Evaluation of Machine Learning-Based Propensity Score Estimation: A Benchmarking Observational Analysis Against a Randomized Trial

Kaicheng Wang [1], Lindsey Rosman [2], Haidong Lu [3]

[1] Yale Center for Analytical Sciences, Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut

[2] Division of Cardiology, Department of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

[3] Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut

Corresponding Author

Kaicheng Wang, MD MPH

300 George St STE 511

New Haven, CT 06510

kaicheng.wang@yale.edu

Abstract

Machine learning (ML) approaches for propensity score estimation are increasingly used with the expectation of improving covariate balance and reducing bias, but their validity in selecting appropriate confounders remains controversial. In this study, we estimated the effectiveness of sacubitril/valsartan versus angiotensin-converting enzyme inhibitor and angiotensin receptor blocker on all-cause mortality among heart failure patients with implantable cardioverter defibrillators in the U.S. Department of Veterans Affairs from 2016 to 2020. We compared results from traditional logistic regression- and ML-based propensity score methods and benchmarked them against the PARADIGM-HF randomized trial. The estimate from logistic regression with *a priori* confounder selection (HR = 0.93, 95% CI 0.61 - 1.42; 27-month RR = 0.87, 95% CI 0.59 - 1.21) most closely aligned with the trial result (HR = 0.81; 95% CI 0.61 - 1.06). In contrast, generalized boosting models did not outperform traditional logistic regression, and may amplify bias when combined with a data-driven confounder selection (HR = 0.63, 95% CI 0.31 - 1.30; RR = 0.61, 95% CI 0.33 - 1.04). Our findings suggest that ML-based propensity scores may introduce overadjustment bias and underscore the importance of subject-matter knowledge in causal inference with high-dimensional real-world data.

INTRODUCTION

Randomized trials are the gold standard for evaluating the efficacy of an intervention. However, many causal questions cannot be feasibly addressed through randomized trials. Fortunately, observational studies using target trial emulation framework[1] have demonstrated comparable results[2, 3], and provided timely and potentially more generalizable evidence to inform clinical practice. These studies often use propensity score matching or inverse probability weighting to balance measured confounders between treatment groups, thereby approximating baseline randomization. Nonetheless, these methods are subject to biases.[4] One critical assumption for valid causal effect estimation is conditional exchangeability[5], meaning that the potential outcome is independent of the treatment assignment, given measured confounders. Although the assumption of no unmeasured confounding cannot be guaranteed without randomization, conditional exchangeability can be approximately achieved by including as many potential confounders as possible in the propensity model.

In light of the challenges posed by confounding and high-dimensional data, machine learning (ML) approaches have gained popularity in recent literature for propensity score estimation.[6-9] ML algorithms are expected to improve covariate balance between exposure groups and reduce bias.[10] Indeed, algorithms, such as random forest, have appealing advantages in feature selection and the ability to capture non-linear relationship and complex interactions without imposing parametric assumptions. These properties often lead to better discrimination and calibration in predictive models compared to traditional regression-based methods, extrapolating arguments for their promising use in propensity score estimation[11], especially with high-dimensional electronic health record (EHR) data.

Yet, the application of ML in propensity score estimation remains controversial. ML algorithms use loss functions to prioritize predictive accuracy which may lead to overfitting rather than bias reduction[12, 13]. Furthermore, data-driven selection of adjustment variables, without explicit control over confounder selection based on subject-matter knowledge, can result in

overadjustment bias[14] by inadvertently including colliders, instrumental variables, or other covariates that should not be adjusted for in causal effect estimation.

Although previous studies have evaluated the performance of ML-based propensity score estimation, most have relied on simulation rather than real-world examples due to the absence of known ground truths necessary to directly quantify bias.[10, 15, 16] However, the benchmarking approach[17-19] offers an opportunity to determine whether observational analyses yield results consistent with those from an existing index randomized trial addressing the same research question. Specifically, in the context of propensity score estimation, benchmarking can help assess, in real-world settings, whether results obtained using ML-based methods more closely align with, or departure from, the findings of an index trial compared to those from parametric regression models.

Here we conducted observational analyses through target trial emulation and leveraged the benchmarking to assess the validity of ML-based method to estimate propensity scores. Specifically, we estimated the effectiveness of sacubitril/valsartan vs. angiotensin-converting enzyme inhibitor and angiotensin receptor blocker (ACEI/ARB) on all-cause mortality among heart failure (HF) patients with implantable cardioverter defibrillators (ICDs) in the U.S. Department of Veterans Affairs (VA) using traditional logistic regression-based and ML-based propensity score methods, and benchmarked these estimates against the result from an existing index randomized trial.

METHODS

Index Randomized Trial

PARADIGM-HF[20] was a multicenter, randomized, double-blind trial comparing the efficacy of sacubitril/valsartan vs. enalapril on the risk of death or HF hospitalization among patients with chronic HF (NCT01035255). A secondary analysis among 1,243 patients with an ICD or cardiac resynchronization therapy defibrillator (CRT-D) for primary prevention of sudden cardiac death

reported that sacubitril/valsartan was associated with a 19% lower risk of all-cause mortality compared to enalapril (HR = 0.81; 95%CI 0.61 - 1.06).[21]

Target Trial Specification and Observational Emulation

Study Design

We specified a hypothetical trial similar to the PARADIGM-HF trial and emulated the target trial using data from the VA National Cardiac Device Surveillance Program (NCDSP) and Corporate Data Warehouse (CDW) among veterans with an ICD or CRT-D who were prescribed sacubitril/valsartan or any ACEIs/ARBs between January 1, 2016 and September 30, 2020. This study was approved by the Research & Development Committee of the VA Connecticut Healthcare System and a waiver of informed consent was granted (IRBNet# 1582934). Key target trial protocol components are summarized in *Supplemental Table 1*.

Eligibility Criteria

We included patients aged 18 years and older diagnosed with heart failure using International Classification of Diseases 9th and 10th versions codes, with a left ventricular ejection fraction (LVEF) ⩽ 35% and implanted with an ICD or CRT-D between January 1, 2010 and September 30, 2020. Patients must have a B-type natriuretic peptide (BNP) ⩾ 150 pg/mL or N-terminal pro B-type natriuretic peptide (NT-proBNP)⩾ 600 pg/mL, and received at least one guideline-directed medical therapy, including β-blocks, mineralocorticoid receptor antagonists (MRAs) and sodium-glucose transport protein 2 inhibitors (SGLT2i). Patients with left ventricular assist devices or had a history of heart transplant were excluded.

Treatment Strategies

We identified sacubitril/valsartan and ACEIs/ARBs prescriptions from CDW outpatient pharmacy domain. The baseline index date (i.e., time zero) for sacubitril/valsartan was the first prescription of sacubitril/valsartan. Similarly, the baseline index date for ACEI/ARB was the first fill in the study period after all eligibility criteria were met. Adherence was defined as receiving a refill within 30 days after the end of supply. The 2016 guideline[22] recommended replacing

ACEIs/ARBs with an angiotensin receptor/neprilysin inhibitor (ARNI, sacubitril/valsartan) to further reduce morbidity and mortality, with de novo treatment with ARNI recommended since 2022[23]. Most patients in this study were initially prescribed an ACEI/ARB and later switched to sacubitril/valsartan. To account for potential selection bias from prevalent user[24] and to be consistent with the run-in period design from the trial, prevalent users were defined as patients who had received an ACEI/ARB in the past 180 days when switching to sacubitril/valsartan or refilling an ACEI/ARB.[25]

Outcome and Follow-up

The primary outcome was all-cause mortality, identified from the VA Death ascertainment file, passively surveilled through VA Master Person Index[26], administrative data, and Social Security Administration, with a 98.3% sensitivity and 97.6% exact agreement with National Death Index.[27] Therefore, we assumed no misclassification of the outcome and no loss to follow-up. Patients were followed from the baseline until the date of death, 51 months after baseline, or administrative end of the study by September 30, 2020, whichever happened first.

Potential Confounders Based on Subject-Matter Knowledge

Variable definitions are included in *Supplemental Table 2.* We used *a priori* [28] approach to identify potential confounders based on subject-matter knowledge and previous literature. We assumed treatment strategies were exchangeable conditional on baseline confounders, including trial sequence number, demographics (age, gender, race, ethnicity), social determinants of health (cigarette use, area deprivation index [ADI]), healthcare-seeking behavior (flu vaccination or HF hospitalization in the past 12 months), clinical characteristics (body mass index [BMI], systolic blood pressure [SBP], LVEF, estimated glomerular filtration rate [eGFR], serum/plasma potassium, diagnosis of atrial fibrillation, cardiomyopathy, chronic kidney disease [CKD], chronic obstructive pulmonary disease [COPD[, diabetes, hypertension, depression, history of myocardial infraction [MI], peripheral artery disease [PAD] and valvular heart disease [VHD]), prescriptions (β-blocker, MRA, SGLT2i, diuretics and digoxin), CRT, user type (new vs. prevalent user) and time since HF diagnosis.

Additional Variables for a Data-Driven Approach

Additionally, to mimic a data-driven approach for confounder selection, we included predictors not meeting the definition of a confounder, but predicting treatment assignment or the death outcome. These variables (C^*) included residence (urban vs. rural), employment status, diagnoses such as anemia, cerebral vascular conditions, deep vein thrombosis and pulmonary embolism (DVT/PE), dementia, dyslipidemia, influenza, non-pathological fracture, malignancy and obstructive sleep apnea (OSA), laboratory measures (hemoglobin and albumin), and time since device implantation.

Statistical Analyses

Baseline characteristics were summarized as frequencies and percentages, means and standardized deviations (SDs), or medians and interquartile ranges (IQRs) for non-normally distributed variables. Missing data for continuous covariates (ADI, hemoglobin, serum/plasma potassium and albumin) were imputed with median values, and categorical variables with missing data were coded as "Unknown". Data were managed using SQL Server Management Studio 19.0, and statistical analyses were conducted using R version 4.4.1.

We performed an intention-to-treat analysis examining the effectiveness of sacubitril/valsartan vs. ACEI/ARB on all-cause mortality. Observational analog of intention-to-effect was estimated by adjusting for baseline confounders using stabilized inverse probability of treatment weights (IPTW). Propensity score was estimated using logistic regression. Covariate balance was assessed using standardized mean difference (SMD), with a value less than 0.1 indicating good balance between treatment groups. Cumulative incidence curves were estimated using IPTW-weighted pooled logistic regression models, with an indicator for treatment strategy, a flexible time-varying intercept, and a treatment-by-time interaction term. Since the median follow-up in PARADIGM-HF trial was 27 months, we compared 27-month risk differences (RDs) and risk ratios (RRs). Confidence intervals (CIs) for cumulative incidence rates, RDs, and RRs were computed nonparametrically using 500 bootstrapping samples. Hazard ratios (HRs) were obtained using

pooled logistic regression models without the interaction term, and the corresponding CIs were estimated through the robust variance estimator.

Next, we estimated the propensity score with generalized boosted model (GBM)[29, 30], a commonly used machine learning algorithm, with the same set of *a priori* confounders. The model began with a shrinkage parameter of 0.1 and 1,000 iterations. Hyperparameters, including shrinkage (0.3, 0.1, 0.05, 0.01 and 0.005), interaction.depth (2, 3, 5 and 7) and n.minobsinnode (10, 20, 30, 50, 75, 100 and 150) were tuned based on 10-fold cross-validation errors. In addition, to explore the impact of data-driven confounder selection approach, we repeated the analysis with additional predictors in the GBM.

Finally, we informally benchmarked the estimates from the target trial emulation using different propensity score estimation methods against the intention-to-treat result from the secondary analysis[21] of the PARADIGM-HF trial, and assessed whether the GBM-based propensity score estimation yielded results more closely aligned with the trial findings.

RESULTS

We identified a study sample of 583 patients prescribed sacubitril/valsartan and 1,864 patients prescribed an ACEI/ARB between January 1, 2016 and September 30, 2020 (*Table 1 & Figure 1*). Compared to those receiving ACEI/ARB, the sacubitril/valsartan group was younger (mean [SD], 68.6 [8.6] vs. 70.2 [8.2] years), had a lower proportion of White individuals (68.3% vs. 74.0%). The sacubitril/valsartan group had a higher proportion of patients diagnosed with cardiomyopathy (84.4% vs. 78.5%), depression (40.5% vs. 36.4%), and prescribed MRAs (48.9% vs. 35.6%), diuretics (68.4% vs. 60.4%) and digoxin (11.5% vs. 9.0%), and a lower proportion of patients with CRT-D (44.6% vs. 49.6%), any HF hospitalizations (44.6% vs. 54.8%) or influenza vaccine (71.7% vs. 75.3%) in the past 12 months, compared to the ACEI/ARB group. Additionally, patients receiving sacubitril/valsartan had longer duration since the initial diagnose of HF (median [IQR], 80.0 [38.5 – 147.5] vs. 62.0 [14.0-125.0] months) and higher proportion of

prevalent users (93.1% vs. 83.6%). The median [IQR] follow-up time was 16.0 [8.0 – 25.0] months for the sacubitril/valsartan group and 27.0 [14.0 – 41.0] months in the ACEI/ARB group.

Estimates From Logistic Regression-Based Propensity Score and A Priori Confounder Selection

The distribution of propensity scores is visualized in *Figure 1A*. After applying the IPTW, baseline covariates were well balanced (*Supplemental Figure 1*). The 27-month cumulative incidence of all-cause mortality (*Figure 1B*) was 9.3% (95% CI, 6.2 – 12.9%) in the sacubitril/valsartan group compared to 10.7% (95% CI, 9.2 – 12.2%) in the ACEI/ARB group, leading to an RD of -1.4% (95% CI, -4.6 – 2.2%), an RR of 0.87 (95% CI, 0.59 – 1.21), and an average HR of 0.93 (95% CI, 0.61 – 1.42) during the entire follow-up period (*Table 2*).

Estimates From GBM-Based Propensity Score and A Priori Confounder Selection

The distribution of GBM-based propensity scores is visualized in *Figure 1C.* After applying the weights, the SMDs were greater than 0.1 for cardiomyopathy (0.113), diuretics (0.112) and user type (0.129). Variable importance is visualized in *Supplemental Figure 2A*. Gender, ethnicity, baseline hypertension and prescription of β-blockers and SGLT2i were not used in any boosted trees. The 27-month cumulative incidence of all-cause mortality (*Figure 1D*) was 10.3% (95% CI, 9.1 – 18.9%) in the sacubitril/valsartan group vs. 10.7% (95% CI, 8.8 – 12.1%) in the ACEI/ARB group, corresponding to an RD of -0.4% (95% CI, -1.4 – 9.1%), an RR of 0.96 (95% CI, 0.87 – 1.93), and an overall HR of 0.97 (95% CI, 0.68- 1.37). After additionally adjusted for cardiomyopathy, diuretic prescriptions and user type in the weighted pooled logistic regression model, the RR (0.95; 95% CI, 0.92 – 1.97) and HR (0.97; 95% CI 0.68 - 1.38) did not change substantially.

Estimates From GBM-Based Propensity Score and a Data-Driven Confounder Selection

After including additional predictors in the GBM for a data-driven approach, the histogram of propensity scores demonstrated poor overlapping (*Figure 1E*). The median [IQR] predicted probability of receiving sacubitril/valsartan was 0.934 [0.629 - 0.996] in the sacubitril/valsartan group and 0.042 [0.022 – 0.087] in the ACEI/ARB group. Among these additional predictors,

time since device implant, employment status, residency, hemoglobin and albumin levels, and diagnosis of anemia, malignancy, cerebral vascular conditions, OSA and non-pathological fractures were deemed predictive of treatment assignment by the algorithm (*Supplemental Figure 2B*). The 27-month RR (*Figure 1F*) was 0.61 (95% CI, 0.33 - 1.04) with an overall HR of 0.63 (95% CI, 0.31 - 1.30).

Informal Benchmarking

Compared to PARADIGM-HF participants, our sample had higher risk profiles. They were older (mean [SD]: 69.8 [8.3] vs. 64.6 [9.9]), had fewer females (1.6% vs. 13.5%), more Black patients (20.1% vs. 6.8%), lower baseline LVEF (26.6 [6.1] vs. 27.3 [6.3]), higher prevalence of hypertension (97.3% vs. 67.5%), and lower use of guideline-directed medical therapy such as β-blockers (92.4% vs. 96.6%) and MRAs (38.7% vs. 56.5%).

The point estimation and CI from the logistic regression-based propensity score analysis (HR = 0.93, 95% CI 0.61 - 1.42; RR = 0.87, 95% CI 0.59 - 1.21) aligned most closely with the trial result (HR = 0.81; 95% CI 0.61 - 1.06), while point estimates from a GBM-based model using the same set of *a priori* confounders were closer to the null (HR = 0.97, 95% CI 0.68 - 1.37; RR = 0.92, 95% CI 0.74 - 1.68). When using a data-driven GBM-based approach, which incorporated non-confounders predictive of treatment assignment (e.g., employment status, time since implantation) or empirically associated with mortality (e.g., albumin, hemoglobin), and the estimate was biased toward the alternative (HR = 0.63, 95% CI 0.31 - 1.30; RR = 0.61, 95% CI 0.33 - 1.04).

DISCUSSION

In this observational study, we compared the effectiveness of sacubitril/valsartan vs. ACEI/ARB among HF patients with ICDs within the VA between January 1, 2016 and September 30, 2020, using traditional logistic regression-based and ML-based propensity score methods. We benchmarked these estimates against the result from the PARADIGM-HF randomized trial. The

estimate derived from *a priori* confounder selection strategy and logistic regression-based propensity score aligned most closely with the trial result. In contrast, estimates from GBM-based propensity score models were not superior to the traditional method. Notably, when the feature selection function of GBM was used to identify potential confounders, the algorithm incorporated non-confounders, and the resulting estimate was biased toward the alternative.

A good predictive model aims to maximize discrimination between events and non-events. In theory, if an ML model almost perfectly predicted intervention or placebo assignment with an area under the curve of 0.999, the propensity score distributions for each treatment group would be nearly no overlapping (as seen in *Figure 2E*). This likely violates the positivity assumption[31-33], which requires that all patients have a non-zero probability to receive either treatment at every level of the confounders. Such divergent propensity score distribution will make it difficult to find matches or can lead to extreme weights, ultimately resulting in biased effect estimates and inflated variance. To this end, while ML models such as GBM often achieved superior predictive performance[34, 35] due to their flexibility in modeling non-linearities and high-order interactions[36], it does not translate into less biased causal estimates.

Additionally, using feature selection functions in ML is problematic to identify confounders with high-dimensional data. In unsupervised or black-box ML models, investigators often had little control over which covariates are selected. Confounders that are weak predictors of treatment may be omitted, resulting in residual confounding. For instance, in our case, β-blockers and SGLT-2i, both Class 1A[23] recommended therapies known to reduce HF hospitalizations and mortality, were excluded from the ML-based propensity model, possible due to their prevalence.

Furthermore, two potential overadjustment biases[14] could be identified in the data-driven GBM model (*Figure 3*): (A). overadjustment for a collider (albumin level) and (B). overadjustment for a near-instrumental variable (time since implantation). Proteinuria, an indication for initiating ACEI/ARB, can lead to hypoalbuminemia, while systemic inflammation and other disease

processes affecting albumin are independently associated with mortality.[37] Information on proteinuria, inflammation (unmeasured confounder) and underlying causes of low albumin (unknown confounder) is limited in the EHR data. Consequently, adjusting for the collider albumin C_1^* opens a bias pathway ($E \leftarrow U_1 \rightarrow C_1^* \leftarrow U_2 \rightarrow D$) between treatment E and death D. Meanwhile, clinical decisions regarding HF management, including pharmacological augmentation with sacubitril/valsartan, are predominantly based on HF severity and responses to prior therapies, such as ICD or CRT-D devices which are reserved for HF patients with reduced ejection fraction (HFrEF) who did not respond to standard guideline-directed medical therapy. As a result, as shown in *Figure 3B*, time since implantation C_2^* may serve as an imperfect proxy for unmeasured HF severity U_3, with no expected causal effect of C_2^* on E. Despite this, the GBM algorithm identified time since implantation as the strongest predictor of sacubitril/valsartan and used it to estimate the propensity score. Using this predicted probability as the propensity score imposes causal effects of $C_2^* \rightarrow E$, thereby rendering it a shortcut feature[38, 39] as described in the ML literature. This substitution could potentially introduce a noncausal pathway if an unadjusted variable U_4, whether unmeasured or unknown, exists as a common cause of both C_2^* and E, may also constitute overadjustment for collider. Moreover, the strong predictability of C_2^* raises concerns about its behavior resembling that of a near-instrumental variable[40, 41], and the bias amplified through overadjustment for a near-instrumental variable and potentially overadjustment for a collider may outweighs the precision gained by including this variable through its weak association to the outcome. These forms of overadjustment biases may also explain why some prior studies, often simulation-based or secondary analysis of randomized trials, found better performance with GBM and deep learning, as those studies either restricted to pre-specified confounding [42-44] or ignore the scenarios in the presence of overadjustment. It also highlights that, while ML algorithms can help mitigate concerns about statistical model misspecification (i.e., incorrect functional forms and relationship between exposure, covariates and outcome), they cannot automatically address causal model misspecification (i.e., identify a sufficient set of confounders and avoid different types of overadjustment biases), as causal model specification requires subject-matter knowledge.

This study has several limitations. First, we only conducted informal benchmarking without transportability analyses from the trial population to the VA study sample. While the PARADIGM-HF trial reported a 19% reduction in all-cause mortality with sacubitril/valsartan vs. an ACEI (enalapril) among HF patients with ICDs, this efficacy may not be directly transportable due to heterogeneity of covariates distribution that may modify the treatment effect. However, without access to individual-level trial data, we could not formally transport the trial result to our study sample.[45] Additionally, the New York Heart Association (NYHA) Functional Classifications, a key eligibility criteria in the trial, was not captured in our EHR data. Neither could we differentiate the indications for ICD implantation between primary (prevent sudden cardiac death in HFrEF patients) vs. secondary prevention (patients with prior sustained ventricular arrhythmia), though secondary prevention indications (e.g., channelopathies) would be rare in the VA due to military service exclusions. Furthermore, BNP and NT-proBNP were not routinely measured, and their ordering itself may indicate decompensation, limiting our ability to replicate the trial population head-to-head. Lastly, the magnitude of the intention-to-treat effect may vary substantially depending on the degree of treatment adherence. While 82.4% of participants randomized to the sacubitril/valsartan arm remained on the assigned treatment and completed the trial over a median of 27 months[46], only 52.1% of patients in our sample adhered to sacubitril/valsartan therapy throughout the follow-up. However, given the high comorbidity burden and lower treatment adherence in our sample, we expected the real-world effectiveness of sacubitril/valsartan to be attenuated compared to the efficacy observed in the trial. This also suggests that comparing the adherence-adjusted per-protocol effects between the trial and the observational emulation may be more appropriate for benchmarking.[47] Further, our analysis did not aim to compare the performance of different ML algorithms. Rather, we used GBM, the most commonly used ML methods in the literature, to illustrate the paradoxical goal of predictive vs. causal models, and the source of potential biases when propensity models were misunderstood as predictive models.

In summary, this empirical benchmark study found that ML-based propensity score estimation do not outperform logistic regression-based propensity score method with *a priori* confounder selection. In fact, a data-driven approach that uses ML-based feature selection to construct propensity scores may result in residual confounding and potential overadjustment bias. When the study aim is causal rather than predictive, logistic regression with flexible terms of continuous confounders (e.g., splines) may mitigate model misspecification due to non-linearity. If ML is used to explore unknown predictors of treatment in high-dimensional data, confounder selection should be informed by subject-matter knowledge and accompanied by diagnostics (e.g., propensity score distribution, balance assessments). Incorporating doubly robust ML estimator coupled with cross fitting could safeguard against bias if one of the component is correctly specified, and partial dependent plots between covariates and predicted outcome could help identify interactions.[48] Lastly, probabilistic bias analysis should be considered to quantify bias from residual confounding. After all, ML is not a panacea.

Funding
Dr. Rosman's effort was sponsored by grants from the National Heart, Lung, and Blood Institute (K23HL141644). The remaining authors received no specific financial support.

Conflict of Interest
LR and KW are listed as inventors on a patent for a machine-learning-based risk prediction model using data from implantable or wearable cardiac devices (U.S. Provisional Patent Application 63/684842). LR has received research grants from Boston Scientific and served on a Biotronik advisory board. She also reports consultancy fees from Pfizer and Biotronik. HL has no conflict of interest to declare.

Disclaimer
The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; the U.S. Department of Health and Human Services; or the U.S. Department of Veterans Affairs.

Data Availability

The data used for this study cannot be shared publicly due to privacy regulations, but are available with an approved study protocol by the U.S. Department of Veterans Affairs (https://www.virec.research.va.gov/). All SQL queries and analytical code are available at https://github.com/kw375/ML_PS.

Reference

1. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*. 2016;183(8):758-764.
2. Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Avoidable flaws in observational analyses: an application to statins and cancer. *Nature medicine*. 2019;25(10):1601-1606.
3. Zuo H, Yu L, Campbell SM, Yamamoto SS, Yuan Y. The implementation of target trial emulation for causal inference: a scoping review. *Journal of Clinical Epidemiology*. 2023;162:29-37.
4. Hubbard RA, Gatsonis CA, Hogan JW, Hunter DJ, Normand S-LT, Troxel AB. Target trial emulation" for observational studies—potential and pitfalls. *N Engl J Med*. 2024;391(21):1975-1977.
5. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*. 2008;168(6):656-664.
6. Yang S, Orlova Y, Park H, et al. Cardiovascular Safety of Anti-CGRP Monoclonal Antibodies in Older Adults or Adults With Disability With Migraine. *JAMA neurology*. 2025;82(2):132-141.
7. O'Brien WJ, Schweizer ML, Strymish J, et al. Propensity score-weighted analysis of postoperative infection in patients with and without preoperative urine culture. *JAMA network open*. 2024;7(3):e240900-e240900.
8. Brito JP, Herrin J, Swarna KS, et al. GLP-1RA Use and Thyroid Cancer Risk. *JAMA Otolaryngology–Head & Neck Surgery*. 2024;
9. Jones N, Shih M-C, Healey E, et al. Use of Machine Learning to Assess the Management of Uncomplicated Urinary Tract Infection. *JAMA Network Open*. 2025;8(1):e2456950-e2456950.
10. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in medicine*. 2010;29(3):337-346.
11. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*. 2010;63(8):826-833.
12. Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *American Journal of Epidemiology*. 2023;192(9):1536-1544.
13. Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology*. 2021;32(3):393-401.
14. Lu H, Cole SR, Platt RW, Schisterman EF. Revisiting overadjustment bias. *Epidemiology*. 2021;32(5):e22-e23.
15. Pirracchio R, Petersen ML, Van Der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology*. 2015;181(2):108-119.
16. Dorie V, Hill J, Shalit U, Scott M, Cervone D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. 2019;

17. Wang SV, Russo M, Glynn RJ, et al. A Benchmark, Expand, and Calibration (BenchExCal) Trial Emulation Approach for Using Real-World Evidence to Support Indication Expansions: Design and Process for a Planned Empirical Evaluation. *Clinical Pharmacology & Therapeutics*. 2025;

18. Dahabreh IJ, Matthews A, Steingrimsson JA, Scharfstein DO, Stuart EA. Using trial and observational data to assess effectiveness: trial emulation, transportability, benchmarking, and joint analysis. *Epidemiologic reviews*. 2024;46(1):1-16.

19. Matthews AA, Dahabreh IJ, Fröbert O, et al. Benchmarking observational analyses before using them to address questions trials do not answer: an application to coronary thrombus aspiration. *American journal of epidemiology*. 2022;191(9):1652-1665.

20. McMurray JJ, Packer M, Desai AS, et al. Angiotensin–neprilysin inhibition versus enalapril in heart failure. *New England Journal of Medicine*. 2014;371(11):993-1004.

21. Rohde LE, Chatterjee NA, Vaduganathan M, et al. Sacubitril/Valsartan and Sudden Cardiac Death According to Implantable Cardioverter-Defibrillator Use and Heart Failure Cause. *JACC: Heart Failure*. 2020;8(10):844-855. doi:doi:10.1016/j.jchf.2020.06.015

22. Yancy CW, Jessup M, Bozkurt B, et al. 2016 ACC/AHA/HFSA focused update on new pharmacological therapy for heart failure: an update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. *Journal of the American College of Cardiology*. 2016;68(13):1476-1488.

23. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Journal of the American College of Cardiology*. 2022;79(17):e263-e421.

24. Webster-Clark M, Ross RK, Lund JL. Initiator types and the causal question of the prevalent new-user design: a simulation study. *American journal of epidemiology*. 2021;190(7):1341-1348.

25. Tan NY, Sangaralingham LR, Sangaralingham SJ, Yao X, Shah ND, Dunlay SM. Comparative effectiveness of sacubitril-valsartan versus ACE/ARB therapy in heart failure with reduced ejection fraction. *JACC: Heart Failure*. 2020;8(1):43-54.

26. Department of Veterans Affairs. VHA Directive 1906: Data Quality Requirements for Health Care Identity Management and Master Person Index Functions. 2020. Accessed 4/8/2025. https://www.va.gov/vhapublications/ViewPublication.asp?pub_ID=8787

27. Sohn M-W, Arnold N, Maynard C, Hynes DM. Accuracy and completeness of mortality data in the Department of Veterans Affairs. *Population health metrics*. 2006;4:1-8.

28. Staerk C, Byrd A, Mayr A. Recent methodological trends in Epidemiology: No need for data-driven variable selection? *American Journal of Epidemiology*. 2024;193(2):370-376.

29. Setodji CM, McCaffrey DF, Burgette LF, Almirall D, Griffin BA. The right tool for the job: choosing between covariate-balancing and generalized boosted model propensity scores. *Epidemiology*. 2017;28(6):802-811.

30. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*. 2004;9(4):403.

31. Zhou Y, Matsouaka RA, Thomas L. Propensity score weighting under limited overlap and model misspecification. *Statistical methods in medical research*. 2020;29(12):3721-3756.

32. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*. 2019;188(1):250-257.

33. Westreich D, Cole SR. Invited commentary: positivity in practice. *American journal of epidemiology*. 2010;171(6):674-677.

34. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA network open*. 2020;3(1):e1918962-e1918962.

35. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC heart failure*. 2021;8(1):106-115.

36. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal*. 2017;38(23):1805-1814.

37. Friedman AN, Fadem SZ. Reassessment of albumin as a nutritional marker in kidney disease. *Journal of the American Society of Nephrology*. 2010;21(2):223-230.

38. Geirhos R, Jacobsen J-H, Michaelis C, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*. 2020;2(11):665-673.

39. Bellamy D, Hernán MA, Beam A. A structural characterization of shortcut features for prediction. *European Journal of Epidemiology*. 2022;37(6):563-568.

40. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American journal of epidemiology*. 2006;163(12):1149-1156.

41. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*. 2011;174(11):1213-1222.

42. Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*. 2010;15(3):234.

43. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*. 2013;32(19):3388-3414.

44. Karim ME. Can supervised deep learning architecture outperform autoencoders in building propensity score models for matching? *BMC Medical Research Methodology*. 2024;24(1):167.

45. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. *American journal of epidemiology*. 2017;186(8):1010-1014.

46. A Multicenter, Randomized, Double-blind, Parallel Group, Active-controlled Study to Evaluate the Efficacy and Safety of LCZ696 Compared to Enalapril on Morbidity and Mortality in Patients With Chronic Heart Failure and Reduced Ejection Fraction. National Library of Medicine (US). https://clinicaltrials.gov/study/NCT01035255

47. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med*. 2017;377(14):1391-1398.

48. Wright R. Interpreting black-box machine learning models using partial dependence and individual conditional expectation plots. *Exploring SAS® Enterprise Miner Special Collection*. 2018;1950

Table 1. Baseline Characteristics of Heart Failure Patients with Implantable Cardioverter Defibrillators Prescribed Sacubitril/Valsartan or ACEI/ARB in the U.S. Department of Veterans Affairs, 2016 – 2020.

	Treatment Strategy		Overall (n = 2,447)
	Sacubitril/Valsartan (n = 583)	ACEI/ARB (n = 1,864)	
Age, mean (SD)	68.6 (8.6)	70.2 (8.2)	69.8 (8.3)
Gender, men (%)	574 (98.5)	1834 (98.4)	2408 (98.4)
Race (%)			
White	398 (68.3)	1380 (74.0)	1778 (72.7)
Black	148 (25.4)	344 (18.5)	492 (20.1)
Asian	3 (0.5)	4 (0.2)	7 (0.3)
American Indian/Alaska Native	4 (0.7)	11 (0.6)	15 (0.6)
Native Hawaiian/Pacific Islander	3 (0.5)	13 (0.7)	16 (0.7)
Multi-Race	2 (0.3)	11 (0.6)	13 (0.5)
Unknown	25 (4.3)	101 (5.4)	126 (5.1)
Ethnicity (%)			
Hispanics	23 (3.9)	73 (3.9)	96 (3.9)
Non-Hispanics	544 (93.3)	1728 (92.7)	2272 (92.8)
Unknown	16 (2.7)	63 (3.4)	79 (3.2)
Area deprivation index , mean (SD)	60.2 (24.2)	60.6 (24.6)	60.5 (24.5)
Missing	2	24	26
Cigarette smoking			
Never	130 (22.3)	419 (22.5)	549 (22.4)
Current	231 (39.6)	765 (41.0)	996 (40.7)
Former	221 (37.9)	668 (35.8)	889 (36.3)
Unknown	1 (0.2)	12 (0.6)	13 (0.5)
Body mass index (mean (SD))	30.4 (6.2)	29.4 (5.9)	29.6 (6.0)
LVEF, %, mean (SD)	25.8 (6.1)	26.8 (6.1)	26.6 (6.1)
Systolic blood pressure, mm Hg, mean (SD)	120.9 (17.8)	121.9 (18.5)	121.7 (18.3)
Creatinine, mg/dL, median [IQR]	1.3 [1.1, 1.5]	1.2 [1.0, 1.5]	1.2 [1.0, 1.5]
eGFR, mL/min/1.73 m^2, mean (SD)	63.5 (18.4)	64.7 (20.5)	64.4 (20.0)
Potassium, mEq/L, median [IQR]	4.2 [3.9, 4.6]	4.2 [3.9, 4.5]	4.2 [3.9, 4.5]
Missing	1	4	5
BNP, pg/mL, median [IQR]	453.6 [228.0, 960.9]	415.0 [226.2, 942.0]	429.1 [226.9, 949.6]

	Treatment Strategy		Overall (n = 2,447)
	Sacubitril/Valsartan (n = 583)	**ACEI/ARB (n = 1,864)**	
< 150	29 (5.0)	85 (4.6)	114 (4.7)
Not done	130 (22.3)	477 (25.6)	607 (24.8)
NT-proBNP, pg/mL, median [IQR]	1935.0 [985.0, 3796.0]	2234.5 [1126.0, 4797.2]	2155.0 [1074.0, 4435.2]
< 600	15 (2.6)	35 (1.9)	50 (2.0)
Not done	324 (55.6)	1104 (59.2)	1428 (58.4)
Comorbidity (%)			
Atrial fibrillation	314 (53.9)	1015 (54.5)	1329 (54.3)
Cardiomyopathy	492 (84.4)	1464 (78.5)	1956 (79.9)
Chronic kidney disease	259 (44.4)	804 (43.1)	1063 (43.4)
COPD	283 (48.5)	864 (46.4)	1147 (46.9)
Diabetes	338 (58.0)	1045 (56.1)	1383 (56.5)
Hypertension	567 (97.3)	1782 (95.6)	2349 (96.0)
Depression	236 (40.5)	679 (36.4)	915 (37.4)
Myocardial infraction	154 (26.4)	498 (26.7)	652 (26.6)
Peripheral artery disease	186 (31.9)	616 (33.0)	802 (32.8)
Valvular heart disease	186 (31.9)	599 (32.1)	785 (32.1)
Treatment (%)			
Cardiac resynchronization therapy	260 (44.6)	925 (49.6)	1185 (48.4)
β-Blockers	536 (91.9)	1724 (92.5)	2260 (92.4)
MRAs	285 (48.9)	663 (35.6)	948 (38.7)
SGLT2i	17 (2.9)	29 (1.6)	46 (1.9)
Diuretics	399 (68.4)	1125 (60.4)	1524 (62.3)
Digoxin	67 (11.5)	168 (9.0)	235 (9.6)
Influenza vaccine (%)	418 (71.7)	1404 (75.3)	1822 (74.5)
Any HF hospitalization (%)	260 (44.6)	1022 (54.8)	1282 (52.4)
Time since HF, mo, median [IQR]	80.0 [38.5, 147.5]	62.0 [14.0, 125.0]	68.0 [19.0, 130.5]
User type, prevalent (%)	543 (93.1)	1558 (83.6)	2101 (85.9)
Additional predictors			
Time since implantation, mo, median [IQR]	22.0 [8.0, 45.0]	2.0 [1.0, 3.0]	3.0 [1.0, 11.0]
Anemia	239 (41.0)	771 (41.4)	1010 (41.3)
Malignancy	94 (16.1)	402 (21.6)	496 (20.3)
DVT/PE	32 (5.5)	69 (3.7)	101 (4.1)
Dementia	20 (3.4)	111 (6.0)	131 (5.4)
Influenzas	26 (4.5)	43 (2.3)	69 (2.8)
Non-pathological fracture	86 (14.8)	274 (14.7)	360 (14.7)

	Treatment Strategy		Overall (n = 2,447)
	Sacubitril/Valsartan (n = 583)	**ACEI/ARB (n = 1,864)**	
Cerebral vascular conditions	98 (16.8)	319 (17.1)	417 (17.0)
Albumin, g/dL, median [IQR]	3.0 [3.0, 4.0]	3.0 [3.0, 4.0]	3.0 [3.0, 4.0]
Missing	6	12	18
Hemoglobin, g/dL, median [IQR]	13.0 [12.0, 14.0]	13.0 [12.0, 14.0]	13.0 [12.0, 14.0]
Missing	4	32	36

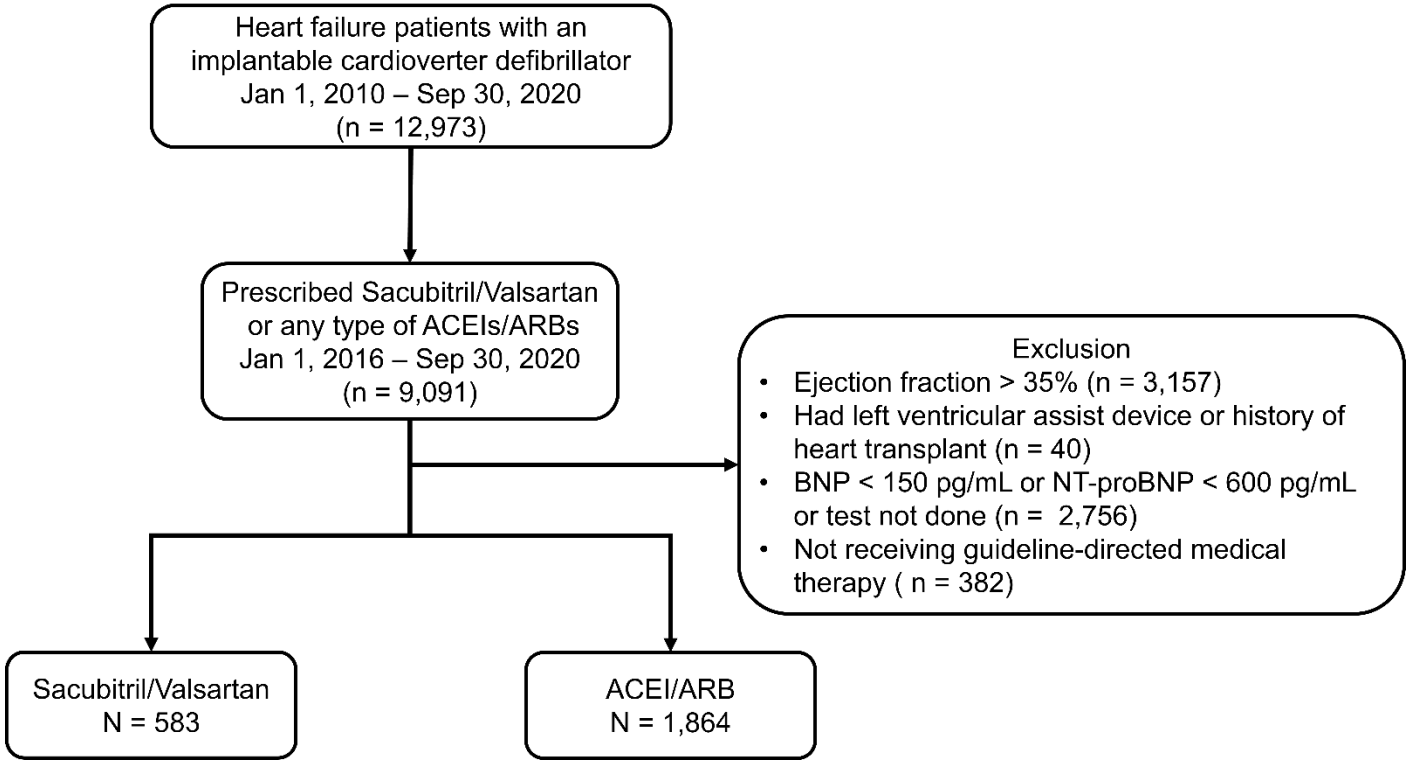
Abbreviations: ACEI/ARB, angiotensin converting enzyme inhibitor and angiotensin II receptor blocker; BNP, brain natriuretic peptide; COPD, chronic obstructive pulmonary disease; CRT, DVT/PE, deep vein thrombosis and pulmonary embolism; eGFR, estimated glomerular filtration rate; HF, heart failure; IQR, interquartile range; LVEF, left ventricular ejection fraction; MRAs: mineralocorticoid receptor antagonists; NT-proBNP, N-terminal pro-brain natriuretic peptide; SGLT2i, sodium-glucose cotransporter 2 inhibitors; SD, standard deviation.

Table 2. Intention-to-Treat Estimates of Cumulative Incidence, Risk Difference, Risk Ratio and Hazard Ratio for All-Cause Mortality from an Emulated Target Trial of Sacubitril/Valsartan versus ACEI/ARB, Benchmarked Against the Hazard Ratio from the Index Randomized Trial.

	27-months cumulative incidence		RD	RR	HR
	Sacubitril/Valsartan, % (95% CI)	ACEI/ARB, % (95% CI)	(95% CI)	(95% CI)	(95% CI)
PARADIGM-HF trial					0.81 (0.61, 1.06)
Emulated target trial					
Logistic regression & *a priori*	9.3 (6.2, 12.9)	10.7 (9.2, 12.2)	-1.4 (-4.6, 2.2)	0.87 (0.59, 1.21)	0.93 (0.61, 1.42)
GBM & *a priori*	10.3 (9.1, 18.9)	10.7 (8.8, 12.1)	-0.4 (-1.4, 9.1)	0.96 (0.87, 1.93)	0.97 (0.68, 1.37)
GBM & *a priori*, additional adjustment	8.0 (6.5, 17.6)	8.4 (5.5, 11.4)	-0.4 (-0.8, 8.1)	0.95 (0.92, 1.97)	0.97 (0.68, 1.38)
GBM & data-driven	6.5 (3.6, 11.2)	10.8 (9.2, 12.4)	-4.2 (-7.6, 0.4)	0.61 (0.33, 1.04)	0.61 (0.30, 1.23)

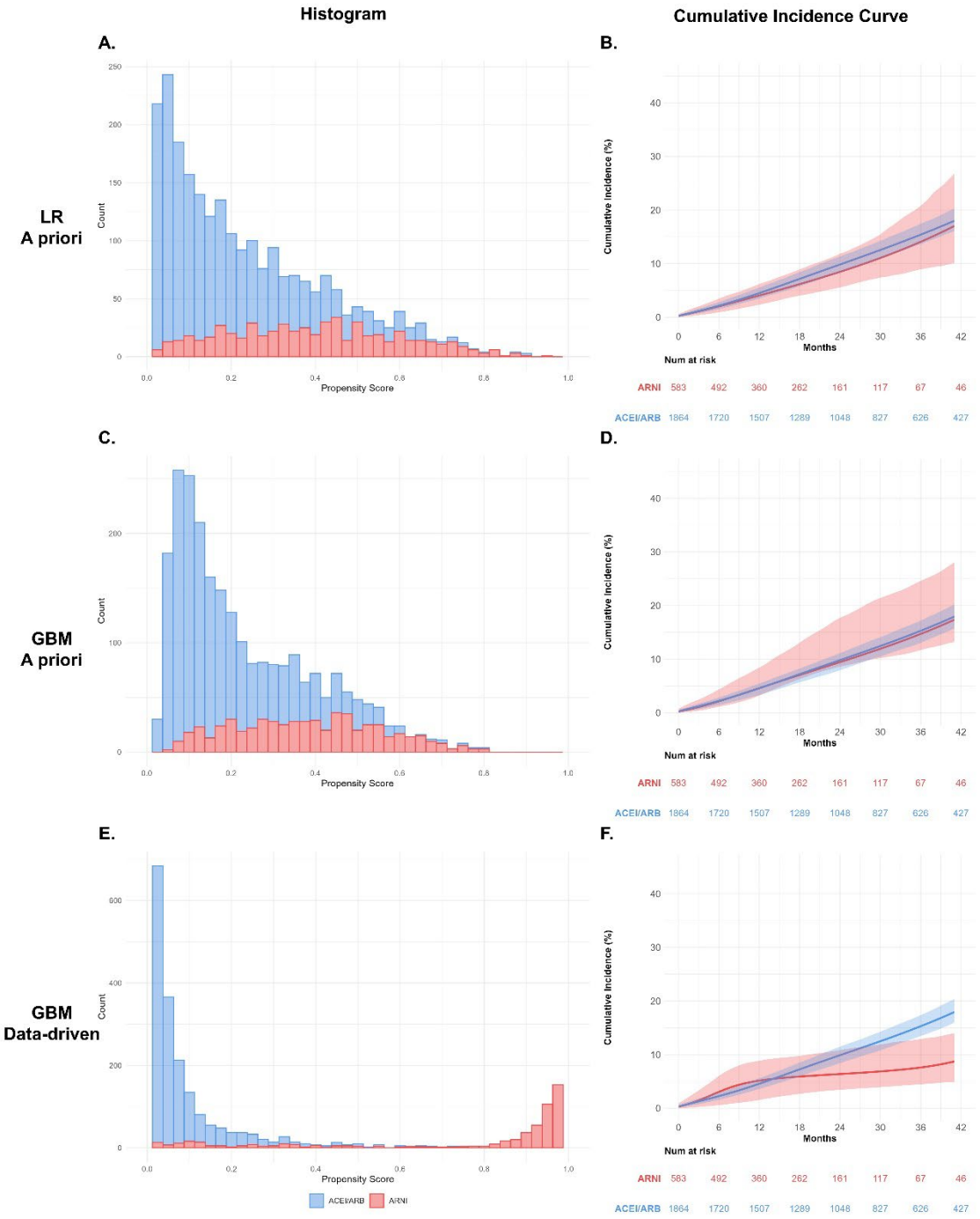
Abbreviations: ACEI/ARB, angiotensin converting enzyme inhibitors and angiotensin II receptor blockers; CI, confidence interval; GBM, generalized boosted model; HR, hazard ratio; logistic regression; RD, risk difference; RR, risk ratio.

Figure 1. Study Flow Chart



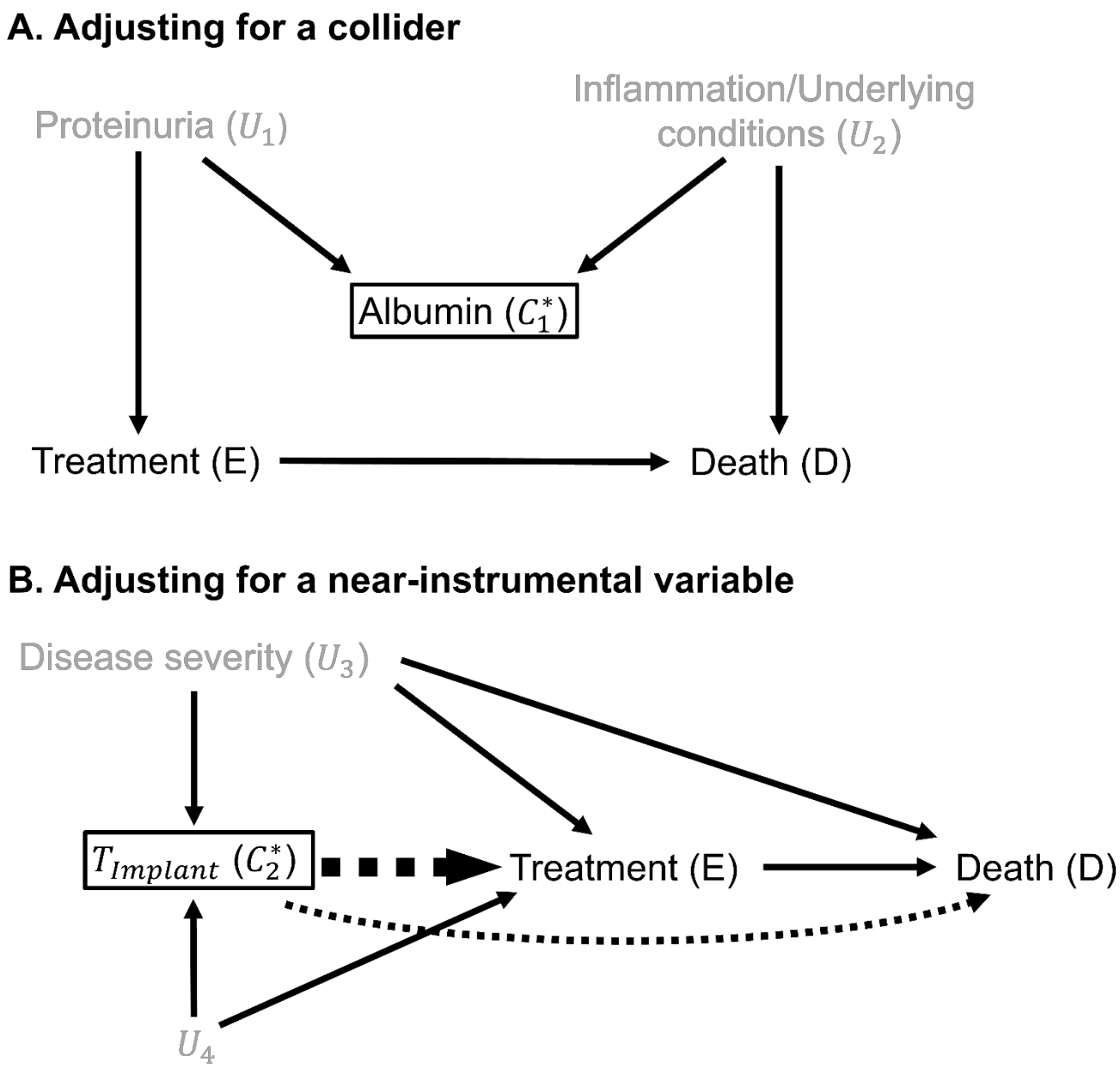
Abbreviations: ACEIs/ARBs, angiotensin converting enzyme inhibitors and angiotensin II receptor blockers; BNP, brain natriuretic peptide; NT-proBNP, N-terminal pro-brain natriuretic peptide.

Figure 2. Propensity Score Distributions and Intention-to-Treat Cumulative Incidence Curves by Analytical Approach.



Abbreviations: ACEI/ARB, angiotensin converting enzyme inhibitors and angiotensin II receptor blockers; ARNI, angiotensin receptor/neprilysin inhibitor; GBM, generalized boosted model; LR logistic regression.

Figure 3. Directed Acyclic Graphs Illustrating Overadjustment Biases in the Data-Drive GBM-Based Propensity Score Approach.



Black boxes indicate variables included in the propensity score model. Gray text represents unmeasured or unknown confounders. Dashed arrows denote non-causal paths implied by the data-driven GBM-based approach with arrow line thickness indicate strength of the association. Variables labeled C_i^* represent GBM-imposed confounders.