Computational Phenomapping of Randomized Clinical Trials to Enable Assessment of their Real-world Representativeness and Personalized Inference

Phyllis M. Thangaraj MD PhD,[1] Evangelos K. Oikonomou MD DPhil,[1] Lovedeep S. Dhingra MBBS,[1] Arya Aminorroaya MD MPH,[1] Rahul Jayaram BS,[1] Marc A. Suchard MD PhD,[2,3] Rohan Khera MD MS[1,4,5,6]

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

[2]Department of Biostatistics, Fielding School of Public Health, University of California, 650 Charles E. Young Drive S, Los Angeles, CA 90095, USA.

[3]Departments of Computational Medicine and Human Genetics, David Geffen School of Medicine at UCLA, University of California, 695 Charles E. Young Drive S, Los Angeles, CA 90095, USA.

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

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

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

Manuscript Type: Original Research
Words:
Abstract: 375
Text: 3393

***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

KEY POINTS

Question: How can we examine the multi-dimensional generalizability of randomized clinical trials (RCT) to real-world patient populations?

Findings: We demonstrate a novel phenotypic distance metric comparing an RCT to real-world populations in a large multicenter RCT of heart failure patients and the corresponding patients in multisite electronic health records (EHRs). Across 63 pre-randomization characteristics, pairwise assessments of members of the RCT and EHR cohorts were more discordant from each other than between members of the EHR cohort (median standardized mean difference 0.200 [0.037-0.410] vs 0.062 [0.010-0.130]), with a majority (55%) of RCT participants closer to each other than any individual EHR patient. The approach also enabled the quantification of expected real world outcomes based on effects observed in the RCT.

Meaning: A multidimensional phenotypic distance metric quantifies the generalizability of RCTs to a given population while also offering an avenue to examine expected real-world patient outcomes based on treatment effects observed in the RCT.

ABSTRACT

Importance: Randomized clinical trials (RCTs) are the standard for defining an evidence-based approach to managing disease, but their generalizability to real-world patients remains challenging to quantify.

Objective: To develop a multidimensional patient variable mapping algorithm to quantify the similarity and representation of electronic health record (EHR) patients corresponding to an RCT and estimate the putative treatment effects in real-world settings based on individual treatment effects observed in an RCT.

Design: A retrospective analysis of the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist Trial (TOPCAT; 2006-2012) and a multi-hospital patient cohort from the electronic health record (EHR) in the Yale New Haven Hospital System (YNHHS; 2015-2023).

Setting A multicenter international RCT (TOPCAT) and multi-hospital patient cohort (YNHHS).

Participants: All TOPCAT participants and patients with heart failure with preserved ejection fraction (HFpEF) and ≥1 hospitalization within YNHHS.

Exposures: 63 pre-randomization characteristics measured across the TOPCAT and YNNHS cohorts.

Main Outcomes and Measures: Real-world generalizability of the RCT TOPCAT using a multidimensional phenotypic distance metric between TOPCAT and YNHHS cohorts. Estimation of the individualized treatment effect of spironolactone use on all-cause mortality within the YNHHS cohort based on phenotypic distance from the TOPCAT cohort.

Results: There were 3,445 patients in TOPCAT and 11,712 HFpEF patients across five hospital sites. Across the 63 TOPCAT variables mapped by clinicians to the EHR, there were larger differences between TOPCAT and each of the 5 EHR sites (median SMD 0.200, IQR 0.037-0.410) than between the 5 EHR sites (median SMD 0.062, IQR 0.010-0.130). The synthesis of these differences across covariates using our multidimensional similarity score also suggested substantial phenotypic dissimilarity between the TOPCAT and EHR cohorts. By phenotypic distance, a majority (55%) of TOPCAT participants were closer to each other than any individual EHR patient. Using a TOPCAT-derived model of individualized treatment benefit from spironolactone, those predicted to derive benefit and receiving spironolactone in the EHR cohorts had substantially better outcomes compared with predicted benefit and not receiving the medication (HR 0.74, 95% CI 0.62-0.89).

Conclusions and Relevance: We propose a novel approach to evaluating the real-world representativeness of RCT participants against corresponding patients in the EHR across the full multidimensional spectrum of the represented phenotypes. This enables the evaluation of the implications of RCTs for real-world patients.

Words: 375

INTRODUCTION

Randomized clinical trials (RCTs) are the standard for defining optimal care practices, but quantifying their generalizability to real-world patients remains challenging.[1–4] Underrepresentation and under-enrollment of key patient demographic and clinical subpopulations contribute to this gap, leading to decreased external validity of RCT treatment effect outcomes in these populations. [5–12] The generalizability of RCTs across real-world populations relies on their external validity.[3] Prior studies assessing the external validity of RCTs, however, have been unable to capture the complete profile of patients, relying instead on comparing populations across a few covariates or one covariate at a time.[12–15] For example, hypothetically, if an RCT had an equal gender distribution, but all men had renal disease, and all women had diabetes, a real-world cohort with similar gender composition but with renal disease and diabetes present split equally between genders would be indistinguishable on univariate comparisons of gender, diabetes, or renal dysfunction. This example, while an extreme case, does not even account for the complex relationships across all covariates. Therefore, a multi-dimensional phenotypic representation of cohorts is needed to adequately evaluate representativeness between RCT cohorts and real-world populations.

To address this, we leveraged participant-level data from a large, phase 3 RCT of heart failure with preserved ejection fraction (HFpEF).[16,17] The inherently heterogeneous patient profiles of HFpEF provide an ideal use case for multi-dimensional phenotypic representation, which we define as *phenomapping*.[17–19] In the RCT, the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist Trial (TOPCAT), spironolactone did not significantly lower the risk of major adverse cardiovascular events.[20] Subsequent analyses of

TOPCAT have shown heterogeneous treatment effects across participants, requiring an evaluation of the extent to which the RCT cohort is representative of real-world patients.[20–22]

Identifying the complex phenotypic profile of patients with HFpEF in a real-world cohort can quantify the generalizability of TOPCAT across patient populations. We leverage the population of HFpEF patients captured in the electronic health record (EHR) at 5 hospital sites in a large, geographically dispersed health system to demonstrate a strategy to define the representativeness of RCT for both patient characteristics and anticipated real-world treatment effects.

METHODS

Study Populations

The first study population, the TOPCAT trial, was a multi-center international RCT that enrolled patients between 2006 and 2012 and evaluated the effect of spironolactone compared with placebo on the incidence of the combined cardiovascular outcome of death from cardiovascular cause, myocardial infarction, stroke, aborted cardiac arrest, and hospitalization for decompensated heart failure among patients with HFpEF. Details of TOPCAT (ClinicalTrials.gov identifier: NCT00094302) have been previously published.[20] The study enrolled 3445 individuals ≥ 50 years from North America, South America, Georgia, and Russia with left ventricular ejection fraction (LVEF) ≥ 45%, one sign and one symptom of heart failure, and at least one hospitalization for heart failure in the preceding 12 months. Alternatively, those without a hospitalization but with an elevated B-type natriuretic peptide were also included.

The second data source was the Yale New Haven Health System, a large health system that includes several hospitals and associated primary care locations with diverse racial and

socioeconomic demographics across Connecticut and Rhode Island. We focused on patients admitted with heart failure to one of the Connecticut sites between January 2015 through April 2023. The study included patients across 5 sites within 4 geographically distinct hospitals: Yale New Haven Hospital York Street Campus (YNHH YSC) and Yale New Haven Hospital St. Raphael's Campus (YNHH SRC), Greenwich Hospital (GH), Bridgeport Hospital (BH), and Lawrence + Memorial Hospital (LMH). Of note, YNHH YSC and YNHH SRC are located in New Haven, while the other sites are located in other cities/towns in Connecticut. The health system uses an Epic EHR system with data organized in Epic Clarity®, a SQL database management system. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines were followed.

EHR Heart Failure Cohort Derivation

Mapping an RCT to an EHR cohort study requires the identification of a cohort that best emulates the eligibility criteria of the study. We extracted patients with at least one hospital admission with a heart failure International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) code, representing codes with the root I50. We curated patient encounters with an echocardiogram within six months of the hospitalization and an LVEF of 45% or above and without any prior echocardiogram with an LVEF below 45%. Hospital sites included YNHH YSC, YNHH SRC, BH, GH, and LMH, referred to collectively as the EHR cohorts.

Curation and Mapping of RCT variables to EHR cohorts

We extracted baseline demographics, conditions, procedures, vital signs, medications, laboratory values, and echocardiogram variables from participant-level data for TOPCAT. We selected structured TOPCAT variables with <50% missingness, therefore representing conditions that

were captured in a majority of participants. We excluded covariates that provided redundant information or did not have a corresponding definition in the EHR and variables that were related to the trial logistics or its timeline (eTable 1). In all, 63 covariates, 1 treatment arm indicator, and 1 outcome of time to composite cardiovascular mortality were included (Figures 1a and 1b, Table 1, eTable 2).

Two clinicians (PT and EKO) collaboratively defined the computational phenotype of each of the 65 TOPCAT variables and outcomes to be deployed in the EPIC Clarity® extracts to map the TOPCAT variables. These included tables summarizing structured data such as conditions, procedures, laboratory values, and medications and semi-structured data such as echocardiogram reports from the EHR. For those with multiple hospitalizations, a random hospitalization encounter was chosen as the start date, and outpatient medications, procedures, conditions, vital signs, and laboratory values were included, if available, to simulate a baseline phenotypic profile similar to the TOPCAT participants (Online Supplement).

Each data category required a separate rule-based mapping function with variable ICD-10-CDM code or variable name string-search (Supplemental Methods, Figure 1b, eTable 4). The primary outcome in the EHR cohorts was all-cause death, representing events occurring in the health system and supplemented with out-of-hospital death data from the CT death index. Events were counted beginning 30 days and ending 5 years after index hospitalization to assess extended use of spironolactone. Those patients without an outcome after 5 years were censored to match follow up time in TOPCAT, and patients who died within 30 days of index hospitalization were removed from outcome analysis in order to further match patients with lower acuity to the TOPCAT participants.

TOPCAT and EHR cohorts were pre-processed separately with continuous and binary categorical values. Pre-processing included imputation of missing values, removal of variables with high collinearity, and winsorization of outliers, which followed our previously described methods[19] (Supplemental Methods).

Phenotypic Distance Metric to Evaluate Representation Distance Between Cohorts

We defined a metric to summarize the differences across a cohort's complex multivariate differences by calculating a dissimilarity distance across covariates. We combined distances across the covariate landscape using Gower's distance, which is a similarity metric that incorporates mixed-type (categorical and continuous) data (Supplemental Methods).[23] We weighted each covariate distance by its prognostic significance, defined by the beta coefficient of a univariate Cox proportional hazards model to predict the hazard of the combined cardiovascular outcome in the TOPCAT control arm. We refer to this weighted Gower's distance as the "*phenotypic distance*". Larger cohorts were subsampled to have the same number of individuals as smaller cohorts for comparison.

To quantitatively assess the distance between cohorts, we compared the median Gower's distance distribution within a cohort with the distribution between two cohorts. The Gower's distance, however, is a dimensionless scaled metric of relative distance, so defining this distance between the TOPCAT and EHR cohorts is not directly interpretable. To address this, we defined the ratio of median phenotypic distances between two cohorts and within one cohort, named the *phenotypic distance metric (PDM)*, which quantifies the indexed dissimilarity when comparing individuals between two cohorts. A ratio greater than 1 represented a larger difference between cohorts than within one cohort.

In sensitivity analyses, we evaluated median differences in TOPCAT and the five hospital cohorts comparing the median distances within and between subgroups, including the TOPCAT spironolactone and placebo arms, and the 5 EHR based cohorts.

Individual EHR Patient Representation in RCT

We defined the position of each EHR patient within the phenotypic distribution of TOPCAT patients across prognostically relevant covariates. For this, we recalculated the weighted Gower's distance of each EHR cohort patient from the TOPCAT cohort covariates most prognostic for the combined cardiovascular outcome. Next, we defined an index TOPCAT participant, the one most phenotypically representative of TOPCAT participants, based on the shortest phenotypic distance to all other TOPCAT participants. We estimated the representation of each EHR cohort patient in TOPCAT by calculating their percentile of phenotypic distance from the index TOPCAT participant. Specifically, using the TOPCAT cohort distribution of phenotypic distance from the index TOPCAT participant, we determined the percentile of each EHR cohort patient within this distribution, representing the position of each EHR patient within the phenotypic distribution of TOPCAT patients.

RCT-derived clinical effect estimates against treatment patterns in the EHR

Based on our prior work[19,24], we also used TOPCAT to define a personalized treatment effect estimate as a function of patient covariates (Online Supplement). We then calculated each EHR patient's estimated individualized hazard ratios (iHRs) based on the model developed in TOPCAT. We compared the all-cause mortality outcome of those EHR patients with an expected benefit from spironolactone (iHR<1) across strata of those receiving and not receiving spironolactone in the clinical setting.

Statistical Analysis

We summarized categorical variables by number and proportion present in each group and continuous variables as mean and standard deviation or median and interquartile range. Categorical variables were compared between the two cohorts using a chi-squared test and continuous variables using Welch's two-sided t-test.[25] We also calculated the standardized mean difference for each covariate between each cohort pair and the median standardized mean difference with IQR for each TOPCAT-EHR cohort pair and each pair of EHR cohorts.[25]

We calculated the PDM as described above with interquartile values of the metric. We depicted the qualitative difference between the TOPCAT cohort and the EHR cohort by projecting the phenotypic distances onto a dimensionality reduction method called uniform manifold approximation and projection (UMAP).[19,26,27] The method projects the high-dimensional dataset onto two dimensions by ensuring points are closest to their nearest neighbors while also attempting to preserve the global representation of each point in the manifold.

Group-level outcome hazard rate differences were evaluated using Cox proportional hazards models over five years with the treatment arm as an independent covariate. We adjusted for age, sex, diabetes mellitus, and prior heart failure hospitalization, representing covariates that were adjusted for in the reported analyses of TOPCAT. All statistical tests were 2-sided with a pre-specified Type 1 Error rate of 0.05.

The Yale Institutional Review Board reviewed this study, and a waiver of consent was granted because it was a retrospective study of medical records.

RESULTS

Populations Characteristics

The TOPCAT trial had 3445 participants with a median age of 69 (61-76, 25-75% IQR) years and included 1775 (52%) women. Of the trial population, 1722 (50%) were assigned to the spironolactone arm and 1723 (50%) to the placebo control arm. The EHR cohorts included 30,858 patients with a diagnosis of heart failure. Of these, 12,548 (41%) had one or more hospitalizations with a principal or a secondary diagnosis of heart failure (91,404 hospitalizations overall) and had at least one echocardiogram across either inpatient or outpatient settings, demonstrating an LVEF ≥45% and no prior echocardiograms with an LVEF <45%. There were 11,712 patients included in the final EHR cohorts who had at least one echocardiogram within six months of their hospital admission, similar to the TOPCAT inclusion criteria. Among the index hospitalization randomly chosen for each patient, 3588 (30%) patients were treated at YNHH YSC, 3435 (29%) at YNHH SRC, 2637 (22%) at BH, 591(5%) at GH, and 1461(12%) at LMH. Across these sites, the median age of patients ranged from 76 (IQR 65-85) years to 84 (IQR 76-90) years. The proportion of women ranged from 54% to 61%. Overall, the EHR cohorts were older, had more women, and had a higher proportion of minorities compared with TOPCAT (Table 1).

Similarity and representation between RCT and real-world EHR cohorts

Both by qualitative UMAP visualizations (eFigure 1) and quantitative similarity distance comparisons (Table 2), there was a phenotypic separation between the TOPCAT and EHR cohorts (median phenotypic distance of 0.23 (IQR 0.21-0.27, Figure 2b). These differences are in contrast to the treatment and placebo arms within the TOPCAT cohort, as well as pairs of sites in the EHR cohort that were less different (median phenotypic distance of 0.20 (IQR 0.18-0.23), Figures 2a and 2c). Acknowledging that phenotypic differences exist within each cohort simply due to the variation of phenotypic profiles, we indexed the similarity between cohorts to the

similarity of phenotypic distances within the reference cohort, thus defining the ratio as the PDM. Using this approach, we confirmed the phenotypic similarity between the treatment and placebo arms of TOPCAT (PDM of 0.99 (IQR 0.98-1.0)) and estimated a median phenotypic distance metric between the EHR cohorts of (PDM of 0.98 (IQR 0.96-1.0)). In contrast, the PDM between the pooled EHR cohorts and the TOPCAT population was 10% above 1, at 1.1 (1.1-1.2). With regards to real-world individual representation within TOPCAT, we compared all EHR patients to the most representative TOPCAT participant, or the individual closest in phenotypic distance to all other participants, termed the index TOPCAT participant. All patients in the EHR cohorts were further in phenotypic distance from the index TOPCAT participant than 55% of the TOPCAT participants. In addition, the average patient in the EHR cohorts was further from the index TOPCAT patient than 80% of the TOPCAT patients.

Outcomes in real-world cohorts based on expected therapeutic effects in the RCT

Based on the individualized treatment effectiveness model developed in TOPCAT, 10,519 of the 10,548 EHR patients with an outcome were predicted to benefit from spironolactone use (iHR of <1). Of these, 1,119 (11%) were receiving the medication before outcome measurement. Within the EHR cohorts, patients with a high predicted benefit from spironolactone who also were on spironolactone had lower mortality than those who were not on spironolactone (Online Supplement). This was statistically significant even after adjusting for age, sex, a history of diabetes mellitus, and a prior heart failure hospitalization, with an adjusted hazard ratio for risk to the first occurrence of all-cause mortality of 0.83 (95% confidence interval 0.74-0.94, p-value <0.001) (Figure 3, eTable 6).

DISCUSSION

In this study, we demonstrate a strategy to quantify the multi-dimensional representation gap between an RCT and real-world patients in the EHR by presenting an approach to map computable phenotypes of RCT participants to real-world clinical populations. We propose a quantitative metric that computes information across all available covariate axes to define a unifying similarity score across cohorts, and apply the score with individualized outcome information in the RCT to identify EHR patients most likely to benefit from the RCT intervention. The PDM quantified differences between TOPCAT and the EHR cohorts while quantifying similarities between the 5 EHR cohorts or between TOPCAT treatment arms. In addition, all individual EHR patients were further in phenotypic distance from a representative TOPCAT participant compared to the majority of TOPCAT participants, supporting quantification of the representation gap. We also demonstrate that based on an individualized treatment response model developed among TOPCAT participants, 99% of EHR patients with HFpEF were those who would be expected to benefit from spironolactone use, and that within these patients, those on spironolactone experienced a 26% reduction in major cardiovascular outcomes.

Our study builds upon the literature evaluating the representativeness of RCTs for real-world settings. Prior approaches have focused on defining differences across a limited number of covariate axes and have also largely been used on one-to-one comparisons across singular covariate or solely categorical data, with the information across covariate comparisons interpreted qualitatively.[12–15,28–30] Our approach addresses these challenges by deploying a strategy that goes beyond describing differences between the RCT and the EHR on individual features and instead quantifies the differences across multiple axes. This is innovative as it overcomes the fallacy of comparing the average distribution of covariates across the entire

population against the averages across another population, which ignores stark differences in various clinical subpopulations that are not identified by focusing on the average. In addition, we apply a quantitative phenotypic distant metric across five sites within four different hospitals that demonstrate the flexibility of the application across different cohorts.

Our study has important clinical implications. When clinicians assess whether a patient would benefit from a particular treatment, they often refer to relevant RCTs for intervention in medical practice. They must also, however, look at the patient in front of them and determine how effectively the study translates to care. They ascertain (1) how generalizable is the study result to my patient? and (2) was my patient well-represented in this trial? The clinician considers not only demographics or comorbidities but also the entire picture of the patient. Our approach addresses each aspect of these decision quantitatively and interpretatively by providing the phenotypic distance metric to assess where the patient in question lies along the phenotypic distribution of the RCT participants and to suggest whether the patient may benefit from the intervention based on phenotypic similarity to the RCT participants.

Our study uses TOPCAT as an example of RCT to demonstrate a process to capture the complex multidimensional picture of each RCT participant and to compare them directly to real-world patients such as those in the EHR. We describe a method to predict benefits in the real-world population based on individual patient characteristics and covariates deemed important by RCT individuals and also assess their representation in the RCT. Our quantification of the large representation gap between TOPCAT and EHR patients with HFpEF suggest an overarching need to assess representation from a multi-dimensional perspective during trial implementation and interim analysis.[31] This further supports an increased role of tracking trial recruitment against real-world populations.

Another observation from our work is the intensive nature of mapping data from clinical trials to real-world populations in the EHR. We highlight the challenges and variability of translating RCT cohorts and study variables to the EHR setting, suggesting another impedance to translating RCT evidence to our patients. Common data models such as the Observational Medical Outcomes Partnership standardize EHR mapping to make this research easier to accomplish and apply across multinational institutions.[32] As trials increasingly become pragmatic, there is an urgent need to computably define RCT conditions within the context of the EHR since the manual approach to identify key features will represent a challenge for trial operations.

There are limitations that merit consideration. Representation of real-world populations in the EHR is a unique challenge since the data represent a snapshot of time when the patient interacts with the medical system, and the patients seeking care likely represent a subset of the HFpEF population. The patients included, however, represented five sites with unique and diverse patient populations, thus maximizing the possible landscape of patients with HFpEF. Second, we chose a set of covariates available across the RCT and the EHR, which may not be a fully representative set of conditions that differ between individuals. However, features captured in large RCTs are often comprehensive, and our clinician-led approach designed a strategy to map many of these conditions using text phrases, billing codes, and all possible patient encounters in the system. We also confirmed the robustness of our approach with sensitivity analyses that focused on prognostically relevant conditions with similar results. Third, we chose Gower's distance given its flexibility with modeling both categorical and continuous data and its ability to weigh some conditions more relative to the others. Although this represents one of many Euclidean distances appropriate for assessing differences across covariates, it has been

found superior to the other methods in identifying phenotypic differences.[33] Finally, outcome risk ratios calculated in the EHR patients with HFpEF are susceptible to both ascertainment and unmeasured confounding by indication. We mitigated the bias by adjusting for the same covariates as seen in TOPCAT and focusing only on those patients predicted to have benefit from spironolactone use.

We propose a novel approach to evaluating the real-world representativeness of RCT participants against corresponding patients in the EHR across the full multidimensional spectrum of the represented phenotypes. This enables the evaluation of the implications of RCTs for real-world patients.

DATA AVAILABILITY

The TOPCAT cohort is publicly available through the National Heart, Lung, and Blood Institute Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) The TOPCAT dataset is available at https://biolincc.nhlbi.nih.gov/studies/topcat/. The Yale electronic health record cohorts are not available due to the use of patient data.

COMPETING INTERESTS

Dr. Thangaraj, Dr. Oikonomou, and Dr. Khera are coinventors of a provisional patent not related to the current work (63/606,203). Dr. Khera is an Associate Editor of JAMA and receives research support, through Yale, from the Blavatnik Foundation, Bristol-Myers Squibb, Novo Nordisk, and BridgeBio. He is a coinventor of U.S. Provisional Patent Applications 63/177,117, 63/428,569, 63/346,610, 63/484,426, 63/508,315, 63/580,137, 63/562,335, and a co-founder of Ensight-AI, Inc and Evidence2Health, LLC. Dr. Oikonomou is an academic co-founder of Evidence2Health LLC, and has been a consultant for Caristo Diagnostics, Ltd and Ensight-AI, Inc. He is a co-inventor in patent applications (US17/720,068, 63/619,241, 63/177,117, 63/580,137, 63/606,203, 63/562,335,WO2018078395A1, WO2020058713A1) and has received royalty fees from technology licensed through the University of Oxford. Dr. Suchard receives grants and contracts from the US Food & Drug Administration, the US Department of Veterans Affairs and Johnson & Johnson, all outside the scope of this work.

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 grants from the National Institutes of Health (5T32HL155000-03 and 1F32HL170592-01, respectively).

REFERENCES

1. 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.

2. Spieth PM, Kubasch AS, Penzlin AI, Illigens BM-W, Barlinn K, Siepmann T. Randomized controlled trials - a matter of design. *Neuropsychiatr Dis Treat*. 2016;12:1341–1349.

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

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

5. 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.

6. Reza N, Gruen J, Bozkurt B. Representation of women in heart failure clinical trials: Barriers to enrollment and strategies to close the gap. *Am Heart J Plus*. 2022;13.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

12. Rogers JR, Hripcsak G, Cheung YK, Weng C. Clinical comparison between trial participants and potentially eligible patients using electronic health record data: A generalizability assessment method. *J Biomed Inform*. 2021;119:103822.

13. 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.

14. Sen A, Chakrabarti S, Goldstein A, Wang S, Ryan PB, Weng C. GIST 2.0: A scalable multi-trait metric for quantifying population representativeness of individual clinical studies. *J Biomed Inform*. 2016;63:325–336.

15. Weng C, Li Y, Ryan P, et al. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Appl Clin Inform*. 2014;5:463–479.

16. Dunlay SM, Roger VL, Redfield MM. Epidemiology of heart failure with preserved ejection fraction. *Nat Rev Cardiol*. 2017;14:591–602.

17. Shah SJ. Innovative Clinical Trial Designs for Precision Medicine in Heart Failure with Preserved Ejection Fraction. *J Cardiovasc Transl Res*. 2017;10:322–336.

18. Shah SJ, Katz DH, Selvaraj S, et al. Phenomapping for Novel Classification of Heart Failure With Preserved Ejection Fraction. *Circulation*. 2015;131:269–279.

19. 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.

20. Pitt B, Pfeffer MA, Assmann SF, et al. Spironolactone for Heart Failure with Preserved Ejection Fraction. *N Engl J Med*. 2014;370:1383–1392.

21. Pfeffer MA, Claggett B, Assmann SF, et al. Regional Variation in Patients and Outcomes in the Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT) Trial. *Circulation*. 2015;131:34–42.

22. Cohen JB, Schrauben SJ, Zhao L, et al. Clinical Phenogroups in Heart Failure With Preserved Ejection Fraction: Detailed Phenotypes, Prognosis, and Response to Spironolactone. *JACC Heart Fail*. 2020;8:172–184.

23. Gower JC. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*. 1971;27:857–871.

24. 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.

25. Pollard TJ, Johnson AEW, Raffa JD, Mark RG. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open*. 2018;1:26–31.

26. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [statML]*. 2018.

27. 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.

28. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. *NPJ Digit Med*. 2020;3:67.

29. Belkin MN, Blair JE, Shah SJ, Alenghat FJ. A composite metric for predicting benefit from spironolactone in heart failure with preserved ejection fraction. *ESC Heart Fail*. 2021;8:3495–3503.

30. Wang SV, Schneeweiss S, RCT-DUPLICATE Initiative, et al. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. *JAMA*. 2023;329:1376–1385.

31. Oikonomou EK, Thangaraj PM, Bhatt DL, et al. An explainable machine learning-based phenomapping strategy for adaptive predictive enrichment in randomized controlled trials. *medRxiv*. 2023. Published onlineNovember 1, 2023. https://doi.org/10.1101/2023.06.18.23291542.

32. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci U S A*. 2016;113:7329–7336.

33. Moore JH, Li X, Chang J-H, et al. SynTwin: A graph-based approach for predicting clinical outcomes using digital twins derived from synthetic patients. *Pac Symp Biocomput*. 2024;29:96–107.

Figures

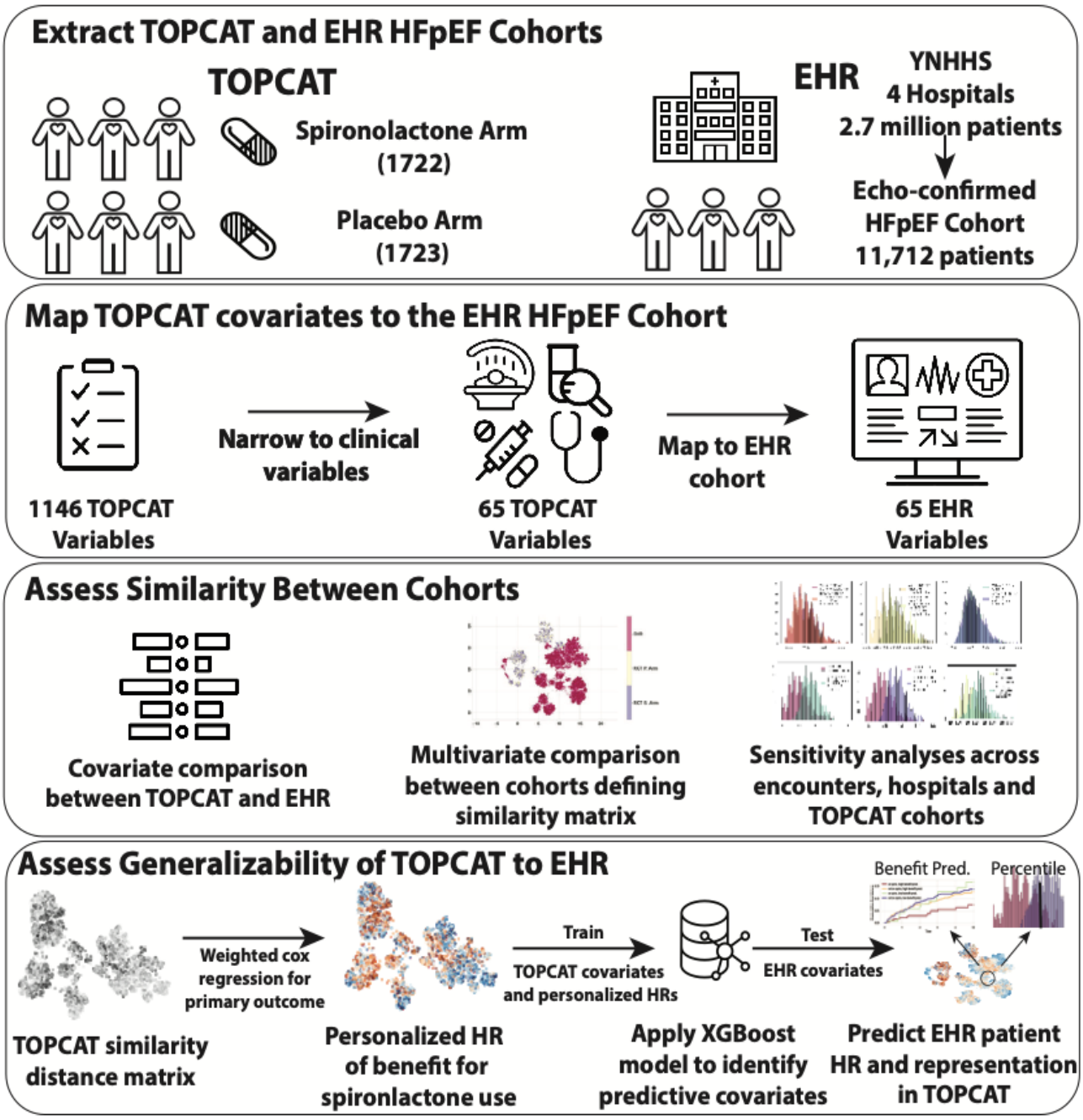

Figure 1: Study Overview A: Depiction of the extraction of TOPCAT and EHR patients with HFpEF. B: Second Panel: Narrowing and acquisition of TOPCAT variables to 65 clinically relevant variables with rule-based mapping to EHR variables. C: Third panel: univariate and multivariate covariate comparison between TOPCAT and EHR patients with HFpEF across multiple sensitivity analyses comparing the different TOPCAT treatment arms and those with and without prior HF hospitalizations with first and last admission encounters within the EHR patients with HFpEF, and two definitions of EHR patients with HFpEF derived by EF. Fourth

panel: Assessing the generalizability of TOPAT by deriving personalized hazard ratios for spironolactone use in the TOPCAT participants, training an extreme gradient boosting classifier on the most important covariates for determining these HRs and applying it to the EHR patients with HFpEF, comparing composite cardiovascular outcomes in the EHR patients with HFpEF stratified by predicted personalized benefit and being on spironolactone, and finally assessing the representation of EHR patients with HFpEF in the TOPCAT population by calculating the percentile distance each EHR patient is from the median TOPCAT participant. Abbreviations: EHR: Electronic Health Record, HR: Hazard Ratio, YNHHS: Yale New Haven Hospital System

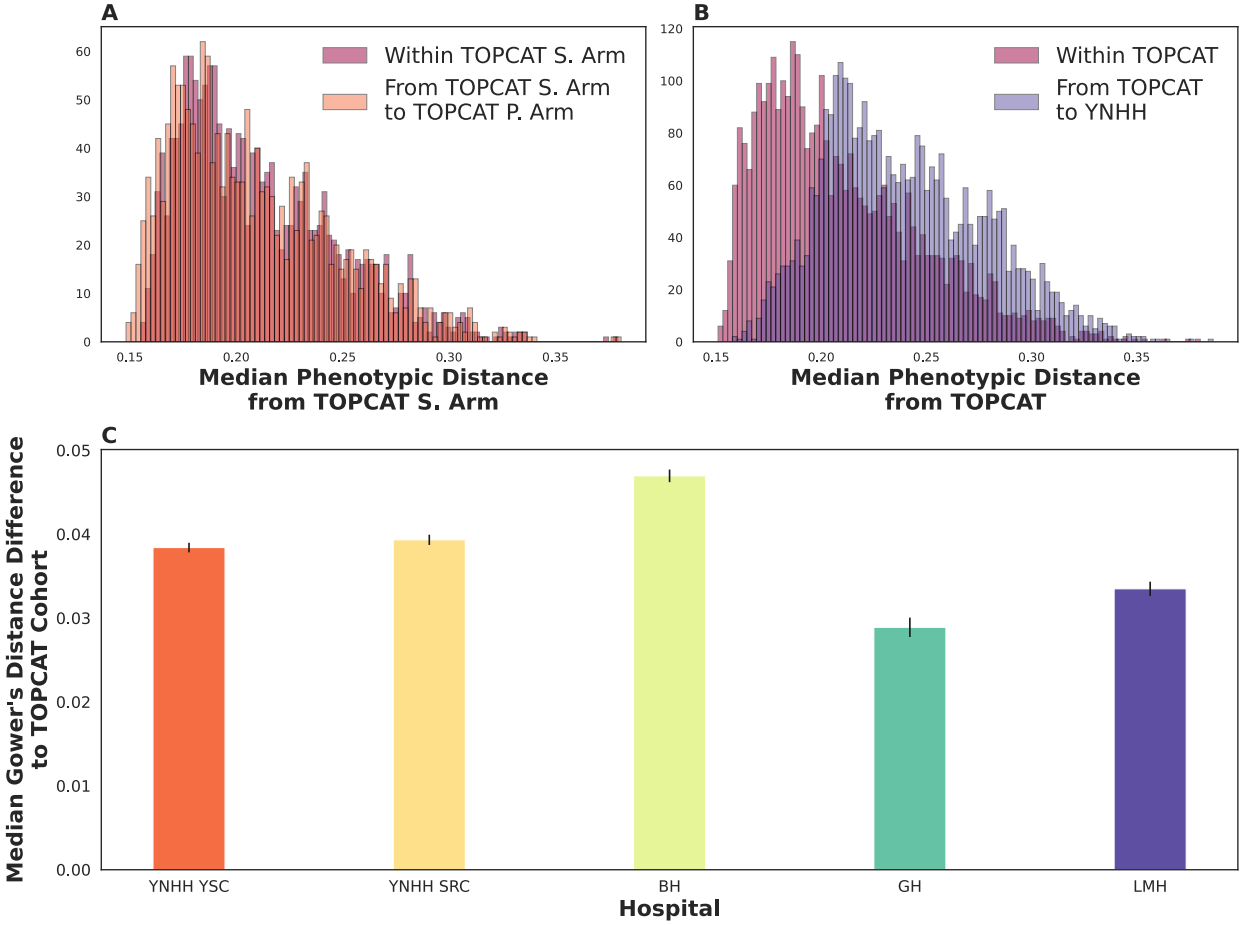

Figure 2: Median Phenotypic Distance within TOPCAT subgroups and between TOPCAT and EHR subgroups. A: Histogram of median phenotypic distance between TOPCAT subgroups: the TOPCAT spironolactone (S.) Arm (dark red), and the TOPCAT Placebo (P.) Arm (dark orange). B: Histogram of median phenotypic distance between the TOPCAT cohort (Dark red) and EHR cohort (Purple). C: Median phenotypic distance difference of each subgroups from the entire EHR cohort median phenotypic distance. Orange is York Street Campus (YNHH YSC), Yellow is St. Raphael's Campus (SRC), Lime green is Bridgeport Hospital (BH), Green in Greenwich Hospital (GH), and purple is Lawrence and Memorial Hospital (LMH).

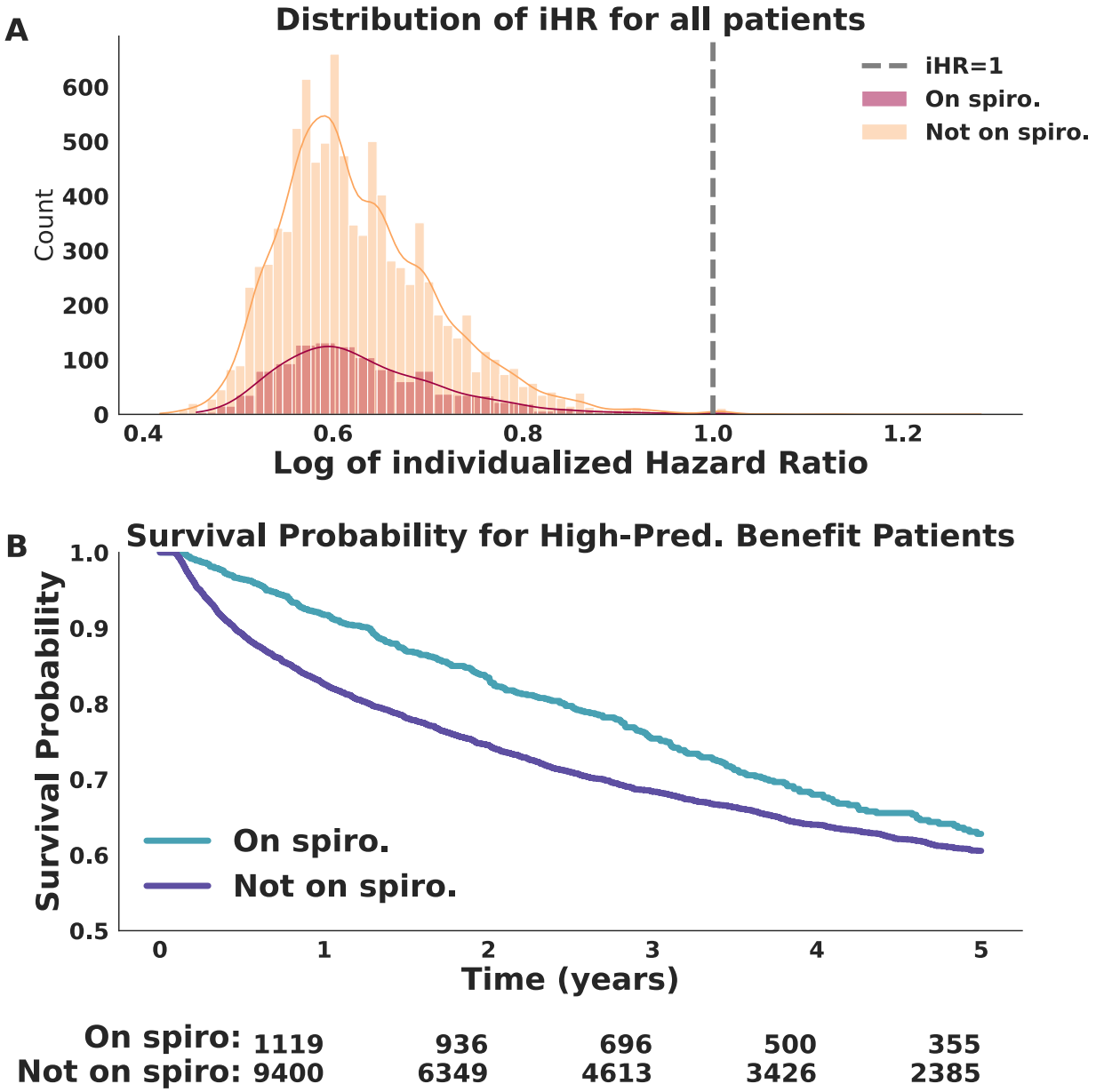

Figure 3: Distribution of individualized hazard ratios of EHR cohort and Kaplan-Meier curve of EHR patients with high predicted spironolactone benefit. (A) Distribution of log of individualized hazard ratios of time to cardiovascular event stratified by spironolactone use predicted for the EHR outcome cohort. The vertical line represents an individualized hazard ratio of 1, and values to the left of the dotted line represent high predicted benefit of spironolactone use. Red represents patients on spironolactone, orange represents patients not on spironolactone. (B) Survival probability versus time to event of composite cardiovascular outcome in EHR patients predicted to have high benefit with spironolactone stratified by being on spironolactone

(blue) or not on spironolactone (purple). Text below each plot represents number at risk for the specified group. Abbreviations: Spiro. stands for spironolactone, iHR stands for individualized hazard ratio for time to cardiovascular event stratified by spironolactone use, High-Pred stands for High Predictive.

TABLES

Table 1: Selected Population Characteristics of the TOPCAT participants compared with EHR patients with HFpEF in 5 different hospital sites.

Covariate	TOPCAT (N=3445)	YNHH YSC (N=3588)	YNHH SRC (N=3435)	BH (N=2637)	LMH (N=1461)	GH (N=591)	P-value
Age—median years (25-75% IQR)	69 [61,76]	76 [65,85]	80 [69,88]	79 [68,87]	79 [69,87]	84 [76,90]	<0.001
Female sex N (%)	1775 (52)	1947 (54)	2102 (61)	1530 (58)	798 (55)	327 (55)	<0.001
Race or ethnicity N (%)							
Asian	19 (0.6)	36 (1.0)	30 (0.9)	30 (1.1)	11 (0.8)	11 (1.9)	0.021
Black	302 (8.8)	664 (19)	538 (15)	493 (19)	100 (6.8)	23 (3.9)	<0.001
Other	70 (2.0)	235 (6.5)	198 (5.8)	327 (12)	109 (7.5)	57 (9.6)	<0.001
White	3062 (90)	2802 (78)	2562 (75)	1806 (68)	1245 (85)	502 (85)	<0.001
Hispanic or Latino	321 (9.3)	232 (6.8)	232 (6.8)	349 (13.2)	87 (6.0)	38 (6.4)	<0.001
Left Ventricular Ejection Fraction median % (25-75% IQR)	56 [51,61]	62 [56,67]	62 [56,67]	61 [57,65]	62 [58,68]	59 [55,64]	<0.001
Selected Vital Signs							
Systolic Blood Pressure median mmHg (25-75% IQR)	130 [120,140]	127 [112,144]	127 [114,144]	133 [119,149]	131 [118,145]	129 [113,146]	<0.001
BMI median kg/m^2 (25-75% IQR	31 [27,36]	29 [24,35]	29 [24,36]	29 [25,35]	29 [25,36]	26 [23,31]	<0.001
Selected Conditions and Procedures							
Atrial Fibrillation	1214 (35.2)	1882 (52.5)	1788 (52.1)	1365 (51.8)	856 (58.6)	375 (63.5)	<0.001
Atrial Fibrillation	500 (14.5)	1882 (52.5)	1788 (52.1)	1365 (51.8)	856 (58.6)	375 (63.5)	<0.001
Percutaneous Coronary Intervention	500 (14.5)	333 (9.3)	211 (6.1)	221 (8.4)	116 (7.9)	33 (5.6)	<0.001
Selected Medications							
Anti-hypertensives	3419 (99.2)	3257 (90.8)	3125 (91.0)	2394 (90.8)	1315 (90.0)		<0.001
Beta Blocker	2679 (77.8)	2028 (56.5)	1871 (54.5)	1600 (60.7)	759 (52.0)		<0.001

P-value was based on Kruskal-Wallis test for the continuous variables and chi-square test for categorical variables across groups. Vital signs of EHR patients were either taken during an outpatient visit or were at discharge from hospitalization if outpatient values were missing. Left

Ventricular Ejection Fraction of EHR patients were measured during an echocardiogram within 6 months of a hospital admission, similar to TOPCAT. Abbreviations: N: number of patients, bpm:beats per minute, IQR: interquartile range, BMI: body mass index, and mmHg: millimeters of mercury.

Table 2: Median of Median Phenotypic Distance and Median Phenotypic Distance Metric Across Cohorts

A. Baseline Cohort (Below)	Comparison Cohort (Right)	**Median Phenotypic Distance (IQR 25-75%)**						
		TOPCAT S. Arm	**TOPCAT P. Arm**	**YNHH YSC**	**YNHH SRC**	**BH**	**GH**	**LMH**
TOPCAT S. Arm		0.20 (0.18-0.23)	0.20 (0.018-0.23)	0.23 (0.21-0.27)	0.24 (0.21-0.27)	0.24 (0.22-0.27)	0.23 (0.20-0.26)	0.23 (0.20-0.26)
TOPCAT P. Arm			0.20 (0.18-0.23)	0.23 (0.21-0.26)	0.23 (0.21-0.26)	0.24 (0.22-0.27)	0.23 (0.21-0.25)	0.23 (0.20-0.26)
YNHH YSC				0.20 (0.18-0.22)	0.20 (0.18-0.23)	0.21 (0.19-0.23)	0.19 (0.17-0.22)	0.19 (0.17-0.22)
YNHH SRC					0.21 (0.19-0.24)	0.20 (0.19-0.23)	0.19 (0.17-0.23)	0.19 (0.17-0.22)
BH						0.21 (0.19-0.25)	0.20 (0.18-0.24)	0.20 (0.18-0.24)
GH							0.20 (0.18-0.22)	0.18 (0.17-0.21)
LMH								0.19 (0.17-0.21)
B. Baseline Cohort		**Phenotypic Distance Metric (IQR 25-75%)**						
TOPCAT S. Arm			0.99 (0.98, 1.0)	1.1 (1.0,1.2)	1.2 (1.1,1.2)	1.2 (1.1,1.3)	1.1 (1.0-1.2)	1.1 (1.0-1.2)
TOPCAT C. Arm				1.2 (1.1, 1.2)	1.2 (1.1, 1.2)	1.2 (1.1-1.3)	1.2 (1.1-1.3)	1.1 (1.0-1.2)
YNHH YSC					0.99 (0.98, 1.0)	1.0 (1.0-1.0)	0.96 (0.92-1.0)	0.98 (0.96-1.0)
YNHH SRC						1.0 (1.0-1.1)	0.96 (0.91-1.0)	0.97 (0.96-1.0)
BH							0.95 (0.88-1.0)	0.94 (0.91-0.99)
GH								0.98 (0.95-1.0)

(A) Median of Median Phenotypic Distance- Median of median phenotypic distance between baseline cohort (first column) and comparator cohort (first row). (B) Median phenotypic distance metric. The ratio of the median phenotypic distance between the baseline and comparator cohort and the median phenotypic distance within the baseline cohort. Abbreviations: BH: Bridgeport Hospital, GH: Greenwich Hospital, and LMH: Lawrence and Memorial Hospital, IQR: interquartile range, YNHH YSC: York Street Campus, and YNHH SRC: St. Raphael's Campus.