Framework for Research in Equitable Synthetic Control Arms

Nafis Neehal, M.S.[1], Vibha Anand, Ph.D.[2], Kristin P. Bennett, Ph.D.[1]
[1]Rensselaer Polytechnic Institute, Troy, NY;
[2]Center for Computational Health, IBM T.J. Watson Research Center, Cambridge, MA

Abstract

Randomized Clinical Trials (RCTs) measure an intervention's efficacy, but they may not be generalizable to a desired target population if the RCT is not equitable. Thus, representativeness of RCTs has become a national priority. Synthetic Controls (SCs) that incorporate observational data into RCTs have shown great potential to produce more efficient studies, but their equity is rarely considered. Here, we examine how to improve treatment effect estimation and equity of a trial by augmenting "on-trial" concurrent controls with SCs to form a Hybrid Control Arm (HCA). We introduce FRESCA – a framework to evaluate HCA construction methods using RCT simulations. FRESCA shows that doing propensity and equity adjustment when constructing the HCA leads to accurate population treatment effect estimates while meeting equity goals with potentially less "on-trial" patients. This work represents the first investigation of equity in HCA design that provides definitions, metrics, compelling questions, and resources for future work.

Introduction

A set of methodologies for supplementing randomized control trials (RCTs) with non-randomized evidence arms based on real-world data (RWD), e.g. health records or prior trials, has been rapidly emerging[1]. Synthetic control arms (SCAs) are of particular interest when there are challenges or difficulties in the recruitment of concurrent controls (CC) due to ethical and practical considerations in rare diseases[2]. The use of hybrid control arms (HCAs) combining both concurrent and synthetic controls offers the potential of preserving the benefits of randomization while strengthening evidence by integrating historical trial data. Augmenting SCA increases sample size, potentially increasing the statistical power and reducing the estimation variance without increasing the trial length or cost. Proper construction of HCAs is challenging because the distribution of the RCT and electronic health records (EHR) data typically differ. Thus, methods like propensity score matching are used to construct the HCAs.

Here, we introduce another benefit of HCAs, namely improving the equity of RCTs. We define "equity" of an RCT as the similarity between an RCT population and an investigator-defined target population with respect to a set of attributes of interest. We have used the terms "equity", "representativeness" and "generalizability" synonymously throughout our paper. RCTs may not be generalizable to clinical practice if the trial population is not representative of the population for which the treatment is intended. To address this, the National Institutes of Health, FDA, and health policymakers have made RCT equity a high priority. Trial designers must address the equity of subgroups defined by protected attributes (e.g. age, gender, and race/ethnicity). Subgroups are defined in terms of attributes that classify the population with a specific disease into groups for which we would like to see parity in terms of health outcomes received. Here, we examine how HCA can be used to achieve these equity goals for all subgroups in RCTs.

RCTs are designed to provide unbiased estimates of "Sample" average treatment effects, but they still may not be representative. A "Target" dataset can be used to estimate the prevalence of subgroups defined by protected attributes in the target population. Then, the RCT sample can be equity "adjusted" using resampling or reweighing to provide population treatment effects estimates for the target population[3]. Our goal is to estimate the target population treatment effect using the treatment arm and a hybrid control arm consisting of both concurrent controls (CC) and synthetic controls (SC). The resulting hybrid control arms must be created with "Propensity Adjustment" methods to correctly estimate the counterfactual outcomes of the treatment population and "Equity Adjustment" to create a population-specific treatment effect estimate for a specific target population while preserving equity.

To evaluate our current and future approaches for HCA construction, we create the "Framework for Research in Equitable Synthetic Control Arms (FRESCA)" shown in Fig 1. FRESCA has five main functions: Cohort Generation, Scenario Simulation, Treatment Effect and Equity Estimation, Target Subgroup Rates Calculation, and Assessment. We illustrate how FRESCA works on hypothetical trial scenarios inspired by the SPRINT (Systolic Blood Pressure Intervention Trial)[4,5]. FRESCA simulates four samples: Treatment Arm (TA), Concurrent Controls (CC), Biased

External Control (EC) Sample, and the Target Population. Together these four samples create a "scenario" representing a hypothetical clinical trial. To imitate randomization, we first take unbiased samples of the SPRINT trial subjects to create the TA and CC Cohorts. To represent different types of RWD, we take a biased sample of SPRINT to create the Biased EC Sample from which the SCs are selected to augment the RCT data. As in prior studies[6], we define subgroups with respect to three protected attributes (i.e. age, gender, race/ethnicity) and use NHANES[7] (National Health and Nutrition Examination Survey) to estimate the target population rates for these subgroups. We create HCA for this scenario using various Propensity and Equity based Hybrid Control Adjustment (PEHCA) methods. We estimate the target population Hazard Ratio (PHR) based on SPRINT's primary outcome using Cox proportional hazard modeling. Then we estimate the quality of the HR estimate and the equity of the HCA, i.e. how representative the HCA is of the target population.

This paper makes the following contributions:

- We identify and define the issue of RCT not being equitable with respect to a target population when using HCA consisting of CC and SC.
- We develop the FRESCA framework for empirically investigating and evaluating methods for creating equitable HCAs. FRESCA introduces metrics for assessing the accuracy of the treatment effect estimates and equity.
- We define and investigate an HCA construction strategy that can be used with combinations of propensity and equity adjustment methods. FRESCA can be used with any desired propensity and equity adjustment method.
- We empirically estimate the performance of these HCA methods with three sets of Biased EC, created from RWD, representing different degrees of bias. We demonstrate that equity is a problem for HCA construction methods, and adjusting for both propensity and equity results in equitable HCA with accurate population treatment effects estimation, especially for RWD that are biased with respect to the trial population.
- We explore the impact of CC size on equity and treatment effect estimation. The empirical results suggest that propensity and equity adjustment methods for HCA construction can potentially produce an accurate and equitable treatment effect estimate even with use of smaller CC cohorts than the RCT originally recruited.

This work is an empirical study of the importance of equity in designing SC arms which provides a foundation for future research in equitable HCA. Several methods exist for augmenting SC population into the trial population. Most of them used standard propensity score matching to select the appropriate external control populations to augment the trial. Some of them used propensity to be treated[9,8], others used propensity to be in the trial[10,11]. But none of them had considered how to adjust for equity with respect to a target population. The existing methods for generating target population estimates from RCT data typically involved re-weighting of the trial population outcomes in different ways[3,12]. But none of these have shown how to extend their methods when augmenting synthetic controls to the trial population. In future work, FRESCA could incorporate these other SCA and population treatment estimation methods to create many new methods for equitable SC arms and evaluate their effectiveness. We leave a more theoretical exploration of the subject as well as other HCA algorithms to future work.

Methodology

Problem Definition We formally define the problem using the potential outcomes framework. Let three distinct finite samples be drawn from an infinite population satisfying the RCT eligibility criteria. Y_{ist} represents the potential outcome of subject i assigned to sample s and treatment t. Here $s = 0$ indicates membership in the target population, $s = 1$ is membership in the RCT population, $s = 2$ is membership in the observed RWD, and $s = 3$ indicates membership in an adjusted sample possibly consisting of both RCT and observed EHR data. For simplicity, we assume that each on-trial subject is randomly assigned to treatment or control. We assume S_i is a sample indicator variable, taking on value s and T_i is a treatment indicator variable. We assume that each subject has covariates that may affect the assignment to the four samples S_i and the outcome Y_i. The "protected attributes" (e.g. age, gender, and race/ethnicity) are a subset of these covariates. Let effect(Y_{11}, Y_{10}) be the hazard ratio or some other measure of difference in treatment effect. The Sample Hazard Ratio (SHR) for the RCT data is $SHR = E(\text{effect}(Y_{11}, Y_{10})|S =$

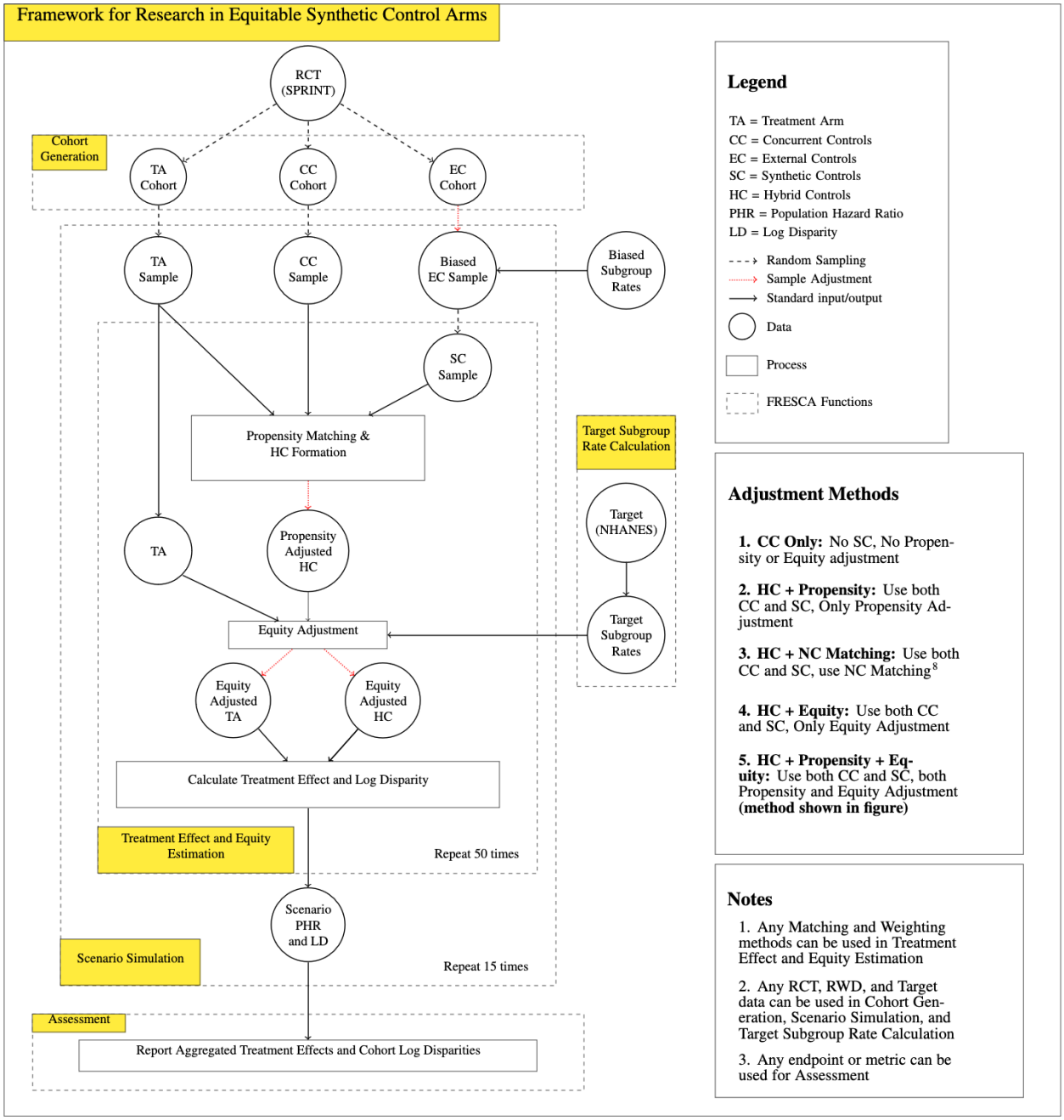


Figure 1: Functions and Data Flow in FRESCA

1). Randomization allows the treatment effect to be directly computed using standard methods using RCT TA and CC arms. Our goal is to estimate the Population Hazard Ratio (PHR) for a specified target population (s=0), which is defined as $PHR = E(\text{effect}(Y_{11}, Y_{10})|S = 0)$ which requires adjustment. If the RCT sample is not representative of the target population, then we say that RCT is not "equitable" with respect to the target population. For example, $P(Asian|S = 0) \neq P(Asian|S = 1)$. This "inequity" may result in inaccurate treatment effect estimation for the target population, i.e. $PHR \neq SHR$. Using the RCT and target data, PHR methods work by estimating the target rates for desired subgroups defined over the sensitive attributes, and then effectively creating a matching "equity adjusted" sample $s = 3$. PHR is estimated by calculating the treatment effect on the adjusted sample. The "equity"

Table 1: Distribution of Protected Attributes in FRESCA Cohorts, Biased External Control Samples, and NHANES Target Subgroup Rates

Attributes	Cohorts		Biased EC Samples			Target Subgroup Rate (NHANES)
	TA (N=4234)	CC (N=2000)	EC (unbiased) (N=2200)	Biased EC (Veteran) (N=2200)	Biased EC (High Risk) (N=2200)	
Age Group						
40-59	923 (21.8%)	438 (21.9%)	455 (20.7%)	1344 (61.1%)	334 (15.2%)	31.2%
59+	3311 (78.2%)	1562 (78.1%)	1745 (79.3%)	856 (38.9%)	1866 (84.8%)	68.8%
Gender						
Female	1499 (35.4%)	691 (34.6%)	763 (34.7%)	617 (28.0%)	631 (28.7%)	55.4%
Male	2735 (64.6%)	1309 (65.5%)	1437 (65.3%)	1583 (72.0%)	1569 (71.3%)	44.6%
Race or Ethnicity						
Hispanic	479 (11.3%)	225 (11.3%)	233 (10.6%)	391 (17.8%)	389 (17.7%)	10.0%
NH Asian	42 (1.0%)	15 (0.8%)	15 (0.7%)	149 (6.8%)	138 (6.3%)	3.9%
NH Black	1232 (29.1%)	616 (30.8%)	647 (29.4%)	448 (20.4%)	468 (21.3%)	12.0%
NH White	2451 (57.9%)	1128 (56.4%)	1291 (58.7%)	1180 (53.6%)	1172 (53.3%)	69.3%
Other	30 (0.7%)	16 (0.8%)	14 (0.6%)	32 (1.5%)	33 (1.5%)	4.8%

of the adjusted sample with respect to the target can be quantified using "log disparity" (defined below) for each subgroup[6]. We use a causal estimation method (propensity score matching) to adjust EC data to construct the HC arm. Propensity matching estimates $p(T|X)$ and then effectively used a new matched control sample $s = 3$ for estimation of the HR. The question examined in this work is how can we use both RCT (s=1) and RWD data (s=2), to create a new sample (s=3) that is equitable, i.e. representative of the target population, and that correctly estimates PHR. For simplicity of presentation, we assume that the treatment is new and does not occur in the RWD data. Thus the goal is to create an appropriate SCA from the observed data and use it to construct an HCA consisting of both RCT and RWD subject data. Our method performs both equity adjustment and matching to accurately estimate the PHR.

We do an empirical simulation to estimate the potential pitfalls and solutions of our proposed approach and we leave a more formal analysis of this scenario to future work. Our proposed novel framework FRESCA enables researchers to examine the effectiveness of HCA formation strategies using simulations inspired by actual clinical trials. For this paper, we use the published SPRINT RCT data to simulate the RCT ($s = 1$) and the biased RWD data ($s = 2$), and NHANES to determine the target population ($s = 0$). We use combinations of propensity and equity adjustment methods to create adjusted samples ($s = 3$). FRESCA clearly shows that doing propensity matching alone can result in inequitable sampling. Similarly, doing just equity adjustments can result in inaccurate PHR estimates. Thus we propose an approach that performs both matching and equity adjustment. We demonstrate empirically that it may be possible to use smaller CC patient cohorts while maintaining equitable estimates with accurate PHR.

Data As in Qi et al[6], we define the target population based on a nationally representative hypertensive population sample from NHANES 2015–2016. We seek representativeness of our RCT based on three protected attributes: Age Group (40-59, 59+), Gender (Male, Female), and Race/Ethnicity (Non-Hispanic Black, Non-Hispanic White, Non-Hispanic Asian, Hispanic, Other). The target rates for each subgroup are calculated using survey-weighted analysis.

We use SPRINT[4] from BioLINCC as the RCT data. SPRINT was a randomized, single-blinded treatment trial with participants randomized to two different treatments. Eligible participants were adults 50+ years with high systolic blood pressure with an increased risk of cardiovascular disease but without diabetes or a history of stroke. After removing patients' missing baseline data, SPRINT contains 4234 treated and 4200 control patients. We analyzed time to the primary composite outcome of myocardial infarction, acute coronary syndrome not resulting in myocardial infarction, stroke, acute decompensated heart failure, or death from cardiovascular causes. To estimate the treatment effect, we use Cox's proportional hazards regression on the TA and HCA as implemented in the "Survival" R package. We report the Hazard Ratio and 95% confidence interval.

As described in Fig. 1, FRESCA splits SPRINT data into the TA, CC, and EC Cohorts. Then an adjusted bootstrapped sample is made to create three distinct control populations representing the different scenarios (unbiased, Veteran, and

Table 2: Distribution of Additional Covariates in FRESCA Cohorts and Biased External Control Samples.

Attributes	Cohorts		Biased EC Samples		
	TA (N=4234)	CC (N=2000)	EC (unbiased) (N=2200)	Biased EC (Veteran) (N=2200)	Biased EC (High Risk) (N=2200)
Educational Status					
College grad and above	1673 (39.5%)	844 (42.2%)	869 (39.5%)	994 (45.2%)	885 (40.2%)
Below HSG	83 (2.0%)	32 (1.6%)	44 (2.0%)	41 (1.9%)	69 (3.1%)
HSG/GED	698 (16.5%)	319 (16.0%)	353 (16.0%)	295 (13.4%)	371 (16.9%)
Some college/TS	1780 (42.0%)	805 (40.3%)	934 (42.5%)	870 (39.5%)	875 (39.8%)
Smoker?					
No smoke	3633 (85.8%)	1718 (85.9%)	1907 (86.7%)	1800 (81.8%)	1811 (82.3%)
Smoke	601 (14.2%)	282 (14.1%)	293 (13.3%)	400 (18.2%)	389 (17.7%)
Fasting Glucose level					
Glucose 100-125	1648 (38.9%)	791 (39.6%)	841 (38.2%)	917 (41.7%)	915 (41.6%)
Glucose<100	2453 (57.9%)	1147 (57.4%)	1293 (58.8%)	1211 (55.0%)	1235 (56.1%)
Glucose>=126	133 (3.1%)	62 (3.1%)	66 (3.0%)	72 (3.3%)	50 (2.3%)
Total Cholesterol					
High TC	1524 (36.0%)	735 (36.8%)	770 (35.0%)	823 (37.4%)	745 (33.9%)
Normal TC	2710 (64.0%)	1265 (63.3%)	1430 (65.0%)	1377 (62.6%)	1455 (66.1%)
Average of 3 sitting Systolic BP					
SBP 130-139	1547 (36.5%)	733 (36.7%)	840 (38.2%)	932 (42.4%)	783 (35.6%)
SBP>=140	2687 (63.5%)	1267 (63.4%)	1360 (61.8%)	1268 (57.6%)	1417 (64.4%)
Has Clinical or Subclinical CVD					
No	3404 (80.4%)	1625 (81.3%)	1750 (79.5%)	1815 (82.5%)	1194 (54.3%)
Yes	830 (19.6%)	375 (18.8%)	450 (20.5%)	385 (17.5%)	1006 (45.7%)
Framingham Risk Score					
High	894 (21.1%)	438 (21.9%)	429 (19.5%)	294 (13.4%)	984 (44.7%)
Moderate	3340 (78.9%)	1562 (78.1%)	1771 (80.5%)	1906 (86.6%)	1216 (55.3%)
Serum Creatinine					
Mean (SD)	1.18 (4.02)	1.05 (0.319)	1.06 (0.328)	1.16 (0.411)	1.11 (0.358)
Estimated GFR within past 6 months					
Disease	1104 (26.1%)	486 (24.3%)	579 (26.3%)	874 (39.7%)	686 (31.2%)
Normal	3130 (73.9%)	1514 (75.7%)	1621 (73.7%)	1326 (60.3%)	1514 (68.8%)

High Risk). Table 1 shows that the distribution of the protected attributes used for equity analysis for the FRESCA cohorts and the corresponding target rate from NHANES can significantly vary from each other. They exhibit inequities when compared with the corresponding target rate from NHANES. Table 2 shows that the rates of other SPRINT covariates vary significantly across cohorts and the target. The attributes in Table 1 and Table 2 are used for propensity matching. To emulate a typical trial, only the protected attributions in Table 1 are used for equity adjustment.

Adjustment Methods We use standard propensity score matching to balance the distributional differences between the synthetic control population and the trial population[13] using the "MatchIt" R package. For equity adjustment and creation of the Biased EC Cohorts, we use Iterative Proportional Fitting with sampling (IPF)[14] as implemented in the "IPFR" R package. IPF is used to adjust marginal distributions reported in a given source dataset to match marginal distributions in the target dataset[15]. Note that any desired method can be used for equity or propensity adjustments.

Metrics For assessing the equity, we use log disparity (LD) here defined as the absolute difference between the log odds of a subgroup in a given sample (e.g. Females in HCA) and the log odds of a subgroup being in the target (i.e. Females in NHANES). LD (without the absolute value) was originally designed to quantify the representativeness of randomized clinical trials[6]. As done previously for clinical trials, LD values between 0 and 0.22 are considered to be

equitable. We calculate LD individually for each possible subgroup and take the median overall sample scenarios since empty subgroups may have infinite LD. The Cohort Log Disparity (CLD) is defined as the mean of all the LDs (absolute value) calculated for all the attributes in Table 1. CLD provides an assessment of the overall representativeness of the cohort with respect to the target population.

FRESCA Framework We designed FRESCA to help answer the question of how well do HCA construction methods work. The goal is to assess how well propensity and equity adjustment methods calculate the population treatment effect in terms of accuracy of the estimated Population HR (PHR) and equity. As shown in Fig 1, FRESCA has five main functions: Cohort Generation, Scenario Simulation, Target Subgroup Rates Calculation (described above), Treatment Effect and Equity Estimation, and Assessment. In FRESCA, we describe them for SPRINT, NHANES, and PHR, but FRESCA can easily generalize to other RCT and RWD data and other types of endpoints respectively.

FRESCA Cohort Generation partitions SPRINT to create TA, CC, and EC cohorts, from which many random scenarios with similar characteristics can be drawn. We create five different cohorts. TA Cohort ($N = 4234$) contains all the treatment group patients. All the available control patients are randomly partitioned into CC Cohort($N = 2000$) and EC Cohort ($N = 2200$). FRESCA simulates hybrid clinical trial "scenarios" by drawing unbiased samples to create TA Sample (size N_{TA}) and CC Sample (size N_{CC}) from their respective cohort. FRESCA adjusts the Biased EC Sample using IPF to match the desired Biased Subgroup Rates. We hypothesized that bias in the EC Sample could affect the estimation of PHR and equity. Thus we adjusted EC Cohort to create three different Biased EC Samples: 1) Unbiased: Data is sampled according to the original SPRINT trial distribution. 2) Veterans: The distribution of Gender and Race/Ethnicity matches US Army Veteran Population[16] (Gender: 72.6% Male, 27.4% Female, Race/Ethnicity: Non-Hispanic Black: 20.2%, Non-Hispanic White: 54%, Hispanic: 17.2%, Non-Hispanic Asian: 6.9%, Others: 1.7%). We also bias Age Group (40-59: 60%, 59+: 40%) and Estimated GFR (Normal: 60%, Disease: 50%) variable to create a healthier population. 3) High Risk: Distribution of Gender and Race/Ethnicity is set to match the US Army Veteran Population(Gender: 72.6% Male, 27.4% Female, Race/Ethnicity: Non-Hispanic Black: 20.2%, Non-Hispanic White: 54%, Hispanic: 17.2%, Non-Hispanic Asian: 6.9%, Others: 1.7%). We also bias CVD History (Yes: 45%, No: 55%) and Framingham Risk Score (FRS) variable (High: 45%, Moderate: 55%) to create a sicker population. The three different EC distributions are intentionally created to produce biased EC sample groups. This simulates the situation in which the EHR dataset used to construct EC has a significantly different distribution than the trial (here SPRINT) and target (here NHANES) populations.

FRESCA Scenario Generation creates hypothetical clinical trials. The TA Sample and CC Sample represent the patients enrolled in the trial with random assignment. The Biased EC Sample ($N = 2200$) represents the potentially biased RWD available for SCA construction. FRESCA can create many different experimental scenarios. The user can select the size of all the samples to experiment with. Throughout all our experiments here, we use a TA sample size (N_{TA}) of 2000 and vary the CC sample size (N_{CC}) to be 0, 500, 1000, 1500, and 2000.

FRESCA Treatment Effect and Equity Estimation then estimates the population treatment effect using Propensity and Equity based Hybrid Control Adjustment (PEHCA). The goals are to construct an HCA that is matched and balanced with respect to TA, and to make both TA and HCA representative of the target. Assuming that the user specifies a 1:1 randomization (i.e. the number of patients in the TA and HCA should be equal), PEHCA uses the following strategy. - 1) PECHA randomly draws a SC Sample of size $N_{SC} = N_{TA} - N_{CC}$ from SC Sample. HCA is created by merging SC Sample and CC Sample. 2) Propensity matching is used to create the "Propensity Adjusted HCA". 3) Both TA and HCA are equity adjusted to match the target population. 4) PHR is estimated using Cox Regression and LD is calculated using the propensity and equity adjusted TA and HC. 5) This process is repeated $N = 50$ times with different random SC Samples and the mean PHR is returned as the PHR estimate.

To make the propensity adjustments, we define a variable Z, where $Z = 1$ for patients in TA and CC samples and $Z = 0$ for patients in Biased EC Sample. A standard binary logistic regression model is trained with TA sample, CC sample, and all of Biased EC Sample to learn the desired underlying propensity function. So for CC sample patients, the score is $P(Z = 1|X = x)$ where X are the covariates used for propensity adjustment; and $Z = 0$ would be for all the Biased EC Sample patients. This score defines how likely a patient is to be in the trial. Based on this score, all SC Samples are matched (with replacement) against TA samples. We then form HCA by combining matched SC Samples with CC. After the propensity adjustments are made, we make equity adjustments on both TA and the propensity-

Table 3: Comparison of Population Hazard Ratio and Log Disparity across different methods. We show this for with 2000 and 1000 patients in TA and CC sample cohort respectively. Bold (*****) symbol in Hazard Ratio column indicates estimated HR being significantly different ($p < 0.05$) from "Ground Truth" Target PHR. Bold (†) symbol in Log Disparity column indicates measured Log Disparity not being within equitable range. ($LD > 0.22$)

Population Reference	Control Population	Adjustment Method	Hazard Ratio [95% Confidence Interval]	Cohort Log Disparity [95% Confidence Interval]
High Risk	CC	None	0.753 [0.692, 0.815]	**0.881 [0.807, 0.983]**†
	HC	Propensity	0.737 [0.689, 0.785]	**0.843 [0.792, 0.900]**†
	HC	NC Matching	0.747 [0.681, 0.819]	**0.893 [0.821, 0.950]**†
	HC	Equity	**0.599 [0.525, 0.671]***	0.007 [0.004, 0.013]
	HC	Propensity + Equity	0.752 [0.668, 0.837]	0.007 [0.003, 0.011]
Veterans	CC	None	0.753 [0.692, 0.815]	**0.881 [0.807, 0.983]**†
	HC	Propensity	0.751 [0.718, 0.784]	**0.769 [0.719, 0.820]**†
	HC	NC Matching	0.742 [0.704, 0.793]	**0.797 [0.758, 0.911]**†
	HC	Equity	0.728 [0.662, 0.793]	0.007 [0.004, 0.011]
	HC	Propensity + Equity	0.751 [0.687, 0.816]	0.007 [0.004, 0.011]
Unbiased	CC	None	0.753 [0.692, 0.815]	**0.881 [0.807, 0.983]**†
	HC	Propensity	0.741 [0.697, 0.784]	**0.889 [0.841, 0.937]**†
	HC	NC Matching	0.753 [0.684, 0.821]	**0.874 [0.815, 0.943]**†
	HC	Equity	0.772 [0.701, 0.844]	0.006 [0.004, 0.012]
	HC	Propensity + Equity	0.771 [0.698, 0.843]	0.006 [0.003, 0.012]
Target Population	All Controls	Equity	0.798 [0.781, 0.817]	0.031 [0.022, 0.040]

adjusted HC. We use IPF to match these two cohorts to target. IPF returns weights for each patient in each of these two cohorts. We denote W_{IPF_TA} and W_{IPF_HC} as the weight vectors for TA and HC respectively. We then take a bootstrap sample of both TA and HC with corresponding IPF weights maintaining the same corresponding sizes. This concludes the step for equity adjustment. Once these adjustments are made, we estimate the treatment effect (PHR) and measure the equity value (LD).

FRESCA Assessment provides methods for evaluating different strategies for HCA construction on different types RWD. Here we investigate four different strategies for HCA construction where we combine propensity and equity adjustment methods to create 4 different combinations. We also compare these results to the existing NC Matching method[8]. The five methods are described in more detail in Fig 1. We then evaluate the effect of each adjustment procedure for three sets of biased external controls as defined above. FRESCA assesses the accuracy of the PHR estimated by different HCA construction methods by comparing it to the "ground-truth" Target PHR value for any statistical significance differences. This Target PHR value is defined to be the HR value obtained by doing only equity adjustment on the whole SPRINT dataset using all of the original target and control subjects. In this case, the whole SPRINT dataset is divided into two groups containing all the treated patients in one group and all the control patients in another, without splitting this control cohort any further. Then equity adjustments are done on these two populations to match NHANES and estimate the Target PHR from that. The final Target PHR is estimated by taking a mean over 50 bootstrap samples of same size of these two populations. To examine equity, we investigate if the reported Log Disparity value is within [0, 0.22] according to the 80% rule[6].

For this paper, we generated and examined 225 different scenarios. We generated five different Cohorts with random seeds and each of those had three different Biased EC Cohorts. Then for each of these, Scenario Generation was applied with the number of CC Samples (N_{CC}) set to 0, 500, 1000, 1500, and 2000. For each Scenario with the same settings (for a fixed TA Sample Size, CC Sample Size, and a fixed Biased EC Sample), we repeat this process 15 times. We report the mean over these 15 estimates of mean PHR values along with 95% confidence intervals in Table 3. For equity, we report the mean value of 15 median estimates of LD. We report CLD value defined above calculated from these 15 estimates in Table 3 along with 95% confidence interval.

Results

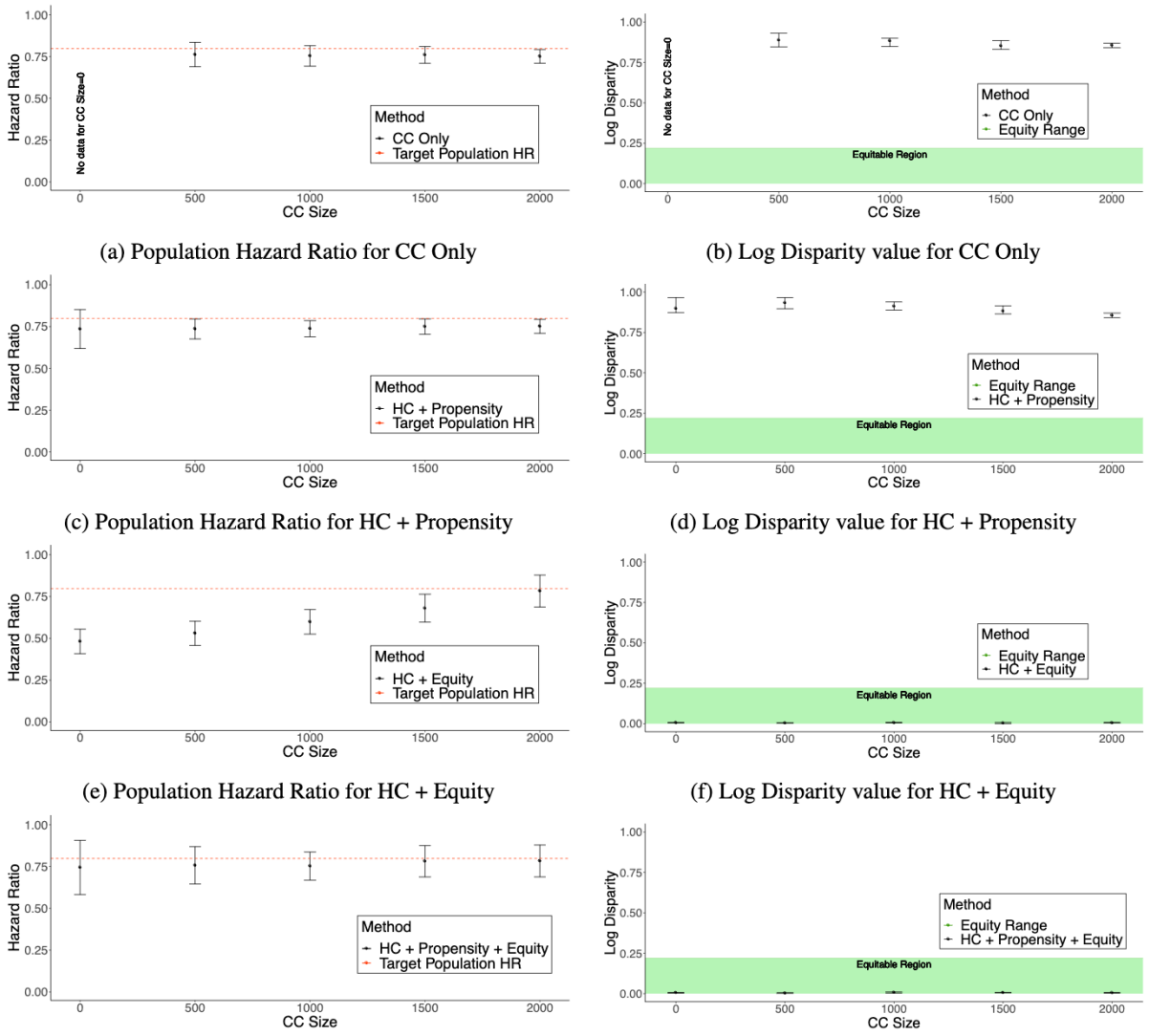
Comparison of Different HCA Construction Methods We evaluate how the treatment effect and equity vary for different methods for HCA construction. In Table 3, we compare the mean hazard ratios and mean of log disparity values for gender, race/ethnicity, and age subgroups with corresponding 95% confidence intervals for all four methods we described, along with the performance of one existing HCA construction method (NC Matching [8]). We compare the performance of these methods against the Target PHR created using all RCT data and examine equity using Log Disparity.

We observe that for all three Biased EC Cohorts, failing to perform equity adjustment led to inequitable trials as measured by LD. In all cases, propensity adjustment alone led to PHR estimates that were not significantly different from the Target PHR, both in the cases of NC Matching and the propensity adjustment method we followed. Similar outcomes were observed with just using CC and making no propensity or equity adjustments. In the "High Risk" Cohort, equity adjustment alone produced an inaccurate PHR estimate which indicates that only performing equity adjustment may not be sufficient to estimate PHR accurately. However, propensity adjustment combined with equity adjustment achieved an accurate estimation of target population HR with a representative control population in all cases.

Examination of Variation in CC Size We further examine how estimated PHR (for whole population) and LD (for Gender=Female subgroup) vary with CC Sample size for all four HCA construction methods. For each HCA construction method, we present the variations in mean PHR on the left and median LD for Females on the right with 95% confidence intervals for measures in Fig 2. Fig 2 (a,c), show that the methods without equity adjustment do not achieve acceptable equity values for all CC sample sizes while methods with equity adjustment (e,g) do. The equity for Race/Ethnicity is even worse for "CC Only". For some of the low CC sizes (e.g. 100, 200), the Non-Hispanic Asian or Others subgroups have no patients in CC. There are no PHR estimates and LD values available for CC Size = 0 since all subgroups are empty. In Fig 2(a) for "CC Only", the estimated PHRs are accurate since they include the Target PHR within the 95% confidence interval. Note the decreasing variance of HR as the CC size grows. In Fig 2(c) for "HC + Propensity", the PHR estimate is accurate and the variance is improved since SC is used to augment the sample. In Fig 2(e) for "HC + Equity", the PHR estimate is significantly off but improves as CC size increases (or equivalently as SC size decreases). In Fig 2(g) for "HC + Propensity + Equity", we see that combining propensity and equity adjustment finds accurate PHRs with roughly constant variance for all CC Sample sizes. Our findings presented in Fig 2 remain consistent with our conclusion from Table 3 for all CC Sizes and demonstrate that by making both propensity and equity adjustments, we can still get an accurate estimate of PHR and preserve equity. Remarkably, this holds even if we use CC Sizes 50% less than the size of the original trial. We leave choosing the optimal CC Size selection and designing equitable HCA trials as future research problems.

Conclusion

We present a novel framework, FRESCA, to empirically investigate new HCA construction methods to estimate target-specific population treatment effects. We introduced a strategy for HCA creation with propensity and equity based adjustment methods. We examined five different combinations of propensity and equity adjustment methods and assessed the robustness and overall performance of these approaches across three different EC Samples with different degrees of bias. We demonstrated that constructing HCA using propensity adjustment alone can lead to problems with equity. Hence propensity and equity adjustments may be necessary for HCA construction to make accurate treatment effect estimates while ensuring equity, especially when the RWD is highly biased. We also empirically showed that by using propensity and equity based hybrid control adjustment, it may be possible to make accurate PHR estimates with equity with fewer concurrent controls than originally recruited in the RCT. We have currently explored only one method for propensity adjustment and one method for equity adjustment, and one existing HCA construction algorithm. In future, FRESCA can be used with any standard covariate balancing methods like Coarsened Exact Matching [17] and other equity adjustment methods like re-weighting along with additional sets of protected attributes to adjust such as geographic distribution or income etc. New and more sophisticated propensity and equity based HCA methods could be developed and evaluated using FRESCA too. Also, we showed results using SPRINT and NHANES as the RCT and target population respectively, but FRESCA can be readily extended to evaluate any trial or target population desired, and will allow using real EHR/RWD data sources to be used as external control. Furthermore, we only examined time-to-event survival data as our primary endpoint, but FRESCA provides the utility to easily examine



(a) Population Hazard Ratio for CC Only

(b) Log Disparity value for CC Only

(c) Population Hazard Ratio for HC + Propensity

(d) Log Disparity value for HC + Propensity

(e) Population Hazard Ratio for HC + Equity

(f) Log Disparity value for HC + Equity

(g) Population Hazard Ratio for HC + Propensity + Equity

(h) Log Disparity value for HC + Propensity + Equity

Figure 2: Population hazard ratio and equity for increasing concurrent control sample sizes using "High Risk" cohort for biased External Controls. Log Disparity is shown for Gender=Female subgroup only.

other types of endpoints too. In future, we plan to provide a more theoretical exploration of PEHCA methods. We leave further exploration of all of these issues for future work. To conclude, this is an important first investigation into equity issues during HCA construction that identified many promising research directions. FRESCA being a modular framework provides the utility for our research peers to investigate other HCA approaches and test metrics to compare results. We hope this work will significantly contribute to efficient and effective clinical trial design processes in the future.

Acknowledgments

This work was primarily funded by IBM Research. This manuscript was prepared using Systolic Blood Pressure Intervention Trial (SPRINT) study research materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the SPRINT or the NHLBI. Thanks to John Erickson for help with code optimization.

References

1. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Statistics in medicine. 2014;33(7):1242-58.

2. Li C, Ferro A, Mhatre SK, Lu D, Lawrance M, Li X, et al. Hybrid-control arm construction using historical trial data for an early-phase, randomized controlled trial in metastatic colorectal cancer. Communications Medicine. 2022;2(1):90.

3. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. Journal of the Royal Statistical Society Series A (Statistics in Society). 2015:757-78.

4. Group SR. A randomized trial of intensive versus standard blood-pressure control. New England Journal of Medicine. 2015;373(22):2103-16.

5. Drawz PE, Pajewski NM, Bates JT, Bello NA, Cushman WC, Dwyer JP, et al. Effect of intensive versus standard clinic-based hypertension management on ambulatory blood pressure: results from the SPRINT (Systolic Blood Pressure Intervention Trial) ambulatory blood pressure study. Hypertension. 2017;69(1):42-50.

6. Qi M, Cahan O, Foreman MA, Gruen DM, Das AK, Bennett KP. Quantifying representativeness in randomized clinical trials using machine learning fairness metrics. JAMIA Open. 2021;4(3):ooab077.

7. Cntrs. for Disease Control and Prevention. National Cnt. for Health Statistics. National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention;. Accessed: 2022-02-01. `https://www.cdc.gov/nchs/nhanes/index.htm`.

8. Yuan J, Liu J, Zhu R, Lu Y, Palm U. Design of randomized controlled confirmatory trials using historical control data to augment sample size for concurrent controls. Journal of biopharmaceutical statistics. 2019;29(3):558-73.

9. Stuart EA, Rubin DB. Matching with multiple control groups with adjustment for group differences. Journal of Educational and Behavioral Statistics. 2008;33(3):279-306.

10. Yin X, Mishra-Kalyan PS, Sridhara R, Stewart MD, Stuart EA, Davi RC. Exploring the potential of external control arms created from patient level data: a case study in non-small cell lung cancer. Journal of Biopharmaceutical Statistics. 2022;32(1):204-18.

11. Harton J, Segal B, Mamtani R, Mitra N, Hubbard RA. Combining real-world and randomized control trial data using data-adaptive weighting via the on-trial score. Statistics in Biopharmaceutical Research. 2022:1-13.

12. Ling AY, Montez-Rath ME, Carita P, Chandross K, Lucats L, Meng Z, et al. A critical review of methods for real-world applications to generalize or transport clinical trial findings to target populations of interest. arXiv preprint arXiv:220200820. 2022.

13. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research. 2011;46(3):399-424.

14. Deming WE, Stephan FF. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics. 1940;11(4):427-44.

15. Lomax N, Norman P. Estimating population attribute values in a table:“get me started in” iterative proportional fitting. The Professional Geographer. 2016;68(3):451-61.

16. National Center for Veterans Analysis and Statistics, USDVA;. Accessed: 2022-02-01. `https://www.va.gov/vetdata/veteran_population.asp`.

17. Iacus SM, King G, Porro G. Causal inference without balance checking: Coarsened exact matching. Political Analysis. 2012;20(1):1-24.