Preventive Medicine Reports 24 (2021) 101572

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

Developing a synthetic control group using electronic health records: Application to a single-arm lifestyle intervention study

Yi-An Ko [a,*], Zhenchao Chen [a], Chang Liu [b], Yingtian Hu [a], Arshed A. Quyyumi [c], Lance A. Waller [a], Melinda Higgins [d], Thomas R. Ziegler [e], Kenneth L. Brigham [f], Greg S. Martin [f]

[a] Department of Biostatistics and Bioinformatics, Emory University Rollins School of Public Health, Atlanta, GA, United States
[b] Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA, United States
[c] Division of Cardiology, Department of Medicine, Emory University School of Medicine, Atlanta, GA, United States
[d] Emory University School of Nursing, Atlanta, GA, United States
[e] Division of Endocrinology, Metabolism and Lipids, Department of Medicine, Emory University School of Medicine, Atlanta, GA, United States
[f] Division of Pulmonary, Critical Care and Allergy, Department of Medicine, Emory University School of Medicine, Atlanta, GA, United States

ARTICLE INFO

Keywords:
Electronic medical record
Doubly robust
Pseudo control
Controlled trials

ABSTRACT

The electronic health records (EHR) infrastructure offers a tremendous resource for identifying controls who match the characteristics of study participants in a single-arm trial. The objectives are to (1) demonstrate the feasibility of curating a synthetic control group for an existing study cohort through EHR data extraction and (2) evaluate the effect of a lifestyle intervention on selected cardiovascular health metrics. A total of 711 university employees were recruited between 2008 and 2012 to participate in a health partner intervention to improve cardiovascular health and were followed for five years. Data of nearly 8000 eligible subjects were extracted from the EHR to create a synthetic control cohort during the same study period. To minimize confounding, crude comparison, exact matching, propensity score matching, and doubly robust estimation were used to compare the selected cardiovascular health metrics at 1 and 5 years of follow-up. Blood pressure and body mass index improved in the intervention group compared to the EHR synthetic controls. The findings of changes in lipid measurements were somewhat unexpected. When analyzing the subgroup without lipid-lowering medications, the intervention group exhibited better control of cholesterol levels over time than did our synthetic controls. Some measurements in the EHR system may be more robust for synthetic selection than others. EHR synthetic controls can provide an alternative to estimate intervention effects appropriately in single-arm studies for these measurements.

1. Introduction

Recent years have seen a significant increase in lifestyle intervention programs around the globe expanding to various groups in clinics, institutions, and communities (Lotfaliany et al., 2020; Ferrara et al., 2020; van Dammen et al., 2018; Keyserling, 2016; Kandula, 2015; Keyserling et al., 2014). The general goals of such programs are to improve or maintain health and to minimize health care use. Evaluations of these intervention programs often do not receive great attention since many are not planned for research investigations. The key elements of intervention are largely based on prior study evidence rather than serving as areas of research evaluation. These lifestyle programs are commonly developed to ensure feasibility and effective use of resources. Nevertheless, when there is a need to evaluate the program, the pre-post comparison is usually considered to be satisfactory for the program provider. The gold-standard randomized controlled trial (RCT) designs are generally not used, as randomizing participants to a control group seems inappropriate in this type of setting. Also, such RCT design requires additional resources, (at least) doubles the study expense through the development of a control arm, and often results in substantial dropouts, especially among those assigned to the control group.

Although a pre-post study design has temporality to suggest that the outcome is caused by the intervention, a major limitation is lack of a control group to account for potential changes in the outcomes over

* Corresponding author at: Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road N.E., Atlanta, GA 30322, United States.
E-mail address: yi-an.ko@emory.edu (Y.-A. Ko).

https://doi.org/10.1016/j.pmedr.2021.101572
Received 17 May 2021; Received in revised form 16 September 2021; Accepted 23 September 2021
Available online 4 October 2021

time. As such, the observed changes or improvements cannot be fully attributed to the intervention program (Thiese, 2014). This complication is amplified in studies with long follow-up periods where the outcome measure can fluctuate with time.

Several studies have utilized historical controls to assess treatment effects to reduce bias from pre-post comparisons, decrease patient burden, and to save costs (Desai et al., 2013; Gökbuget et al., 2016; Mendell et al., 2016; Viele, 2014; Peddada et al., 2007). When it is not feasible to identify appropriate controls from previous trials, the electronic health record (EHR) infrastructure offers a tremendous resource. By leveraging EHR data, one can sample a hypothetical, or "synthetic" control cohort from the same study population that matches the characteristics of the study participants. Recent evolution of standardized EHR data collection and the availability of advanced statistical methods and data science tools enable researchers to evaluate the efficacy of an intervention program in a cost-effective way.

The objectives of this study were to evaluate the feasibility of curating a synthetic control group through EHR data extraction for a lifestyle intervention program and to investigate the effect of the program. Our goal was to explore the effect of having a personalized health partner who provided continued lifestyle consultation on cardiovascular health metrics in terms of blood pressure, body mass index (BMI), serum cholesterol concentrations, and 10-year Atherosclerotic Cardiovascular Disease (ASCVD) risk score (Grundy, 2019). We hypothesized that those coaching with a health partner would improve or maintain health over time compared to those who received no intervention or regular care via primary physicians or annual check-ups.

2. Methods

2.1. Predictive health study participants

A total of 711 employees of Emory University (Atlanta, Georgia) were recruited as part of the Emory/Georgia Tech Predictive Health Initiative in 2008–2012 via random sampling among employees who had been employed $\geq$ 2 years with university-sponsored insurance (Al Mheid et al., 2016; Rask et al., 2011). The program focused on maintaining health and evaluating the impact of health partner intervention on cardiovascular health. Study participants were generally healthy without uncontrolled or acute illness. At the baseline visit, each subject was assigned a health partner who provided continued counseling throughout the study. These health partners were specifically trained to utilize the participant's health data and collaboratively generate a goal to establish a personalized action plan promoting a healthy lifestyle (Brigham, 2010). Individuals were followed for five years, with visits at 6 and 12 months, followed by annual visits. During each visit, blood pressure (average of three measures) and BMI were measured, and blood samples were taken after an overnight fast to monitor health status. Measurements of serum concentrations of glucose, total cholesterol, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), and triglycerides were analyzed by Quest Diagnostics Lipid Panel. Baseline disease diagnosis (including hypertension, diabetes, and dyslipidemia) was obtained through questionnaires, combined with self-reported medication prescriptions reviewed by two cardiologists. The study was approved by the institutional review board of Emory University. Written informed consent was obtained from all participants.

2.2. Synthetic control group derived from the EHR

Medical records data for eligible individuals during clinical visits were extracted from the Emory Healthcare Data Warehouse from 2008 to 2017 (since 2012 was the last year of Predictive Health study recruitment). The inclusion criteria were Emory employees with at least one billed encounter at a primary care / family medicine clinic within the Emory system between 2008 and 2012 (i.e., the study recruitment period). Employees were identified through Emory sponsored insurance plans. We implemented the same exclusion criteria as the study using ICD-10 codes, including psychosocial disorders, congestive heart failure, vascular and cognitive disease (e.g., peripheral vascular disease, cerebrovascular disease, hemiplegia/paraplegia, dementia), rheumatic disease, renal disease, moderate or severe liver disease, pregnancy, any malignancy (including lymphoma and leukemia except malignant neoplasm of skin), AIDS/HIV, substance abuse, and any hospitalization within 1 year prior to the baseline date (identified along with medications that are only prescribed during hospitalization).

A total of 434 Predictive Health study participants found in the EHR database were further removed by matching full name and date of birth (Ko et al., 2020). For those who met the eligibility criteria, the following data elements were obtained at each hospital encounter between 2012 and 2017 along with visit date: birth date, sex, race, height, weight, systolic and diastolic blood pressure measurements, education, smoking status, ICD-10 codes, lab results (glucose, total cholesterol, HDL, LDL, and triglycerides), and medication prescriptions. Diagnoses of hypertension, diabetes, and dyslipidemia were derived via a combination of ICD-10 codes and medication prescriptions. We assumed no disease if no relevant record within 3 years prior to the baseline year was found. Given that fasting is generally not required for blood draws during hospital visits, we did not restrict time of blood draws for laboratory data. Multiple measurements of blood pressure within a single visit day were summarized using the median to avoid possible outliers or typos in the medical records. Individuals that had any measurement outside the data range of the Predictive Health study were excluded. Additionally, individuals with only one clinical encounter or a maximum follow-up time $<$ 180 days were excluded. The study was approved by the institutional review board of Emory University.

2.3. Statistical analysis

Baseline characteristics were summarized as means or medians and standard deviations or interquartile ranges for continuous variables and as counts and percentages for categorical variables. Baseline age, sex, race, education, smoking, BMI, blood pressure, and laboratory values were compared between the Predictive Health study participants and those obtained from the EHR using two-sample *t*-test, Mann-Whitney *U* test, and Chi-square test, as appropriate. The follow-up time was compared using two-sample *t*-test.

The cardiovascular health metrics included systolic and diastolic blood pressures at resting status, BMI, total cholesterol, HDL, LDL, triglycerides, and ASCVD risk score (Lloyd-Jones, 2019). To summarize the change in each health metric over 5 years of follow-up for each individual with varying numbers of measurements, a linear mixed-effects model was used to obtain a temporal "slope" estimate for each individual. Specifically, each metric was regressed on time (years) in the mean model along with subject-specific random intercept and random time slope to account for within-subject correlation. The best linear unbiased predictors (including the fixed and random effects of time) were extracted, which were used as a simple summary of individual estimated change per year. As most of the improvement in these metrics was found in the first year in the Predictive Health study, we additionally focused on changes within 1 year by calculating the difference between 1-year follow-up and baseline visits. Since not everyone in the EHR cohort would have data available at exactly 1 year from the baseline time point, we included those who had a follow-up observation between 9 months and 15 months.

For each cardiovascular health metric, to examine if there were any differences in the slope across 5 years and the change in 1 year between the two groups while considering covariate imbalance, we adopted 4 analysis strategies as follows: (1) crude comparison without covariate adjustment, (2) exact matching, (3) propensity score matching, (4) doubly robust estimation. Since the EHR data were extracted from subjects who met the study inclusion criteria, we initially compared

their slope estimates with the Predictive Health group using a two-sample *t*-test. Since several characteristics appeared to be different between the two groups and there were a large number of control individuals to choose from, for each Predictive Health subject, we identified a matched control by matching age (within 5 years), sex, race (African American or other, given that African American is a known risk factor for our outcomes), smoking history (yes, no), as well as baseline diagnosis of diabetes, hypertension, and dyslipidemia. Next, with the same covariates, a propensity score was calculated for each individual using a logistic regression model (Rubin, 2007). Matched controls were selected using nearest neighbor with 1:1 matching of propensity scores. A paired *t*-test was used to compare the slope estimates between the Predictive Health group and the matched controls. To further account for covariate imbalance between the two groups, we performed doubly robust estimation that combines an outcome regression (linear regression) with the propensity score model (via inverse probability weighting) to estimate the causal effect of the intervention on an outcome. The advantage is that only one of the two models need be correctly specified to obtain an unbiased effect estimator (Funk, et al., 2011). In particular, the R package 'drtmle' that allows an adaptive estimator of the propensity score was used to implement these methods (Benkeser, et al., 2017; van der Laan, 2014). In both models, we also included the baseline value of the outcome (e.g., baseline BMI for regression of BMI) in addition to the aforementioned baseline covariates (including age, sex, race, smoking history, diabetes, hypertension, and dyslipidemia). Finally, to delineate the intervention effect on changes in lipids without medication effects, we identified a subset of individuals who were never prescribed with lipid-lowering medications at baseline or during the study period and repeated the same analysis procedure. R 4.0.2 was used for analysis.

3. Results

3.1. Data availability and baseline characteristics

Table 1 shows the numbers of available observations in the EHR system for each outcome metric. Baseline demographic and clinical characteristics of the Predictive Health study participants and the EHR cohort are summarized in Table 2. Out of the original 711 Predictive Health participants, 599 were included for analysis because of opt-outs and lack of follow-up data. In the EHR cohort, 4497 had at least two lab measurements with a follow-up time of 6 months of longer. Overall, the distribution of sex was similar with more females (64–67%) present in both groups. The majority of the Predictive Health study participants were Caucasian (73% vs. 55% in the EHR cohort) and 22% were African American (vs. 37–38% in the EHR cohort). The Predictive Health group appeared to have prevalent diagnoses of hypertension (34%), diabetes (11%), and dyslipidemia (17%), whereas the corresponding prevalence values were 16–21%, 4–6%, 11–18% in the EHR cohort. The follow-up time was similar (median: 3–3.5 years).

3.2. Changes in 5 years of follow-up

Fig. 1 shows the 5-year trajectories of systolic and diastolic blood pressures, BMI, serum total cholesterol, HDL, LDL, triglyceride concentrations, and ASCVD risk score for the two groups based on the estimates presented in Table 3. Table 3 displays the number of individuals used for each outcome measure depending on data availability and summarizes the individual slope estimates (derived from the best linear unbiased predictors). Given a variety of EHR data availability, the sample size varied across different metrics. On average, the Predictive Health group exhibited decreases in BMI, diastolic blood pressure, total cholesterol, LDL, and triglycerides and minimum increases in systolic blood pressure and HDL each year. In contrast, the synthetic controls demonstrated increases in the averages of all metrics except LDL and triglycerides. The ASCVD score increased by approximately 0.3% per year in both groups.

Fig. 2 shows the estimated difference in the averaged change per year (based on the slope estimate for each individual over 5 years of follow-up) between the Predictive Health study and the synthetic control group for each cardiovascular health measure using the crude, exact matching, nearest neighbor (based on propensity score matching), and the doubly robust estimation methods. The findings (in terms of statistical significance) were generally consistent across four sample selection and estimation methods (Supplementary Table 1). Compared to the synthetic control group, the Predictive Health study intervention group demonstrated significant reductions in the average BMI, a lower increase in systolic and a lower diastolic blood pressure, and a significantly greater decrease in total cholesterol and triglyceride levels over 5 years, Fig. 3.

Table 1
Data captured in electronic health records (EHR) from the Emory data warehouse for individuals who met the study inclusion criteria.

Data element	Number of observations	Number of individuals	Year
ICD code	20,257,707	165,997	2005–2017
Lab (cholesterols)	2,298,689	72,462	2008–2017
Vitals (blood pressures)	4,898,374	104,158	2008–2017
Height, weight, BMI	2,972,027	104,097	2008–2017
Date of birth, sex, race, education	107,897	107,897	2008–2017

Table 2
Baseline characteristics of the Predictive health study participants and the EHR synthetic control cohort.

Variable	Predictive Health (N = 599)	Synthetic Control[a] (N = 7548)	Synthetic Control[b] (N = 4497)
Age (years)	49 (11)	41 (12)	45 (11)
Sex (female)	395 (66%)	5063 (67%)	2856 (64%)
Race			
African American	130 (22%)	2893 (38%)	1665 (37%)
Caucasian	437 (73%)	4153 (55%)	2477 (55%)
Other	32 (5%)	502 (7%)	355 (8%)
Hypertension	204 (34%)	1209 (16%)	959 (21%)
Diabetes	63 (11%)	307 (4%)	276 (6%)
Dyslipidemia	101 (17%)	852 (11%)	824 (18%)
History of smoking	34 (6%)	16 (0.2%)	9 (0.2%)
Body mass index (kg/m^2)	27.6 (6.1)	27.7 (6.4)	–
Systolic blood pressure (mmHg)	120.5 (15.5)	120.6 (15.7)	–
Diastolic blood pressure (mmHg)	76.1 (10.6)	76.8 (10.4)	–
Total Cholesterol (mg/dL)	194.732 (36.2)	–	186.7 (34.9)
High-density lipoprotein (mg/dL)	63.898 (18.2)	–	51.0 (15.2)
Low-density lipoprotein (mg/dL)	110.650 (31.7)	–	114.4 (30.4)
Triglycerides (mg/dL)	101.326 (56.4)	–	106.5 (66.4)
ASCVD score (%)[c]	4.3 (5.4)	–	4.2 (4.7)
Follow-up time (years)	3.0 [2.0, 5.0]	3.5 [2.3, 4.2]	3.0 [2.0, 3.8]
Number of repeated measurements	5 [3,6]	6 [4, 9]	3 [2, 4]

Mean (standard deviation), frequency count (percentage), and median [lower quartile, upper quartile] are presented.

[a] Individuals obtained from the EHR with available vital data, including body mass index and blood pressure measurements.

[b] Individuals obtained from the EHR with available lab data, including total cholesterol, high-density lipoprotein, low-density lipoprotein, and triglyceride concentration measurements.

[c] 10-year ASCVD risk applies to individuals age 40–79, so the sample sizes are 2293 and 492, respectively.

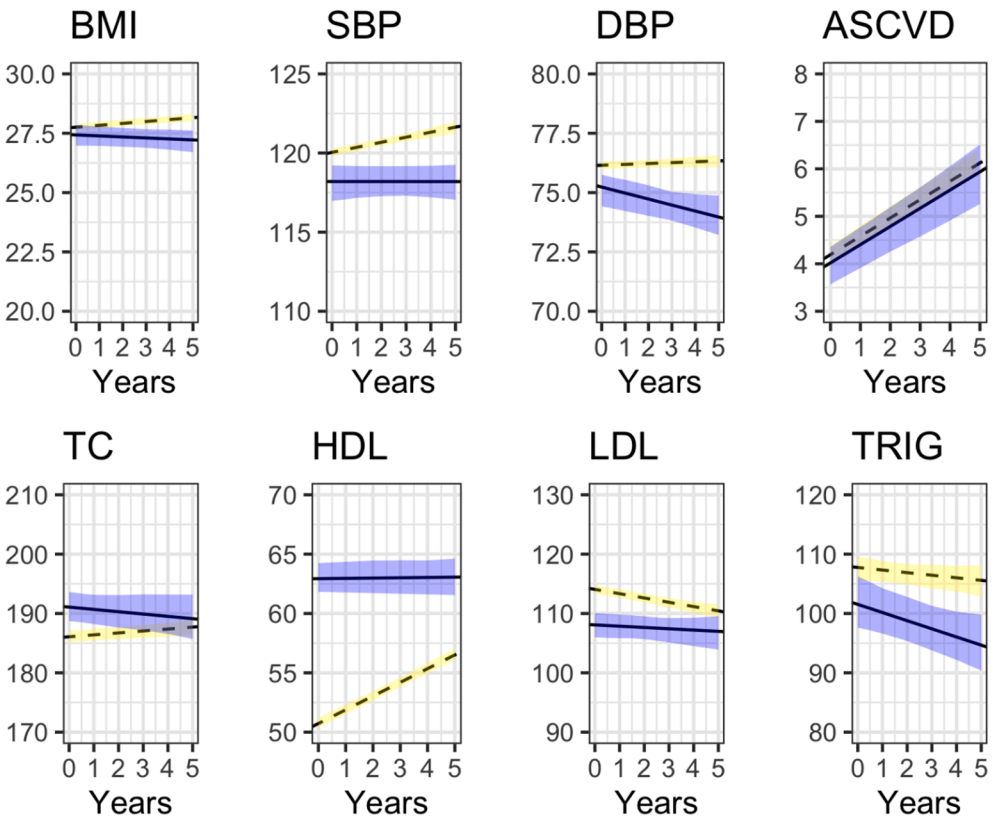

Fig. 1. Estimated mean values and the corresponding 95% confidence bands of cardiovascular health measures across 5 years of follow-up for the Predictive Health study (solid line) and the synthetic control group derived from the electronic health records (dashed line). BMI = body mass index (kg/m^2), SBP = systolic blood pressure (mmHg), DBP = diastolic blood pressure (mmHg), TC = total cholesterol (mg/dL), HDL = high-density lipoprotein (mg/dL), LDL = low-density lipoprotein (mg/dL), TG = triglycerides (mg/dL), ASCVD = 10-year risk estimator for atherosclerotic cardiovascular disease (%).

Table 3
Summary statistics of the estimated changes per year across 5 years of follow-up.

Variable	Statistic	Predictive Health	Synthetic Control
Body mass index (kg/m^2)	N	599	7599
	Mean (SD)	−0.040 (0.318)	0.077 (0.469)
	Median [Q1, Q3]	−0.042 [−0.141, 0.090]	0.068 [−0.100, 0.260]
Systolic blood pressure (mmHg)	N	600	8112
	Mean (SD)	0.013 (0.322)	0.317 (0.648)
	Median [Q1, Q3]	−0.004 [−0.132, 0.117]	0.285 [−0.028, 0.619]
Diastolic blood pressure (mmHg)	N	600	8111
	Mean (SD)	−0.272 (0.437)	0.047 (0.401)
	Median [Q1, Q3]	−0.280 [−0.499, −0.065]	0.046 [−0.163, 0.252]
Total Cholesterol (mg/dL)	N	589	4578
	Mean (SD)	−0.405 (0.968)	0.319 (0.918)
	Median [Q1, Q3]	−0.442 [−0.833, 0.017]	0.313 [−0.098, 0.739]
High-density lipoprotein (mg/dL)	N	590	4534
	Mean (SD)	0.026 (0.676)	1.160 (0.651)
	Median [Q1, Q3]	−0.007 [−0.287, 0.285]	1.138 (0.807, 1.485)
Low-density lipoprotein (mg/dL)	N	590	4501
	Mean (SD)	−0.222 (0.949)	−0.732 (0.670)
	Median [Q1, Q3]	−0.233 [−0.620, 0.161]	−0.742 [−1.021, −0.423]
Triglycerides (mg/dL)	N	589	4526
	Mean (SD)	−1.489 (0.966)	−0.540 (1.048)
	Median [Q1, Q3]	−1.548 [−1.834, −1.212]	−0.675 [−0.992, −0.273]
ASCVD score (%)	N	492	2293
	Mean (SD)	0.370 (0.386)	0.391 (0.349)
	Median [Q1, Q3]	0.309 [0.202, 0.428]	0.328 [0.227, 0.449]

SD = standard deviation, Q1 = lower quartile, Q3 = upper quartile.

Even the overall HDL and LDL levels were improved in both groups, the increase in HDL and decrease in LDL levels were greater in the synthetic control group. In a subset aged 40–79 years at the time of measurement, we found a smaller increase in the ASCVD score in the intervention than the synthetic control group that appeared to be more prominent using the doubly robust method (-0.064%, 95% CI = [−0.126%, −0.002%]), but this difference was not significant using other methods.

3.3. Changes in 1 year of follow-up

Fig. 3 shows the estimated difference in the change approximately within 1 year (ranging from 9 months to 15 months of follow-up) between the Predictive Health study and the synthetic control groups for each cardiovascular health measure using the crude, exact matching, nearest neighbor (based on propensity score matching), and the doubly robust estimation methods. Comparing the intervention group with the synthetic controls, the estimated differences in the changes in BMI and systolic blood pressure within 1 year were −0.292 kg/m^2 (95% CI = [−0.454, −0.130]) and −1.478 mmHg (95% CI = [−2.445,−0.511]), respectively, using the doubly robust method. The results were similar when using the crude and matching comparison methods (Supplementary Table 2). In contrast, the 1-year changes in diastolic blood pressure and lipid levels as well as ASCVD were generally not significantly different between the two groups. Regardless, using the doubly robust method, we found the reduction in triglycerides were greater and the increase in ASCVD was less in the Predictive Health group. Compared with the synthetic control group, the differences were estimated to be −5.796 mg/dL (95% CI = [−9.718, −1.874]) and −0.246% (95% CI = [−0.460, −0.033]), respectively.

3.4. Subgroup without lipid-control medication prescriptions

Given the well-known efficacy of lipid-lowering medications, we investigated the changes between groups among those who were not

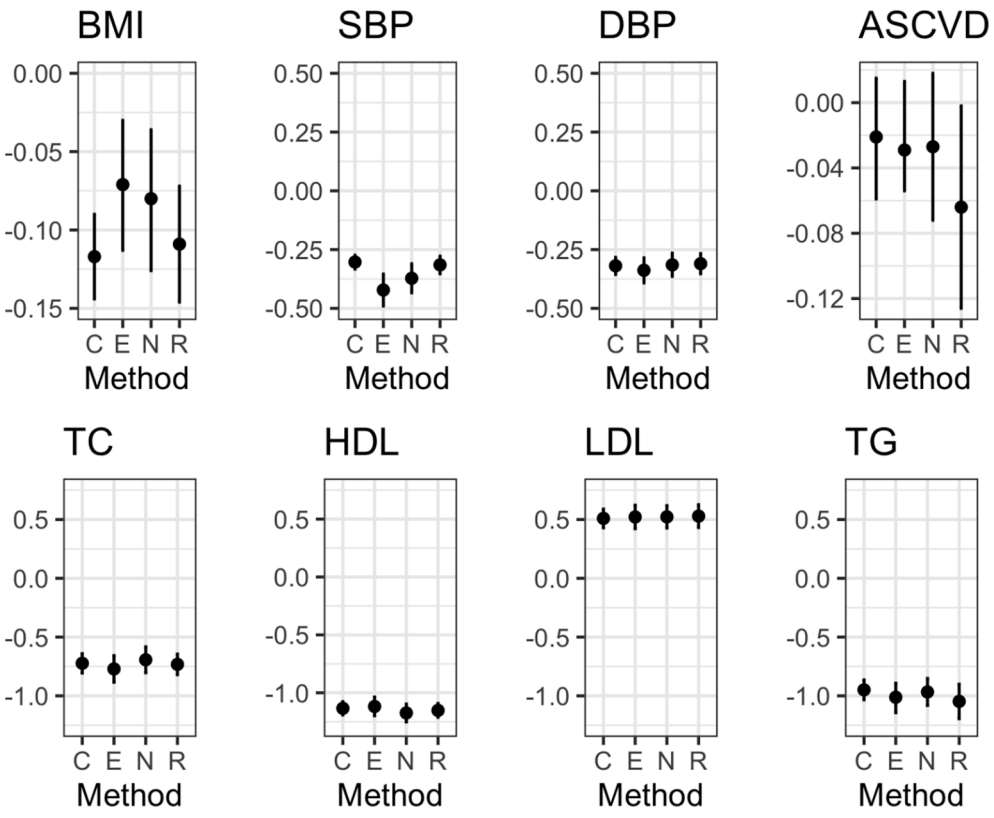

Fig. 2. Estimated difference (and the corresponding 95% confidence interval) in the averaged change per year (over 5 years of follow-up) between the Predictive Health study and the synthetic control group (derived from the electronic health records) for each cardiovascular health measure using four comparison methods. BMI = body mass index (kg/m^2), SBP = systolic blood pressure (mmHg), DBP = diastolic blood pressure (mmHg), TC = total cholesterol (mg/dL), HDL = high-density lipoprotein (mg/dL), LDL = low-density lipoprotein (mg/dL), TG = triglycerides (mg/dL), ASCVD = 10-year risk estimator for atherosclerotic cardiovascular disease (%), C = crude, E = exact matching, N = nearest neighbor (based on propensity score matching), R = doubly robust.

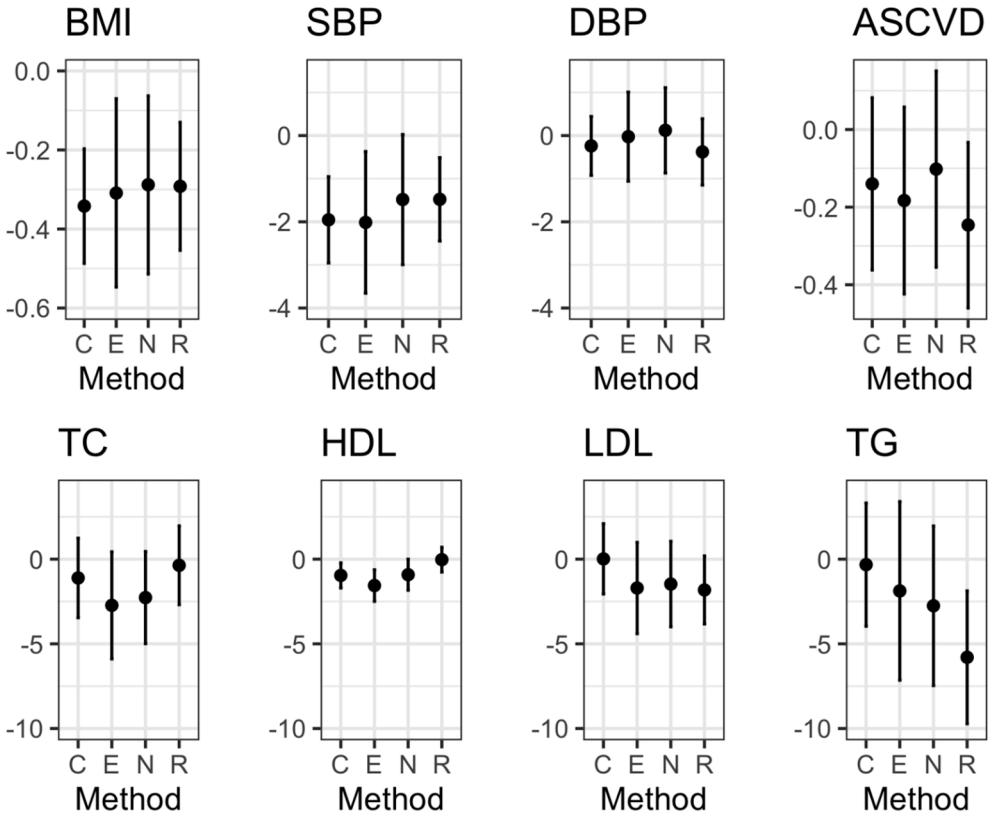

Fig. 3. Estimated difference in the change (and the corresponding 95% confidence interval) within 1 year of follow-up between the Predictive Health study and the synthetic control group (derived from the electronic health records) for each cardiovascular health measure using four comparison methods. BMI = body mass index (kg/m^2), SBP = systolic blood pressure (mmHg), DBP = diastolic blood pressure (mmHg), TC = total cholesterol (mg/dL), HDL = high-density lipoprotein (mg/dL), LDL = low-density lipoprotein (mg/dL), TG = triglycerides (mg/dL), ASCVD = 10-year risk estimator for atherosclerotic cardiovascular disease (%), C = crude, E = exact matching, N = nearest neighbor (based on propensity score matching), R = doubly robust.

prescribed with any lipid-lowering medications at baseline and during the study period. Approximately 14% in the EHR group and 18% in the Predictive Health group were excluded from this analysis. Table 4 shows summarizes the individual slope estimates (derived from the best linear unbiased predictors) in the lipid-control medication-free subgroup. Fig. 4 shows the estimated differences in the changes in the four lipid

Table 4
Summary statistics of the estimated changes per year among individuals without any lipid-lowering prescriptions in 5 years and 1 yearSD = standard deviation, Q1 = lower quartile, Q3 = upper quartile.

	Statistic	Predictive Health	Synthetic Control
5 years of follow-up			
Total Cholesterol (mg/dL)	N	491	3214
	Mean (SD)	−0.492 (0.700)	1.257 (0.577)
	Median [Q1, Q3]	−0.516 [−0.786, −0.173]	1.244 [0.938, 1.550]
High-density lipoprotein [mg/dL)	N	492	3175
	Mean (SD)	0.020 (0.605)	1.207 (0.615)
	Median [Q1, Q3]	−0.004 [−0.263, 0.253]	1.183 [0.885, 1.508]
Low-density lipoprotein [mg/dL)	N	490	3154
	Mean (SD)	−0.386 (0.675)	0.082 (0.328)
	Median [Q1, Q3]	−0.385 [−0.667, −0.098]	0.069 [−0.091, 0.242]
Triglycerides [mg/dL)	N	490	3167
	Mean (SD)	−1.240 (1.506)	−0.142 (1.007)
	Median [Q1, Q3]	−1.336 [−1.751, −0.874]	−0.277 [−0.596, 0.126]
1 year of follow-up[a]Total Cholesterol [mg/dL)	N	486	905
	Mean (SD)	−5.463 (23.328)	−0.849 (19.948)
	Median [Q1, Q3]	−3.500 [−13.375, 5.875]	0.000 [−11.833, 10.673]
High-density lipoprotein [mg/dL)	N	486	890
	Mean (SD)	−1.388 (7.638)	−0.191 [7.678)
	Median [Q1, Q3]	−1.000 [−4.500, 2.000]	−0.805 [−5.033, 3.989]
Low-density lipoprotein [mg/dL)	N	485	885
	Mean (SD)	−3.818 (20.043)	−0.788 (18.195)
	Median [Q1, Q3]	−2.000 [−10.125, 6.125]	0.567 [−10.172, 9.663]
Triglycerides [mg/dL)	N	485	888
	Mean (SD)	−1.567 (27.412)	−0.423 (45.333)
	Median [Q1, Q3]	−1.000 [−10.500, 9.000]	0.000 [−16.919, 20.184]

[a]The synthetic control cohort included those who had a follow-up observation between 9 months and 15 months from their baseline time point.

measurements between the Predictive Health study and the synthetic control group for over 5 years and in 1 year. Compared to the synthetic control group, the intervention group demonstrated a significantly greater decrease or less increase in all cholesterol levels over 5 years. The estimated differences in the changes in total cholesterol, HDL, LDL, and triglycerides were −1.772 (95% CI = [−1.833, −1.712]), −1.194 (95% CI = [−1.254,-1.134]), −0.466 (95% CI = [−0.526, −0.406]), and −1.222 (95% CI = [−1.397, −1.048]) mg/dL, respectively, using the doubly robust method (Supplementary Table 3). Using the crude and matching comparison approaches, the results were consistent. The 1-year findings, however, were not consistent across different analysis approaches, suggesting weaker evidence for differences between groups in 1 year.

4. Discussion

EHR system has great potential for identifying synthetic control cohorts for single-arm studies. The empirical application, however, is plagued by implementation challenges and ambiguous choices of valid controls. Our study provided a practical roadmap to curate an EHR synthetic control cohort and demonstrates the utility for evaluating intervention effects when changes over time (e.g., aging) must be accounted for. However, we also found that the utility of the approach was greatest for consistent measures with little dependence on particular features of the medical encounter (e.g., fasting vs. non-fasting measures of lipids). We illustrated four analysis approaches with the goal of minimizing bias due to potential cofounding in such observational study.

Overall, we found that blood pressure, BMI, total cholesterol, LDL, and triglycerides improved (relatively) in the intervention group. The findings are mostly consistent with a *meta*-analysis of changes expected with lifestyle interventions (Zhang et al., 2017). Comparing 5-year with 1-year results, the observed effect size was relatively small, suggesting the improvements were mainly achieved in the first year. Our previous investigation had found that the changes occurred within the first 6–12 months and most Predictive Health study participants were able to maintain their ideal health status over time (Al Mheid et al., 2016). Though the frequency of contact with the health partner can be adjusted based on the participant's needs, it had no significant impact on the improvement of health outcomes.

In light of these observations, findings comparing the changes in HDL levels between the two groups appeared to be unexpected, possibly due to differences in time of measurement, laboratory assay, and fasting status (Ko et al., 2020; Cooper et al., 2002). However, when analyzing the subgroup without lipid-lowering medication prescriptions, the intervention group exhibited better control of cholesterol levels than the EHR synthetic controls. As such, BMI and blood pressure measurements may be considered more robust tools for synthetic control-based research in the EHR database, whereas measurements of cholesterols are often more sensitive to both endogenous and exogenous factors. Caution needs to be made when analyzing study-based lipid data with other lipid measures obtained from a different source. Nevertheless, our subgroup analysis provided a valuable opportunity to evaluate the intervention effect on natural deterioration of cholesterol levels over time among individuals with normal or borderline cholesterol levels.

A recent review assessed by the National institute for Health and Care Excellence identified 22 studies using external controls to estimate comparative clinical effectiveness (Anderson et al., 2019). Of these, 13 utilized published RCT data, and 6 utilized observational data. Our

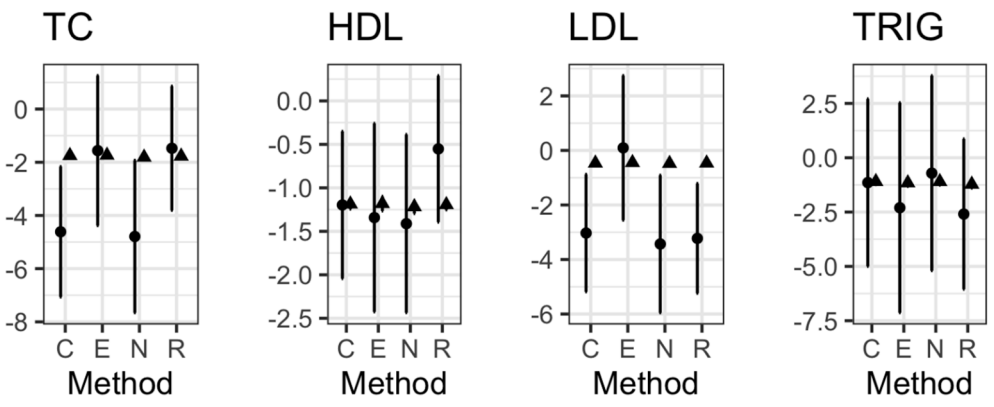

Fig. 4. Estimated difference (and the corresponding 95% confidence interval) in the averaged change in cholesterol lipids per year over 5 years of follow-up (triangle) and within 1 year (solid circle) between the Predictive Health study and the synthetic control group (derived from the electronic health records) using four comparison methods. The 95% confidence intervals for the 5-year results were too small to show in the figure. TC = total cholesterol (mg/dL), HDL = high-density lipoprotein (mg/dL), LDL = low-density lipoprotein (mg/dL), TG = triglycerides (mg/dL), C = crude, E = exact matching, N = nearest neighbor (based on propensity score matching), R = doubly robust.

major contribution is the proposed workflow to select samples from EHR that enable adjustment for the natural changes/declines observed in routine health care for the evaluation of a single-arm study. We illustrated different comparison methods that can be considered as potential analysis recommendations for future studies. Using matched controls may be considered as the most straightforward and sensible approach in practice. On the other hand, using regression adjustment or the proposed doubly robust estimation makes most efficient use of data, which, in turn, can generate more powerful results.

Our study has limitations. First, using EHR data to derive a synthetic control group is only applicable when the same outcome variables and covariates are available. Thus, different groups of individuals were used in analysis of different outcome metrics. Second, Emory employees in the synthetic control cohort were identified using insurance plan recorded in the EHR system, and we assumed that all employees were included. However, it is possible that a portion of employees who chose to use other insurance plans were not captured. Similarly, spouses covered by the same insurance plan may have been mistakenly included in our control cohort pool. Third, the reliability of the estimated differences in various outcome changes may depend on the time interval between repeated measurements. Such intervals may affect how the underlying trajectories over time are estimated. In our analysis, we simply assumed and used a constant slope estimate for each individual to make comparisons. This change may not be constant over time on the individual level in either group, especially when any new treatment is prescribed during the study period. We have attempted to report analysis results of 5-year and 1-year follow-up as well as in a subgroup without statin prescriptions. Fourth, studies generally have higher validity when a concurrent control group is used to validate synthetic controls (Thorlund, 2020). Though there was technically not a concurrent control group for the Predictive Health intervention, we were able to identify 72% of the study participants who were also present in the EHR system during the same study period. Based on this subgroup, we compared their data from the EHR with those from the Predictive Health study and found that BMI and blood pressure measurements were considerably consistent (Ko et al., 2020). Lastly, the estimated effects may be compromised or biased in the presence of unmeasured confounders. Moreover, since the participation of this health partner program was entirely voluntary, the findings may not be generalized to all university employees.

While EHR data are maintained by healthcare systems for mostly administrative purposes, they provide a unique opportunity for researchers to develop a specific cohort to fill research gaps. Randomization is ideal to obtain an unbiased estimate of intervention effect, but when a control group is not available, identifying EHR synthetic controls may provide an alternative to estimate intervention effect appropriately. Further studies are indicated to explore other ways to develop study-specific cohorts from the EHR and similar data bases.

Funding

This work is supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378 and supported in part by 2020 Synergy II Nexus Award (Differentially-Private, Synthetic Controls for the Center for Health Discovery and Well-Being (CHDWB) Cohort: Data Science to Assess Health, Wellness and Disease) from the Woodruff Health Science Center of Emory University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Woodruff Health Science Center, Emory University, or the National Institutes of Health.

CRediT authorship contribution statement

Yi-An Ko: Conceptualization, Supervision, Formal analysis, Writing - original draft, Writing - review & editing. **Zhenchao Chen:** Data curation, Formal analysis. **Chang Liu:** Data curation, Formal analysis, Writing - review & editing. **Yingtian Hu:** Data curation, Formal analysis. **Arshed A. Quyyumi:** Conceptualization, Resources, Funding acquisition, Writing - review & editing. **Lance A. Waller:** Conceptualization, Methodology, Resources, Writing - review & editing. **Melinda Higgins:** Writing - review & editing. **Thomas R. Ziegler:** Writing - review & editing. **Kenneth L. Brigham:** Writing - review & editing. **Greg S. Martin:** Resources, Project administration, Funding acquisition, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Chad Robichaux for extracting electronic health records data, and Jane Clark for distributing the study data sets.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2021.101572.

References

Al Mheid, I., Kelli, H.M., Ko, Y.-A., Hammadah, M., Ahmed, H., Hayek, S., Vaccarino, V., Ziegler, T.R., Gibson, G., Lampl, M., Alexander, R.W., Brigham, K., Martin, G.S., Quyyumi, A.A., 2016. Effects of a Health-Partner Intervention on Cardiovascular Risk. J. Am. Heart Assoc. 5 (10).

Anderson, M., Naci, H., Morrison, D., Osipenko, L., Mossialos, E., 2019. A review of NICE appraisals of pharmaceuticals 2000–2016 found variation in establishing comparative clinical effectiveness. J. Clin. Epidemiol. 105, 50–59.

Benkeser, D., et al., 2017. Doubly robust nonparametric inference on the average treatment effect. Biometrika 104 (4), 863–880.

Brigham, K.L., 2010. Predictive health: the imminent revolution in health care. J. Am. Geriatr. Soc. 58 (Suppl 2), S298–302.

Cooper, G.R., Myers, G.L., Kimberly, M.M., Waymack, P.P., 2002. The effects of errors in lipid measurement and assessment. Curr Cardiol Rep 4 (6), 501–507.

Desai, J.R., Bowen, E.A., Danielson, M.M., Allam, R.R., Cantor, M.N., 2013. Creation and implementation of a historical controls database from randomized clinical trials. J. Am. Med. Inform. Assoc. 20 (e1), e162–e168.

Ferrara, A., Hedderson, M.M., Brown, S.D., Ehrlich, S.F., Tsai, A.-L., Feng, J., Galarce, M., Marcovina, S., Catalano, P., Quesenberry, C.P., 2020. A telehealth lifestyle intervention to reduce excess gestational weight gain in pregnant women with overweight or obesity (GLOW): a randomised, parallel-group, controlled trial. Lancet Diabetes Endocrinol. 8 (6), 490–500.

Funk, M.J., et al., 2011. Doubly robust estimation of causal effects. Am. J. Epidemiol. 173 (7), 761–767.

Gökbuget, N., Kelsh, M., Chia, V., Advani, A., Bassan, R., Dombret, H., Doubek, M., Fielding, A.K., Giebel, S., Haddad, V., Hoelzer, D., Holland, C., Ifrah, N., Katz, A., Maniar, T., Martinelli, G., Morgades, M., O'Brien, S., Ribera, J.-M., Rowe, J.M., Stein, A., Topp, M., Wadleigh, M., Kantarjian, H., 2016. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood Cancer J. 6 (9), e473.

Grundy, S.M., et al., 2019. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation 139 (25), E1046–E1081.

Kandula, N.R., et al., 2015. Translating a heart disease lifestyle intervention into the community: the South Asian Heart Lifestyle Intervention (SAHELI) study; a randomized control trial. BMC Public Health 15, 1064.

Keyserling, T.C., et al., 2016. A community-based lifestyle and weight loss intervention promoting a Mediterranean-style diet pattern evaluated in the stroke belt of North Carolina: the Heart Healthy Lenoir Project. BMC Public Health 16, 732.

Keyserling, T.C., Sheridan, S.L., Draeger, L.B., Finkelstein, E.A., Gizlice, Z., Kruger, E., Johnston, L.F., Sloane, P.D., Samuel-Hodge, C., Evenson, K.R., Gross, M.D., Donahue, K.E., Pignone, M.P., Vu, M.B., Steinbacher, E.A., Weiner, B.J., Bangdiwala, S.I., Ammerman, A.S., 2014. A comparison of live counseling with a web-based lifestyle and medication intervention to reduce coronary heart disease risk: a randomized clinical trial. JAMA Intern. Med. 174 (7), 1144.

Ko, Y.-A., Hu, Y., Quyyumi, A.A., Waller, L.A., Voit, E.O., Ziegler, T.R., Lampl, M., Martin, G.S., Shimosawa, T., 2020. Comparison of physical examination and

laboratory data between a clinical study and electronic health records. PLoS One 15 (7), e0236189.
Lloyd-Jones, D.M., et al., 2019. Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology (vol 73, pg 3153, 2019). J. Am. Coll. Cardiol. 73 (24), 3234.
Lotfaliany, M., Mansournia, M.A., Azizi, F., Hadaegh, F., Zafari, N., Ghanbarian, A., Mirmiran, P., Oldenburg, B., Khalili, D., 2020. Long-term effectiveness of a lifestyle intervention on the prevention of type 2 diabetes in a middle-income country. Sci. Rep. 10 (1) https://doi.org/10.1038/s41598-020-71119-2.
Mendell, J.R., Goemans, N., Lowes, L.P., Alfano, L.N., Berry, K., Shao, J., Kaye, E.M., Mercuri, E., 2016. Longitudinal effect of eteplirsen versus historical control on ambulation in Duchenne muscular dystrophy. Ann. Neurol. 79 (2), 257–271.
Peddada, S.D., Dinse, G.E., Kissling, G.E., 2007. Incorporating historical control data when comparing tumor incidence rates. J. Am. Stat. Assoc. 102 (480), 1212–1220.
Rask, K.J., Brigham, K.L., Johns, M.M.E., 2011. Integrating Comparative Effectiveness Research Programs Into Predictive Health: A Unique Role for Academic Health Centers. Acad. Med. 86 (6), 718–723.
Rubin, D.B., 2007. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat. Med. 26 (1), 20–36.
Thiese, M.S., 2014. Observational and interventional study design types; an overview. Biochem. Med. (Zagreb) 24 (2), 199–210.
Thorlund, K., et al., 2020. Synthetic and External Controls in Clinical Trials - A Primer for Researchers. Clin. Epidemiol. 12, 457–467.
van Dammen, L., Wekker, V., van Oers, A.M., Mutsaerts, M.A.Q., Painter, R.C., Zwinderman, A.H., Groen, H., van de Beek, C., Muller Kobold, A.C., Kuchenbecker, W.K.H., van Golde, R., Oosterhuis, G.J.E., Vogel, N.E.A., Mol, B.W.J., Roseboom, T.J., Hoek, A., Stepto, N.K., 2018. Effect of a lifestyle intervention in obese infertile women on cardiometabolic health and quality of life: A randomized controlled trial. PLoS One 13 (1), e0190662.
van der Laan, M.J., 2014. Targeted Estimation of Nuisance Parameters to Obtain Valid Statistical Inference. Int. J. Biostat. 10 (1), 29–57.
Viele, K., et al., 2014. Use of historical control data for assessing treatment effects in clinical trials. Pharm. Stat. 13 (1), 41–54.
Zhang, X., Devlin, H.M., Smith, B., Imperatore, G., Thomas, W., Lobelo, F., Ali, M.K., Norris, K., Gruss, S., Bardenheier, B., Cho, P., Garcia de Quevedo, I., Mudaliar, U., Jones, C.D., Durthaler, J.M., Saaddine, J., Geiss, L.S., Gregg, E.W., Barengo, N.C., 2017. Effect of lifestyle interventions on cardiovascular risk factors among adults without impaired glucose tolerance or diabetes: A systematic review and meta-analysis. PLoS One 12 (5), e0176436.