A Novel Digital Twin Strategy to Examine the Implications of Randomized Control Trials for Real-World Populations

Phyllis M. Thangaraj[1*], Sumukh Vasisht Shankar[1*], Sicong Huang[3], Girish, Nadkarni[4,5], Bobak Mortazavi[3], Evangelos K. Oikonomou[1], and Rohan Khera[1,2,6,7]

[1]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

[2]Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT

[3]Department of Computer Science and Engineering, Texas A&M University, College Station, TX

[4]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[5]The Division of Data Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[6]Section of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT

[7]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA

[*]Contributed Equally

***Address for correspondence:**

Rohan Khera, MD, MS

195 Church St, 6th Floor, New Haven, CT 06510

203-764-5885; rohan.khera@yale.edu; @rohan_khera

Abstract:
Randomized clinical trials (RCTs) are essential to guide medical practice; however, their generalizability to a given population is often uncertain. We developed a statistically informed Generative Adversarial Network (GAN) model, RCT-Twin-GAN, that leverages relationships between covariates and outcomes and generates a digital twin of an RCT (RCT-Twin) conditioned on covariate distributions from a second patient population. We used RCT-Twin-GAN to reproduce treatment effect outcomes of the Systolic Blood Pressure Intervention Trial (SPRINT) and the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Blood Pressure Trial, which tested the same intervention but had different treatment effect results. To demonstrate treatment effect estimates of each RCT conditioned on the other RCT patient population, we evaluated the cardiovascular event-free survival of SPRINT digital twins conditioned on the ACCORD cohort and vice versa (SPRINT-conditioned ACCORD twins). The conditioned digital twins were balanced by the intervention arm (mean absolute standardized mean difference (MASMD) of covariates between treatment arms 0.019 (SD 0.018), and the conditioned covariates of the SPRINT-Twin on ACCORD were more similar to ACCORD than a sprint (MASMD 0.0082 SD 0.016 vs. 0.46 SD 0.20). Most importantly, across iterations, SPRINT conditioned ACCORD-Twin datasets reproduced the overall non-significant effect size seen in ACCORD (5-year cardiovascular outcome hazard ratio (95% confidence interval) of 0.88 (0.73-1.06) in ACCORD vs median 0.87 (0.68-1.13) in the SPRINT conditioned ACCORD-Twin), while the ACCORD conditioned SPRINT-Twins reproduced the significant effect size seen in SPRINT (0.75 (0.64-0.89) vs median 0.79 (0.72-0.86)) in ACCORD conditioned SPRINT-Twin). Finally, we describe the translation of this approach to real-world populations by conditioning the trials on an electronic health record population. Therefore, RCT-Twin-GAN simulates the direct translation of RCT-derived treatment effects across various patient populations with varying covariate distributions.

BACKGROUND

Randomized control trials (RCTs) generate evidence that defines optimal clinical practices, but their generalizability to real-world patient populations is often challenging to quantify.[1,2] This is a concern because RCTs often have underrepresentation from several demographic and clinical subpopulations[3–7] and varying treatment effects among individuals with certain characteristics.[8–10] These considerations are critical to translating information from RCTs to real-world patient populations,[11,12] but no strategies exist to evaluate how they may affect the applicability to patients in these settings.

The differences across RCTs testing similar interventions with discrepant treatment effects are one of the manifestations of potential issues arising out of the generalizability of interventions tested in RCTs.[13–19] For example, the Systolic Blood Pressure Intervention Trial (SPRINT) was a treatment intervention RCT that showed improved cardiovascular outcomes with intensive blood pressure control.[13] In contrast, the Action to Control Cardiovascular Risk in Diabetes Blood Pressure (ACCORD) trial did not find improved cardiovascular outcomes with the same intervention.[14] Among the explanations posited for these discrepant findings, differences in population composition and event rates are frequently suggested,[20–23] but despite experimental evidence available from two trials, there is no quantitative strategy to evaluate these assertions explicitly. Therefore, while it is critical to evaluate whether the effects observed in an RCT population generalize to a second population – either a planned second RCT or a general population of patients with the condition – it is currently challenging to examine these effects across the complex differences across multiple population characteristics.

Digital twins represent a strategy to create a synthetic representation of complex systems that replicates their underlying structure. Specifically, digital twin synthesis through deep

generative models such as generative Adversarial Networks (GANs) can integrate multiple patient population characteristics by constructing a synthetic cohort that accurately represents their covariate distributions. While GANs have been utilized to estimate individual treatment effects, their potential for evidence translation across patient populations has not been explored.[24–27] Conditional GANs (CGAN) enable the generation of synthetic datasets that allow for the conditioning of covariates with a second population distribution.[28,29] We hypothesize that applying this model to an RCT with the conditioning of a second population will estimate the treatment effects of the original RCT in the new patient population.

We present RCT-Twin-GAN, a generative framework that combines clinical knowledge and the statistically informed architecture to create a digital twin of an RCT conditioned on the characteristics of another patient population to assess for generalizability of treatment effect. To demonstrate the ability of the digital twin to replicate treatment effects in the conditioning or target population, we first compared two RCTs, SPRINT and ACCORD, with similar interventions but disparate treatment effects on cardiovascular outcomes. We created a digital twin of each of the 2 RCTs conditioned on covariate distributions of the other and evaluated whether the RCT-Twins reproduced the treatment effect of the conditioning or target cohort. Finally, we describe the cardiovascular outcomes of SPRINT and ACCORD digital twins conditioned on characteristics of patients in the electronic health record (EHR), introducing the role of RCT-Twins in estimating RCT treatment effects in real-world populations.

RESULTS

Study Populations

The study developed digital twins of two RCTs. The first, SPRINT, was a treatment intervention study to test whether intensive blood pressure control (goal systolic blood pressure less than 120 mmHg) versus standard care (goal systolic blood pressure less than 140 mmHg) reduced major cardiovascular events. The trial consisted of 9361 participants (median age 67 (61 to 76 (25-75% IQR, and 3332 (36%) women). Patients with prior stroke, diabetes mellitus, and a recent heart failure exacerbation had been excluded from the study. The patients in SPRINT were followed for a median of 3.26 years for the first occurrence of any of the primary composite outcome components of myocardial infarction, acute coronary syndrome, stroke, heart failure, or death from cardiovascular causes.

Our study built a SPRINT digital twin with a population representation of another RCT with the same intervention, the ACCORD trial, a double factorial RCT of participants with type 2 diabetes mellitus and cardiovascular disease. We specifically leveraged the blood pressure management component of the ACCORD trial, wherein half of the participants were randomized to intensive versus standard care blood pressure control, with the same treatment goals as those in the SPRINT trial. ACCORD consisted of 4733 participants (median age 62, IQR, 58-67, and 2258 [48%] women). ACCORD median follow-up time was 4.7 years for the primary composite outcome of myocardial infarction, stroke, or death for cardiovascular cause.

We also incorporated two cohorts from the Yale New Haven Hospital Health System Electronic Health Record (EHR), a large healthcare system including several hospitals with diverse racial and socioeconomic demographics across Connecticut and Rhode Island. Two sets of patients with hypertension, one without and the other with diabetes were identified to broadly represent populations included in SPRINT and ACCORD, respectively, to estimate the treatment effects found in the two RCTs on corresponding real-world patient populations. We included

4,000 randomly selected patients from each cohort for the final SPRINT EHR and ACCORD EHR cohorts. The SPRINT EHR cohort had a median age of 73 years (IQR, 61 to 84) and 2069 (52%) women), while the ACCORD EHR cohort had a median age of 71 (61 to 80 IQR) and 2032 (51%) women).

Development of RCT Digital Twins Conditioned on a Second Patient Population

We developed CGAN models to create digital twin datasets of an RCT conditioned on covariate distributions from a second patient population. We first built a SPRINT digital twin (SPRINT-twin) trained on the SPRINT cohort without a second conditioning cohort. We then built a SPRINT digital twin conditioned on the ACCORD participant population ($\text{SPRINT}_{\text{ACCORD}}$-Twin) with the intention of reproducing the ACCORD primary outcome in a SPRINT digital twin (Figure 1). To implement this, we applied the Conditional inputs for Direct Acyclic Tabular Generative Adversarial Networks (CiDATGANs), a conditional tabular GAN that uses a directed acyclic graph to assign relationships between pre-randomized covariates.[29,30] The directed acyclic graph ensures clinically relevant connections are introduced between covariates and prevents the weighting of spurious correlations between covariates. To condition the digital twins on the other RCT population, we mapped 33 equivalent covariates between SPRINT and ACCORD.

Our directed acyclic graph contained 16 representative covariates, including age at randomization, body mass index (BMI), current smoker, family history of cardiovascular disease, glomerular filtration rate (GFR), low-density lipoprotein (LDL) cholesterol level, left ventricular hypertrophy (LVH), previous myocardial infarction (MI), race, systolic blood pressure and heart rate, sex, statin use, and symptoms of angina, to minimize the number of variables needed to build a digital twin of the RCT cohort (Figure 2). In addition to the covariates, outcome, time to

outcome, and treatment arm were included in the graph. Two expert clinicians identified directed clinical relationships between the covariates and outcomes. Overall, 71 connections within the DAG were identified.

After training CiDATGAN with the DAG and SPRINT cohort data, the created SPRINT-Twin (the non-conditioned SPRINT twin) reproduced the distributions of the original covariates in SPRINT (Supplementary Table 6,7). In addition, all covariates were balanced between the intervention and standard care groups in SPRINT-Twin, as evidenced by absolute standardized mean differences of less than 0.1 and a mean absolute standardized difference (MASMD) of 0.019 (SD 0.018) between treatment groups across all covariates (Figure 3a). This was similar to the MASMD between treatment arms of SPRINT, 0.013 (SD 0.013).

After developing the CiDATGAN with the DAG informed by the SPRINT cohort, we conditioned the generator with covariate distributions from the ACCORD, choosing 10 covariates most dissimilar between the two population distributions. This was done to enable the ACCORD conditioning to specifically evaluate the hypothesis that the differences in findings between the two trials were at least partly mediated by the differences in population characteristics. The included binary and continuous covariates, in the order of increasing dissimilarity between cohorts, were black race, history of previous MI, female sex, statin use, LVH, BMI, seated heart rate, age at randomization, family history of CVD, and GFR. The standardized mean differences between $\text{SPRINT}_{\text{ACCORD}}$-Twin and ACCORD for these covariates all had standardized mean differences of less than 0.1, with a MASMD of 0.0082 (SD 0.016), while the MASMD between SPRINT and ACCORD of the same covariates was 0.46 (SD 0.20) (Figure 3b).

Estimating the Primary Cardiovascular Outcome in SPRINT-Twin

We confirmed the differences in reported outcomes in the SPRINT and ACCORD in the trial in our trial datasets, with a significant reduction in cardiovascular events in SPRINT's intervention arm compared with control (hazard ratio 0.75 [0.64-0.89 95% CI, p<0.001]), without a significant reduction in a similar primary outcome in ACCORD (hazard ratio 0.88 [0.73 to 1.06 95% CI, p=0.20]). In the SPRINT-Twin without conditioning, the median hazard ratio across 10 generated SPRINT-Twin datasets was 0.73 (CI 0.61-0.87), with the 10 replications performed to ensure the reproducibility of the findings. This was comparable to the HR of 0.75 in the SPRINT trial. Similarly, the ACCORD-Twin without conditioning replicated the primary results of the ACCORD trial, with a median HR of 0.89 (CI 0.79-1.0) comparable to the HR of 0.88 of the ACCORD trial.

Estimating the Primary Cardiovascular Outcome

We then demonstrated the ability of RCT-Twins to replicate the known treatment effects of a second population. The $SPRINT_{ACCORD}$-Twin – the SPRINT-Twin that was conditioned on ACCORD. We found the median hazard ratio of 10 $SPRINT_{ACCORD}$-Twin datasets was 0.87 (CI 0.68-1.13), this time comparable to the HR of 0.88 of the ACCORD trial (Figure 4a). In contrast, in 10 replicated digital twins of the ACCORD cohort conditioned on covariate distributions in SPRINT ($ACCORD_{SPRINT}$-Twin), reproduced the significant effect size seen in SPRINT (HR 0.75) with a median hazard ratio of 0.79 (CI 0.72-0.86) (Figure 4b).

Estimating the Treatment Effect of SPRINT and ACCORD in the EHR

In a descriptive substudy, we demonstrated the ability to estimate SPRINT and ACCORD treatment effect outcomes in patient populations reflecting a large US health system, YNHHS. The same 10 conditioning covariates used for conditioning SPRINT and ACCORD against each other, were computably extracted from the YNHHS EHR by clinician experts to define

covariates in the corresponding EHR cohorts. In the digital twin of SPRINT conditioned on the corresponding EHR cohort ($SPRINT_{EHR}$-Twin), we confirmed the replication of RCT features, including covariate balance across treatment and control arms (MASMD 0.03 (SD 0.03), Supplementary Figure 1. In this $SPRINT_{EHR}$-Twin the median cardiovascular outcome HR was 0.84 (95% CI, 0.64-1.09) across the 10 replications. Similarly, the $ACCORD_{EHR}$-Twin replicated both RCT features and EHR covariate distributions, with a median cardiovascular outcome HR of 0.94 (CI 0.8-1.1).

DISCUSSION

We present RCT-Twin-GAN, a deep generative model that utilizes clinical knowledge of covariate relationships to synthesize a digital twin of an RCT with selected covariate distributions from a second population distribution, which could be another RCT cohort to a general patient population reflected in an EHR. We found that RCT-Twin-GAN created digital twins that replicate the fundamental feature of RCTs, i.e., balanced covariates across treatment intervention arms, but with conditioning able to reflect the covariate distributions to mirror this second population's distribution. Moreover, in a positive control experiment in a 2-RCT system where treatment effects were known from well-conducted experiments but were discordant across the RCTs, the RCT-Twins conditioned on covariates from the opposing RCT replicated the results observed in other RCT, demonstrating the value of the approach in examining the effect of population characteristics on study outcomes. We also demonstrate that the approach is flexible to these characteristics drawn from any population, thereby enabling a quantitative evaluation of an RCT's potential treatment effects in populations that differed from those included in the trial.

Our work has built upon the established need to quantify generalizability of RCTs to new populations.[31] Prior methods, such as standardization of event rates, allow adjustment by single variables, which groups patients together by singular stratification.[32] Others have used distance metrics and decision tree machine learning techniques to represent the complex interplay of covariates and characterize the heterogeneity of treatment effect.[8–10,23,33–35] Our method complements these by building digital twins for each patient, drawing from the multiple covariate distribution and outcomes of each population to create complex subgroups within the conditioning population and allowing granularized treatment effect estimates. Statistical methods to assess heterogeneous treatment effects across populations have generally focused on equalizing baseline characteristics between populations using propensity score matching, but this scores one variable at a time, thereby ignoring multi-variable differences across patients, and does not consider effect modifiers.[36] We incorporate the distributions of multiple mutual pre-randomization covariates available across datasets to ensure representation across multivariate axes. In addition, we utilize clinician expertise to identify connections between covariates and build digital twins modeling the complex interplay of effect modifiers and outcomes.

Our application of digital twins introduces a novel approach for evidence translation across populations. Discordant randomized clinical trials can muddy the development of guidelines, but assessing population-level response could provide generalizable information needed to elucidate to which patient populations the guidelines apply. When a patient does not fit the population enrolled in the trial, assessment of the trial effect estimates with a general population, such as patients in the EHR or a registry, could similarly inform clinicians of possible differences in treatment effect from the original patient populations. Health systems could determine the likely treatment effect of an intervention in their patient population to better

contextualize their patient outcomes with the intervention by developing population-wide digital twins. This effort to use general real-world evidence to establish the efficacy of interventions has major regulatory support from agencies such as the US Food and Drug Administration.[37]

A unique feature of RCT-Twin-GAN is incorporating both rigorous statistical methods and clinical knowledge to build digital twins of RCTs with representative covariate balance and effect modifier information. The directed acyclic graph structure weights clinically relevant relationships between covariates and outcomes and removal of spurious correlations, which would otherwise be included in the GAN. Our ability to reproduce treatment effect estimates from the conditioning cohort by sampling its covariate distributions relies on the inference of important correlations between covariates during GAN training and digital twin generation. Although prior digital twin studies have focused on individual patient twin generation and supplementing RCTs with synthetic patients, our study builds upon these by estimating treatment effect across different patient populations. Measuring the hazard ratios of treatment effect outcomes as an evaluation metric provided valuable insights into the fidelity of the synthetic dataset in simulating clinical trial outcomes and treatment responses.

There are limitations to consider. First, RCT-Twin-GAN uses a select set of variables to build the digital twin. We chose a smaller set of covariates to maximize efficiency and showed that even with this small number of representative variables, we can build a digital twin that successfully replicates treatment effect estimates. Second, our model relies on outside input for identifying correlations between covariates, but we believe this can be considered a strength that clinical expertise can be imbued into the model to reduce the weight of spurious correlations inherent in data. Third, this is a post-hoc analysis of RCTs, but we show the ability of digital twins to mirror covariate characteristics and treatment effects found in SPRINT and ACCORD.

Fourth, we only applied RCT-Twin-GAN the SPRINT - ACCORD pair because it was the only paired trial testing the same intervention with different results available through a public domain, the National Heart, Lung, and Blood Institute Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). As the data are publicly available, further research can build upon this example, and we further anticipate applying our model to other examples. Fifth, modeling real-world patients in the EHR can be challenging since the data represents a snapshot of patients who seek care, but we choose patients from a diverse tertiary care system to maximize the breadth of the general population identified. In addition, the EHR covariates had to be operationally defined by experts to be analogous to the criteria used in RCT, but this is a descriptive study that shows different covariate distributions can be modeled. Finally, the true effect estimates in the EHR populations are unknown, and those estimated by RCT-Twin-GAN should not inform care but rather give an idea of discordance or concordance with the original RCT population.

We have introduced a new application of GANs to build synthetic cohorts by creating an RCT digital twin reflective of different patient populations, including similar RCTs and real-world patients found in the EHR. Our study demonstrates a way to evaluate the generalizability of an RCT to the general population by embedding covariate distributions that are more representative of real-world populations. This amplifies the effects for those more frequently seen in clinical practice. Overall, our model contributes significantly to the evidence supporting the development of an RCT digital twin that more authentically mirrors real-world populations, thereby enhancing inference for real-world patients.

METHODS

Data Source and Patient Populations

SPRINT and ACCORD Cohorts

The Systolic Blood Pressure Intervention Trial (SPRINT) was an open-label RCT testing difference in major cardiovascular outcomes between intensive blood pressure management of less than 120 mm Hg versus less than 140 mm Hg. From 2010-2013, at 102 clinical sites across the United States, 9361 participants were recruited who were at least 50 years old, had a systolic blood pressure between 130 and 180 mm Hg, and had increased cardiovascular event risk, including cardiovascular disease with the exception of stroke, chronic kidney disease, Framingham 10 year cardiovascular risk score of 15% or greater, and advanced age over 75. Patients with diabetes mellitus or stroke were excluded.

The Action to Control Cardiovascular Risk in Diabetes Blood Pressure (ACCORD BP) Trial was part of the ACCORD trial in which 4733 of the participants were randomly assigned to the same intensive vs standard blood pressure control as SPRINT in addition to intensive or standard glycemic control in a 2 by 2 factorial design. From 2001 to 2005, at 77 clinical sites across the United States and Canada, participants with type 2 diabetes mellitus, a glycated hemoglobin level of 7.5% or greater, and either age 40 or older with cardiovascular disease or age 55 or older with risk factors for cardiovascular disease and anatomical evidence of longstanding hypertension or diabetes such as albuminuria or left ventricular hypertrophy. Patients with a BMI over 45, a creatinine over 1.5 mg/dL, or serious illness were excluded.

EHR cohorts

The two EHR cohorts were extracted from patients within the Yale New Haven Health System (YNHHS), a tertiary healthcare system with hospitals and outpatient locations that serve diverse socioeconomic, geographic, and demographic populations across Connecticut and Rhode Island.

The study was reviewed by Yale Institutional Review Board and deemed exempt as it uses retrospective data. We sampled 100,000 adult patients to identify EHR cohorts similar to SPRINT and ACCORD. We first filtered the cohort to those with an ICD-10-CDM code for hypertension (Supplementary Table 2). Out of these patients, we filtered for patients with an ICD-10-CDM code for type 2 diabetes mellitus (Supplementary Table 2). Patients with both hypertension and type 2 diabetes mellitus were considered for the ACCORD EHR cohort. The remaining hypertension patients that did not have type 2 diabetes mellitus were considered for the SPRINT EHR cohort. Out of 100,710 sampled EHR patients, 30,972 had a hypertension diagnosis, and 8,840 of these patients also had type 2 diabetes mellitus. The participants with hypertension and type 2 diabetes mellitus were considered for the ACCORD EHR cohort, and the participants with hypertension but no type 2 diabetes mellitus were considered for the SPRINT EHR cohort. After excluding patients who did not have values for continuous covariates and patients above the age of 110, the SPRINT EHR cohort had 5,218 patients, while the ACCORD EHR cohort had 4,676 patients. We sampled 4000 patients each for the ACCORD EHR and SPRINT EHR cohorts with values for all conditioning covariates. We further excluded patients who had continuous values out of range of the training cohort of SPRINT or ACCORD (Supplementary Table 4), leading to a final 3,130 patients in the SPRINT EHR cohort and 2,731 patients in the ACCORD EHR cohort. Of note, the choice of the covariates was governed by primary analysis focused on shared covariates between SPRINT and ACCORD. In real-world translations, a different covariate set shared between a development and target population can be selected.

Covariate extraction for SPRINT, ACCORD, and the EHR

In order to condition the SPRINT digital twin (SPRINT-Twin) on equivalent ACCORD covariates ($SPRINT_{ACCORD}$-Twin), we mapped 33 equivalent covariates between the two cohorts, which included demographics such as age, gender, and race, conditions, and social history, such as smoking history, family history of cardiovascular disease, hyperlipidemia, and prior myocardial infarction, medications such as taking aspirin or statins, procedures such as coronary revascularization, and laboratory values and vital signs such as glucose, GFR, and seated systolic blood pressure (Supplementary Table 1). We also included outcome, time to outcome, and treatment arm assignment. We limited the maximum time to outcome to five years, censoring all subsequent outcomes.

In order to build the DAG, we identified 16 representative variables of those mapped between SPRINT and ACCORD including a family history of cardiovascular disease, race, symptoms of angina, seated systolic blood pressure and heart rate, LDL cholesterol, GFR, BMI, LVH, statin use, female sex, current smoker, previous myocardial infarction (MI), and age at randomization (Figure 2). Since BMI was considered a binary variable in ACCORD (above or below 32 kg/m^2), we used a similar definition in SPRINT. Variables related to exclusion criteria of at least one of the cohorts were not included in the conditioning of the model or constructing the DAG. These included glucose and type 2 diabetes mellitus. First, expert clinicians determined clinically relevant pairs between covariates and outcomes, and then a data-driven iterative process was conducted in which the network model recommended pairs based on correlations between the covariates in the data. The expert clinicians then determined whether to add them to the DAG. The arrows' direction pointed from the independent covariate to the dependent covariate. No arrow pointed to the treatment arm covariate, labeled "Group", since this assignment was independent of all covariates. All covariates and the "Group" pointed to the

“Outcome” and “Time to Outcome” covariates since all covariates and treatment arm assignment were thought to influence the outcome (Supplementary Table 3, Figure 2). We used the 10 covariates with the largest mean absolute standardized difference between SPRINT and ACCORD as the conditioning covariates in order to condition from the covariate distributions most representative of the second cohort population. These 10 covariates included black race, prior MI, female sex, statin use, LVH, BMI, heart rate, age, family history of CVD, and GFR.

Since we sought to condition on the EHR populations as well, we extracted the 10 conditioning covariates established in the prior analysis from the EHR as well. Only patients with all conditioning covariates were included. Specifically, these cohorts required a value for race, sex, glomerular filtration rate, heart rate, and BMI. The other binary covariates of previous MI, statin use, LVH, and family history of CVD were considered absent if not found in the patient’s EHR. Age was calculated on October 1, 2023 (EHR query date), unless they were deceased, where we used the death date to define their last known age.

Design of the RCT-Twin-GAN Model

RCT-Twin-GAN is a Generative Adversarial Network model, which is a deep learning model rooted in game theory that pits a generator, the neural network that creates synthetic data, against a discriminator, the neural network that determines whether the data it is trained on is synthetic or real. The minimization of the discrimination between real and synthetic data allows for the GAN to make realistic digital twins of the data on which it is trained.[38] The neural networks are comprised of Long Short Term Memory (LSTM) cells, which are structured to retain information from prior inputs in addition to the current variable input. Since the GAN was initially built to produce synthetic images, this has been adapted to accurately synthesize tabular data such as

EHR.[27,28,39] The incorporation of a conditioning parameter enables tabular digital twins to define covariate distributions sampled from a second cohort.[28,29,39]

RCT-Twin-GAN is based on the architecture of CiDATGAN, which is an extension of DATGAN with the additional feature of conditioning covariates with distributions from an alternate population.[29,30] The DATGAN model employs a unique feature allowing the user to feed to the generator causal relationships between covariates and outcomes of the original training cohort via a Directed Acyclic Graph (DAG). Continuous and categorical variables from the original dataset are encoded from tabular data at the discriminator input. During training, the generator utilizes Gaussian noise along with the DAG covariate relationships and the conditioned covariates of the original dataset to generate synthetic data. The discriminator is trained to differentiate between the generated data and the original cohort. The generator combines Gaussian noise and attention vectors of the LSTM cells to produce synthetic values and transforms conditional inputs using a dense layer. During the sampling phase, the generator receives Gaussian noise and inputs of the conditioned covariates from the second cohort. It generates the complementary set of variables based on these inputs and combines them with the conditional inputs to produce the final dataset.

Once the DAG was built, we encoded the continuous columns based on the min-max values of the covariate in the training dataset (Supplementary Table 4) winsorized to remove outlier values below the 2.5% and 97.5% percentiles. The CiDATGAN was then trained with the DAG and encoded dataset to generate the synthetic dataset. We performed a hyperparameter grid search with different batch sizes and epochs to find the best parameters to generate synthetic data with similar outcomes to the original dataset (Supplementary Table 5).

Application of RCT-Twin-GAN in SPRINT, ACCORD, and the EHR

We first used RCT-Twin-GAN to build a SPRINT-Twin, which created a DAG based on the SPRINT cohort, trained the DATGAN architecture on the SPRINT cohort, and sampled the synthetic data generation for 10 iterations. We then built a $SPRINT_{ACCORD}$-Twin, which again built a DAG from the SPRINT cohort and trained on the SPRINT cohort but was conditioned on the ACCORD cohort. This meant that the DAG was modified to remove connections going to the conditioning covariates, and in the sampling phase, Gaussian noise and the conditioned covariate distributions from the ACCORD cohort were inputs for the generator in order to create the final synthetic dataset. We sampled this process for 10 iterations to make 10 $SPRINT_{ACCORD}$-Twin datasets. We repeated this process to create $ACCORD_{SPRINT}$-Twin datasets by replacing ACCORD to be the training cohort and SPRINT to be the conditioning cohort.

We repeated this training and conditioning process using the EHR cohorts as well. Specifically, we trained RCT-Twin-GAN on the SPRINT cohort, conditioned on the SPRINT-EHR cohort, and sampled 10 times to create $SPRINT_{EHR}$-Twins. We similarly made $ACCORD_{EHR}$-Twins with ACCORD and the ACCORD-EHR cohort.

Analysis of Cohort Representation in Digital Twins

To determine whether RCT-Twin-GAN created digital twins that are balanced by treatment arm, we calculated the mean absolute standardized mean difference (MASMD) of all covariates of the digital twins stratified by treatment arm assignment. A value of less than 0.1 was considered adequate balance, consistent with convention when assessing the success of propensity score matching.[40] To assess the representation of the conditioning cohort in the synthetic digital twin, we also calculated the ASMD between SPRINT and ACCORD for each covariate, SPRINT-Twin and ACCORD, and $SPRINT_{ACCORD}$-Twin and ACCORD.

We also ensured all the digital twins were made up of only synthetic rows, meaning there was no duplication of original or conditional rows using a previously recognized digital twin evaluation by Synthetic Data Vault.[41]

Estimation of Treatment Effect on Cardiovascular outcomes in the digital twins

In order to assess the ability of RCT-Twin-GAN to estimate RCT treatment effect outcomes in populations other than the original RCT, we calculated the hazard ratio of cardiovascular outcomes stratified by treatment arms in each of the digital twin datasets using cox proportional hazard models. We utilized hazard ratios to evaluate the comparative risks of events over time between different treatment groups within the synthetic data. This analytical approach allowed us to gauge the effectiveness of the synthetic dataset in accurately representing the underlying dynamics of treatment effects and event occurrences observed in real-world scenarios.

We reported the median hazard ratio and 95% confidence intervals for the 10 SPRINT-Twin, $SPRINT_{ACCORD}$-Twin, and $ACCORD_{SPRINT}$-Twin digital twins. In order to demonstrate the ability to estimate treatment effect outcomes in a variety of cohorts, we calculated the hazard ratio and 95% confidence intervals of cardiovascular outcomes of the $SPRINT_{EHR}$-Twins and $ACCORD_{EHR}$-Twins as well.

Statistical Analysis

Categorical variables were summarized as numbers with percentages, and continuous variables were summarized as median with 25% and 75% interquartile ranges (IQR) or mean with standard deviation (SD). Covariate distributions were compared using mean absolute standardized mean difference and standard deviations and graphed as love plots comparing datasets. Data was winsorized at 2.5% and 97.5% percentiles to remove outliers. Survival analysis was conducted using unadjusted cox proportional hazard models with p values

calculated after 5 years and presented as Kaplan Meier survival curves. Hazard ratios across digital twins and SPRINT and ACCORD cohorts were presented as forest plots with 95% confidence interval error bars. Analyses were conducted using python 3.9, with packages specified in the supplement.

Funding:

The study is supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (R01HL167858). Dr. Thangaraj and Dr. Oikonomou are also supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health (5T32HL155000-03 and 1F32HL170592-01, respectively).

Disclosures:

The authors Dr. Thangaraj, Mr. Shankar, Dr. Oikonomou, and Dr. Khera are coinventors of a provisional patent related to the current work (63/606,203). Dr. Oikonomou is a co-inventor of the U.S. Patent Applications 63/508,315 63/177,117, a cofounder of Evidence2Health (with Dr. Khera), and has previously served as a consultant to Caristo Diagnostics Ltd (outside the present work). Dr. Nadkarni is a founder of Renalytix, Pensieve, and Verici and provides consultancy services to AstraZeneca, Reata, Renalytix, Siemens Healthineer, and Variant Bio, and serves a scientific advisory board member for Renalytix and Pensieve. He also has equity in Renalytix, Pensieve, and Verici. Dr. Mortazavi reported receiving grants from the National Institute of Biomedical Imaging and Bioengineering, National Heart, Lung, and Blood Institute, US Food and Drug Administration, and the US Department of Defense Advanced Research Projects Agency outside the submitted work. In addition, B.J.M. has a pending patent on predictive

models using electronic health records (US20180315507A1). Dr. Khera is an Associate Editor of JAMA. He receives support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under awards R01HL167858 and K23HL153775) and the Doris Duke Charitable Foundation (under award 2022060). He also receives research support, through Yale, from Bristol-Myers Squibb, Novo Nordisk, and BridgeBio. He is a coinventor of U.S. Pending Patent Applications 63/562,335, 63/177,117, 63/428,569, 63/346,610, 63/484,426, 63/508,315, and 63/606,203. He is a co-founder of Ensight-AI, Inc. and Evidence2Health, health platforms to improve cardiovascular diagnosis and evidence-based cardiovascular care.

References

1. Averitt AJ, Ryan PB, Weng C, Perotte A. A conceptual framework for external validity. *J Biomed Inform.* 2021;121:103870.

2. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet*. 2005;365:82–93.

3. Filbey L, Zhu JW, D'Angelo F, et al. Improving representativeness in trials: a call to action from the Global Cardiovascular Clinical Trialists Forum. *Eur Heart J*. 2023;44:921–930.

4. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials*. 2015;16:495.

5. Ranganathan M, Bhopal R. Exclusion and inclusion of nonwhite ethnic minority groups in 72 North American and European cardiovascular cohort studies. *PLoS Med*. 2006;3:e44.

6. Sardar MR, Badri M, Prince CT, Seltzer J, Kowey PR. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA Intern Med*. 2014;174:1868–1870.

7. DeFilippis EM, Echols M, Adamson PB, et al. Improving Enrollment of Underrepresented Racial and Ethnic Populations in Heart Failure Trials: A Call to Action From the Heart Failure Collaboratory. *JAMA Cardiol*. 2022;7:540–548.

8. Oikonomou EK, Spatz ES, Suchard MA, Khera R. Individualising intensive systolic blood pressure reduction in hypertension using computational trial phenomaps and machine learning: a post-hoc analysis of randomised clinical trials. *Lancet Digit Health*. 2022;4:e796–e805.

9. Oikonomou EK, Suchard MA, McGuire DK, Khera R. Phenomapping-Derived Tool to Individualize the Effect of Canagliflozin on Cardiovascular Risk in Type 2 Diabetes. *Diabetes Care*. 2022;45:965–974.

10. Oikonomou EK, Van Dijk D, Parise H, et al. A phenomapping-derived tool to personalize the selection of anatomical vs. functional testing in evaluating chest pain (ASSIST). *Eur Heart J*. 2021;42:2536–2548.

11. Patel HC, Hayward C, Dungu JN, et al. Assessing the Eligibility Criteria in Phase III Randomized Controlled Trials of Drug Therapy in Heart Failure With Preserved Ejection Fraction: The Critical Play-Off Between a “Pure” Patient Phenotype and the Generalizability of Trial Findings. *J Card Fail*. 2017;23:517–524.

12. Lim YMF, Molnar M, Vaartjes I, et al. Generalizability of randomized controlled trials in heart failure with reduced ejection fraction. *Eur Heart J Qual Care Clin Outcomes*. 2022;8:761–769.

13. SPRINT Research Group, Wright JT Jr, Williamson JD, et al. A Randomized Trial of Intensive versus Standard Blood-Pressure Control. *N Engl J Med*. 2015;373:2103–2116.

14. ACCORD Study Group, Cushman WC, Evans GW, et al. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *N Engl J Med*. 2010;362:1575–1585.

15. Carson JL, Brooks MM, Hébert PC, et al. Restrictive or Liberal Transfusion Strategy in Myocardial Infarction and Anemia. *N Engl J Med*. 2023;389:2446–2456.

16. Ducrocq G, Gonzalez-Juanatey JR, Puymirat E, et al. Effect of a Restrictive vs Liberal Blood Transfusion Strategy on Major Cardiovascular Events Among Patients With Acute Myocardial Infarction and Anemia: The REALITY Randomized Clinical Trial. *JAMA*. 2021;325:552–560.

17. Joosten LPT, van Doorn S, van de Ven PM, et al. Safety of Switching from a Vitamin K Antagonist to a Non-Vitamin K Antagonist Oral Anticoagulant in Frail Older Patients with Atrial Fibrillation: Results of the FRAIL-AF Randomized Controlled Trial. *Circulation*. 2023. Published onlineAugust 27, 2023. https://doi.org/10.1161/CIRCULATIONAHA.123.066485.

18. Granger CB, Alexander JH, McMurray JJV, et al. Apixaban versus Warfarin in Patients with Atrial Fibrillation. *N Engl J Med*. 2011;365:981–992.

19. Jane-wit D, Horwitz RI, Concato J. Variation in results from randomized, controlled trials: stochastic or systematic? *J Clin Epidemiol*. 2010;63:56–63.

20. Krakoff LR. A tale of 3 trials: ACCORD, SPRINT, and SPS3. What happened? *Am J Hypertens*. 2016;29:1020–1023.

21. Chobanian AV. Hypertension in 2017-what is the right target? *JAMA*. 2017;317:579–580.

22. Huang C, Dhruva SS, Coppi AC, et al. Systolic blood pressure response in SPRINT (Systolic Blood Pressure Intervention Trial) and ACCORD (Action to Control Cardiovascular Risk in Diabetes): A possible explanation for discordant trial results. *J Am Heart Assoc*. 2017;6.

23. Laffin LJ, Besser SA, Alenghat FJ. A data-zone scoring system to assess the generalizability of clinical trial results to individual patients. *Eur J Prev Cardiol*. 2019;26:569–575.

24. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature*. 2021;592:629–633.

25. Ge Q, Huang X, Fang S, et al. Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection. *Front Genet*. 2020;11:585804.

26. Yoon J, Jordon J, Van Der Schaar M. Ganite: Estimation of individualized treat- ment effects using generative adversarial nets. 2018. Accessed November 9, 2023. https://openreview.net/pdf?id=ByKWUeWA-.

27. Li J, Cairns BJ, Li J, Zhu T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digit Med*. 2023;6:98.

28. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. *arXiv [csLG]*. 2019.

29. Lederrey G, Hillel T, Bierlaire M. ciDATGAN: Conditional Inputs for Tabular GANs. *arXiv [csLG]*. 2022.

30. Lederrey G, Hillel T, Bierlaire M. DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data. *arXiv [csLG]*. 2022.

31. He Z, Tang X, Yang X, et al. Clinical Trial Generalizability Assessment in the Big Data Era: A Review. *Clin Transl Sci*. 2020;13:675–684.

32. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Stratification for confounding--part 2: direct and indirect standardization. *Nephron Clin Pract*. 2010;116:c322-5.

33. Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical Value of Predicting Individual Treatment Effects for Intensive Blood Pressure Therapy. *Circ Cardiovasc Qual Outcomes*. 2019;12:e005010.

34. Brantner CL, Nguyen TQ, Tang T, Zhao C, Hong H, Stuart EA. Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects. *Stat Med*. 2024. Published onlineJanuary 25, 2024. https://doi.org/10.1002/sim.9955.

35. Raghavan S, Josey K, Bahn G, et al. Generalizability of heterogeneous treatment effects based on causal forests applied to two randomized clinical trials of intensive glycemic control. *Ann Epidemiol*. 2022;65:101–108.

36. Degtiar I, Rose S. A Review of Generalizability and Transportability. *Annual Review of Statistics and Its Application*. 2023;10:501–524.

37. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER) U.S. Food and Drug Administration. Framework for FDA's Real World Evidence Program. *US Food & Drug Administration*. 2018. Accessed March 6, 2024. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

38. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks. *arXiv [statML]*. 2014.

39. Zhao Z, Kunar A, Van der Scheer H, Birke R, Chen LY. CTAB-GAN: Effective Table Data Synthesizing. *arXiv [csLG]*. 2021.

40. Normand ST, Landrum MB, Guadagnoli E, et al. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *J Clin Epidemiol*. 2001;54:387–398.

41. Patki N, Wedge R, Veeramachaneni K. The synthetic data vault. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2016.

Figures

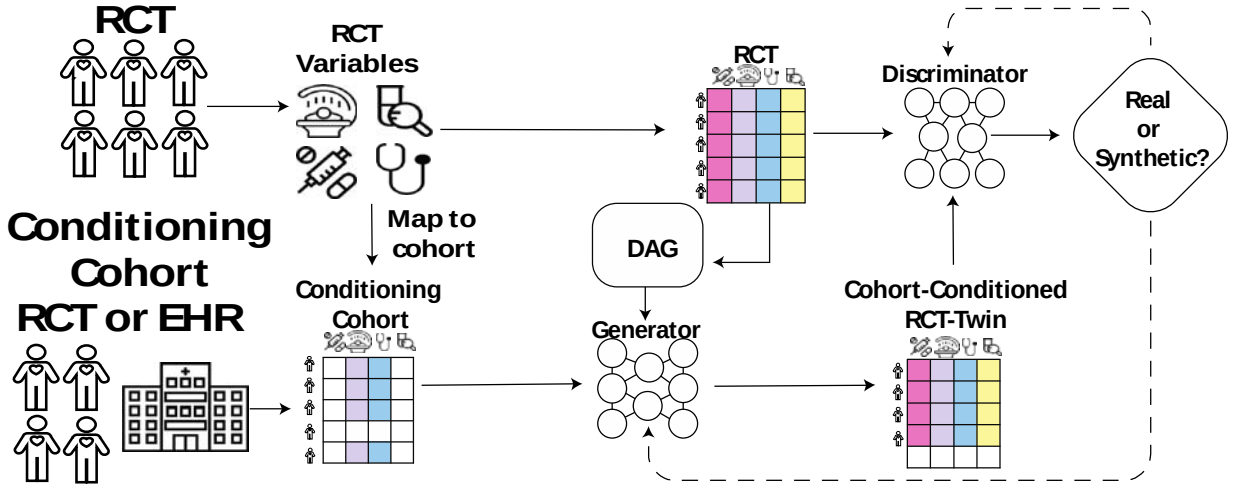

Figure 1: Graphical Abstract of RCT-Twin-GAN Model. The original cohort is a randomized control trial (RCT) and variables from this cohort are mapped to a conditioning cohort, examples of which are another RCT or a patient cohort in the electronic health record (EHR). Variables from the original RCT cohort are mapped to the conditioning cohort. The original RCT variables are input to the discriminator neural network. The directed acyclic graph (DAG) is made up of the original cohort covariates and inputed to the generator, along with select covariates from the conditioning cohort and gaussian noise. The cohort-conditioned RCT twin is generated by the generator and serves as the second input to the discriminator, which has to discriminate between the original RCT data and the cohort-conditioned RCT-Twin. Abbreviations: RCT: Randomized Control Trial, EHR: electronic health record, and DAG: Directed Acyclic Graph.

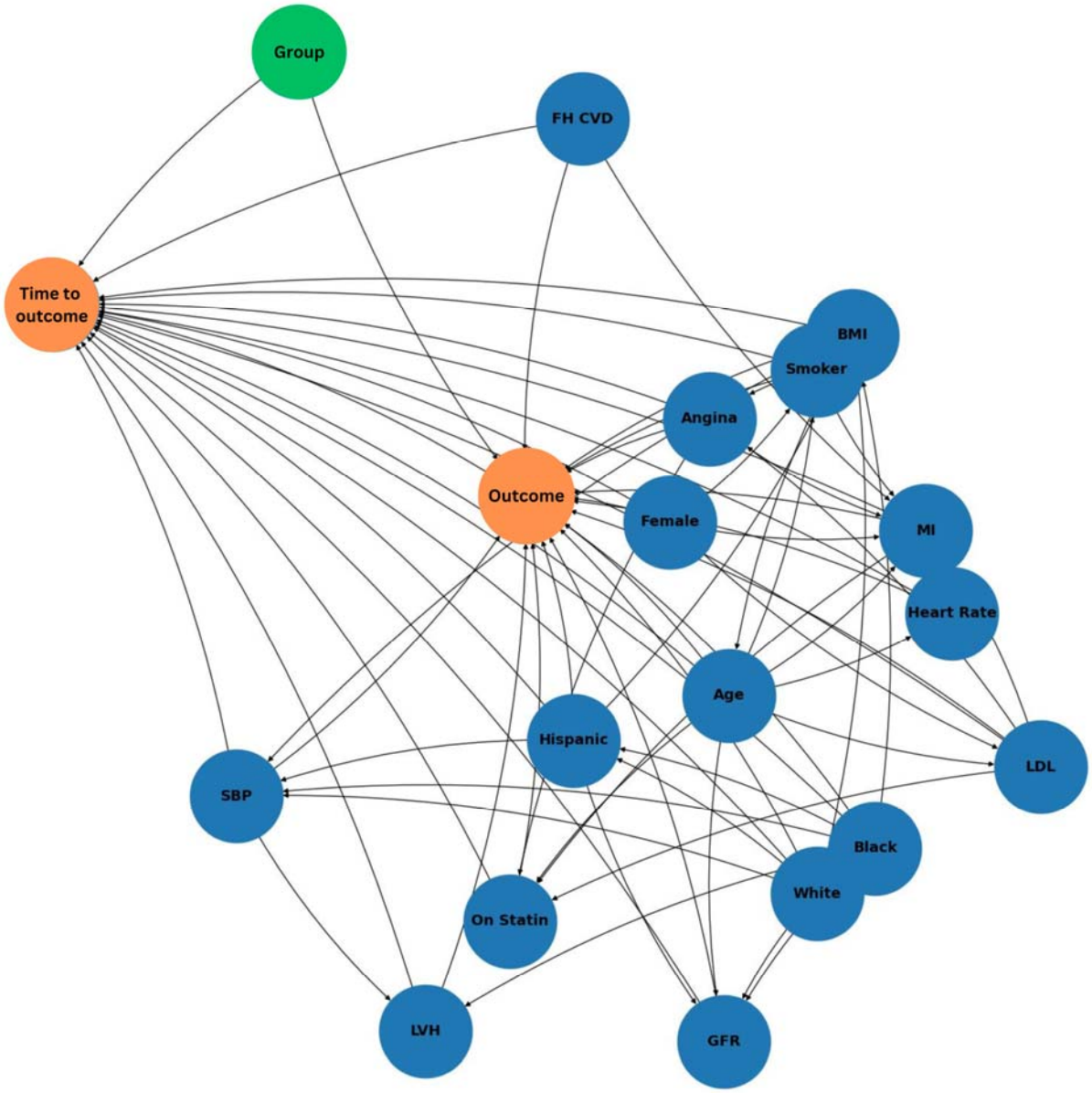

Figure 2: Directed Acyclic Graph of RCT-Twin-GAN**.** Directed relationships between covariates (in blue), time to outcome and outcome (in orange), and treatment arm designation ("Group") in green. Abbreviations: FH: Family History, CVD: Cardiovascular disease, BMI: Body Mass Index, GFR: Glomerular Filtration Rate, LVH: Left ventricular hypertrophy, LDL: low-density lipoprotein, MI: Myocardial infarction, SBP: Systolic Blood Pressure.

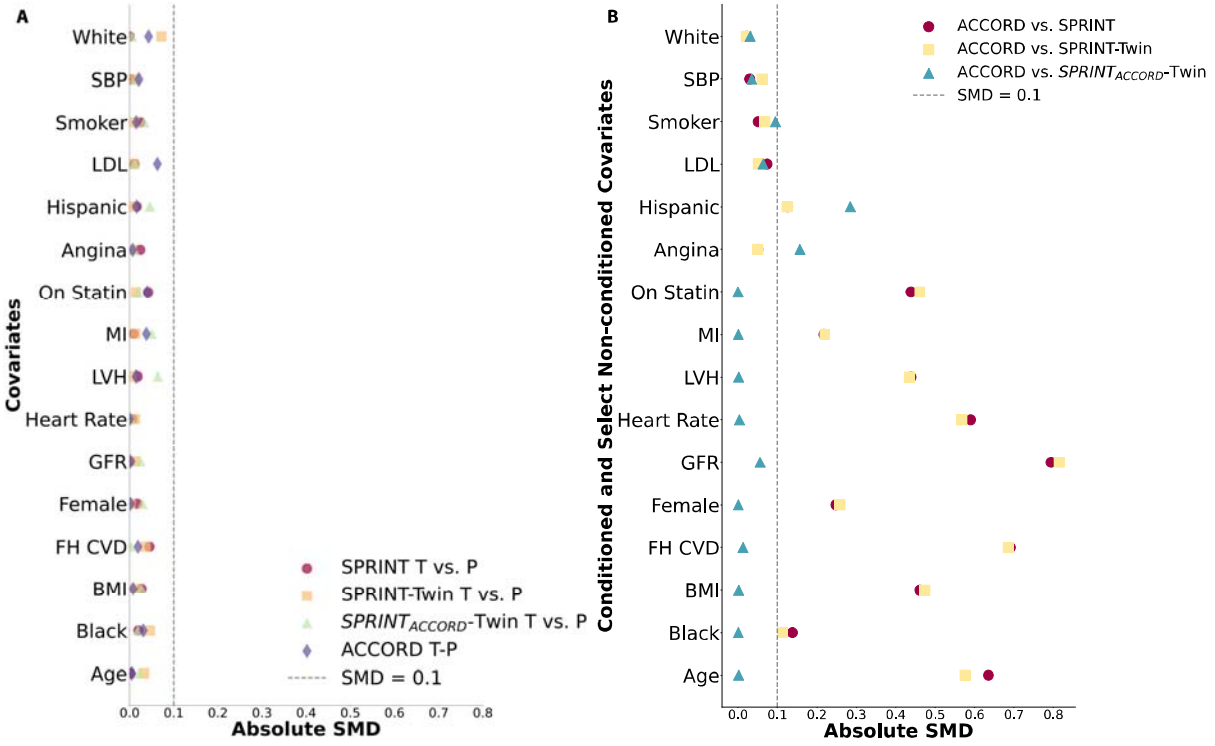

Figure 3: (A) Absolute standardized mean difference (ASMD) between treatment and placebo arms of RCTs and digital twins and (B) Absolute standardized mean difference (ASMD) between ACCORD and other datasets. (A) Markers include SPRINT (red circle), SPRINT-Twin (orange square), $SPRINT_{ACCORD}$-Twin (green triangle), and ACCORD (purple diamond). (B) Dark red circle represents ASMD between SPRINT and ACCORD, yellow square represents ASMD between SPRINT-Twin and ACCORD, and blue triangle represents ASMD between $SPRINT_{ACCORD}$-Twin and ACCORD. The dotted line represents an ASMD of 0.1. The conditioning covariates included Age, Black, BMI, FH CVD, Female, GFR, Heart Rate, LVH, MI, and On Statin. Abbreviations: T: Treatment Arm, P: Placebo Arm, FH: Family History, CVD: Cardiovascular disease, BMI: Body Mass Index, GFR: Glomerular Filtration Rate, LVH: Left ventricular hypertrophy, LDL: low-density lipoprotein, MI: Myocardial infarction, SBP: Systolic Blood Pressure, SMD: Standardized Mean Difference.

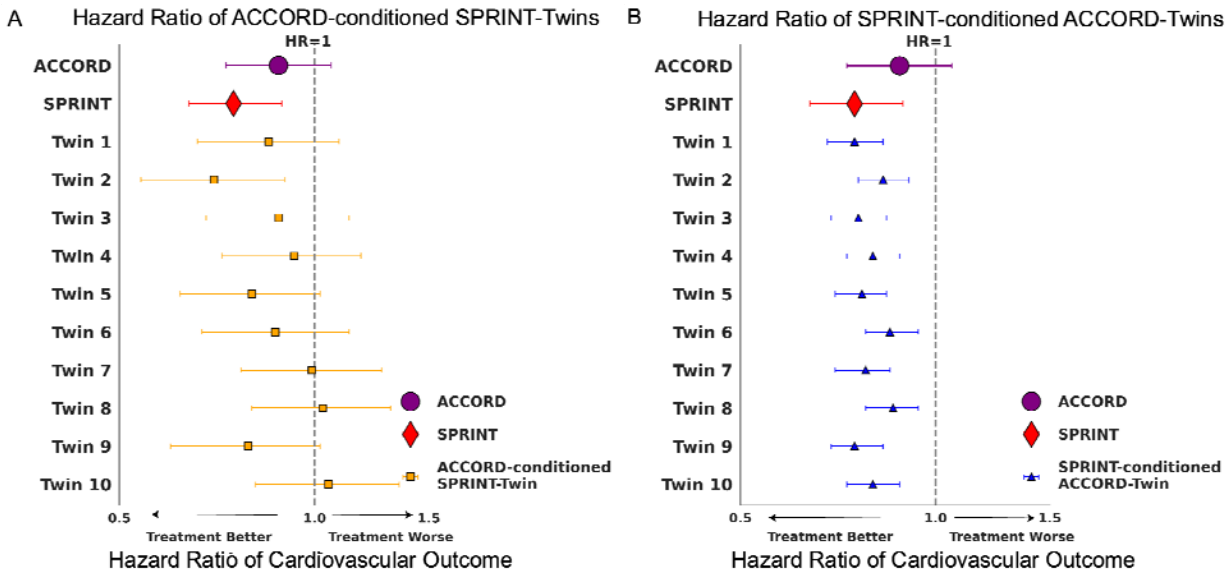

Figure 4: Hazard ratio of intensive blood pressure lowering on major cardiovascular outcomes in (A) the $SPRINT_{ACCORD}$-Twin datasets and (B) the $ACCORD_{SPRINT}$-Twin datasets. In both graphs, the purple circle is the ACCORD hazard ratio and 95% confidence interval, the red diamond is the SPRINT hazard ratio and 95% confidence interval, and the grey dotted line represents a hazard ratio of 1. In (A), orange squares are the hazard ratio of major cardiovascular outcome predicted for each twin run of ACCORD-conditioned SPRINT twins with 95% confidence intervals and in (b) the blue triangles are SPRINT-conditioned ACCORD twins. Abbreviations: MACE: Major Cardiovascular Outcomes, $SPRINT_{ACCORD}$-Twin: ACCORD-conditioned SPRINT Twin, $ACCORD_{SPRINT}$-Twin: SPRINT-conditioned ACCORD twin.

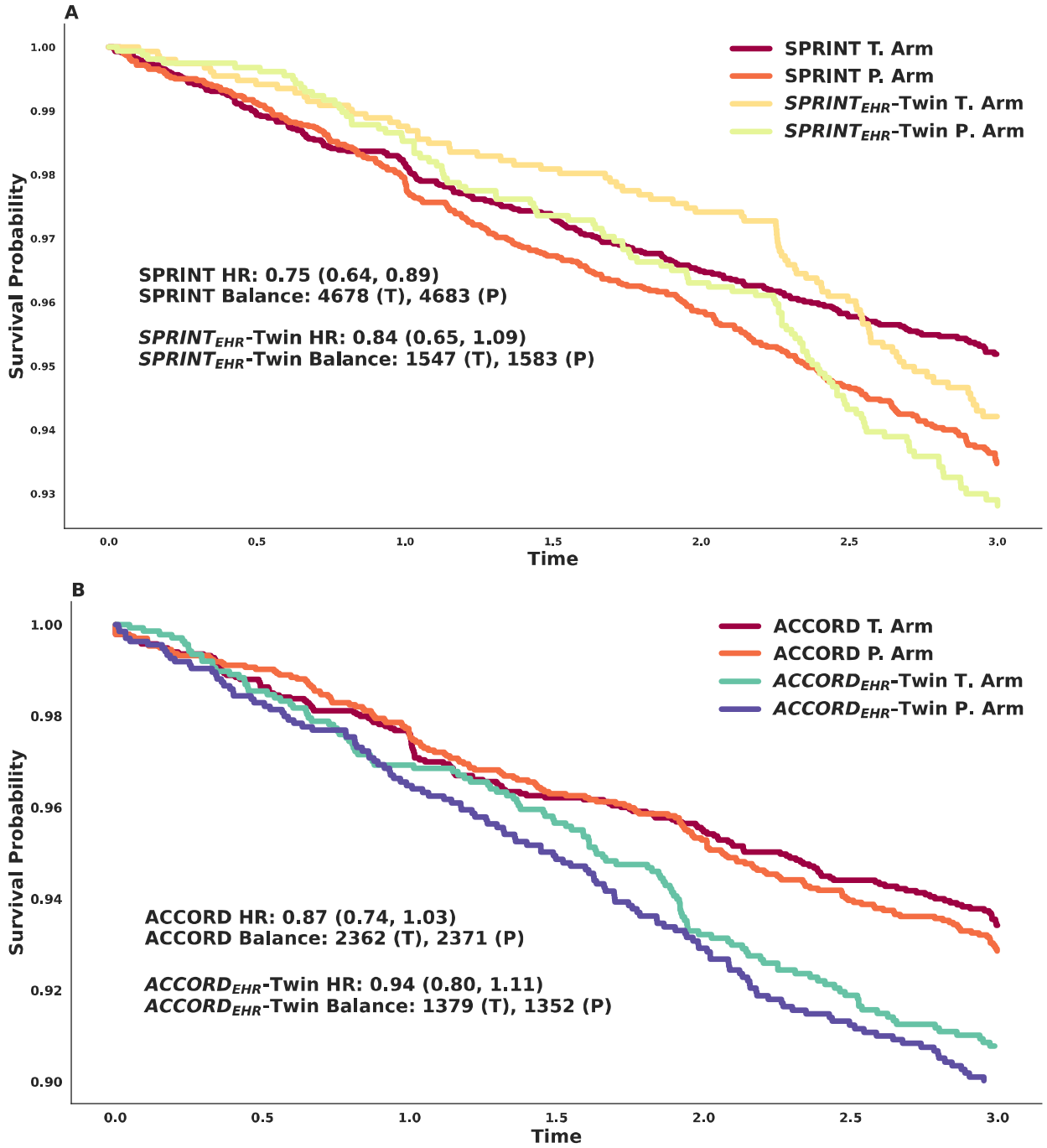

Figure 5: Representative Kaplan-Meier curves of digital twins conditioned on EHR data. (A) Kaplan-Meier curves of SPRINT treatment and placebo arms along with EHR-conditioned SPRINT treatment and placebo arms, (B) Kaplan Meier curves of ACCORD treatment and placebo arms along with EHR-conditioned ACCORD treatment and placebo arms. Hazard ratios (HR) and 95% confidence intervals displayed along with the treatment and placebo balance of the original cohorts and digital twins. Abbreviations: T: Treatment Arm, P. Placebo arm, EHR: Electronic Health Record, SPRINT$_{EHR}$-Twin: SPRINT conditioned on EHR digital twin, ACCORD$_{EHR}$Twin: ACCORD conditioned on EHR digital twin.