67 (0.47)	0.080
Insurance type						
Medi-Cal	0.44 (0.5)	0.46 (0.50)	0.052	0.44 (0.50)	0.44 (0.50)	0.016
Medicare FFS	0.32 (0.47)	0.29 (0.45)	0.052	0.32 (0.47)	0.33 (0.47)	0.039
Other/Unknown	0.07 (0.26)	0.07 (0.25)	0.018	0.07 (0.26)	0.06 (0.24)	0.030

Table 3.2. Predictors prior to the implementation of the ETOH-P protocol.

Predictors	Treated	Synthetic	Sample Mean
avg_age	51.854	50.856	52.517
prop_hispanic	0.096	0.074	0.062
prop_nonhispanic	0.244	0.247	0.209
prop_unketh	0.008	0.012	0.012
prop_otherrace	0.006	0.003	0.003
prop_asian	0.056	0.048	0.026
prop_black	0.058	0.066	0.027
prop_multirace	0.004	0.004	0.003
prop_unkrace	0.073	0.077	0.060
prop_white	0.152	0.136	0.164
prop_male	0.154	0.145	0.125
prop_female	0.193	0.188	0.157
prop_unksex	0.000	0.000	0.000
prop_otherins	0.010	0.013	0.010
prop_public	0.102	0.095	0.075
prop_medicare	0.086	0.078	0.077
prop_private	0.115	0.114	0.096
prop_unkins	0.034	0.033	0.024
prop_uninsured	0.000	0.000	0.000
prop_ama	0.000	0.000	0.000

Table 3.3. Hospital weights for synthetic EMC in approximating the proportion of patients who leave AMA.

Predictors	Treated
ABSMC Ashby	0.167
ABSMC Summit	0.288
CPMC Davies Campus	0.037
Delta Medical Center	0.167
Memorial Hospital Los Banos	0.000
Memorial Medica Ctr Modesto	0.002
Novato Community Hospital	0.000
Peninsula Medical Center	0.000
Sutter Amador Hospital	0.000
Sutter Auburn Faith Hospital	0.003
Sutter Coast Hospital	0.001
Sutter Davis Hospital	0.150
Sutter Lakeside Hospital	0.000
Sutter Medical Center Sacramento	0.015
Sutter Roseville Medical Center	0.000
Sutter Santa Rosa Regional Hospital	0.001
Sutter Solano Medical Center	0.007
Sutter Tracy Communimty Hospital	0.312

Table 3.4. Hospital weights for synthetic EMC in approximating the proportion of patients with an ICU admission.

Predictors	Treated
ABSMC Ashby	0.000
ABSMC Summit	0.000
CPMC Davies Campus	0.000
Delta Medical Center	0.561
Memorial Hospital Los Banos	0.000
Memorial Medica Ctr Modesto	0.000
Novato Community Hospital	0.000
Peninsula Medical Center	0.000
Sutter Amador Hospital	0.000
Sutter Auburn Faith Hospital	0.000
Sutter Coast Hospital	0.040
Sutter Davis Hospital	0.150
Sutter Lakeside Hospital	0.000
Sutter Medical Center Sacramento	0.000
Sutter Roseville Medical Center	0.000
Sutter Santa Rosa Regional Hospital	0.249
Sutter Solano Medical Center	0.000
Sutter Tracy Communimty Hospital	0.000

Table 3.5. Hospital weights for synthetic EMC in approximating the average LOS.

Predictors	Treated
ABSMC Ashby	0.247
ABSMC Summit	0.009
CPMC Davies Campus	0.000
Delta Medical Center	0.052
Memorial Hospital Los Banos	0.000
Memorial Medica Ctr Modesto	0.000
Novato Community Hospital	0.000
Peninsula Medical Center	0.205
Sutter Amador Hospital	0.000
Sutter Auburn Faith Hospital	0.000
Sutter Coast Hospital	0.136
Sutter Davis Hospital	0.000
Sutter Lakeside Hospital	0.058
Sutter Medical Center Sacramento	0.084
Sutter Roseville Medical Center	0.000
Sutter Santa Rosa Regional Hospital	0.177
Sutter Solano Medical Center	0.019
Sutter Tracy Communimty Hospital	0.012

Table 3.6. Mean squared prediction error (MSPE) in the pre-intervention period.

Outcome	MSPE
Leaving AMA	0.000000045
ICU Admission	0.00000018
Average LOS	1.34

Chapter 4 - A Synthetic Control Approach to Evaluating ELEVATE-DP

Abstract

Purpose:
To apply the synthetic control method (SCM) to evaluate the impact of the GLB program on mean weight (kg) and mean body mass index (BMI) at the clinic-level.

Methods: We utilized the SCM to conduct a retrospective secondary analysis of electronic health record (EHR) data from the ELEVATE-DP study in order to evaluate the impact of the GLB program offered at 20 different Sutter Health clinics. Using data from 2005 to 2017 and a pool of control clinics that did not offer the GLB program during the study period, we constructed a synthetic control comparison group for the GLB clinics. We used a series of permutation tests to assess whether our results could have been due to chance.

Results
A total of 26 control clinics were included in the synthetic control. While the outcome trends for the actual GLB clinics remained elevated over the synthetic control during the period after program implementation, we did not find a significant effect of the GLB program on either mean weight or mean BMI.

Conclusions
This study provides valuable insight into the feasibility and applicability of SCM applied to evaluating a health care system program. Future studies should explore methodological adjustments to SCM that will take into account programs that do not uniformly impact the entire population of interest.

Abbreviations

BMI – Body Mass Index
ELEVATE-DP – Evaluation of the Lifestyle Intervention Adopted for Clinical Practice for Diabetes Prevention
DPP – Diabetes Prevention Program
LCP – Lifestyle Change Program
CDC – Centers for Disease Control and Prevention
EHR – Electronic Health Record
GLB – Group Lifestyle Balance

ICD – International Classification of Disease
SCM – Synthetic Control Method
IGT – Impaired glucose tolerance

4.1. Background

Diabetes and GLB program background

In the United States, more than 34 million people have diabetes, and over 88 million US adults have prediabetes, both of which increases the risk of developing a heart disease and stroke[65]. The CDC estimates that in 2017, the total cost of diagnosed diabetes was $327 billion. With rates of obesity increasing and much of the country leading a sedentary lifestyle, the burden of diabetes will also continue to increase.

The Diabetes Prevention Program (DPP) was a 27-center randomized clinical trial to determine whether a lifestyle intervention program or pharmacological therapy could help prevent or delay the onset of diabetes in individuals who had impaired glucose tolerance (IGT). The DPP found that an intensive lifestyle intervention program reduced the incidence of diabetes by more than half[66–69]. However, intensive lifestyle interventions like DPP are difficult to implement and sustain within busy healthcare systems, thus there is a need for real-world adaptations of the DPP lifestyle intervention.

To address this, and to accomplish the widespread implementation of the DPP results, the CDC established the National DPP in order to develop an evidence-based, comprehensive training curriculum to deliver a year-long lifestyle change program to people with diagnosed prediabetes or those at high risk for developing Type 2 diabetes[70]. The curriculum focuses on lifestyle change and the importance of at least 150 minutes of moderate physical activity per week, healthy eating and weight loss of 5-7% over the 1-year program period. Since establishing the National DPP, there have been numerous applications of translating the DPP into group-based lifestyle interventions within the general community, with most studies finding clinically meaningful reductions in weight among program participants[68,69,71].

At Sutter Health, the Group Lifestyle Balance (GLB) program is also modeled off the original DPP intervention, consisting of a year-long, group-based curriculum divided into three phases. To our knowledge, there is no previous work that utilizes SCM to evaluate a group-based lifestyle change program (LCP). As the method focuses on determining group-level effects, we are ultimately interested in understanding the impact of the GLB program at the clinic-level.

Synthetic control method

SCM is an approach to program evaluation in which one or a small number of units are subject to intervention, and a comparative control unit is constructed such that the outcomes of the control units are weighted to construct the counterfactual outcome of

the treated unit(s) in the absence of the treatment. This method has previously been used to analyze political and economic effects following large-scale events, state-level policy changes, health systems reforms, nutritional interventions, climate changes, and even the current COVID-19 pandemic[2,4,8,17,37,38,40,60] where it is difficult to find a single comparison unit that best approximates the relevant characteristics of the treated unit; indeed, a combination of units often provides a better comparison unit than any single unit alone.

SCM offers another tool for program evaluation, in which time series for the unit of interest in the period before the intervention are used to make predictions about what future trends would look like without the intervention. No extrapolation is required as weights are required to be non-negative and sum to 1, and the weights are calculated and chosen without seeing the post intervention data, reducing the risk of cherry picking or p-hacking. The contribution of each control unit to the counterfactual is made explicit and offers transparency in the selection of the best counterfactual[18]. Additionally, SCM provides a visual representation of the impact of the intervention and how it varies over time, as well as a clear visualization of the actual discrepancy between the treated unit and the convex combination of untreated units, something that propensity score methods do not provide. Plots are produced that display what the observed outcome looks like compared to what would be expected in the absence of the intervention. Lastly, predictions from SCM may serve as input to the propensity score model, by acting as a baseline counterfactual.

The main analysis sought to examine the effectiveness of the Sutter Health GLB program in comparison to usual care in a real-world healthcare setting (ELEVATE-DP study). An electronic health record (EHR)-based, propensity score matched analysis was performed and found that compared with usual-care patients (N=965,265), participants in the program (N=3,156) had a greater odds of attaining clinically meaningful weight loss at 12-months and 24-months of follow-up, as well as a greater odds of blood pressure control at 12-months of follow-up[72]. While the GLB program is administered at the individual patient level, the health equity leadership committee at Sutter Health are also interested in the clinic-level effects of the program, and thus in the present analysis, we are interested in detecting any clinic-level changes as a result of the GLB program. Our hypothesis is that the in-person, group-based GLB program will support patients to lose weight and make behavioral modifications that will ultimately lead to improved weight management and reduced incidence of Type 2 Diabetes. As such, we would expect that the GLB would lead to a decrease in mean weight and BMI at the clinics that implemented the GLB program.

4.2. Methods

Study Design and Setting

This is a retrospective EHR-based observational study conducted at Sutter Health, a large, private, and not-for-profit community-based health care delivery system in Northern California that provides medical services across 130 ambulatory clinics and 24

acute care hospitals, including 22 ED sites. All Sutter Health clinics and hospitals are linked by a single electronic health record system (Epic, Verona, WI). Sutter Health has approximately 11 million ambulatory visits, 870,000 ED visits and 200,000 hospital discharges annually. This study was approved by the Sutter Health Institutional Review Board (IRB) and the UC Berkeley Committee for the Protection of Human Subjects (CPHS), with a Health Insurance Portability and Accountability Act (HIPAA) waiver of authorization and informed consent.

GLB Program

Sutter Health offers a group-based, 12-month structured LCP, known as the Group Lifestyle Balance (GLB) program, an adaptation of the original Diabetes Prevention Program (DPP) intervention. The GLB program targets individuals who are non-diabetic, overweight/obese individuals, at least 18 years of age, with a diagnosis of pre-diabetes and/or metabolic syndrome but is open to all patients who are at risk of a cardiometabolic event[72]. The program has been offered at 20 clinics since 2010, and consists of a 12 month in-person, group-based format led by trained lifestyle coach, with the primary goal of enhancing self-efficacy through social support and mastery of self-regulation skills (e.g., goal setting, self-monitoring).

The GLB program is composed of three phases[72]:

- The core phase (months 1-3) focuses on weight loss and behavioral goal setting through weekly sessions for the first 12 weeks.
- The transition phase (months 4-6) continues the focus on weight loss and behavioral goals through four bimonthly/monthly sessions.
- The support phase (months 7-12) consists of a session once a month to (1) facilitate continued behavior change through an iterative guided mastery process, (2) to foster self-efficacy and independence, and (3) to reinforce problem-solving and behavior maintenance skills.

The goal of the GLB program is to enhance self-efficacy through social support and gradual mastery of self-regulation skills (e.g., goal setting, self-monitoring).

Exposure

If the clinic implemented the GLB program, it will be considered a “treated” clinic. Clinics in which GLB was not available will be considered “control” clinics.

Covariates

We will utilize the following hospital-level covariates:

- Mean weight
- Mean BMI
- Proportion of the patient population that is male

- Mean age
- Median Household Income based on hospital zip code
- Proportion of the patient population that is covered by each type of health insurance (Preferred Provider Organization (PPO)/ Fee-for-service (FFS), health maintenance organization (HMO), Medicare, Medicaid, Other, Unknown)
- Mean diastolic blood pressure
- Mean systolic blood pressure
- Proportion of the population who are smokers

Outcome measures

The two primary outcomes of interest were mean weight and mean BMI.

Statistical Analyses

Data extraction and statistical analyses were performed using R statistical software package v3.6.3 and Microsoft SQL Server Management Studio 13.0.15700.28.

We used SCM developed by Abadie, Diamond and Hainmueller[17] to assess the impact of the GLB program at the clinic-level. This method uses a weighted combination of clinics to create a "synthetic" version of the treated clinic(s) that estimates the expected trends in mean weight and mean BMI in the GLB clinics had they never implemented the program. SCM then compares these to the observed trends to quantify the GLB program's impact on weight and BMI during 2010-2017. The difference between the observed and expected values is the effect of interest.

To construct the synthetic control, we used pre-program outcome and covariate data from 2005 to 2010. While most examples in the synthetic control literature utilize one treatment unit, we had 20 clinics that implemented the GLB program. Thus, we chose to group these clinics together into one treated clinic, as suggested by Abadie et al., 2010.

Traditional large sample inferential techniques are not appropriate in this setting due to the small number of units; however, exact inferential techniques, such as a placebo/permutation test may be used to assess how unusual an effect would be if it were due to chance and thus provide context for the effect size. As advised in prior literature, we obtained inference for these estimates by using a series of placebo (permutation) tests, in which we applied the synthetic control method for each control clinic, as though it were the one that had implemented the GLB program. The estimated effect of the actual GLB clinic can then be compared to the size of these other effect estimates. Typically, the permutation test results are compared for clinics in which pre-program trends are well predicted by the synthetic control. Thus, we compared the effect of the actual GLB clinics to all control clinics, as well as just those with five times the MSPE observed for the actual GLB clinics, because clinics with a poorly matched synthetic control might appear to have more extreme differences resulting from an artifact of poor prediction.

4.3. Results

Synthetic Control results

A total of 26 clinics comprised the pool of possible donor control clinics. These were the only clinics that had sufficient outcome and covariate data during the pre-program period to construct a synthetic control. Additionally, we chose to discard data from 2002-2004 due to large amounts of missing data. Both mean weight and mean BMI were on the rise during the study period (**Figure 4.1 and 4.2**). The covariates used to calculate the weights for the synthetic control series are listed in **Table 4.1**, along with their values in GLB clinics, the mean values for the rest of the clinics, and their values in the mean weight and mean BMI synthetic controls.

The clinics included in the synthetic controls are listed along with their corresponding weight in **Tables 4.2 and 4.3**. The weights indicate that mean weight and mean BMI trends for GLB clinics are best reproduced by a combination of many clinics, with only a few receiving zero weight. We used the mean squared prediction errors (MSPE) to measure fit between the treated unit and its synthetic control during the pre-intervention period. The MSPE in the pre-program period for both outcomes are listed in **Table 4.4**. A small MSPE indicates that the synthetic control approximates well the actual outcome trend during the pre-program period. A comparison of outcome trends in GLB clinics vs non-GLB clinics (**Figures 4.1 and 4.2**) illustrates that a simple average of the control units does not closely approximate the outcomes of GLB clinics. Thus, we sought to explore if the construction of a synthetic control would provide a better "counterfactual" to the GLB clinic trend.

Overall, the pre-program mean weight and mean BMI levels appear to be well approximated by the synthetic control (**Figures 4.3 and 4.4**). Actual trends of mean weight diverged from the synthetic control in 2012, increasing and peaking around 2016, while remaining elevated over the control for most of the post-program period (**Figure 4.3**). The differences in mean weight between the actual GLB clinics and its synthetic control from 2010 to 2017 indicate that there was an average increase in mean weight during the program of approximately 4.2 kg (**Figure 4.5**). The permutation test suggests that this effect is not significant compared to other clinics (**Figures 4.7 and 4.8**), when comparing to all control units, as well as just those control units with an MSPE < 5 times that of the actual GLB clinics. The ratio of post-program MSPE to pre-program MSPE for the actual GLB clinics was approximately 54. Only one other clinic had a ratio larger than this. Despite this, we obtained a p-value of 0.07 from the permutation tests and thus, the effect of the program on mean weight was not significant.

Actual trends of mean BMI remain relatively steady across the study period and start to diverge from the synthetic control around 2012. The actual mean BMI trend was elevated over the synthetic control for most of the post-program period (**Figure 4.4**).

The differences in mean BMI between the actual GLB clinics and its synthetic control from 2010 to 2017 indicate that there was an average increase in mean BMI during the post-program period of approximately 1.64 (**Figure 4.6**). This did not appear to be a significant effect as the permutation tests indicate that the observed effect for the actual GLB clinics was not unusually large compared to the distribution of effects from other clinics (**Figures 4.9 and 4.10**). Additionally, the ratio of post-program MSPE to pre-program MSPE for the actual GLB clinics does not stand out from other clinics (**Figures 4.11 and 4.12**); one clinic had a higher ratio than the actual GLB clinics.

4.4. Discussion

We did not find evidence that the GLB program was associated with a decrease in mean weight and mean BMI relative to the synthetic control. For both outcomes, the average difference between the actual GLB clinics and the synthetic control during the post-program period was positive. If our hypothesis about the impact of the GLB program was correct, we would have expected to see a reduction in our outcomes of interest, and thus the gap (treated – synthetic) would be a negative number. However, as the average difference between the two series remained positive during the entire study period, we saw the opposite effect than what we would have expected.

Based on previous studies of diabetes prevention and education programs, we expected to see a decrease in weight and BMI as a result of the GLB program[73]. However, our results suggest that weight and BMI increased during the program period. In fact, the results of this synthetic control system-level analysis contradict the primary individual-level analysis of ELEVATE-DP, which found that participants in the GLB program had a greater odds of clinically meaningful weight loss through 24 months and BP control through 12 months, compared to usual care patients. This reversal of effect may be attributable to the fact that the GLB program is open to all Sutter Health patients and patients are specifically referred to the GLB program if they have elevated CMD risk (including those who are overweight/obese, pre-diabetic, or with metabolic syndrome). This may lead patients to choose GLB clinics outside of their typical healthcare utilization patterns. If this self-selection into GLB clinics were true, it would inflate the mean weight and mean BMI values at GLB clinics, making it appear as if the program had no effect, or an effect in the opposite direction than hypothesized. Additionally, because the majority of patients pay out-of-pocket for the GLB program, and costs vary between affiliates, the program may select for specific socio-demographic characteristics. It is also possible that small effects are difficult to detect as the number of patients impacted by the program is small relative to the overall patient population. If outcome trends are volatile, it is possible that the impact of “small” programs with effects that are of a magnitude similar to the volatility of the outcome will be difficult to detect[18].

Our study benefits from several strengths. We utilized a large sample of patients and encounter visits from a diverse pool of patients that Sutter Health services. We used five

years of pre-program data to establish outcome trends in the GLB clinics prior to the program implementation. We chose to use the synthetic control method in order to improve our control group selection. As **Figures 4.1-4.2** showed, a simple average of all control clinics did not closely mimic the outcome trends of the GLB clinics, thus we attempted to construct a synthetic control group that might better approximates these trends. Another central motivation for using the synthetic control method was that it provides researchers with a quantitative tool to select appropriate comparison groups. In our analysis, many control clinics received weight in the synthetic control. **Tables 4.2 and 4.3** make explicit the contribution of each comparison clinic to the counterfactual of interest. Traditional regression analysis fails to provide such a transparent list, as typically all units contribute to the regression fit.

We also acknowledge several limitations in our study. First, we had limited ability to replicate all of the relevant characteristics of the GLB clinics in the control clinics given the available data. Additional clinic-level covariate data, as well as additional years of pre-program data would greatly benefit the ability to construct the best synthetic control possible. Additionally, while the synthetic control method allowed us to capture the impact of the GLB program, it is non-specific and thus does not allow us to test the specific mechanism that led to the change. It is possible that factors or events that we did not control for during the study period, while unrelated to the GLB program, affected our outcomes of interest. To explore this further, future analyses should incorporate techniques such as a negative control, which estimates the effect of an exposure on an outcome that it should be plausibly impact, but which may be affected by a confounding factor[44]. If an effect is still observed, it can be assumed that there is confounding or bias present. Additionally, it is likely that some of the control clinics had their own diabetes prevention or support programs to some effect. If this were the case, the effect of the GLB program may appear to be attenuated (effect is underestimated).

While the core components of the GLB program are similar across sites, variation in how the program was implemented within each affiliate based on existing infrastructure, resources and workflow exist, and may have an impact on the results of our evaluation. In particular, the GLB program was not rolled out to all clinics at the same time; it began in 2010 and was rolled out to more clinics in subsequent years. Lastly, we acknowledge that missing data for our study was not missing at random; not all hospitals within the Sutter system have been using the EpicCare EHR system for all the pre-intervention years. The last hospital to come online with EHR was in 2015. Therefore, data from some hospitals may be incomplete and missing, thus leading us to throw out much of the available data. Additionally, because Sutter Health is an open-network healthcare system, patients may obtain care outside the system, and thus not all clinical effectiveness and healthcare utilization outcomes may be captured. We will assume that any under-reporting or misclassification of these outcomes is non-differential across clinics, which would bias our estimates towards the null.

4.5. Conclusions

Health care systems are highly interested in evaluating the impact of programs on patient outcomes. The synthetic control method provides another tool for evaluation, in which only aggregate level data over multiple years is required. While our results did not indicate that the GLB program had its intended effect, this analysis still provides valuable information on how SCM may be applied in a health-care systems context. Future studies should take into account the open nature of the GLB program. Unlike a law or policy that affects all individuals in the group, programs implemented within a health care system are often voluntary and target a specific population at risk for specific conditions, and thus pose additional methodological considerations that may require extensions to the traditional SCM.

4.6. Figures

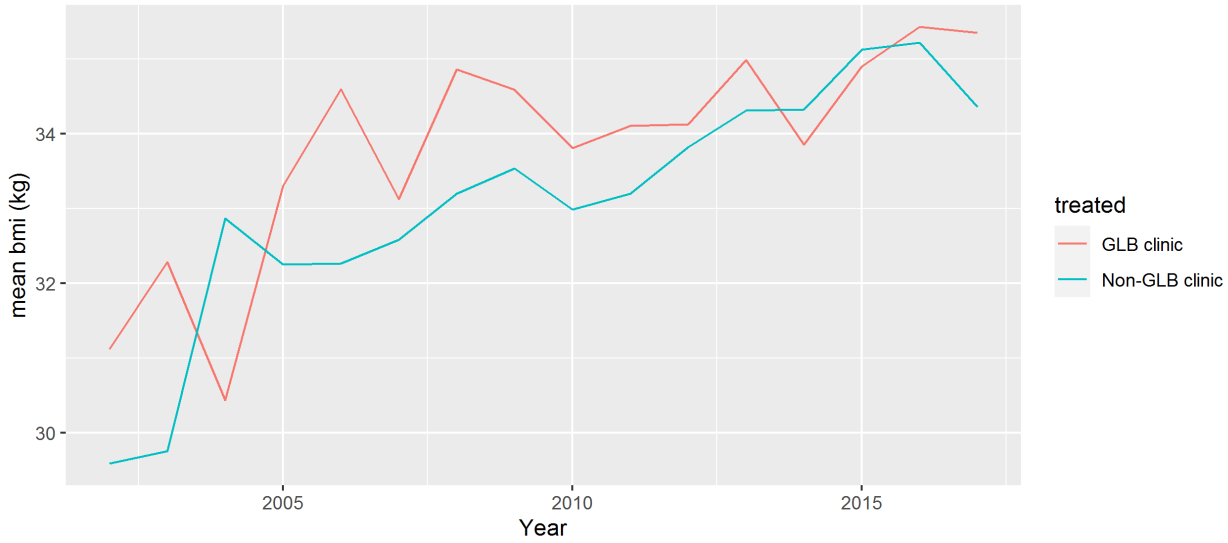


Figure 4.1. Trends in mean weight (kg), GLB clinics vs non-GLB clinics (2002-2017).

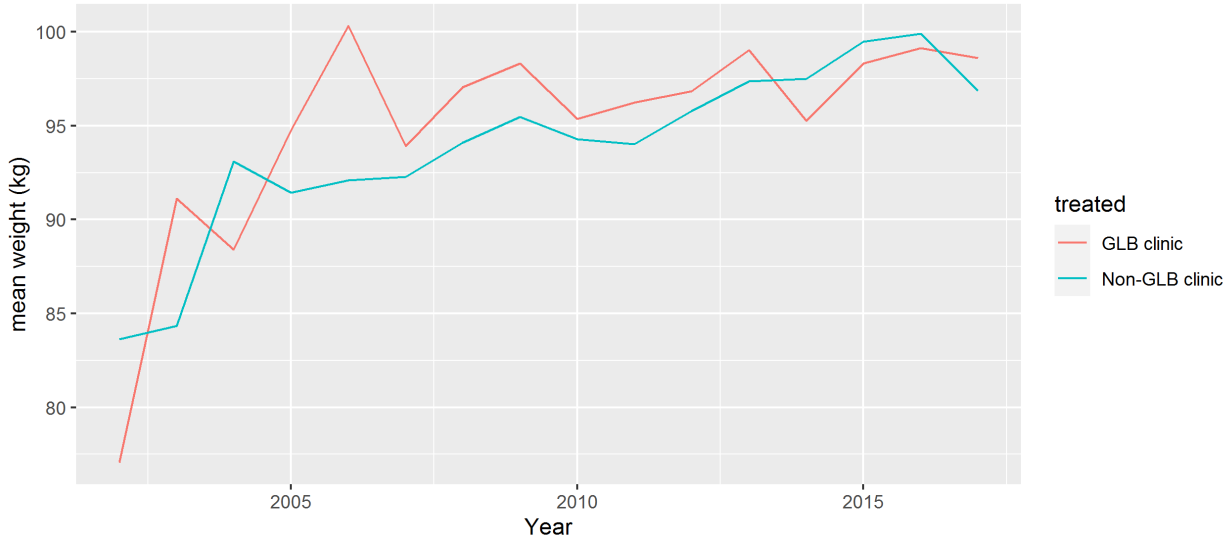


Figure 4.2. Trends in mean BMI, GLB clinics vs non-GLB clinics (2002-2017).

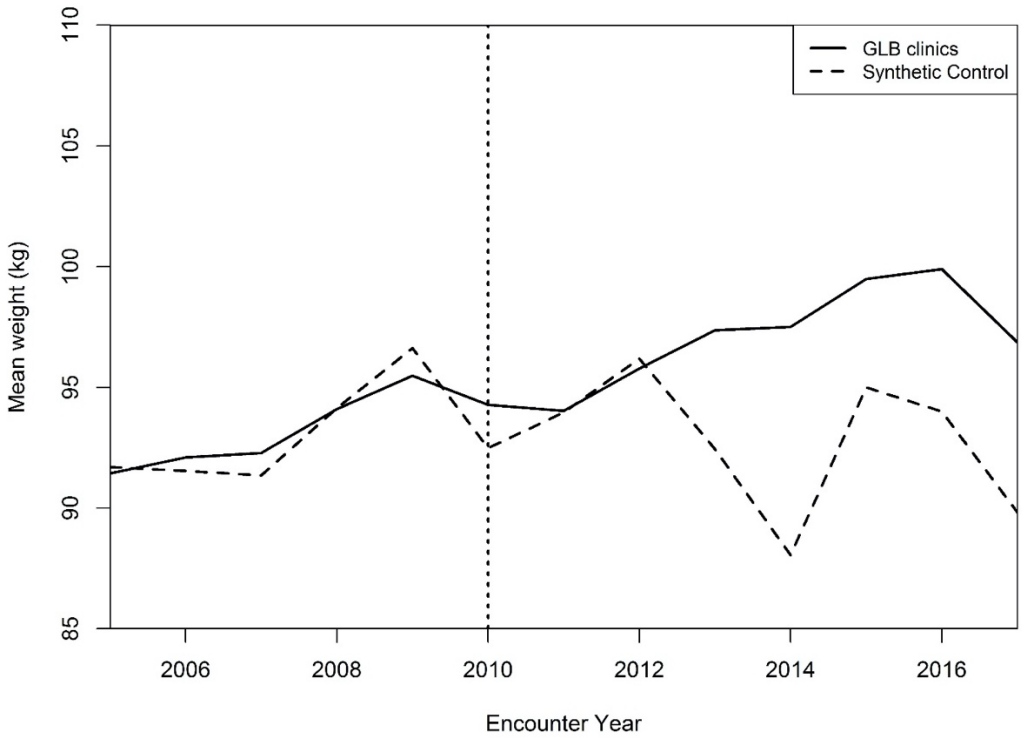


Figure 4.3. Trends in mean weight (kg), GLB clinics vs Synthetic GLB, 2005 – 2017.

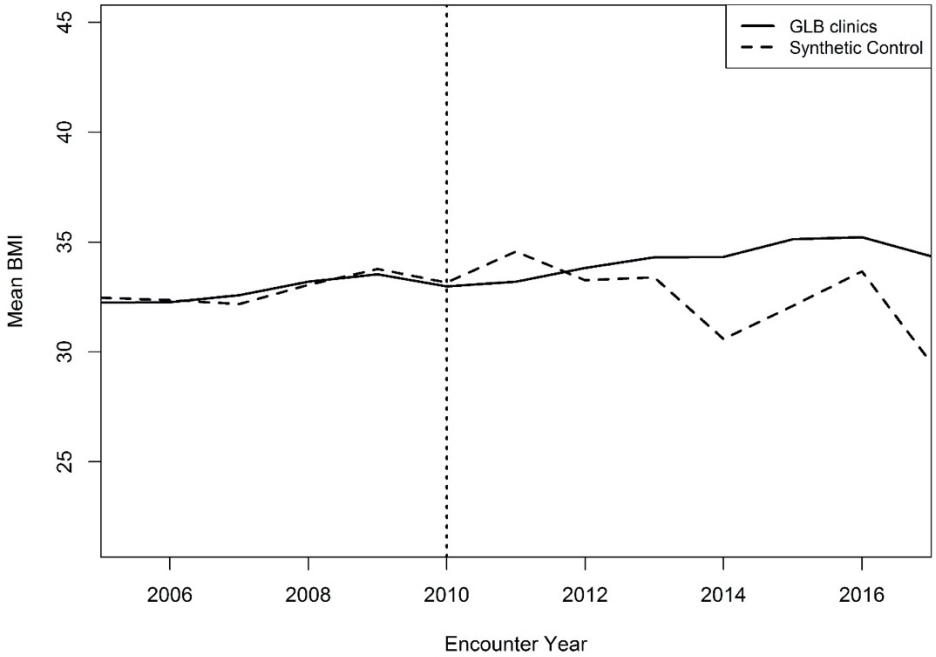


Figure 4.4. Trends in mean BMI, GLB clinics vs Synthetic GLB, 2005 – 2017.

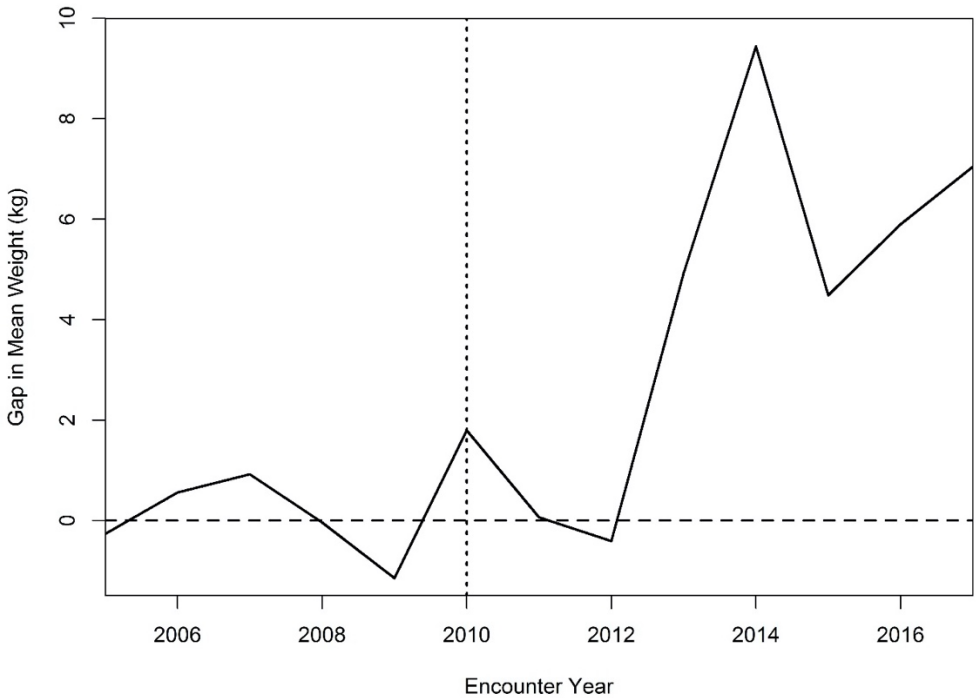


Figure 4.5. Gap in mean weight (kg) between GLB clinics and Synthetic GLB, 2005 – 2017.

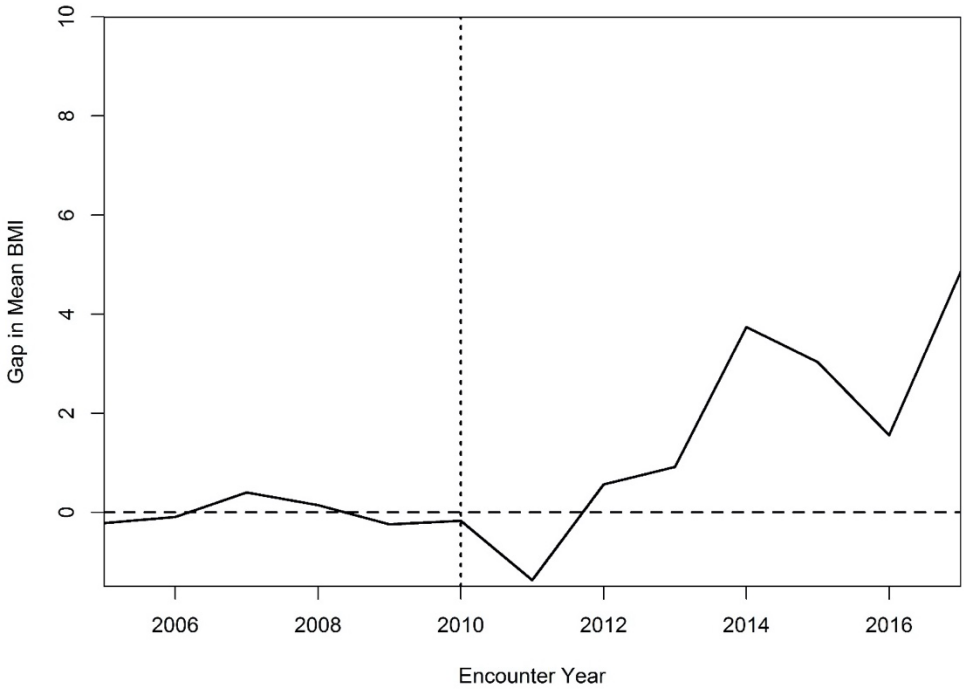


Figure 4.6. Gap in mean BMI between GLB clinics and Synthetic GLB, 2005 – 2017.

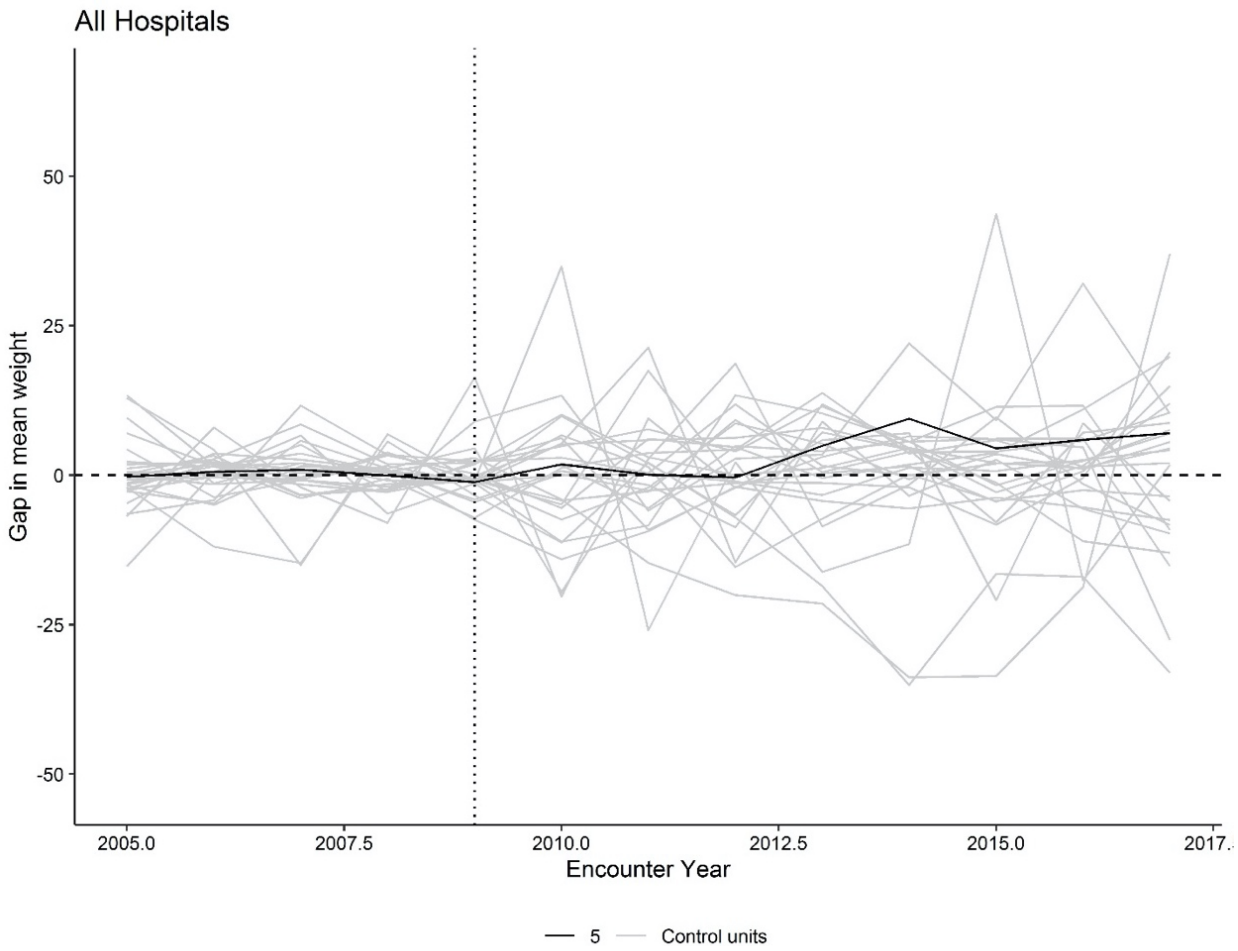


Figure 4.7. Difference in mean weight between control clinics and GLB clinics and their respective synthetic controls, all clinics from 2005 - 2017. The superimposed black line denotes the gap estimated for the actual GLB clinics.

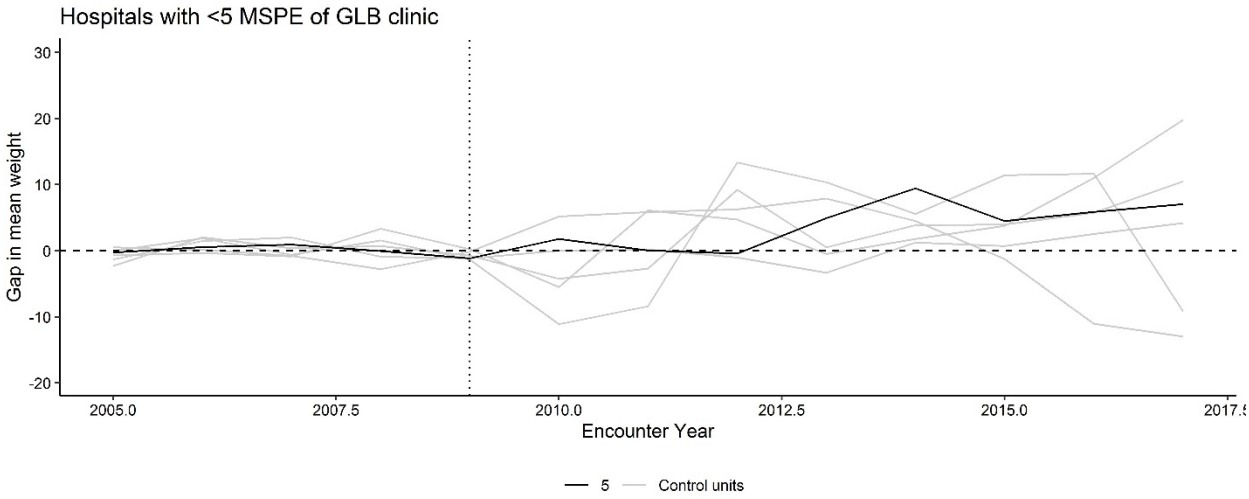


Figure 4.8. Difference in mean weight between control clinics and GLB clinics and their respective synthetic controls, clinics with <5 MSPE of the actual GLB clinics from 2005 - 2017. The superimposed black line denotes the gap estimated for the actual GLB clinics.

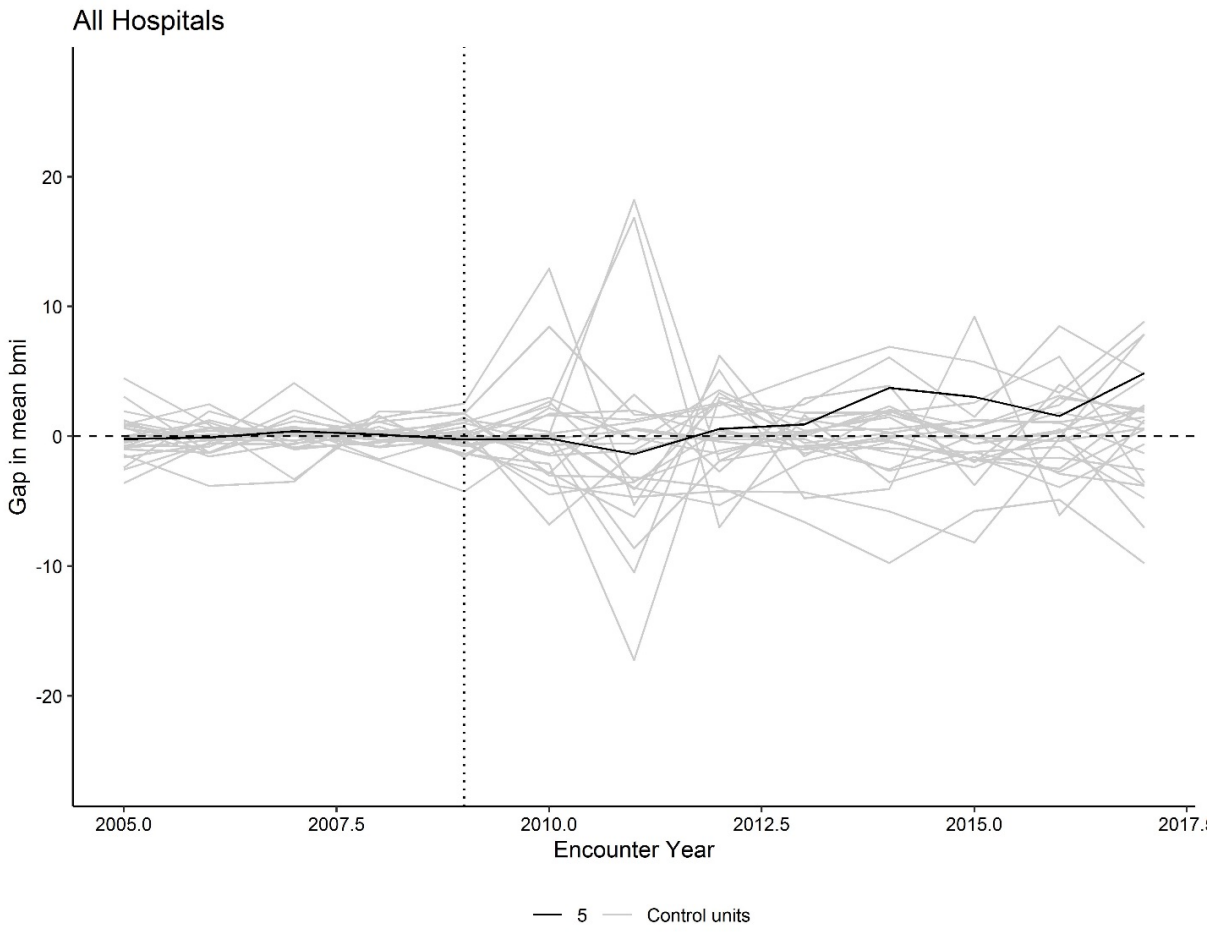


Figure 4.9. Difference in mean BMI between control clinics and GLB clinics and their respective synthetic controls, all clinics from 2005 - 2017. The superimposed black line denotes the gap estimated for the actual GLB clinics.

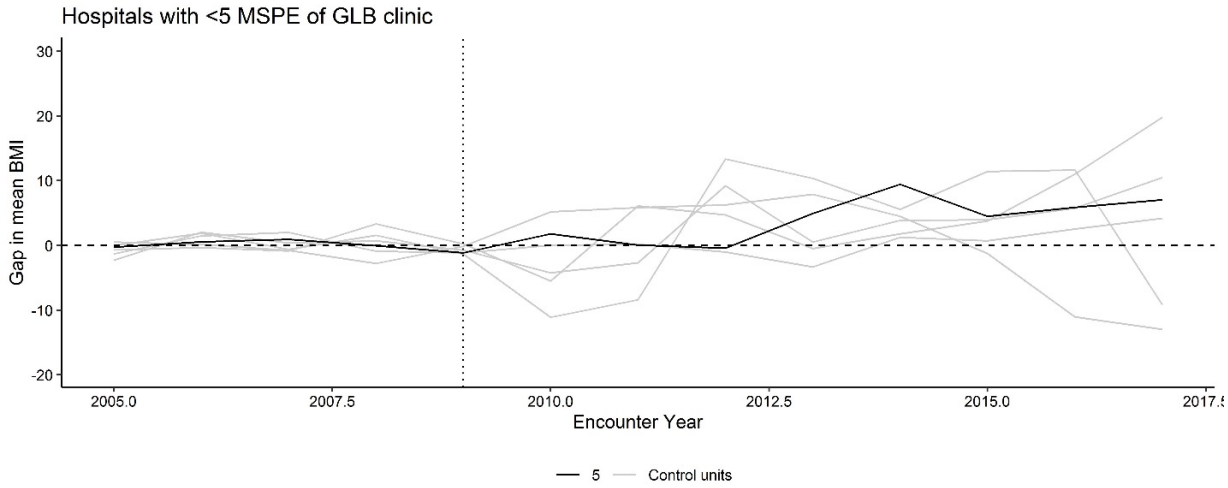


Figure 4.10. Difference in mean BMI between control clinics and GLB clinics and their respective synthetic controls, clinics with <5 MSPE of GLB clinics from 2005 - 2017. The superimposed black line denotes the gap estimated for the actual GLB clinics.

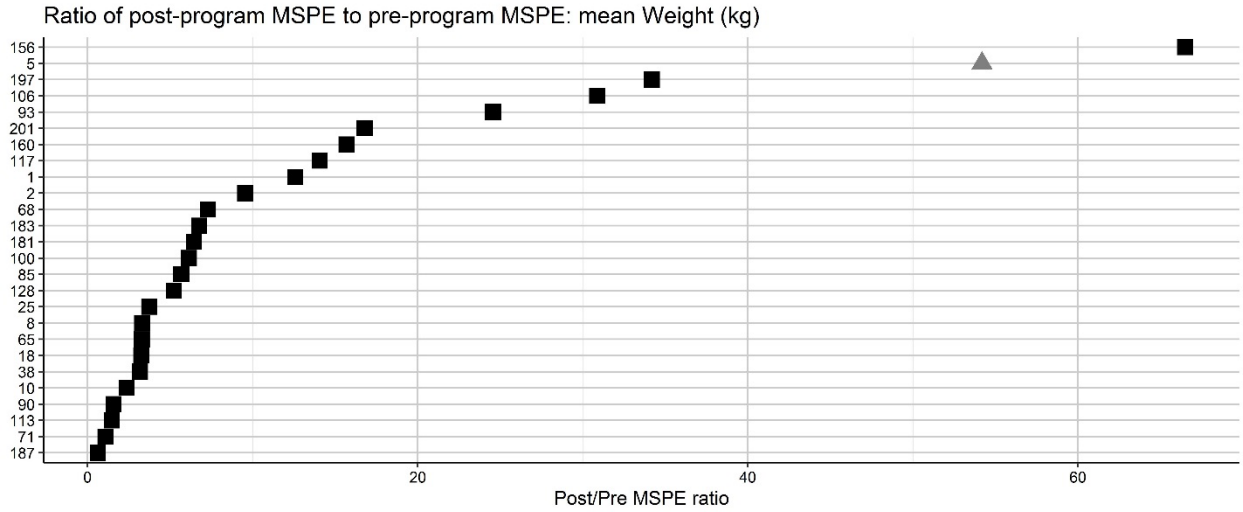


Figure 4.11. Ratios of post intervention MSPE to pre-intervention MSPE for mean weight (kg) for GLB clinics and all the clinics in the control group.

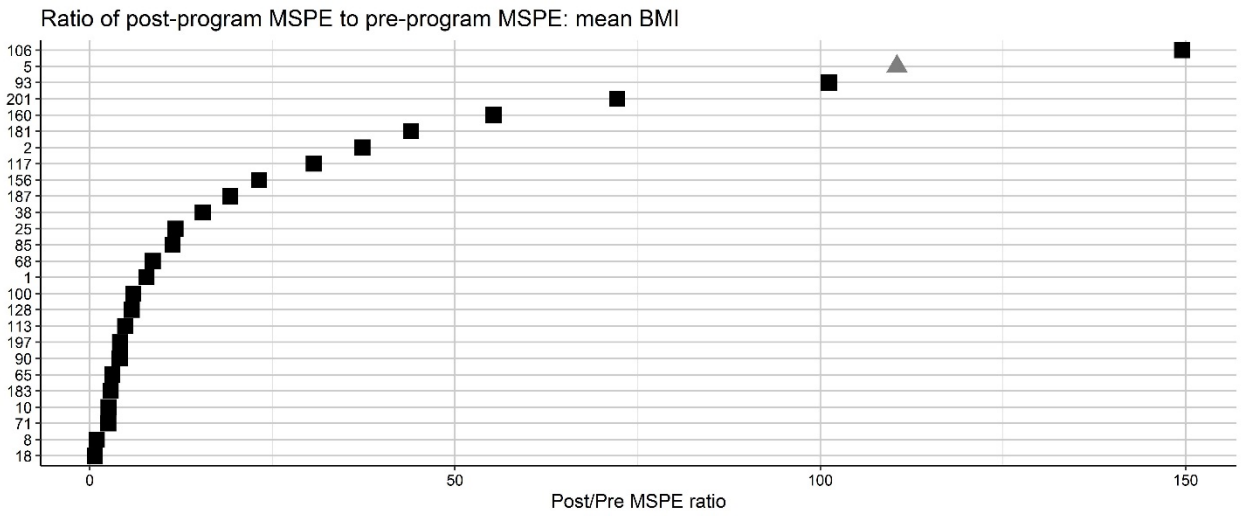


Figure 4.12. Ratios of post intervention MSPE to pre-intervention MSPE for mean BMI for GLB clinics and all the clinics in the control group.

4.7. Tables

Table 4.1. Comparison of mean covariate levels in GLB clinics, its synthetic control, and control clinics.

Predictors	Treated	Synthetic	Sample Mean
Age (mean)	55.004	54.955	53.642
Median Household Income (mean)	82478.872	82867.202	74123.171
Diastolic BP (mean)	76.528	76.774	77.717
Systolic BP (mean)	127.700	129.578	128.584
Male	0.457	0.455	0.440
Medi-Cal	0.006	0.007	0.003
Medicare	0.064	0.090	0.035
Other	0.533	0.585	0.883
PPO	0.186	0.165	0.043
HMO	0.211	0.152	0.037
Smoker	0.276	0.279	0.289
Weight	93.081	93.075	95.409
BMI	32.765	32.767	33.643

Table 4.2. Clinics included in the synthetic control along with their weights, mean weight (kg).

Clinic	Weight
1	0.010
10	0.005
100	0.001
106	0.001
113	0.002
117	0.342
128	0.186
156	0.026
160	0.000
18	0.001
181	0.001
183	0.000
187	0.001
197	0.240
2	0.001
201	0.002
25	0.003
38	0.011
65	0.003
68	0.000
71	0.001
8	0.067
85	0.001
90	0.001
93	0.094

Table 4.3. Clinics included in the synthetic control along with their weights, mean BMI.

Clinic	Weight
1	0.010
10	0.127
100	0.022
106	0.012
113	0.004
117	0.010
128	0.172
156	0.000
160	0.000
18	0.003
181	0.007
183	0.000
187	0.008
197	0.201
2	0.007
201	0.007
25	0.008
38	0.080
65	0.018
68	0.112
71	0.006
8	0.139
85	0.006
90	0.016
93	0.025

Table 4.4. Mean squared prediction error (MSPE) in the pre-program period.

Outcome	MSPE
Mean weight (kg)	0.51
Mean BMI	0.06

Chapter 5 - Conclusion

5.1. Key Findings

Findings from Chapter 2: Evaluation of the Advancing Health Equity Asthma Program Utilizing Propensity Score Matching and the Synthetic Control Method

Objective: to evaluate the impact of the AHE asthma program on 30-day and 90-day return to the ED for any reason and for breathing difficulty related difficulties, from January 1, 2019 to February 29, 2020.

In 2016, Sutter Health's ABSMC found that 649 patients utilized the ED at total of 877 times for asthma-related reasons, resulting in an HEI value of 1.5. This was largely driven by a disproportionate number of elderly Black/AA patients compared to other racial groups. Because of this, the AHE Asthma program was implemented to address the specific needs of Black/AA patients who utilized the ED for asthma-related reasons. The program brings culturally appropriate community based primary care, asthma education about the disease and medication self-management, and high-touch and high-tech real-time counseling services. We evaluated the program using propensity score matching and synthetic controls in order to assess individual-level and hospital-level impact of the program.

Based on the propensity score matched analysis, we found that the odds of returning to the ED for breathing difficulty related reasons among program participants was 1.28 (0.99, 1.64) times the odds among non-participants. This effect was mirrored when stratified by COPD status. For 30-days returns to the ED for any reason, we saw results that were consistent with the direction of effect we had hypothesized, however the estimates were not significant. The odds ratio for 30-day return to the ED for any reason was 0.85 (0.62 – 1.16) comparing program participants and control patients. The odds ratio associated with 30-day returns for any reason was 0.63 (0.34 – 1.20) and 0.89 (0.61 – 1.29) for those with a history of COPD and those without a history of COPD, respectively. For 90-day return to the ED for any reason, program participants had 1.06 (0.83 – 1.06), 0.91 (0.54 – 1.54), and 1.09 (0.82 – 1.45) times the odds compared to control patients for the whole cohort, those with COPD and those without COPD, respectively.

Using SCM, we did not find a significant effect of the program on reducing the number of 30-day and 90-day returns to the ED for any reason. We had hypothesized that the introduction of the AHE Asthma program would allow program participants to better manage their asthma symptoms and thus reduce the burden of utilizing the ED. However, contrary to our hypothesis, we found that outcome trends at both ABSMC sites were consistently higher than their synthetic control during the post-intervention period. At ABSMC Ashby, there was an average increase during the program of

approximately 0.5 returns and 1.2 returns for 30-day and 90-day returns respectively. At ABSMC Summit, the average increase during the program was about 0.72 returns and 2.29 returns for 30-day and 90-day returns respectively. Based on the results of placebo tests, the gaps for both ABSMC Ashby ABSMC Summit for either outcome did not appear to be unusually large, even when restricting the placebo tests to only those hospitals that had <2 times the MSPE of the treated unit. Examining the ratio of post-program MSPE to pre-program MSPE further supported this conclusion as the ratio for the ABSMC hospitals did not stand out from ratios for the control hospitals. However, for both outcomes, outcome trends at ABSMC Ashby and ABSMC Summit did not appear to be well approximated by their respective synthetic controls, evidenced both visually and by the relatively large MSPE, and thus the results during the post-intervention period are not reliable.

Findings from Chapter 3: A Propensity score and Synthetic Control Approach to Evaluating a New Protocol for Patients at Risk for Alcohol Withdrawal Syndrome

Objective: to assess the impact of a new protocol for treating individuals at risk for developing Alcohol Withdrawal Syndrome on leaving against medical advice, risk of ICU admission, and hospital length of stay from April 1, 2019 to March 2, 2020.

Current practice for treating patients who are at risk for developing Alcohol Withdrawal Syndrome varies greatly both within the Sutter Health system and the overall medical community. Historically, the protocol at EMC is to assess patients using the CIWA score and treat symptoms of AWS with benzodiazepines. While the CIWA tool has been shown to have high validity and inter-rater reliability, it requires the clinical team to wait for the patient to show signs and symptoms of withdrawal prior to starting treatment. However, as recent studies have shown that delirium increases hospital acquired conditions, length of stay, cost of care, death and long-term cognitive impairments, the question of why treatment can't be initiated sooner to prevent AWS symptoms before they begin is pertinent. Thus, in April of 2019, EMC implemented a new treatment protocol (ETOH-P) for patients at risk of developing AWS. The program consists of two protocols, one aimed at preventing AWS (prophylaxis protocol) and the other at treating AWS earlier in its course with targeted pharmaceuticals (active withdrawal protocol). We utilized propensity score matched methods and a synthetic control analysis to assess the impact of the ETOH-P program on risk of leaving AMA, average hospital length of stay and risk of ICU admission at both the patient-level and hospital-level, respectively.

The average hospital length of stay among the whole matched cohort, ETOH-P patients and control patients was 7.16 days, 7.32 days and 7.13 days respectively. 10.3% of ETOH-P patients left the hospital AMA compared to 7.3% of control patients who left AMA. 109 ETOH-P patients (35.2%) had an ICU admission, compared to 435 control patients (35.1%).

We found that patients who received the ETOH-P protocol had 1.47 (0.96, 2.27) times the odds of leaving AMA compared to those who did not receive the protocol, however, this was not statistically significant. For ICU admissions, patients who received the

ETOH-P protocol had 1.00 (0.77, 1.31) times the odds of ICU admissions compared to those who did not receive the protocol, though also not statistically significant. Patients on the ETOH-P protocol had, on average, 0.19 (0.11, 0.26) more days in the hospital, compared to those not on the ETOH protocol, and this difference was statistically significant.

Overall, the pre-intervention trends for leaving AMA and ICU admission did not appear to be well approximated by their respective synthetic controls; in fact, the EMC curve crosses at multiple points with the synthetic control curve during the pre-intervention period. It does appear that the synthetic control for ICU admissions provides a better approximation of actual trends. The EMC curve crosses at multiple points with the synthetic control curve during the pre-intervention period for all three outcomes. The differences in the proportion of patients who leave AMA between the actual EMC and its synthetic control indicate that there was an average increase during the program period of approximately 0.0003. There was an average increase in the proportion of patients admitted to the ICU of 0.0012 and an increase in the average hospital length of stay of 2.03 days.

Based on the placebo tests, the estimated gap for EMC did not appear to be unusually large compared to the distribution of gaps for the other hospitals in the donor pool, for any of the outcomes. The results from the ratio of post- to pre-intervention MSPE and p-value were consistent with the placebo plots, which indicated that the effect of the ETOH-P program on leaving AMA and ICU admission was not significant, but hospital LOS was significant.

Findings from Chapter 4: A Synthetic Control Approach to Evaluating the ELEVATE-DP Study

Objective: to evaluate the impact of the Group Lifestyle Balance (GLB) program on mean weight and mean BMI from 2010 to 2017.

The landmark Diabetes Prevention Program (DPP) randomized clinical trial found that an intensive lifestyle intervention program was effective at reducing the incidence of diabetes by more than half. Sutter Health's real-world adaptation of the LCP is the GLB program, consisting of a year-long group-based curriculum modeled off the original DPP. The primary evaluation of the effectiveness, adoption, implementation and maintenance of the GLB program has previously be conducted; an EHR-based, propensity score matched analysis was performed and found that compared with usual-care patients, participants in the program had a greater odds of attaining clinically meaningful weight loss at 12-months and 24-months of follow-up, as well as a greater odds of blood pressure control at 12-months of follow-up. In this analysis, we were interested in whether this same effect would be evident at the clinic-level, and thus we used synthetic controls to study the impact of the GLB program on mean weight and mean BMI from 2010 to 2017.

Overall, the pre-program mean weight and mean BMI levels appeared to be well approximated by the synthetic control. The differences in mean weight between the actual GLB clinics and its synthetic control from 2010 to 2017 indicate that there was an average increase in mean weight during the program of approximately 4.2 kg. The permutation test suggests that this effect is not significant compared to other clinics when comparing to all control units, as well as just those control units with an MSPE < 5 times that of an actual GLB clinic. Actual trends of mean BMI remain relatively steady across the study period and start to diverge from the synthetic control around 2012. The actual mean BMI trend was elevated over the synthetic control for most of the post-program period. The differences in mean BMI between the actual GLB clinics and its synthetic control from 2010 to 2017 indicate that there was an average increase in mean BMI during the post-program period of approximately 1.64. Permutation tests indicate that this was also not significant as the observed effect for the actual GLB clinics was not unusually large compared to the distribution of effects from other clinics.

5.2. Discussion

SCM was first developed and formalized by Abadie, Diamond, and Hainmueller[37] to study the impact of a 1988 large-scale tobacco control program that was implemented in California and Abadie and Gardeazabal[2] to investigate the economic effect of terrorist conflict in the Basque Country of Spain. While the method originated from comparative case studies in economics, in which an intervention effect is assessed by comparing the aggregate-level outcomes of the treated unit to a group of units that are similar, but unaffected by the treatment. However, it is often difficult to identify a single unit that is the best comparison group, and comparative case studies lack a formalized, systematic method for selecting comparison units. Since then, SCM has been used as an evaluation tool to study programs and interventions at the group-level in many different settings and disciplines. In recent years, SCM has been applied to study key policy issues around tobacco control and smoking policies[74–77], gun control and right-to-carry laws[78–81], health care delivery and insurance reform programs[82–89], carbon emissions and deforestation[90–93], and even drought[40]. Synthetic controls have also been used to study important recent events such as the legalization of same sex marriage[94], vaccine introductions[95–97], marijuana[98,99], the adoption of open data among research publications[100], and the COVID-19 pandemic[8,10,13,101].

Boutell et al. (2018)[102] surveyed the use of SCM in health research settings and provided a very valuable and thorough introduction of SCM to public health researchers. They found 38 health-related studies utilizing SCM to study topics spanning health finance and health systems reform, health industry reform (ban on trans fat, mobile phone bans, alcohol licensing hours, etc.), taxation policies (tax on sugar-sweetened drinks, cigarettes, etc.), nutritional interventions, and health welfare reforms[103]. In general, SCM has proven to be a valuable addition to the program evaluation toolbox, particularly in the observational study design setting, however the method is underused

in public health research. Some key assumptions are made in all of these health-related applications of SCM. First, although weights are systematically calculated by the method to ensure the most appropriate synthetic control, it is assumed that the treated unit(s) is sufficiently similar to the control units in the donor pool. Second, the method assumes that there is no contamination or spillover of the intervention into any potential control units. Third, there must be no external shocks to the pool of control units. As much of the existing literature notes, and as we discuss below, these assumptions must be carefully considered within the health system setting.

A common theme highlighted throughout many of these prior applications is that synthetic controls can be readily applied when there is aggregate-level data available, particularly if it is publicly available; for example, an assessment of the impact of a sugar-sweetened beverage tax on employment used monthly employment count data from the Bureau of Labor Statistics[104], and a study looking at the effects of border shutdowns on the spread of the COVID-19 outbreak used confirmed case counts from the World Health Organization (WHO) situation reports[13]. While we did not find any applications of SCM to study asthma or AWS programs, we found one paper that examined the impact of sugar and processed food imports in 172 countries on average BMI. The study found that overtime, imports of sugar and processed food was associated with an increase in average BMI in Fiji compared to the synthetic control group[105].

Compared to traditional regression methods, SCM offers several advantages. Synthetic control estimators avoid extrapolation beyond the data because the weights are restricted to be positive and sum to one. Regression weight in comparison may be outside the [0,1] range, allowing for extrapolation outside the bounds of the data. Because the synthetic control is a weighted average of potential control units, readers can assess for themselves, the relative contribution of each control unit to the synthetic control. Those with expert knowledge of the potential control units can quickly assess the validity of the synthetic control based on the units that are given weight. Additionally, the audience can judge for themselves similarities and differences in outcome and predictor values between the treated unit and the synthetic control. In further contrast with regression methods, construction of the synthetic control does not require access to post-intervention outcomes. Synthetic control weights can be calculated using just pre-intervention data before even seeing any post-intervention data. This is advantageous as it allows you to make decisions about study design, such as identifying the most appropriate comparison group without knowledge of how they might affect the conclusions of the study[18]. Furthermore, SCM provides a nice visual representation of the impact of the intervention and how it varies over time, as well as a clear visualization of the actual discrepancy between the treated unit and the convex combination of control units, something that alternatives such as propensity score methods do not provide. Plots are produced that display what the observed outcome looks like compared to what would be expected in the absence of the intervention; the sharp divergence of the outcome trends can provide a strong visual aid in telling a story about program impacts.

This dissertation describes three separate applications of the synthetic control method to evaluate 1) an outpatient asthma program in which participants are recruited into the program by a program coordinator based on eligible characteristics, 2) an inpatient hospitalization protocol for patients who present to the ED who are at risk for AWS, and 3) an outpatient diabetes program available to all Sutter Health patients who are at cardiometabolic risk. In doing so, this dissertation contributes to the literature by being the first study, to our knowledge, to utilize SCM to evaluate an inpatient or outpatient health system program. While the focus of this dissertation is the synthetic control analysis, we also performed an individual-level analysis of the asthma and ETOH-P programs utilizing propensity score matched methods; a previous study undertook a propensity score matched analysis of the GLB program. In conducting these analyses, we learned several key considerations for utilizing SCM as an evaluation tool within this setting. We divide these into data acquisition and data preparation considerations and methodological considerations.

Data acquisition and data preparation

1. **SCM requires complete aggregate-level data for outcomes and predictors during the pre-intervention period.** As chapters 2-4 demonstrated, the SCM method requires aggregate-level data for the outcomes and predictors for the treated unit and all potential control units. However, even more important is having sufficient pre-intervention data for all outcomes and predictors. This is because the credibility of a synthetic control estimator depends largely on its ability to reproduce the outcome trajectory of the treated unit during the pre-intervention period. Abadie, Diamond and Hainmueller (2010) discuss this by showing that bias in the synthetic control estimator is bounded by a function that is inversely proportion to the number of pre-intervention periods, therefore necessitating a large pre-intervention period.
2. **Sufficient pre-intervention data are necessary.** We found that in order to construct a synthetic control, complete outcome data was required for all time points during the pre-intervention period. Because of this, we ended up throwing out large portions of the data; this was particularly true for the GLB program evaluation where we discarded multiple years of data due to too much missingness. Furthermore, additional years of data for all three applications would have strengthened the match during the pre-intervention period as it may have shed light on any seasonal trends. We utilized data from Sutter Health's EHR system, which began onboarding of hospitals to their EHR system in 2012, but wasn't fully implemented at every hospital until 2015, thus hospital data availability prior to the EHR system is fragmented and sparse. Future uses of SCM data within a health-care system should take into consideration the fact that data availability may pose a challenge to obtaining a close fit during the pre-intervention period.

3. **Sufficient post-intervention data are also necessary.** Programs implemented within health systems are usually not introduced at one point in time, in fact they are generally implemented over a period of time, either because they are being

rolled out to multiple clinics (such as with GLB) or because the program slowly ramps up as logistics and kinks are sorted out. Thus, unlike prior application of SCM that evaluate a law or policy that goes into effect at one specific point in time, health system programs do not have one specific start point. Because of this, it is all the more important to have post-intervention data sufficiently into the future to identify program effects that are expected to take some time. In our evaluations of the asthma program and ETOH program, both had short post-intervention periods of around one year. It is possible that the outcomes we chose for these evaluations were too "downstream" of the intervention to be detected within such a short period of time, such that the mechanism between intervention and the outcomes of interest were too variable and subject to many outside influences. A longer post-intervention window would allow for a more complete picture of the intervention trend.

Methodological considerations

4. **While SCM is motivated by the use of a data-driven approach to control selection, there still existing subjective decisions during the data preparation process.** For example, we chose to aggregate all the GLB clinics together into one "treated" unit. As the synth() procedure could only accept a complete and balanced dataset, we had to make subjective decisions based on data availability and amount of missingness; these decision included which years of data to use and which control units to keep within the donor pool.
5. **Size of the effect and volatility of the outcome.** As made evident throughout this dissertation, SCM is an evaluation tool that can be used to estimate the effect of an intervention on an aggregate level. In most cases, the intervention is implemented at only one unit or a few units. Thus, it is possible that the effects of some interventions are too small to be detectable at the aggregate level. This is particularly true if the outcome of interest is highly volatile, and the trend of the outcome is highly variable over time. Intervention effects that are a smaller or similar magnitude to the volatility of the outcome will be indistinguishable from the normal variability of the outcome[18]. As we saw with outcomes in the asthma evaluation (30-day and 90-day return to the ED) and ETOH-P evaluation (proportion of patients leaving the hospital AMA, proportion of patients who had an ICU admission, and the average hospital length of stay), the trends fluctuated during the pre-period, crossing over at multiple time points with the synthetic control trend line. It could be that changes to these outcomes were too small to detect at the hospital-level; one possibility being that the uptake of the program was small relative to the overall population at those hospitals.
6. **No interference or spillover effects.** The organization of a health care system such as Sutter Health appears to be a seemingly suitable setting to apply SCM, as it is an integrated network of care centers with various hospital locations; when programs are implemented only at one hospital and not at others within the same health care system, SCM should be considered as a tool for program evaluation. However, it is important to carefully consider the potential control hospitals to be included in the donor pool. Hospitals with very dissimilar

characteristics to the treated hospital should be excluded from the possible donor pool. This is because, although weights are restriction to avoid extrapolation, it is still possible for interpolation biases to exist if the synthetic control has to average away large differences in order to closely match the treated unit[18]. Additionally, control hospitals should be eliminated from the donor pool if they themselves implemented a similar program during the study period. If these hospitals were to be included in the donor pool, it could attenuate the effect of the program we observe in the treated unit. Likewise, outcomes in control hospitals that are affected by the intervention should also be disregarded as this "spillover" effect could also bias the results towards the null, making it appear that the program is less effective than it actually was.

7. **Patients are free to seek care where they want.** One final consideration when using SCM within a health system such as Sutter Health is that patients are free to choose hospital locations from which to seek care. If the program is highly successful at having a beneficial effect on its targeted outcomes, patients may choose to seek care at the hospital where the new program is implemented, even if it is not the hospital they normally go to. We believe that this was the case for the GLB program. As DPP was a landmark study that gained national attention and awareness, once the GLB program began at Sutter Health, it is likely that patients heard about the program and actively choose to seek care at a GLB program clinic. As patients targeted for GLB are those who are overweight/obese and at high risk for cardiometabolic events, it may have appeared that our outcomes of interest increased during the program period, when actually average weight and BMI at GLB program clinics was actually increasing because patients were self-selecting into the program. Furthermore, it is possible that patients who choose to enroll in these programs are systematically different from those who do not participate. This form of self-selection bias is highly likely and may obscure any true effects of the intervention
8. **SCM is nonspecific.** While SCM allows us to capture the impact of an intervention on our outcomes of interest, it does not test the specific mechanism for how the intervention led to change, thus SCM is non-specific and does not allow us to distinguish between co-occurring events that could affect the outcome. One possible way in which to address this would be to use a series of negative controls, an outcome that the intervention should not plausibly impact, but which may be affected by a confounding factor. If an effect of the intervention is observed when using a negative control, it can be assumed that there is some level of confounding or bias present that needs to be taken into account.

Despite these issues to consider, the application of SCM within Sutter Health is promising and benefits heavily from understanding the context of the program. The impact of health-system programs on patient outcomes has a high potential to be mediated through a range of individual, social, cultural and organizational factors. Failure to account for these contextual effects makes it difficult to determine whether nonsignificant effects of the program on patient outcomes is truly due to an ineffective intervention or other reasons. In our SCM analysis, we were not able to control for these complexities; in most cases, we had data on sex, age, race and ethnicity, type of health

insurance and limited medical history. Additional data on patients that might affect health-seeking behaviors as well as contextual program factors may help to shed light on whether or not the intervention was truly ineffective or not. Furthermore, as a health-care system is in essence a nesting of multiple levels that interact with each other (patients within providers, within hospitals, within neighborhoods, etc.) SCM may not be the most appropriate method to evaluate these programs. Instead, multilevel modeling should be considered as a potential alternative, as multi-level statistical models account for the nested and clustered structures present in health systems[106]. Additionally, the use of mixed-methods is also highly appropriate for evaluating health-care system programs, as it would add more to the understanding of specific mechanisms for change and how variations in program implementation might impact the evaluation results.

One of the primary issues in our applications of SCM was the difficulty in obtaining pre-intervention fit in the outcome trends between the treated unit and the synthetic control. While we discuss above that selecting appropriate controls for the donor pool may improve this fit, there have also been very recent methodological extensions of SCM that address this issue. In 2019, Ben-Michael et al. proposed the augmented synthetic control method (ASCM), which can be applied in settings where good pre-intervention match between the treatment unit and the synthetic control is infeasible. Ben-Michael et al. uses a ridge-regularized linear regression as the outcome model to directly control pre-intervention fit while relaxing the non-negative weight restriction that the original SCM uses[107]. There have been many additional advances related to SCM that are discussed in methodological detail elsewhere[18], but we highlight a select few here. Abadie and L'Hour (2019)[108] discuss situations in which there is not one unique solution to finding the synthetic control that best reproduces the outcome of the treated unit. In such cases, a synthetic control estimator that penalizes the pairwise discrepancies between characteristics of the treated units and characteristics of the donor pool units may be used. This extension is useful in situations where there are multiple treated units, such as our GLB program, and there may not be one unique solution that minimizes the distance between treated and control units. Doudchenko and Imbens (2017)[109] propose a generalization of SCM that uses elastic net regression such that weights can be negative and not sum to 1, allowing for a permanent additive difference between the treated unit and the synthetic controls, similar to a DiD approach. Lastly, Arkhangelsky et al. (2019)[110] present a new perspective on SCM using a weighted least squares regression estimator with time fixed effects and weights for both unit and time, termed a synthetic difference in differences (SDID) estimator, and can also be generalized to cases with multiple treated units and multiple treated periods.

Lastly, this exercise in the application of SCM within a health care setting reminds of the importance of reproducible science and transparency. The recent narrative of the "reproducibility in science crisis" has sparked increased efforts across all fields to make research more transparent and results reproducible[111]. The modern approach to incorporating reproducibility into the scientific process incorporates steps such as registration of protocols and pre-analysis plans, internal replication with blinded analyses and sharing data and code publicly at the time of publication[112]. As previously

mentioned, the development of SCM stemmed from the need for a formalized systematic approach to choosing comparison units in comparative case studies. As such, SCM may be a champion for reproducibility and transparency efforts as it allows researchers to clearly present which units have been selected to contribute to the synthetic control, along with their respective weights. This may be determined beforehand, before ever seeing any post-intervention outcomes, and can be specified in a pre-analysis plan that may be registered before conducting any analyses or drawing conclusions. Additionally, much of health systems research involves confidential patient data, and thus individual-level datasets cannot be easily shared. SCM is a valuable tool in this setting as it utilizes aggregate-level data; researchers may remove any personal-identifying information from the datasets and thus share data and code at the time of publication. Efforts such as these may help to bolster confidence that health programs implemented within health care systems are improving patient outcomes, and thus improve reproducibility and transparency within the field.

5.3 Conclusions

The analyses presented in this dissertation are, to our knowledge, the only applications of SCM to evaluate health programs within an integrated health-care delivery system like Sutter Health. We sought to understand if SCM could be utilized as an evaluation tool to study health programs targeted at specific disease condition. While we did not find a significant effect of any of the programs when utilizing SCM (or traditional propensity score matching methods), we did gain some key takeaways for future applications. Primarily, we learned that there are significant data requirements as well as methodological considerations to contemplate prior to undertaking an SCM analysis. Sufficient aggregate-level outcome and predictor data for the treated units and pool of control units is necessary during the pre-intervention period as well as post-intervention period. Within health care delivery systems, these are important considerations as often 1) multiple EHR systems are used or implemented at different times, resulting in dis-jointed data or insufficient time series data over a long period of time, and 2) implementation of the program occurs over a period of time, potentially resulting in delayed effects, if any. We had to make decisions about which units and time periods to include in the analysis, resulting in areas of subjective decision-making. Methodologically, outcome volatility, interference and spillover between units, patients' path to seeking care and the mechanism of change should be considered beforehand in the context of a health care system. We recommend continued use of SCM within Sutter Health; with the above lessons in hand, we believe SCM is an efficient evaluation method that provides valuable results for hospital-level decision making.

References

1. Athey, S. & Imbens, G. W. The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives* **31**, 3–32 (2017).
2. Abadie, A. & Gardeazabal, J. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review* **93**, 113–132 (2003).
3. Abadie, A., Diamond, A. & Hainmueller, J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *null* **105**, 493–505 (2010).
4. Kreif, N. *et al.* Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units. *Health Economics* **25**, 1514–1528 (2016).
5. Yu, V. *et al.* Hospital admissions after vertical integration of general practices with an acute hospital: a retrospective synthetic matched controlled database study. *Br J Gen Pract* **70**, e705–e713 (2020).
6. Bauhoff, S. The effect of school district nutrition policies on dietary intake and overweight: a synthetic control approach. *Econ Hum Biol* **12**, 45–55 (2014).
7. Goin, D. E., Rudolph, K. E. & Ahern, J. Impact of drought on crime in California: A synthetic control approach. *PLOS ONE* **12**, e0185629 (2017).
8. Cole, M. A., Elliott, R. J. R. & 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* **76**, 553–580 (2020).
9. Poppe, A. Impact of the Healthcare System, Macro Indicator, General Mandatory Quarantine, and Mask Obligation on COVID-19 Cases and Death in Six Latin American Countries: An Interrupted Time Series Study. *Front Public Health* **8**, 607832 (2020).
10. Mitze, T., Kosfeld, R., Rode, J. & Wälde, K. Face masks considerably reduce COVID-19 cases in Germany. *PNAS* **117**, 32293–32301 (2020).
11. Dave, D., McNichols, D. & Sabia, J. J. The contagion externality of a superspreading event: The Sturgis Motorcycle Rally and COVID-19. *Southern Economic Journal* **87**, 769–807 (2021).
12. Bennett, M. All things equal? Heterogeneity in policy effectiveness against COVID-19 spread in chile. *World Development* **137**, 105208 (2021).
13. Kang, N. & Kim, B. The Effects of Border Shutdowns on the Spread of COVID-19. *J Prev Med Public Health* **53**, 293–301 (2020).
14. Cho, S.-W. (Stanley). Quantifying the impact of nonpharmaceutical interventions during the COVID-19 outbreak: The case of Sweden. *The Econometrics Journal* **23**, 323–344 (2020).
15. Clarke, G. M., Conti, S., Wolters, A. T. & Steventon, A. Evaluating the impact of healthcare interventions using routine data. *BMJ* **365**, l2239 (2019).

16. Diamond, A. & Hainmueller, J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program AU - Abadie, Alberto. *Journal of the American Statistical Association* **105**, 493–505 (2010).

17. Abadie, A., Diamond, A. & Hainmueller, J. Comparative Politics and the Synthetic Control Method. *American Journal of Political Science* **59**, 495–510 (2015).

18. Abadie, A. Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature* doi:10.1257/jel.20191450.

19. Loftus, P. A. & Wise, S. K. Epidemiology of asthma. *Curr Opin Otolaryngol Head Neck Surg* **24**, 245–249 (2016).

20. Akinbami, L. J., Moorman, J. E. & Liu, X. Asthma prevalence, health care use, and mortality: United States, 2005-2009. *Natl Health Stat Report* 1–14 (2011).

21. Gupta, R. S., Springston, E. E. & Weiss, K. B. Eliminating asthma disparities: is there evidence of progress? *Curr Opin Pulm Med* **15**, 72–78 (2009).

22. Lieu, T. A. *et al.* Racial/ethnic variation in asthma status and management practices among children in managed medicaid. *Pediatrics* **109**, 857–865 (2002).

23. AAFA. Asthma Disparities - Reducing Burden on Racial and Ethnic Minorities | AAFA.org. https://www.aafa.org/asthma-disparities-burden-on-minorities.aspx.

24. Laditka, J. N. & Laditka, S. B. Race, Ethnicity and Hospitalization for Six Chronic Ambulatory Care Sensitive Conditions in the USA. *Ethnicity & Health* **11**, 247–263 (2006).

25. Getahun, D., Demissie, K. & Rhoads, G. G. Recent trends in asthma hospitalization and mortality in the United States. *J Asthma* **42**, 373–378 (2005).

26. Krishnan, J. A. *et al.* Race and sex differences in consistency of care with national asthma guidelines in managed care organizations. *Arch Intern Med* **161**, 1660–1668 (2001).

27. Pressman, A. *et al.* Measuring Health Equity for Ambulatory Care Sensitive Conditions in a Large Integrated Health Care System: The Development of an Index. *Health Equity* **3**, 92–98 (2019).

28. American Hospital Association. *AN ANALYTIC APPROACH TO ADVANCING HEALTH EQUITY*. https://d3oxih60gx1ls6.cloudfront.net/f6b7290d-751c-46c2-80f8-970cb87f1b9b/e302be3d-99fa-46d7-858b-36a6585c290f_Sutter_Health_EOC_Case_Study.pdf (2019).

29. ROSENBAUM, P. R. & RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).

30. Austin, P. C. A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. *Multivariate Behav Res* **46**, 119–151 (2011).

31. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* **46**, 399–424 (2011).

32. Austin, P. C. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery* **134**, 1128-1135.e3 (2007).

33. Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* **42**, 1--28 (2011).

34. Ho, D. E., Imai, K., King, G. & Stuart, E. A. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* **15**, 199–236 (2007).

35. Zhang, Z., Kim, H. J., Lonjon, G., Zhu, Y. & Group, written on behalf of A. B.-D. C. T. C. Balance diagnostics after propensity score matching. *Annals of Translational Medicine* **7**, 16–16 (2019).

36. Olmos, A. & Govindasamy, P. Propensity Scores: A Practical Introduction Using R. *Journal of MultiDisciplinary Evaluation* **11**, 68–88 (2015).

37. Diamond, A. & Hainmueller, J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program AU - Abadie, Alberto. *Journal of the American Statistical Association* **105**, 493–505 (2010).

38. Bauhoff, S. The effect of school district nutrition policies on dietary intake and overweight: A synthetic control approach. *Economics & Human Biology* **12**, 45–55 (2014).

39. Kiesel, K. & Villas-Boas, S. B. Can information costs affect consumer choice? Nutritional labels in a supermarket experiment. *International Journal of Industrial Organization* **31**, 153–163 (2013).

40. Goin, D. E., Rudolph, K. E. & Ahern, J. Impact of drought on crime in California: A synthetic control approach. *PLOS ONE* **12**, e0185629 (2017).

41. Synth.pdf.

42. Hahn, J. & Shi, R. Synthetic Control and Inference. *Econometrics* **5**, 52 (2017).

43. SCtools.pdf.

44. Lipsitch, M., Tchetgen, E. T. & Cohen, T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology* **21**, 383–388 (2010).

45. Carlson, R. W. *et al.* Alcohol withdrawal syndrome. *Crit Care Clin* **28**, 549–585 (2012).

46. Rehm, J. *et al.* Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders. *The Lancet* **373**, 2223–2233 (2009).

47. Li, T.-K., Hewitt, B. G. & Grant, B. F. Alcohol use disorders and mood disorders: A National Institute on Alcohol Abuse and Alcoholism perspective. *Biological Psychiatry* **56**, 718–720 (2004).

48. Bouchery, E. E., Harwood, H. J., Sacks, J. J., Simon, C. J. & Brewer, R. D. Economic Costs of Excessive Alcohol Consumption in the U.S., 2006. *American Journal of Preventive Medicine* **41**, 516–524 (2011).

49. Pirmohamed, M. *et al.* The burden of alcohol misuse on an inner-city general hospital. *QJM* **93**, 291–295 (2000).

50. White, T., Lavoie, S. & Nettleman, M. D. Potential Cost Savings Attributable to Influenza Vaccination of School-aged Children. *Pediatrics* **103**, e73–e73 (1999).

51. Gentilello, L. M., Ebel, B. E., Wickizer, T. M., Salkever, D. S. & Rivara, F. P. Alcohol Interventions for Trauma Patients Treated in Emergency Departments and Hospitals. *Ann Surg* **241**, 541–550 (2005).

52. Hall, W. & Zador, D. The alcohol withdrawal syndrome. *Lancet* **349**, 1897–1900 (1997).

53. Long, D., Long, B. & Koyfman, A. The emergency medicine management of severe alcohol withdrawal. *The American Journal of Emergency Medicine* **35**, 1005–1011 (2017).

54. Bayard, M., McIntyre, J., Hill, K. & Woodside, J. Alcohol Withdrawal Syndrome. *AFP* **69**, 1443–1450 (2004).

55. Grover, S. & Ghosh, A. Delirium Tremens: Assessment and Management. *J Clin Exp Hepatol* **8**, 460–470 (2018).

56. MatchIt.pdf.

57. Zhang, Z. Propensity score method: a non-parametric technique to reduce model dependence. *Ann Transl Med* **5**, (2017).

58. Austin, P. C. Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses. *Int J Biostat* **5**, (2009).

59. Kreif, N. *et al.* Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units. *Health Econ* **25**, 1514–1528 (2016).

60. Kiesel, K. & Villas-Boas, S. B. Can information costs affect consumer choice? Nutritional labels in a supermarket experiment. *International Journal of Industrial Organization* **31**, 153–163 (2013).

61. Abadie, A., Diamond, A. & Hainmueller, J. Synth: An R Package for Synthetic Control Methods in Comparative Case Studies. *Journal of Statistical Software* **42**, 1–17 (2011).

62. Lee, J. A., Duby, J. J. & Cocanour, C. S. Effect of early and focused benzodiazepine therapy on length of stay in severe alcohol withdrawal syndrome. *Clinical Toxicology* **57**, 624–627 (2019).

63. Melkonian, A., Patel, R., Magh, A., Ferm, S. & Hwang, C. Assessment of a Hospital-Wide CIWA-Ar Protocol for Management of Alcohol Withdrawal

Syndrome. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* **3**, 344–349 (2019).

64. Griessbach, A. N., Mueller, B. U., Battegay, E. & Beeler, P. E. The maximum alcohol withdrawal syndrome score associates with worse clinical outcomes—A retrospective cohort study. *Drug and Alcohol Dependence* **205**, 107708 (2019).
65. Diabetes and Prediabetes | CDC. https://www.cdc.gov/chronicdisease/resources/publications/factsheets/diabetes-prediabetes.htm (2020).
66. Ackermann, R. T., Finch, E. A., Brizendine, E., Zhou, H. & Marrero, D. G. Translating the Diabetes Prevention Program into the community. The DEPLOY Pilot Study. *Am J Prev Med* **35**, 357–363 (2008).
67. Diabetes Prevention Program (DPP) Research Group. The Diabetes Prevention Program (DPP): description of lifestyle intervention. *Diabetes Care* **25**, 2165–2171 (2002).
68. Tuomilehto, J. *et al.* Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle among Subjects with Impaired Glucose Tolerance. *http://dx.doi.org/10.1056/NEJM200105033441801* https://www.nejm.org/doi/10.1056/NEJM200105033441801 (2009) doi:10.1056/NEJM200105033441801.
69. Effects of Diet and Exercise in Preventing NIDDM in People With Impaired Glucose Tolerance: The Da Qing IGT and Diabetes Study | Diabetes Care. https://care.diabetesjournals.org/content/20/4/537.
70. Ely, E. K. *et al.* A National Effort to Prevent Type 2 Diabetes: Participant-Level Evaluation of CDC's National Diabetes Prevention Program. *Diabetes Care* **40**, 1331–1341 (2017).
71. Translating the Diabetes Prevention Program Into Practice in the General Community - Helen A. Amundson, Marcene K. Butcher, Dorothy Gohdes, Taryn O. Hall, Todd S. Harwell, Steven D. Helgerson, Karl K. Vanderwood, , 2009. https://journals.sagepub.com/doi/abs/10.1177/0145721709333269?casa_token=moQp8QXap20AAAAA:uMgcLWHUSZTS9pKp0SCL5TBRj4cxbQWem0neJ0leGp1zZtZG1UHBN2Ys1NRKRrmZnDXWMXn8xxREbQ.
72. Romanelli Robert. Propensity Score Matching of ELEVATE-DP. (2019).
73. Kramer, M. K., McWilliams, J. R., Chen, H.-Y. & Siminerio, L. M. A Community-Based Diabetes Prevention Program: Evaluation of the Group Lifestyle Balance Program Delivered by Diabetes Educators. *Diabetes Educ* **37**, 659–668 (2011).
74. Nonnemaker, J. *et al.* Estimating the return on investment of the Bureau of Tobacco Free Florida tobacco control programme from 1999 to 2015. *BMJ Open* **11**, e040012 (2021).
75. Radó, M. K., Lenthe, F. J. van, Sheikh, A. & Been, J. V. Investigating the effects of comprehensive smoke-free legislation on neonatal and infant mortality in Thailand using the synthetic control method. *EClinicalMedicine* **27**, (2020).

76. Sheridan, P., Trinidad, D., McMenamin, S., Pierce, J. P. & Benmarhnia, T. Evaluating the impact of the California 1995 smoke-free workplace law on population smoking prevalence using a synthetic control method. *Prev Med Rep* **19**, (2020).

77. Chelwa, G., Walbeek, C. van & Blecher, E. Evaluating South Africa's tobacco control policy using a synthetic control method. *Tobacco Control* **26**, 509–517 (2017).

78. Bartos, B. J., McCleary, R., Mazerolle, L. & Luengen, K. Controlling Gun Violence: Assessing the Impact of Australia's Gun Buyback Program Using a Synthetic Control Group Experiment. *Prev Sci* **21**, 131–136 (2020).

79. Matthay, E. C. *et al.* Firearm and Nonfirearm Violence After Operation Peacemaker Fellowship in Richmond, California, 1996–2016. *Am J Public Health* **109**, 1605–1611 (2019).

80. Kagawa, R. M. C. *et al.* Repeal of Comprehensive Background Check Policies and Firearm Homicide and Suicide. *Epidemiology* **29**, 494–502 (2018).

81. Rudolph, K. E., Stuart, E. A., Vernick, J. S. & Webster, D. W. Association Between Connecticut's Permit-to-Purchase Handgun Law and Homicides. *Am J Public Health* **105**, e49-54 (2015).

82. Wigley, S., Dieleman, J. L., Templin, T., Mumford, J. E. & Bollyky, T. J. Autocratisation and universal health coverage: synthetic control study. *BMJ* **371**, m4040 (2020).

83. Kendall, M. *et al.* Epidemiological changes on the Isle of Wight after the launch of the NHS Test and Trace programme: a preliminary analysis. *The Lancet Digital Health* **2**, e658–e666 (2020).

84. Zang, X., Zhang, M., Wei, S., Tang, W. & Jiang, S. Impact of public hospital pricing reform on medical expenditure structure in Jiangsu, China: a synthetic control analysis. *BMC Health Services Research* **19**, 512 (2019).

85. Hu, L., Kaestner, R., Mazumder, B., Miller, S. & Wong, A. The Effect of the Affordable Care Act Medicaid Expansions on Financial Wellbeing. *J Public Econ* **163**, 99–112 (2018).

86. Hernaes, O. *Activation Against Absenteeism: Evidence from a Sickness Insurance Reform in Norway*. https://papers.ssrn.com/abstract=3037896 (2017).

87. Olsen, C. B. & Melberg, H. O. Did adolescents in Norway respond to the elimination of copayments for general practitioner services? *Health Economics* **27**, 1120–1130 (2018).

88. Maclean, J. C. & Saloner, B. Substance Use Treatment Provider Behavior and Healthcare Reform: Evidence from Massachusetts. *Health Econ* **27**, 76–101 (2018).

89. Rieger, M., Wagner, N. & Bedi, A. S. Universal health coverage at the macro level: Synthetic control evidence from Thailand. *Social Science & Medicine* **172**, 46–55 (2017).

90. West, T. A. P., Börner, J., Sills, E. O. & Kontoleon, A. Overstated carbon emission reductions from voluntary REDD+ projects in the Brazilian Amazon. *PNAS* **117**, 24188–24194 (2020).

91. Zhang, K., Xu, D., Li, S., Zhou, N. & Xiong, J. Has China's Pilot Emissions Trading Scheme Influenced the Carbon Intensity of Output? *Int J Environ Res Public Health* **16**, (2019).

92. Mao, Y. Does democratic transition reduce carbon intensity? Evidence from Indonesia using the synthetic control method. *Environ Sci Pollut Res Int* **25**, 19908–19917 (2018).

93. Sills, E. O. *et al.* Estimating the Impacts of Local Policy Innovation: The Synthetic Control Method Applied to Tropical Deforestation. *PLOS ONE* **10**, e0132590 (2015).

94. Trandafir, M. Legal recognition of same-sex couples and family formation. *Demography* **52**, 113–151 (2015).

95. Reyburn, R. *et al.* Effect of ten-valent pneumococcal conjugate vaccine introduction on pneumonia hospital admissions in Fiji: a time-series analysis. *Lancet Glob Health* **9**, e91–e98 (2021).

96. Reinbold, G. W. Effects of the Convention on the Rights of the Child on child mortality and vaccination rates: a synthetic control analysis. *BMC International Health and Human Rights* **19**, 24 (2019).

97. Bruhn, C. A. W. *et al.* Estimating the population-level impact of vaccines using synthetic controls. *PNAS* **114**, 1524–1529 (2017).

98. Alcocer, J. J. Exploring the effect of Colorado's recreational marijuana policy on opioid overdose rates. *Public Health* **185**, 8–14 (2020).

99. Laqueur, H. *et al.* The impact of cannabis legalization in Uruguay on adolescent cannabis use. *Int J Drug Policy* **80**, 102748 (2020).

100. Zhang, L. & Ma, L. Does open data boost journal impact: evidence from Chinese economics. *Scientometrics* **126**, 3393–3419 (2021).

101. Bennett, M. All things equal? Heterogeneity in policy effectiveness against COVID-19 spread in chile. *World Development* **137**, 105208 (2021).

102. 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* **72**, 673–678 (2018).

103. 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* **72**, 673–678 (2018).

104. Marinello, S., Leider, J., Pugach, O. & Powell, L. M. The impact of the Philadelphia beverage tax on employment: A synthetic control analysis. *Econ Hum Biol* **40**, 100939 (2021).

105. Lin, T. K., Teymourian, Y. & Tursini, M. S. The effect of sugar and processed food imports on the prevalence of overweight and obesity in 172 countries. *Global Health* **14**, 35 (2018).

106. Alexander, J. A. & Hearld, L. R. Methods and metrics challenges of delivery-system research. *Implementation Sci* **7**, 15 (2012).

107. Ben-Michael, E., Feller, A. & Rothstein, J. The Augmented Synthetic Control Method. *arXiv:1811.04170 [econ, stat]* (2020).

108. L'Hour, A. A. J. A Penalized Synthetic Control Estimator for Disaggregated Data. 40.

109. Doudchenko, N. & Imbens, G. W. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. *arXiv:1610.07748 [stat]* (2017).

110. Arkhangelsky, D., Athey, S., Hirshberg, D., Imbens, G. & Wager, S. *Synthetic Difference In Differences*. w25532 http://www.nber.org/papers/w25532.pdf (2019) doi:10.3386/w25532.

111. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature News* **533**, 452 (2016).

112. Jade Benjamin-Chung. Replication, transparency & reproducibility.

Appendix A

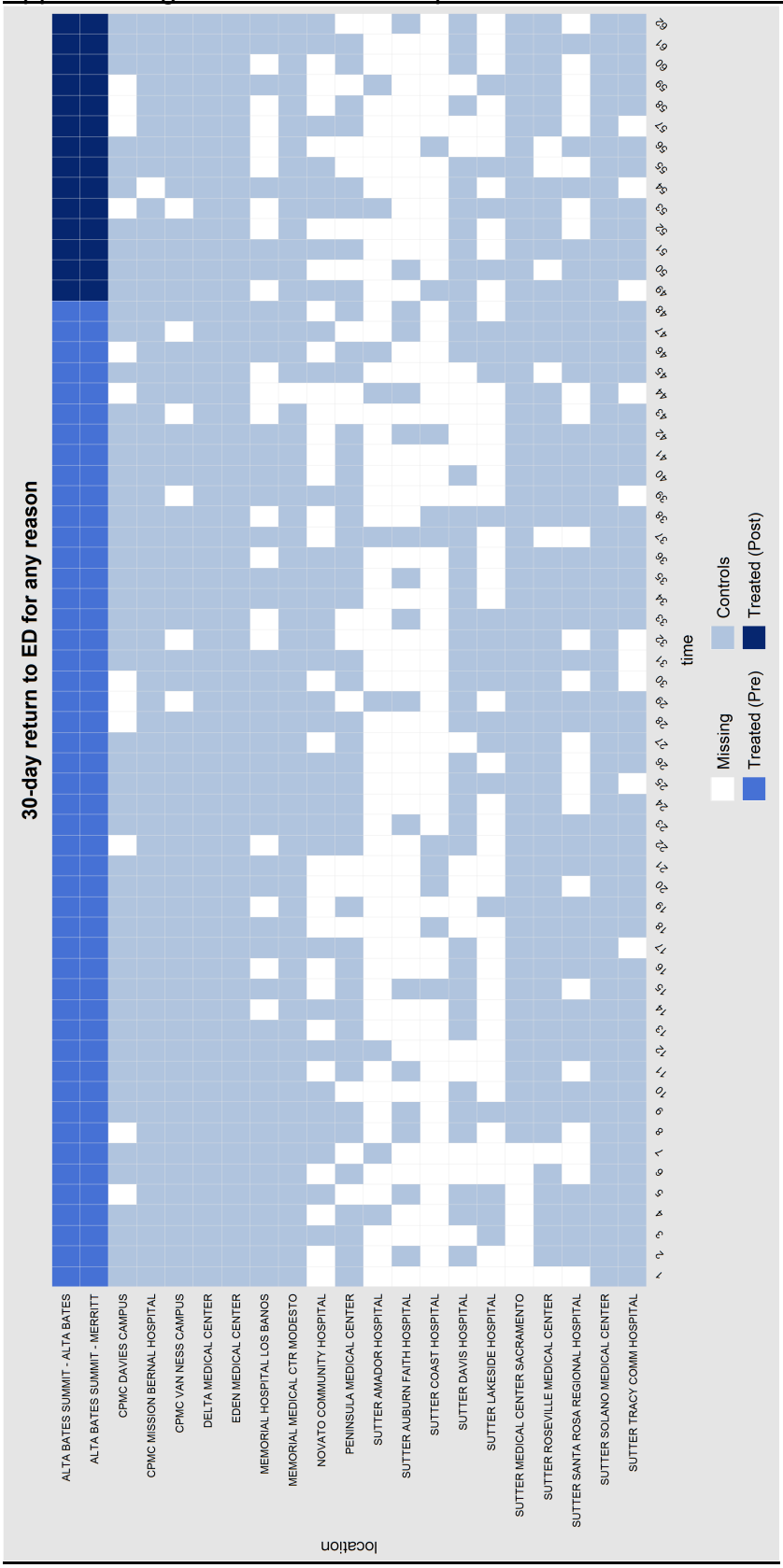
Appendix Table A1. ICD-9/ICD-10 codes for respiratory disease history

ICD9	ICD9 Description	ICD10	ICD10 Description
466	Acute bronchitis	J20.9	Acute bronchitis, unspecified
466.11	Acute bronchiolitis due to respiratory syncytial virus (RSV)	J21.0	Acute bronchiolitis due to respiratory syncytial virus
466.19	Acute bronchiolitis due to other infectious organisms	J21.8	Acute bronchiolitis due to other specified organisms
477.9	Allergic rhinitis, cause unspecified	J30.0	Vasomotor rhinitis
477	Allergic rhinitis due to pollen	J30.1	Allergic rhinitis due to pollen
477.8	Allergic rhinitis due to other allergen	J30.2	Other seasonal allergic rhinitis
477.1	Allergic rhinitis due to food	J30.5	Allergic rhinitis due to food
477.2	Allergic rhinitis due to animal (cat) (dog) hair and dander	J30.81	Allergic rhinitis due to animal (cat) (dog) hair and dander
477.8	Allergic rhinitis due to other allergen	J30.89	Other allergic rhinitis
477.9	Allergic rhinitis, cause unspecified	J30.9	Allergic rhinitis, unspecified
472	Chronic rhinitis	J31.0	Chronic rhinitis
472.2	Chronic nasopharyngitis	J31.1	Chronic nasopharyngitis
472.1	Chronic pharyngitis	J31.2	Chronic pharyngitis
473	Chronic maxillary sinusitis	J32.0	Chronic maxillary sinusitis
473.1	Chronic frontal sinusitis	J32.1	Chronic frontal sinusitis
473.2	Chronic ethmoidal sinusitis	J32.2	Chronic ethmoidal sinusitis
473.3	Chronic sphenoidal sinusitis	J32.3	Chronic sphenoidal sinusitis
473.8	Other chronic sinusitis	J32.4	Chronic pansinusitis
473.8	Other chronic sinusitis	J32.8	Other chronic sinusitis
473.9	Unspecified sinusitis (chronic)	J32.9	Chronic sinusitis, unspecified
490	Bronchitis, not specified as acute or chronic	J40	Bronchitis, not specified as acute or chronic
491	Simple chronic bronchitis	J41.0	Simple chronic bronchitis
491.1	Mucopurulent chronic bronchitis	J41.1	Mucopurulent chronic bronchitis
491.8	Other chronic bronchitis	J41.8	Mixed simple and mucopurulent chronic bronchitis
491.9	Unspecified chronic bronchitis	J42	Unspecified chronic bronchitis
492	Emphysematous bleb	J43.9	Emphysema, unspecified
492.8	Other emphysema	J43.9	Emphysema, unspecified

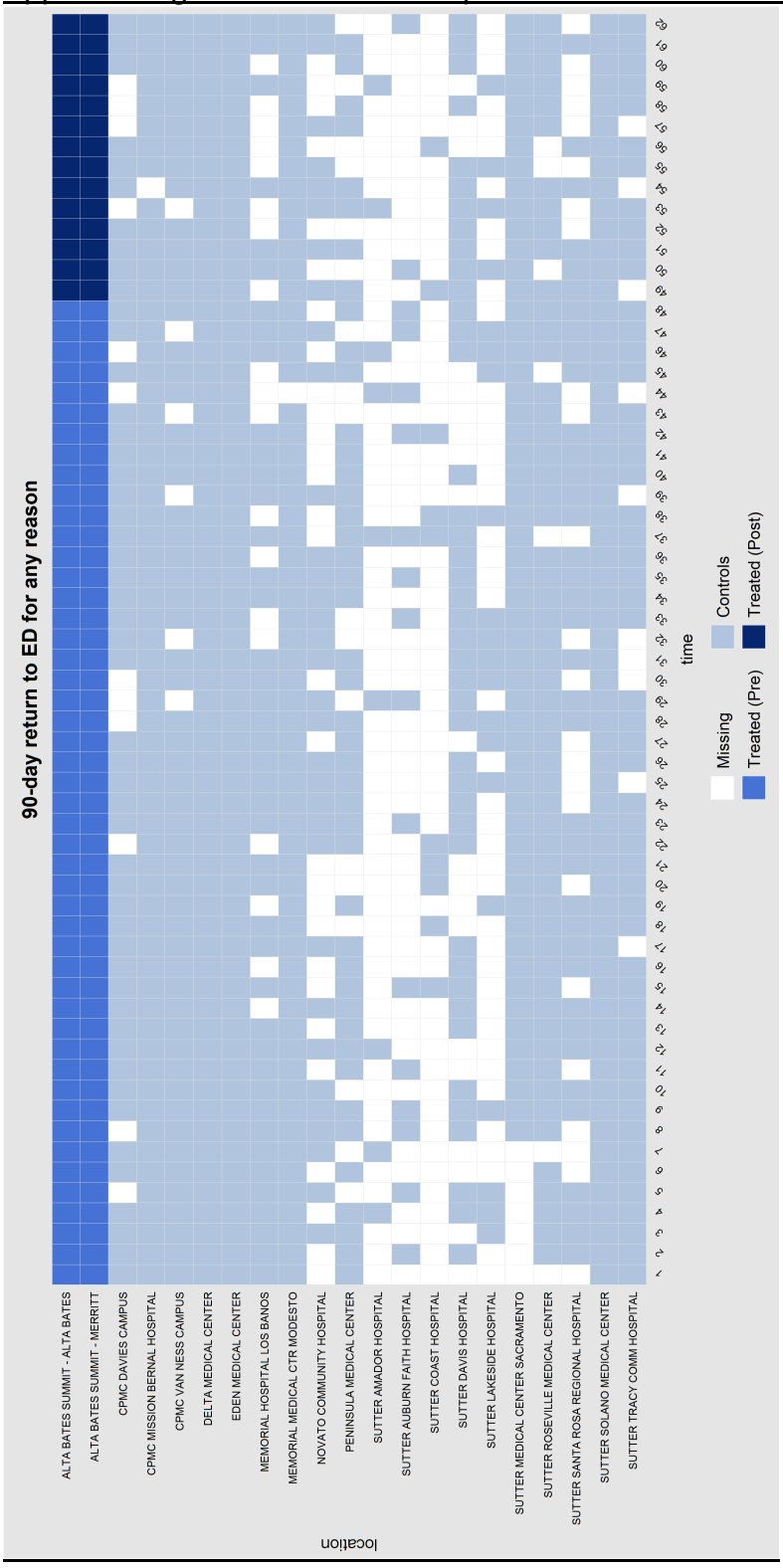
491.22	Obstructive chronic bronchitis with acute bronchitis	J44.0	Chronic obstructive pulmonary disease with acute lower respiratory infection
493.21	Chronic obstructive asthma with status asthmaticus	J44.0	Chronic obstructive pulmonary disease with acute lower respiratory infection
491.21	Obstructive chronic bronchitis with (acute) exacerbation	J44.1	Chronic obstructive pulmonary disease with (acute) exacerbation
493.22	Chronic obstructive asthma with (acute) exacerbation	J44.1	Chronic obstructive pulmonary disease with (acute) exacerbation
491.2	Obstructive chronic bronchitis without exacerbation	J44.9	Chronic obstructive pulmonary disease, unspecified
493.2	Chronic obstructive asthma, unspecified	J44.9	Chronic obstructive pulmonary disease, unspecified
496	Chronic airway obstruction, not elsewhere classified	J44.9	Chronic obstructive pulmonary disease, unspecified
493	Extrinsic asthma, unspecified	J45.20	Mild intermittent asthma, uncomplicated
493.1	Intrinsic asthma, unspecified	J45.20	Mild intermittent asthma, uncomplicated
493.02	Extrinsic asthma with (acute) exacerbation	J45.21	Mild intermittent asthma with (acute) exacerbation
493.12	Intrinsic asthma with (acute) exacerbation	J45.21	Mild intermittent asthma with (acute) exacerbation
493.01	Extrinsic asthma with status asthmaticus	J45.22	Mild intermittent asthma with status asthmaticus
493.11	Intrinsic asthma with status asthmaticus	J45.22	Mild intermittent asthma with status asthmaticus
493.92	Asthma, unspecified type, with (acute) exacerbation	J45.901	Unspecified asthma with (acute) exacerbation
493.91	Asthma, unspecified type, with status asthmaticus	J45.902	Unspecified asthma with status asthmaticus
493.9	Asthma, unspecified type, unspecified	J45.909	Unspecified asthma, uncomplicated
493.81	Exercise induced bronchospasm	J45.990	Exercise induced bronchospasm
493.82	Cough variant asthma	J45.991	Cough variant asthma
493.9	Asthma, unspecified type, unspecified	J45.998	Other asthma
327.2	Organic sleep apnea, unspecified	G47.30	Sleep apnea, unspecified
780.51	Insomnia with sleep apnea, unspecified	G47.30	Sleep apnea, unspecified

780.53	Hypersomnia with sleep apnea, unspecified	G47.30	Sleep apnea, unspecified
780.57	Unspecified sleep apnea	G47.30	Sleep apnea, unspecified
530.11	Reflux esophagitis	K21.0	Gastro-esophageal reflux disease with esophagitis
530.81	Esophageal reflux	K21.9	Gastro-esophageal reflux disease without esophagitis

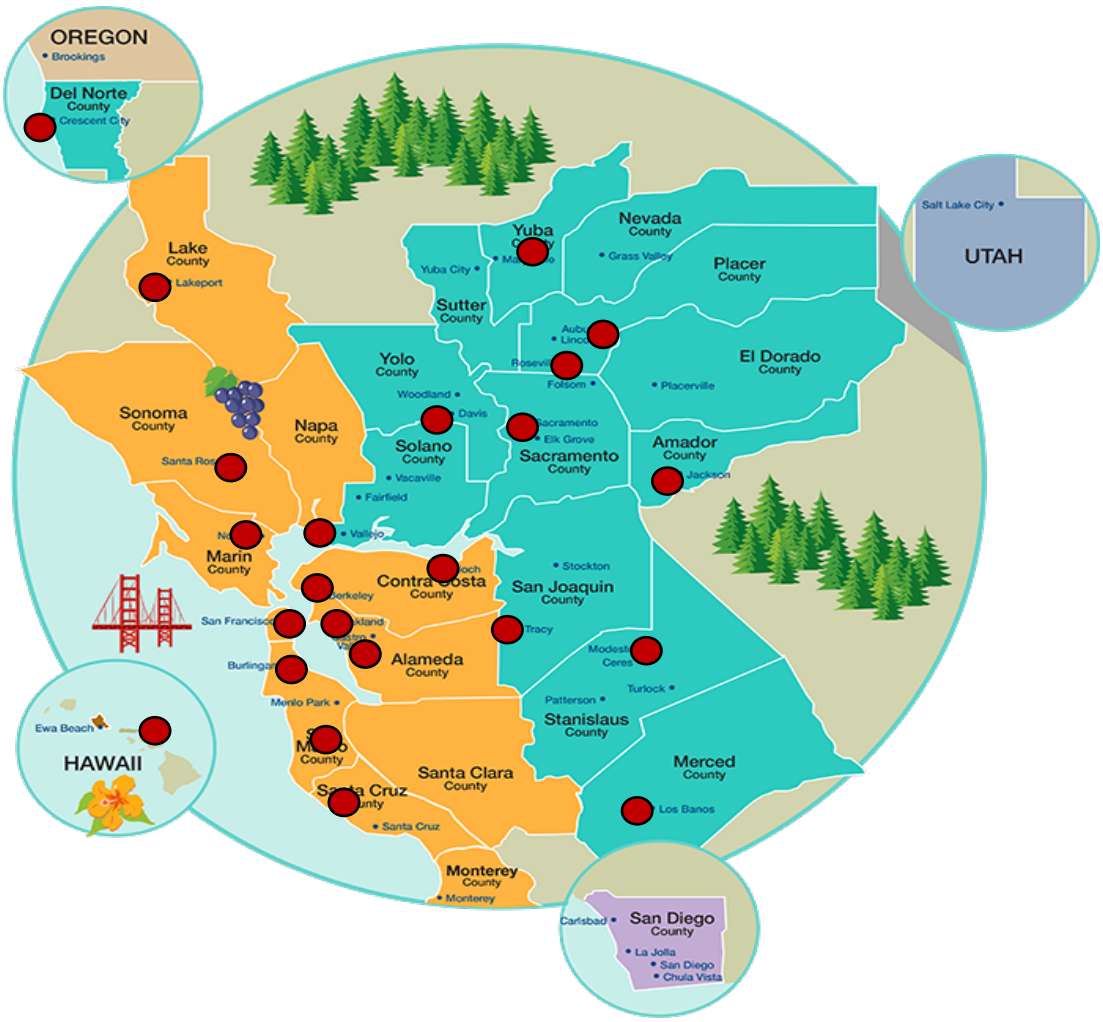
Appendix Figure A1. Panel view plots of 30d returns for SCM analysis



Appendix Figure A2. Panel view plots of 90d returns for SCM analysis



Appendix Figure A3. Summary statistics for Sutter hospitals, include map of service areas



Synthetic control details

In this dissertation, we follow the guidance by Abadie et al., 2010 and define two potential outcomes: Y_{it}^{N} and Y_{it}^{I}, where Y_{it}^{N} is the outcome that would be observed for hospital i at time t if hospital i is not exposed to the intervention, and Y_{it}^{I} the outcome that would be observed if hospital i is exposed to the intervention. We observe Y_{it}^{I} in the post-intervention period for the treated hospital, but Y_{it}^{N} is unobserved for the treated hospital in the post-intervention period. The goal of the synthetic control method is to construct a synthetic control group that yields a reasonable estimate of this missing potential outcome. In doing so, we will be able to estimate the treatment effect of

interest, or the effect of the program on our outcomes of interest for the treated hospital in the post-intervention period:

$$\alpha_{1t} = Y_{it}^{I} - Y_{it}^{N}$$

The program occurs at time period $T_0 + 1$ so that 1, 2, …, T_0 are the pre-program periods and $T_0 + 1$, $T_0 + 2$, …, T are the post-program periods.
The synthetic control should resemble the treated unit in relevant pre-program characteristics and pre-intervention outcomes. Thus, we define U_i as a vector of observed covariates for each unit. These variables should consist of a set of predictors of the outcome. We also define a (T_0 x 1) vector $K = (k_1, \dots, k_{T_0})'$ that denotes some linear combination of pre-intervention outcomes.
Abadie and Gardeazabal (2003) and Abadie et al. 2010 propose that weight W* are chosen such that the synthetic control best approximates the program hospitals with respect to outcome predictors $U_i$and linear combinations of pre-intervention outcomes $\bar{Y}_i^{K_1}, \dots, \bar{Y}_i^{K_M}$.

X_1 will be a (k x 1) vector containing the values of the pre-program characteristics of the treated hospital that we aim to match as closely as possible. **X_0** is a k x J matrix with values of the same variables for the control hospitals. The preprogram characteristics in **X_0** and **X_1** may include pre-intervention values of the outcome(s).

We will utilize the *synth* package to choose the vector W* that minimizes the distance || X1-X0W|| between **X_1** and **X_0** **W**[9]. The synth() (Abadie 2010, Abadie 2019, ADH 2011) function solves for W* by minimizing

$$\|X_1 - X_0 W\| v = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

Where V is a matrix allowing for different weights to be applied to variables in **X_0** and **X_1** based on their predictive power on the outcome. The synth() function allows for a data-drive procedure to choose V*, as proposed in Abadie and Gardeazabal (2003) and Abadie et al. (2010), such that the mean-squared prediction error (MSPE) of the outcome is minimized over the pre-program years (Abadie, Diamond, Hainmueller, 2011).

Large sample inferential techniques are not appropriate in this setting due to the small number of units; however, exact inferential techniques, such as a placebo/permutation test may be used to assess how unusual an effect would be if it were due to change and thus provide context for the effect size. As advised in prior literature (Abadie et al 2010, etc.) we obtained inference for these estimates by using a series of placebo (permutation tests), in which we applied the synthetic control method for each control hospital, as though it were the one that had implemented the program. The estimated effect of the actual treated hospital can then be compared to the size of these other effect estimates. Typically, the permutation test results are compared for hospitals in which pre-program trends are well predicted by the synthetic control.