HYPERGRAPH-BASED DOUBLY ROBUST ESTIMATION FOR CAUSAL INFERENCE OF DRUG COMBINATION EFFECTS IN HEART FAILURE TREATMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Disease management for heart failure with preserved ejection fraction (HFpEF) requires understanding the comparative effectiveness of real-world drug combinations rather than single agents. Standard randomized controlled trials for multi-drug regimens are prohibitively expensive, slow, and often infeasible at scale, motivating the use of causal machine learning methods on large-scale electronic health records (EHRs). However, reliable estimation of treatment effects is challenging due to the high-order multi–drug interactions, strong confounding factors, and patient heterogeneity across sex, age, and other characteristics. Existing causal machine learning methods mostly focus on comparing treatments with pairwise drug combinations, while techniques for multi-drug regimens are less studied. To fill this gap, we propose a **Hyper**graph–based **D**oubly **R**obust framework (**HyperDR**), which represents drug classes as nodes and observed multi-drug regimens as hyperedges, and uses a hypergraph neural network to learn shared representations for both drugs and combinations from cross-sectional EHR data. Our framework jointly trains a propensity score model and an outcome model using cross-entropy losses and an inverse-probability-weighted residual regularizer. Treatment effects are then estimated using an augmented inverse-probability-weighted (AIPW) estimator, which is consistent under standard causal and regularity assumptions if either the propensity scores or conditional outcome means are consistently estimated. On a semi-synthetic benchmark with known ground-truth treatment effects, HyperDR demonstrates improved causal effect recovery by achieving the lowest ATE and CATE RMSE among the evaluated methods. Experiments on two real-world HFpEF cohorts further demonstrate strong factual outcome prediction, while exploratory case studies characterize subgroup-specific treatment rankings.

1 INTRODUCTION

Heart failure with preserved ejection fraction (HFpEF) accounts for approximately 50% of all heart failure cases, and its prevalence is increasing due to aging populations and comorbidities (Tsao et al., 2023; Redfield, 2016; Redfield & Borlaug, 2023; Borlaug & Paulus, 2011). Unlike heart failure with reduced ejection fraction (HFrEF), for which large-scale randomized controlled trials (RCTs) have already established clear multi-drug guideline-directed medical therapies (GDMT) (Heidenreich et al., 2022; McMurray & Packer, 2021; Bloom et al., 2017; Chen et al., 2019), evidence for optimal combination strategies in HFpEF is less established. While sodium-glucose cotransporter-2 (SGLT2) inhibitors are now a recommended therapy (Solomon et al., 2022; Madaan et al., 2016), HFpEF management in real-world practice involves complex polypharmacy. Because HFpEF patients often have multiple comorbidities, such as hypertension, obesity, and diabetes, clinicians routinely prescribe combinations of SGLT2 inhibitors, renin-angiotensin-aldosterone system (RAAS) inhibitors, β-blockers, and diuretics.

Identifying the most effective drug combination for an individual patient is difficult. RCTs cannot feasibly test every possible combination of three or four drug classes (Lienhardt et al., 2011). Additionally, RCTs often exclude patients with severe multimorbidity, limiting their applicability to the broader patient population. Electronic health records (EHRs) provide large sample sizes for analysis

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

but present two main challenges: confounding by indication (where treatment assignment depends on patient severity) and high dimensionality (modeling many possible drug combinations) (Hernán & Robins, 2016).

Causal machine learning has recently emerged as a method to estimate treatment effects from observational data (Feuerriegel et al., 2024; Sanchez et al., 2022; Richens et al., 2020; Schölkopf, 2022). A key challenge lies in estimating counterfactual outcomes while adjusting for confounding. Standard statistical methods, such as inverse probability of treatment weighting (IPTW), often suffer from instability in high-dimensional settings due to insufficient overlap in propensity scores across treatment groups (Cheng et al., 2022). Techniques based on representation learning can balance covariates between treatment groups in the latent space to reduce confounding (Johansson et al., 2016), e.g., TARNet (Shalit et al., 2017) and DragonNet (Shi et al., 2019). Several works extend the methodology from binary treatment to more general treatment settings: TransTEE (Zhang et al., 2023) leverages a transformer backbone to handle continuous or structured treatments via treatment–covariate attention; NCoRE (Parbhoo et al., 2021) explicitly models interactions between different treatment arms; and METO (Liu et al., 2025) introduces a multi-treatment encoding mechanism tailored to drug sequences for hypertension. To better capture structural relationships among treatments, GraphITE (Harada & Kashima, 2021) represents each drug treatment as a pairwise graph and learns treatment embeddings with a GCN (Kipf & Welling, 2017). Recent studies have applied these methods to rank treatments for diabetes (Belthangady et al., 2022) and hypertension (Liu et al., 2025). However, most existing models represent drug combination therapies as simple lists, sequences, or pairwise graphs. These representations may not fully capture higher-order interactions within multi-drug regimens (the combined effect of multiple drugs may differ from the sum of their individual effects) and the shared structure across overlapping regimens, both of which naturally call for a hypergraph formulation.

In this study, we develop a novel hypergraph-based, doubly robust causal deep learning framework for estimating the effects of HFpEF drug combinations on subsequent HFpEF hospitalization. We represent six key drug categories, including sacubitril/valsartan, SGLT2 inhibitors, loop diuretics, mineralocorticoid receptor antagonists (MRAs), RAAS inhibition, and β-blockers, as nodes in a drug hypergraph, and represent each observed multi-drug regimen as a hyperedge. A hypergraph neural network learns shared embeddings for drug nodes and induces representations for drug combinations (hyperedges). These embeddings are then jointly used with patient confounding covariates to parameterize two coupled modules: (i) a propensity score predictor that models the assignment probability of each drug combination treatment, and (ii) an outcome predictor that estimates the risk of HFpEF rehospitalization under each potential combination. We also design a training objective that links the two modules through outcome prediction, propensity supervision, and residual regularization. The fitted models are subsequently combined through AIPW for doubly robust treatment-effect estimation under standard causal and regularity assumptions.

We validate this framework on two HFpEF cohorts — a real-world dataset called *OneFlorida* from OneFlorida Clinical Research Network (Hogan et al., 2022) and a public dataset *MIMIC-IV* (Johnson et al., 2023) to (1) rank drug combinations by estimated causal effects on rehospitalization, (2) characterize patient heterogeneity in optimal combinations across different clinically relevant subgroups, and (3) compare our approach with both deep causal and non-deep baselines. To assess causal treatment-effect recovery against known ground truth, we additionally conduct a Polynomial semi-synthetic experiment using real *OneFlorida* covariates and 28 observed treatment combinations, with simulated treatment assignments and outcomes under observed confounding and higher-order drug interactions. HyperDR achieves the lowest ATE and CATE RMSE among the evaluated methods in this setting, supporting its ability to recover causal treatment effects under the specified simulation mechanism. Our results provide both methodological advances, i.e., a hypergraph-based doubly robust estimator tailored to multi-drug therapy, and clinically interpretable insights into HFpEF pharmacotherapy that potentially inform future guidelines, trial design, and decision support systems.

2 PROBLEM STATEMENT

2.1 PRELIMINARIES

We denote patient **covariates (features)** as $\mathbf{x} \in \mathbb{R}^d$, which are assumed to capture all *observed confounding factors* relevant to treatment assignment and outcomes. These covariates include demographics, comorbidities, laboratory measurements, and vital signs. We denote the **observed outcome** as $y \in \{0, 1\}$. In our heart failure setting, the outcome indicates whether the patient experiences an HFpEF-related hospitalization within a predefined follow-up horizon after treatment.

We define a **drug hypergraph** as $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, where each node $v_i \in \mathcal{V}$ corresponds to a drug class, and each hyperedge $e_j \in \mathcal{E}$ represents a unique drug combination therapy (treatment) composed of multiple drug classes. For example, a hyperedge may represent the combination of an SGLT2 inhibitor, a loop diuretic, and a β-blocker. We denote the total number of drug classes as $K_{\text{drug}} = |\mathcal{V}|$ and the total number of observed combination therapies as $K_{\text{comb}} = |\mathcal{E}|$.

Our dataset consists of a collection of cross-sectional patient snapshots,

$$\mathcal{D} = \left\{(\mathbf{x}_u, T_u, y_u)\right\}_{u=1}^{N},$$

where $T_u \in \mathcal{E}$ denotes the observed treatment (i.e., a drug combination corresponding to a hyperedge in $\mathcal{H}$) for snapshot u, and N is the total number of snapshots across all patients. A list of commonly used notations is in Table 4 in Appendix.

2.2 PROBLEM DEFINITION

Given an observational dataset $\mathcal{D}$ and a predefined drug hypergraph $\mathcal{H}$, our goal is to learn a model f_θ that predicts the potential outcome $y(T)$ for a patient with covariates $\mathbf{x}$ under different drug combination therapies $T \in \mathcal{E}$ (for notational simplicity, we omit index u when there is no ambiguity). Formally, the model produces an estimate $\hat{y}(T) = f_\theta(\mathbf{x}, T; \mathcal{H})$, enabling counterfactual comparison of multiple treatment combinations while adjusting for confounding effects induced by high-dimensional covariates and complex multi-drug interactions.

3 METHODOLOGY

Figure 1 provides an overview of our proposed hypergraph-based doubly robust estimation framework. Our framework falls under the category of doubly robust estimation (Funk et al., 2011). It consists of two main modules: a **propensity module** that takes a patient's covariates as inputs and predicts the probability mass of all possible treatments as a multi-class classification, and an **outcome module** that takes a patient's covariates and a drug combination treatment as inputs and predicts the outcome. The two modules are jointly learned through shared patient embeddings. One major unique design is that our framework models treatments (multi-drug combinations) with a hypergraph, as shown in Figure 1(a), and generalizes the common doubly robust estimation model through a hypergraph neural network to capture the complex high-order interactions between drug agents within and across different kinds of combination therapies.

3.1 HYPERGRAPH-BASED DRUG AND TREATMENT REPRESENTATION

The goal here is to learn the latent representations of different multi-drug combination treatments while capturing high-order drug interactions and overlaps among different drug combinations. Given hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, we learn latent embeddings for both drug classes (nodes) and treatment combinations (hyperedges).

Specifically, we define the drug-treatment incidence matrix $\mathbf{H} \in \{0, 1\}^{K_{\text{drug}} \times K_{\text{comb}}}$, where $\mathbf{H}_{i,j} = 1$ if drug class node v_i belongs to hyperedge (drug combination treatment) e_j, and $\mathbf{H}_{i,j} = 0$ otherwise. Let $\mathbf{W}_e \in \mathbb{R}^{K_{\text{comb}} \times K_{\text{comb}}}$ be a learnable diagonal hyperedge-weight matrix. The node and hyperedge degree matrices are $\mathbf{D}_v(i,i) = \sum_j \mathbf{H}_{i,j}\mathbf{W}_e(j,j)$ and $\mathbf{D}_e(j,j) = \sum_i \mathbf{H}_{i,j}$.

A hypergraph neural network (HGNN) (Feng et al., 2019) layer updates node embeddings via Equation 1:

$$\mathbf{Z} = \sigma\left(\mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{W}_e\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{D}_v^{-1/2}\mathbf{X}\mathbf{W}\right), \tag{1}$$

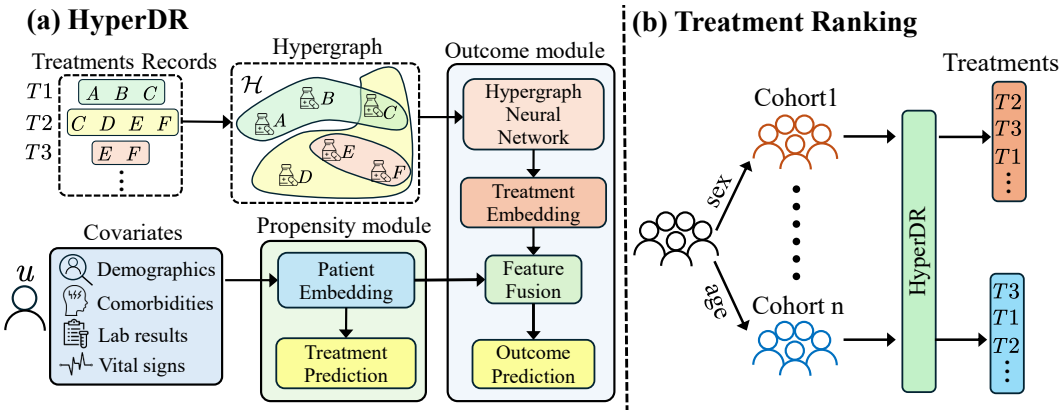


Figure 1: (a) Overview of the HyperDR framework. The architecture integrates a hypergraph neural network for treatment embedding with patient embedding, jointly optimizing propensity and outcome prediction for doubly robust causal inference. (b) The trained HyperDR model is applied to different patient subgroups (*e.g.,* stratified by sex or age) to generate personalized or subgroup-specific treatment rankings based on estimated potential outcomes.

where $\mathbf{X} \in \mathbb{R}^{K_{\text{drug}} \times d}$ is the initial node feature matrix, $\mathbf{W} \in \mathbb{R}^{d \times d_h}$ is trainable, $\sigma(\cdot)$ is a nonlinearity (e.g., ReLU), $\mathbf{Z} \in \mathbb{R}^{K_{\text{drug}} \times d_h}$ are updated node embeddings, and d_h is the latent dimension hyperparameter. We stack L such layers to yield the final drug-class embedding matrix. In order to get treatment (hyperedge) embeddings, we use a permutation-invariant aggregation function. In our implementation, we aggregate via attention. We denote all hyperedge embeddings as $\{\mathbf{h}_{e_j}\}$.

3.2 Shared Representation and Doubly Robust Prediction Modules

For notational simplicity, we drop the snapshot index u throughout this section. Given snapshot-level covariates $\mathbf{x}$ and the corresponding treatment embedding $\mathbf{h}_T$, our model learns a shared covariate representation that feeds into two coupled prediction modules: a propensity-score model and an outcome model.

3.2.1 Propensity-score module.

The propensity-score module estimates the treatment assignment probability given covariates, *i.e.,* the distribution $P_\theta(T \mid \mathbf{x})$ over observed treatment combinations $T \in \{e_j\} \subset \mathcal{E}$. To avoid leakage of treatment information into propensity estimation, this module takes only the covariate representation as input. We use standard dense layers to get patient covariate embeddings. We train the propensity head with the standard multi-class cross-entropy. This propensity estimate will be used both in the doubly robust pseudo-outcome and residual-based regularization term later, thereby coupling treatment assignment modeling with outcome prediction.

3.2.2 Outcome prediction module.

The outcome module predicts the outcome (e.g., hospitalization risk) of a treatment given specific patient covariates. For the treatment $e_j \in \mathcal{E}$, let $\mathbf{r}(\mathbf{x}) \in \mathbb{R}^{d_h}$ denote the patient covariate embedding (obtained via dense layers), and define the joint representation $\mathbf{c}_{e_j} = [\mathbf{r}(\mathbf{x}) \,\|\, \mathbf{h}_{e_j}] \in \mathbb{R}^{2d_h}$ as the concatenation of the patient and treatment embeddings. We model the conditional mean outcome by a sigmoid predictor:

$$\mu_\theta(\mathbf{x}, e_j) = \sigma\big(g_\theta(\mathbf{c}_{e_j})\big), \tag{2}$$

where $g_\theta : \mathbb{R}^{2d_h} \to \mathbb{R}$ is a nonlinear multilayer perceptron with trainable parameters, and $\sigma(\cdot)$ is the sigmoid function. Denoting by $j^\star$ the index of the observed treatment ($T = e_{j^\star}$), the factual prediction is $\hat{y} = \mu_\theta(\mathbf{x}, e_{j^\star})$. We train the outcome head using the standard binary cross-entropy loss.

3.2.3 MODEL TRAINING LOSS WITH RESIDUAL REGULARIZER.

The overall objective combines factual outcome prediction, propensity supervision, and a residual regularizer:

$$\mathcal{L}_{\text{total}} = \frac{1}{N}\sum_{u=1}^{N}\Big(\ell_{\text{out}} + \lambda_1\,\ell_{\text{prop}} + \lambda_2\,\ell_{\text{r}}\Big), \tag{3}$$

where ℓ_{out} and ℓ_{prop} are the loss of the outcome prediction module and the loss of the propensity module, respectively. The residual regularizer ℓ_{r} is defined as:

$$\ell_{\text{r}} = \left(\frac{y-\hat{y}}{\max(\hat{p}(j^\star),\epsilon)}\right)^2,$$

where $\hat{p}(j^\star)$ is the estimated propensity of the observed combination and $\epsilon > 0$ is a small constant for numerical stability. The regularizer penalizes large inverse-probability-weighted factual residuals, with the aim of improving practical stability for infrequent treatment combinations. The hyperparameters λ_1 and λ_2 control the strength of the auxiliary objectives.

3.3 CAUSAL EFFECT ESTIMATION AND TREATMENT RANKING

After training, we use the learned model to estimate causal effects of treatment combinations and to construct treatment rankings.

3.3.1 DOUBLY ROBUST COUPLING VIA PSEUDO-OUTCOMES.

The propensity and outcome modules are coupled through an augmented inverse-probability–weighted (AIPW) pseudo-outcome (Glynn & Quinn, 2010), which provides a doubly robust estimate of the marginal risk for each treatment combination. Specifically, for any $e_j \in \mathcal{E}$, we define

$$\psi(j) = \mu_\theta(\mathbf{x}, e_j) + \frac{\mathbb{I}(j^\star = j)}{\hat{p}(j)}\Big(y - \mu_\theta(\mathbf{x}, e_j)\Big), \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $j^\star$ denotes the index of the observed drug combination treatment corresponding to $T = e_j$, $\hat{p}(j) = P_\theta(e_j \mid \mathbf{x})$ is the propensity estimate from the propensity head, and $\mu_\theta(\mathbf{x}, e_j)$ is the outcome-model prediction under the combination e_j computed by the outcome head (Equation 2).

According to the proven theory of doubly robust estimation (Funk et al., 2011), Equation 4 can be viewed as an outcome-regression baseline $\mu_\theta(\mathbf{x}, e_j)$ augmented by an inverse-probability–weighted residual term. The augmentation corrects systematic bias in μ_θ when the propensity model is accurate, while the residual term vanishes in expectation when the outcome model is accurate. Consequently, under standard causal and regularity assumptions, averaging $\psi(j)$ over snapshots yields a doubly robust estimator of the marginal risk under combination e_j, which is consistent if either $\hat{p}(j)$ or $\mu_\theta(\mathbf{x}, e_j)$ is correctly specified.

3.3.2 MARGINAL RISKS AND AVERAGE TREATMENT EFFECTS.

For each combination $e_j \in \mathcal{E}$, we estimate the marginal risk of HFpEF hospitalization using the AIPW pseudo-outcome:

$$\widehat{R}(j) = \frac{1}{N}\sum_{u=1}^{N}\psi_u(j). \tag{5}$$

Subgroup-specific risks (e.g., by sex, race, or obesity) are obtained by restricting the averages in Equation 5 to snapshots in the subgroup of interest.

3.3.3 POPULATION-LEVEL AND INDIVIDUALIZED RANKINGS.

At the population level, we rank treatment combinations by their estimated marginal risk $\widehat{R}(j)$, with lower values indicating more favorable regimens. For an individual patient profile $\tilde{\mathbf{x}}$ at decision time, we form an individualized ranking by sorting combinations $\{e_j\}$ according to the predicted

conditional risk $\mu_\theta(\tilde{\mathbf{x}}, e_j)$. Here $\widehat{R}(j)$ targets a population-level marginal (causal) risk via AIPW, whereas $\mu_\theta(\tilde{\mathbf{x}}, e_j)$ provides a patient-specific risk score for personalized decision support. Because treatment combinations are represented as hyperedges in a shared drug hypergraph and encoded by a common HGNN, these rankings exploit pharmacologic relationships and higher-order drug–drug interactions, while the AIPW-based estimation improves robustness to misspecification in either the propensity or outcome component.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 DATASETS

We evaluate HyperDR on two EHR datasets: (1) a real-world HFpEF cohort from OneFlorida Clinical Research Network (Hogan et al., 2022) (*OneFlorida*), and (2) the publicly available *MIMIC-IV* dataset (Johnson et al., 2023). The *OneFlorida* dataset contains 5,697 patients and 26,510 snapshots with 76 covariates and 28 observed treatment combinations, while *MIMIC-IV* includes 16,432 patients and 37,421 snapshots with 14 covariates and 42 treatment combinations. Following Section 2.1, patient records are converted into snapshot-level samples consisting of covariates, observed treatment combinations, and binary hospitalization outcomes. Treatments are defined over six HF-relevant drug classes: sacubitril/valsartan, SGLT2 inhibitors, loop diuretics, mineralocorticoid receptor antagonists (MRAs), RAAS inhibition, and β-blockers, and represented as hyperedges in a predefined drug hypergraph.

Table 6 and Table 7 in Appendix present the mapping between treatment identifiers and their corresponding drug-class combinations on the *OneFlorida* and *MIMIC-IV* dataset. Each treatment ID represents a unique observed regimen composed of one or more HFpEF-relevant drug classes. These combinations were extracted from real-world prescription records and reflect clinically prevalent mono- and polypharmacy patterns in routine practice.

4.1.2 BASELINES

We compare HyperDR against three categories of baselines. (i) A traditional statistical reference: Logistic Regression (**LR**). (ii) Deep causal representation models originally developed for single/binary treatment settings: **TARNet** (Shalit et al., 2017), which uses treatment-specific outcome heads; **DragonNet** (Shi et al., 2019), which jointly learns outcome and multi-class propensity models; and **TransTEE** (Zhang et al., 2023), which leverages a Transformer to model treatment–covariate interactions. (iii) Combination-aware methods designed for multi-treatment estimation: **NCoRE** (Parbhoo et al., 2021), which explicitly models cross-treatment interactions via dedicated subnetworks; **METO** (Liu et al., 2025), which encodes multi-drug regimens with propensity-based reweighting; and **GraphITE** (Harada & Kashima, 2021), which represents drug treatments as graphs and learns their embeddings with a graph neural network. Performance is evaluated using Accuracy, Precision, Recall, and F1-score. More experimental setup details are provided in Appendix B.

4.2 FACTUAL OUTCOME PREDICTION ON REAL-WORLD COHORTS

Table 1 reports the classification performance on the *OneFlorida* and *MIMIC-IV* datasets.

On *OneFlorida*, HyperDR achieves the highest Accuracy (91.8%), Recall (81.4%), and F1-score (82.2%) among the evaluated methods. Its Accuracy and F1-score exceed those of the strongest baseline on these metrics, TransTEE (91.2% and 80.5%), by 0.6 and 1.7 percentage points, respectively. TransTEE achieves the highest Precision (83.3%), compared with 82.9% for HyperDR. Among combination-aware baselines, GraphITE achieves the highest F1-score (79.9%), ahead of METO (77.3%) and NCoRE (76.8%).

On *MIMIC-IV*, HyperDR achieves the highest Accuracy (85.7%), Precision (79.8%), and F1-score (83.0%). Its Accuracy and F1-score exceed those of the strongest baseline on these metrics, GraphITE (84.6% and 82.0%), by 1.1 and 1.0 percentage points, respectively. HyperDR's Recall

Table 1: Classification results (%, mean ± standard deviation over three runs) on *OneFlorida* and *MIMIC-IV*. Best results are in bold.

Methods	OneFlorida				MIMIC-IV			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
LR	$84.7_{\pm 0.0}$	$67.0_{\pm 0.0}$	$68.1_{\pm 0.1}$	$67.5_{\pm 0.1}$	$76.5_{\pm 0.0}$	$69.2_{\pm 0.1}$	$75.1_{\pm 0.0}$	$72.0_{\pm 0.0}$
TARNet	$87.9_{\pm 0.2}$	$74.8_{\pm 0.1}$	$73.2_{\pm 0.1}$	$74.0_{\pm 0.1}$	$81.4_{\pm 0.1}$	$76.1_{\pm 0.1}$	$78.5_{\pm 0.1}$	$77.3_{\pm 0.0}$
DragonNet	$89.4_{\pm 0.1}$	$78.5_{\pm 0.1}$	$75.0_{\pm 0.2}$	$76.7_{\pm 0.2}$	$81.6_{\pm 0.1}$	$73.0_{\pm 0.2}$	$86.5_{\pm 0.1}$	$79.2_{\pm 0.1}$
TransTEE	$91.2_{\pm 0.2}$	$\mathbf{83.3_{\pm 0.2}}$	$77.8_{\pm 0.1}$	$80.5_{\pm 0.1}$	$83.9_{\pm 0.1}$	$76.9_{\pm 0.2}$	$85.7_{\pm 0.1}$	$81.1_{\pm 0.1}$
NCoRE	$89.2_{\pm 0.2}$	$77.0_{\pm 0.2}$	$76.6_{\pm 0.2}$	$76.8_{\pm 0.2}$	$82.5_{\pm 0.2}$	$74.8_{\pm 0.2}$	$85.4_{\pm 0.1}$	$79.7_{\pm 0.1}$
METO	$89.4_{\pm 0.1}$	$76.5_{\pm 0.1}$	$78.1_{\pm 0.1}$	$77.3_{\pm 0.1}$	$84.2_{\pm 0.1}$	$77.9_{\pm 0.1}$	$84.8_{\pm 0.0}$	$81.2_{\pm 0.0}$
GraphITE	$90.7_{\pm 0.1}$	$80.4_{\pm 0.2}$	$79.5_{\pm 0.1}$	$79.9_{\pm 0.1}$	$84.6_{\pm 0.1}$	$77.6_{\pm 0.1}$	$\mathbf{86.9_{\pm 0.1}}$	$82.0_{\pm 0.1}$
HyperDR	$\mathbf{91.8_{\pm 0.1}}$	$82.9_{\pm 0.2}$	$\mathbf{81.4_{\pm 0.1}}$	$\mathbf{82.2_{\pm 0.1}}$	$\mathbf{85.7_{\pm 0.1}}$	$\mathbf{79.8}_{\pm 0.1}$	$86.4_{\pm 0.1}$	$\mathbf{83.0_{\pm 0.1}}$

Table 2: Polynomial semi-synthetic results (mean ± standard deviation over three runs; lower is better). Best results are in bold.

Metric	LR	TARNet	DragonNet	TransTEE	NCoRE	METO	GraphITE	**HyperDR**
ATE RMSE ↓ (AIPW)	0.084 ±0.011	0.081 ±0.021	0.062 ±0.016	0.062 ±0.016	0.061 ±0.011	0.068 ±0.002	0.077 ±0.017	**0.039 ±0.005**
CATE RMSE ↓ (outcome)	0.213 ±0.022	0.197 ±0.019	0.167 ±0.014	0.173 ±0.014	0.149 ±0.022	0.187 ±0.054	0.195 ±0.020	**0.083 ±0.009**

(86.4%) is slightly lower than that of GraphITE (86.9%) and DragonNet (86.5%), while its Precision is higher than both (77.6% and 73.0%, respectively).

Overall, HyperDR achieves the highest Accuracy and F1-score on both datasets, supporting its predictive performance for factual hospitalization outcomes in the evaluated cohorts.

4.3 Causal Effect Recovery on Semi-Synthetic Data

Inspired by NCoRE's polynomial treatment-interaction simulation (Parbhoo et al., 2021), we evaluate causal effect recovery on a Polynomial semi-synthetic benchmark constructed from the *OneFlorida* covariates and 28 observed treatment combinations. Treatments and binary outcomes are simulated under observed confounding, with covariate-dependent drug interactions of orders two through five. The known conditional risks provide ground truth for all candidate treatments. Our primary metric is the root mean squared error of average treatment effects estimated by AIPW (ATE RMSE), evaluated over all 378 treatment pairs on the same held-out population. The secondary metric, CATE RMSE, evaluates conditional average treatment effects directly from outcome predictions. Appendix D gives the data-generation process, simulation-specific implementations, training procedure, and metric definitions.

Table 2 shows that HyperDR achieves the lowest error on both metrics. Compared with NCoRE, the strongest baseline on both metrics, HyperDR reduces ATE RMSE from 0.061 to 0.039 and CATE RMSE from 0.149 to 0.083, corresponding to relative reductions of 36.7% and 44.2%, respectively. These results support the evaluated model's ability to recover population-average and covariate-conditional treatment effects under the specified simulation mechanism. They do not establish that the estimated effects in the original clinical cohorts are causally identified.

4.4 Ablation Study

We conduct an ablation study to assess the contribution of the proposed hypergraph-based treatment representation module. On the *OneFlorida* cohort, we compare HyperDR with two variants that replace the treatment encoder while keeping all other components unchanged (patient encoder, propensity/outcome heads, and the training objective): (i) **HyperDR-MLP**, which treats each observed drug combination as an unstructured category and learns its representation with a multilayer perceptron (MLP) (Gardner & Dorling, 1998), and (ii) **HyperDR-GCN**, which models drug classes

Table 3: Ablation results (%) on *OneFlorida*. HyperDR-MLP and HyperDR-GCN use MLP and GCN treatment encoders, respectively.

Metric	HyperDR-MLP	HyperDR-GCN	HyperDR
Acc	89.0	91.2	**91.8**
Pre	77.2	81.8	**82.9**
Rec	75.8	80.4	**81.4**
F1	76.5	81.1	**82.2**

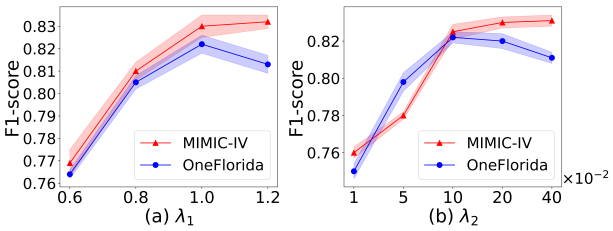


Figure 2: Sensitivity experiments of hyperparameters (a) λ_1 and (b) λ_2. Curves show the mean F1-score over three runs; shaded bands indicate standard deviation.

with a pairwise graph convolutional network (GCN) (Kipf & Welling, 2017) and aggregates node embeddings to obtain combination representations.

Table 3 shows that HyperDR achieves the highest values across all four reported metrics on the *OneFlorida* cohort, with Accuracy of 91.8% and F1-score of 82.2%. Replacing the hypergraph encoder with an unstructured MLP reduces Accuracy to 89.0% and F1-score to 76.5%; the pairwise GCN variant achieves Accuracy of 91.2% and F1-score of 81.1%. HyperDR therefore exceeds the MLP and GCN variants in Accuracy by 2.8 and 0.6 percentage points, respectively. These results support the contribution of the hypergraph encoder to predictive performance in the evaluated setting.

4.5 Sensitivity Analysis

We examined HyperDR's sensitivity to two key hyperparameters in the total training objective: propensity loss weight λ_1 and residual regularization weight λ_2. As illustrated in Figure 2(a), increasing λ_1 from 0.6 to 1.0 improves F1-score on both datasets. At $\lambda_1 = 1.2$, the *OneFlorida* score decreases, whereas the *MIMIC-IV* score increases slightly. We use $\lambda_1 = 1.0$; this value achieves the highest plotted score on *OneFlorida*, while the highest plotted *MIMIC-IV* score occurs at 1.2. Figure 2(b) varies λ_2 from 1×10^{-2} to 40×10^{-2}. F1-score increases on both datasets up to 10×10^{-2}. Beyond this value, the *OneFlorida* score decreases, while the *MIMIC-IV* score continues to increase slightly. The highest plotted scores therefore occur at different λ_2 values for the two datasets. These results describe sensitivity of predictive performance to the loss weights; they do not directly measure inverse-probability weight stability or causal-estimation variance.

4.6 Case Study: Treatment Ranking Analysis

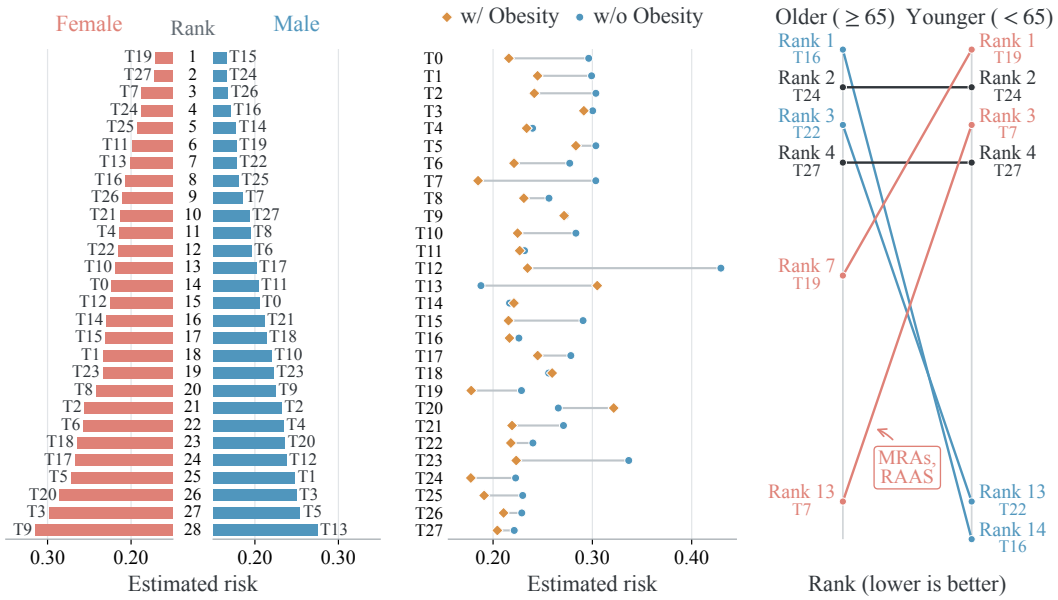
We stratify the snapshot-level *OneFlorida* cohort by sex, obesity, and age, and rank observed treatment combinations by their estimated marginal risk of HFpEF hospitalization (lower is better), using the doubly robust procedure in Section 3.3. These exploratory comparisons do not by themselves establish treatment-effect modification or drug-class-wide differences in benefit; patient composition and estimation uncertainty may also contribute to subgroup differences.

4.6.1 Sex-stratified treatment ranking

In Figure 3(a), T19 (β-blocker + RAAS + sacubitril/valsartan) and T27 (loop diuretic + RAAS + sacubitril/valsartan) rank first and second in female patients, compared with sixth and tenth in male patients. In male patients, T15 (SGLT2 monotherapy), T24 (β-blocker + MRAs + RAAS + SGLT2), T26 (β-blocker + RAAS + SGLT2), and T16 (RAAS + SGLT2) rank first through fourth, respectively; their corresponding ranks in female patients are seventeenth, fourth, ninth, and eighth. T25 (loop diuretic + SGLT2) ranks fifth in female patients and eighth in male patients.

4.6.2 Obesity-stratified treatment ranking

Figure 3(b) shows relatively small between-group risk differences for T16 and T26 compared with T7, T12, T13, and T23. T24 has lower estimated risk in patients with obesity, with a larger group difference than T16 or T26. T7, T12 (β-blocker + loop diuretic + MRAs), and T23 (MRAs + SGLT2) also have lower estimated risk in patients with obesity, whereas T13 (loop diuretic + MRAs) has



(a) Sex (b) Obesity (c) Age

Figure 3: Demographic subgroup comparisons on *OneFlorida*. (a) Sex-stratified global ranking of all 28 treatments (T0–T27). (b) Estimated marginal hospitalization risk across all treatments for patients with obesity (w/ Obesity) and without obesity (w/o Obesity); lower risk indicates better outcomes. (c) Rank shifts for selected treatments between older (≥ 65) and younger (< 65) patients. Red and blue indicate treatments ranked higher in younger and older patients, respectively; black indicates unchanged ranks.

higher estimated risk. T12 and T13 thus show opposite subgroup directions despite both containing loop diuretics and MRAs.

4.6.3 Age-stratified treatment ranking

In Figure 3(c), T24 and T27 rank second and fourth, respectively, in both older (≥ 65) and younger (< 65) patients. T16 ranks first in older patients and fourteenth in younger patients, while T22 (β-blocker + SGLT2) ranks third and thirteenth. Conversely, T19 ranks seventh in older patients and first in younger patients, and T7 (MRAs + RAAS) ranks thirteenth and third, respectively.

5 Conclusion and Future Work

In this paper, we formulated the comparative effectiveness of multi-drug HFpEF treatment as an observational analogue of a randomized trial, and proposed HyperDR, a hypergraph-based doubly robust framework that represents drug classes and their combinations as hyperedges and jointly trains propensity and outcome models. Experiments on the *OneFlorida* and *MIMIC-IV* cohorts show that HyperDR achieves the highest Accuracy and F1-score among the evaluated methods for hospitalization outcome prediction. Case studies additionally provide treatment rankings and subgroup-specific patterns for further investigation.

Future work will extend HyperDR from static combinations to ordered treatment strategies, explicitly modeling the initiation, escalation, and discontinuation sequence of drug classes as dynamic treatment regimes over a hypergraph action space. Additional directions include cross-site validation on heterogeneous health systems, integrating prior RCT and guideline evidence into the hypergraph, and incorporating uncertainty quantification and sensitivity analysis for unmeasured confounding to further strengthen its use in decision support and trial design.

ETHICS STATEMENT

This study was approved by the Institutional Review Board (IRB) of our institution. The findings are intended to inform further research and should not be interpreted as clinical recommendations without further validation.

AI USE STATEMENT

In this work, generative AI tools were used to assist with manuscript writing and language editing, including improving clarity and readability and proofreading for grammatical errors. Generative AI tools were not used to formulate the research questions or hypotheses, design the methodology or experiments, generate or process data, implement the methods, derive mathematical claims, or interpret the results. All AI-assisted text was reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 describe the problem formulation and model, while Section 4 and Appendix B document the clinical experimental setup. Appendix D describes the semi-synthetic data-generating process, data splits, and evaluation protocol. *MIMIC-IV* is available through PhysioNet to credentialed users who complete the required training and sign the data use agreement. We will release the source code upon acceptance.

REFERENCES

Chinmay Belthangady, Stefanos Giampanis, Ivana Jankovic, Will Stedden, Paula Alves, Stephanie Chong, Charlotte Knott, and Beau Norgeot. Causal deep learning reveals the comparative effectiveness of antihyperglycemic treatments in poorly controlled diabetes. *Nature Communications*, 13(1):6921, 2022.

Michelle W Bloom, Barry Greenberg, Tiny Jaarsma, James L Januzzi, Carolyn SP Lam, Aldo P Maggioni, Jean-Noel Trochu, and Javed Butler. Heart failure with reduced ejection fraction. *Nature reviews Disease primers*, 3(1):17058, 2017.

Barry A Borlaug and Walter J Paulus. Heart failure with preserved ejection fraction: pathophysiology, diagnosis, and treatment. *European heart journal*, 32(6):670–679, 2011.

Yei-Tsung Chen, Lee Lee Wong, Oi Wah Liew, and Arthur Mark Richards. Heart failure with reduced ejection fraction (hfref) and preserved ejection fraction (hfpef): the diagnostic value of circulating micrornas. *Cells*, 8(12):1651, 2019.

Chao Cheng, Fan Li, Laine E Thomas, and Fan Li. Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. *American journal of epidemiology*, 191(6):1140–1151, 2022.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 3558–3565, 2019.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7): 761–767, 2011.

Matt W Gardner and Stephen R Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

Adam N Glynn and Kevin M Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56, 2010.

Shonosuke Harada and Hisashi Kashima. Graphite: Estimating individual effects of graph-structured treatments. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 659–668, 2021.

Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun, Monica M Colvin, Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, 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. *Circulation*, 145 (18):e895–e1032, 2022.

Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

William R Hogan, Elizabeth A Shenkman, Temple Robinson, Olveen Carasquillo, Patricia S Robinson, Rebecca Z Essner, Jiang Bian, Gigi Lipori, Christopher Harle, Tanja Magoc, et al. The oneflorida data trust: a centralized, translational research data infrastructure of statewide scope. *Journal of the American Medical Informatics Association*, 29(4):686–693, 2022.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=SJU4ayYgl`.

Christian Lienhardt, Sharlette V Cook, Marcos Burgos, Victoria Yorke-Edwards, Leen Rigouts, Gladys Anyo, Sang-Jae Kim, Amina Jindani, Don A Enarson, Andrew J Nunn, et al. Efficacy and safety of a 4-drug fixed-dose combination regimen compared with separate drugs for treatment of pulmonary tuberculosis: the study c randomized controlled trial. *Jama*, 305(14):1415–1423, 2011.

Ruoqi Liu, Lang Li, and Ping Zhang. Estimating the treatment effects of multiple drug combinations on multiple outcomes in hypertension. *Cell Reports Medicine*, 6(2), 2025.

Tushar Madaan, Mohd Akhtar, and Abul Kalam Najmi. Sodium glucose cotransporter 2 (sglt2) inhibitors: Current status and future perspective. *European Journal of Pharmaceutical Sciences*, 93:244–252, 2016.

John JV McMurray and Milton Packer. How should we sequence the treatments for heart failure and a reduced ejection fraction? a redefinition of evidence-based medicine. *Circulation*, 143(9): 875–877, 2021.

Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. Ncore: Neural counterfactual representation learning for combinations of treatments. *arXiv preprint arXiv:2103.11175*, 2021.

Margaret M Redfield. Heart failure with preserved ejection fraction. *New England Journal of Medicine*, 375(19):1868–1877, 2016.

Margaret M Redfield and Barry A Borlaug. Heart failure with preserved ejection fraction: a review. *Jama*, 329(10):827–838, 2023.

Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923, 2020.

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pp. 765–804. 2022.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Scott D Solomon, John JV McMurray, Brian Claggett, Rudolf A de Boer, David DeMets, Adrian F Hernandez, Silvio E Inzucchi, Mikhail N Kosiborod, Carolyn SP Lam, Felipe Martinez, et al. Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction. *New England Journal of Medicine*, 387(12):1089–1098, 2022.

Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Andrea Z Beaton, Amelia K Boehme, Alfred E Buxton, et al. Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93, 2023.

YiFan Zhang, Hanlin Zhang, Zachary Chase Lipton, Li Erran Li, and Eric Xing. Exploring transformer backbones for heterogeneous treatment effect estimation. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=1kl4YM2Q7P.

Table 4: Main mathematical notations.

Notation	Description
u	Snapshot index of patients
N	Total number of snapshots
$\mathcal{D}$	Observational dataset $\{(\mathbf{x}_u, T_u, y_u)\}_{u=1}^{N}$
$\mathbf{x}_u$	Snapshot covariates
T_u	Observed treatment combination
y_u	Binary outcome
K_{drug}	Number of drug classes
K_{comb}	Number of observed treatment combinations
$\mathcal{H}$	Drug hypergraph $(\mathcal{V}, \mathcal{E})$
e_j	j-th treatment combination (hyperedge)
$\hat{p}_u(j)$	Propensity $P_\theta(e_j \mid \mathbf{x}_u)$
$\mu_\theta(\mathbf{x}_u, e_j)$	Predicted risk under combination e_j
$\psi_u(j)$	AIPW pseudo-outcome for combination e_j
$\widehat{R}(j)$	Estimated marginal (AIPW) risk under e_j
ℓ_{out}	Outcome loss
ℓ_{prop}	Propensity loss
ℓ_{r}	Residual regularizer
λ_1, λ_2	Weights for ℓ_{prop} and ℓ_{r}

A Math Notation

A list of commonly used notations is in Table 4.

B Experimental Setup Details

B.1 Dataset Preprocessing

We conducted experiments on two electronic health record (EHR) datasets. The first is a real-world HFpEF cohort extracted from a regional medical network, referred to as the *OneFlorida* dataset. The second dataset is *MIMIC-IV* (Johnson et al., 2023), a publicly available critical-care database.

For both datasets, longitudinal patient records were transformed into snapshot-level observational samples following the procedure described in Section 2.1. Each snapshot consists of a covariate vector $\mathbf{x}_u$ summarizing demographic, clinical, and laboratory measurements, an observed treatment combination T_u, and a binary outcome y_u indicating HFpEF-related hospitalization within a fixed follow-up horizon. All snapshots were constructed to ensure temporal ordering between covariates, treatments, and outcomes.

Treatments were defined over six clinically relevant drug classes, including sacubitril/valsartan, SGLT2 inhibitors, loop diuretics, mineralocorticoid receptor antagonists (MRAs), RAAS inhibitors, and β-blockers. Each snapshot's treatment T_u is represented as a set of drug-class indices, and all distinct observed regimens form hyperedges in the predefined drug hypergraph. This representation enables shared learning across related treatment patterns and facilitates modeling of higher-order drug interactions.

B.2 Dataset Statistics

Table 5 reports baseline demographic, clinical, laboratory, and treatment characteristics across decile-based subgroups on the *OneFlorida* dataset. Continuous variables are reported as mean (standard deviation), and categorical variables as counts (percentages). Compared with the bottom group, the top group has a higher mean left ventricular ejection fraction (60.6% versus 58.9%), lower BNP (360.5 versus 373.7 pg/mL), and lower troponin I (611.8 versus 793.6 ng/L), but higher NT-proBNP (3773.7 versus 3327.2 pg/mL). The top group also has lower estimated GFR (53.4 versus 56.3 mL/min/1.73m^2), higher blood urea nitrogen (23.4 versus 21.3 mg/dL), and higher creatinine

Table 5: Baseline characteristics by decile group.

Characteristic	Top 10 Decile	10–50 Decile	Bottom 50 Decile	p-value
Age (years)	65.8 (12.6)	65.1 (11.9)	65.6 (12.5)	0.038
Length of Stay (days)	0.2 (0.4)	0.1 (0.3)	0.2 (0.8)	<0.001
Body Weight (kg)	100.9 (26.1)	99.7 (24.6)	99.6 (24.1)	0.078
Diastolic Blood Pressure (mmHg)	74.4 (9.6)	74.8 (9.4)	75.9 (10.4)	<0.001
Systolic Blood Pressure (mmHg)	136.0 (16.8)	136.1 (16.5)	137.9 (18.5)	<0.001
Body Mass Index (kg/m^2)	33.3 (9.6)	33.4 (9.4)	33.2 (9.4)	0.440
Heart Rate (beats/min)	77.7 (11.2)	77.6 (11.6)	76.8 (11.9)	<0.001
Left Ventricular Ejection Fraction (%)	60.6 (8.3)	58.9 (5.8)	58.9 (6.1)	<0.001
Hemoglobin A1c (%)	6.7 (1.5)	6.7 (1.6)	6.7 (1.6)	0.722
Lymphocyte Count (10^9/L)	1.7 (0.8)	1.7 (0.8)	1.8 (0.8)	<0.001
B-type Natriuretic Peptide (pg/mL)	360.5 (362.1)	382.1 (303.3)	373.7 (273.3)	0.021
N-terminal proBNP (pg/mL)	3773.7 (7667.7)	3510.6 (7066.4)	3327.2 (6114.2)	0.004
Troponin I (ng/L)	611.8 (1086.0)	799.8 (1337.0)	793.6 (1305.1)	<0.001
Total Cholesterol (mg/dL)	158.1 (41.0)	157.1 (43.7)	159.7 (42.9)	0.001
LDL Cholesterol (mg/dL)	83.8 (32.1)	83.2 (34.5)	85.5 (34.3)	<0.001
HDL Cholesterol (mg/dL)	49.4 (15.1)	48.9 (15.6)	48.8 (15.1)	0.142
Triglycerides (mg/dL)	133.2 (73.1)	135.4 (82.1)	136.2 (87.3)	0.280
Serum Creatinine (mg/dL)	1.5 (1.6)	1.6 (1.8)	1.4 (1.5)	<0.001
Blood Urea Nitrogen (mg/dL)	23.4 (13.5)	22.8 (13.7)	21.3 (12.3)	<0.001
Serum Sodium (mmol/L)	138.9 (3.0)	139.0 (3.1)	139.2 (3.1)	<0.001
Serum Potassium (mmol/L)	4.2 (0.4)	4.2 (0.5)	4.2 (0.5)	0.905
Serum Albumin (g/dL)	3.9 (0.5)	3.9 (0.5)	3.9 (0.5)	<0.001
Estimated GFR (mL/min/1.73m^2)	53.4 (23.0)	54.0 (24.3)	56.3 (23.2)	<0.001
Smoking Status	9.0 (0.3)	9.0 (0.0)	9.0 (0.5)	<0.001
Male Sex	846 (38.75%)	1267 (35.52%)	7949 (38.29%)	0.005
Diabetes Mellitus	1515 (69.40%)	2706 (75.86%)	14525 (69.97%)	<0.001
Hypertension	2147 (98.35%)	3552 (99.58%)	20599 (99.22%)	<0.001
Obesity	1694 (77.60%)	2840 (79.62%)	15673 (75.50%)	<0.001
Statin Therapy	1101 (50.44%)	1985 (55.65%)	10402 (50.11%)	<0.001
Insulin Therapy	750 (34.36%)	1317 (36.92%)	6010 (28.95%)	<0.001

(1.5 versus 1.4 mg/dL). These mixed directions do not support a uniform characterization of the top group as having a more favorable cardio-renal profile. Mean length of stay rounds to 0.2 days in both the top and bottom groups, so the displayed means do not establish a longer stay in the bottom group. Diabetes, hypertension, and obesity are common across all three groups, while their reported proportions and those of statin and insulin therapy differ across groups.

B.3 MIMIC-IV Preprocessing

Table 7 presents the mapping between treatment identifiers and their corresponding drug-class combinations on the *MIMIC-IV* dataset.

B.4 Baselines

We compared HyperDR with representative baseline methods from three methodological categories: traditional statistical models, deep causal representation learning approaches originally designed for single/binary treatment settings, and combination-aware deep models for multi-treatment estimation.

- **Logistic Regression (LR).** A traditional statistical baseline that models hospitalization risk using observed covariates and treatment indicators, serving as a reference for assessing the benefits of nonlinear representation learning.
- **TARNet** (Shalit et al., 2017). A deep causal representation learning model that learns a shared patient representation and uses treatment-specific outcome heads to estimate potential outcomes under different treatment assignments.
- **DragonNet** (Shi et al., 2019). A joint outcome–propensity network that simultaneously learns balanced patient representations and treatment assignment probabilities through targeted regularization.

Table 6: Mapping from treatment ID to drug-class combinations on the *OneFlorida* dataset.

Treatment ID	Treatment combination
0	RAAS
1	β-blockers
2	β-blockers, RAAS
3	β-blockers, Loop diuretics, RAAS
4	Loop diuretics
5	β-blockers, Loop diuretics
6	Loop diuretics, RAAS
7	MRAs, RAAS
8	MRAs
9	β-blockers, Loop diuretics, MRAs, RAAS
10	β-blockers, MRAs, RAAS
11	β-blockers, MRAs
12	β-blockers, Loop diuretics, MRAs
13	Loop diuretics, MRAs
14	Loop diuretics, MRAs, RAAS
15	SGLT2
16	RAAS, SGLT2
17	β-blockers, Loop diuretics, RAAS, Sacubitril/Valsartan
18	β-blockers, Loop diuretics, MRAs, RAAS, Sacubitril/Valsartan
19	β-blockers, RAAS, Sacubitril/Valsartan
20	RAAS, Sacubitril/Valsartan
21	β-blockers, Loop diuretics, RAAS, SGLT2
22	β-blockers, SGLT2
23	MRAs, SGLT2
24	β-blockers, MRAs, RAAS, SGLT2
25	Loop diuretics, SGLT2
26	β-blockers, RAAS, SGLT2
27	Loop diuretics, RAAS, Sacubitril/Valsartan

- **TransTEE** (Zhang et al., 2023). A Transformer-based treatment effect estimator that uses attention mechanisms to model interactions between patient covariates and treatments, enabling flexible handling of continuous and structured treatment types. We adapt it to our multi-treatment classification setting by treating each observed drug combination as a discrete treatment token.
- **NCoRE** (Parbhoo et al., 2021). A neural counterfactual representation learning model explicitly designed for combination treatment settings. It models cross-treatment interactions through dedicated interaction subnetworks that are activated when multiple treatments are applied simultaneously.
- **METO** (Liu et al., 2025). A combination-aware deep model that explicitly encodes multi-drug regimens as sequences and incorporates propensity-based reweighting to mitigate confounding in observational data.
- **GraphITE** (Harada & Kashima, 2021). A graph-based treatment effect estimator that represents each drug treatment as a graph and learns treatment embeddings using a graph neural network.

All baseline methods were tuned using validation data and implemented under comparable training, optimization, and early-stopping protocols to ensure fair comparison.

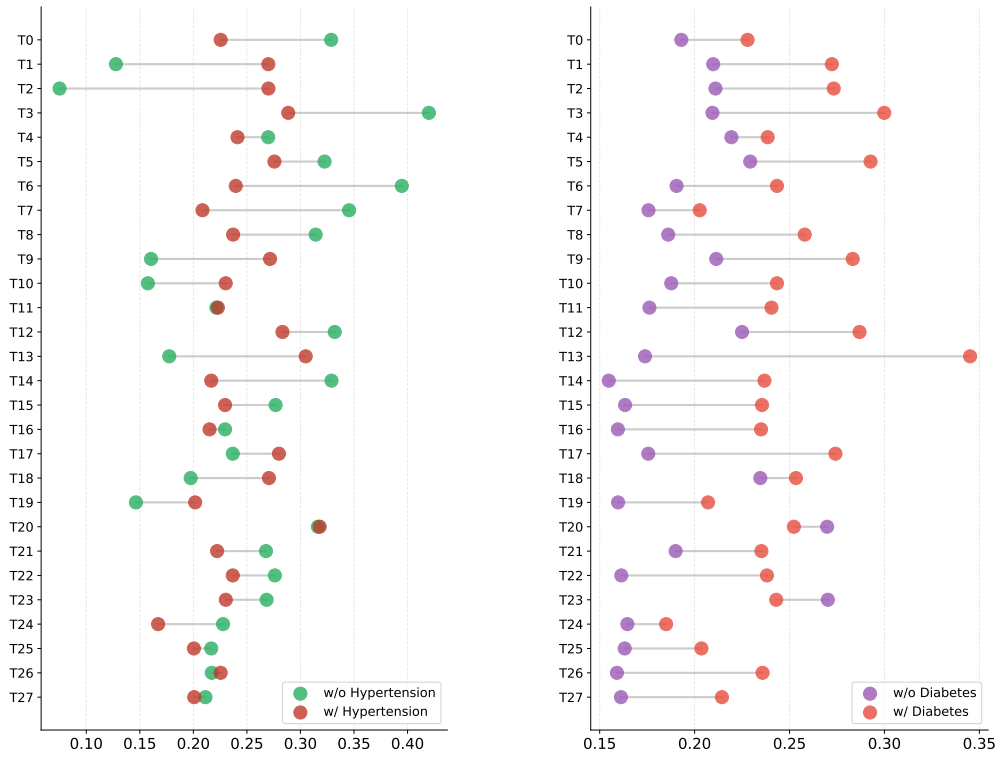
B.5 Implementation Details

The clinical experiments were implemented in PyTorch and repeated three times. Models were trained using a batch size of 256 for up to 150 epochs with early stopping based on validation performance (patience = 10). We set the number of layers L to 2. We used the Adam optimizer with a learning rate of 5×10^{-5} and weight decay of 1×10^{-4}. The treatment embedding dimension was

Table 7: Mapping from treatment ID to drug-class combinations on the *MIMIC-IV* dataset.

Treatment ID	Treatment combination
0	β-blockers, Loop diuretics
1	β-blockers, Loop diuretics, RAAS
2	Loop diuretics
3	β-blockers
4	β-blockers, Loop diuretics, MRAs
5	β-blockers, Loop diuretics, MRAs, RAAS
6	β-blockers, RAAS
7	Loop diuretics, RAAS
8	Loop diuretics, MRAs
9	RAAS
10	β-blockers, MRAs, RAAS
11	Loop diuretics, MRAs, RAAS
12	β-blockers, MRAs
13	MRAs, RAAS
14	MRAs
15	β-blockers, Loop diuretics, MRAs, SGLT2
16	β-blockers, Loop diuretics, MRAs, RAAS, Sacubitril/Valsartan
17	β-blockers, Loop diuretics, RAAS, Sacubitril/Valsartan
18	β-blockers, Loop diuretics, SGLT2
19	Loop diuretics, MRAs, SGLT2
20	β-blockers, Loop diuretics, MRAs, RAAS, SGLT2
21	Loop diuretics, RAAS, Sacubitril/Valsartan
22	β-blockers, RAAS, Sacubitril/Valsartan
23	β-blockers, Loop diuretics, MRAs, RAAS, SGLT2, Sacubitril/Valsartan
24	Loop diuretics, MRAs, RAAS, SGLT2
25	β-blockers, Loop diuretics, RAAS, SGLT2
26	Loop diuretics, SGLT2
27	β-blockers, MRAs, RAAS, Sacubitril/Valsartan
28	β-blockers, Loop diuretics, RAAS, SGLT2, Sacubitril/Valsartan
29	Loop diuretics, RAAS, SGLT2
30	Loop diuretics, MRAs, RAAS, Sacubitril/Valsartan
31	β-blockers, SGLT2
32	RAAS, Sacubitril/Valsartan
33	β-blockers, MRAs, SGLT2
34	Loop diuretics, RAAS, SGLT2, Sacubitril/Valsartan
35	MRAs, RAAS, SGLT2
36	SGLT2
37	β-blockers, MRAs, RAAS, SGLT2, Sacubitril/Valsartan
38	β-blockers, RAAS, SGLT2, Sacubitril/Valsartan
39	Loop diuretics, MRAs, RAAS, SGLT2, Sacubitril/Valsartan
40	β-blockers, MRAs, RAAS, SGLT2
41	RAAS, SGLT2

set to 128, and categorical features were embedded into 64-dimensional vectors, followed by hidden layers of size 256 with a dropout rate of 0.1. These experiments were conducted on one NVIDIA B200 GPU with 16 AMD EPYC 7742 CPU cores and 256 GB of system memory.



(a) Hypertension (b) Diabetes

Figure 4: Estimated marginal hospitalization risk across all 28 treatments on *OneFlorida*, comparing patients with and without (a) hypertension and (b) diabetes. The horizontal axes report estimated risk; lower values indicate better outcomes.

C Additional Results

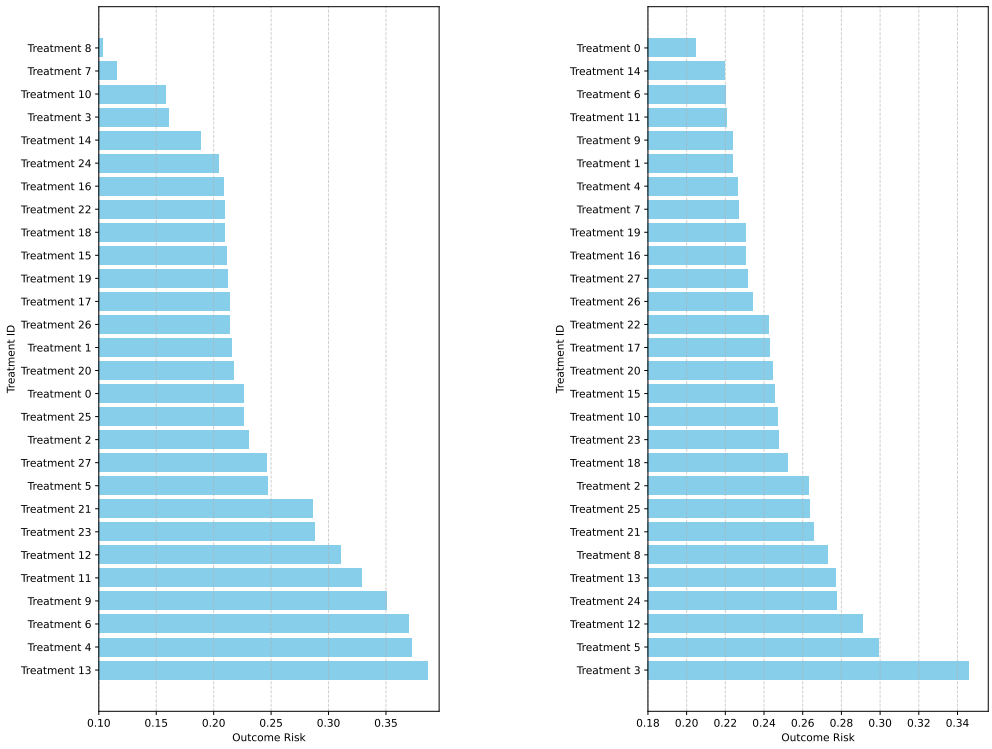
We extend the subgroup analysis in Section 4.6 to hypertension, diabetes, and racial and ethnic groups in *OneFlorida*. These comparisons describe subgroup-specific risks; they do not by themselves establish differences in treatment benefit or a uniform pattern for an entire drug class.

C.1 Hypertension-stratified treatment ranking

In Figure 4(a), T2 (β-blocker + RAAS) has higher estimated risk in patients with hypertension, whereas T3 (β-blocker + loop diuretic + RAAS), T6 (loop diuretic + RAAS), and T7 (MRAs + RAAS) have lower risk. T16 (RAAS + SGLT2) and T26 (β-blocker + RAAS + SGLT2) show smaller between-group differences than T2, T3, and T6. T24 (β-blocker + MRAs + RAAS + SGLT2) has relatively low risk in both groups, with a lower estimate in patients with hypertension.

C.2 Diabetes-stratified treatment ranking

Estimated risk is higher in patients with diabetes for most regimens, including T3, T7, T9 (β-blocker + loop diuretic + MRAs + RAAS), T10 (β-blocker + MRAs + RAAS), and T13 (loop diuretic + MRAs); T13 shows the largest plotted between-group difference (Figure 4(b)). T20 (RAAS + sacubitril/valsartan) and T23 (MRAs + SGLT2) show the opposite direction, with lower estimated risk in patients with diabetes. T24 has relatively low risk in both groups and a smaller difference than T16 or T26, whose estimated risks are low in patients without diabetes but visibly higher in those with diabetes.



(a) Hispanic (b) Non-Hispanic Black

Figure 5: Global rankings of all 28 treatment combinations on *OneFlorida* for (a) Hispanic and (b) Non-Hispanic Black patients, ordered by estimated marginal hospitalization risk (lower is better). Horizontal scales differ across panels.

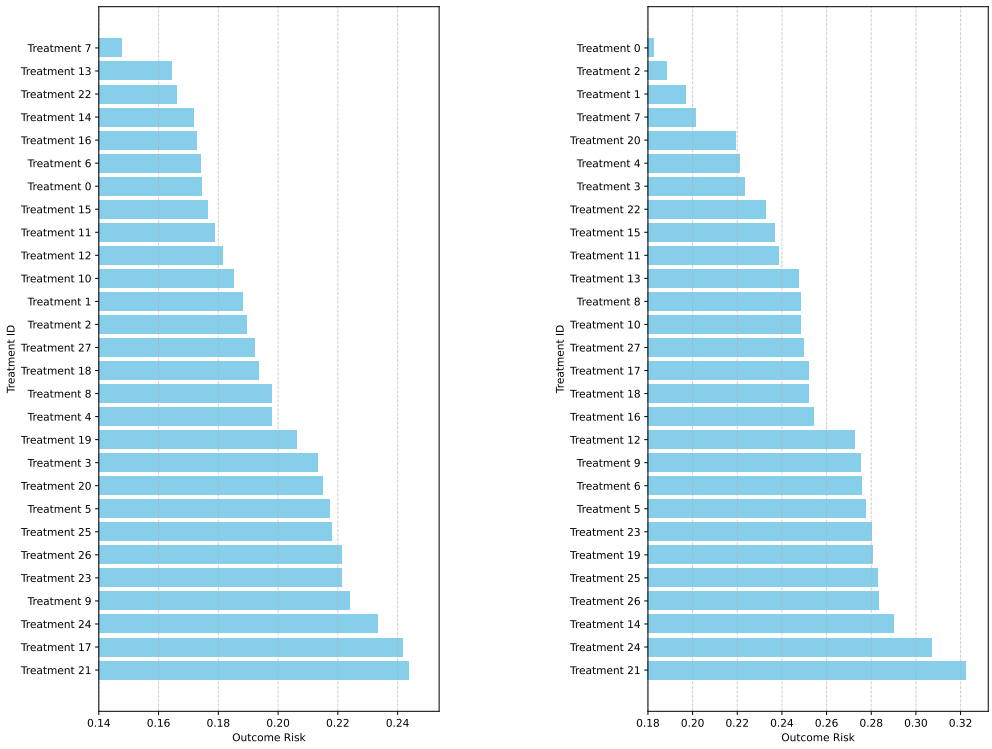
C.3 Race- and ethnicity-stratified treatment ranking

The leading regimens differ across the four groups in Figures 5 and 6. In Hispanic patients, T8 (MRAs), T7, and T10 rank first through third, with T3 fourth and T13 last. In Non-Hispanic Black patients, T0 (RAAS), T14 (loop diuretic + MRAs + RAAS), and T6 rank first through third, while T7 ranks eighth, T12 (β-blocker + loop diuretic + MRAs) twenty-sixth, and T3 last. The first three positions in Non-Hispanic White patients are occupied by T7, T13, and T22 (β-blocker + SGLT2). In the Other Races subgroup, T0, T2, and T1 (β-blocker) rank first through third, with T7 fourth.

The rankings of T24 and T26 vary substantially across groups. T24 ranks sixth in Hispanic patients but twenty-fifth, twenty-sixth, and twenty-seventh in Non-Hispanic Black, Non-Hispanic White, and Other Races patients, respectively. For T26, the ranks in these four groups are thirteenth, twelfth, twenty-third, and twenty-fifth.

The panels use different horizontal scales. The Hispanic group has the widest range of estimated risks and the Non-Hispanic White group the narrowest. Maximum estimated risk is highest in Hispanic patients, followed by Non-Hispanic Black, Other Races, and Non-Hispanic White patients. The width of a plotted range alone does not establish how uncertain or stable the rankings are.

These rankings do not establish intrinsic racial or ethnic differences in drug efficacy or identify the clinical or social mechanisms behind the observed patterns. Subgroup sample sizes, treatment support, and uncertainty in within-group treatment contrasts need to be assessed before drawing conclusions about differential benefit or making clinical recommendations.



(a) Non-Hispanic White

(b) Other racial/ethnic groups

Figure 6: Global rankings of all 28 treatment combinations on *OneFlorida* for (a) Non-Hispanic White patients and (b) patients of other racial/ethnic groups, ordered by estimated marginal hospitalization risk (lower is better). Horizontal scales differ across panels.

D Polynomial Semi-synthetic Causal Evaluation

This experiment evaluates treatment-effect recovery when the conditional risks under all candidate treatments are known. We retain real covariates and candidate drug combinations, but simulate treatment assignments and binary outcomes. The original hospitalization labels are neither used for fitting nor treated as causal ground truth. This appendix specifies the Polynomial experiment reported in Table 2.

D.1 Cohort, Treatment Space, and Data Splitting

The source cohort comprises 5,697 patients and 26,510 snapshots from *OneFlorida*. The semi-synthetic experiment uses 57 source covariates (24 continuous and 33 categorical), distinct from the 76-covariate description of the clinical experiment in Section 4.1.1. Patient identifiers, timestamps, existing treatment IDs, treatment and other medication indicators, and fields denoting outcomes or follow-up are excluded from the model inputs. Records with no candidate medication are excluded, and patient–snapshot-time keys are checked for uniqueness.

Let $z_t \in \{0,1\}^6$ denote the membership vector of candidate treatment t. The fixed candidate set contains $K = 28$ observed nonempty combinations, each containing one to five medication classes. Combinations are encoded lexicographically by their membership vectors; these simulation IDs need not match the clinical figure IDs. The candidate set uses medication membership only, without outcomes or simulated truth. Evaluation enumerates this same set, so it does not test extrapolation to unseen combinations or to all 2^6 possible subsets.

For each repetition, patients are randomly assigned to training, validation, and test groups in a 60:20:20 ratio, keeping all snapshots of a patient together. The specified training fraction retains 25% of the initial training patients, yielding 855 training, 1,139 validation, and 1,139 test patients; the remaining 2,564 patients are unused. Snapshot counts vary with the patient split. All target risks average over test snapshots, rather than assigning equal weight to each patient.

Additional preprocessing is fitted on the retained training group. Continuous variables are median-imputed and standardized using training means and standard deviations; all-missing columns use zero for imputation, and near-zero standard deviations are set to one. Categorical vocabularies are learned from training observations, with missing or unseen values mapped to an unknown category. Concatenating the continuous and one-hot categorical features gives a design vector, which is transformed elementwise as $x = \tanh(\text{design}) \in \mathbb{R}^d$. Models and the simulator use this same observed input. The source data already incorporate earlier clinical preprocessing; the training-only restriction applies to the additional transformations specified here.

D.2 Known Risks and Confounded Treatment Assignments

For each repetition, draw $W \in \mathbb{R}^{d\times 7}$ with $W_{jr} \sim \mathcal{N}(0, 1/d)$. The second argument of each normal distribution below denotes its variance. Using the retained training observations, compute the componentwise mean c and standard deviation s of xW, flooring s at 0.05, and define

$$h(x) = \tanh\{(xW - c)/s\}, \qquad h(x) = (h_0(x), \ldots, h_6(x)). \tag{6}$$

Draw medication main effects $b_j \sim \mathcal{N}(0, 0.3^2)$ for $j = 1, \ldots, 6$. For every medication subset A of size two through five occurring in at least one candidate combination, let $f_A(t) = \prod_{j\in A} z_{tj}$. Denote this collection by $\mathcal{S}$; it contains 14 pairwise, 15 three-way, seven four-way, and one five-way terms. Draw $\gamma_A \sim \mathcal{N}(0, 0.7^2/|A|)$ and $Q_{jA} \sim \mathcal{N}(0, 1/6)$. The Polynomial outcome mechanism is

$$\begin{aligned}
\eta_0(x,t) &= 0.8h_0(x) + \sum_{j=1}^{6}[b_j + 0.4h_j(x)]z_{tj}, \\
\eta(x,t) &= \eta_0(x,t) + \sum_{A\in\mathcal{S}}\left[\gamma_A + 0.45\tanh\left\{\sum_{j=1}^{6} h_j(x)Q_{jA}\right\}\right] f_A(t), \\
\mu_t(x) &= \text{sigmoid}\{b_0 + \eta(x,t)\}.
\end{aligned} \tag{7}$$

Thus, higher-order interactions enter the logit, can have either sign, and vary with observed patient characteristics. The latent features and generator coefficients are not supplied to the fitted models.

To generate confounding, let n_t be the original clinical count of treatment t in the retained training snapshots and set $\pi_t = \sqrt{n_t + 1} / \sum_{s=1}^{K} \sqrt{n_s + 1}$. Independently draw $U \in \mathbb{R}^{7 \times K}$ with $U_{rt} \sim \mathcal{N}(0, 1/7)$ and set $q_t(x) = \tanh\{h(x)^\top U_{\cdot t}\}$. The true assignment probabilities are

$$e_t(x) = \frac{0.20}{K} + 0.80 \, \frac{\exp\{\log \pi_t + 2q_t(x)\}}{\sum_{s=1}^{K} \exp\{\log \pi_s + 2q_s(x)\}}. \tag{8}$$

Assignments and outcomes share $h(x)$, producing observed confounding, while the uniform mixture ensures $e_t(x) \geq 0.20/K$. This mechanism does not introduce unmeasured confounding.

Calibrate b_0 by 70 bisection steps on $[-40, 40]$ so that the expected factual event rate on the retained training covariates satisfies

$$\frac{1}{N_{\text{train}}} \sum_{i \in \text{train}} \sum_{t=1}^{K} e_t(x_i) \mu_t(x_i) = 0.20. \tag{9}$$

Then sample $T_i \sim \text{Categorical}\{e_1(x_i), \ldots, e_K(x_i)\}$ and $Y_i \sim \text{Bernoulli}\{\mu_{T_i}(x_i)\}$. Realized event rates may differ from 20%. Each snapshot contributes one factual treatment and outcome. Conditional on the covariates and sampled mechanism, these draws are independent across snapshots. The complete risk and propensity arrays are retained separately as oracle data for evaluation.

D.3 Model Implementations and Training

The baselines follow the method implementations used in Table 1. HyperDR follows the treatment encoder and prediction modules described in Sections 3.1 and 3.2, with the simulation training settings specified below.

With width $d_h = w = 32$, the patient encoder is a $d \to 2w \to w$ multilayer perceptron with ELU activation and dropout. Treatment embeddings $\mathbf{h}_{e_t} \in \mathbb{R}^w$ are obtained by stacking the HGNN layers in Equation 1, followed by attention aggregation over the drug classes in each combination. Following Equation 2, the outcome module concatenates the patient and treatment embeddings and predicts

$$\widehat{\mu}_t(x) = \mu_\theta(x, e_t) = \sigma(g_\theta([\mathbf{r}(x) \,\|\, \mathbf{h}_{e_t}])), \tag{10}$$

where g_θ is a nonlinear multilayer perceptron with dimensions $2w \to w \to 1$, and $\sigma(\cdot)$ is the sigmoid function. The propensity-score module takes only the shared covariate representation $\mathbf{r}(x)$ as input and outputs probabilities over the K candidate treatments using softmax.

The simulation training configuration uses AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, dropout 0.1, gradient-norm clipping at 5, and at most 100 epochs. The overall objective is the batch mean of

$$\text{BCE}(Y, \widehat{\mu}_T(x)) + \alpha \, \text{CE}(T, \widehat{e}(x)) + \rho \left\{ \frac{Y - \widehat{\mu}_T(x)}{\max(\widehat{e}_T(x), 0.001)} \right\}^2, \tag{11}$$

with $\alpha = 1$ and $\rho = 10^{-4}$. The residual term backpropagates through both predictions. Outcome and propensity checkpoints minimize validation factual binary cross-entropy and treatment cross-entropy, respectively, and may come from different epochs. Training stops when neither criterion improves by more than 10^{-6} for 12 consecutive epochs. A propensity temperature from $\{1, 1.5, 2\}$ is subsequently chosen by validation cross-entropy. Oracle errors do not enter model selection.

For separately trained outcome models, the common propensity procedure compares a smoothed empirical treatment prior, multinomial logistic regression with $C \in \{0.01, 0.1, 1\}$, and a $d \to w \to K$ multilayer perceptron. The latter two also use validation-based temperature selection. The candidate with the lowest validation treatment cross-entropy is shared across these outcome models. Models with a propensity-score module use its fitted treatment probabilities for AIPW estimation. Accordingly, the main comparison assesses the full estimation procedure rather than isolating the treatment encoder alone.

D.4 Estimands and Evaluation Metrics

For the same N held-out test snapshots, define the true marginal risk $R_t = N^{-1} \sum_i \mu_t(x_i)$ and average treatment effect $\Delta_{ab} = R_a - R_b$. These contrast assigning the same population to treatment a or b, rather than comparing observed treatment-group event rates. With models fixed after

validation-based selection, estimate

$$\widehat{R}_t^{\mathrm{AIPW}} = \frac{1}{N}\sum_{i=1}^{N}\left[\widehat{\mu}_t(x_i) + \frac{\mathbf{1}(T_i = t)\{Y_i - \widehat{\mu}_t(x_i)\}}{\max\{\widehat{e}_t(x_i), 0.001\}}\right]. \tag{12}$$

Test factual treatments and outcomes enter this final estimate, but are not used to fit or select the models. The estimate is not normalized by the sum of weights, and the final risks are not clipped to $[0, 1]$. The propensity floor is a numerical stabilization choice and must be distinguished from the unmodified estimator's theoretical properties.

Let $P = K(K-1)/2 = 378$. The primary metric is

$$\mathrm{RMSE}_{\mathrm{ATE}}^{\mathrm{AIPW}} = \sqrt{\frac{1}{P}\sum_{a<b}\left[(\widehat{R}_a^{\mathrm{AIPW}} - \widehat{R}_b^{\mathrm{AIPW}}) - (R_a - R_b)\right]^2}. \tag{13}$$

For conditional effects, define $\tau_{i,ab} = \mu_a(x_i) - \mu_b(x_i)$ and $\widehat{\tau}_{i,ab} = \widehat{\mu}_a(x_i) - \widehat{\mu}_b(x_i)$. The secondary metric is

$$\mathrm{RMSE}_{\mathrm{CATE}}^{\mathrm{outcome}} = \sqrt{\frac{1}{NP}\sum_{i=1}^{N}\sum_{a<b}(\widehat{\tau}_{i,ab} - \tau_{i,ab})^2}. \tag{14}$$

This metric evaluates conditional mean risk differences directly from outcome predictions; it does not add an individual AIPW residual or evaluate realized binary potential-outcome differences. Both errors are reported on the probability scale, with lower values indicating better recovery. The CATE metric alone is not a test of double robustness.

D.5 Execution and Aggregation

We perform three repetitions, each following the same sequence:

1. Split patients and retain the labeled training subset; fit the additional covariate transformations using that subset only.
2. Sample the Polynomial outcome and assignment mechanisms, calibrate the intercept, and generate factual treatments and outcomes. Store oracle risks separately from the training inputs.
3. Fit all methods on the same observed data, choose checkpoints and calibration using validation data, and freeze the fitted predictors.
4. Predict all 28 candidate treatments for every test snapshot, compute AIPW risks and both effect-error metrics, and compare them with the stored truth on those same rows.
5. Aggregate the per-repetition metrics as mean and standard deviation, retaining the complete set of repetitions.

The simulator seed controls the patient split, mechanism, treatment/outcome draws, and model initialization through separate random streams. The independent repetition unit for aggregation is the seed, not a treatment pair or a snapshot. Table 2 reports the mean and sample standard deviation across the three repetitions; no statistical-significance claim is made from these summaries. The evaluation uses a patient-disjoint holdout design rather than cross-fitting. Its conclusions concern recovery under this specified observed-confounding mechanism and fixed candidate set, rather than identification of the original clinical effects or performance under unmeasured confounding.