An explainable machine learning-based phenomapping strategy for adaptive predictive enrichment in randomized controlled trials

Original research

Evangelos K Oikonomou[1], Phyllis M. Thangaraj[1], Deepak L. Bhatt[2], Joseph S Ross[3], Lawrence H Young[1], Harlan M. Krumholz[1,4], Marc A Suchard[5], Rohan Khera[1,4*]

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

[2] Mount Sinai Heart, Icahn School of Medicine at Mount Sinai Health System, New York, NY, USA

[3] Section of General Internal Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

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

[5] Departments of Computational Medicine and Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA

Manuscript type: original research

Word count (Introduction, Results, Discussion): 3572 words

Main figures: 5

Funding: National Heart, Lung, and Blood Institute of the National Institutes of Health (under award K23HL153775 to RK) & Doris Duke Charitable Foundation (under award 2022060 to RK)

Disclosures: see full Disclosure statement in pages 17-18

Brief title: Machine learning for adaptive trial enrichment

***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 controlled trials (RCT) represent the cornerstone of evidence-based medicine but are resource intensive. We propose and evaluate a novel machine learning (ML) strategy of adaptive predictive enrichment through computational trial phenomaps to optimize RCT enrollment. In simulated group sequential analyses of two large cardiovascular outcomes RCTs of (1) a therapeutic drug (pioglitazone versus placebo; Insulin Resistance Intervention after Stroke or IRIS), and (2) a disease management strategy (intensive versus standard systolic blood pressure reduction in SPRINT), we constructed dynamic phenotypic representations to infer profiles of patients benefiting from the intervention versus control during interim trial analyses and examined their association with study outcomes. Across three interim analyses, our strategy learned dynamic phenotypic signatures predictive of individualized cardiovascular benefit in each arm. By conditioning a prospective candidate's probability of enrollment on their predicted benefit, we estimate that our approach would have enabled a reduction in the final trial size across five simulations (IRIS: -18 ± 4.7%, p=0.008; SPRINT: -27.4 ± 3.4%, p=0.002), while preserving the original average treatment effect (IRIS: hazard ratio of 0.71 ± 0.01 for pioglitazone vs placebo, vs 0.76 in the original trial; SPRINT: hazard ratio of 0.72 ± 0.01 for intensive vs standard systolic blood pressure, vs 0.75 in the original trial; all comparisons with p<0.01). This adaptive framework has the potential to maximize RCT enrollment efficiency.

INTRODUCTION

Large randomized controlled trials (RCTs) represent the cornerstone of evidence-based medicine and are the scientific and regulatory gold standard.[1,2] Despite their strengths, they are often both resource and time intensive.[3] The required investments are particularly large for RCTs evaluating the effects of novel therapies on major clinical endpoints, such as mortality or acute cardiovascular events, among patients with chronic cardiometabolic or other disorders.[4–6] Modern pivotal trials have been facing exponentially rising costs as more patients and clinic visits are needed to prove a treatment effect, with the median cost of pivotal clinical trials per approved drug estimated at $48 million and a median cost of $41413 per patient enrolled (2005-2017). Across studies, the largest single factor driving cost was the number of patients required to establish the treatment effects.[5,6] With a growing pipeline of potential new therapeutics, there is a need to explore alternative methods of conducting RCTs which would increase their efficiency, maintain their robustness, and provide high-quality evidence to ensure patient safety and regulatory compliance.[7,8]

Adaptive trials have been proposed as a potential solution,[9,10] as highlighted in a recent United States Food and Drug Administration (FDA) statement.[11] Adaptive trials allow prospectively planned modifications to the study design of clinical trials based on accumulating data from patients already enrolled in the trial.[12] The ability to adjust the trial to new information has the theoretical advantage of improving statistical efficiency while ensuring safety by detecting early harm or lack of effectiveness.[12–14] However, defining the ways in which to adapt a trial *a priori* remains challenging.

We have recently described a machine learning (ML) method that leverages the phenotypic diversity of patients in RCTs and the random allocation of the intervention to define signatures of individualized treatment effects.[15–17] Our method is based on an approach that creates a multidimensional representation of an RCT population across all pre-randomization features ("phenomap") and extracts signatures that define consistent benefit or risk from each study arm. This approach has been validated retrospectively across several RCTs,[15–17] however its utility in an adaptive trial design has not been explored.

In the present study, we evaluate an adaptive approach that uses study arm effect differences for similar patients grouped on their complex phenotypic features, to design predictively enriched clinical trials. We demonstrate the application of this hypothesis using individual participant data from two large cardiovascular outcomes trials; a double-blind, placebo-controlled, randomized trial of a drug (pioglitazone, as studied in the Insulin Resistance Intervention after Stroke [IRIS] trial,[18] and a disease management strategy (intensive versus standard blood pressure reduction in the Systolic Blood Pressure Intervention Trial [SPRINT]),[19] assessing the effects of our proposed approach on the efficiency of detecting primary efficacy, safety end points, as well as, the final trial composition.

RESULTS

Study population

The study was designed as a post hoc simulation of real-world clinical trial data from IRIS[18] and SPRINT (**Fig. 1A)**.[19] The detailed protocol, study population demographics, and results have been previously reported and are further summarized in the **Methods**.[18,19] Briefly, IRIS included 3876 patients, 40 years of age or older, with a recent ischemic stroke or a transient ischemic attack (TIA) during the 6 months before randomization, who did not have diabetes mellitus but had evidence of insulin resistance based on a homeostasis model assessment of insulin resistance (HOMA-IR) index score of 3.0 or greater. Participants were randomly assigned in a 1:1 ratio to receive either pioglitazone or a matching placebo. Enrollment occurred over 2899 days (7.9 years), from February 2005 until January 2013, with the final study report including 3876 patients (median age 62 [55-71] years, n=1338 (34.5%) women). Participants were followed for a median of 4.7 [3.2-5.0] years for the primary

endpoint of fatal or non-fatal stroke or myocardial infarction, which occurred in 175 of 1939 (9.0%) participants in the pioglitazone group and 228 of 1937 (11.8%) participants in the placebo group.

The SPRINT trial enrolled 9361 participants (median age 67 [61-76] years, n=3332 [35.6%] women) with a systolic blood pressure (SBP) of 130-180 mm Hg as well as an additional indicator of cardiovascular risk, with random assignment to targeting an SBP of less than 120 mm Hg (intensive treatment arm) versus less than 140 mm Hg (standard treatment arm). Patients with diabetes mellitus, prior stroke, or dementia were excluded. To simulate a longer enrollment period similar to IRIS, we modeled a steady recruitment years over 5 (vs 2.4 years in the original study) and used the original follow-up data with a median period of 3.2 [2.7-3.8] years, for the primary composite outcome of time-to-first myocardial infarction/acute coronary syndrome, stroke, acute decompensated heart failure, or cardiovascular mortality, which occurred in 243 of 4678 (6.8%) participants in the intensive and 319 of 4683 (5.2%) participants in the standard arm.

Defining a group sequential trial design

We defined a group sequential design, with three total interim analyses planned before the final analysis, which was performed once all primary events had occurred in the original trial. Each analysis examined intervention superiority, assuming a power of 80% and a one-sided type I error of 0.025,[20] providing adjusted significance levels (alpha) at each analysis timepoint based on the O'Brien-Fleming and alpha-spending Pocock methods (**Extended Data Table S1**). For IRIS, we assumed that the primary outcome would occur in 11.8% vs 9% of the placebo- and pioglitazone-treated arms. This was based on the observed outcomes given the amendments made in the original trial during its course. For SPRINT, we assumed the respective primary outcome would occur in 6.8% vs. 5.4% of the standard and intensive arms, per the original power calculations. We defined the interim analysis timepoints based on the occurrence of the first 50, 100 and 150 primary outcome events in the original trial. In IRIS, this corresponded to 1121, 1582, and 1905 days after the first patient was randomized. In the simulated SPRINT analysis, this corresponded to 613 days, 921 days, and 1137 days after the patient was randomized (**Fig. 1B-C)** in the trial. In both trials, all interim analyses were performed during periods of active trial enrollment.

Learning machine learning, phenomapping-derived signatures of treatment benefit

At each of the 3 interim analyses, we adapted and implemented an ML algorithm that learned signatures of individualized treatment response to the intervention arm (pioglitazone, intensive SBP reduction) versus control based on data available at that time. The algorithm (**Fig. 2**) is described in detail in the **Methods (subsections 4 & 5)**, and is based on our prior work.[15–17]

Briefly, for each interim analysis, baseline characteristics of participants were defined based on participant assessments before randomization (summarized in **Extended Data Tables S2-3**).[17] Participants recruited up until that stage were randomly split into training/cross-validation (50%) and testing sets (50%). In the training set, baseline data were pre-processed and used to define a phenomap, which represented the phenotypic architecture of the recruited population at that timepoint across all axes of baseline variance.

Through iterative analyses centered around each unique individual and weighted for each individual participant's location in the phenotypic space,[17] we defined individualized estimates of the effects of the studied intervention, as compared to control, for the primary outcome. Subsequently, using extreme gradient boosting and the Boruta SHAP (Shapley additive explanations) feature selection algorithm we built an ML framework to identify key features that collectively determined a phenotypic signature (algorithm) predictive of these individualized estimates.

The predictive algorithm (relying exclusively on pre-randomization features) was then applied in the testing set. We assessed for evidence of heterogeneous treatment effects (defined based on a *p*

value for interaction of <0.2) by dichotomizing the population into two groups based on their predicted response, avoiding major imbalance in our subgroups by restricting the smallest group size to 20% of the population.

If there was potential evidence of heterogeneity based on this analysis, sample size calculations were updated at that interim analysis timepoint by revising the expected effect size (under the assumption of predictive enrichment) at the original power and alpha levels to ensure that predictive enrichment would maintain sufficient power at a sample size equal to or smaller than the originally planned one. We performed sample size calculations assuming various levels of predictive enrichment, which ranged from enrolling 50% to 95% of all remaining candidates, in 5% increments. If there were several levels that met these criteria, we ultimately chose the predictive enrichment level that minimized the required sample.

Once a predictive model had been generated and our analysis in the testing set had met criteria for possible heterogeneity in the treatment effect with sufficient power for the primary outcome, we chose to proceed with predictive enrichment. Over the subsequent period (time between the last and next interim analyses) the model was prospectively applied to all prospective trial candidates. For example, a model trained at interim analysis timepoint #1 was applied to individuals screened between the interim analysis timepoints #1 and #2 to furnish a probability of enrollment, with all original trial participants during this period considered eligible candidates. For all candidates, the probability of being enrolled was conditioned on their predicted individualized benefit, ultimately enriching the population at the level defined during the last interim analysis sample size calculation. Alternatively, if there was no evidence of heterogeneous treatment effect, or the proposed enrichment in enrollment would not be adequately powered at a sample size equal to or lower than the originally planned one, we proceeded as originally planned and continued with standard enrollment for that period without predictive enrichment. Given the stochastic nature of the algorithm, all simulations were repeated r=5 times.

Primary (efficacy) outcomes

IRIS: In the IRIS trial, phenomapping performed at the pre-specified timepoints identified baseline features, such as use of antihypertensive therapies, smoking, hypercholesterolemia, the patient's reported sex and prior history of stroke or transient ischemic attack as possible predictors of treatment effect heterogeneity for pioglitazone (**Extended Data Fig. S1.**). We confirmed that adaptive enrichment did not impact the random assignment of the study participants to pioglitazone versus placebo, compared to the original trial (p=1.00, **Extended Data Table S4).**

A strategy of selective enrollment conditioned on individualized estimates of pioglitazone benefit was associated with a significant reduction in the final sample size of -18.4±4.7% across all simulations (3162±180 vs 3876 participants in the original trial, $p_{\text{one-sample t-test}}$=0.008), with point estimates of greater benefit for pioglitazone versus placebo on the primary outcome (hazard ratio estimates of 0.71±0.01 vs 0.76 in the original trial) which retained statistical significance at the end of the study across all simulations (p-values of 0.003±0.001, all <0.025) (**Fig. 3**).

To assess the sensitivity of our approach to spurious confounding in the data, we repeated the analysis after randomly shuffling the baseline covariates of the study population. In this analysis, there was a similar average treatment effect (HR 0.76 for pioglitazone versus placebo, 95% CI 0.62-0.93, p=0.007), but given the simulated random association of outcomes within each study arm with the baseline covariates, we aimed to eliminate potential heterogeneous treatment effects. As opposed to our adaptive analysis, we observed that our approach did not recommend predictive enrichment in IRIS (**Fig. 3**).

SPRINT: In SPRINT, phenomapping performed at the pre-specified timepoints highlighted several features, such as female sex, a history of chronic kidney disease, anginal symptoms, and left ventricular hypertrophy as pre-randomization features predictive of treatment effect heterogeneity (**Extended Data Fig. S2.**). The random assignment of the study participants to the intensive versus standard SBP reduction arm remained balanced across all adaptive simulations and the original trial (p=0.26, **Extended Data Table S5).**

A strategy of selective enrollment conditioned on individualized estimates of benefit from intensive versus standard SBP control was associated with a significant reduction in the final sample size of -27.4 ± 3.4% across all simulations (6793±317 vs 9361 participants in the original trial, $p_{\text{one-sample t-test}}$=0.002), with point estimates consistently demonstrating greater benefit from the intensive vs standard SBP reduction on the primary outcome (point hazard ratio estimates of 0.72 ± 0.01 vs 0.75 in the original trial) which retained statistical significance at the end of the study across all simulations (p-values of 0.001±0.001, all <0.025) (**Fig. 4**). In contrast, in a sensitivity analysis where baseline covariates were randomly shuffled within each treatment arm, we observed that our algorithm did not result in enrichment at the end of the trial, with no significant decrease in the final study size or primary effect estimates (p=0.07 and 0.74, respectively) (**Fig. 4**)

Final population demographics

The distribution of the baseline demographics at the end of each simulation is summarized in **Extended Data Tables S4 and S5**. Across five adaptive simulations in each trial, no key demographic population (i.e. men or women, or any specific racial/ethnic group) were excluded from the final analysis, and only differences observed were in risk profiles that were predictively enriched.

Secondary (safety) outcomes

To ensure that predictive enrichment based on the projected benefit for the primary outcome is not offset by an increase in risk, we performed longitudinal tracking at each interim analysis timepoint by defining a hierarchy of key outcomes, i.e. all-cause mortality, followed by non-fatal MACE components, and then, for IRIS, hospitalization events, heart failure events, and bone fractures, and for SPRINT serious adverse events, analyzed based on the win ratio. As shown in **Fig. 5**, predictive enrichment was not associated with a significant change in the relative hazard of a hierarchical safety endpoint between the intervention and control arms, compared to the original trials.

Absolute versus relative risk enrichment

To explore whether adaptive enrichment was associated with preferential enrollment of individuals at higher risk of the primary outcome, we performed subgroup analyses stratified based on the simulation strategy and the period during which a patient was enrolled. Except simulations after the last interim analysis timepoint in SPRINT, we observed no significant difference in the primary outcome event rates between the original and adaptive simulations (**Extended Data Fig. S3 & S4**). This finding suggests that predictive enrichment focuses on the effect of therapy as opposed to the underlying risk of participants.

DISCUSSION

We present an ML-driven algorithm for adaptive predictive enrichment in RCTs that relies on phenome-wide, computational trial maps to define individualized signatures of treatment benefit. Using participant-level data from the IRIS and SPRINT trials, we find that these signatures can adaptively modify the enrollment strategy in two trials without selectively excluding any demographic, or clinical group. We demonstrate across multiple simulations that this approach adaptively enriched

for individuals most likely to benefit from the intervention and reduced the required sample size by 18 to 28% while preserving the overall efficacy and safety signal of the intervention. These findings propose a new paradigm to maximize the efficiency and safety of clinical trials through dynamic data-driven inference.

The importance of maximizing the efficiency of clinical trial design through innovative methods has been recently embraced by numerous societies and agencies, including the FDA,[11] and the European Medicines Agency (EMA).[21] Among several innovations in this space, adaptive trials and trial enrichment strategies have attracted significant attention. Adaptive trials describe a study design that permits flexibility in various aspects of the trial process, from sample size to enrollment criteria and treatment schedules, based on accumulating trial data.[13] The tenet of this approach is that trial adaptation can minimize the size and costs of a study, as well as the potential risk to trial participants. Enrichment strategies, which are often integrated into adaptive trial designs, steer a trial toward a patient population that is most likely to respond to a given treatment.[13,22,23] Use of these methods has been proposed across various conditions, including heart failure with preserved ejection fraction,[24] neurodegenerative & psychiatric conditions,[25,26] or kidney disease.[27]

To improve the precision of RCTs, prior studies have implemented various approaches. For instance, risk-based *prognostic enrichment* relies on the selective recruitment of individuals at high risk of a given condition, thus increasing the statistical power of a study for a given sample size and the chance of demonstrating a high absolute treatment effect. This can be achieved through risk algorithms,[23,25,28,29] and imaging or circulating biomarkers with prognostic value.[22,30,31] A representative example in the cardiovascular field is the use of coronary artery calcium scoring to enrich for individuals at high risk of adverse cardiovascular outcomes.[32] This approach, however, may affect the generalizability of a trial's findings and does not evaluate whether treatment allocation to specific patient phenotypes aligns with those most likely to benefit. To address this issue, *predictive enrichment* focuses on the individualized effects of the intervention within the context of a patient's unique phenotypic profile.[33] This can be achieved by defining biomarkers or parameters that describe mechanistic pathways through which an intervention exerts its beneficial effects,[34] such as through molecular or proteomic profiling.[14,35] Yet these mechanisms are not always known *a priori*, and molecular analyses are costly, and therefore ineligible for use in large studies. More recently, N-of-1 trials, which involve perioding switching between treatment arms such that each individual functions as their own control,[36] have been proposed as a potential mechanism to personalize treatment effects. Despite their promise, this trial design is not applicable to large, phase III trials powered against hard clinical endpoints, such as mortality. Similarly, response adaptive randomization,[37] or sequential multiphase adaptive randomized trials (SMART), which allow patients who do not respond to an initial course of treatment to be re-randomized to a separate arm,[38] interfere with the random assignment of treatment, a hallmark of RCTs.

In this context, our method bridges an adaptive trial design with predictive enrichment through machine learning-driven insights into phenotypes associated with individualized treatment effects. First, our method relies on an algorithm that can be defined *a priori* and remains independent of investigator input or biases for the duration of the study. Second, it does not require prior knowledge on potential phenotypic determinants of heterogeneous treatment effects but rather learns those in real-time as data accumulate in each study. Third, it operates in a stochastic fashion and does not prevent any specific population subgroups from enrolling in the trial, but rather adaptively conditions the probability of enrollment on the predicted treatment benefit. This ensures that the results remain broadly applicable to the original patient population for which the trial was designed. Of note, the algorithm may be further customized to ensure that traditionally under-represented groups are not excluded from an ongoing trial. Fourth, by modeling the relative treatment effect, our algorithm provides predictive enrichment without restricting enrollment to the individuals most likely

to experience the primary endpoint. Fifth, in simulations where there is an absence of heterogeneity in the treatment effect, the model appropriately does not recommend predictive enrichment. Finally, it respects the random treatment assignment throughout the trial since the predictive algorithm is trained on information collected before randomization. As a result, the treatment assignment is independent of the baseline characteristics that define a participant's probability of benefit and by extent, probability of being enrolled in the trial.

Our analysis carries certain limitations that merit consideration. First, our work here represents a post hoc analysis of a real-world RCTs. However, IRIS was chosen among other trials, given that it illustrates frequent challenges faced in cardiovascular outcome trials, including but not limited to slow enrollment (~7-8 years), need for a large study group (~almost 4000 patients) and long prospective follow-up (~5 years). Similarly, SPRINT models a large cardiovascular trial of a strategy, rather than a specific medication. Second, even though the FDA has embraced the need for innovative technologies in clinical trial design and interpretation, regulatory concerns remain, such as defining the product label for a therapy that is based on an adaptive trial design. It should be noted, however, that our proposed design does not explicitly incorporate new exclusion or inclusion criteria but rather modifies the probability that a patient fulfilling all original inclusion criteria will be enrolled at the later stages of the trial. This needs to be done with caution to ensure equity while maximizing the benefit-risk ratio for all potential patient phenotypes. Notably, post hoc analyses of completed trials have demonstrated variation across sites in the characteristics as well as outcomes of the enrolled populations, highlighting existing variations despite specific protocols and enrollment criteria.[39] Third, a predictively enriched, accelerated study design could hinder our ability to identify safety signals for rare events or derive inference from traditional subgroup analyses, including populations with traditionally low rates of events. Finally, although the algorithm aims to introduce explainability to its predictions through SHAP, reliance on broad phenotypic features is often a surrogate for underlying biological, functional, or anatomical differences. Therefore, the association between the identified baseline features and the effect of the studied treatment is not necessarily causal. Nonetheless, our algorithm remains versatile and can theoretically incorporate additional high-dimensional patient features, such as genetic, genomic, or imaging biomarkers, that may be collected as part of a trial.

CONCLUSIONS

We hereby describe and implement an ML-guided algorithm for adaptive, predictive enrichment of RCTs based on individualized signatures of treatment benefit derived from computational trial phenomaps. In a post hoc analysis of two large cardiovascular outcome trials powered for clinical endpoints as primary outcomes, our proposed strategy of predictive enrichment based on ML-derived insights estimates that it is possible to achieve a consistent and robust reduction in the required sample size while conserving the study's power to detect significant average treatment effects. This is achieved through real-time predictive enrichment which is independent of a patient's baseline absolute risk and modifies the trial's baseline phenotypic composition in a standardized way, thus ensuring a trial's efficiency, safety, power, and generalizability.

METHODS

1. Original trial design

IRIS trial: The Insulin Resistance Intervention after Stroke (IRIS) trial recruited patients at least 40 years of age with a recent ischemic stroke or a transient ischemic attack (TIA) during the 6 months prior to randomization, who did not have diabetes mellitus at the time of enrollment but had evidence of insulin resistance based on a homeostasis model assessment of insulin resistance (HOMA-IR) index score of 3.0 or greater. Participants were randomly assigned in a 1:1 ratio to receive either pioglitazone or matching

placebo (with dose up-titration as specified in the original trial report).[18] Patients were contacted every 4 months, and participation ended at 5 years or at the last scheduled contact before July 2015.

SPRINT trial: Systolic Blood Pressure Intervention Trial (SPRINT) enrolled 9361 participants, 50 years of age or older, with a systolic blood pressure (SBP) of 130-180 mm Hg with or without antihypertensive drug treatment as well as an additional indicator of cardiovascular risk. These included clinical or subclinical cardiovascular disease, chronic kidney disease, 10-year risk Framingham Risk Score of cardiovascular disease of 15% or higher or age of 75 years or older. Patients with diabetes mellitus, prior stroke, or dementia were excluded from this trial. Participants were enrolled between 2010-2013 at 102 clinical sites in the U.S.[19]

2. Study characteristics and outcomes

In accordance with the primary outcome of the original trials, we focused on a composite of first fatal or nonfatal stroke or fatal or nonfatal myocardial infarction as the primary outcome for IRIS, and a composite of myocardial infarction, acute coronary syndrome not resulting in myocardial infarction, stroke, acute decompensated heart failure, or death from cardiovascular causes for SPRINT. Definitions were concordant with those used in the original trial reports.[18,19] All outcomes and selected safety events were adjudicated by the members of independent committees in a blinded fashion for each of the trials.

3. Design of a group sequential, adaptive trial experiment

We designed a simulation algorithm to test the hypothesis that interim ML-guided analyses of computational trial phenomaps can adaptively guide the trials' enrollment process and maximize their efficiency while reducing their final/required size. The tenet of this approach is that ML photomapping-derived insights can steer the recruitment towards patients who are more likely to benefit from the intervention. For this, we defined three interim analysis timepoints, with the final analysis occurring once all primary events had been reported. It should be noted that the original power calculation for IRIS had assumed higher event and faster enrollment rates than the ones that were observed during the course of the trial, thus prompting serial amendments to the trial protocol, including an extension of recruitment and an increase in the study size (from 3136 patients initially to 3936 patients). In a post-hoc fashion, knowing that the primary outcome occurred in 228 of 1937 participants in the placebo arm (~11.8% rate) and 175 of 1939 participants in the pioglitazone arm (~9.0%), we simulated power calculations in a *post-hoc* manner, this time assuming a superiority trial design with a one-sided α of 0.025 (see the "*power calculations*" section in the *Methods*). We defined the timepoint at which 50, 100 and 150 total primary outcome events had been recorded in the original trial as the timepoint for our first, second and third interim analysis timepoints, respectively. In SPRINT we assumed that the respective primary outcome would occur in 6.8% vs 5.4% of the standard and intensive arms, and for consistency, defined the interim analysis timepoints based on the occurrence of the first 50, 100 and 150 primary outcome events.

4. Overview of the predictive enrichment approach

During the first enrollment period of the simulation (study onset until first interim analysis) we enrolled all trial participants, similar to the original trials, without any restrictions or modifications in the enrollment process. Beginning at the first interim analysis timepoint, participants recruited up until that stage were randomly split into training/cross-validation (50%) and testing sets (50%). In the training set, baseline data were pre-processed and used to define a phenomap (see "*Machine learning trial phenomapping*" below), which represented the phenotypic architecture of the population across all axes of baseline variance. Through iterative analyses centered around each unique individual and weighted for each individual participant's location in the phenotypic space, we defined individualized estimates of the effects of the studied intervention, as compared to the control arm, for the primary outcome.

Subsequently, we built a ML framework to identify key features that collectively determined a phenotypic signature (algorithm) predictive of these individualized estimates (explained in **Methods 5;**

Machine learning trial phenomapping). The algorithm was then tested in the testing set, assessing for evidence of potential heterogeneous treatment effects by dichotomizing the population into two groups based on their predicted response. To avoid imbalanced groups or identifying extreme outliers of responders or non-responders, the smallest subgroup size was set at 20%. We then analyzed the presence of heterogeneity in the observed effect estimates between the two subgroups in the testing set by calculating the *p* value for interaction of treatment effect. Given that testing was done using just a half of the observations collected at each interim analysis timepoint, we defined a threshold of $p_{\text{interaction}}<0.2$ as our criterion for possible presence of heterogeneity.

If there was potential evidence of heterogeneity based on this analysis, sample size calculations were updated at that interim analysis timepoint by revising the expected effect size (under the assumption of predictive enrichment) at the original power and alpha levels (0.8 and 0.025, respectively, in both trials). This was done to assess whether prospective predictive enrichment and the associated decrease in the projected number of recruited individuals would provide sufficient power at a sample size equal to or smaller than the originally planned one. We performed sample size calculations assuming various levels of predictive enrichment, which ranged from enrolling 50% to 95% of all remaining candidates, in 5% increments (see *Methods 5e* below). If there were several levels that met these criteria, we ultimately chose the predictive enrichment level that minimized the required sample.

Assuming the above, over the subsequent period, the probability of enrollment was conditioned on the anticipated benefit, assessed by applying the most recent model to each potential candidate's baseline characteristics. Alternatively, if there was no evidence of heterogeneous treatment effect, or the proposed enrichment in enrollment would not be adequately powered at the revised sample size, we proceeded as originally planned and continued with standard enrollment for that time-period without predictive enrichment. This process was repeated at each interim analysis timepoint. There was no assessment for futility.

Given the stochastic nature of the algorithm, the simulation was repeated $r=5$ times. For reference, we present the observed outcomes of the full trial population at the same timepoints. To enable direct comparison between the different simulations, the final analysis was performed at the timepoint at which all primary outcome events had occurred in the original trial population.

5. Machine learning trial phenomapping

5a - Data pre-processing: Our analysis included 62 phenotypic features recorded at baseline in IRIS (**Extended Data Table S1**), and 82 baseline features in SPRINT (**Extended Data Table S2**), as per our prior work.[17] At every point, pre-processing steps, including imputation, were performed independently for each patient subset to avoid data leakage. Baseline features with greater than 10% missingness are removed from further analysis. To avoid collinearity of continuous variables, we calculate pairwise correlations across variables, and wherever pairs exceed an absolute correlation coefficient of 0.9, we exclude the variable with the largest mean absolute correlation across all pairwise comparisons. Continuous variables also undergo 95% winsorization to reduce the effects of extreme outliers, whereas factor variables with zero variance are dropped from further processing. Next, we impute missing data using a version of the random forest imputation algorithm adapted for mixed datasets with a maximum of five iterations. Factor variables undergo one-hot encoding for ease of processing with downstream visualization and machine learning algorithms.

5b - Creating a computational trial phenomap: Once the dataset for a given simulation at a specified time-point has been created, we compute a dissimilarity index that classifies individuals based on their detailed clinical characteristics according to Gower's distance. Gower's method computes a distance value for each pair of individuals. For continuous variables, Gower's distance represents the absolute value of the difference between a pair of individuals divided by the range across all individuals. For categorical variables, the method assigns "1" if the values are identical and "0" if they are not. Gower's distance is ultimately calculated as the mean of these terms.[40] At this point, the phenotypic architecture of the trial can be

visualized using uniform manifold approximation and projection (UMAP),[41] a method that constructs a high-dimensional graph and then optimizes a low-dimensional graph to be as structurally similar as possible. UMAP aims to maintain a balance between the local and global structure of the data by decreasing the likelihood of connection as the outwards radius around each data point increases, thus maintaining the local architecture while ensuring that each point is connected to at least its closest neighbor and ensuring a global representation.[41]

5c - Defining individualized hazard estimates: To extract personalized estimates of predicted benefit with pioglitazone or intensive SBP control, versus placebo or standard SBP reduction respectively, for each individual included in each interim analysis, we applied weighted estimation in Cox regression models.[42] With every iteration of this regression around each unique individual, every study participant was assigned unique weights based on the phenotypic (Gower's) distance from the index patient of that analysis. To ensure that patients phenotypically closer to the index patient carried higher cumulative weights than patients located further away, we applied a cubic exponential transformation of the similarity metric, defined as (*1-Gower's distanc*e). These values were further processed through a Rectified Linear Unit (ReLU) function prior to their inclusion as weights in the regression models. This allowed us to simultaneously model an exponential decay function and control the impact of low values (ReLU). From each personalized Cox regression model (fitted for each unique participant with individualized weightings as above), we extracted the natural logarithmic transformation of the hazard ratio (log HR) for the primary outcome for the intervention versus control.

5d - Training a model to predict the individualized benefit based on baseline characteristics: To identify baseline features that are important in determining the personalized benefit of the studied intervention relative to control (described by the individualized log HR), an extreme gradient boosting algorithm (known as XGBoost; based on a tree gradient booster) is fitted with simultaneous feature selection based on thee Boruta and SHAP (SHapley Additive exPlanations) methods. Briefly, the Boruta method creates randomized (permuted) versions of each feature (called "shadow features") before merging the permuted and original data. Once a model is trained, the importance of all original features is compared to the highest feature importance of the shadow features. This process is repeated for n=20 iterations, without sampling. SHAP is added as an approach to explain the output of the ML model, based on concepts derived from game theory. SHAP calculates the average marginal contributions for each feature across all permutations at a local level. With the addition of SHAP analysis, the feature selection further benefits from the strong additive feature explanations but maintains the robustness of the Boruta method.[43,44] The testing data are further split into training and testing sets (with a random 80-20% split). We set our problem as a regression task using root mean squared error as our metric to evaluate our model's accuracy during testing. Before training, the labels (i.e., previously calculated individualized log HR) undergo 95% winsorization to minimize the effects of extreme outliers. First, we fit an XGBoost model using the Boruta algorithm to identify a subset of important baseline features, and then repeat this process to predict the individualized log HR, this time using only the selected features as input. Hyperparameter tuning is achieved through a grid search across 25 iterations (learning rate: [0.01, 0.05, 0.10, 0.15]; maximal depth of the tree: [3, 5, 6, 10, 15, 20]; fraction of training samples that will be used to train each tree: 0.5 to 1.0 by 0.1 increments, number of features supplied to a tree: 0.4 to 1.0 by 0.1 increments; random subsample of columns when every new level is reached: 0.4 to 1.0 by 0.1, number of gradient boosted trees: [100, 500, 1000]). We train the model for a maximum of 1000 rounds, with an early stopping function every 20 rounds once the loss in the validation set starts to increase. The importance of each feature is again visualized using a SHAP plot. SHAP values measure the impact of each variable considering the interaction with other variables. We visualised these using a SHAP summary plot, in which the vertical axis represents the variables in descending order of importance and the horizontal axis indicates the change in prediction (with wider bars along the horizontal axis associated with higher feature importance). The gradient color denotes the

original value for that variable (for instance, for binary variables such as sex, it only takes two colors, whereas for continuous variables, it contains the whole spectrum).

5e - Adaptive trial enrollment: Once a predictive model has been generated at a given interim analysis timepoint, the model is prospectively applied to all trial candidates during the following trial period (time between two interim analyses). For example, a model trained at interim analysis timepoint #1 will be applied to individuals screened between the interim analysis timepoints #1 and #2. Here, all patients that were included and enrolled in the original trial during this period are considered eligible candidates. This approach yields individualized predictions of expected cardiovascular benefit with pioglitazone versus placebo (or intensive versus standard SBP reduction), with these predictions used to condition the probability of a given patient being enrolled in the simulation. Given that the predicted individualized log HR could have both negative (favoring pioglitazone or intensive SBP reduction) and positive values (favoring placebo or standard SBP reduction), predictions were multiplied by -1, normalized to the [0, 1] range. The result (input x) was processed through a sigmoid transformation function with a scaling factor of k=10; ($\frac{1}{(1 + e^{-10(x-(1-z))})}$), where z = the ratio of the responders to non-responders, followed by squared transformation. These numbers were used as sampling weights during the subsequent period to ensure that patients with higher predicted benefit were more likely to be enrolled over the next period. The process was repeated at each interim analysis timepoint.

5f - Power calculations: We simulated a superiority trial design, assuming a power of 80% and type I error of 0.025. We present alpha level adjustments for each time point, adjusted based on the O'Brien-Fleming and alpha-spending Pocock methods (**Extended Data Table S3**). We performed our analyses using the *rpact* package in R, using the expected event rates used in our power calculations above and simulating three interim analyses with four total looks as described above. These analyses are restricted to the primary endpoint.

6. Negative control analysis

To assess the performance of our algorithm in the presence of an identical average treatment effect (ATE) but with absent (or at least randomly distributed) heterogeneous treatment effects, we randomly shuffled the baseline characteristics of each trial. This ensured that any effects of the baseline characteristics on the effectiveness of the intervention would be lost or be due to random variation.

7. Statistical analysis

Categorical variables are summarized as numbers (percentages), and continuous variables as mean ± standard deviation or median with IQR (Q1–Q3) unless specified otherwise. Continuous variables between three or more groups were compared using analysis of variance (ANOVA) or the Kruskal-Wallis test (as appropriate). In contrast, categorical variables between groups are compared by Pearson's chi-squared test. Survival analyses were performed by fitting a Cox regression model for the time-to-primary outcome using the treatment arm as an independent predictor. While estimating individualized treatment effects, each observation was weighted based on the calculated similarity metric to the index patient of each analysis. Between-subgroup analyses for heterogeneity of treatment effect were performed by computing a P value for interaction. Simulation-level counts and point estimates were compared to the respective numbers/counts from the original trial using one-sample t-tests; for the counts of final study participants, alpha was set at 0.025 given the single-sided nature of the test; otherwise, alpha was set at 0.05. We graphically summarized the counts of enrolled participants, primary outcome events, Cox regression-derived effect estimates (unadjusted, for each one of the primary and secondary outcomes), and P values at each one of the interim analysis timepoints (with error bars denoting the standard error of the mean), in addition to the final analysis timepoint. Cumulative incidence curves for the primary outcome stratified by the enrollment period and simulation analysis were graphically presented. Each one was compared to

the original trial subset for the same period using the log-rank statistic. As reviewed above, for the primary outcome, we simulated a superiority trial design, assuming a power of 80% and type I error of 0.025. Statistical tests were two-sided with a level of significance of 0.05, unless specified otherwise. Analyses were performed using Python (version 3.9) and R (version 4.2.3). Reporting of the study design and findings stands consistent with the STROBE guidelines.[45]

DATA AVAILABILITY

The SPRINT data are available through the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) of the National Heart, Lung, and Blood Institute (NHLBI). The IRIS data were made available through communication with the original study investigators.

INSTITUTIONAL REVIEW BOARDS REVIEW

The local ethics committee/IRB (institutional review boards) of Yale University provided a determination of exemption and waived ethical approval for this work (under IRB protocol ID 2000029730, 1/21/2021).

REFERENCES

1. Collins, R., Bowman, L., Landray, M. & Peto, R. The Magic of Randomization versus the Myth of Real-World Evidence. *N. Engl. J. Med.* **382**, 674–678 (2020).
2. Bothwell, L. E. & Podolsky, S. H. The Emergence of the Randomized, Controlled Trial. *N. Engl. J. Med.* **375**, 501–504 (2016).
3. Fogel, D. B. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemp Clin Trials Commun* **11**, 156–164 (2018).
4. Bentley, C. *et al.* Conducting clinical trials-costs, impacts, and the value of clinical trials networks: A scoping review. *Clin. Trials* **16**, 183–193 (2019).
5. Moore, T. J., Zhang, H., Anderson, G. & Alexander, G. C. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. *JAMA Intern. Med.* **178**, 1451–1457 (2018).
6. Moore, T. J., Heyward, J., Anderson, G. & Alexander, G. C. Variation in the estimated costs of pivotal clinical benefit trials supporting the US approval of new therapeutic agents, 2015-2017: a cross-sectional study. *BMJ Open* **10**, e038863 (2020).
7. Janiaud, P., Serghiou, S. & Ioannidis, J. P. A. New clinical trial designs in the era of precision medicine: An overview of definitions, strengths, weaknesses, and current use in oncology. *Cancer Treat. Rev.* **73**, 20–30 (2019).
8. Park, J. J. H. *et al.* Randomised trials at the level of the individual. *Lancet Glob Health* **9**, e691–e700 (2021).
9. Mehta, C. *et al.* Optimizing trial design: sequential, adaptive, and enrichment strategies. *Circulation* **119**, 597–605 (2009).
10. Bhatt, D. L. & Mehta, C. Adaptive Designs for Clinical Trials. *N. Engl. J. Med.* **375**, 65–74 (2016).
11. Center for Drug Evaluation & Research. Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry. *U.S. Food and Drug Administration* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.
12. Pallmann, P. *et al.* Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med.* **16**, 29 (2018).
13. Simon, N. & Simon, R. Adaptive enrichment designs for clinical trials. *Biostatistics* **14**, 613–625 (2013).
14. Fountzilas, E., Tsimberidou, A. M., Vo, H. H. & Kurzrock, R. Clinical trial design in the era of precision medicine. *Genome Med.* **14**, 101 (2022).
15. Oikonomou, E. K. *et al.* A phenomapping-derived tool to personalize the selection of anatomical vs. functional testing in evaluating chest pain (ASSIST). *Eur. Heart J.* **42**, 2536–2548 (2021).
16. Oikonomou, E. K., Suchard, M. A., McGuire, D. K. & Khera, R. Phenomapping-Derived Tool to Individualize the Effect of Canagliflozin on Cardiovascular Risk in Type 2 Diabetes. *Diabetes Care* **45**, 965–974 (2022).
17. Oikonomou, E. K., Spatz, E. S., Suchard, M. A. & 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* **4**, e796–e805 (2022).
18. Kernan, W. N. *et al.* Pioglitazone after Ischemic Stroke or Transient Ischemic Attack. *N. Engl. J. Med.* **374**, 1321–1331 (2016).
19. SPRINT Research Group *et al.* A Randomized Trial of Intensive versus Standard Blood-Pressure Control. *N. Engl. J. Med.* **373**, 2103–2116 (2015).
20. Meurer, W. J. & Tolles, J. Interim Analyses During Group Sequential Clinical Trials. *JAMA: the journal of the American Medical Association* vol. 326 1524–1525 (2021).

21. Collignon, O. *et al.* Adaptive designs in clinical trials: From scientific advice to marketing authorisation to the European Medicine Agency. *Trials* **19**, (2018).
22. Kerr, K. F. *et al.* Evaluating biomarkers for prognostic enrichment of clinical trials. *Clin. Trials* **14**, 629–638 (2017).
23. Jering, K. S. *et al.* Improving clinical trial efficiency using a machine learning-based risk score to enrich study populations. *Eur. J. Heart Fail.* **24**, 1418–1426 (2022).
24. Shah, S. J. Innovative Clinical Trial Designs for Precision Medicine in Heart Failure with Preserved Ejection Fraction. *J. Cardiovasc. Transl. Res.* **10**, 322–336 (2017).
25. Podichetty, J. T. *et al.* Application of machine learning to predict reduction in total PANSS score and enrich enrollment in schizophrenia clinical trials. *Clin. Transl. Sci.* **14**, 1864–1874 (2021).
26. Ezzati, A. *et al.* Application of predictive models in boosting power of Alzheimer's disease clinical trials: A post hoc analysis of phase 3 solanezumab trials. *Alzheimers. Dement.* **8**, e12223 (2022).
27. Lazzareschi, D. *et al.* Overcoming barriers in the design and implementation of clinical trials for Acute Kidney Injury: a report from the 2020 Kidney Disease Clinical Trialists meeting. *Nephrol. Dial. Transplant* (2022) doi:10.1093/ndt/gfac003.
28. Scott, J. V., Garnett, C. E., Kanwar, M. K., Stockbridge, N. L. & Benza, R. L. Enrichment Benefits of Risk Algorithms for Pulmonary Arterial Hypertension Clinical Trials. *Am. J. Respir. Crit. Care Med.* **203**, 726–736 (2021).
29. Tam, A., Laurent, C., Gauthier, S. & Dansereau, C. Prediction of Cognitive Decline for Enrichment of Alzheimer's Disease Clinical Trials. *J Prev Alzheimers Dis* **9**, 400–409 (2022).
30. Lin, Y., Shih, W. J. & Lu, S.-E. Two-stage enrichment clinical trial design with adjustment for misclassification in predictive biomarkers. *Stat. Med.* **38**, 5445–5469 (2019).
31. Renfro, L. A., Mallick, H., An, M.-W., Sargent, D. J. & Mandrekar, S. J. Clinical trial designs incorporating predictive biomarkers. *Cancer Treat. Rev.* **43**, 74–82 (2016).
32. Cainzos-Achirica, M. *et al.* Rationale and pathways forward in the implementation of coronary artery calcium-based enrichment of randomized trials. *Am. Heart J.* **243**, 54–65 (2022).
33. Li, J. *et al.* A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies. *Biometrics* **72**, 877–887 (2016).
34. Kent, D. M. *et al.* Heterogeneity of Treatment Effects in an Analysis of Pooled Individual Patient Data From Randomized Trials of Device Closure of Patent Foramen Ovale After Stroke. *JAMA* **326**, 2277–2286 (2021).
35. Chan, M. Y. *et al.* Prioritizing Candidates of Post-Myocardial Infarction Heart Failure Using Plasma Proteomics and Single-Cell Transcriptomics. *Circulation* **142**, 1408–1421 (2020).
36. Lillie, E. O. *et al.* The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Per. Med.* **8**, 161–173 (2011).
37. Wang, Y., Carter, B. Z., Li, Z. & Huang, X. Application of machine learning methods in clinical trials for precision medicine. *JAMIA Open* **5**, ooab107 (2022).
38. Collins, L. M., Murphy, S. A. & Strecher, V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *Am. J. Prev. Med.* **32**, S112-8 (2007).
39. Greene, S. J. *et al.* Influence of Clinical Trial Site Enrollment on Patient Characteristics, Protocol Completion, and End Points: Insights From the ASCEND-HF Trial (Acute Study of Clinical Effectiveness of Nesiritide in Decompensated Heart Failure). *Circ. Heart Fail.* **9**, (2016).
40. Gower, J. C. A general coefficient of similarity and some of its properties. *Biometrics* 857–871 (1971).
41. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 (2018).

42. Schemper, M., Wakounig, S. & Heinze, G. The estimation of average hazard ratios by weighted Cox regression. *Stat Med* **28**, 2473–2489 (2009).
43. Chieregato, M. *et al.* A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. *Sci. Rep.* **12**, 4329 (2022).
44. *Boruta-Shap: A Tree based feature selection tool which combines both the Boruta feature selection algorithm with shapley values*. (Github).
45. von Elm, E. *et al.* Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* **335**, 806–808 (2007).

ACKNOWLEDGEMENTS
The authors would like to thank Walter N. Kernan (Department of Internal Medicine, Yale School of Medicine) for facilitating access to the IRIS data.

DECLARATIONS
E.K.O and R.K. are co-inventors of the U.S. Patent Applications 63/508,315 & 63/177,117 and co-founders of Evidence2Health, a health analytics company to improve evidence-based cardiovascular care. E.K.O. reports a consultancy and stock option agreement with Caristo Diagnostics Ltd (Oxford, U.K.), unrelated to the current work. R.K. received support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award K23HL153775) and the Doris Duke Charitable Foundation (under award, 2022060). R.K. further receives research support, through Yale, from Bristol-Myers Squibb. He is also a coinventor of U.S. Pending Patent Applications 63/428,569 and 63/346,610, unrelated to the current work. D.L.B. discloses the following relationships - Advisory Board: Angiowave, Bayer, Boehringer Ingelheim, Cardax, CellProthera, Cereno Scientific, Elsevier Practice Update Cardiology, High Enroll, Janssen, Level Ex, McKinsey, Medscape Cardiology, Merck, MyoKardia, NirvaMed, Novo Nordisk, PhaseBio, PLx Pharma, Regado Biosciences, Stasys; Board of Directors: Angiowave (stock options), Boston VA Research Institute, Bristol Myers Squibb (stock), DRS.LINQ (stock options), High Enroll (stock), Society of Cardiovascular Patient Care, TobeSoft; Chair: Inaugural Chair, American Heart Association Quality Oversight Committee; Consultant: Broadview Ventures, Hims; Data Monitoring Committees: Acesion Pharma, Assistance Publique-Hôpitaux de Paris, Baim Institute for Clinical Research (formerly Harvard Clinical Research Institute, for the PORTICO trial, funded by St. Jude Medical, now Abbott), Boston Scientific (Chair, PEITHO trial), Cleveland Clinic (including for the ExCEED trial, funded by Edwards), Contego Medical (Chair, PERFORMANCE 2), Duke Clinical Research Institute, Mayo Clinic, Mount Sinai School of Medicine (for the ENVISAGE trial, funded by Daiichi Sankyo; for the ABILITY-DM trial, funded by Concept Medical), Novartis, Population Health Research Institute; Rutgers University (for the NIH-funded MINT Trial); Honoraria: American College of Cardiology (Senior Associate Editor, Clinical Trials and News, ACC.org; Chair, ACC Accreditation Oversight Committee), Arnold and Porter law firm (work related to Sanofi/Bristol-Myers Squibb clopidogrel litigation), Baim Institute for Clinical Research (formerly Harvard Clinical Research Institute; RE-DUAL PCI clinical trial steering committee funded by Boehringer Ingelheim; AEGIS-II executive committee funded by CSL Behring), Belvoir Publications (Editor in Chief, Harvard Heart Letter), Canadian Medical and Surgical Knowledge Translation Research Group (clinical trial steering committees), CSL Behring (AHA lecture), Cowen and Company, Duke Clinical Research Institute (clinical trial steering committees, including for the PRONOUNCE trial, funded by Ferring Pharmaceuticals), HMP Global (Editor in Chief, Journal of Invasive Cardiology), Journal of the American College of Cardiology (Guest Editor; Associate Editor), K2P (Co-Chair, interdisciplinary curriculum), Level Ex, Medtelligence/ReachMD (CME steering committees), MJH Life Sciences, Oakstone CME (Course Director, Comprehensive Review of Interventional Cardiology), Piper Sandler, Population Health Research Institute (for the COMPASS operations committee, publications committee, steering committee, and USA national co-leader, funded by Bayer), Slack Publications (Chief Medical Editor, Cardiology Today's Intervention), Society of Cardiovascular Patient Care (Secretary/Treasurer), WebMD (CME steering committees), Wiley (steering committee); Other: Clinical Cardiology (Deputy Editor), NCDR-ACTION Registry Steering Committee (Chair), VA CART Research and Publications Committee (Chair); Patent: Sotagliflozin (named on a patent for sotagliflozin assigned to Brigham and Women's Hospital who assigned to Lexicon; neither I nor Brigham and Women's Hospital receive any income from this patent); Research Funding: Abbott, Acesion Pharma, Afimmune, Aker Biomarine, Alnylam, Amarin, Amgen, AstraZeneca, Bayer, Beren, Boehringer Ingelheim, Boston Scientific, Bristol-Myers Squibb, Cardax,

CellProthera, Cereno Scientific, Chiesi, CinCor, Cleerly, CSL Behring, Eisai, Ethicon, Faraday Pharmaceuticals, Ferring Pharmaceuticals, Forest Laboratories, Fractyl, Garmin, HLS Therapeutics, Idorsia, Ironwood, Ischemix, Janssen, Javelin, Lexicon, Lilly, Medtronic, Merck, Moderna, MyoKardia, NirvaMed, Novartis, Novo Nordisk, Otsuka, Owkin, Pfizer, PhaseBio, PLx Pharma, Recardio, Regeneron, Reid Hoffman Foundation, Roche, Sanofi, Stasys, Synaptic, The Medicines Company, Youngene, 89Bio; Royalties: Elsevier (Editor, Braunwald's Heart Disease); Site Co-Investigator: Abbott, Biotronik, Boston Scientific, CSI, Endotronix, St. Jude Medical (now Abbott), Philips, SpectraWAVE, Svelte, Vascular Solutions; Trustee: American College of Cardiology; Unfunded Research: FlowCo, Takeda. H.M.K. works under contract with the Centers for Medicare & Medicaid Services to support quality measurement programs, was a recipient of a research grant from Johnson & Johnson, through Yale University, to support clinical trial data sharing; was a recipient of a research agreement, through Yale University, from the Shenzhen Center for Health Information for work to advance intelligent disease prevention and health promotion; collaborates with the National Center for Cardiovascular Diseases in Beijing; receives payment from the Arnold & Porter Law Firm for work related to the Sanofi clopidogrel litigation, from the Martin Baughman Law Firm for work related to the Cook Celect IVC filter litigation, and from the Siegfried and Jensen Law Firm for work related to Vioxx litigation; chairs a Cardiac Scientific Advisory Board for UnitedHealth; was a member of the IBM Watson Health Life Sciences Board; is a member of the Advisory Board for Element Science, the Advisory Board for Facebook, and the Physician Advisory Board for Aetna; and is the co-founder of Hugo Health, a personal health information platform, and co-founder of Refactor Health, a healthcare AI-augmented data management company. M.A.S. reports institutional grant support from the US National Institutes of Health, US Food and Drug Administration, and US Department of Veteran Affairs; personal consulting fees from Janssen Research and Development and Private Health Management; and institutional grant support from Advanced Micro Devices, outside the scope of the submitted work. All other authors declare no competing interests.

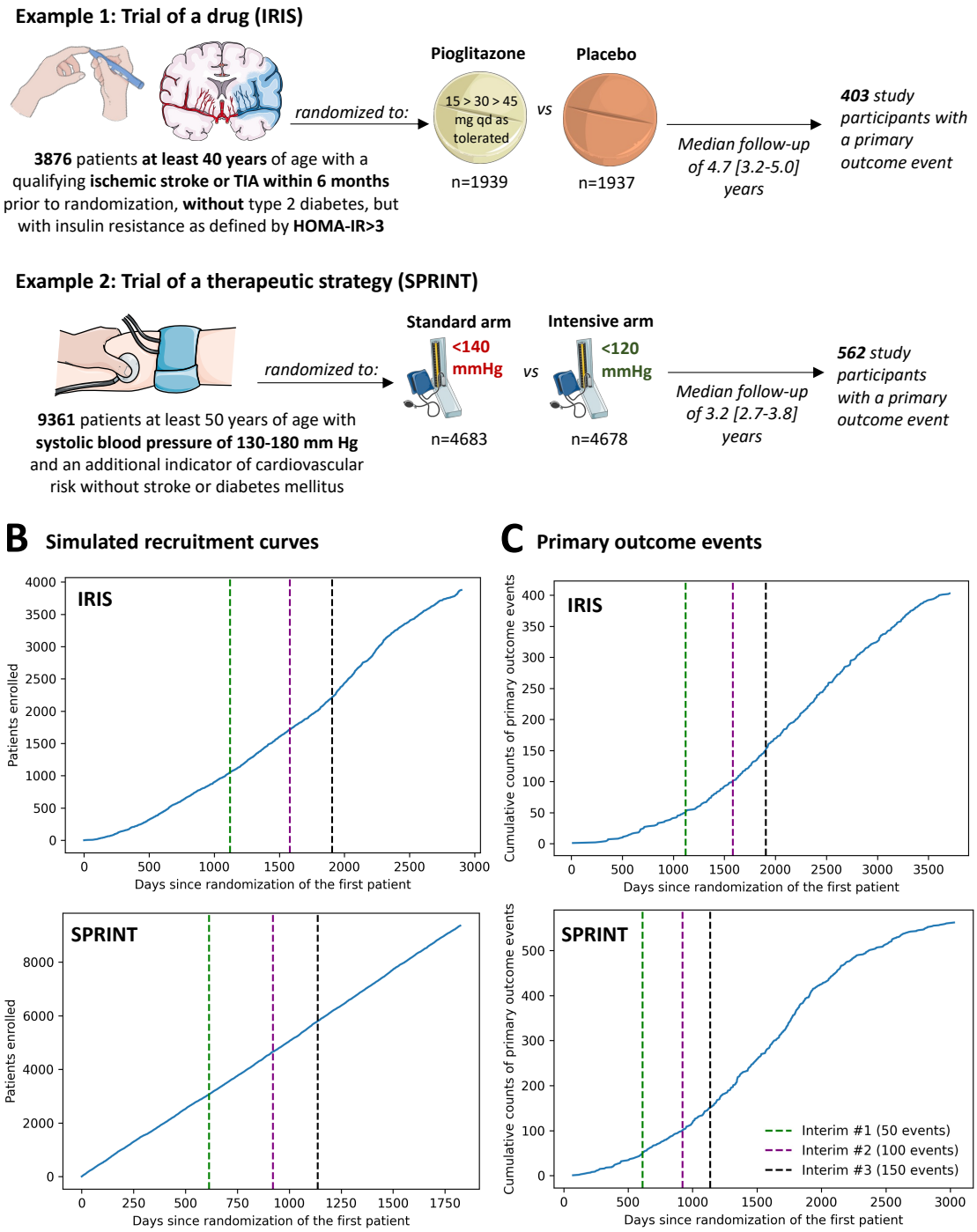

Fig. 1 | Summary of the study design. (A) Visual summary of the original IRIS and SPRINT trials. **(B)** Cumulative recruitment & **(C)** primary outcome event numbers across different simulated trial timepoints. HOMA-IR: homeostasis model assessment of insulin resistance; IRIS: Insulin Resistance Intervention after Stroke; SPRINT: Systolic Blood Pressure Intervention Trial; TIA: transient ischemic attack.

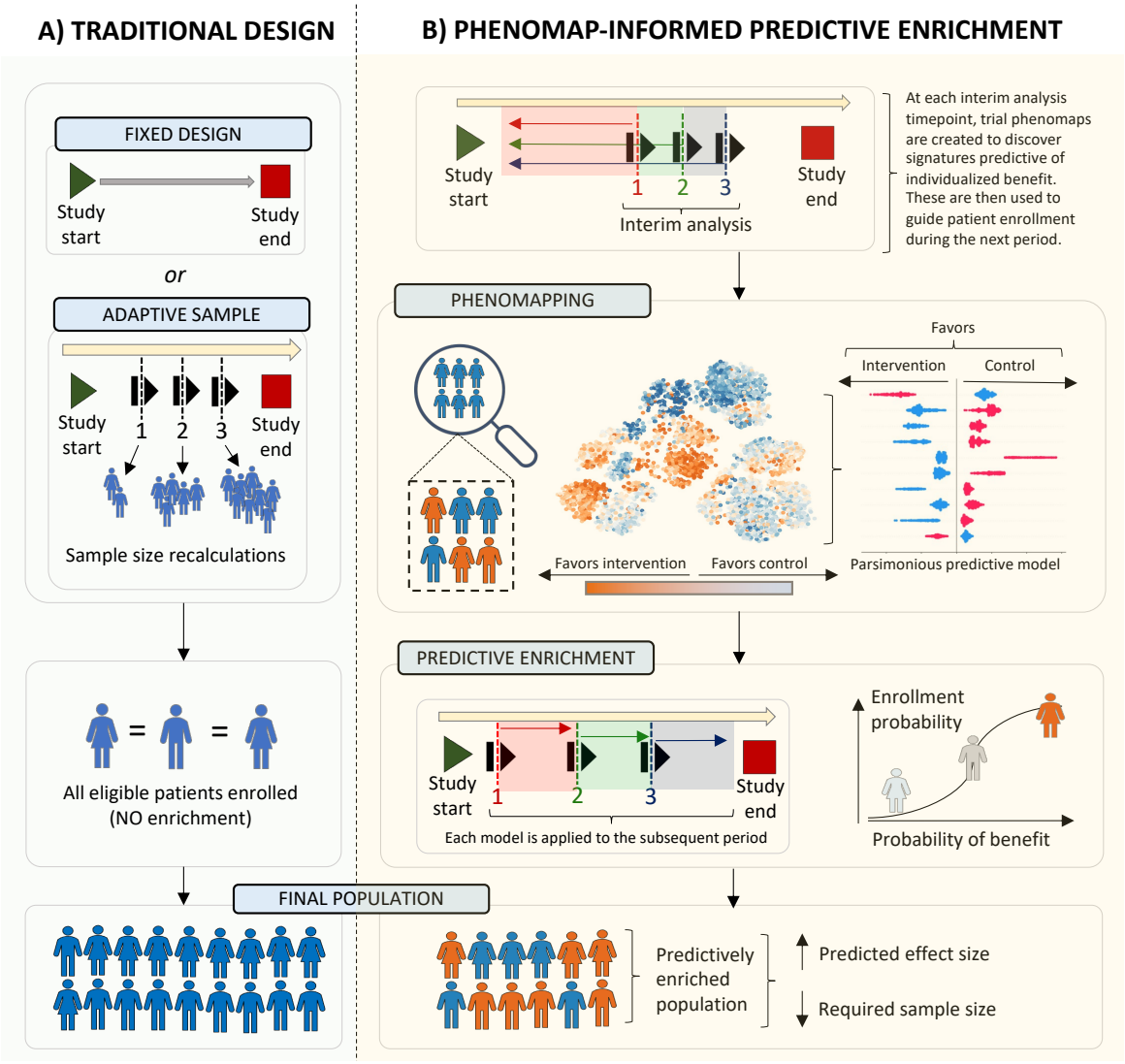

Fig. 2 | Overview of the algorithm. (A) Traditional clinical trials commonly use a fixed design approach, with a pre-defined total sample size goal and inclusion criteria. In some cases, interim analyses are pre-defined to adaptively modify the target study size based on revised power calculations, however without modification of the target population. **(B)** We hereby propose an approach of adaptive predictive enrichment through phenomapping-derived signatures of individualized benefit. At each pre-defined interim analysis timepoint, the observations and events collected up until that point are randomly split into a training and testing set. In the training set, a trial phenomap is created that represents a representation of the phenotypic similarities across all recorded baseline features. This allows the estimation of weighted average treatment effects by analyzing the observed outcomes from the phenotypic angle of each individual participant. This is followed by training of an extreme gradient boosting algorithm that links pre-randomization features to the observe treatment effect heterogeneity. If there is evidence of possible heterogeneity, power calculations are revised based on the observed effect sizes and, assuming the required final sample is lower than the originally estimated one, predictive enrichment occurs over the following period. During this time, the probability of enrollment for each prospective candidate is conditioned on their estimated benefit, whereas treatment assignment remains completely randomized. The process is repeated at each interim analysis timepoint, where a completely new model is trained.

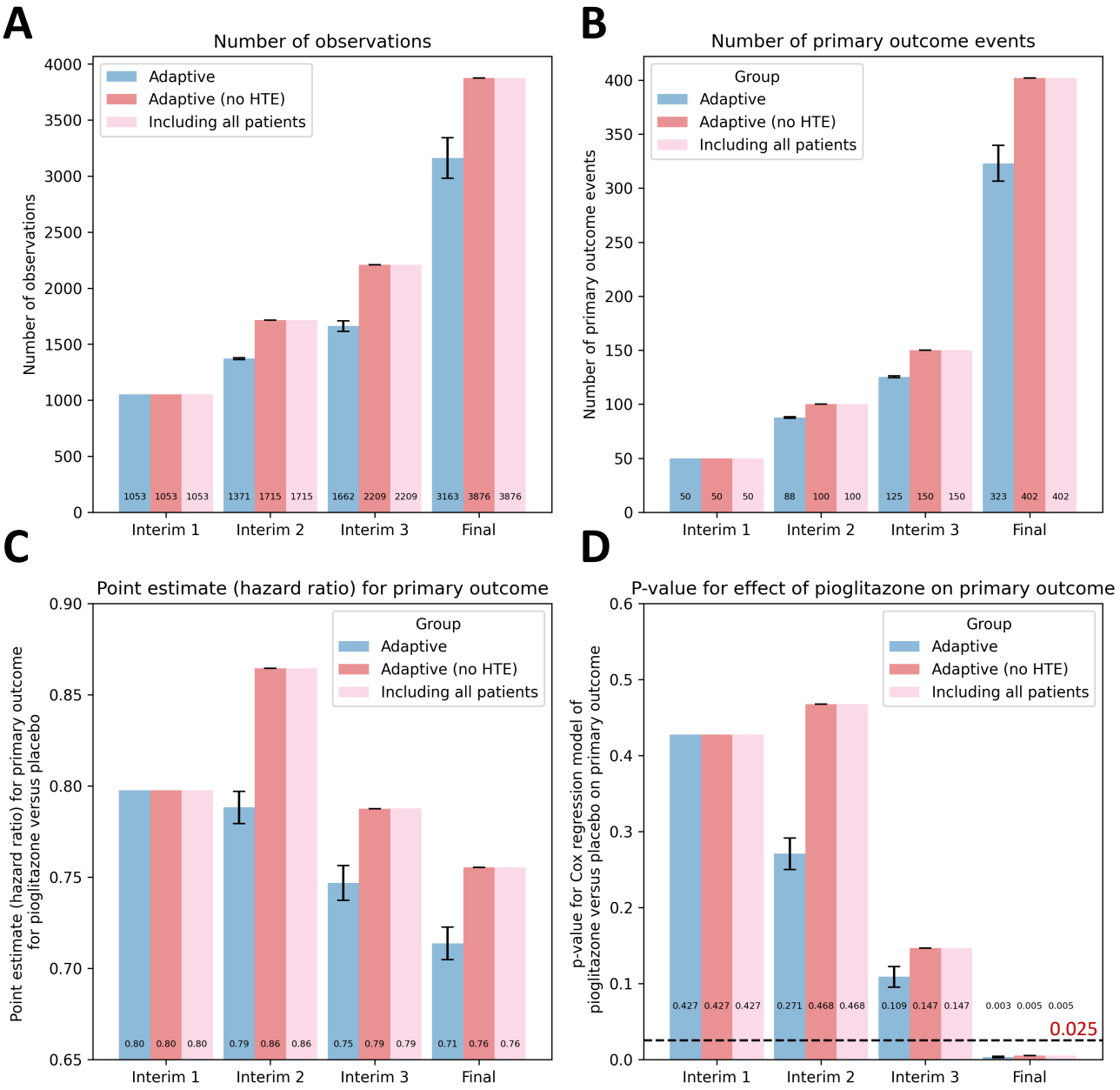

Fig. 3 | Primary outcome results - IRIS. (A) Cumulative patient enrollment, (B) cumulative primary outcome events, (C) Cox regression-derived effect estimates (Hazard Ratios) for pioglitazone versus placebo, and (D) corresponding P values. The error bars denote the standard error of mean across n=5 adaptive simulations. HTE: heterogeneous treatment effect; IRIS: Insulin Resistance Intervention after Stroke.

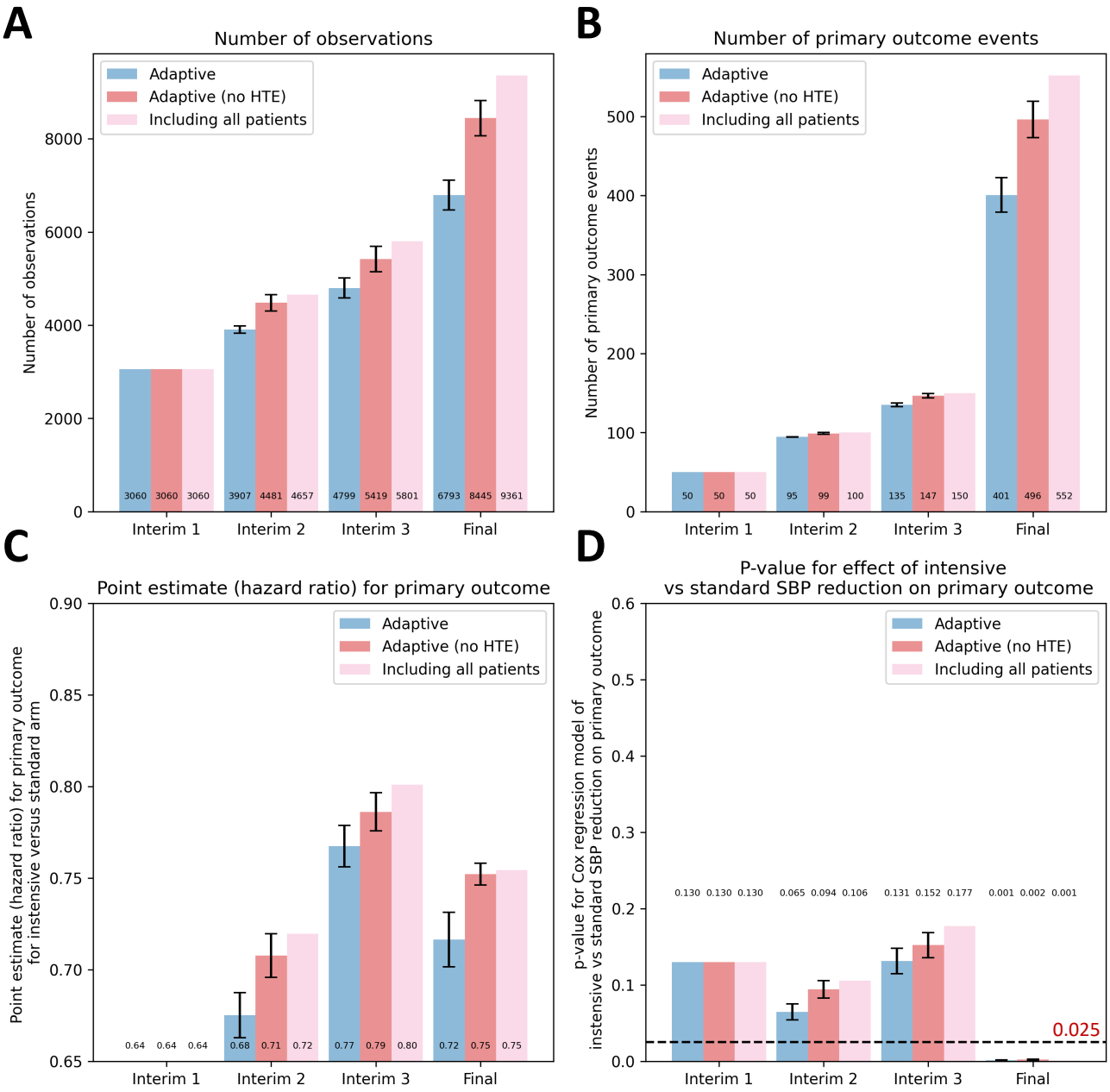

Fig. 4 | Primary outcome results - SPRINT. (A) Cumulative patient enrollment, **(B)** cumulative primary outcome events, **(C)** Cox regression-derived effect estimates (Hazard Ratios) for intensive versus standard systolic blood pressure reduction, and **(D)** corresponding P values. The error bars denote the standard error of mean across n=5 adaptive simulations. HTE: heterogeneous treatment effect; SPRINT: Systolic Blood Pressure Intervention Trial.

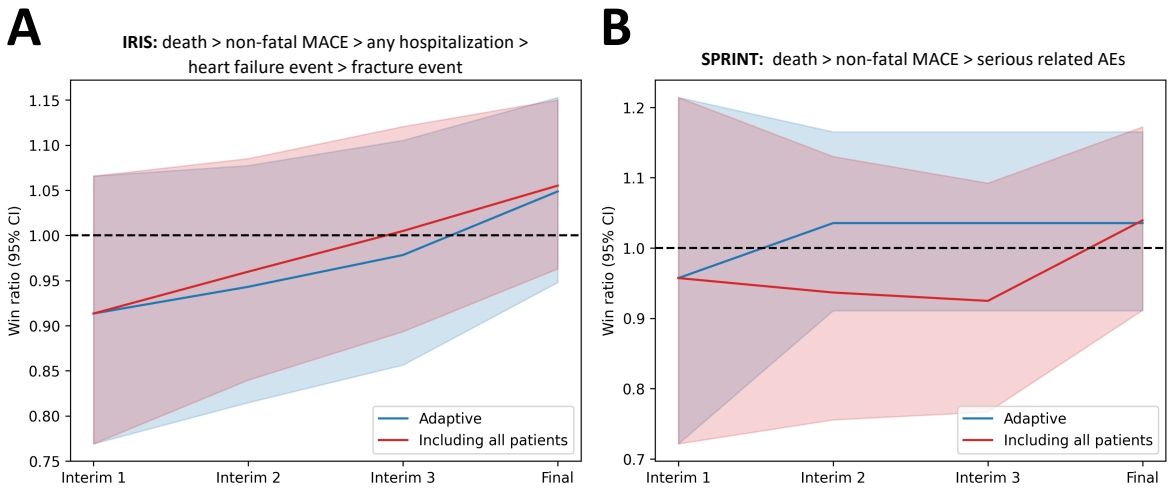

Fig. 5 | Secondary outcome (safety) results. Win ratio with corresponding 95% confidence intervals across the pre-defined interim analysis timepoints. **(A)** Example from IRIS with outcome events ranked as follows: all-cause mortality, followed by non-fatal MACE components, and then all-cause hospitalizations, heart failure events and bone fractures; **(B)** Example from SPRINT: all-cause mortality, followed by non-fatal MACE components, and then serious adverse events. The lines correspond to the win ratio point estimate with shaded areas denoting the 95% confidence interval (for the adaptive trial design the point and upper and lower ends of the confidence interval were averaged across the five simulations). IRIS: Insulin Resistance Intervention after Stroke; SPRINT: Systolic Blood Pressure Intervention Trial; TIA: transient ischemic attack.