CODE - XAI: Construing and Deciphering Treatment Effects via Explainable AI using Real-world Clinical Data.

Mingyu Lu[1,3], Ian Covert[1,3], Nathan J. White[2,3,*], and Su-In Lee[1,3,*]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington

[2]Department of Emergency Medicine, University of Washington
[3]Resuscitation Engineering Science Unit, University of Washington
[*]indicates co-senior authorship

Abstract

Understanding which features drive the treatment effect has long been a complex and critical question in clinical decision-making. Significant advances have been made in estimating treatment effects, particularly with Conditional Average Treatment Effect (CATE) models. These models account for confounding variables, e.g. age, and gender, thereby capturing heterogeneity in treatment effects. However, identifying the specific features driving these effects remains largely unexplored. To bridge these gaps, we propose CODE-XAI, a framework that interprets CATE models using Explainable AI (XAI) to perform feature discovery. CODE-XAI provides feature attribution at individual and cohort levels, enhancing our understanding of treatment responses. We benchmark these XAI methods using real-world clinical data, demonstrating their effectiveness in uncovering feature contributions and enabling cross-cohort analysis, advancing precision medicine and scientific discovery.

Introduction

Quantifying why an intervention affets a given result is a quintessential issue researchers face in numerous high-stake applications [1, 2]. In medicine, healthcare professionals use available evidence to decide which treatments could improve an individual patient's health[2]. In this context, randomized controlled clinical trials (RCTs) are considered the gold standard[3] for establishing the presence of a significant treatment effect of an intervention. More precisely, RCTs establish the average treatment effect (ATE) of an intervention when applied to a well-defined cohort of patients sharing similar characteristics. Treatment randomization is further used to effectively control for potentially confounding variables, e.g. age, gender, or health status[3], thus isolating the treatment effect of the chosen intervention.

Numerous frameworks and approaches, including those based on neural networks [4–8] have been developed to address the challenge of treatment effect estimation. Among these frameworks are Conditional Average Treatment Effect (CATE) models, which enhance treatment effect estimates by conditioning on observed covariates[2, 9]. Despite making progress, these approaches are primarily evaluated on (semi)synthetic datasets that fail to capture the full complexity of real-world disease dynamics [8]. More critically, these methods are tailored for optimal treatment effect estimation, and fall short of answering two vital questions: (1) *which feature drives the treatment effect?* and (2) *why do individual responses to treatments vary?* Such factors are diverse and complex and differ across cohorts, so simply measuring the treatment effect is insufficient to identify them.

Attempts to understand why treatments differ for individuals or subgroups, and thus maximize their real world application, have traditionally been relegated to secondary subgroup analyses of RCTs that lack the power to drive changes in clinical practice [10]. Traditional subgroup analyses focus on ATE differences across patients based on predefined covariates, for example, male versus female [2, 11, 12]. While providing cohort-level feature importance for predefined covariates, subgroup analysis fails to provide insights into how treatments can affect individuals, or enable cross-cohort comparisons [10]. Subgrouping also typically relies on categorical variables or categorizing continuous variables because as data dimensionality increases or continuous variables are introduced, the number of potential subgroups grows exponentially, making comparisons unwieldy [13]. Therefore, despite advances in modeling, limited progress has been made in identifying the key features contributing to average treatment effects, thus limiting the application of knowledge gained from RCTs to a wider range of situations and individual patients. [14, 15]. This

knowledge chasm becomes particularly apparent during real world clincial decision making, where clinicians must synthesize and apply average treatment effects derived from RCT cohorts towards individual patients, even when they are different from those studied in the clinical trial.

To this end, we propose *CODE-XAI, a framework that discovers feature that drives treatment effects by interpreting CATE models using Explainable AI (XAI)* [16, 17]. In particular, *local explanation* methods[18], such as Integrated Gradient (IG) [19] and Shapley values [20, 21], can address the issue of which feature drives the treatment effect for a given individual. These methods are favorable because they decompose the treatment effect (i.e., CATE model's output) into each feature's contribution directly without grouping or feature conversion [22], Figure 1 (b). Additionally, they enable feature attribution on the individual level in a usable way, enhancing our understanding of why certain individuals may respond more favorably to treatment than others. At the cohort level, individual attributions can be aggregated to provide a global explanation, allowing us to understand the impact of features on treatment effects. Additionally, we employed an ensemble approach to enhance uncertainty quantification in treatment effects and ensure reliable attribution scores.

Furthermore, to facilitate clinical translation, we first introduced *benchmarking techniques that assess both CATE and XAI methods* and demonstrated that Shapley value [20], outperforms other local attribution methods. Using Shapley values, the features identified by CODE-XAI aligned closely with key features reported in existing studies on the cohort level. Moreover, we propose a *novel subpopulation analysis* on various baselines to uncover clinical feature interactions and resolve conflicting results across different trials. We then tested CODE-XAI against the two most common hurdles present when applying RCTs to real-world practice, differences in patient characteristics, and alternative clinical practice settings, helping to decipher conflicting results from previous studies.

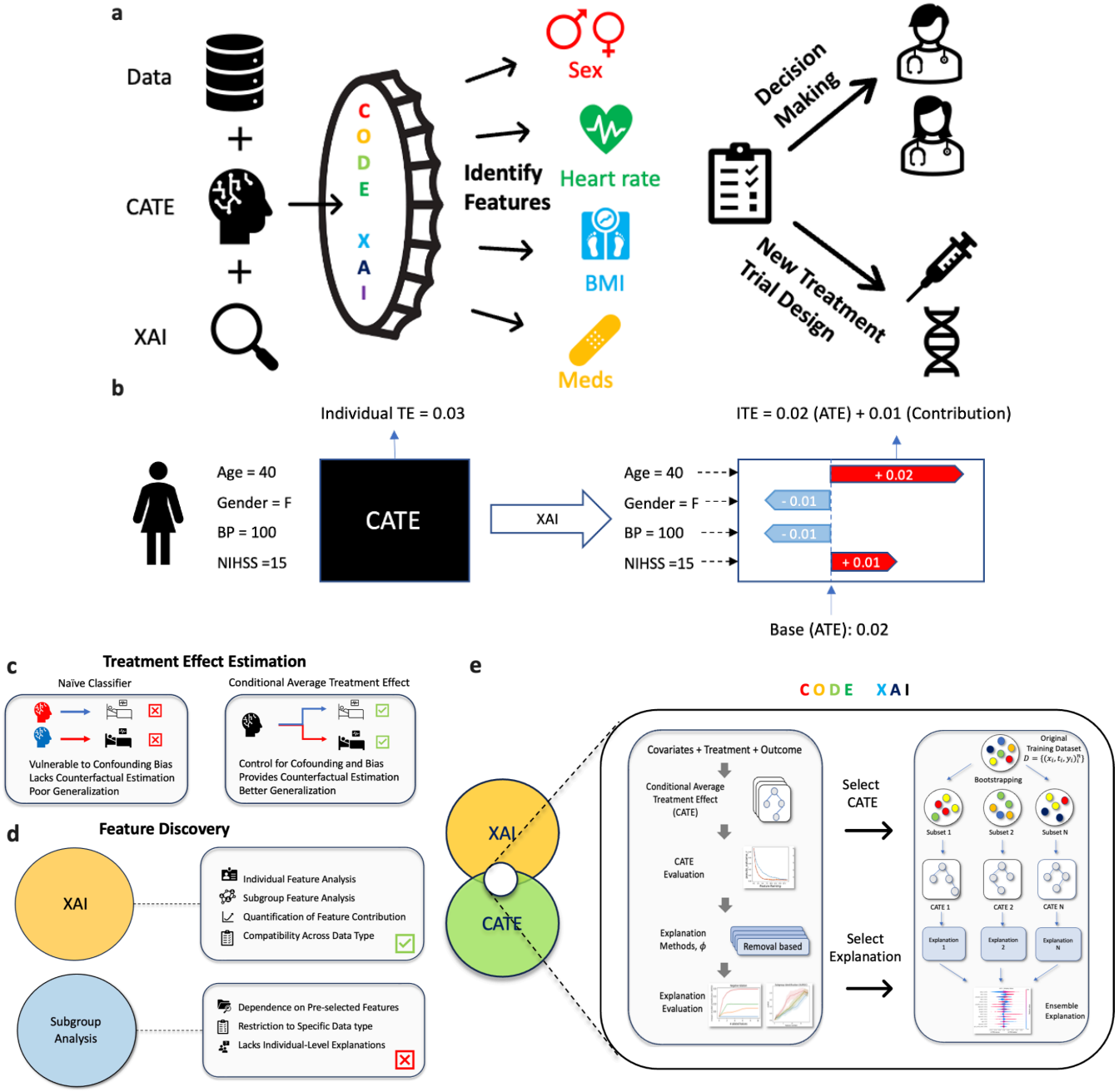

Fig. 1 | Overview of the CODE-XAI Framework. **(a)** Concept figure of the framework. **(b)** Individual explanations through XAI. **(c)** Treatment effect estimation: trade-offs between plugin estimates and conditional average treatment effect (CATE). **(d)** Feature discovery analysis: subgroup analysis vs XAI methods. **(e)** CODE-XAI overview, evaluation of CATE and explanation methods, and explanation of the selected model with ensemble Shapley.

Results

Benchmarking CATE and XAI on Real-World Clinical Data

We examine the performance of both CATE models and their corresponding explanation (feature attribution) in real-world clinical data. We first train CATE models for each cohort, including IST3[23], CRASH-2[24], ACCORD[25], and SPRINT[26], and we obtain explanations with methods described in Section 0.2. Details of cohort description, datasets, and model implementations are in Appendix S1.

Estimating Real-World Treatment Effects with Ensemble CATEs

We first trained CATE models to estimate treatment effects using data from four well-known randomized control trials: IST-3, CRASH-2, SPRINT, and ACCORD [23–26]. To ensure accurate explanations, we select the best-performing models according to their pseudo-outcome surrogate (Appendix S4.2), finding that X-learner, a two-stage regression estimator [27], outperforms other models in IST-3, CRASH-2, and SPRINT, while Doubly Robust Learner (DR-Learner), a two-stage learner utilizing doubly robust estimation[7], performs best in ACCORD (Table S4).

In well-controlled RCTs, randomization minimizes confounding, so when there are no significant effect modifiers, the average of CATE estimates should align with the average treatment effect (ATE) [2]. We therefore compared the ensemble ATE estimates to the reported ATE in each trial. For IST-3 and CRASH-2, the CATE models produced estimates closely aligned with the reported outcomes [23, 24] (see Table 1). However, in the blood pressure control trials, i.e., SPRINT and ACCORD, the CATE models provided higher ATE estimates than those reported: 1.6% for SPRINT and 1.2% for ACCORD, compared to 0.54% and 0.22%, respectively. Moreover, the analysis reveals that SPRINT, which demonstrated the efficacy of intensive blood pressure control, and the ACCORD study, which showed no significant treatment effects, align with the reported findings. These findings show that ensemble CATE models capture treatment effects at the cohort level, particularly in trials with substantial treatment effects.

Cohort	Predicted ATE (95% CI)	Reported ATE (%)
CRASH-2	1.1 (0.2 - 1.9)	1.5
IST-3	2.0 (0.3 - 4.0)	2.0
SPRINT	1.6 (0.8 - 2.4)	0.54
ACCORD	1.2 (-0.3 - 2.4)	0.22

Table 1 | Comparison of predicted Average Treatment Effect (ATE) estimates from CATE models (with 95% confidence intervals) and reported ATE values (primary outcome differences) across four clinical trials.

Enhanced Consistency in Feature Attributions through Ensemble Models

We next demonstrate the importance of interpreting ensemble models over single models. We compare cosine similarities of feature attributions, i.e. Shapley value. for models trained with different random initializations. With different random initialization, explanations from single models exhibit low similarity and high variance, with scores of 0.13, 0.15, 0.15, and 0.21. In contrast, ensemble models provide more consistent and robust explanations, as shown in Figure 2(b-top). The average similarity of Shapley values within the ensemble increases from 0.6 with 10 models to 0.8 with 20 models, highlighting the enhanced reliability and consistency of feature attributions achieved through the ensemble approach (Figure 2(b-middle)).

Benchmarking XAI Methods on Real-World Clinical Data

We then evaluated the performance of ensemble explanations using various local explanation methods. While ablation studies, which systematically add or remove features [28], are common, they are computationally expensive for ensembles. Instead, we evaluated the explanations using our proposed distillation benchmark test with global features (Section 0.3). As shown in Figure 2(c), both Shapley-mean and IG-mean consistently demonstrate lower distillation loss across the SPRINT, ACCORD, and IST-3 datasets under various feature budgets. In CRASH-2 dataset, the performance of all methods is comparable except for Saliency. Our findings also suggest that the same explanation method, when using a population mean as the baseline (e.g., Shapley-mean), provides more reliable results compared to using constant baseline values (e.g., Shapley-0), as shown in Figure 2(c).

Additionally, We also present the best-performing methods and their identified top five features across different datasets in Table S6. In the CRASH-2 dataset, IG-mean identifies injury type, gender, age, and GCS score as the most

important factors influencing treatment effects, while Saliency highlights heart rate, respiratory rate, and capillary refill time as the key features.

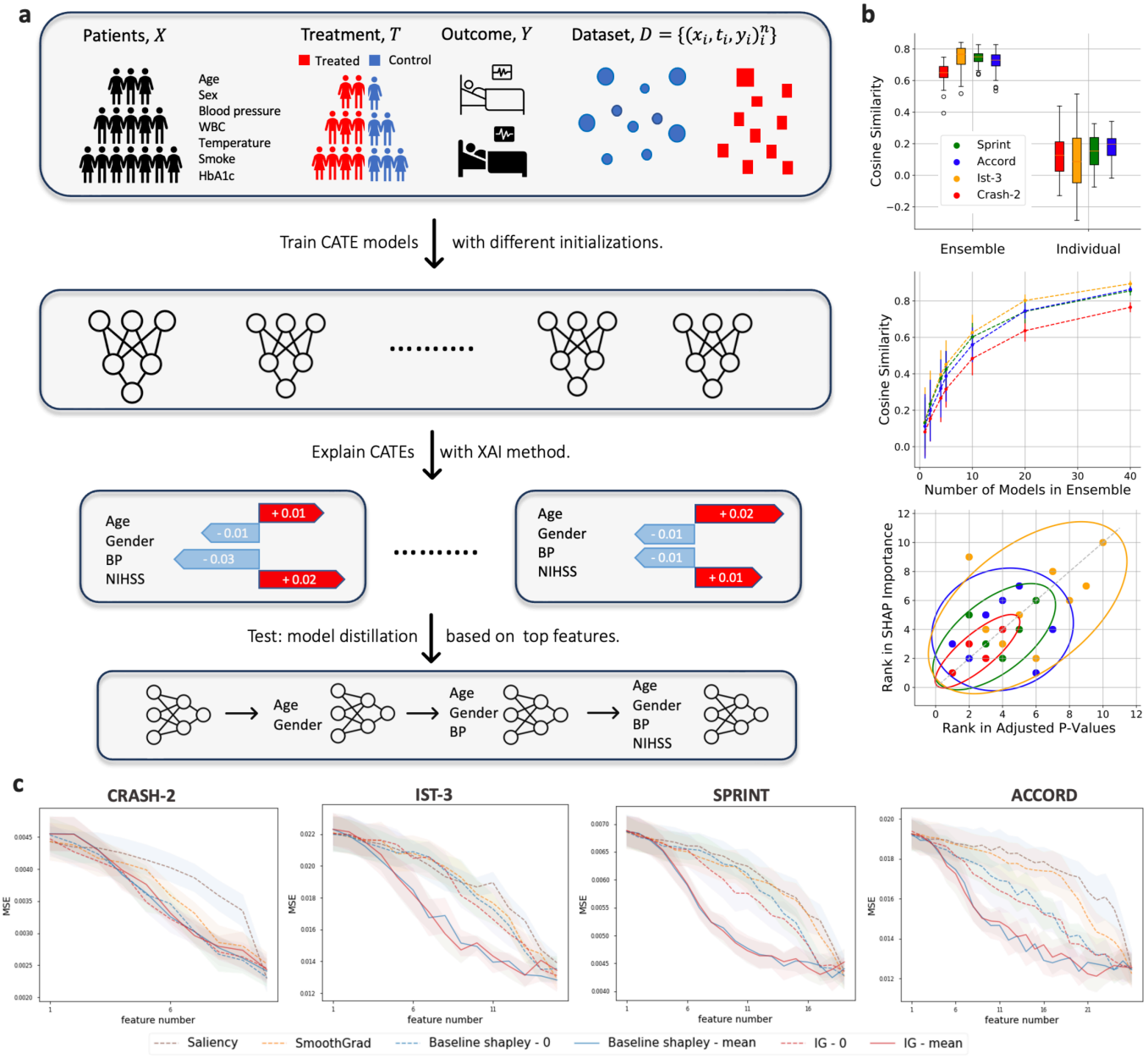

Fig. 2 | Results of Examining Ensemble Explanation.(a) *Evaluation and Explanation generation procedure* of CODE-XAI. Ensemble CATE models are trained with patients' data and different initializations. Features obtained through CATEs and XAI methods are used for follow-up evaluation. (b): (top) Comparison of cosine similarity between explanations from ensembles (40 models in an ensemble) and individual models.(middle) Model in an ensemble and its cosine similarity between explanations. (bottom) Comparison of interaction p-value rank and Shapley value rank with 95% confidence ellipses. (c) Knowledge distillation performance across datasets. The x-axis denotes the feature count of student models, and the y-axis represents their performance metrics: Mean Squared Loss (MSE).

Insights by Explaining Ensemble CATEs with Shapley Value

Here, we demonstrate how to leverage the best-performing feature attribution method, Shapley values, to analyze clinical trials, highlighting its advantages over traditional subgroup analysis. For the remainder of this section, we refer to aggregated Shapley values from ensemble CATE models simply as Shapley values.

Global Feature Identification: Shapley Values versus RCT Findings

To evaluate the effectiveness of Shapley values in cohort-level feature discovery, we compared feature rankings based on Shapley values[1] with those reported in the original studies using Spearman's rank correlation [29]. For these studies, reported interaction p-values[2] were used as proxies for feature ranking [30]. Our findings show a significant correlation between Shapley rankings and reported features, with values of 0.8, 0.54, and 0.6 for the CRASH-2, IST-3, and SPRINT, respectively (Table 2). In contrast, the ACCORD study shows a low correlation (0.05), which is expected, as no significant features were reported [25]. We also conduct additional experiments to demonstrate that Shapley value outperforms other local attribution methods in identifying key features in semi-synthetic environments (Appendix S4.1).

Dataset	Correlation (Corr)	p-value	Number of Reported Features
CRASH-2	0.80	0.11	4
IST-3	0.54	0.09	10
SPRINT	0.60	0.12	6
ACCORD	0.05	0.90	7

Table 2 | Correlation between ranks based on Shapley value and interaction p-values across RCTs.

IST3: Analyzing Features' Contribution to rt-TPA Treatment Effect through Shapley Value

Here, we analyze clinical features in IST-3, a clinical trial that assesses the efficacy of intravenous rt-PA in acute ischaemic stroke patients. Unlike traditional subgroup analysis, which requires dividing patients into subgroups and calculating risk or odds ratios, Shapley values allow direct analysis of feature impact at both individual and cohort levels. Shapley values provide individual-level explanations [16, 20] by breaking down the total treatment effect into contributions from each feature for every patient (Figure 3(a)).

In Figure 3(b), the upper force plot shows an example patient who experienced a treatment effect of 11%, significantly above the average treatment effect (ATE) of 1.6%. The red bars represent features that contribute positively to the treatment effect, including a high NIHSS score, TACI, and usage of anti-platelet within 48 hours; the blue bars indcate features that reduce the treatment effect, including atrial fibrillation history and higher systolic blood pressure. Conversely, the lower force plot shows a male patient with low NIHSS scores and PACI syndrome, whose treatment effect decreased by 11%.

Figure 3(d) illustrates individual feature attributions and each feature's global ranking across cohorts, showing that the NIH Stroke Scale (NIHSS), a neurological exam for stroke evaluation, is the most influential factor affecting rt-PA efficacy. Without categorizations or creating numerous subgroups, we can easily examine the impact of continuous features. For example, Shapley value indicates that patients with higher NIHSS, depicted by the red cluster, contribute to treatment effects positively when administered TPA, in contrast to those with lower NIHSS scores, marked by the blue cluster. This observation is consistent with prior research [23, 31], which also identified a significant interaction between NIHSS scores and tPA treatment effectiveness.

Additionally, the second most impactful feature is the type or syndrome of the stroke. As shown in Figure 3(d), rt-TPA exhibits enhanced benefits for patients diagnosed with TACI and PACI, a finding consistent with the original IST-3 study and reported in several stroke-related studies [32]. Our findings also reveal that factors such as receiving an anti-platelet drug within 48 hours and infarction history significantly affect the effect of rt-TPA, which previous studies have also discovered [32, 33].

IST-3: Subgroup Analysis with Shapley Value

We now extend the analysis to multiple features and identify subgroups that are more susceptible to rt-TPA treatment. For instance, in Figure 3(c), we analyze gender and NIHSS and their combined influence on treatment effect. We

[1] For each feature, we aggregated the absolute values of local attributions across all individuals in the cohort and took the average to obtain the global (cohort) explanation.

[2] A lower interaction p-value indicates a higher likelihood of a feature being a treatment effect modifier.

observe that with the same NIHSS scores, males and females exhibit different treatment efficacy. In male patients (red dots) with lower NIHSS scores (< 15), rt-TPA appears less effective, whereas its effectiveness increases in males with higher NIHSS scores (> 15).

To obtain deeper insights into the contributions of specific features within a particular subgroup, we modify the baseline used in Shapley value calculations (Section 0.3). We thereby compare male individuals or female individuals to male or female baselines by adjusting our research question to: *Which features are important for males or for females compared to other males or females?* In this case, the significance of gender is no longer present.

Within the male population, while the NIHSS score remains the most critical feature, the order of importance of other features shifts; see Figure S10(b)). Conversely, when analyzing female patients against a female baseline, the significance of NIHSS diminishes, and TACI emerges as the most influential feature, followed by anti-platelet usage, Figure S10(b). Interestingly, although most feature trends remain consistent when using the population baseline, the effects of pre-stroke anti-platelet therapy differ between genders. Its usage seems to counteract the benefits of rt-TPA in males while enhancing its effects in female patients. This finding is consistent with several studies that emphasize the positive impact of anti-platelet therapy on women, as reported by [34].

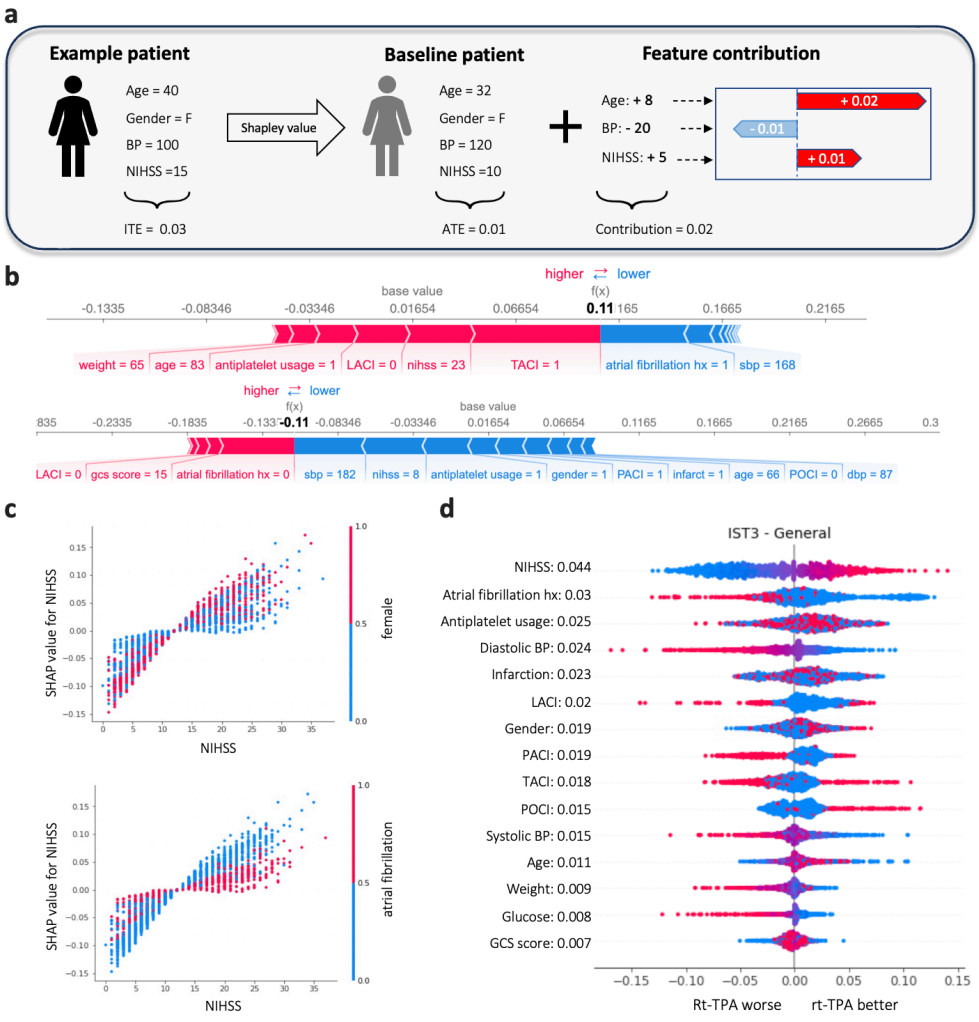

Fig. 3 | Analyzing the IST-3 Study with Shapley Values: (a) Decomposing feature contributions for an example patient with Shapley value. (b) Shapley values for example individuals, where red indicates positive attributions and blue represents negative attributions. (c) Combined Shapley values (left y-axis) and feature values pairs (x-axis and right y-axis) of NIHSS with gender (top) and atrial fibrillation (bottom). For binary features, the red dot indicates a feature value of 1, while blue indicates 0. (d) IST-3 summary plot showing features on the y-axis sorted by mean absolute Shapley values and on the x-axis by their corresponding Shapley values. Colors indicate feature values, with red for higher and blue for lower.

Deciphering Treatment Effects When Patients are Different

A common reason why RCTs cannot be applied to more general populations is due to variation in patient characteristics that influence treatment effects. To address this issue, We stress-tested CODE-XAI's ability to identify key differences in patient characteristics driving alternative treatment outcomes in the setting of intensive blood pressure management using two notable RCTs. The SPRINT trial showed that intensive blood pressure management reduced cardiovascular events and mortality in high-risk, non-diabetic patients, whereas the ACCORD trial found no significant benefit when the same treatment was applied to patients with type 2 diabetes[25, 26].

Discrepancies in Predictive Features

We first compared the top features affecting treatment outcomes in both trials. Interestingly, despite overall similarities between the cohorts, the top features affecting the treatment effect for each trial were quite different. In the SPRINT trial, *age* was the most significant factor influencing blood pressure control, followed by gender, statin usage, chronic kidney disease history (CKD), and cardiovascular (CVD) history; see Figure 4(a-bot). Conversely, in ACCORD, the most significant feature affecting the treatment effect was a *history of CVD*, followed by gender, aspirin use, number of antihypertensive medications, and an individual's ethnicity.

Additionally, when examining the identified features' clusters, the SPRINT trial showed a clear effect of feature pairs, e.g., age and CVD history or age and gender Figure 4(c-bottom, d-bottom). However, such effects were absent in the ACCORD trials. In some cases, the combined effect of features seems to be reversed, e.g., in glucose level and aspirin usage; see Figure 4(b-top).

Analyzing ACCORD with a SPRINT Baseline

Using CODE-XAI, we directly addressed the question of *Which features are important for ACCORD individuals compared to the SPRINT population?* We achieved this by simply substituting the baseline with an example individual from the SPRINT cohort (Appendix S3.2).

Upon reassessing the top features from both cohorts and reanalyzing the feature rankings, we observed that *fasting glucose (fpg) emerged as a prominent feature in ACCORD, but it ranked 14th among the 18 clinical features in SPRINT*; see Figure S11 (a). By identifying fasting glucose as a key treatment effect, CODE-XAI correctly and independently identified the underlying key patient characteristic, i.e. the presence of diabetes, most likely driving the difference in treatment effect between the two trials. Moreover, CODE-AXI independently provided a clear and usable treatment metric (fasting glucose) for clinicians seeking to manage blood pressure in diabetic patients.

To further investigate the impact of glucose on the effectiveness of blood pressure control in the ACCORD study, we analyzed the treatment uplift using qini scores and uplift scores (Appendix S2.2.1) among patients with varying glucose levels. As we show in Figure 4 (f-left) and Table S7, the uplift score and qini score for the original ACCORD was 3.8×10^{-3} and 2.2×10^{-3}, respectively, significantly lower than the SPRINT studies, i.e., 7.5×10^{-2} and 3.9×10^{-2}, respectively. However, when excluding patients with glucose levels exceeding 300 mg/dL (the maximum observed value in the SPRINT cohort), the average treatment effect of ACCORD increased by 39.5% for the uplift score and 36.3% for the qini score.

Using CODE-XAI, we thus unravel these conflicting results in trials. Our analysis highlights variances in glucose levels as a potential explanatory factor for the observed disparities in treatment outcomes between the two studies.

Applying CODE-XAI across Clinical Practice Settings.

Here, we test the ability of CODE XAI to identify important features in treatment effects when a proven treatment is applied to a different clinical setting. For this test, we used the treatment of traumatic bleeding after injury using tranexamic acid (TXA), a drug that is used to stabilize blood clots to reduce bleeding after injury. Strong randomized data favor the use of TXA for trauma victims at risk of significant bleeding if given at hospital admission and within 3 hours of injury[24]. Time from injury has emerged as having an important effect on TXA efficacy. So clinical practice has steadily crept towards using this drug at the scene of injury or during transport (pre-hospital), despite the lack of randomized evidence for its efficacy in this alternative practice setting. In this scenario, we asked CODE-XAI to identify *which features were most important for trauma patients when TXA was given in the hospital setting vs. when TXA was given pre-hospital*. Using data made available from CRASH-2 study investigators [24] and our local trauma center registry, we asked CODE-XAI to identify features that determine TXA efficacy when administered in these different clinical practice settings (Appendix S1.3). We then validated the feature selected by CODE-XAI in the new pre-hospital setting by computing the treatment effect gain. We also compared it to features identified during a more recent randomized controlled trial of TXA when given specifically in the pre-hospital setting[35].

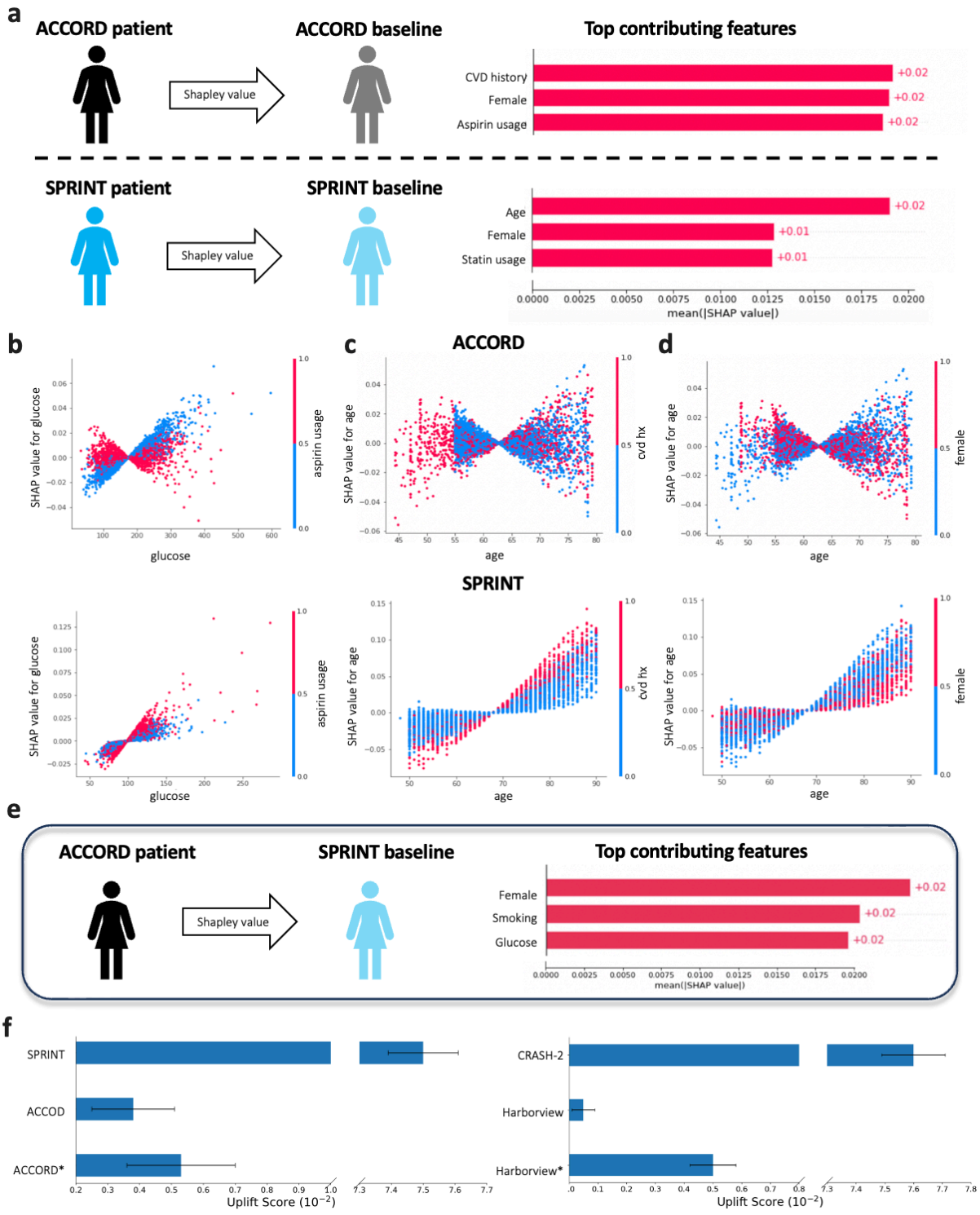

Fig. 4 | (a) Top 3 features based on mean absolute Shapley values in the ACCORD and SPRINT studies. The upper plots show results from the model trained with overlapped subsets, while the lower plots are from separately trained models.; Shapley scatter plots for feature pairs in separately trained models: (b) glucose and aspirin, (c) age and history of CVD, and (d) age and gender, with SPRINT results on top and ACCORD results on the bottom. (e) Analysing Accord with SPRINT baseline and its top contributing features. (f) Uplift score for SPRINT and ACCORD (left); CRASH-2 and Harborview trauma registry(right). (*) 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.

We first compared the top features based on their Shapley values. As shown in Figure S13(a-left), in the pre-hospital settings, the top features were time-to-injury, GCS score, trauma type, and a new effect, *age*. We then examined the treatment effects among different age groups. As shown in Figure 4(f-right) and Table S7, the uplift score and qini score for our pre-hospital cohort are 5×10^{-4} and -5×10^{-4}, respectively. Surprisingly, after excluding patients older than 45 y/o in the pre-hospital settings, the scores increase to 5×10^{-3} and 8×10^{-4}, respectively. This finding indicates that, in the pre-hospital setting, CODE-XAI is identifying age as a new and potentially crucial correlate of TXA efficacy. This result was validated by similar emergence of age as a new treatment effect for TXA efficacy from the PATCH study, a randomized controlled trial of TXA administered to injured patients in the pre-hospital clinical setting [35]. This result highlights the ability of CODE-AXI to identify important treatment effects when randomized clinical trial data are applied towards different clinical practice settings.

Discussion

Estimating treatment effects in medicine is a critical area of research [8]. However, understanding the impact of individual features on treatment effects remains largely explored, particularly in terms of their application and robustness.[14, 28, 36, 37]. We demonstrate that providing a deeper understanding of CATE dynamics with XAI can extend the capability of RCT's to unveil real world clinical insights and support physicians to make better-informed decisions. In doing so, we present a framework, CODE-XAI, that rigorously explains these models, overcoming the hurdles involved in applying randomized controlled trial data toward real-world use in a robust and explainable way.

We first demonstrate that ensemble CATE models can reliably estimate treatment effects using real-world clinical data, comparing them with factual outcomes and benchmarking pseudo-outcomes for model selection [8]. Next, we show that ensemble explanations are more robust than explanations derived from the best single model. To benchmark feature attribution, we propose using knowledge distillation via global explanation to evaluate these methods. This differentiates our method from explanation evaluations that are inefficient for ensemble models [28], or those that rely on unrealistic assumptions about oracle accessibility in real-world scenarios [14]

A natural use case of CODE-XAI is to analyze driving features for treatment effects across various trials in healthcare. We demonstrate how to use the ensemble Shapley value to analyze well-known RCTs [23–26]. Compared to traditional analysis, our approach provides not only subgroup analysis but *individual* analysis without the need to analyze millions of strata [12]. By analyzing local attributions, we observe how individual features can have varying effects on treatment outcomes (Figure 3). Such explanations of patient response differences can be particularly useful for clinical practitioners making individual treatment decisions. Similarly, with features at hand, we identify subgroups that would respond better to certain treatments in real-world settings (Figure 4), which can help researchers identify scientific insights that require further investigation.

CODE-XAI can also untangle conflicting results between trials and identify crucial covariates on the cohort level. We analyze two well-known trials, ACCORD [25] and SPRINT [26], which both evaluated blood pressure control but showed conflicting results, presumably due to differences in trial subject characteristics. Notably, we observe that glucose plays a significant role in the treatment effect, thus independently identifying the key difference between subjects enrolled in the two trials, i.e., the presence of diabetes. In addition, fasting glucose was identified as an important and clinically relevant treatment effect for clinicians to consider when expanding intensive blood pressure control to real world populations. We also investigated how CODE-XAI could inform important treatment effects when translating RCT knowledge across differing clinical practice settings. When examining TXA efficacy across in- and out-of-hospital practice settings, CODE-XAI identified age as a vital treatment effect explaining differences in efficacy. These results suggest that CODE-XAI can help clinicians identify key variations between study cohorts that explain outcome differences despite seemingly overlapping demographics, treatments, and outcomes.

However, the effectiveness of explanations is limited by the performance of the CATE models. While these models are effective in controlled environments such as RCTs, their reliability diminishes when faced with unobserved confounders in observational studies. Such confounders can lead to violations of plausibility assumptions, undermining the efficacy of CATE models [8, 9] and resulting in biased explanations [14]. Therefore, a promising research direction involves developing methods to impute robust attribution scores to mitigate the impact of these confounders. Additionally, integrating causal knowledge with domain expertise has been demonstrated to enhance the accuracy of feature attributions [38].

To conclude, we present a new approach to performing clinical feature discovery by explaining CATE models with XAI. We propose evaluation methods to assess CATE models with XAI in real-world clinical trial. Our framework, CODE-XAI, demonstrates several advantages compared to traditional subgroup analysis, including individual explanation, subpopulation analysis, and cross-cohort examinations. In an era where precision medicine and individualized treatments are taking center stage, understanding the nuances of treatment effects is more crucial than ever.

Methods

This section describes (1) CATE models, (2) XAI methods and ensemble explanation, and (3) evaluation of ensemble explanation. We include detailed descriptions of these topics in Appendix S1 (dataset), S2 (potential outcome framework), and S3 (explanation methods).

0.1 CATE Models

0.1.1 Model Design, Evaluation, and Cross Examination

Under the potential outcome framework [9] (S2), *meta-learners* [14] represent a class of *nonparametric CATE estimation methods*. These methods approach treatment effect estimation for binary treatments as an imputation problem for missing counterfactual outcomes. They simplify the task by decomposing it into multiple sub-regression problems, often termed *pseudo-outcomes* [2], which can be solved using any standard supervised machine learning (ML) methods.

CATE estimation methods include T-Learner [2], X-learner[27], DR-learner[7], and R-learner[39]. These methods estimate CATE by learning nuisance functions η to identify the optimal τ^*

$$\tau^* = \arg\min_{\hat{\tau}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\hat{\tau} - \hat{Y}_{\hat{\eta}}^{\text{pseudo}})^2], \tag{1}$$

where $\hat{Y}_{\hat{\eta}}^{\text{pseudo}}$ is pseudo-outcome loss depending on the learner and $\mathcal{D}$ is the training distribution.

This work uses a diverse range of CATE models, including meta-learners such as S-learner, T-learner, X-learner, DR-learner, and R-learner as well as representation learners like Dragonnet[40], TARNets[41], CFR[5], and DR-CFR. See Appendix S2.1.1 for further details regarding the structures, training procedures, and implementation of these models.

To evaluate CATE models, we employ *pseudo-outcome surrogate criteria* (S2.2) with a 5-fold validation technique. Additionally, to assess model performance across different cohorts, we utilize the Qini curve and Uplift curve (S2.2.1), which base model evaluation on observed treatment outcomes.

0.2 Explaning CATEs with Feature Attribution Methods

Once the best-performing models were identified, we used explainability (XAI) methods[16] to obtain feature contributions, i.e., *explanations*, for CATE treatment effects. XAI methods decompose model output into each feature's contribution on the individual level with respect to a baseline; they effectively address the specific question: *What is the contribution of each feature for an individual compared to the average person within a specific cohort?* Specifically, we choose methods for CATE models that meet specific criteria (S3.1), including Integrated Gradients[19] and Shapley values[20].

Integrated Gradients (IG). IG assigns importance to input features by approximating the integral of a model's gradients from a baseline input to the actual input [19]. For a given trained CATE model τ, the IG attribution for an explicand x, a variable x_i, and a baseline x' is:

$$\text{IG}_i(x, x', \tau) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial \tau(x' + \alpha(x - x'))}{x_i} d\alpha \tag{2}$$

Typically, the zero vector serves as the baseline, denoted as $x' = 0$. This means feature contributions are measured relative to their absence.

Shapley Value. The Shapley value, a concept derived from cooperative game theory, offers a unique approach to feature attributions[20]. For any prediction model, it assigns each feature an importance value by averaging all possible combinations of feature presence or absence. Mathematically, for a CATE model τ, the exact Shapley value for a feature x_i is defined as:

$$\Phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [\tau(x_{S\cup\{i\}}) - \tau(x_S)], \tag{3}$$

where N is the set of all features and S is any subset of N that does not include feature x_i.

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 [21], Baseline Shapley[42] or KernelSHAP[20] is often used. This work experiments with various methods, including Vanilla Gradient

(Saliency), Integrated Gradient (IG) with 0 as the baseline (IG-0), Integrated Gradient (IG) with population mean as the baseline (IG-mean), Baseline Shapley with 0 as the baseline (Shapley-0), and Baseline Shapley with the population mean as the baseline (Shapley-mean). Additional details about these methods are in Appendix S3.

0.2.1 Ensemble-based CATE Estimation and Explanation

Despite the progress in CATE models based on neural networks, their stability in real-world datasets remains an issue due to the inherent randomness encountered during model initialization and training [43]. To address this, we employ an *ensemble* approach [44] within CODE-XAI. We train individual CATE models $\tau_i(x)$ with different random seeds i. The ensemble CATE estimator, τ_e, and its ensemble explanation, ϕ_j, for a feature, j, and an explicand, x, can be computed as:

$$\tau_{\mathrm{e}}(x) = \frac{1}{N}\sum_{i=1}^{N}\tau_i(x) \quad \text{s.t.} \quad \phi_j(\tau_{\mathrm{e}}, x) = \frac{1}{N}\sum_{i=1}^{N}\phi_j(\tau_i, x), \tag{4}$$

where N is the number of models in an ensemble. This method enhances both the model's and explanation's stability by averaging out variability.

0.3 Examining Explainability Methods on CATEs

In this section, we introduce methods that assess the explanations of CATE.

Explanation Robustness Assessment

To evaluate the effect of the number of single models in an ensemble on explanation stability, we first train L ensembles, each with k single models, and then calculate the pairwise cosine similarity of their explanations. Given feature attributions $\phi(\cdot)$ for the l^{th} ensemble, $\tau^k_{e,l}$, composed of k single models, the average cosine similarity $\cos(\theta_k)$ is:

$$\cos(\theta_k) = \frac{1}{L(L-1)}\sum_{l=1}^{L}\sum_{j\neq l}^{L}\frac{\phi(\tau^k_{e,l})\cdot\phi(\tau^k_{e,j})}{\|\phi(\tau^k_{e,l})\|_2\|\phi(\tau^k_{e,j})\|_2}. \tag{5}$$

Examining Ensemble Explanation via Knowledge Distillation

Though ablation studies offer a convenient way to inspect explanation methods, their choice of the baseline can potentially favor particular explanation methods[17, 45]. To address this, we introduce an evaluation approach rooted in *knowledge distillation*[46], wherein the student model is coached to emulate the behavior of the teacher model. However, retraining models using local explanation rankings is resource-intensive given the myriad combinations of feature subsets[28]. We circumvent this by retraining with a global explanation ranking. Intuitively, an optimal explanation method should also highlight impactful features on a *global level*. To quantify the efficacy of an explanation method, we propose using the knowledge distillation loss, $\mathcal{E}_{\mathrm{KD}}$. Formally, this evaluation is defined as

$$\mathcal{E}_{\mathrm{KD}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_s(X^k_i) - \tilde{y}_i)^2, \quad \text{where} \quad \hat{\tau}_s = \arg\min_{\theta}\mathcal{L}(\tau(X^k;\theta), \tilde{y}) \tag{6}$$

where $\hat{\tau}_s$ is a student model, $\mathcal{L}$ is the training loss depending on the types of CATE, X^k represents the top k features ranked by their average absolute attribution scores across training samples, and $\tilde{y}$ is the output from the ensemble (teacher) model, $\hat{\tau}(X)$. If the identified features are predictive of the treatment effect, $\mathcal{E}_{\mathrm{KD}}$ would be low in the testing set.

Our approach shares similarities with the Remove-and-Retrain (ROAR) method; however, in our setting, ROAR requires retraining every model in an ensemble whenever a feature is removed, imposing a heavy computational cost[28]. In contrast, our approach requires only a single student model at every removing step, significantly enhancing computational efficiency. Notably, knowledge distillation is the only way to obtain comparable model performance for an ensemble, as shown in [47]. This approach also bypasses the dilemma when selecting a baseline [17, 45]. Additionally, feature contribution on a global (cohort) level facilitates human evaluation[26, 48].

Global Feature Identification

Alternatively, if the ground-truth explanation or important feature is available, we propose computing *Spearman's rank correlation*[29] rankings derived from the explanation methods and the oracle. Specifically, in the context of the treatment effect, we consider interaction p-values [30] as ground truth. A lower p-value indicates a higher likelihood of a feature being an important factor in the treatment effect. To evaluate an explanation method in identifying important features on the global level, we propose computing *Spearman's rank correlation*[29]

$$\rho(g(\hat{\tau}, \phi), g(p)), \tag{7}$$

where ρ is the Spearman's rank correlation, $g(\hat{\tau}, \phi)$ denotes the global ranking according to the explanation method ϕ and model $\hat{\tau}$, and $g(p)$ indicates the ranking based on interaction p-values.

Data availability

The generation process for synthetic datasets is available on GitHub at `https://github.com/AliciaCurth/CATENets`. The IST-3 dataset is publicly accessible at `https://datashare.ed.ac.uk/handle/10283/1931`. The CRASH-2 dataset can be accessed at `https://freebird.lshtm.ac.uk/index.php/available-trials/`, with treatment allocations available upon request. Both the ACCORD and SPRINT datasets are available upon request at `https://biolincc.nhlbi.nih.gov/home/`.

Code availability

The code for training, inference, and evaluation of the CATE models and XAI methods used in this study will be made publicly available on GitHub upon publication. The code is distributed under the BSD 3-Clause License. The model weights are provided and intended for non-commercial use only.

Acknowledgement

We extend our gratitude to the CRASH-2 investigators for sharing treatment allocation data, and to the researchers in the Lee lab for their valuable discussions.

Funding

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Heckman, J. J. & Vytlacil, E. J. Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics* **6,** 4779–4874 (2007).
2. Hernán, M. A. & Robins, J. M. *Causal inference* 2010.
3. Frieden, T. R. Evidence for health decision making—beyond randomized, controlled trials. *New England Journal of Medicine* **377,** 465–475 (2017).
4. Austin, P. C. & Stuart, E. A. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine* **34,** 3661–3679 (2015).
5. Johansson, F., Shalit, U. & Sontag, D. *Learning representations for counterfactual inference* in *International conference on machine learning* (2016), 3020–3029.
6. Funk, M. J. *et al.* Doubly robust estimation of causal effects. *American journal of epidemiology* **173,** 761–767 (2011).
7. Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497* (2020).
8. Feuerriegel, S. *et al.* Causal machine learning for predicting treatment outcomes. *Nature Medicine* **30,** 958–968 (2024).
9. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66,** 688 (1974).
10. Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Social science & medicine* **210,** 2–21 (2018).
11. Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J. & Drazen, J. M. Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* **357,** 2189–2194 (2007).
12. Brookes, S. T. *et al.* Subgroup analysis in randomised controlled trials: quantifying the risks of false-positives and false-negatives (2001).
13. Sauerbrei, W. & Blettner, M. Interpreting results in 2× 2 tables: part 9 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International* **106,** 795 (2009).
14. Crabbé, J., Curth, A., Bica, I. & van der Schaar, M. Benchmarking heterogeneous treatment effect models through the lens of interpretability. *Advances in Neural Information Processing Systems* **35,** 12295–12309 (2022).
15. Martınez, J. A. *Interpretability for conditional average treatment effect estimation* 2021.
16. Covert, I., Lundberg, S. & Lee, S.-I. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research* **22,** 1–90 (2021).
17. Chen, H., Lundberg, S. M. & Lee, S.-I. Explaining a series of models by propagating Shapley values. *Nature communications* **13,** 4512 (2022).
18. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2,** 56–67 (2020).
19. Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic attribution for deep networks* in *International conference on machine learning* (2017), 3319–3328.
20. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).
21. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **41,** 647–665 (2014).
22. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **23,** 18 (2020).
23. Group, I.-3. C. *et al.* The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. *The Lancet* **379,** 2352–2363 (2012).
24. Roberts, I. *et al.* The CRASH-2 trial: a randomised controlled trial and economic evaluation of the effects of tranexamic acid on death, vascular occlusive events and transfusion requirement in bleeding trauma patients. *Health Technol Assess* **17,** 1–79 (2013).

25. Group, A. S. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *New England Journal of Medicine* **362,** 1575–1585 (2010).
26. Group, S. R. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine* **373,** 2103–2116 (2015).
27. Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* **116,** 4156–4165 (2019).
28. Hooker, S., Erhan, D., Kindermans, P.-J. & Kim, B. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems* **32** (2019).
29. Spearman, C. The proof and measurement of association between two things. (1961).
30. Christensen, R., Bours, M. J. & Nielsen, S. M. Effect modifiers and statistical tests for interaction in randomized trials. *Journal of clinical epidemiology* **134,** 174–177 (2021).
31. De Havenon, A. *et al.* Effect of Alteplase on Ischemic Stroke Mortality Is Dependent on Stroke Severity. *Annals of Neurology* **93,** 1106–1116 (2023).
32. Campbell, B. C., Meretoja, A., Donnan, G. A. & Davis, S. M. Twenty-year history of the evolution of stroke thrombolysis with intravenous alteplase to reduce long-term disability. *Stroke* **46,** 2341–2346 (2015).
33. Zinkstok, S., Vermeulen, M., Stam, J., De Haan, R. & Roos, Y. Antiplatelet therapy in combination with rt-PA thrombolysis in ischemic stroke (ARTIS): rationale and design of a randomized controlled trial. *Cerebrovascular Diseases* **29,** 79–81 (2009).
34. Patti, G. *et al.* Platelet function and long-term antiplatelet therapy in women: is there a gender-specificity? A 'state-of-the-art'paper. *European heart journal* **35,** 2213–2223 (2014).
35. Investigators, P.-T. & the ANZICS Clinical Trials Group. Prehospital Tranexamic Acid for Severe Trauma. *New England Journal of Medicine* (2023).
36. Curth, A. & van der Schaar, M. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. *arXiv preprint arXiv:2302.02923* (2023).
37. Schulam, P. & Saria, S. Reliable decision support using counterfactual models. *Advances in neural information processing systems* **30** (2017).
38. Heskes, T., Sijben, E., Bucur, I. G. & Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems* **33,** 4778–4789 (2020).
39. Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108,** 299–319 (2021).
40. Shi, C., Blei, D. & Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems* **32** (2019).
41. Shalit, U., Johansson, F. D. & Sontag, D. *Estimating individual treatment effect: generalization bounds and algorithms* in *International conference on machine learning* (2017), 3076–3085.
42. Sundararajan, M. & Najmi, A. *The many Shapley values for model explanation* in *International conference on machine learning* (2020), 9269–9278.
43. Mehrer, J., Spoerer, C. J., Kriegeskorte, N. & Kietzmann, T. C. Individual differences among deep neural network models. *Nature communications* **11,** 5725 (2020).
44. Dietterich, T. G. *Ensemble methods in machine learning* in *International workshop on multiple classifier systems* (2000), 1–15.
45. Sturmfels, P., Lundberg, S. & Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill* **5,** e22 (2020).
46. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
47. Allen-Zhu, Z. & Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816* (2020).
48. Lipkovich, I., Dmitrienko, A. & B D'Agostino Sr, R. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine* **36,** 136–196 (2017).
49. Almond, D., Chay, K. Y. & Lee, D. S. The costs of low birth weight. *The Quarterly Journal of Economics* **120,** 1031–1083 (2005).

50. Asuncion, A. & Newman, D. *UCI machine learning repository* 2007.

51. Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition (2019).

52. Abadie, A. & Imbens, G. W. Matching on the estimated propensity score. *Econometrica* **84,** 781–807 (2016).

53. Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20,** 217–240 (2011).

54. Robins, J. M. & Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90,** 122–129 (1995).

55. Alaa, A. & Van Der Schaar, M. *Validating causal inference models via influence functions* in *International Conference on Machine Learning* (2019), 191–201.

56. Van Der Laan, M. J. & Dudoit, S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples (2003).

57. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

58. Shapley, L. S. *et al.* A value for n-person games (1953).

59. Sechidis, K. *et al.* Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics* **34,** 3365–3376 (2018).