scientific reports

Check for updates

OPEN

Synthetic data to boost under-represented patients and create virtual trial cohorts: the RATE-AF case study

Barbara Draghi[1,2]✉, Dima Attal[2], Dipak Kotecha[3,4], Puja Myles[1], Matthew Chapman[3], Asgher Champsi[3], Richard Branson[1], Karina V. Bunting[3], Alastair R. Mobley[3] & Allan Tucker[2]

Clinical trials are essential for medical progress, but in certain circumstances can be constrained by recruitment costs, ethical challenges, and limited diversity in participant representation, reducing the generalisability of findings across all subgroups. In these cases, the development of digital approaches that complement traditional trials could be of particular value. We present a Bayesian network framework for synthetic data generation in clinical trials, designed to (1) boost the representation of small and under-represented subgroups and (2) generate virtual patient cohorts that replicate full trial populations with high fidelity. The framework combines probabilistic modelling with conditional synthetic data generation and is evaluated using data from the RAte control Therapy Evaluation in permanent Atrial Fibrillation (RATE-AF) randomised controlled trial, a study that compared two treatments for rate control (digoxin versus bisoprolol, a beta-blocker) in patients with atrial fibrillation and symptoms of heart failure. The framework was first assessed in a controlled boosting experiment, designed to recover simulated subgroup under-representation within the original cohort, and then extended to an exploratory boosting scenario to examine hypothetical increases in subgroup representation, before being applied to replicate the full trial population. Across these settings, it preserved statistical fidelity and reproduced the analytical results observed in the real data, boosting under-represented subgroups where sufficient data are available, whilst acknowledging limitations of boosting under extreme small-sample scenarios. This study positions synthetic data as a potential digital pathway that, with further development, could be used to support real-world clinical trials where recruitment of some population subgroups may be challenging.

Keywords Synthetic data, Clinical trials, Boosting

Clinical trials are central to the evaluation of medical interventions and the advancement of patient care. Randomised controlled trials represent the core approach to evaluate the efficacy and safety of clinical therapies avoiding selection, confounding and information biases that make observational data unsuitable for informing treatment decisions[1,2]. Yet, the conventional randomised design, in which participants are allocated to intervention or control groups, can sometimes be subject to recruitment challenges. These limitations can contribute to the under-representation of specific subgroups and the emergence of systematic biases within clinical trial populations[3,4].

Among the multiple sources of bias in healthcare data, a notable one is *representation bias*, defined as the systematic under-representation of specific patient subgroups within the study cohort[5,6]. This bias arises when the distribution of key characteristics, such as ethnicity, age, or comorbidity burden in a trial cohort does not reflect that of the broader population to which the results are intended to apply[7,8]. Representation bias in clinical trials may occur due to under-represented minority groups[9]. The lack of diversity in clinical trials can limit the generalisability of findings to diverse patient populations. Differences in pharmacokinetic and pharmacodynamic response exist across racial and ethnic groups, reinforcing the importance of diverse trial populations for ensuring safe and effective treatment across subgroups[10]. Throughout this work, we refer to representation bias

[1]Medicines and Healthcare products Regulatory Agency (MHRA), London, UK. [2]Brunel University London, London, UK. [3]Department of Cardiovascular Sciences, University of Birmingham, Birmingham, UK. [4]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. ✉email: barbara.draghi@mhra.gov.uk; barbara.draghi@brunel.ac.uk

simply as *bias*, unless otherwise specified. We use the term *sensitive attribute* to denote patient characteristics in which bias may manifest, including ethnicity, age and gender, and *sensitive group* to indicate patients sharing a specific value of that attribute (for example, female when gender is the sensitive attribute).

Traditional strategies to mitigate bias, such as targeted recruitment or post hoc statistical reweighting, have demonstrated limited effectiveness, particularly in cases involving small or structurally under-represented groups[11]. Consequently, there is increasing interest in data-centric methods that aim to detect and address these biases in the accrued data. The search for alternatives to conventional randomised controlled trials has intensified across the pharmaceutical and MedTech sectors, with increasing advocacy for the use of synthetic or external arms to overcome the limitations of small and ethically constrained studies. In the UK, these developments align with the recommendations of the Taskforce on Innovation, Growth and Regulatory Reform (TIGRR) and with the Medicines and Healthcare products Regulatory Agency (MHRA) initiatives exploring synthetic data for regulatory science[12].

Within this context, synthetic data generation has emerged as a promising method for producing realistic patient data, with recent advances enabling high-fidelity substitutes for real-world datasets[13–17]. Bayesian networks offer an interpretable and transparent framework for modelling dependencies among clinical variables and generating synthetic patient data. They have been shown to reproduce large-scale primary care datasets with high fidelity while preserving both marginal distributions and clinically meaningful relationships[18,19]. While synthetic data shows potential for replicating large-scale health dataset[18], its application in clinical trials remains largely unexplored. Methods to replicate[20] or simulate clinical trial populations[21] have been proposed, but they typically focus on large cohorts with substantial sample sizes and event rates, which differ substantially from the smaller and more granular population examined in this study[22].

Synthetic data generation has been increasingly proposed for mitigating bias and increasing fairness through deep learning approaches[23–25] and resampling techniques[5,26,27]. Bayesian networks have also been applied to mitigate representation bias, offering an interpretable and probabilistic framework for boosting under-represented groups in healthcare data[28,29]. However, most of these approaches have been developed and validated on large healthcare datasets, and their application to clinical trials, which typically involve smaller sample sizes, remains largely unexplored. In addition, previous work has shown that the fidelity and utility of synthetic data generation deteriorate when subgroup representation becomes too limited, as models fail to recover the underlying population structure[30].

To address these challenges, this study introduces a Bayesian network framework for synthetic data generation and evaluation in clinical trials, designed to achieve two complementary objectives: (1) improving the representation of small and under-represented subgroups through *data boosting*, and (2) generating virtual patient cohorts that replicate full trial populations while preserving the clinical and statistical dependencies observed in the trial data. The framework is specifically tailored to small, structured clinical-trial datasets. The main methodological contribution of this study lies in adapting probabilistic synthetic data generation to this small-data clinical-trial setting by incorporating aggregation strategies in Bayesian network learning to improve structural stability, explicit bias quantification through the coverage metric, and targeted conditional generation rules for subgroup boosting. Unlike prior work, which has largely focused on larger healthcare datasets, the proposed framework is evaluated under controlled subgroup-removal scenarios, including extreme sparsity settings in which training data are reduced to as few as 41 patients. This allows us to characterise the conditions under which synthetic boosting remains reliable and where it begins to fail.

First, the framework enables *data boosting* through conditional generation of synthetic patients, improving the representation of under-represented subgroups while preserving the clinical and statistical dependencies observed in the trial data. Boosting is first validated in a *controlled* setting, where bias is deliberately simulated and subsequently corrected, demonstrating the framework's ability to recover under-represented subgroups without compromising data fidelity or clinical utility. The approach is then extended to an *exploratory* setting, where the original trial is treated as biased relative to a hypothetical balanced scenario in which all ethnic groups are equally represented. Conditional generation is guided by this hypothetical distribution, illustrating how tailored synthetic data generation can support the exploration of fairness-oriented and scenario-based analyses, enabling exploration of how trial generalisability might change under more balanced cohorts. Second, the framework supports *data replication*, generating fully synthetic clinical trial cohorts that could replace or complement a real dataset when recruitment, cost, or ethical constraints limit further enrollment.

Beyond its technical contribution, this study advances a principled methodology for using synthetic data to support real-world clinical trials. By generating synthetic patients where patients recruitment and diversity remain limited, the study investigates the extent to which synthetic data can support real-worlds clinical trials.

Methods

Overview of the framework

We developed a probabilistic framework for generating synthetic clinical trial data using Bayesian networks. The framework was designed to evaluate two complementary applications of synthetic data generation in clinical research: *data boosting* and *data replication*. These two applications operationalise the central objectives of this study: to demonstrate that synthetic data can (1) boost real datasets to address small sample limitations, and (2) generate fully synthetic cohorts that act as virtual patient populations. Both applications follow a structured workflow comprising model training, conditional generation, and evaluation components, described in the following subsections. The terminology for sensitive attributes and groups introduced in the Introduction is used consistently throughout.

Data boosting employs conditional synthetic data generation to increase the representation of under-represented patient subgroups. Two complementary experimental configurations were implemented, the first to validate and the second to explore the boosting approach. In the *controlled boosting* experiment, bias was

deliberately simulated by removing a predefined proportion of patients from selected sensitive groups, and the framework was applied to *boost* their representation through conditional generation. This controlled setting validates the capacity of the model to recover under-represented subgroups while preserving both the statistical structure and clinical plausibility of the data.

The *exploratory boosting* experiment extends this approach to a hypothetical balanced scenario designed to explore how achieving demographic balance could influence fairness in clinical outcomes. Here, the original real-world cohort is treated as biased relative to a hypothetical balanced population defined *a priori* (for example, a demographically balanced population in which all ethnic groups are equally represented). Conditional generation is guided by this hypothetical configuration to generate a rebalanced synthetic cohort, allowing exploration of how demographic equity could affect analytical conclusions. This configuration investigates how tailored synthetic generation can support fairness-oriented and scenario-based analyses, facilitating the exploration of trial generalisability in populations that differ from those originally enrolled and highlighting where additional real-world clinical studies or dedicated sub-trials may be warranted.

Data replication aims to generate a fully synthetic clinical trial cohort that mirrors the real one with high fidelity. A Bayesian network trained on the original trial data is used to synthesise a cohort of matching size and structure, enabling direct comparison of marginal distributions and inter-variable dependencies between real and synthetic data. An overview of the framework and experimental design is shown in Fig. 1.

Model training

Model training formed the generative foundation of the framework. To support both boosting and replication objectives, Bayesian networks were trained to model the joint probability distribution among clinical variables. Network structure and parameters were learned using the Structural Expectation-Maximization (SEM), which alternates between imputing missing values and optimising the network structure, combining the advantages of expectation-maximisation and score-based search. Structure learning was performed using hill climbing guided by the Bayesian Information Criterion (BIC), and parameters were estimated via maximum likelihood. This approach accommodates mixed data types and handles the sparsity typical of clinical trial datasets. Robustness of the trained structure was evaluated through repeated SEM runs and bootstrap aggregation to ensure stability of key dependencies. The final model was reviewed by clinical members of the trial management group, who defined lists of allowed and prohibited paths that were used to refine the model-training settings applied for all subsequent synthetic data generation. A summary of the aggregated Bayesian network structure is provided in the Supplementary Material (Figure S3), where edges are shown according to their frequency across repeated model-learning iterations.

Bias detection and condition formulation

Representation bias was quantified to derive the conditional rules guiding synthetic data generation. Among established bias metrics[31], we adopted *coverage*[32] as the principal measure, as it directly quantifies how well a dataset reflects a reference population. Coverage was defined as the ratio between the proportion of a sensitive group in the dataset and its corresponding proportion in the target population. For each sensitive group, the number of synthetic patients required to achieve a target coverage value was estimated and used to construct conditional sampling rules applied during data generation. In the controlled boosting experiment, coverage was computed relative to the original real cohort to quantify the magnitude of simulated under-representation and determine the samples required for recovery, whereas in the exploratory boosting experiment, it was computed relative to a hypothetical balanced distribution defined *a priori* (for example, a demographically balanced cohort).

Synthetic data generation

Synthetic datasets were generated from the trained Bayesian network using either random or conditional sampling, depending on the experimental setting. For data replication, patient records were drawn randomly from the joint probability distribution encoded by the Bayesian network, reproducing the overall statistical structure of the trial population and generating a fully synthetic cohort matched in size and composition to the real one. For data boosting, sampling was performed conditionally on predefined sensitive-group attributes identified during bias detection. The generation process iteratively sampled new patient records consistent with the conditional probabilities specified by the network, increasing the representation of the target subgroup until the desired coverage or population composition was achieved.

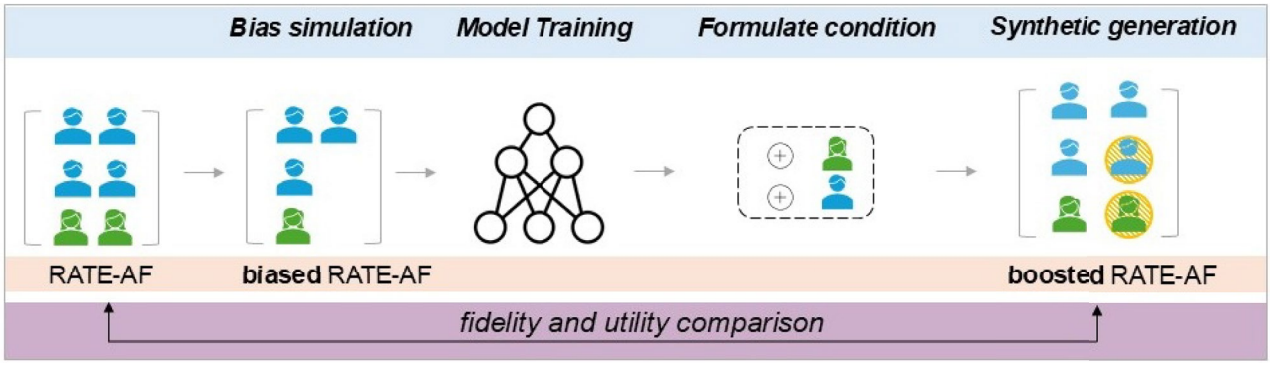
Synthesis pre- and post-processing

Pre- and post-processing procedures were designed to ensure that the Bayesian network could model clinical trial data realistically and that the resulting synthetic datasets retained meaningful variability and scale. Before model training, continuous variables were discretised using quantile-based binning to allow their representation within a mixed-type Bayesian network while preserving their underlying distributional properties. Variables that were deterministic or derived from others were excluded to prevent the introduction of artificial dependencies and to maintain a coherent probabilistic structure. Following synthetic data generation, discretised variables were transformed back to continuous form by resampling within their original bin intervals to restore the natural granularity of clinical measurements.

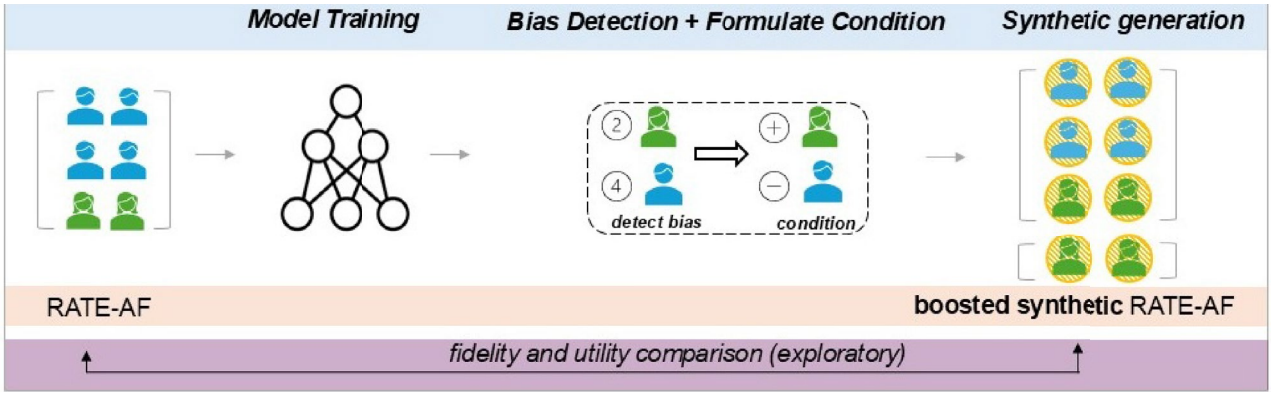
Data and experimental settings

Data

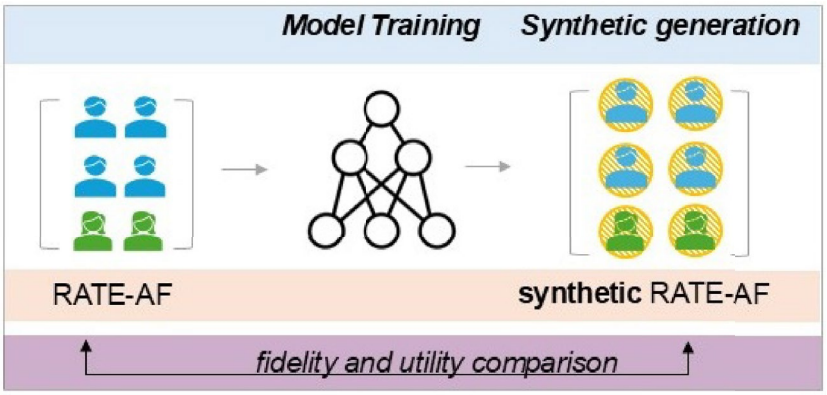
The framework was evaluated using data from the RAte control Therapy Evaluation in permanent Atrial Fibrillation (RATE-AF) trial, a UK-based, randomised, open-label, blinded end-point study that compared low-



(a) Controlled boosting. Bias is simulated by removing patients from the RATE-AF dataset and corrected through conditional synthetic data generation.



(b) Exploratory boosting. The original cohort is treated as biased relative to a hypothetical balanced scenario, and conditional generation rebalances the data toward this target.



(c) Replication. The Bayesian network model is used to reproduce the entire clinical trial cohort.

Fig. 1. Overview of the proposed framework. The framework comprises two applications of synthetic data generation in clinical trials, namely data boosting and data replication, implemented across three experimental settings: (**a**) controlled boosting, (**b**) exploratory boosting, and (**c**) replication. Blue and green icons represent real patients; icons with a yellow circular background denote synthetic patients. The lower bar illustrates fidelity and analytical-utility comparisons performed in each setting.

dose digoxin with bisoprolol for rate control in patients with permanent atrial fibrillation and symptoms of heart failure[33,34]. The trial was embedded within the UK National Health Service and conducted in partnership with a patient and public involvement team[35]. After written informed consent, 160 participants aged 60 years or older with a clinical need for heart rate control were recruited from primary care sites and hospitals in the West Midlands between 2016 and 2018.

The data comprises 160 patients and 70 clinical features covering demographics, comorbidities, treatment allocation, physiological measurements, and outcomes, which have previously been published in full[36,37]. For this study, a subset of clinically relevant variables was selected in consultation with the trial management group (Table 1). Participants were asked to self-declare their ethnicity based on the code list for the 2011 UK Census.

Feature	Description	Values used in analysis
Ethnicity	Self-declared ethnicity, coded per 2011 UK Census categories	Recoded as White, Asian, and Black (see Table S10 in Supplementary Materials)
Age	Age at baseline (years).	Continuous
Gender	Self-reported gender.	Male; Female
Creatinine	Serum creatinine (μmol/L).	Continuous
mEHRA	Modified European Heart Rhythm Association symptom classification for atrial fibrillation measured at baseline and follow-up visits.	1: no symptoms; 2a: mild symptoms from AF; 2b: moderate troubling symptoms; 3: severe symptoms affecting daily activity; 4: disabling symptoms preventing normal activity
NYHA	New York Heart Association Functional Classification for Heart Failure measured at baseline and follow-up visits.	I: no limitation of physical activity II: slight limitation of physical activity III: marked limitation of physical activity IV: unable to carry out any physical activity
Type II Diabetes	Presence of Type II diabetes mellitus.	No/Yes
NT-proBNP	N-terminal pro B-type natriuretic peptide (pg/mL).	Continuous
Heart Rate	Heart rate measured at baseline and follow-up visits (bpm).	Continuous
SBP	Baseline systolic blood pressure (mmHg).	Continuous
LVEF	Estimated left-ventricular ejection fraction (%).	<40% (reduced); 40–49% (mildly reduced); ≥50% (preserved)
ECG HR	Heart rate measured via 12-lead electrocardiogram (ECG) at baseline and follow-up visits (bpm).	Continuous
Adverse Events	Any adverse event during 12 months.	No/Yes
Cardiovascular Events	Adjudicated cardiovascular event during 12 months.	No/Yes
Composite Clinical Outcome	presence of either adverse events or cardiovascular events during 12 months.	No/Yes

Table 1. Summary of clinical features used in evaluation.

Ethnicity data are collected to monitor for health inequalities in the UK National Health Service. However, participants were able to refuse disclosure of ethnicity. Ethnicity codes were grouped into three categories: White, Asian, and Black (see Table S10 in Supplementary Materials). All methods were performed in accordance with relevant guidelines and regulations; the methodology developed in this study was applied to anonymised RATE-AF clinical trial data that were collected under appropriate ethical and regulatory approvals and used in accordance with those approvals (see Ethics Declarations for details).

Experimental settings

Ethnicity was selected as the sensitive attribute following expert consultation and exploratory analysis of the RATE-AF dataset, which showed under-representation of Asian and Black participants compared with the Birmingham population reported by the Office for National Statistics (ONS)[38]. Three experimental settings were implemented to evaluate the framework across progressively broader applications: controlled boosting, exploratory boosting, and replication.

In the *controlled boosting* setting, bias was simulated by randomly removing 30%, 60%, or 80% of patients from the sensitive group. The framework was then applied to generate synthetic patients until the original population composition was recovered. This configuration served as a validation exercise, testing the framework's ability to recover under-represented subgroups under increasing degrees of bias.

The *exploratory boosting* setting extended this approach to a broader scenario in which the original RATE-AF cohort was considered biased relative to a hypothetical balanced scenario. An idealised distribution with equal representation across ethnic groups was defined *a priori* as an expert-informed exploratory benchmark, chosen to provide a simple and interpretable counterfactual scenario for examining how subgroup balance might affect predictive fairness and analytical conclusions. This target was not intended to represent the underlying source population or a normative definition of fairness. Conditional generation was guided by this composition to synthesise a rebalanced cohort and explore how subgroup rebalancing could influence fairness in clinical outcomes.

Finally, the *replication* setting evaluated the framework's ability to reproduce a virtual, synthetic copy of the entire clinical trial dataset. The Bayesian network was used to generate, through random sampling, a fully synthetic cohort of equivalent size and structure of the original trial.

Evaluation

Evaluation focused on assessing how effectively the proposed framework met its two central objectives: improving the representation of under-represented patient subgroups and generating virtual trial cohorts. The evaluation plan was structured across three experimental settings (controlled boosting, exploratory boosting, and replication), each addressing a distinct methodological objective and therefore requiring different evaluation dimensions.

In the controlled boosting experiments, the primary goal was to test whether synthetic augmentation could recover under-represented subgroups without compromising the statistical and clinical properties of the data. Fidelity and utility were therefore used as the main validation criteria. In the exploratory boosting experiments, the focus shifted to examining whether conditional generation could rebalance the cohort toward a hypothetical balanced scenario and how this affected predictive performance and fairness across sensitive groups. Finally, the

replication experiments aimed to determine whether the framework could reproduce the statistical structure and analytical results of the original trial dataset, again prioritising fidelity and utility.

All experiments were repeated 500 times and results were averaged to ensure stability and reproducibility. The following sections describe the metrics and analytical procedures used for each evaluation dimension.

Fidelity Fidelity is defined as the extent to which synthetic data preserve the statistical properties of the original dataset. In this study, it was assessed using the Kruskal-Wallis rank-based nonparametric test across demographic characteristics, clinical biomarkers, physiological measurements, disease classification scores, and comorbidity indicators, listed in Table 1. Concordance between real and synthetic distributions provided a quantitative measure of fidelity. In the controlled boosting setting, fidelity is evaluated at the level of the final boosted dataset to reflect the primary objective of assessing its suitability for downstream analyses. To complement this, we also compared removed real records with their recovered synthetic counterparts, with results provided in the Supplementary Material.

Utility Utility was evaluated by replicating the original RATE-AF statistical analyses to determine whether synthetic data reproduced clinical findings and treatment effects observed in the real data. The evaluation included both continuous and categorical endpoints used in the trial, encompassing changes in NYHA functional class, heart rate, mEHRA score, NT-proBNP levels, and the occurrence of adverse and cardiovascular events (See Table 1 for details).

Consistent with the RATE-AF analytical framework, clinical outcomes were modelled using multivariable regression approaches adjusted for key baseline covariates. A one or more class improvement in NYHA functional class at 12 months was modelled using logistic regression, adjusting for treatment arm, age at randomisation, gender, baseline mEHRA score, and baseline LVEF. ECG heart rate at 12 months was analysed using a linear regression model with the same covariates, also including baseline ECG heart rate. A two or more class improvement in mEHRA class was also modelled using logistic regression with the same independent variables as for NYHA class improvement.

NT-proBNP at 12 months, due to its non-normal distribution, was analysed using log-transformed linear regression with independent variables including treatment arm, gender, age at randomisation, baseline NT-proBNP, baseline mEHRA, and baseline ejection fraction. Results were back-transformed and reported as geometric mean ratios. Safety outcomes, including adverse and cardiovascular events, were modelled using logistic regression. In all analyses, the bisoprolol arm was used as the reference treatment, and results were reported as adjusted odds ratios or mean differences with corresponding 95% confidence intervals.

Bias and Fairness Bias was quantified using the coverage metric described in the subsection "Bias Detection and Condition Formulation", with ethnicity specified as the sensitive attribute. In the boosting experiments, coverage was calculated before and after augmentation to measure recovery toward the target population.

Fairness was evaluated in the exploratory boosting setting using a Naïve Bayes classifier trained to predict the composite clinical outcome. The model used the clinical and demographic variables listed in Table 1 as predictors (excluding outcome components), with ethnicity used only to define the sensitive groups for fairness evaluation. Because RATE-AF is a small clinical-trial dataset, fairness was assessed using repeated bootstrap evaluation rather than a fixed train/test split. At each iteration, a bootstrap sample was used for model training and the corresponding out-of-bootstrap observations were used for testing. Group-level metrics (Equal Opportunity, Predictive Rate Parity, and Accuracy Parity) were calculated from model predictions[39,40]. Fairness values were summarised across iterations as means with 95% confidence intervals. Selecting fairness metrics in healthcare presents particular challenges because attributes such as ethnicity, age, and gender may legitimately influence outcomes rather than act solely as confounders[41]. Accordingly, the chosen metrics capture complementary aspects of predictive equity, assessing error-rate balance (Equal Opportunity), calibration (Predictive Rate Parity) and overall performance consistency (Accuracy Parity) across sensitive groups. Because no independent hold-out set was available, this procedure provides an internal assessment of fairness within the observed dataset rather than external validation of fairness in unseen populations.

Results

We evaluated the framework across three progressively broader experimental settings, *controlled boosting*, *exploratory boosting*, and *replication* of RATE-AF. These configurations represent a methodological progression from targeted correction of subgroup bias to fully synthetic replication of the clinical trial dataset. The following sections present the results obtained from these evaluations, repeated across 500 independent iterations to ensure robust conclusions, with confidence intervals summarising variability across iterations where applicable.

Boosting patients in RATE-AF

Data boosting experiments investigated whether conditional synthetic data generation could recover under-represented patient subgroups identified in the RATE-AF dataset while maintaining the statistical and analytical properties of the original trial, in line with the first aim of the study.

Controlled boosting. The controlled boosting experiment was designed to validate the ability of the framework to recover under-represented subgroups (i.e., restoring the original cohort composition after simulated removal) while preserving both data fidelity and clinical utility in the final boosted dataset. Table 2 summarises the configuration of each simulated scenario, including the number of participants removed to simulate bias, those retained for model learning, and those generated through conditional boosting. Figures 2, 3, 4 present the results obtained by simulating different degrees of subgroup under-representation (removal of 30%, 60%, and 80% of patients from a given ethnic group) and regenerating the missing representation through conditional synthetic data generation. In all figures, each column label (e.g., *Asian 30%*) refers to the boosted RATE-AF dataset obtained after removing 30% of patients from the Asian subgroup and subsequently restoring them through synthetic data generation to recover the original population composition.

Subgroup	Bias %	n patients				
		Original RATE-AF	Subgroup total	Removed	Remaining	Recovered
White	30%	160	149	45	115	45
White	60%	160	149	89	71	89
White	80%	160	149	119	41	119
Asian	30%	160	8	2	158	2
Asian	60%	160	8	5	155	5
Asian	80%	160	8	6	154	6
Black	30%	160	3	1	159	1
Black	60%	160	3	2	158	2
Black	80%	160	3	2	158	2

Table 2. Summary of simulated bias and boosting scenarios in RATE-AF (Controlled Boosting). Each scenario specifies the degree of subgroup under-representation simulated (*Bias %*), the number of patients before and after removal, and the number of synthetic patients generated to restore representation. All counts represent averages across 500 iterations. "Original RATE-AF" indicates the full trial cohort (160 patients). "Subgroup total" denotes the number of patients in each subgroup before simulated removal. "Removed" indicates the absolute number of patients removed from the subgroup. "Remaining" corresponds to the dataset used for model training after simulated subgroup bias, and "Recovered" to the number of synthetic patients generated to restore the full cohort

Fidelity. Figure 2 shows the extent to which the boosted datasets preserved the statistical distributions of the original RATE-AF variables. For each variable, the Kruskal-Wallis test compared real and boosted datasets, with high p-values indicating no significant difference. The heatmap reports mean p-values averaged over 500 iterations, providing an overview of distributional preservation for each variable and scenario. Green cells indicate preserved distributions ($p > 0.05$), and red cells mark detectable distributional shifts. Most variables retained high fidelity across all boosting levels, with only minor deviations observed under the 80% removal scenario. Extended results are reported in the Supplementary Material, Table S1,S2 and S3. In addition, fidelity analyses comparing removed real patients with their recovered synthetic counterparts are provided in the Supplementary Material (Section S1.1, Figure S1).These results further support the findings in Fig. 2, indicating that the distributions of the recovered synthetic records are largely consistent with those of the corresponding removed real records, with no systematic evidence of statistically significant differences, although increased variability is observed in higher-removal scenarios.

Utility. Figures 3 and 4 summarise whether boosted datasets preserved the analytical significance of treatment effects estimated in the original RATE-AF models. For each outcome model, the p-value of the treatment-allocation term was averaged across 500 iterations. Preservation was defined as maintaining the same significance pattern and direction of effect as the baseline model, i.e., the model fitted on the real RATE-AF data. Figure 3 displays these results as a heatmap: blue cells indicate the baseline models, green cells show scenarios where the boosted dataset preserves the baseline significance status and red cells indicate scenarios where the significance status changes. Here, 'significance' refers specifically to whether the treatment allocation term (digoxin vs. bisoprolol) is statistically significant or non-significant in the outcome model. Figure 4 presents the corresponding averaged effect estimates and 95% confidence intervals as forest plots. Effect measures are reported as odds ratios (OR) for binary outcomes, ratios of geometric means (RGM) for log-transformed continuous outcomes, and adjusted mean differences (AMD) for linear outcomes; the vertical dashed line denotes the null effect (OR = 1, RGM = 1, AMD = 0). Extended results are provided in the Supplementary Material, Table S4,S5 and S6.

Subgroup-level utility analyses, stratified by ethnic group, are also reported in the Supplementary Materials (Figure S1). These analyses applied the same utility evaluation procedure within each subgroup to assess reproducibility of treatment effects in the boosted datasets.

Exploratory boosting. The exploratory boosting experiment was designed to evaluate whether the framework could rebalance the original RATE-AF cohort toward a hypothetical balanced scenario (i.e., equal representation across White, Asian, and Black participants) while improving both subgroup representation and predictive fairness. The resulting dataset, referred to as the *boosted synthetic RATE-AF* cohort, represents a conceptual scenario rather than a real-world population, used to explore how balanced subgroup composition might influence representation and predictive fairness within clinical trial data. Evaluation focused on two complementary aspects: coverage, quantifying subgroup representation relative to a target population, and fairness, assessing parity of predictive performance across sensitive groups. Baseline characteristics stratified by ethnicity for both Real and Boosted Synthetic data are provided in Supplementary Table S4, for readers interested in detailed distributional comparisons.

Coverage. Table 3 reports coverage values for each ethnic group before and after boosting. Coverage values below 1 indicate under-representation relative to the hypothetical balanced scenario, and values above 1 indicate over-representation. In the original RATE-AF data, Asian and Black participants were under-represented (coverage = 0.2 and 0.6, respectively), while White participants were over-represented (coverage = 2.8). After boosting, coverage values converged to 1.0 across all groups, indicating full recovery of the target population balance.

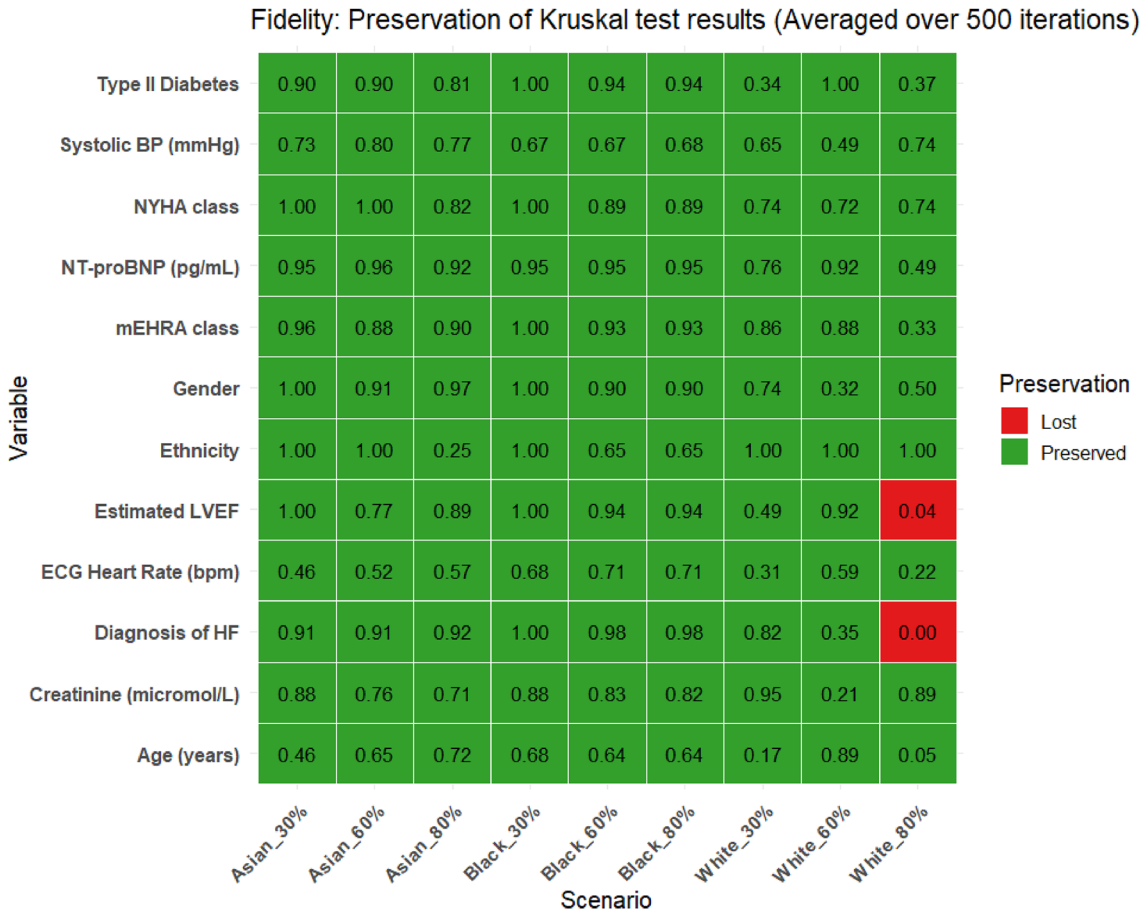


Fig. 2. Fidelity analysis under controlled boosting. Each column (e.g., *White 30%*) corresponds to a *boosted RATE-AF dataset* generated for a specific bias-simulation scenario (e.g., *White_30%* indicates that 30% of White patients were removed from the original cohort and subsequently generated with synthetic data). The heatmap displays mean *p*-values from the Kruskal–Wallis test cross 500 iterations, with higher values indicating closer alignment between the real and boosted distributions (i.e., higher fidelity). Green cells indicate no significant difference from the original RATE-AF distribution, and red cells indicate detectable shifts. Most clinical variables retained high fidelity across all boosting levels, with only minor deviations at 80%.

Fairness. The same table presents fairness metrics, Equal Opportunity (EO), Predictive Rate Parity (PRP), and Accuracy Parity (AP), computed for prediction of the composite clinical outcome across ethnic groups. Fairness was measured relative to the White reference group, with values closer to 1 indicating parity in model performance. In the original RATE-AF data, fairness estimates for Asian participants were generally close to parity, whereas estimates for the Black subgroup were unstable or not estimable for some metrics, reflecting the extremely small subgroup size. After boosting, fairness metrics moved closer to parity across all groups, with EO, PRP, and AP all close to 1.0 in the boosted synthetic RATE-AF cohort. These results indicate that, within this exploratory bootstrap-based evaluation, boosting improved both subgroup representation and the consistency of predictive performance across ethnic groups under the hypothetical balanced scenario.

Replication of RATE-AF

The replication experiment was designed to evaluate whether the framework could reproduce the complete RATE-AF clinical trial data while preserving both the statistical structure and analytical properties of the original data, in line with the second aim of the study. Evaluation focused on two complementary aspects: fidelity, assessing the alignment of statistical distributions between real and synthetic data, and utility, measuring the extent to which synthetic data preserved the analytical conclusions of the trial.

Fidelity. Figure 5 presents the distributional fidelity between the real and synthetic RATE-AF datasets. Each panel compares the distributions of the baseline variables in the real (blue) and synthetic (yellow) cohorts. For continuous variables (left), densities were generated from mean and standard deviation values aggregated over 500 iterations, while for categorical variables (right), grouped bar charts depict average proportions across iterations. Kruskal-Wallis *p*-values quantify alignment between real and synthetic distributions, where $p > 0.05$ indicates no statistically significant differences between groups. Comprehensive descriptive statistics, including mean, median and 95% confidence intervals across 500 replications, are provided in Supplementary Table S8.

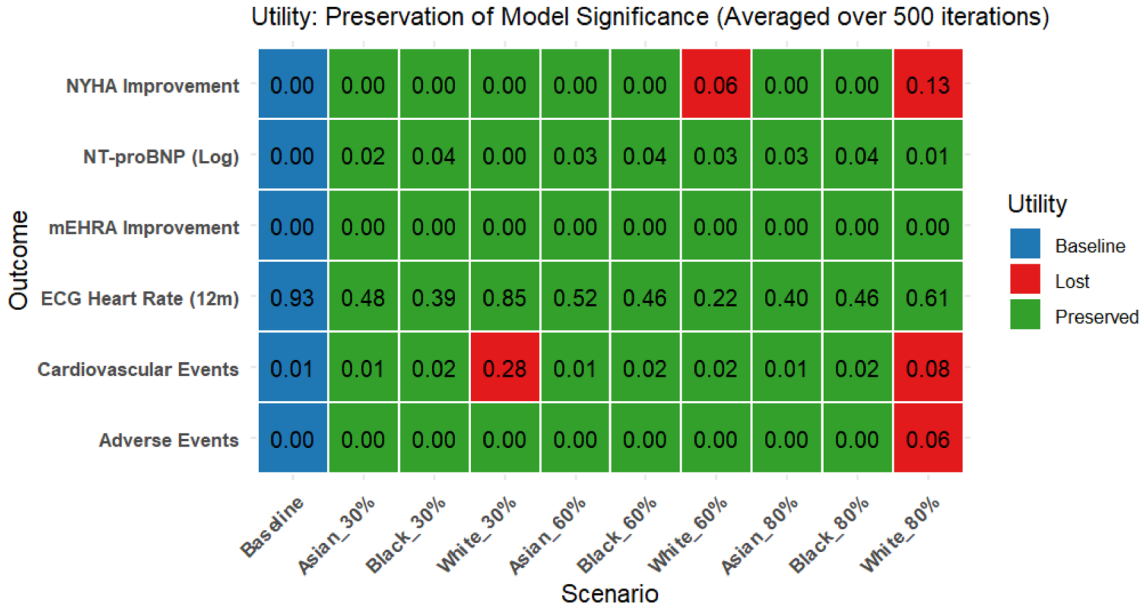


Fig. 3. Utility analysis under controlled boosting. Each column (e.g., *White 30%*) corresponds to a *boosted RATE-AF dataset* generated for a specific bias-simulation scenario (e.g., *White_30%* indicates that 30% of White patients were removed from the original cohort and subsequently generated with synthetic data). Each cell reports the average p-value associated with the treatment-allocation term across 500 iterations. p-values in the boosted datasets are expected to match the baseline behaviour: values ≥ 0.05 indicate non-significant treatment effects, and < 0.05 indicate significant effects. Blue cells (Baseline) denote results from the real RATE-AF dataset, green cells indicate preservation of significance in the corresponding boosted datasets, red cells a loss of significance.

Utility. Figure 6 summarises adjusted treatment effect estimates for six clinical outcomes at 12 months, comparing Digoxin versus Bisoprolol across the real and synthetic RATE-AF data. Outcomes included NYHA class improvement ($\geq$1 category), mEHRA class improvement ($\geq$2 categories), adverse events, cardiovascular events, ECG heart rate and NT-proBNP levels (See Table 1 for more details). Effect measures are reported as odds ratios (OR) for binary outcomes, ratios of geometric means (RGM) for log-transformed continuous outcomes and adjusted mean differences (AMD) for linear outcomes; the vertical dashed line denotes the null effect (OR = 1, RGM = 1, AMD = 0).

For functional improvement, real data showed significantly greater odds of NYHA class improvement with Digoxin, closely matched by the synthetic estimate. Similarly, improvements in mEHRA were consistent between datasets. Safety outcomes were also preserved: the odds of adverse events were significantly lower for Digoxin in both datasets, as were cardiovascular events. Continuous outcomes demonstrated concordant but slightly attenuated effects. Differences in ECG heart-rate were negligible. NT-proBNP levels were significantly reduced with Digoxin in both datasets.

Preservation of utility results was defined as concordance in both statistical significance and direction of treatment effect compared to the model fitted on the real RATE-AF data. Collectively, these findings indicate high concordance between real and synthetic RATE-AF data, with preserved direction and magnitude of treatment effects in all evaluated outcomes.

Discussion

This study introduced and empirically validated a Bayesian network framework for synthetic data generation in clinical trials, designed to meet two complementary objectives: (1) boosting small and under-represented patient subgroups, and (2) generating fully synthetic cohorts that replicate the original trial populations. Applied to the RATE-AF randomised trial, the framework explored conditional probabilistic generation in three progressively broader experimental settings: controlled boosting, exploratory boosting, and full replication of a real-world clinical trial dataset.

In the context of boosting, controlled experiments demonstrated that conditional synthetic generation can successfully recover under-represented subgroups in the RATE-AF dataset while maintaining the statistical and analytical properties of the original trial at the overall dataset level, with important limitations for extremely small subgroups where subgroup-specific information is insufficient for reliable modelling. As shown in Fig. 2, fidelity remained high in nearly all scenarios, with deviations only under the *White 80%* configuration. According to Table 2, this setting involved the removal of 119 of 160 patients, leaving only 41 for model learning. This extreme reduction in sample size limited the information available to estimate conditional dependencies, preventing the Bayesian network from reconstructing the full joint structure of the data reliably, and performance deteriorated as expected.

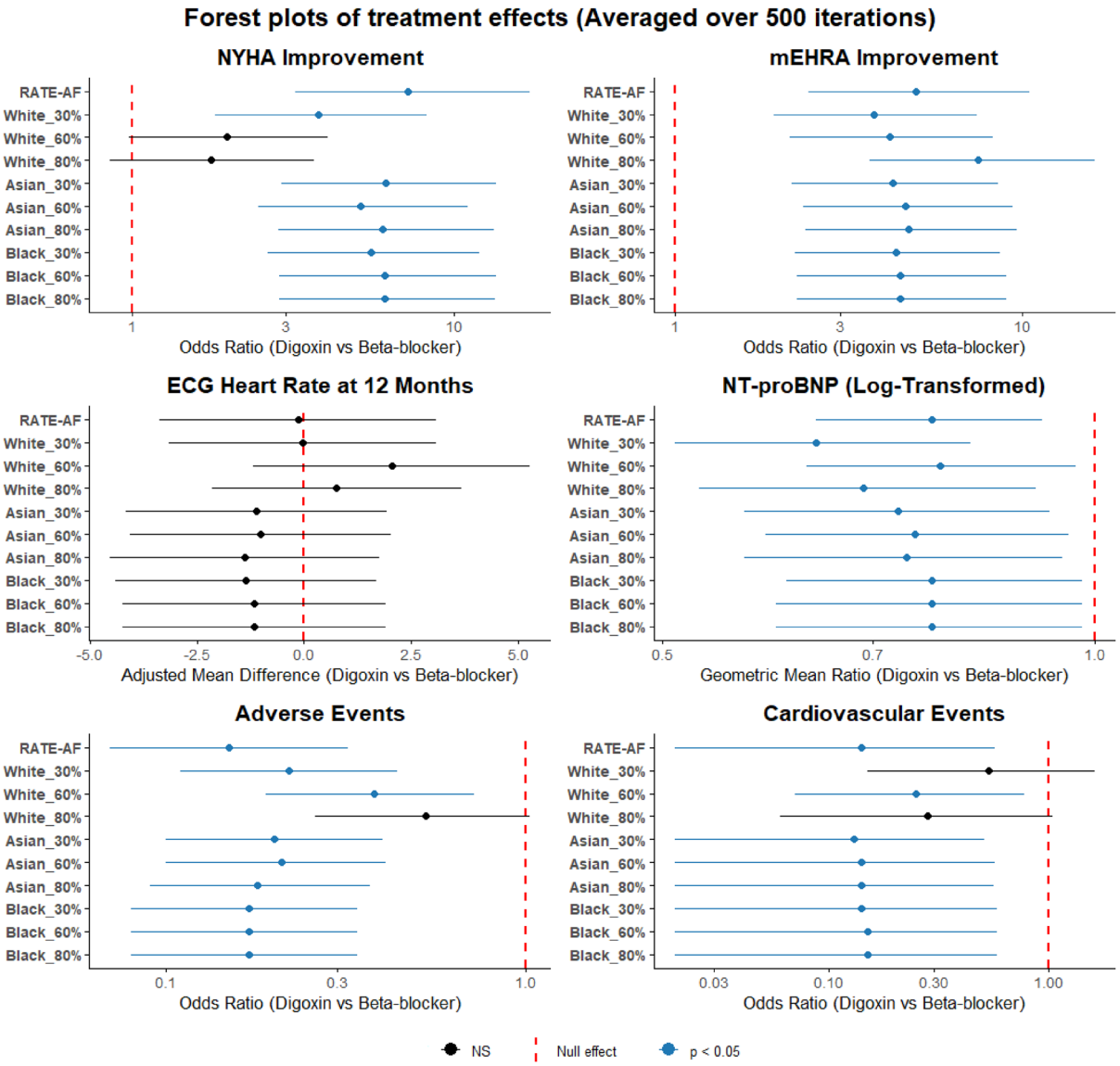


Fig. 4. Forest plots of averaged treatment effects under controlled boosting. Each line (except the top row, *RATE-AF*) corresponds to a boosted RATE-AF dataset generated for a specific bias-simulation scenario (e.g., *White_30%* indicates that 30% of White patients were removed from the original cohort and subsequently generated with synthetic data). Points indicate averaged adjusted treatment-effect estimates and horizontal lines their 95% confidence intervals, computed across 500 iterations. The vertical dashed line marks the null effect (OR = 1, RGM = 1, AMD = 0). The top row (*RATE-AF*) represents treatment effects estimated from the real RATE-AF dataset, used as the baseline reference for all boosted scenarios. Blue markers denote statistically significant effects ($p < 0.05$); black markers indicate non-significance (NS).

By contrast, the *Asian 80%* and *Black 80%* scenarios showed no comparable decline in fidelity, not because the subgroups themselves are large, but because only a small number of records were removed (six and two, respectively), leaving more than 150 patients overall to support structural learning. This suggests that fidelity begins to deteriorate only when the total amount of information available to the model becomes too sparse, as seen in the White 80% scenario, whereas scenarios retaining more than 150 patients still provide enough evidence to estimate the joint distribution accurately, even if few subgroup examples remain. However, although this preserves overall fidelity, it also means that subgroup-specific patterns may be inferred largely from the majority population when minority groups are extremely small, an aspect that may limit generalisability and is, therefore, acknowledged in the limitations. This limitation is particularly relevant for extremely small subgroups, such as Asian (n=8) and Black (n=3) participants in the RATE-AF dataset. In these cases, the model does not have sufficient data to learn subgroup-specific conditional relationships. Consequently, while the boosting procedure can restore subgroup counts, it cannot recover subgroup-specific information that is not present in the original data. As such, the proposed approach is not suitable for deriving subgroup-specific conclusions in cases of extreme under-representation without incorporating external data sources.

Metric	Dataset	White	Asian	Black
Coverage	RATE-AF	2.8	0.2	0.6
	Boosted RATE-AF	1.0	1.0	1.0
EO (Equal Opportunity)	RATE-AF	1.0 (1.0–1.0)	1.1 (1.0–1.2)	–
	Boosted RATE-AF	1.0 (1.0–1.0)	1.0 (0.9–1.2)	1.0 (0.9–1.1)
PRP (Predictive Rate Parity)	RATE-AF	1.0 (1.0–1.0)	1.1 (1.0–1.1)	0.0 (0.0–0.0)
	Boosted RATE-AF	1.0 (1.0–1.0)	1.0 (1.0–1.0)	1.0 (1.0–1.0)
AP (Accuracy Parity)	RATE-AF	1.0 (1.0–1.0)	0.8 (0.8–0.9)	0.9 (0.8–1.0)
	Boosted RATE-AF	1.0 (1.0–1.0)	1.0 (1.0–1.1)	1.0 (1.0–1.1)

Table 3. Coverage and fairness metrics under exploratory boosting. Coverage values below 1 indicate under-representation relative to the hypothetical balanced scenario, and values above 1 indicate over-representation. Fairness metrics–Equal Opportunity (EO), Predictive Rate Parity (PRP), and Accuracy Parity (AP)–were computed for prediction of the composite clinical outcome across ethnic groups relative to the White reference group. Values closer to 1 indicate parity in model performance across groups. Fairness values are reported as means with 95% confidence intervals across bootstrap iterations

Analytical utility followed a similar pattern (Figs. 3, 4). Treatment-effect estimates and significance patterns were preserved in most boosted RATE-AF datasets. Key outcomes, including NYHA and mEHRA improvement, NT-proBNP reduction, and cardiovascular event odds, maintained consistent direction and approximate magnitude of effect. Loss of significance was observed in the *White 60%* and *White 80%* scenarios, where the limited number of observations restricted regression estimation and increased uncertainty. In a few instances, the boosted datasets also showed shifts in treatment-effect estimates that diverged from the original null effect, particularly within the *White* subgroup under higher bias levels. For example, in the *White 60%* and *White 80%* configurations, odds ratios for outcomes such as mEHRA improvement fluctuated in different directions relative to baseline, a behaviour that emerged only when the remaining real-data sample for that subgroup became very small (approximately 71 or 41 patients), leading to unreliable estimations.

Although overall treatment effects were robustly replicated, subgroup-specific analyses for the *Asian* and *Black* boosted datasets could not be computed with confidence because of the very limited sample sizes, occasionally resulting in non-estimable parameters (Supplementary Figure S1), underscoring the difficulty of validating utility in very small subgroups. Overall, these findings indicate that boosting can recover subgroup representation and preserve analytical validity across a range of simulated bias conditions at the overall dataset level, provided that sufficient real information remains available for model learning. Model performance deteriorates only when information density falls below a critical threshold, as illustrated by the *White 80%* scenario, defining the practical limits of recovery under extreme small-sample conditions (e.g. using only 71 and 41 patients to train a probabilistic model).

Building on the controlled experiments, the framework was next applied in an exploratory setting to investigate the implications of boosting under a hypothetical balanced population scenario. Unlike the controlled setting, which evaluates recovery of subgroup representation within the original cohort after simulated removal, this exploratory configuration extends subgroup representation beyond that observed in RATE-AF and should therefore be interpreted as a scenario-based analysis rather than a directly validated real-world augmentation. In this configuration, conditional generation was guided by a hypothetical balanced composition with equal representation across ethnic groups, representing an exploratory scenario rather than a real-world population. The equal-representation target was used as a simple counterfactual benchmark to examine how substantial rebalancing might affect predictive fairness and analytical conclusions, and should not be interpreted as a population-representative target or as a universal definition of fairness. This enabled exploration of how extending subgroup representation beyond the distributions observed in the original RATE-AF cohort may affect fairness and predictive consistency across sensitive groups. The resulting boosted synthetic data achieved ideal coverage (coverage = 1.0 for all groups), indicating that the proposed framework can be configured to generate cohorts with predefined demographic compositions. Because synthetic boosting does not introduce new empirical information beyond that contained in the original trial, exploratory analyses based on heavily boosted synthetic cohorts may yield statistically significant findings that were not present in the observed data and should not be interpreted as clinical evidence. Instead, such findings may serve as model-based signals for further review by clinicians, with any clinical interpretation requiring careful expert assessment and validation against real-world data.

Fairness metrics also indicated improvement across sensitive groups. In the original RATE-AF data, Asian and Black participants were under-represented, and fairness estimates for the Black subgroup were unstable or not estimable for some metrics because of the extremely limited subgroup size. After boosting, Equal Opportunity, Predictive Rate Parity, and Accuracy Parity were all close to 1 across ethnic groups, indicating more consistent predictive behaviour in the rebalanced synthetic cohort. These results could be particularly relevant for researchers developing machine-learning models from clinical trial data, such as the composite clinical outcome prediction model examined in this study, where limited subgroup representation can introduce bias and reduce reliability in model outcomes. By generating synthetic cohorts that reflect a hypothetical balanced scenario, we illustrated how boosting could improve fairness in predictive models across sensitive groups and support equitable assessment of digital-health algorithms.

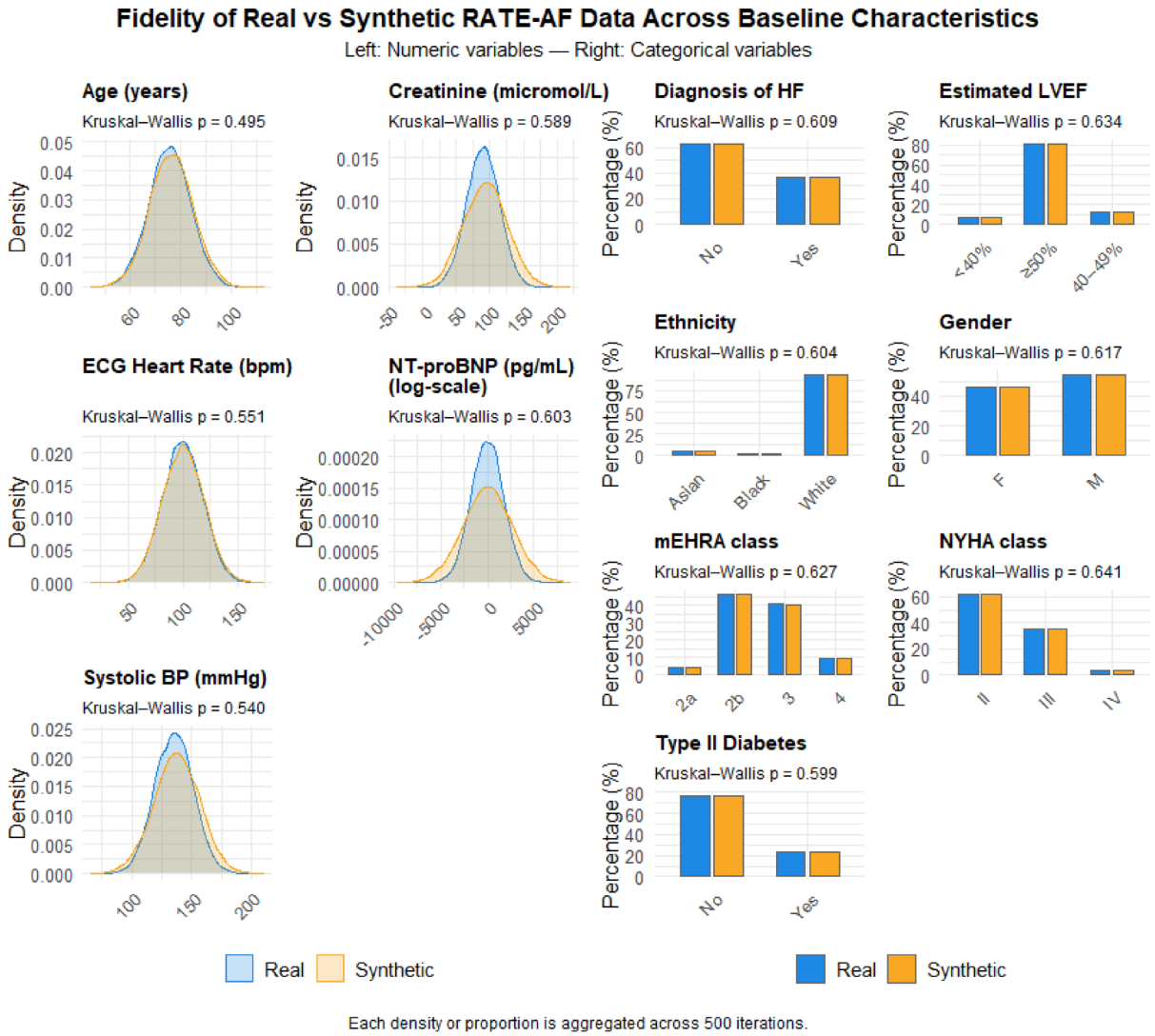


Fig. 5. Distributional fidelity of real vs synthetic RATE-AF data across baseline characteristics. Continuous variables (left) are shown as density plots and categorical variables (right) as grouped bar charts. Each density and proportion represents the average across 500 synthetic iterations. Kruskal–Wallis *p*-values quantify similarity between real and synthetic distributions. Category encodings for all variables are reported in Table 1. Baseline characteristics were closely reproduced, confirming high distributional fidelity of the replicated datasets.

Replication experiments demonstrated that synthetic data can replicate clinical trial datasets with high fidelity and utility. In the synthetic RATE-AF data, baseline characteristics were statistically similar compared to those in the real data (all Kruskal-Wallis $p > 0.5$), with continuous variables such as age, heart rate and systolic blood pressure showing nearly identical distributions and categorical variables reproduced with matching proportions (Fig. 5). Clinical outcomes were equally well preserved (Fig. 6). Synthetic data reproduced the direction and approximate magnitude of all treatment effects observed in the real trial, maintaining statistical significance for the same endpoints. Improvements in NYHA and mEHRA classes and reductions in NT-proBNP were consistently detected, while lower odds of adverse and cardiovascular events with Digoxin remained evident. Minor attenuation in continuous outcomes reflected expected smoothing from probabilistic sampling rather than structural distortion, confirming that the underlying clinical associations were maintained.

These findings indicate that synthetic cohorts can emulate the statistical behaviour of a clinical trial, generating virtual patient populations that mirror real trial data with statistical and analytical coherence.

Limitations and future work

Several limitations must be acknowledged when interpreting the findings of this study. Although controlled boosting experiments illustrated high overall fidelity in most simulated bias scenarios, these results require careful interpretation for subgroups with extremely limited representation. In scenarios where very small minority groups were boosted (e.g. Asian and Black patients), only a few records were removed from the original

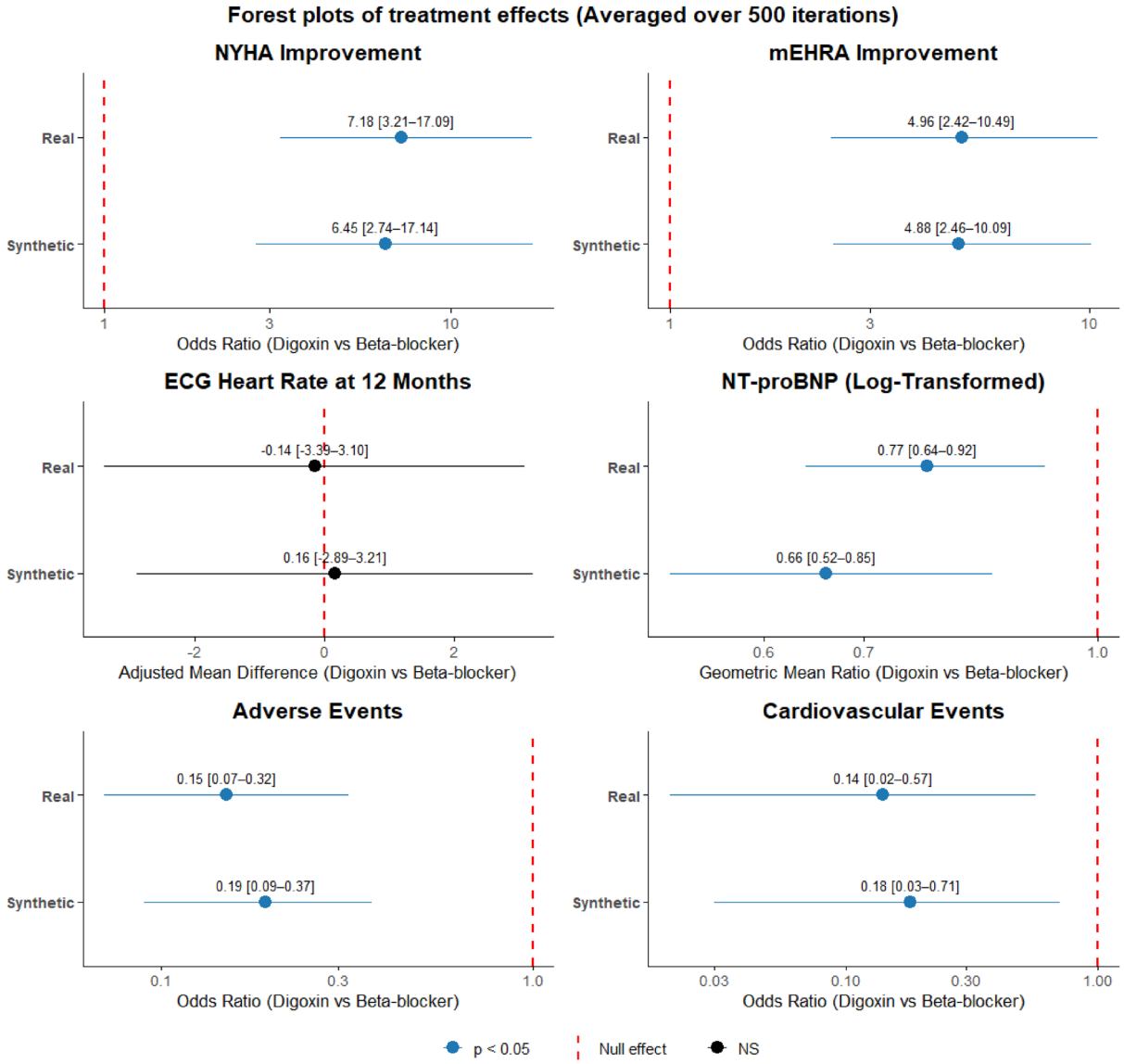


Fig. 6. Comparison of treatment effects across Real and Synthetic RATE-AF data. Forest plots display adjusted estimates for six key clinical outcomes at 12 months. Points indicate averaged adjusted treatment-effect estimates and horizontal lines their 95% confidence intervals, computed across 500 iterations. The vertical dashed line marks the null effect (OR = 1, RGM = 1, AMD = 0). Blue markers denote statistically significant effects ($p < 0.05$); black markers indicate non-significance.

dataset, leaving more than 150 observations to support model learning. This allowed the Bayesian network to maintain stable conditional estimates, but the learnt dependencies were necessarily dominated by the majority population. As a result, the model's ability to capture subgroup-specific structure is restricted when real-world evidence for these groups is extremely sparse, limiting the generalisability of subgroup-level inferences unless supplemented by external data. Therefore, in the absence of additional external data, the framework should not be used to infer subgroup-specific effects for extremely small populations, as synthetic augmentation in these cases reflects extrapolation from the majority population rather than evidence grounded in the subgroup itself.

Related to this, a further limitation arises when the remaining real-data sample for a subgroup becomes too small to support reliable model learning. In the simulated scenarios with highest bias (White 60% and 80%), where only 71 or 41 patients remained, the treatment-effect estimates became unstable and occasionally diverged from the original null effect. This indicates that synthetic boosting is only meaningful when a sufficient amount of real information is still available to inform subgroup structure, and that there exists a minimum-information threshold below which boosting becomes unreliable.

Extremely small subgroup sizes also constrained the evaluation of subgroups analytical utility. Treatment effects for Asian and Black participants could not be reliably estimated in the controlled setting, and parameters were occasionally non-estimable. Although exploratory boosting produced larger synthetic subgroups, these analyses lack real-world ground truth for validation. Since the RATE-AF trial did not include enough Asian

or Black participants to estimate subgroup-specific clinical responses, the alignment between synthetic and observed effects necessarily reflects information learnt from the majority population rather than evidence of genuine subgroup-level behaviour. We acknowledge that it is difficult to quantify the clinical value of subgroup effects in synthetic populations beyond what is supported by the original trial data, especially for groups represented by fewer than ten individuals in the real data. Accordingly, statistically significant results arising in heavily boosted synthetic subgroups should be interpreted with caution and regarded as hypothesis-generating signals for clinical review rather than evidence of subgroup-specific effects.

A similar limitation applies to fairness evaluation, where some metrics for the Black subgroup in the original RATE-AF data were unstable or not estimable because the subgroup contained too few observations to support reliable prediction-based assessment. In addition, fairness was evaluated without an independent hold-out set. Although repeated bootstrap evaluation is appropriate for a small dataset such as RATE-AF and provides an internal measure of stability, it does not constitute external validation. The reported fairness metrics should therefore be interpreted as exploratory estimates of within-dataset predictive parity rather than definitive evidence of fairness in unseen populations.

Furthermore, although synthetic replication preserved baseline characteristics, clinical outcomes, and treatment-effect estimates, these results should be interpreted strictly within the scope of the modelled variables and endpoints. Statistical fidelity does not guarantee broader clinical interpretability, and it remains uncertain whether synthetic cohorts capture meaningful behaviour beyond patterns encoded in the original trial. The evaluation strategy adopted here was intentionally tailored to the clinical-trial use case and the statistical analysis priorities of the RATE-AF study, and should therefore be interpreted as use-case-specific rather than as a general-purpose benchmarking of synthetic data quality.

The evaluation was also based on a single clinical trial with a modest sample size, which limits generalisability and constrains the assessment of extreme bias scenarios. Learning probabilistic structure from such limited evidence revealed occasional clinical inconsistences that required model review with the RATE-AF investigators, underscoring the importance of domain expertise and highlighting practical constraints of purely data-driven learning when sample sizes are small.

Finally, these findings should be interpreted within the context of the RATE-AF population. Participant selection, limited diversity, and homogeneous trial setting restrict the generalisability of the results. These limitations define the current scope of the framework and highlight opportunities for further methodological refinement and broader validation in different clinical settings.

Building on this study, future research will focus on integrating external real-world data sources to strengthen the underlying probabilistic model and address information gaps in the clinical trial. In particular, data from the Clinical Practice Research Datalink (CPRD)[42] will be used to update the Bayesian network model, incorporating conditional relationships that are poorly represented or absent in the RATE-AF data. By enriching the evidence base beyond a single trial, this integration will allow us to examine how external information influences inferences and whether it can support cautious extrapolation of findings to populations not adequately represented in the original study.

Further work is also needed to determine whether the synthetic cohorts generated by the framework carry meaningful clinical value beyond their statistical properties. Analyses by clinical experts will be required to understand the extent to which synthetic cohorts can inform clinical interpretation and where their role should remain purely analytical. In addition, future work should formally characterise the minimum-information threshold required for reliable boosting. Sensitivity analyses across multiple datasets and varying sample sizes will be essential to establish guidance on the size of real data needed to guarantee reliable boosting.

Finally, these investigations will help further clarify how synthetic data should be positioned within the clinical trials ecosystem. Synthetic data cannot replace the need for real clinical trials, particularly when assessing treatment effects in under-represented populations. However, boosted synthetic datasets may help identify where additional real-world studies or targeted sub-trials are warranted, especially if boosted data suggest signals or subgroup patterns that diverge from those observed in the original trial. Establishing the boundary between statistical insight and clinical evidence will be essential for defining an appropriate and responsible use of synthetic data in clinical trials.

At a broader level, an important question for future research is to quantify the extent to which synthetic data can meaningfully support digital health equity, ensuring that participation in evidence generation is not constrained by geography, recruitment barriers, or demographic imbalances.

Data availability

The data that support the findings of this study are available from the University of Birmingham but restrictions apply to the availability of these data, which were used under a data sharing agreement for the current study and are therefore not publicly available. Data are however available from the authors upon reasonable request and with permission of the University of Birmingham and the RATE-AF Steering Committee, which reviews requests on behalf of the trial sponsor. Release of any data will be subject to a data use agreement and may require additional approval from a Research Ethics Committee.

Code availability

All code required to reproduce the experiments in this study is available at https://github.com/barbaradraghi/synthetic-RATEAF. The repository includes complete R scripts for the controlled, exploratory, and replication experiments, together with the supporting utility functions and documentation. The *bnlearn* package (v4.8.1) is used for Bayesian Network modelling[43]. A schema-only illustrative dataset (`RATEAF-dummy.csv`) is included in the `data/` folder to indicate the expected input structure. Researchers with approved access to the

original RATE-AF clinical trial dataset (for example, RATE-AF.dta) can place the data in this folder and the scripts will run without modification.

Received: 13 December 2025; Accepted: 26 June 2026
Published online: 05 August 2026

References

1. Ziff, O. et al. Safety and efficacy of digoxin: Systematic review and meta-analysis of observational and controlled trial data. *BMJ* https://doi.org/10.1136/bmj.h4451 (2015).
2. Champsi, A. Digoxin in patients with heart failure and/or atrial fibrillation: A systematic review and meta-analysis of 5.9 million patient years of follow-up. In *SSRN 2023*. https://doi.org/10.2139/ssrn.4544769.
3. Wang, X. & Kattan, M. Cohort studies: Design, analysis, and reporting. *Chest* **158**, S72–S78. https://doi.org/10.1016/j.chest.2020.03.014 (2020).
4. Oertelt-Prigione, S. & Turner, B. Tackling biases in clinical trials to ensure diverse representation and effective outcomes. *Nat. Commun.* **15**, 1407. https://doi.org/10.1038/s41467-024-45718-w (2024).
5. Chakraborty, J., Majumder, S. & Menzies, T. Bias in machine learning software: Why? How? What to do? In *Proceedings of 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. https://doi.org/10.1145/3468264.3468537 (ACM, 2021).
6. Ferrara, E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Science* https://doi.org/10.3390/sci6010003 (2023).
7. Shahbazi, N., Lin, Y., Asudeh, A. & Jagadish, H. Representation bias in data: A survey on identification and resolution techniques. *ACM Comput. Surv.* **55**, 1–39. https://doi.org/10.1145/3588433 (2023).
8. De-Arteaga, M. *et al.* Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287572 (ACM, 2019).
9. Versavel, S. et al. Diversity, equity, and inclusion in clinical trials: A practical guide from the perspective of a trial sponsor. *Contemp. Clin. Trials* **126**, 107092. https://doi.org/10.1016/j.cct.2023.107092 (2023).
10. Ramamoorthy, A., Pacanowski, M., Bull, J. & Zhang, L. Racial/ethnic differences in drug disposition and response: Review of recently approved drugs. *Cli. Pharmacol. Ther.* **97**, 263–273. https://doi.org/10.1002/cpt.61 (2015).
11. Li, F. et al. Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *J. Biomed. Inform.* **138**, 104294. https://doi.org/10.1016/j.jbi.2023.104294 (2023).
12. Taskforce on Innovation, Growth and Regulatory Reform. Independent report: Taskforce on innovation, growth and regulatory reform (tigrr). Tech. Rep., institutionHM Government (2021).
13. Jordon, J. *et al.* Synthetic data—What, why and how? Preprint at arXiv:2205.03257 (2022).
14. Espinosa, E. & Figueira, A. On the quality of synthetic generated tabular data. *Mathematics* https://doi.org/10.3390/math11153278 (2023).
15. Murtaza, H. et al. Synthetic data generation: State of the art in health care domain. *Comput. Sci. Rev.* **48**, 100546. https://doi.org/10.1016/j.cosrev.2023.100546 (2023).
16. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R. & Rankin, D. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* **493**, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053 (2022).
17. El Emam, K., Mosquera, L., Fang, X. & El-Hussuna, A. An evaluation of the replicability of analyses using synthetic health data. *Sci. Rep.* **14**, 57207. https://doi.org/10.1038/s41598-024-57207-7 (2024).
18. Wang, Z., Myles, P. & Tucker, A. Generating and evaluating synthetic UK primary care data: Preserving data utility & patient privacy. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, 126–131. https://doi.org/10.1109/CBMS.2019.00036 (organizationIEEE, 2019).
19. Tucker, A., Wang, Z., Rotalinti, Y. & Myles, P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digit. Med.* **3**, 1–13. https://doi.org/10.1038/s41746-020-00353-9 (2020).
20. Petrakos, N. Z., Moodie, E. E. M. & Savy, N. A framework for generating realistic synthetic tabular data in a randomized controlled trial setting. arXiv:2501.17719 (2025).
21. Chandra, N. K., Sarkar, A., de Groot, J. F., Yuan, Y. & Müller, P. Bayesian nonparametric common atoms regression for generating synthetic controls in clinical trials. arXiv:2201.00068 (2023).
22. Azizi, Z., Zheng, C., Mosquera, L., Pilote, L. & El Emam, K. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* https://doi.org/10.1136/bmjopen-2020-043497 (2021).
23. Xu, D., Yuan, S., Zhang, L. & Wu, X. Fairgan: Fairness-aware generative adversarial networks. arXiv:1805.11202 (2018).
24. van Breugel, B., Kyono, T., Berrevoets, J. & van der Schaar, M. Decaf: generating fair synthetic data using causally-aware generative networks. In *Proceedings of 35th International Conference on Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.5555/3540261.3541963 (Curran Associates Inc., addressRed Hook, NY, USA, 2021).
25. Hameed, M., Qureshi, A. & Kaushik, A. Bias mitigation via synthetic data generation: A review. *Electronics* https://doi.org/10.3390/electronics13193909 (2024).
26. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357. https://doi.org/10.1613/jair.953 (2002).
27. He, H., Bai, Y., Garcia, E. A. & Li, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969 (2008).
28. Draghi, B., Wang, Z., Myles, P. & Tucker, A. Identifying and handling data bias within primary healthcare data using synthetic data generators. *Heliyon* **10**, e24164. https://doi.org/10.1016/j.heliyon.2024.e24164 (2024).
29. Wang, Z., Draghi, B., Rotalinti, Y., Lunn, D. & Myles, P. High-fidelity synthetic data applications for data augmentation. In Domínguez-Morales, M., Civit-Masot, J., Muñoz-Saavedra, L. & Damaševičius, R. (eds.) *Deep Learning*, chap. 7. https://doi.org/10.5772/intechopen.113884 (IntechOpen, 2024).
30. Juwara, L., El-Hussuna, A. & El Emam, K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns* **5**, 100946. https://doi.org/10.1016/j.patter.2024.100946 (2024).
31. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**, 115. https://doi.org/10.1145/3457607 (2021).
32. Asudeh, A., Jin, Z. & Jagadish, H. V. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 554–565. https://doi.org/10.1109/ICDE.2019.00056 (2019).
33. Kotecha, D. et al. A review of rate control in atrial fibrillation, and the rationale and protocol for the rate-AF trial. *BMJ Open* **7**, e015099. https://doi.org/10.1136/bmjopen-2016-015099 (2017).
34. Jones, J. et al. Importance and assessment of quality of life in symptomatic permanent atrial fibrillation: Patient focus groups from the rate-AF trial. *Cardiology* **145**, 666–675. https://doi.org/10.1159/000511048 (2020).
35. Bunting, K. V., Stanbury, M., Tica, O. & Kotecha, D. Transforming clinical research by involving and empowering patients-the rate-AF randomized trial. *Eur. Heart J.* **42**, 2411–2414. https://doi.org/10.1093/eurheartj/ehab098 (2021).

36. Kotecha, D. et al. Effect of digoxin vs bisoprolol for heart rate control in atrial fibrillation on patient-reported quality of life: The rate-AF randomized clinical trial. *JAMA* **324**, 2497–2508. https://doi.org/10.1001/jama.2020.23138 (2020).
37. Abdali, Z. et al. Cost-effectiveness of digoxin versus beta blockers in permanent atrial fibrillation: The rate control therapy evaluation in permanent atrial fibrillation (rate-AF) randomised trial. *Heart* **111**, 362–369. https://doi.org/10.1136/heartjnl-2024-324761 (2025).
38. Office for National Statistics. Ethnicity and national identity in england and wales: 2011. (2012).
39. Verma, S. & Rubin, J. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7. https://doi.org/10.1145/3194770.3194776 (2018).
40. Caton, S. & Haas, C. Fairness in machine learning: A survey. *ACM Comput. Surv.* **56**, 166. https://doi.org/10.1145/3616865 (2024).
41. Anahideh, H., Nezami, N. & Asudeh, A. Finding representative group fairness metrics using correlation estimations. *Expert Syst. Appl.* **262**, 125652. https://doi.org/10.1016/j.eswa.2024.125652 (2025).
42. Clinical Practice Research Datalink (CPRD). https://www.cprd.com/.
43. Scutari, M. Learning bayesian networks with the bnlearn r package. *J. Stat. Softw.* **35**, 1–22. https://doi.org/10.18637/jss.v035.i03 (2010).

Acknowledgements

We thank the RPF Grant Steering Group for their guidance and support throughout this study. We particularly acknowledge the contributions of Dr. David Brown, Prof. Feng Dong, Prof. Kazem Rahimi, Dr. Rebecca Ghosh and Dr. Xiaoxia Wang who served as members of the Steering Committee. Their oversight was invaluable in ensuring the rigour and relevance of this work. We also gratefully acknowledge the input of our public representative, Sandra Haynes, Mary Stanbury and Jaqueline Jones for their critical role in aligning this research with patient perspectives.

Author contributions

B.D. and A.T. designed the study, with B.D. developing the code to generate and evaluate synthetic data, and D.A. supporting the code development. P.M. and R.B. provided overall project guidance. D.K., A.C., and M.C. provided clinical expertise, developed the statistical analysis, reviewed the results, and contributed to the description and interpretation of findings. A.R.M. and K.V.B. providing governance overview and data extraction. B.D. wrote the manuscript, and all authors reviewed the manuscript and approved the final version.

Funding

This work was funded by the Regulators' Pioneer Fund (RPF) grant call number 3 (Using High-fidelity Synthetic Data as synthetic control arms and to boost sample sizes in clinical trials). The RATE-AF trial was funded by the National Institute for Health and Care Research (NIHR; CDF-2015-08-074). The authors were additionally supported by the British Heart Foundation/University of Birmingham Accelerator Award (AA/18/2/34218), the NIHR Birmingham Biomedical Research Centre (NIHR203326), the Medical Research Council Health Data Research UK (HDRUK/CFC/01) and the Cook & Wolstenholme Charitable Trust.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

Anonymised clinical trial data from the RATE-AF study were made available to the authors through a data sharing agreement with the University of Birmingham. All data were handled in line with the original ethical approvals and consent. RATE-AF was approved by the East Midlands–Derby research ethics committee (16/EM/0178) and Health Research Authority (IRAS 191437), funded by the UK National Institute for Health and Care Research (CDF-2015-08-074), and registered with clinicaltrials.gov (NCT02391337) and clinicaltrialsregister.eu (2015-005043-13).

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-60252-z.

Correspondence and requests for materials should be addressed to B.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© Crown 2026