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

Determining which features drive the treatment effect for individual patients has long been a complex and critical question in clinical decision-making. Evidence from randomized controlled trials (RCTs) are the gold standard for guiding treatment decisions. However, individual patient differences often complicate the application of RCT findings, leading to imperfect treatment options. Traditional subgroup analyses fall short due to data dimensionality, type, and study design. To overcome these limitations, we propose CODE-XAI, a framework that interprets Conditional Average Treatment Effect (CATE) models using Explainable AI (XAI) to perform feature discovery. CODE-XAI provides feature attribution at the individual subject level, enhancing our understanding of treatment responses. We benchmark these XAI methods using semi-synthetic data and RCTs, demonstrating their effectiveness in uncovering feature contributions and enabling cross-cohort analysis, advancing precision medicine and scientific discovery.

Introduction

Quantifying the influence of an intervention on 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]. Randomized controlled clinical trials (RCTs) are the current gold standard for determining treatment effects [3]. However, applying such evidence towards treatment decisions for individual patients can be complicated by deviations in patient characteristics and clinical practice settings that differ from the strictly controlled conditions enforced during RCTs. As a result, clinicians are left guessing if the treatment identified in the RCT will benefit an individual patient when they differ in some way from those studied.

Attempts to understand why treatments are effective, and thus maximize their application, have traditionally been relegated to secondary objectives of RCTs that lack the power to drive changes in clinical practice. Subgroup analysis focuses on treatment outcome differences across patients based on observed covariates[2, 4, 5]. However, as data dimensionality increases, the number of potential subgroups increases exponentially, quickly overwhelming their application to patients and practice in the real world. [6, 7]. Subgrouping also typically relies on categorical variables while many features are continuous, and converting continuous features to categorical variables can lead to loss of important information, difficulty in determining the number and boundaries of categories, and risk of false discovery [7]. Moreover, it requires balanced treatment and control allocation within each subgroup, complicating the analysis of features or subgroups not accounted for in the original trial design [6, 8]. Finally, subgroup analysis fails to both provide insights into how individual characteristics affect treatment efficacy and to allow cross-cohort comparisons, even among groups with similar treatments or features.

To more effectively understand and quantify treatment effects, researchers have developed Conditional Average Treatment Effect (CATE) models [9]. CATE models aim to adjust for imbalances between control and treatment groups and leverage observed covariates to enhance the estimation of treatment effects. Numerous proposed approaches [10–13] address the question of *how* the treatment affects the outcome. However, these methods are tailored for optimal prediction, and do not inform robust feature or subgroup discovery. They fall short of answering two vital questions related to *why* estimations drive specific outcomes: (1) *which feature drives the treatment effect?* and (2) *why do individual responses to treatments vary?* Such factors differ across cohorts and are diverse and complex, so simply

measuring the treatment effect is insufficient to identify them. Thus, the need to *interpret* these CATE models provides a unique opportunity to answer these important questions.

To overcome these deficiencies, we propose *CODE-XAI, a framework that discovers feature that drives treatment effects by interpreting CATE models using Explainable AI (XAI)* [14, 15].In particular, *local explanation* methods[16], such as Integrated Gradient (IG) [17] and Shapley values [18, 19], 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 [20]. 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.

To obtain reliable attribution scores from CODE-XAI, we employed an *ensemble approach* and introduced *benchmarking techniques that assess both CATE and XAI methods*. Moreover, we propose a *novel subpopulation analysis using Shapley values* 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 RCT's to real world practice, differences in patient characteristics and alternative clinical practice settings. Finally, we demonstrate that CODE-XAI can successfully distill RCT treatment effects to the level of the individual patient.

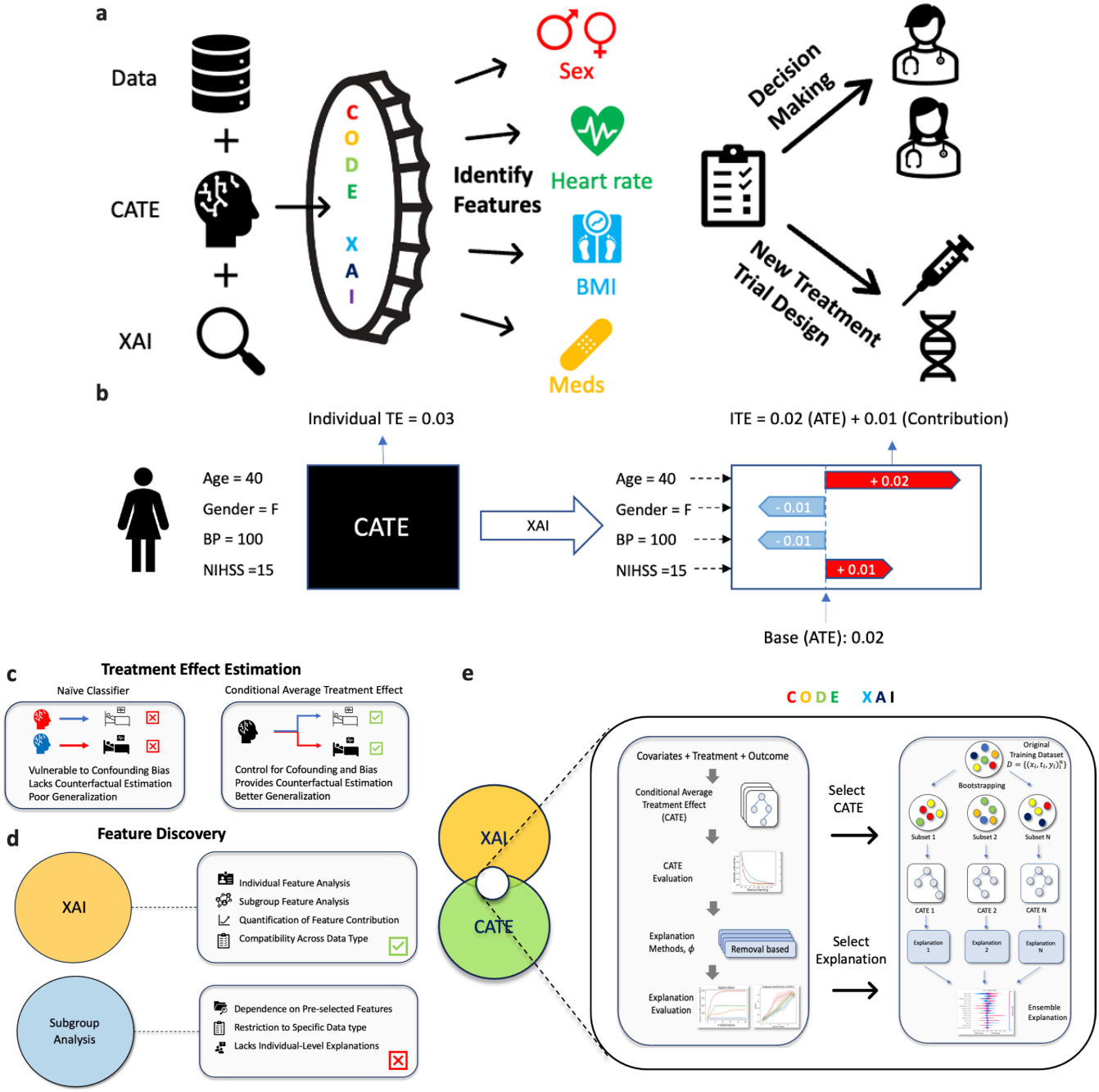

Fig. 1 | Overview of the CODE-XAI Framework. **(a)** Concept figure of the framework. **(b)** Individual explanations through XAI. **(c)** Treatment effect estimation: trade-offs between plugin estimates and conditional average treatment effect (CATE). **(d)** Feature discovery analysis: subgroup analysis 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 next examine the performance of both CATE models and their explanation in real-world datasets. We first train CATE models for each cohort, including IST3[21], CRASH-2[22], ACCORD[23], and SPRINT[24], and we obtain explanations with methods described in Section 0.2. Details of cohort description, datasets, and model implementations are in Appendix S1. We also conduct additional experiments in semi-synthetic environments to examine each explanation method (S4.1).

Estimating Real-World Treatment Effect with Ensemble CATEs

To obtain an accurate explanation, we first train CATE models to emulate treatment effects from four well-known randomized control trials [21–24]. We select the best-performing models according to their pseudo-outcome surrogate (Appendix S4.2), finding that X-learner outperforms other models in IST-3, CRASH-2, and SPRINT, while DR-learner performs best in ACCORD (Table S4).

Table 1 presents an ensemble estimate of the average treatment effect (ATE) for each cohort, including uncertainty estimates. CATE estimates for IST-3 and CRASH-2 are consistent with their reported findings [21, 22]. Interestingly, for the blood pressure control trials, i.e., SPRINT and ACCORD, the CATE model provides more optimistic estimates, showing improvements of 1.6% and 1.2% in primary outcomes, respectively, compared to 0.54% and 0.22% reported originally. The CATE estimation for SPRINT also demonstrates better ATE compared to ACCORD.

Cohort	CATE	Reported Findings
	Average Treatment Effect (%)	
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 between estimated Average Treatment Effect (ATE) from CATE and reported primary outcomes difference between treatment and control groups in four trials.

Enhanced Robustness in Explanations with Ensemble Models

We next demonstrate the importance of interpreting ensemble models over a single model. By measuring cosine similarities between explanations (0.3), single-model explanations exhibit low similarity and high variance, with scores of 0.13, 0.15, 0.15, and 0.21, as depicted in Figure 2(b-top). In contrast, ensemble explanations display greater consistency and robustness. As shown in Figure 2(b-middle), the average explanation similarity (Shapley value) within the ensemble increases from 0.6 with 10 models to 0.8 with 20 models, highlighting the enhanced reliability and consistency of explanations achieved through the ensemble approach.

Benchmarking XAI Methods on Real-World Clinical Data

To examine the obtained explanation, a commonly used approach is an ablation study, where features are systematically added or removed based on their importance ranking.[25]. However, individual-based ablation studies suffer from baseline selection bias (S4.3) and are computationally expensive for ensemble models. Instead, we propose a distillation technique (0.3) that focuses on global explanations, training student models on globally ranked features to emulate the CATE model's outputs across varying feature budgets (0.3).

As we show in Figure 2(c), both Shapley-mean and IG-mean consistently demonstrate lower distillation loss (mean squared error) across the SPRINT, ACCORD, and IST-3 datasets under various feature budgets. In contrast, within the CRASH-2 dataset, the performance of all methods is comparable except for Saliency, likely due to the dataset containing only 10 features, simplifying the task of identifying influential features. Our proposed evaluation shows that explanation methods with a population mean as the baseline outperform constant baselines.

In Table S6, we show the best methods and their top 5 features across different RCTs. In the CRASH-2 dataset, the top 5 features identified by IG-mean as important factors to treatment effect are injury type, gender, age, and gcs score; in contrast, Saliency ranks heart rate, respiratory rate, and capillary refill time as the top features.

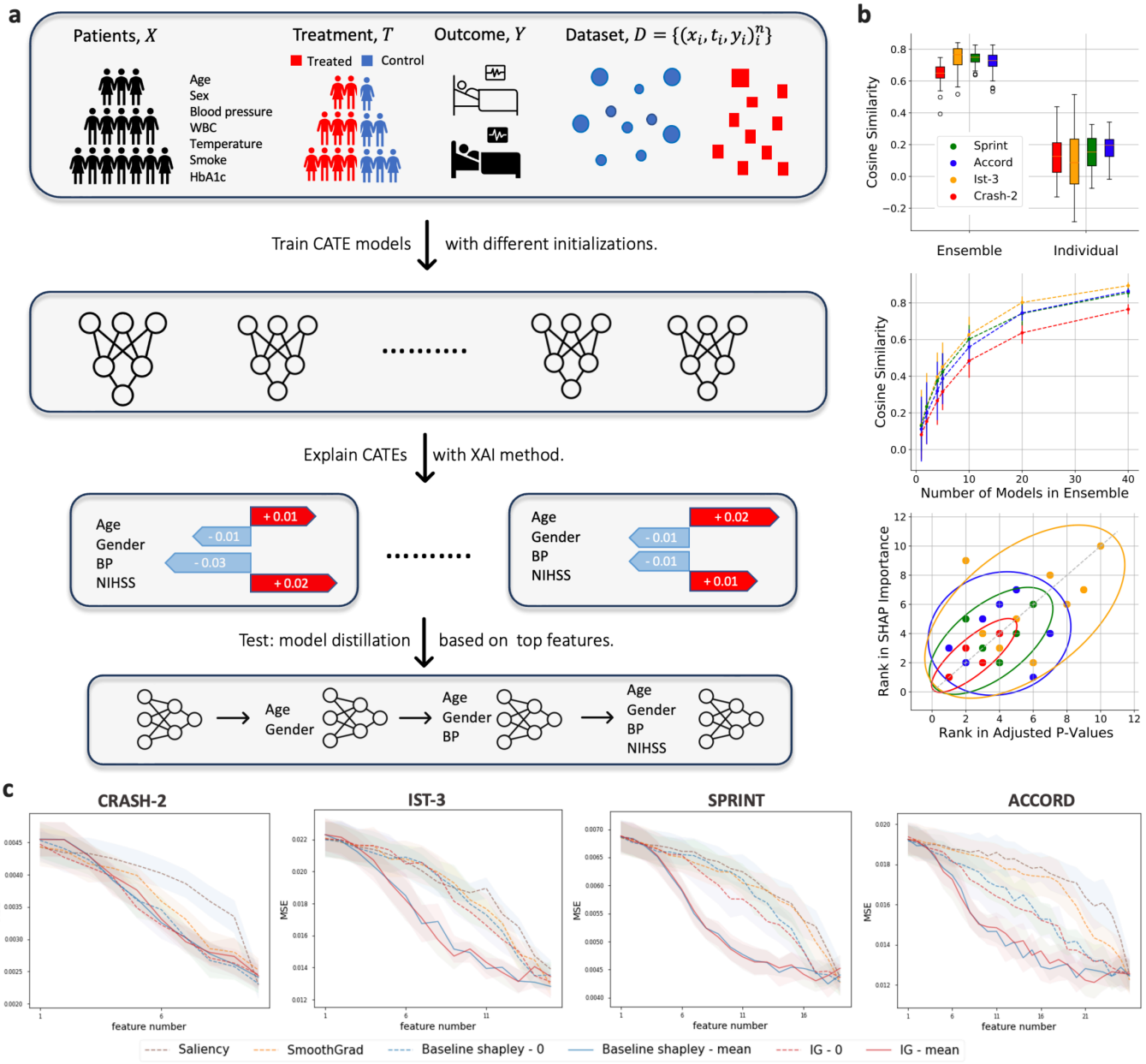

Fig. 2 | Results of Examining Ensemble Explanation.(a) *Evaluation and Explanation generation procedure* of CODE-XAI. Ensemble CATE models are trained with patients' data and different initializations. Features obtained through CATEs and XAI methods are used for follow-up evaluation. (b): (top) Comparison 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 CATE with Shapley Value

We now show how to use feature attributions obtained from CATE models to analyze clinical trials and their advantages relative to traditional subgroup analysis. Since gradient-based attribution is difficult to interpret[17], we employ Shapley value explanation methods to analyze clinical cohorts.

Global Feature Identification: Shapley Values versus RCT Findings

To assess the effectiveness of Shapley values in feature discovery, we compute the Spearman's rank correlation [26] between the global explanation from Shapley values and the features reported in original studies. For RCTs, we employ reported interaction p-values as proxies for feature ranking [27].

As shown in Table 2 and Figure 2(b-bottom), a significant correlation between Shapley rankings and reported features is observed, 0.8, 0.54, and 0.6 in CRASH-2, IST-3, and SPRINT, respectively, where CATE models accurately predict treatment effects. However, the correlation is low, 0.05, in the ACCORD study. This is expected given that no significant features have been reported[23], and the explanation would be less reliable when the CATE model struggles, Table 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 in RCT studies. A lower p-value indicates a higher likelihood of a feature being a treatment effect modifier.

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. Compared to traditional subgroup analysis, which requires subgrouping and computing risk or odds ratios, Shapley values enable direct analysis of feature impact at both individual and group levels. They provide *individual* explanations [14, 18] by decomposing the total treatment effect into each feature's contribution for every individual.

In Figure 3 (b), the upper force plot shows an example patient who experienced an increased survival probability of 11%, significantly above the ATE, which is 1.6%. The red bar indicates features that contribute positively to the treatment effect, including a high NIHSS score, TACI, and usage of anti-platelet within 48 hours; the blue bar indicates features that reduce the treatment effect, including atrial fibrillation history and higher systolic blood pressure. Conversely, the individual in the bottom force plot, a male patient with low NIHSS scores and PACI, had a treatment effect diminished by 11%.

On the cohort level, we analyze feature importance in IST3-trial by averaging their Shapley value across the cohort. Results show that the NIH Stroke Scale (NIHSS), a neurological examination for stroke evaluation, is the most influential feature affecting rt-TPA's efficacy; see Figure 3(c). Further, without categorizations or creating numerous subgroups, we can easily examine the impact of continuous features. The Shapley plot indicates that patients with higher NIHSS, depicted by the red cluster, demonstrate a pronounced improvement in treatment outcomes when administered TPA, in contrast to those with lower NIHSS scores, marked by the blue cluster. This observation is consistent with prior research [21, 28], which also identified a significant interaction between NIHSS scores and tPA treatment effectiveness.

Notably, the second most impactful feature is the type or syndrome of the stroke. In Figure 3(c), 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 [29]. 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 [29, 30].

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 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 [31].

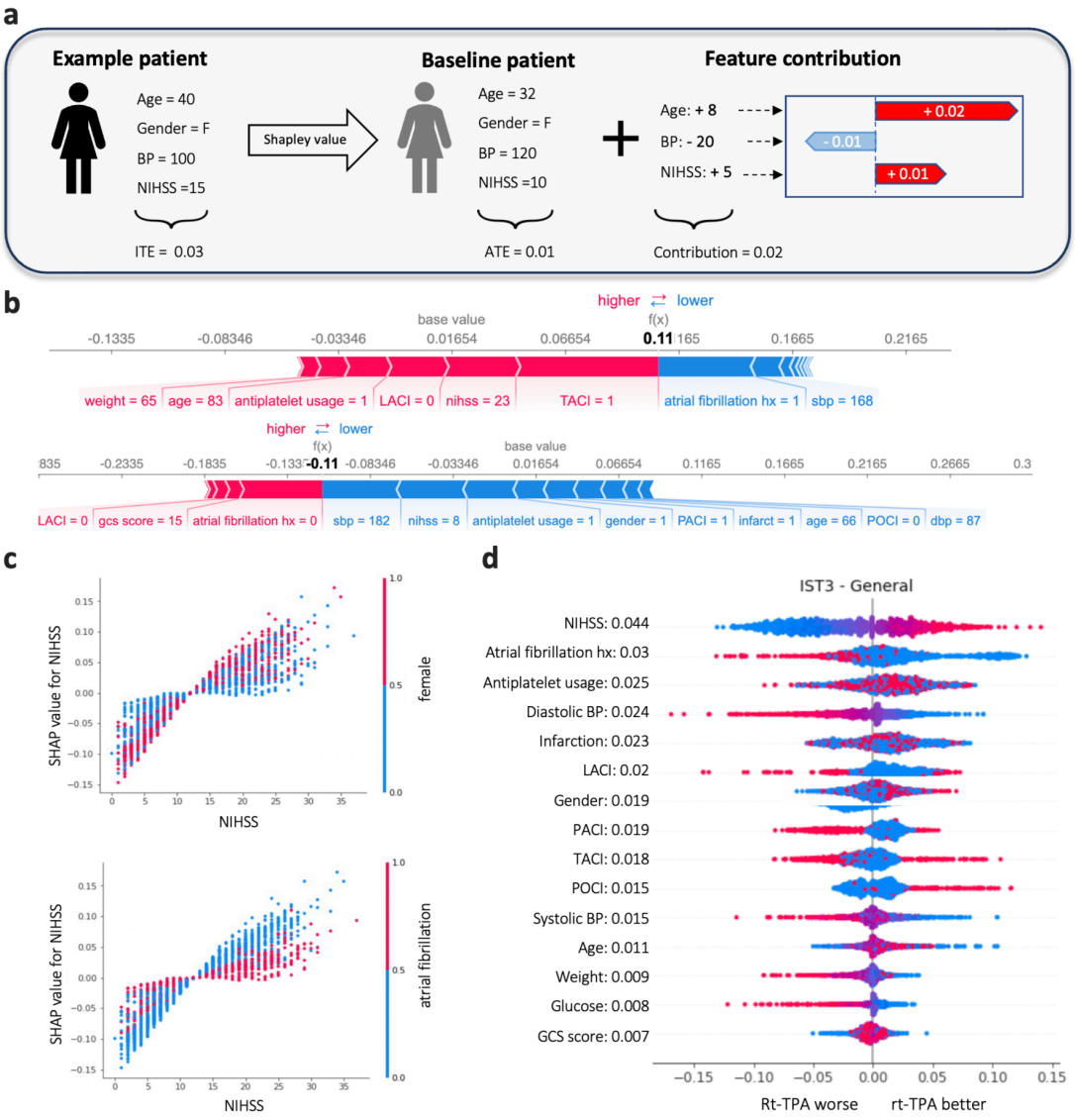

Fig. 3 | Analyzing the IST-3 Study with Shapley Values: (a) Decomposing feature contributions for an example patient with Shapley value. (b) Shapley values for example individuals, where red indicates positive attributions and blue represents negative attributions. (c) Combined Shapley values (left y-axis) 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[23, 24].

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 (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 (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[22]. 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 [22] 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[32].

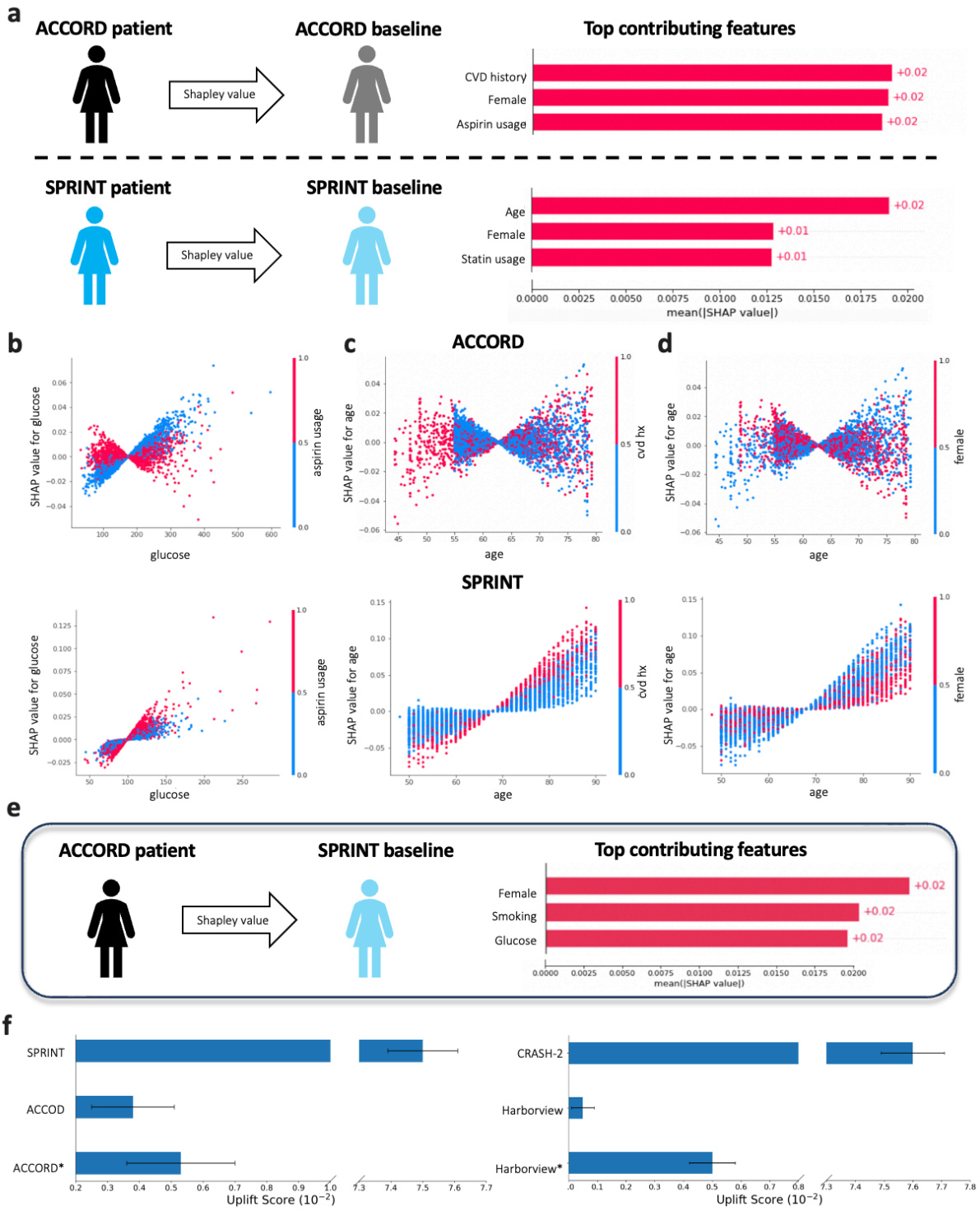

Fig. 4 | (a) Top 3 features based on mean absolute Shapley values in the ACCORD and SPRINT studies. The upper plots show results from the model trained with overlapped subsets, while the lower plots are from separately trained models.; Shapley scatter plots for feature pairs in separately trained models: (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 [32]. 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

Using explainable AI (XAI) in the life sciences continues to expand [33–35], however, its application, robustness, validity, and trustworthiness remain largely unexplored [25, 36–38]. 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 showcase that ensemble CATE models can reliably estimate treatment effects using real-world clinical data by comparing with factual outcomes and benchmarking pseudo-outcomes for model selection [39]. We then demonstrate that an ensemble explanation is more robust than the best single model. However, since examining explanations from an ensemble is not straightforward, we highlight the importance of global explanations and propose using knowledge distillation to benchmark feature attribution methods. This differentiates our method from those reliant on unrealistic assumptions regarding oracle accessibility in real-world scenarios [38], benchmark tests susceptible to inherent biases [15], or evaluations that are inefficient for ensemble models [25].

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 [21–24]. Compared to traditional analysis, our approach provides not only subgroup analysis but *individual* analysis without the need to analyze millions of strata [5]. By analyzing individual features, we observe how a single feature 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. We analyze two well-known trials, ACCORD [23] and SPRINT [24], 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. Though these models work well in controlled settings such as RCTs, it is difficult to obtain a reliable CATE model from observational studies with imbalanced treatment assignments. In the presence of unobserved confounders, the identifiability assumption would be violated, invalidating CATE model efficacy [9, 39] and leading to biased explanations [38]. Therefore, a promising research direction involves developing methods to impute robust attribution scores to mitigate selection bias. Additionally, some works incorporate causal knowledge to enhance the accuracy of feature attribution [40], but this assumption is often impractical in real-world experiments.

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* [38] 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[41], DR-learner[13], and R-learner[42]. 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[43], TARNets[44], CFR[11], 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[14] 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[17] and Shapley values[18].

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 [17]. 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[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 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 [19], Baseline Shapley[45] or KernelSHAP[18] 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 detail about these methods is 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 [46]. To address this, we employ an *ensemble* approach [47] 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 baseline can potentially favor particular explanation methods[15, 48]. To address this, we introduce an evaluation approach rooted in *knowledge distillation*[49], 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[25]. 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

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

where τ_s is a student model, 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[25]. 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 [50]. This approach also bypasses the dilemma when selecting a baseline [15, 48]. Additionally, feature contribution on a global (cohort) level facilitates human evaluation[24, 51].

Global Feature Identification

Alternatively, if the ground-truth explanation or important feature is available, we propose computing *Spearman's rank correlation*[26] rankings derived from the explanation methods and the oracle. Specifically, in the context of the treatment effect, we consider interaction p-values [27] 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*[26]

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

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/`.

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. 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).
5. Brookes, S. T. *et al.* Subgroup analysis in randomised controlled trials: quantifying the risks of false-positives and false-negatives (2001).
6. Druckman, J. N. & Green, D. P. *Advances in experimental political science: Subgroup Analysis: Pitfalls, Promise, and Honesty* (Cambridge University Press, 2021).
7. 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).
8. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* **22,** 1359–1366 (2011).
9. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66,** 688 (1974).
10. 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).
11. Johansson, F., Shalit, U. & Sontag, D. *Learning representations for counterfactual inference* in *International conference on machine learning* (2016), 3020–3029.
12. Funk, M. J. *et al.* Doubly robust estimation of causal effects. *American journal of epidemiology* **173,** 761–767 (2011).
13. Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497* (2020).
14. 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).
15. Chen, H., Lundberg, S. M. & Lee, S.-I. Explaining a series of models by propagating Shapley values. *Nature communications* **13,** 4512 (2022).
16. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2,** 56–67 (2020).
17. Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic attribution for deep networks* in *International conference on machine learning* (2017), 3319–3328.
18. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).
19. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **41,** 647–665 (2014).
20. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **23,** 18 (2020).
21. 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).
22. 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).
23. Group, A. S. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *New England Journal of Medicine* **362,** 1575–1585 (2010).

24. Group, S. R. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine* **373,** 2103–2116 (2015).
25. 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).
26. Spearman, C. The proof and measurement of association between two things. (1961).
27. 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).
28. De Havenon, A. *et al.* Effect of Alteplase on Ischemic Stroke Mortality Is Dependent on Stroke Severity. *Annals of Neurology* **93,** 1106–1116 (2023).
29. 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).
30. 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).
31. 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).
32. Investigators, P.-T. & the ANZICS Clinical Trials Group. Prehospital Tranexamic Acid for Severe Trauma. *New England Journal of Medicine* (2023).
33. Oikonomou, E. K., Spatz, E. S., Suchard, M. A. & Khera, R. Individualising intensive systolic blood pressure reduction in hypertension using computational trial phenomaps and machine learning: a post-hoc analysis of randomised clinical trials. *The Lancet Digital Health* **4,** e796–e805 (2022).
34. Moon, I. *et al.* Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. *Nature Medicine,* 1–11 (2023).
35. Ling, Y., Upadhyaya, P., Chen, L., Jiang, X. & Kim, Y. Emulate randomized clinical trials using heterogeneous treatment effect estimation for personalized treatments: methodology review and benchmark. *Journal of Biomedical Informatics,* 104256 (2022).
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. 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).
39. Feuerriegel, S. *et al.* Causal machine learning for predicting treatment outcomes. *Nature Medicine* **30,** 958–968 (2024).
40. 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).
41. 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).
42. Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108,** 299–319 (2021).
43. Shi, C., Blei, D. & Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems* **32** (2019).
44. Shalit, U., Johansson, F. D. & Sontag, D. *Estimating individual treatment effect: generalization bounds and algorithms* in *International conference on machine learning* (2017), 3076–3085.
45. Sundararajan, M. & Najmi, A. *The many Shapley values for model explanation* in *International conference on machine learning* (2020), 9269–9278.
46. Mehrer, J., Spoerer, C. J., Kriegeskorte, N. & Kietzmann, T. C. Individual differences among deep neural network models. *Nature communications* **11,** 5725 (2020).
47. Dietterich, T. G. *Ensemble methods in machine learning* in *International workshop on multiple classifier systems* (2000), 1–15.

48. Sturmfels, P., Lundberg, S. & Lee, S.-I. Visualizing the impact of feature attribution baselines. *Distill* **5,** e22 (2020).

49. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

50. Allen-Zhu, Z. & Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816* (2020).

51. 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).

52. Almond, D., Chay, K. Y. & Lee, D. S. The costs of low birth weight. *The Quarterly Journal of Economics* **120,** 1031–1083 (2005).

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

54. 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).

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

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

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

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

59. 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).

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

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

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