Supplementary Information

S1 Datasets

S1.1 Semi-synthetic Data

In our synthetic data environment, we replicate the data generation process described in [38]. This approach generates semi-synthetic data based on the Twins, News, and ACIC2016 datasets. It's important to note that our primary focus is on varying confounding settings. Within these settings, treatment assignment is determined by:

$$\pi(x) = \text{sigmoid}(\omega_\pi * \phi(x)) \tag{8}$$

In this equation, ω_π stands for the confounding (propensity) scale, and $\phi(x)$ represents the confounding function. The confounding is either predictive, with $x = I_{pred}$, prognostic, defined by $x = I_{prog}$, or irrelevant, where $x = I_{irrelevant}$.

Twins. The Twins dataset [52] encompasses 11,400 twin births in the USA from 1989 to 1991. It includes 39 covariates—both continuous and categorical, related to parents, pregnancy, and birth details.

News. The News dataset comprises 10,000 randomly sampled news items, each characterized by 2,858-word counts [11, 53]. The dataset is processed with Principal Component Analysis, with the first 100 principal components serving as continuous covariates for each item.

ACIC2016. The ACIC2016 dataset consists of data from the Collaborative Perinatal Project provided as part of the Atlantic Causal Inference Competition (ACIC2016) [54][1]. It consists of 55 mixed (continuous and categorical) features for 2,200 patients.

S1.2 Real-World RCTs

IST-3 (The Third International Stroke Trial) [21] was designed to ascertain if a broader range of stroke patients could benefit from treatment with recombinant tissue plasminogen activator (rt-PA) when given within 6 hours after a stroke. The trial was randomized and involved 3,035 patients with acute ischemic stroke. Participants were subjected to either an rt-PA intervention or standard treatment without rt-PA. Functional outcomes were assessed using the Oxford Handicap Scale (modified Rankin scale) at 6 months. The trial found that in certain populations, the benefits of early thrombolytic treatment (with rt-PA) outweigh the risks. The data is available at `https://datashare.ed.ac.uk/handle/10283/1931`.

CRASH-2 (Clinical Randomisation of an Antifibrinolytic in Significant Haemorrhage) [22] had the objective to evaluate the effect of early administration of tranexamic acid (TXA) on mortality, surgical intervention, and transfusion requirements in trauma patients with, or at risk of, significant bleeding. The randomized trial encompassed over 20,000 trauma patients. The intervention was TXA versus placebo. The major outcomes of interest were death in hospital within 4 weeks post-injury, need for blood transfusion, and surgical interventions. The major discovery was that early administration of TXA reduced all-cause mortality without increasing the risk of vascular occlusive events. This dataset is available at `https://freebird.lshtm.ac.uk/index.php/available-trials/`.

SPRINT (Systolic Blood Pressure Intervention Trial) [24] [2] aimed to assess the effects of intensive versus standard blood pressure control on cardiovascular outcomes and mortality in the general population. It was a randomized controlled trial with 9,361 adults who had a systolic blood pressure of 130 mm Hg or higher and at least one additional cardiovascular disease risk factor. The participants underwent either an intensive treatment targeting systolic blood pressure below 120 mm Hg or standard treatment targeting below 140 mm Hg. The primary composite outcome included myocardial infarction, non-myocardial infarction acute coronary syndrome, stroke, heart failure, or death from cardiovascular causes. The pivotal finding was that intensive blood pressure control reduced the rate of the primary composite outcome and death from any cause compared to standard treatment.

[1] `https://github.com/AliciaCurth/CATENets`

[2] For ACCORD and SPRINT, both are available at `https://biolincc.nhlbi.nih.gov/home/` upon request.

ACCORD (Action to Control Cardiovascular Risk in Diabetes)[23] was set out to study the effects of intensive blood pressure control and various blood glucose and lipid interventions in type 2 diabetic patients. This set of randomized controlled trials involved 10,251 adults with type 2 diabetes. Interventions included intensive blood pressure control (targeting below 120 mm Hg) versus standard blood pressure control (targeting below 140 mm Hg); other arms of ACCORD also addressed glucose and lipid control strategies. The primary outcomes were nonfatal myocardial infarction, nonfatal stroke, or death from cardiovascular causes. Intensive blood pressure control did not significantly reduce the rate of the primary composite outcome compared to standard treatment; however, intensive glucose control reduced the rate of certain outcomes but increased mortality and certain lipid strategies showed benefits.

S1.3 Observational Data

Harborview pre-hospital TXA cohort The emergency medicine datasets used in this study were gathered over 13 years (2007-2020) and encompass 14,463 emergency department admissions. It is the only retrospective electronic health record (EHR) data that we used in our study. We excluded patients under the age of 18 and patients with hypotension, and we curated a clinical cohort with patients prescribed tranexamic acid (TXA) among trauma patients and the corresponding control group. The corresponding cohort group is selected based on propensity score matching [55]. The cohort consists of 240 patients with 120 in the treated group and 120 in the control group. We selected variables available in the pre-hospital setting, including trauma type, demographic information (age, sex), and pre-hospital vital signs (blood pressure, heart rate, respiratory rate). The outcome was each patient's survival. This dataset is not publicly available due to patient privacy concerns.

S1.4 Data Preprocessing

For the IST-3 and CRASH-2 datasets, we utilized features that were randomized at the baseline measurement. For the SPRINT and ACCORD datasets, we incorporated both the features randomized at the baseline measurement and additional clinical features. These additional features include the number of medications for blood pressure control, cardiovascular medications, history of cardiovascular disease, and history of chronic kidney disease.

For data preprocessing, we excluded features with over 90% missing values. For those features with missing values but below the threshold, we imputed the missing data using the empirical mean, with `SimpleImputer` from the *scikit-learn* package. Subsequently, all features were normalized to lie within the range [0, 1].

S2 Potential Outcome Framework

Using the Neyman–Rubin potential outcome framework[9], suppose a superpopulation $\mathcal{P}$ gives rise to a sample of N independent random variables, represented as $(Y_i(0), Y_i(1), X_i, W_i) \sim \mathcal{P}$. Here: - $X_i \in \mathbb{R}^d$ denotes a d-dimensional feature vector. - $W_i \in \{0, 1\}$ signifies the treatment assignment, with the specific meaning to be clarified later. - $Y_i(0)$ and $Y_i(1)$ are the potential outcomes for unit i when assigned to the control and treatment groups, respectively. From this, we denote the Average Treatment Effect (ATE), as:

$$ATE := \mathbb{E}[Y(1) - Y(0)] \tag{9}$$

A core challenge in causal inference is the inability to observe both potential outcomes for each unit. For each unit, either the outcome under control ($W_i = 0$) or under treatment ($W_i = 1$) is observed, but not both. Given this, the observed data is:

$$\mathcal{D} = (X_i, W_i, Y_i)_{1 \le i \le N} \tag{10}$$

To decide the treatment for a new individual i with covariate x_i, we aim to estimate its Individual Treatment Effect (ITE), $D_i = Y_i(1) - Y_i(0)$. However, since D_i remains unobserved and is not identifiable without robust assumptions, we focus on estimating the Conditional Average Treatment Effect (CATE), $\tau(x)$:

$$\tau(x) := \mathbb{E}[D|X = x] = \mathbb{E}[Y(1) - Y(0)|X = x] \tag{11}$$

Note that the optimal estimator for CATE is also the best for ITE in terms of Mean Squared Error (MSE) [41]. We aim for estimators that minimize the Expected Mean Squared Error (EMSE) for CATE estimation. This metric is also called the *precision in estimating heterogeneous effects (PEHE).*[56].

$$\tau^* = \arg\min_{\hat{\tau}} \mathbb{E}[(\hat{\tau}(x) - \tau(x))^2] \tag{12}$$

To be able to identify the causal effects from observational data, we make the standard assumptions for both domains.
Assumption 1. *(Unconfoundedness)* There are no unobserved confounders, such that the treatment assignment and POs are conditionally independent given the covariates:

$$Y(1), Y(0) \perp\!\!\!\perp W|X \tag{13}$$

Assumption 2. *(Overlap)* for all $x \in \mathbb{R}^d$,

$$0 < p(W = 1|X = x) < 1 \tag{14}$$

S2.1 CATE Models

S2.1.1 Meta Learners

In this section, we provide the background of meta-learners mentioned in the main text.

S-learner (Single Learner) The S-learner trains a single model on both treated and control units.

$$\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$$

where $\hat{\mu}(x, w)$ is the prediction of the model given covariate x and treatment w.

T-learner (Two Learner) [41] introduced the T-learner, an approach that creates distinct regression functions for each treatment group and computes the differences. The T-learner trains separate models for treated and control units.

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

where $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ are the predictions for treated and control units, respectively.

X-learner [41] also introduce the X-learner, a two-step regression estimator that uses each data point twice. The X-learner is designed to leverage both the treated and control groups for estimating heterogeneous treatment effects. To start, models $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$ are fitted on the treated and control groups respectively. Using these models, treatment effects for the treated group are calculated as $\hat{\tau}_1(X) = Y - \hat{\mu}_0(X)$ and for the control group as $\hat{\tau}_0(X) = \hat{\mu}_1(X) - Y$. Subsequently, counterfactual outcomes for the treated group are estimated using g_0 to get $\hat{\tau}_0(X)$ and for the control group using g_1 to get $\hat{\tau}_1(X)$. The final treatment effect estimate integrates these results using the propensity score, $e(X) = P(W = 1|X)$

$$\hat{\tau}(X) = e(X) \cdot \hat{\tau}_1(X) + (1 - e(X)) \cdot \hat{\tau}_0(X) \tag{15}$$

DR-learner (Doubly Robust Learner) [13] introduced the DR-learner, or doubly robust learner, a two-step procedure that employs the formula for the doubly robust augmented inverse propensity weighted (AIPW) estimator [57] as a pseudo-outcome in a two-step regression framework. The DR-learner combines propensity score weighting with regression adjustment.

$$\hat{\tau}(x) = \mathbb{E}[Y \mid X = x, W = 1] - \mathbb{E}[Y \mid X = x, W = 0] - \frac{W - e(X)}{e(X)(1 - e(X))}\big(Y - \hat{m}(X, W)\big)$$

where $e(X)$ is the propensity score.

R-learner [42] proposed the R-learner, which demands an estimate of the treatment-unconditional mean. The R-learner utilizes residuals to estimate the heterogeneous causal effect. Given a model, the residuals are computed as $R = Y - \hat{\mu}(X, W)$. The causal effect is then estimated as:

$$R_1 = Y - \hat{\mu}(X, W = 1) \quad R_0 = Y - \hat{\mu}(X, W = 0)$$

The causal effect is deduced from the relation between the residuals and the treatment assignment, given by:

$$\hat{\tau}(X) = \frac{\mathbb{E}[R_1 \cdot W \mid X]}{\mathbb{E}[R_0 \cdot (1 - W) \mid X]}$$

In this approach, the essence is to determine how deviations from the model's predictions (residuals) correlate with the treatment, adjusted by the propensity of receiving the treatment.

S2.1.2 Representation Learners

In addition to meta-learners, recent studies aim to learn a covariate shift function[11], ensuring that the feature representation of both treated and untreated groups have similar distributions. Algorithms such as CFRNet[11], TARNet [44], and Dragonnet [43] are introduced to predict CATE using balanced representations derived from observational data.

Specifically, the balanced learning approach identifies a representation, $h : \mathcal{X} \to \mathcal{R}$, and treatment-specific functions, μ_1 and μ_0, which aim to minimize the PEHE evaluation measure. The model is trained using a loss function that's bounded by PEHE. This function comprises the combined expected factual treated and control losses for outcome regression and the distance between $h(x)$ for given $W = 1$ and $W = 0$ values concerning the covariate shift.

$$\tau(x) := \mathbb{E}[Y(1) - Y(0)|X = h(x)] \tag{16}$$

Dragonnet Dragonnet employs a neural network architecture where both the treatment assignment and the outcome are jointly modeled. The network is structured to provide a representation $h(X)$ of the input covariates that captures the nuances needed to predict both treatment propensity and potential outcomes. By jointly modeling, Dragonnet ensures that the learned representations are informative about both treatment assignments and outcomes.

TARNet (Treatment Agnostic Representation Network) TARNet aims to learn a shared representation $h(X)$ of the covariates for both potential outcomes. Once this representation is determined, it is then passed through two separate outcome models to predict the potential outcomes under treatment and control:

$$\begin{aligned} h(X) &= \text{Shared representation} \\ \hat{Y}_1 &= \mu_1(h(X)) \quad \text{(Treated outcome model)} \\ \hat{Y}_0 &= \mu_0(h(X)) \quad \text{(Control outcome model)} \end{aligned}$$

The shared representation ensures that the outcome predictions are based on a consistent understanding of the covariates.

CFR (Counterfactual Regression) CFR focuses on deriving a shared representation $h(X)$ that can produce balanced representations of treated and control units. This balance ensures that the distributions of the treated and control units in the representation space are similar, minimizing the distributional shift and aiding in counterfactual prediction. With this balanced representation, potential outcomes are then estimated using respective outcome models.

DR-CFR (Doubly Robust Counterfactual Regression) DR-CFR combines the strengths of doubly robust estimation methods with the representation learning of CFR. After learning a balanced representation $h(X)$, DR-CFR not only predicts potential outcomes but also adjusts for discrepancies using propensity scores or outcome residuals, leading to a more robust and accurate estimate of treatment effects.

S2.2 CATE Model Evaluation

Here we introduce different evaluation strategies to validate CATE models.

S2.2.1 Evaluation Metrics

Factual criteria. [36] evaluate models by measuring simple prediction loss only on the observed potential outcomes.

$$\mathcal{E}_Y^{Fact}(\hat{\tau}) = \sqrt{\frac{1}{n}\sum_i^n (Y_i - \hat{\mu}_{A_i}(X_i))^2} \tag{17}$$

Obviously, one obvious disadvantage is that it may wrongly prioritize good fit on the potential outcomes over good CATE fit, resulting in bias [36].

Plug-in criteria. construct surrogates for CATE evaluation by fitting a new CATE estimator $\tilde{\tau}(x)$ on held-out data and using this to compare against the estimates.

$$\mathcal{E}_{\tilde{\tau}}^{Plug}(\hat{\tau}) = \sqrt{\frac{1}{n}\sum_i^n (\tilde{\tau}(X_i) - \hat{\tau}(X_i))^2} \tag{18}$$

However, one obvious disadvantage for a plug-in surrogate is that it would favor models with similar structures e.g. S-Learner and T-Learner, a phenomenon called congenital bias. [36, 38]

Pseudo-outcome surrogate criteria obtain estimate through given auxiliary nuisance estimates $\tilde{\eta} = \tilde{\mu_0}(x), \tilde{\mu}_1(x), \tilde{\pi}(x)$ obtained from the validation data using ML method M, one can construct pseudo-outcomes $Y_{\tilde{\eta}}^{\text{pseudo}}$ for which it holds that for ground truth nuisance parameter $\eta, \mathbb{E}[Y_\eta | X = x] = \tau(x)$ and – instead of using them as regression outcomes as in the learners themselves

$$\mathcal{E}_{Y_{\tilde{\eta}}}^{\text{pseudo}}(\hat{\tau}) = \sqrt{\frac{1}{n}\sum_i^n (Y_{\tilde{\eta}}^{\text{pseudo}} - \hat{\tau}(X_i))^2} \tag{19}$$

Here $Y_{\tilde{\eta}}^{\text{pseudo}}$ can be any pseudo-outcome objectives. In this work, we employ *Influence function*, *DR-Learner* objective, and *R-Learner* objective. Proposed by [13], DR pseudo-outcome employs doubly robust AIPW estimator and is hence unbiased if either propensity, $\tilde{\pi}$, or outcome regressions, $\tilde{u}_1$ and $\tilde{u}_0$, are correctly specified. $\mathcal{E}_{\text{DR}}(\hat{\tau})$ and $Y_{\tilde{\eta}}^{DR}$ are denoted as,

$$\mathcal{E}_{Y_{\tilde{\eta}}}^{\text{DR}}(\hat{\tau}) = \sqrt{\frac{1}{n}\sum_i^n (Y_{\tilde{\eta}}^{DR} - \hat{\tau}(X_i))^2} \tag{20}$$

$$Y_{\tilde{\eta}}^{DR} = (\frac{W}{\tilde{\pi}(X)} - \frac{(1-W)}{1-\tilde{\pi}(X)})Y + [(1 - \frac{W}{\tilde{\pi}(X)})\tilde{\mu}_1(X) - (1 - \frac{1-W}{1-\tilde{\pi}(X)}\tilde{u}_0(X)] \tag{21}$$

The R-learner objective of [42], which requires an estimate of the treatment-unconditional mean $\mu(x) = E[Y|X = x]$, relies on a similar idea and can also be used for the selection task, resulting in the criterion.

$$\mathcal{E}_{Y_{\tilde{\eta}}}^{\text{R}}(\hat{\tau}) = \sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \tilde{\mu}(X_i)) - \hat{\tau}(X_i)(W_i - \tilde{\pi}(X_i))^2} \tag{22}$$

The influence function objective [58], categorized as surrogate pseudo-outcome [36], leverages Talyor-like expansion to approximate the PEHE.

$$\mathcal{E}^{\text{IF}}(\hat{\tau}) = (1-B)\tilde{\tau}(X)^2 + BY(\tilde{\tau}(X) - \hat{\tau}(X)) - D(\tilde{\tau}(X) - \hat{\tau}(X))^2 + \hat{\tau}(X)^2 \tag{23}$$

where $D = W - \tilde{\pi}(X), B = 2A(A - \tilde{\pi}(X))C^{-1}$ and $C = \tilde{\pi}(X)(1 - \tilde{\pi}(X))$, and $\tilde{\tau}(X)$ is a plug-in estimate. Surrogate pseudo-outcome is an unbiased estimator for CATE and more robust to congenital bias [36]. On the theoretical side, the analysis in [59] shows that – under certain technical conditions – minimizing the surrogate validation estimate is a consistent model selection rule.

Qini Score and Uplift score. While pseudo-outcome surrogate provides an estimation for model selection, evaluating the performance of the chosen model on a different cohort can be difficult. As it requires training a second CATE model and nuisance functions. Therefore, the Qini curve is often used to evaluate model performance across different cohorts. It is defined as follows:

$$\text{Qini Curve}(\phi, \hat{\tau}) = \left(\frac{n_{t=1,y=1}(\phi(\hat{\tau}))}{N_{t=1}} - \frac{n_{t=0,y=1}(\phi(\hat{\tau}))}{N_{t=0}} \right) \tag{24}$$

where $\phi(\hat{\tau})$ is the fraction of the population treated (in either treatment $t = 1$ or control $t = 0$) ordered by the uplift (treatment effect) from the model, $\hat{\tau}$. N_t represent the total individual count in each group. Similarly, the uplift score can also be used to evaluate CATE model but only considering the treatment uplift within ϕ population. The uplift score is calculated as:

$$\text{Uplift Curve}(\phi, \hat{\tau}) = \left(\frac{n_{t=1,y=1}}{N_{t=1}}(\phi(\hat{\tau})) - \frac{n_{t=0,y=1}}{N_{t=0}}(\phi(\hat{\tau})) \right) \tag{25}$$

A baseline method is a model that cannot distinguish positive and negative uplift, and rank patients at random[3]. To quantify uplift/treatment effect from a given model, it is common to measure the area under the curve (AUROC) between the Qini/Uplift curve (intervention assignment based on model predictions) and random choice.

S2.2.2 Model Implementation

For CATE model implementations, we used PyTorch for all models, which were sourced from the CATENets Python package[4]. Each estimated function (i.e., $\hat{\mu}_w(x)$, $\hat{\pi}(x)$, and $\hat{\tau}(x)$) was constructed with equal numbers of hidden layers and units. Specifically, all functions had 2 hidden layers with 100 units each, along with a final prediction layer. For architectures like TARNet and CFRNet, the representation Φ and the outcome heads h_w each contained 1 hidden layer. We used dense layers, integrating the ReLU activation function. We use dense layers with the ReLU activation function. All models are trained using the Adam optimizer with a learning rate 10^{-4} batch size of 1024 and early stopping on the validation set (which represents 30 % of the initial training set).

[3]Note that this is different from the random treatment assignment in RCT.

[4]`https://github.com/AliciaCurth/CATENets`

S3 Interpretability (XAI) Methods

S3.1 Key Explanation Properties for CATE Models

Considering the CATE setting, [38] suggests the following key properties for explanation methods, $a_i, i \in [d]$, in a CATE model.

Sensitivity. The covariates that do not affect the CATE model are given zero contribution. More formally, if for some $i \in [d]$ we have $\hat{\tau}(x) = \hat{\tau}(x_{-i})$ for all $x \in \mathcal{X}$, then $a_i(\hat{\tau}, x) = 0$ for all $x \in \mathcal{X}$.

Completeness. Summing the importance scores gives the shift between the CATE and a baseline. More formally, for all $x \in \mathcal{X}$, we have:

$$\sum_{i=1}^{d} a_i(\hat{\tau}, x) = \hat{\tau}(x) - b, \tag{26}$$

where $b \in R$ is a constant baseline. In this way, each importance score a_i can be interpreted as the contribution from covariate i of$x$ to have a CATE that differs from the baseline b. Note that the choice of the baseline differs from one method to another. For instance, the baseline for SHAP[18] is the average treatment effect: $b = \mathbb{E}_{X\sim P}[\hat{\tau}(X)]$.

Linearity. The importance score of a covariate is linear with respect to a black-box function. If the CATE model $\hat{\tau}$ is written in terms of the estimated potential outcomes $\hat{\mu_1} - \hat{\mu_0}$, it can be written as $a_i(\hat{\mu_1}, x) - a_i(\hat{\mu_0}, x)$. This makes it easy to differentiate prognostic and predictive covariates. If x_i is a prognostic covariate, then $a_i(\hat{\tau}, x)$ is zero. If x_i is a predictive covariate, then $a_i(\hat{\tau}, x)$ is not zero.

Model Agnosticism. The feature importance score can be computed for all CATE model$\hat{\tau} : \mathcal{X} \to \mathcal{Y}$. Some methods only work with a restricted family of models, which prevents them from being model agnostic. For example, Gradient-based explanation methods [37, 60] only work for $\hat{\tau}$ that is differentiable with respect to its input.

Implementation Invariance. The feature attribution would be the same for two functionally equivalent models. This means that if we have two CATE models $\hat{\tau}_1$ and $\hat{\tau}_2$ such that $\hat{\tau}_1 = \hat{\tau}_2$ for all $x \in \mathcal{X}$, this implies that $a_i(\hat{\tau}_1, x) = ai(\hat{\tau}_2, x) =$ for all $x \in \mathcal{X}$ and all $i \in [d]$.

S3.2 Explanation Methods

In this study, we utilize explanation methods that fulfill the following criteria, including *sensitivity, completeness, linearity*, and *implementation invariance*. Accordingly, we employ Integrated Gradients[17] and Shapley values[18]. Although Vanilla Gradient (Saliency)[60] doesn't satisfy the completeness criterion, it's served as a basic baseline. All the feature importance methods are implemented using Pytorch and Captum Python package[5].

Vanilla Gradient The vanilla gradient, often referred to as the saliency map[60], is frequently used as a benchmark for different explanation techniques. It is derived from the gradient of the output function in relation to the input features. Specifically, within CATE models, it assesses how the estimated treatment effect changes with respect to a given feature x_i.

$$\nabla x_i \tau(x) = \frac{\partial \tau(x)}{x_i} \tag{27}$$

Integrated Gradients Integrated Gradients (IG) assigns importance to input features by approximating the integral of a model's gradients from a baseline input to the actual input [17]. This method provides a holistic view of feature contributions, satisfying key properties like completeness. The IG attribution for an explicand x, a variable x_i, and a baseline x' is:

$$\mathrm{IG}_i(x, x', \tau) = (x_i - x'_i) \int_{\alpha=0}^{1} \frac{\partial \tau(x' + \alpha(x - x'))}{x_i} d\alpha \tag{28}$$

Typically, the zero vector serves as the baseline, denoted as $x' = 0$. This means feature contributions are measured relative to their absence. By integrating over a path, IG provides a comprehensive insight into a feature's importance, overcoming the limitations of simple gradient-based approaches.

[5] `https://captum.ai/`

Shapley Value Sampling Shapley Value, a concept borrowed from cooperative game theory, offers a unique approach to feature attributions[18]. For any prediction model, it assigns each feature an importance value by averaging all possible combinations of feature presence or absence. Mathematically, for a prediction model τ, the exact Shapley value for a feature x_i is defined as:

$$\Phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [\tau(x_{S \cup \{i\}}) - \tau(x_S)] \tag{29}$$

Where N is the set of all features and S is any subset of N that does not include feature x_i. This equation evaluates the contribution of the feature x_i by contrasting the prediction with and without the feature over all possible combinations.

However, computing the exact Shapley value can be computationally intensive, especially for models with a large number of features. Therefore, in practice, an approximation method like Shapley Value Sampling [19] is often used. Given a feature set N, and for a particular feature x_i, the sampled Shapley value is estimated as:

$$\hat{\Phi}_i(x) = \frac{1}{M} \sum_{m=1}^{M} \left(\tau(x_{S_m \cup \{i\}}) - \tau(x_{S_m}) \right) \tag{30}$$

Where M is the number of sampled orderings and S_m is a random subset of N without feature x_i in the m^{th} sampled ordering. By sampling, we approximate the Shapley value, making it feasible for practical applications while maintaining a close approximation to the exact value.

Baseline Shapley To obtain explanation/feature attribution for CATE models, we used the Shapley value [18, 61], which is the average expected marginal contribution of adding one feature to the treatment effect after all possible combinations of features have been considered. More formally, the Shapley value takes as input a set function $v : 2^N \to R$. The Shapley value produces attributions s_i for each player $i \in N$ that add up to $v(N)$. The Shapley value of a player i is given by:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{31}$$

In other words, the Shapley value of player i is the average weighted average of its marginal contribution $v(S \cup \{i\}) - v(S)$. Since computing the exact shapley value is intractable, several approximation methods have been proposed [18, 19]. Also, different choices of set function could lead to different Shapley values $\phi(i)$. One choice is to define $v_{f,x}(S)$ as the conditional expected model output on a data point when only the features in S are known [45]

$$v_{f,x}(S) = \mathbb{E}[f(X)|X_S = x_s] \tag{32}$$

However, computing conditional distribution, $X_{\bar{S}}|X_S$, is usually difficult. Empirically, methods such as KernelSHAP[18] and shapley value sampling[19], conditional expectations are estimated by assuming feature independence; samples of the features in $\bar{S} = D \setminus S$ are drawn from the marginal joint distribution of these variables.

$$v_{f,x}(S) = \mathbb{E}_{X_{\bar{S}}|X_S}[f(x_s, X_{\bar{S}})] \approx \mathbb{E}_{X_{\bar{S}}}[f(x_s, X_{\bar{S}})] \tag{33}$$

Moreover, we also leverage the assumption of model linearity to simplify the computation of the expected values [18]. Empirically, we approximate the marginal distribution with an empirical mean of features as our baseline.

$$v_{f,x}(S) = f(x_s, \mathbb{E}[X_{\bar{S}}]) \approx f(x_s, \frac{1}{n} \sum X_{\bar{S}}]) \tag{34}$$

This approach is called Baseline Shapley (BShap)[45] and is more computationally efficient. More generally, this approach models a feature's absence using its value in the baseline x'.

$$v_{f,x}(S) = f(x_s, x'_{N \setminus S}) \tag{35}$$

Another extension from baseline Shapley is Random Baseline Shapley (RBShap)[15, 45] where the baseline $X_{\bar{S}}$ is drawn randomly according to the data distribution D. When propagating Shapley values, it approximates equation 33. However, we observed that this approach requires a significant amount of time to reach convergence. Therefore, in this work, we utilize BShap to explain CATE models.

S3.3 Examination of Explanation Method in CATE

Predictive Features Identification with Ground Truth

Following the works of [38, 62], within a semi-synthetic environment, the health outcome, denoted as μ_W, can be expressed as a function of the patient characteristics X and the treatment W:

$$\mu_W(X) = \sum_j \alpha_j I_{prog,j} + \sum_k \beta_k I_{pred,k} W \tag{36}$$

Where $I_{prog,j}$ and $I_{pred,k}$ denote the j^{th} and k^{th}component of X that are prognostic and predictive covariates respectively. α_j and β_k are the corresponding coefficients. Under this assumption, the treatment effect can then be defined as:

$$\tau(X) = \mu_1(X) - \mu_0(X) = \sum_k \beta_k I_{pred,k} \tag{37}$$

Therefore, in the context of CATE estimation, an explanation method should differentiate between predictive and prognostic covariates. To evaluate this, we can compute the average proportion of attribution correctly assigned to the predictive covariates:

$$Attr_{pred} = \frac{1}{|D_{test}|} \sum_{X \in D_{test}} \frac{\sum_{i \in I_{pred}} |a_i(\hat{\tau}, X)|}{\sum_{i \in D_{test}} |a_i(\hat{\tau}, X)|} \tag{38}$$

In this equation, a_i denotes the attribution score for a given feature i. This metric principally assesses the ability of an explanation method to differentiate between predictive and prognostic factors. While actual explanations or ground truths are absent in counterfactual predictions, this approach can offer guidance when choosing which explanation method is best suited for real-world datasets.

Ablation and Insertion and Deletion with Pseudo-outcomes

In practical scenarios, it's usually impractical to obtain oracle features, therefore, to overcome this, one way is to perform an ablation test with CATE models. The ablation test is a commonly used evaluation technique for attribution values [25]. Its concept involves replacing features with baseline feature values based on their attributions to measure their impact on the evaluation metric. This can be iteratively described using modified versions of the original explicands. For a given mixture of explicands, denoted by $X^e \in \mathbb{R}^{n_e,m}$, the attribution score is represented as $\phi(f, X^e)$. The ablation study is characterized by three distinct parameters: (1) feature ordering, (2) an imputation sample, represented as $x^b \in \mathbb{R}^m$, and (3) an evaluation metrics.

$$X^{e,0} = X^e \tag{39}$$

$$X^{e,k} = X^e \odot I_k(\phi) + X^b \odot (1 - I_k(\phi)), \quad \forall k \in 1 \dots m. \tag{40}$$

Where $X^b := [x^b \dots x^b]^T$ and $I_k(\phi(f, X^e)) = \arg\max_{k,\text{axis}=1}(\phi(f, X^e))$. The latter yields an indicator matrix the same size as G. A value of 1 indicates its position among the maximum k elements across a specific axis. The operation $\odot$ represents the Hadamard product. Further, to assess the results of the ablation test, we compute the mean CATE output. When k features are ablated, this average is given by:

$$\frac{1}{N} \sum_{i=1}^{N} \hat{\tau}(X_i^{e,k}) \tag{41}$$

To contextualize the ablation test using treatment effects, imagine these features represent different features in a clinical trial, and their attributions (ϕ indicate the importance of each feature in predicting a particular effect of treatment, say patient recovery. When employing ablation: (1) Positive ablation: If we ablate (remove) features with the most positive attributions (those believed to have the most significant positive effect), we'd anticipate a substantial drop in, for example, patient recovery rates (mean model output). As we progressively remove additional features, this rate will continue to decline but at a diminishing pace. In this scenario, steeper declines imply that our attributions of features' contributions are accurate. (2) Negative ablation: on the other hand, when removing features with the most negative attributions (those believed to hinder recovery), we'd expect recovery rates to increase significantly. More steep inclines would indicate better attributions, suggesting that the dropped features were indeed important.

In addition to CATE output, we can also leverage pseudo-outcome surrogates **??**, measuring the estimated PEHE when features are removed.

$$\mathcal{E}^{\text{pseudo}}_{Y_{\tilde{\eta}}}(\hat{\tau}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{\tau}(X_i^{e,k}) - Y^{\text{pseudo}}_{\tilde{\eta}})^2 \tag{42}$$

The term $Y_{\tilde{\eta}}^{\text{pseudo}}$ refers to the pseudo-outcome, such as the R or DR objectives, further detailed in **??**.

In addition to analyzing positive or negative ablations, insertion and deletion tests provide another way to assess performance. To conduct an insertion test, start with a baseline where no features are present and then gradually add features based on their ranking. Similarly, when assessing recovery rates, positive attributions should improve them while negative ones should have a negative impact.

S4 Additional Experiment Results

S4.1 Identifying Important Features to CATE in Semi-synthetic Environment

In this section, we evaluate explanation methods' abilities in identifying important features in semi-synthetic environments S1 where we have access to oracle predictive features. These tasks include (1) identification of predictive features, (2) ablation studies with pseudo-outcomes, and (3) knowledge distillation with identified features.

Feature Identification Decrease as Confounding Effect Increases

Within the synthetic environment, we evaluated the performance of different explanation methods, Appendix S3.3. Our results indicate that the Shapley value with the mean baseline consistently outperforms Integrated Gradients (IG) across various confounding effects in all datasets. However, as confounding effects intensify, the performance of all methods declines in tandem with the decreasing performance of the CATE model, Figure S5 (a).

Insertion & Deletion Procedure with Pseudo-outcomes

We examine explanation methods with proposed evaluation task –ablation tests utilizing pseudo-outcomes. Notably, despite disparities between ϵ_{pseudo} and ϵ, their abilities in differentiating the performance of explanation methods are consistent. Results of ablation studies with ϵ_{pseudo} align with the findings from the one with true ϵ, as shown in Figure S5 (b).

For the insertion procedure, where we start with a baseline and sequentially insert features based on their local rankings, both ϵ and ϵ_{pseudo} stabilize at approximately 8 features for IG, BShap - 0, and BShap - mean. In contrast, they settle at around 13 features for the vanilla gradient. On the other hand, in the deletion procedure, which is characterized by the stepwise removal of features starting from the most crucial, we find that IG, BShap - 0, and BShap - mean reach a steady state at around 6 features. In comparison, the saliency gradients exhibit stability a bit later, approximately at the 10-feature.

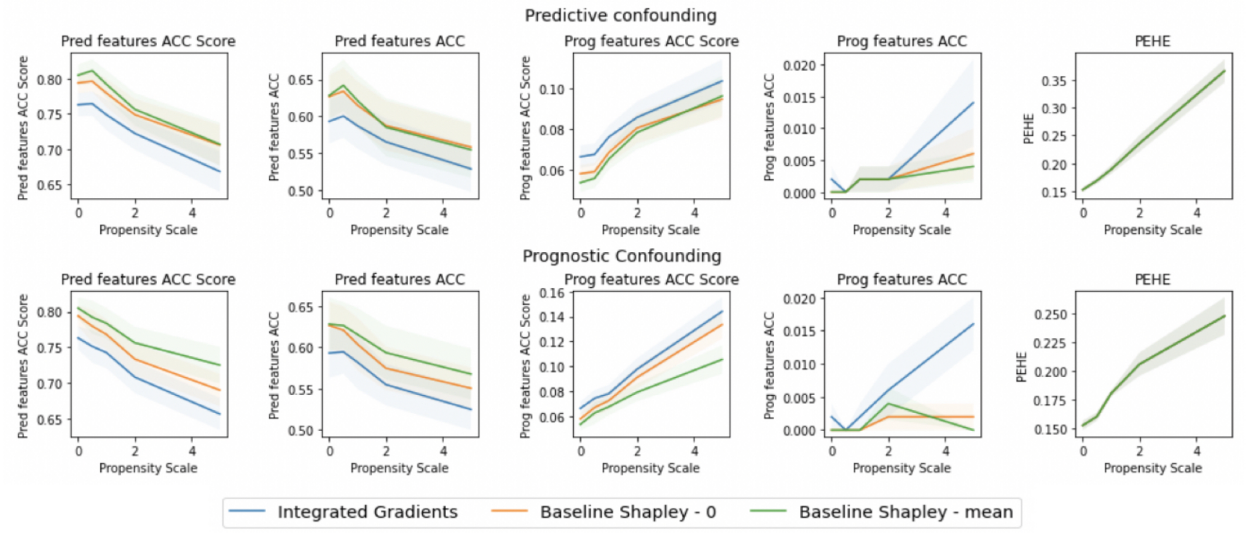
However, it's worth noting that in semi-synthetic setups, the performance differential between IG and baseline Shapleys is almost negligible. Furthermore, similar to findings in S5 (a), the performance shapley values with different baselines are indistinguishable in semi-synthetic contexts.

Knowledge Distillation as a Benchmark Test

In this section, we utilize model distillation as a means to assess various explanation methods in terms of global feature identification. Specifically, every student model adopts a 2-layer MLP architecture and the same hidden dimension. For each explanation method, we select the top features based on the average absolute attribution score. The results of this approach are presented in Table S3, showing the performance of the student models on the test set across synthetic environments with different confounding effects.

To ensure the robustness of our results and to identify potential inconsistencies in the student model, we begin by training the student with a feature count, n, equivalent to the full set of predictive features utilized during data generation. The results, as depicted in Table S3, indicate Mean Squared Errors (MSE) of 0.007, 0.028, and 0.003 for the Twins, ACIC, and News datasets respectively, across all the explanation methods evaluated.

However, when the number of input features is halved from the original set of predictive features, an expected decline in the student models' predictive performance emerges. Notably, the saliency method lags behind its counterparts IG, BShap, and BShap across all datasets. These observations are aligned with our ablation studies presented in S5. An intriguing finding is that, despite BShap's superior performance in identifying predictive features, Fig S5 (a), there's a negligible difference among most explanation methods regarding global feature selection, except the vanilla gradient approach.



(a)

(b)

Fig. S5 | (a) Identifying Predictive & Prognostic Features in confounding environment with semi-synthetic data: ACIC dataset comprises 20 predictive (10 pred_0 + 10 pred_1) and 10 prognostic features among a total of 55. The x-axis represents the (confounding) propensity scale, ω, while the y-axis depicts the percentage of features identified. **(b)Analyzing Insertion and Deletion via Pseudo-outcome surrogates**: This illustrates insertion (left panel) and deletion (right panel) analyses, employing both the R-Learner objective, $\mathcal{E}^{\text{R}}_{Y_{\tilde{\eta}}}(\hat{\tau})$(top), and the DR-Learner objective, $\mathcal{E}^{\text{DR}}_{Y_{\tilde{\eta}}}(\hat{\tau})$ (bottom). The ground truth, ϵ, in the ACIC dataset is used with a propensity scale of $\omega = 0.5$.

Datasets	Twins			ACIC			News		
Num of features	2	4	8	3	5	10	5	10	20
Saliency	0.045	0.040	0.007	0.096	0.077	0.028	0.010	0.008	0.003
IG	0.031	0.022	0.007	0.069	0.049	0.028	0.011	0.007	0.003
BShap - 0	0.031	0.021	0.007	0.069	0.050	0.028	0.011	0.007	0.003
BShap - mean	0.031	0.022	0.007	0.069	0.050	0.028	0.011	0.007	0.003

Table S3 | Knowledge Distillation Performance: Average MSE of student models relative to the original model, based on various feature counts in synthetic contexts. Environments encompass predictive and prognostic confounding with propensity scales $\omega = \{0, 0.5, 1, 2, 5, 10\}$. Features are selected at rates of $\frac{n}{4}$ and $\frac{n}{2}$, with n being the total feature count for semi-synthetic data: 8 (Twins), 10 (ACIC), and 20 (News).

S4.2 CATE Evaluation in Clinical Datasets

In evaluating the Conditional Average Treatment Effect (CATE), we employed the pseudo-outcome surrogate approach, Appendix S2.2. The experimental results are aggregated from 50 iterations of each model, each initialized with unique random seeds. A significant variance was observed with the influence function-based surrogate outcome metric, $\epsilon_{\texttt{if-pehe}}$. Owing to this instability, our model selection was principally driven by the more robust metrics, ϵ_{DR} and ϵ_R, aligning with the findings of[36].

Our evaluations revealed minimal disparity in terms of the pseudo-outcome surrogate metrics ϵ_{DR} and ϵ_R. Given that a majority of our datasets are sourced from randomized control trials—except the pre-hospital TXA dataset—this is anticipated. Especially in scenarios where the confounding (propensity) scale approaches 0, we observed a consistent identification of significant features across various explanatory methodologies, corroborated by the work of Crabbe et al. [38]. We also observed that, for representation learners, the computational time for explanation is significantly longer than CATE models. Given these considerations, the X-Learner was predominantly selected as the model to be interpreted throughout our experiments.

	XLearner	DRLearner	SLearner	TLearner	RLearner	RALearner	TARNet	DragonNet	CFRNet-0.01
					IST-3				
$\epsilon_{\text{if-pehe}}$	30.698 (31.717)	79.697 (29.885)	11.853 (36.463)	32.703 (35.360)	15.956 (33.668)	13.488 (33.059)	2.796 (30.644)	3.554 (31.203)	3.533 (34.293)
ϵ_R	0.467 (0.011)	0.476 (0.009)	0.461 (0.010)	0.465 (0.011)	0.461 (0.010)	0.462 (0.011)	0.459 (0.010)	0.459 (0.010)	0.459 (0.011)
ϵ_{DR}	0.952 (0.028)	0.969 (0.023)	0.941 (0.025)	0.949 (0.029)	0.941 (0.027)	0.942 (0.027)	0.937 (0.026)	0.937 (0.026)	0.937 (0.027)
					CRASH-2				
$\epsilon_{\text{if-pehe}}$	38.299 (51.343)	56.379 (44.205)	31.871 (43.959)	36.881 (47.693)	26.977 (42.618)	22.135 (45.608)	30.643 (38.574)	30.751 (39.253)	31.786 (37.921)
ϵ_R	0.355 (0.011)	0.358 (0.011)	0.354 (0.010)	0.355 (0.012)	0.353 (0.010)	0.353 (0.010)	0.352 (0.010)	0.352 (0.010)	0.352 (0.010)
ϵ_{DR}	0.672 (0.022)	0.678 (0.021)	0.668 (0.020)	0.670 (0.022)	0.666 (0.020)	0.666 (0.020)	0.663 (0.020)	0.664 (0.020)	0.664 (0.020)
					SPRINT				
$\epsilon_{\text{if-pehe}}$	37.590 (23.453)	51.935 (31.377)	3.369 (25.799)	42.112 (29.610)	2.738 (35.172)	7.261 (23.848)	1.171 (20.816)	1.563 (20.718)	1.230 (20.793)
ϵ_R	0.411 (0.038)	0.431 (0.044)	0.386 (0.040)	0.400 (0.039)	0.389 (0.043)	0.388 (0.040)	0.384 (0.042)	0.384 (0.042)	0.384 (0.042)
ϵ_{DR}	0.859 (0.078)	0.885 (0.088)	0.796 (0.086)	0.820 (0.082)	0.813 (0.089)	0.800 (0.085)	0.796 (0.091)	0.795 (0.090)	0.796 (0.091)
					ACCORD				
$\epsilon_{\text{if-pehe}}$	20.94 (28.55)	71.00 (36.50)	11.49 (28.39)	9.42 (22.96)	10.25 (29.10)	8.58 (26.37)	13.41 (17.82)	12.52 (18.41)	14.37 (16.90)
ϵ_R	0.320 (0.013)	0.331 (0.013)	0.314 (0.013)	0.317 (0.012)	0.315 (0.013)	0.315 (0.013)	0.313 (0.013)	0.313 (0.013)	0.313 (0.013)
ϵ_{DR}	0.627 (0.022)	0.652 (0.023)	0.618 (0.021)	0.623 (0.019)	0.619 (0.021)	0.618 (0.021)	0.615 (0.021)	0.615 (0.021)	0.615 (0.021)
					Pre-hospital TXA				
$\epsilon_{\text{if-pehe}}$	73.161 (185.123)	226.653 (441.105)	5.984 (116.521)	134.012 (411.408)	45.525 (172.293)	9.841 (130.707)	2.796 (30.644)	3.554 (31.203)	3.533 (34.293)
ϵ_R	0.468 (0.050)	0.491 (0.041)	0.465 (0.047)	0.466 (0.053)	0.467 (0.045)	0.465 (0.051)	0.459 (0.010)	0.459 (0.010)	0.459 (0.011)
ϵ_{DR}	5.254 (6.285)	5.283 (6.281)	5.260 (6.294)	5.254 (6.303)	5.251 (6.274)	5.261 (6.288)	0.937 (0.026)	0.937 (0.026)	0.937 (0.027)

Table S4 | CATEs Evaluation on Clinical Datasets with pseudo-outcome surrogates including $\epsilon_{\text{if-pehe}}$, ϵ_{DR} and ϵ_R described in Appendix S2.2. The lower the better.

	XLearner	XLearner(Ensemble)
	IST-3	
$\mathcal{E}_{\text{if-pehe}}$	30.698 (31.7)	**25.3 (22.1)**
$\mathcal{E}_{\text{R}}$	0.467 (0.011)	**0.41 (0.009)**
$\mathcal{E}_{\text{DR}}$	0.952 (0.028)	**0.85 (0.023)**

Table S5 | Comparative Performance of Single vs. Ensemble Models Using Pseudo-outcome Surrogate: in IST-3.

S4.3 Insertion & Deletion Procedure with Pseudo-outcomes - Real-world Data

Influence of Baseline Choices in Ablation Studies: Potential Bias in Evaluation

In this part, we examine explanation methods with popular insertion and deletion experiments S3, extracting relevant features with respect to pseudo-outcomes, ϵ_{pseudo}. The process of insertion and deletion is shown in **??**(a). For the deletion plot, as negative features are removed sequentially based on their contribution, the model's prediction error escalates, and then as positive features are removed, the model's prediction error starts to decrease. With all features removed, the prediction error increases or deviates more from 0. However, we observed that, in contrast to the semi-synthetic environment, this metric is largely affected by the replacement baseline [15] with real-world data, as shown in **??**(a). We can observe that, with different replacement values, one can choose a baseline value that favors a particular explanation method. Similar results can be seen in the ablation studies where the ranking largely depends on the baseline. We conclude that, with real-world data, with different replacement values, it's difficult to determine which explanation methods to use only with ablation studies. Therefore, we propose an alternative approach for evaluation explanation methods.

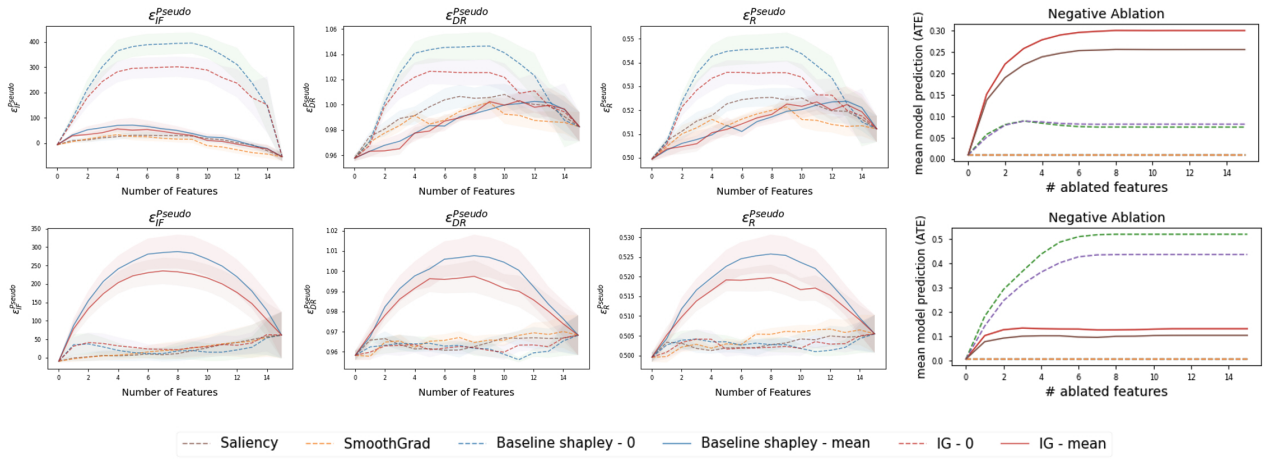
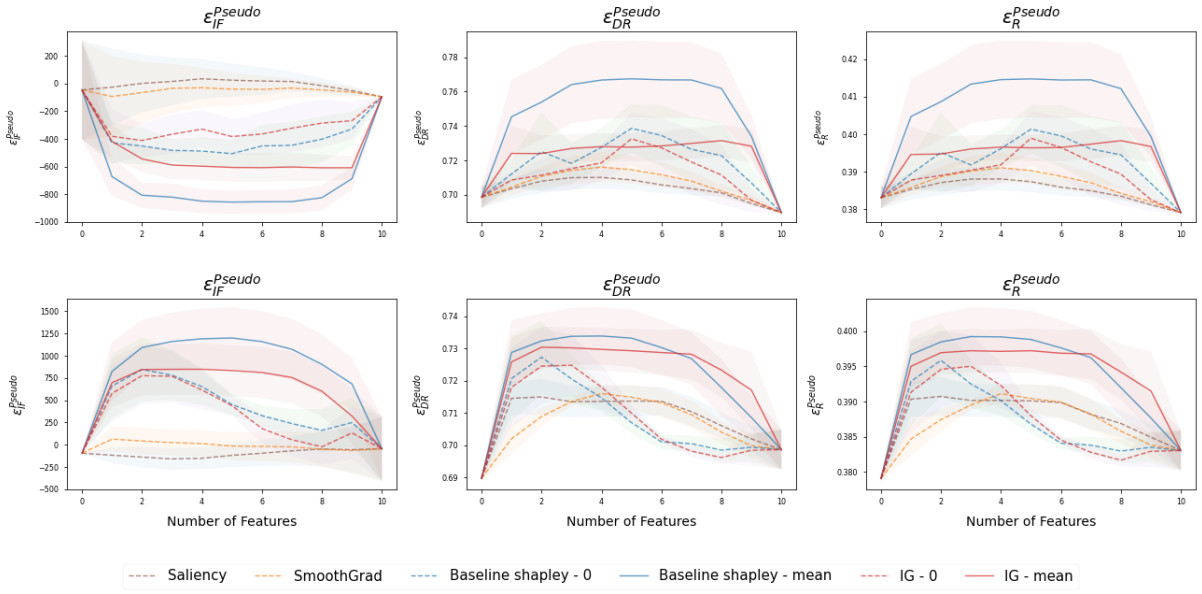
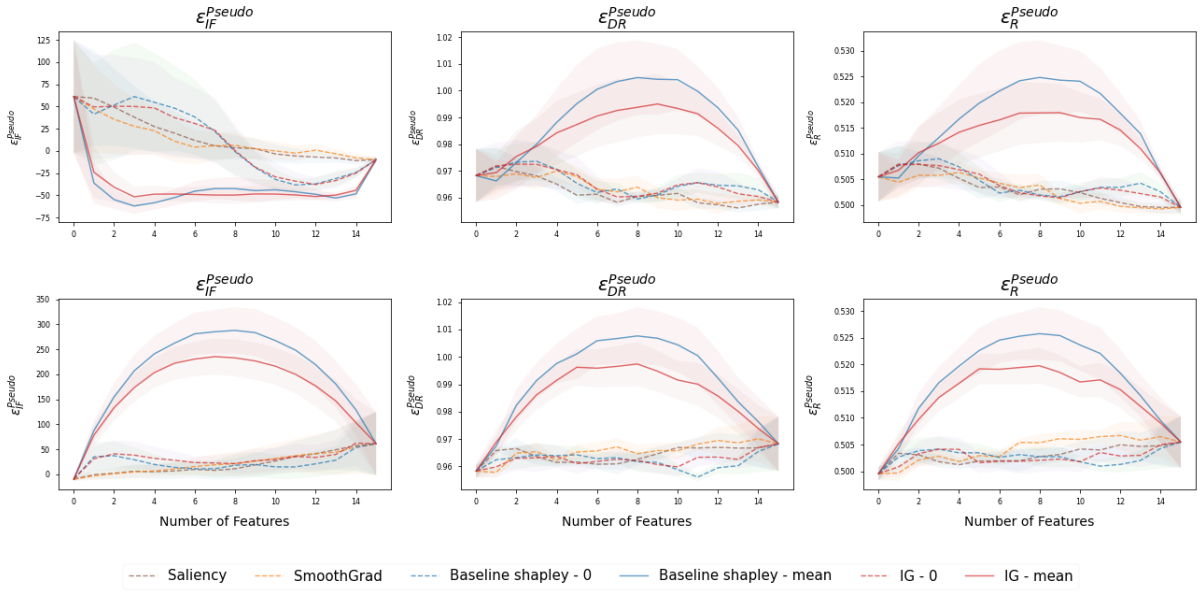


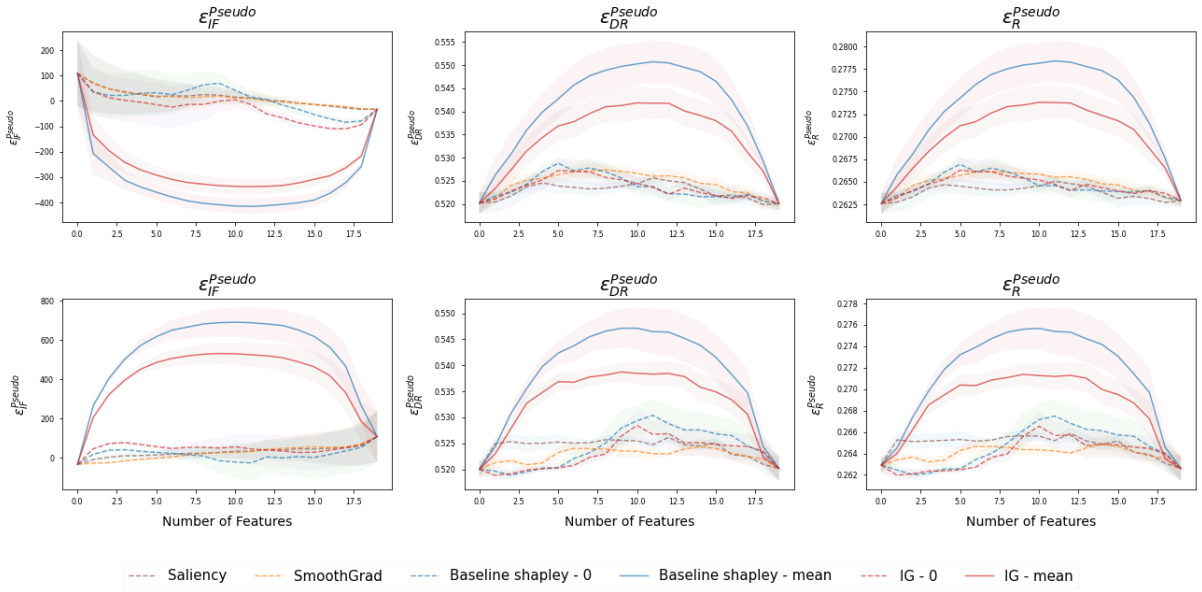
Fig. S6 | Deletion & Ablation with Pseudo-outcome for IST-3: On the left, we present the deletion study, while the right is the negative ablation study. The upper plots utilize the mean from random training data samples as the baseline for both deletion and ablation. The lower plots adopt a zero replacement value. The x-axis denotes the number of removed features, and the y-axis indicates the precision error of the heterogeneous effect (pehe).



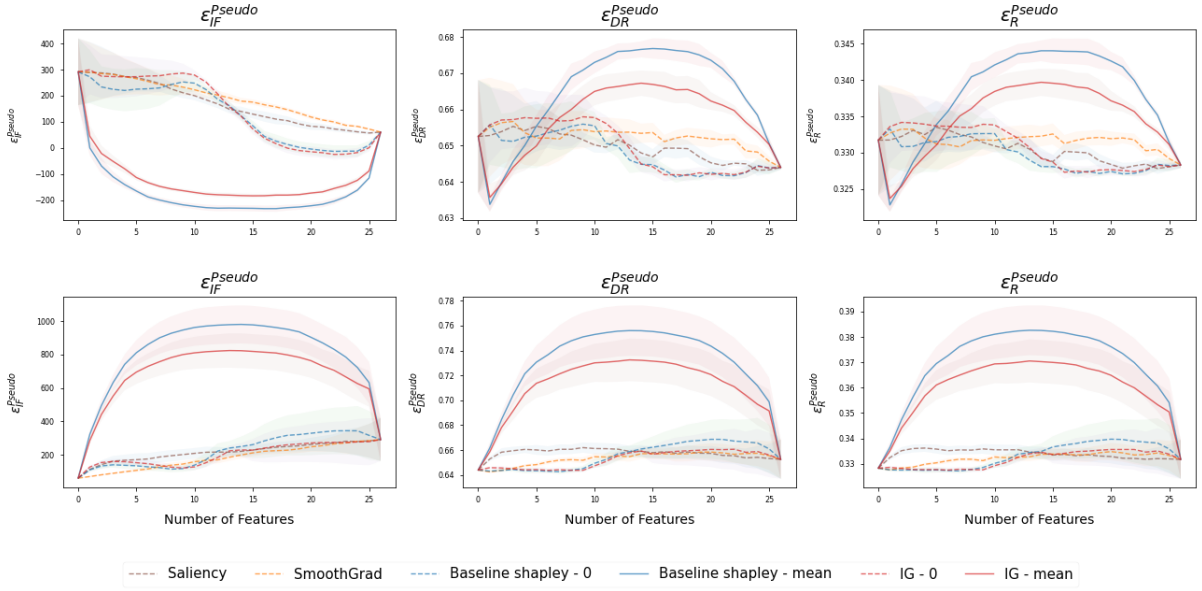
(a) Insertion (top) and deletion (bot) curves with the population mean as baseline and pseudo-outcome in CRASH-2.



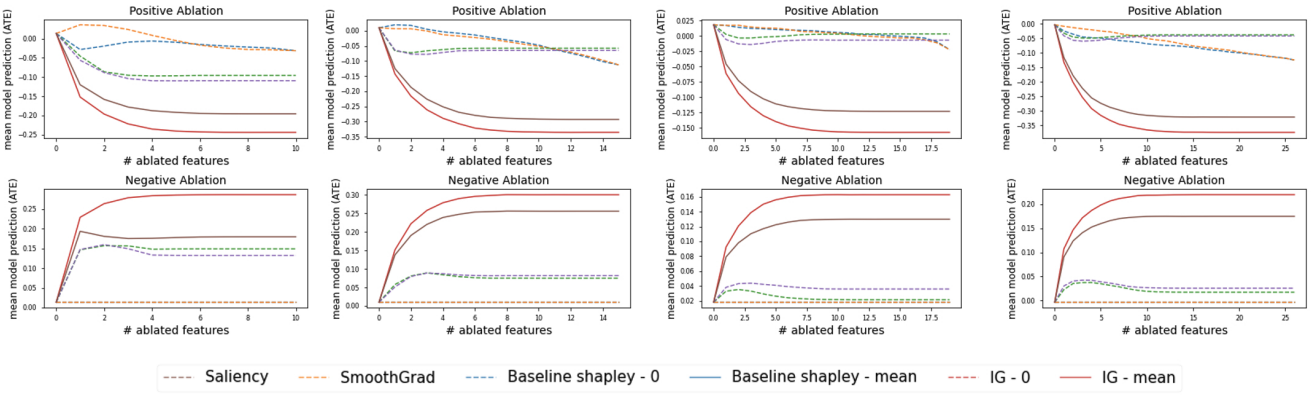
(b) Insertion (top) and deletion (bot) curves with the population mean as baseline and pseudo-outcome in IST-3.



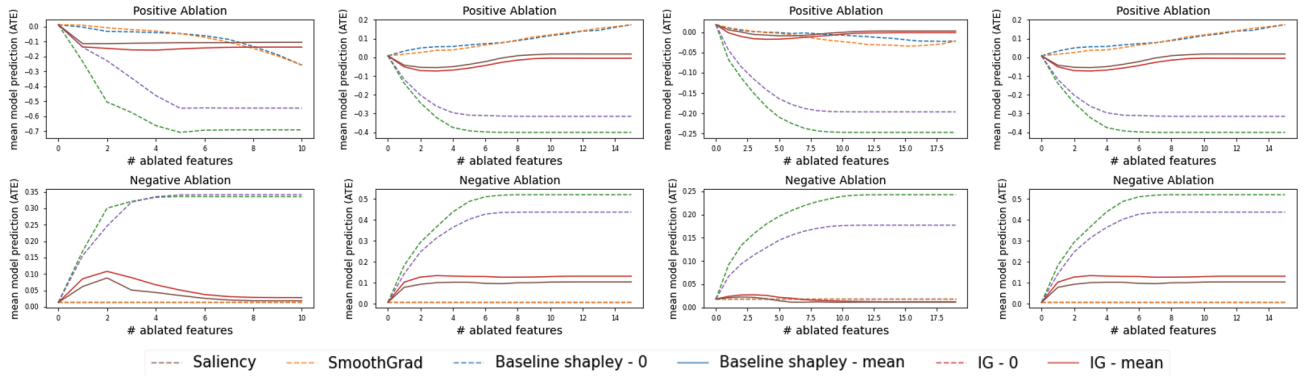
(a) Insertion (top) and deletion (bot) curves with the population mean as baseline and pseudo-outcome in SPRINT.



(b) Insertion (top) and deletion (bot) curves with the population mean as baseline and pseudo-outcome in ACCORD.



(a)



(b)

Fig. S9 | (a): Positive (top) and negative (bot) ablation curve with the population mean as baseline. (b): Positive (top) and negative (bot) ablation curve with zero as baseline.

S4.4 Knowledge Distillation as a Benchmark Test in Clinical Datasets

Explanation Methods	MSE ($\times 10^{-3}$)	Top 5 Features
	CRASH-2	
Saliency	4.0 ($\pm$0.2)	heart rate, injury time, respiratory, capillary refill time, systolic blood pressure
SmoothGrad	3.8 ($\pm$0.2)	heart rate, injury time, respiratory, capillary refill time, systolic blood pressure
Shapley - 0	3.5 ($\pm$0.2)	Injury type, gender, GCS, age, heart rate
Shapley - mean	3.5 ($\pm$0.1)	Injury type, gender, GCS, injury time, age
IG - 0	**3.4 ($\pm$0.2)**	**Injury type, gender, GCS, age, heart rate**
IG - mean	3.5 ($\pm$0.2)	Injury type, gender, injury time, age
	IST-3	
Saliency	19.1 ($\pm$0.64)	gcs, nihss, age, weight, infarct: possibly yes
SmoothGrad	18.8 ($\pm$0.60)	diastolic blood pressure, nihss, age, weight, gender
Shapley - 0	18.7 ($\pm$0.7)	Gcs, gender, age, systolic blood pressure, diastolic blood pressure
Shapley - mean	**17.0 ($\pm$0.64)**	**Nihss, infarct, anti-platelet usage, atrial fibrillation, gender**
IG - 0	18.6 ($\pm$0.66)	gcs, diastolic blood pressure, gender, age, nihss
IG - mean	17.0 ($\pm$0.66)	Nihss, infarct, anti-platelet usage, atrial fibrillation, TACI
	SPRINT	
Saliency	6.1 ($\pm$0.21)	Umaclr, EGFR, TRR, creatine, glucose
SmoothGrad	5.9 ($\pm$0.21)	creatine, EGFR, age, diastolic blood pressure
Shapley - 0	5.9 ($\pm$0.20)	DBP at baseline, SBP at baseline, EGFR, Taking Aspirin or not, age
Shapley - mean	5.1 ($\pm$0.12)	CKD history, statin usage, Age, Cardiovascular history, gender
IG - 0	5.8 ($\pm$0.21)	gcs, diastolic blood pressure, gender, age, nihss
IG - mean	**5.3 ($\pm$0.17)**	**Age, gender, statin, CKD history, Cardiovascular history**
	ACCORD	
Saliency	17.6 ($\pm$0.64)	Trig, EGFR, Potassium, ALT, HDL
SmoothGrad	17.0 ($\pm$0.64)	HDL, potassium, EGFR, alt, Trig
Shapley - 0	16.0 ($\pm$0.57)	Smoking, # of anti-bp agents, DBP, age, LDL
Shapley - mean	**14.7 ($\pm$0.56)**	**Cardiovascular history, gender, aspirin, # of anti-bp agents, race**
IG - 0	15.9 ($\pm$0.60)	# of anti-bp agents, HR, age, SBP at baseline
IG - mean	14.8 ($\pm$0.57)	Cardiovascular history, gender, aspirin, # of anti-bp agents, race

Table S6 | Knowledge distillation performance and top 5 identified features for datasets IST-3, CRASH-2, SPRINT, and ACCORD.

S4.5 IST-3

S4.5.1 Subpopulation Analysis

Here we demonstrate the results of shapley value by only considering a subgroup, male and female population specifically. For example, to understand important features only within the male population, for each feature, we used the average feature of males as the baseline for replacement during shapley calculation, Appendix S3.2.

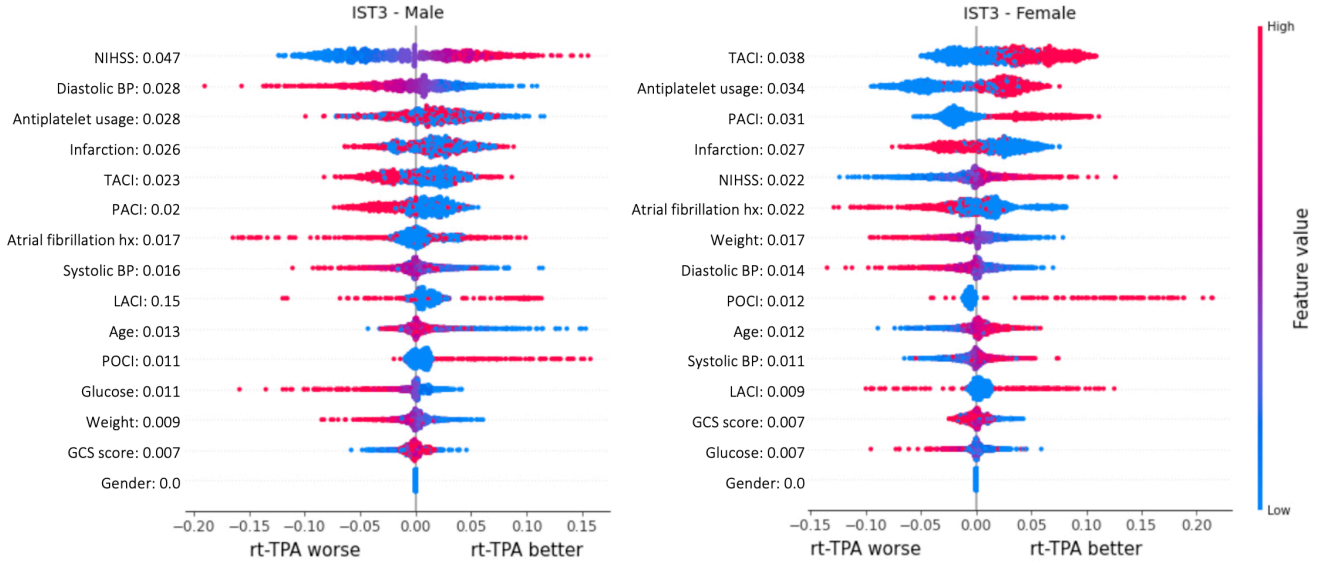
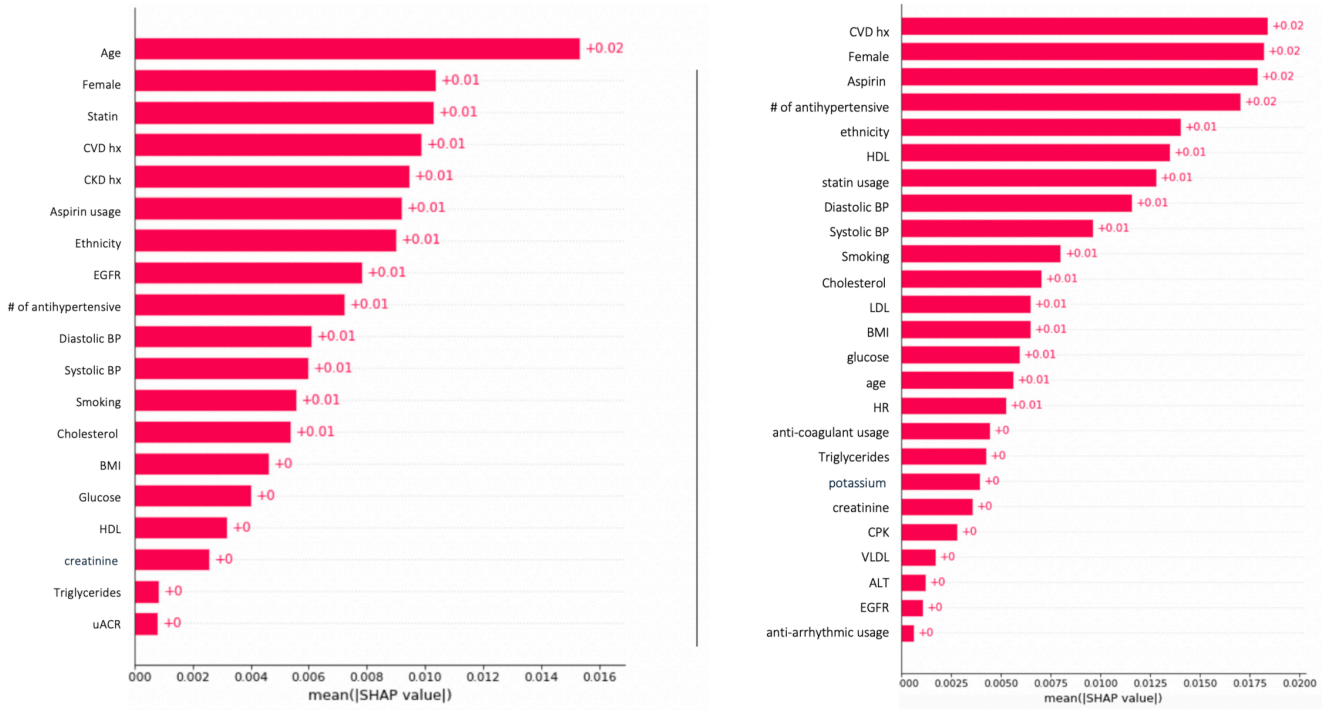


Fig. S10 | Shapley summary plots for the male (left) and female (right) subpopulations with gender-specific baselines.

S4.5.2 ACCORD & SPRINT Additional Results

Here, we present the feature rankings derived from Shapley values for the ACCORD and SPRINT cohorts. The Shapley values are calculated based on their respective population baselines.



(a)

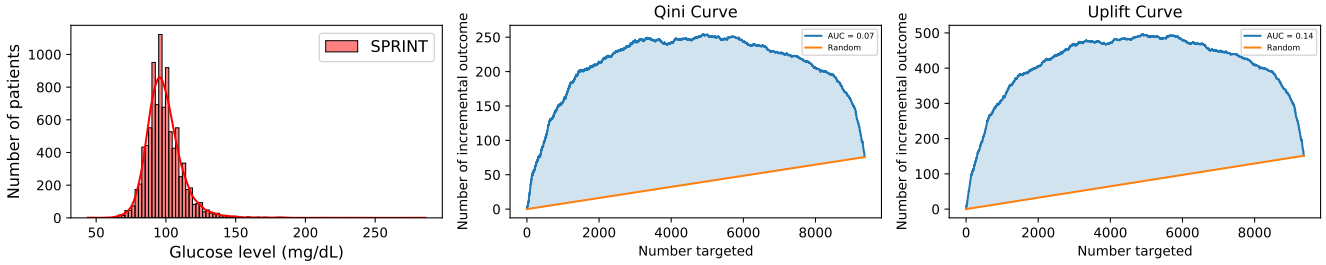
Fig. S11 | **(a)** Feature ranking between SPRINT (right) and ACCORD (left) based on absolute Shapley Values. The Y-axis represents feature names, and the X-axis indicates their average absolute Shapley values.

Dataset	**Uplift Score** ($\times 10^{-2}$)	**Qini Score** ($\times 10^{-2}$)
SPRINT (train)	7.5 ± 0.11	3.9 ± 0.06
ACCORD (test)	0.38 ± 0.17	0.22 ± 0.09
ACCORD*	0.53 ± 0.17	0.30 ± 0.09
CRASH-2 (train)	7.6 ± 0.11	4.0 ± 0.07
Harborview (test)	0.05 ± 0.04	-0.05 ± 0.03
Harborview*	0.50 ± 0.08	0.08 ± 0.05

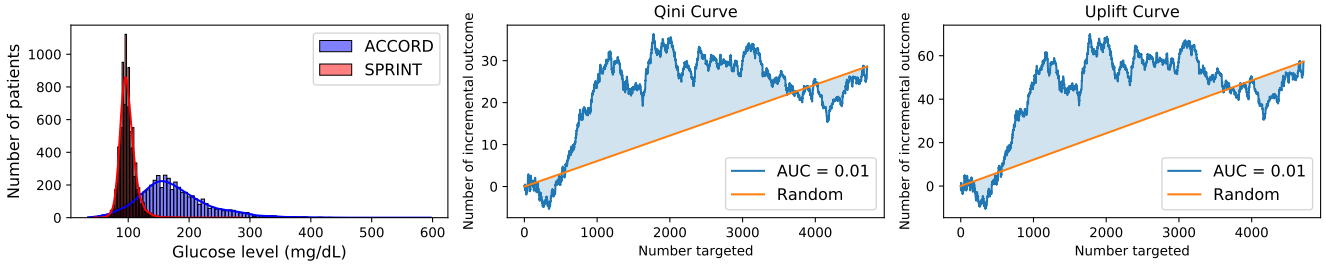
Table S7 | Uplift and Qini scores with 95% confidence intervals for various datasets. (*) denotes datasets excluding individuals with glucose levels greater than 300 mg/dL in ACCORD and patients older than 45 y/o in the Harborview trauma registry.

S4.5.3 Cross validation ACCORD with SPRINT

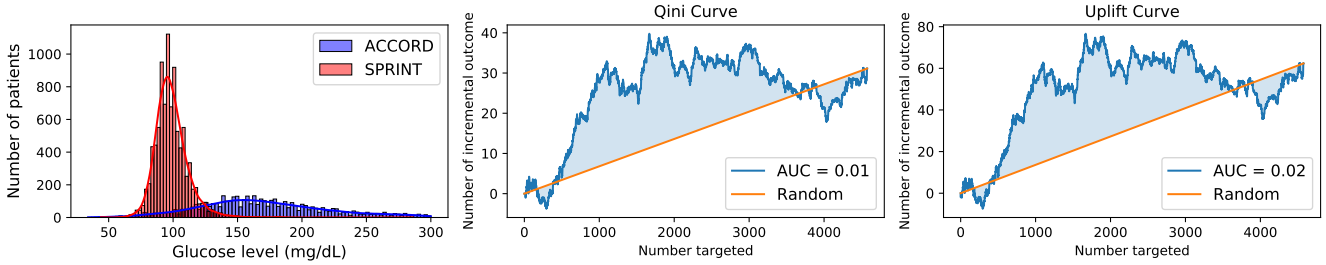
In this section, we demonstrate the CATE model's performance in SPRINT (train) and ACCORD (test) with qini and uplift scores.



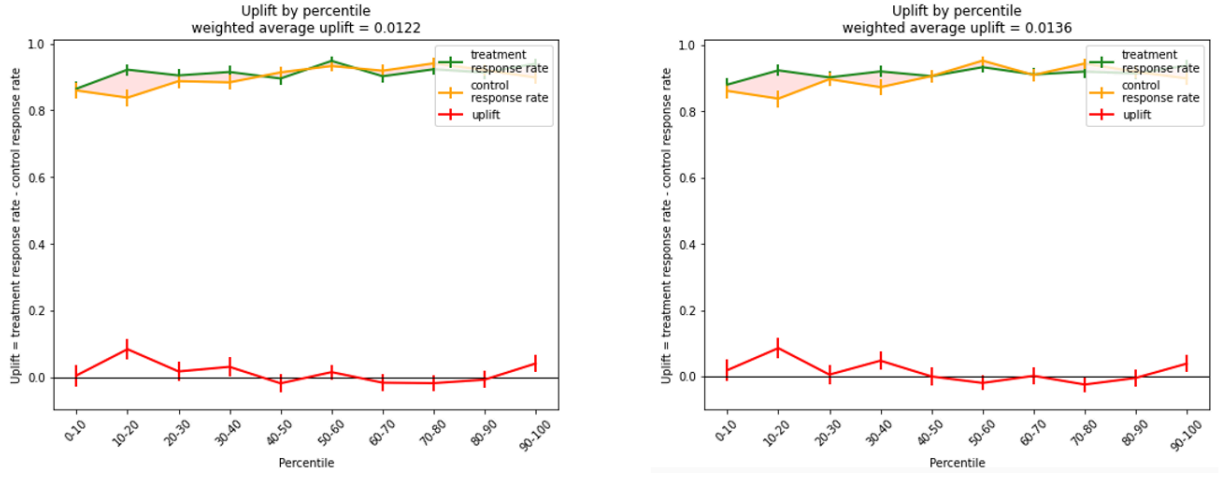
(a) SPRINT



(b) ACCORD: Before capping



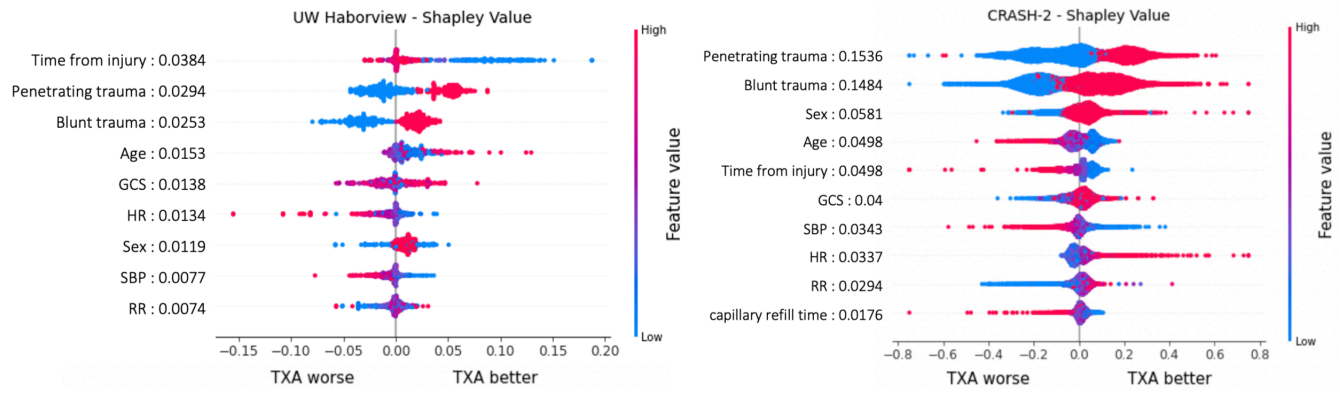
(c) ACCORD: After capping



(d)

Fig. S12 | (a) Glucose distribution, Qini curve, and uplift curve for SPRINT. (b) Distribution of glucose levels, Qini curve, and uplift curve before capping at the maximum SPRINT glucose threshold. (c) Distribution of glucose levels, Qini curve, and uplift curve after capping at the maximum SPRINT glucose threshold. (d) Weighted uplift score, treatment response rate, and control response rate at each percentile before (left) and after (right) capping.

S4.6 Pre-hospital (UW Harborview) & In-hospital (CRASH-2) TXA Additional Results



(a)

(b)

Fig. S13 | **(a)** Shapley summary plot for the Harborview cohort (left), CRASH-2 (right). **(b)** CRASH-2: Explaining sample individuals with different demographics and laboratory values. Red colors denote positive attributions and blue denotes negative attributions.