LIFT - XAI: Leveraging Important Features in Treatment Effects to Inform Clinical Decision-Making via Explainable AI

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

Clinicians rely on evidence from randomized clinical trials (RCTs) to decide on medical treatments for patients. However, RCTs often lack the granularity needed to inform decisions for individual patients or specific clinical scenarios. RCTs report subgroup analysis to find groups of patients who might respond better or worse according to certain characteristics, but RCT subgroups are arbitrarily determined and confined to the stringent limitation of the trial, thus providing limited generalizable information for real world use. Meta-learners, a more advanced class of data models built on conditional average treatment effect (CATE) frameworks, offer a promising solution for estimating treatment effects for individuals. However, identifying the key features that drive treatment effects for individual patients remain largely unexplored, limiting their potential for broadly advancing clinical decision-making. To overcome these barriers, we propose LIFT-XAI, a principled framework that interprets ensemble CATE models using Shapley value to uncover unique features and subgroups contributing to treatment effects in RCTs. We demonstrate that LIFT-XAI identifies the true features driving treatment benefits for both individual patients and subgroup cohorts. We benchmark LIFT-XAI using real-world clinical data, demonstrating its effectiveness in uncovering feature contributions robustly, and enabling cross-cohort analysis, advancing precision medicine—from chronic disease management to acute care.

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 pre-selected and well-defined cohort of patients sharing similar characteristics, Figure 1 (a). Treatment randomization is further used to effectively control for potentially confounding variables, e.g. age, gender, or health status[3], thus ideally isolating the treatment effect of the chosen intervention.

Building on this foundation, the conditional average treatment effect (CATE)[4] can extend these insights by quantifying how treatment effects vary across individuals or subgroups based on their covariates to estimate *counterfactual* outcomes, Figure 1 (b). In particular, recent works have advanced CATE by leveraging neural networks to learn individual treatment effects (ITEs)[5–9]. Despite making progress, these approaches are primarily evaluated on (semi)synthetic or non-clinical datasets that fail to capture the full complexity of real-world population or disease dynamics[9]. 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 cannot 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]. Feature discovery in this context involves examining ATE differences across predefined covariates to uncover factors influencing treatment efficacy[2,11,12]. For example, varying ATEs for patients of different blood pressure (BP) ranges can identify BP as a key feature, Figure 1 (c). 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 in some respects from those studied in the clinical trial.

Recent works have explored the use of *eXplainable AI (XAI)* under the framework of CATE, including post-hoc and ante-hoc approaches[14,16,17]. Self-interpretable models, such as Causal Forests[18] and related variants[19,20], provide built-in measures of variable importance, but are constrained by model design and often underperform neural network–based meta-learners[16]. Global variable removal-based approaches[16,21,22] quantify population-level feature importance but overlook individual-level contributions, thereby restricting their value for personalized applications. On the other hand, local explanation methods[23], such as Shapley values[24,25], are model-agnostic and have shown promise in semi-synthetic evaluations[14,17], but remain largely untested in large-scale real-world clinical datasets. Moreover, most existing approaches are single-model explanations[14], making them highly sensitive to stochasticity in training (e.g., random seeds, hyperparameters)[26]. This instability raises concerns about whether explanations from a single best model can reliably reflect features in the true treatment effect function, $\tau^\star$.

To address these gaps, we propose *LIFT-XAI*, a framework for interpreting ensemble CATE models with Shapley values. Unlike existing single-model approaches, LIFT-XAI combines Shapley value with model ensembling to provide consistent explanations. Shapley values[24,25] are particularly suitable in this setting because they directly decompose the predicted treatment effect (i.e., the CATE model's output) into feature-level contributions without requiring grouping or transformations[27], Figure 1 (d). This enables fine-grained, individual-level attribution that enhances our understanding of why certain individuals may respond more favorably to treatment than others. Moreover, by aggregating these local explanations, LIFT-XAI recovers global-level insights, thereby bridging the gap between personalized and population-level interpretability, Figure 1 (e). Through explaining ensemble CATE, LIFT-XAI delivers reliable attribution scores, offering a robust framework for understanding heterogeneous treatment effects.

Furthermore, to facilitate clinical translation, we benchmarked two core components of LIFT-XAI by assessing both ensemble CATE and Shapley values, and demonstrated that LIFT-XAI not only outperforms other local attribution methods but also surpasses global approaches in identifying feature importance at the cohort level. The features identified by LIFT-XAI aligned closely with key features reported in existing studies on the cohort level. Notably, clustering in the explanation space reveals that patients with similar features may respond to treatment through different underlying mechanisms. Moreover, we propose a *cross-cohort analysis* based on feature attribution to uncover clinical feature interactions and resolve conflicting results across different trials. We then tested LIFT-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.

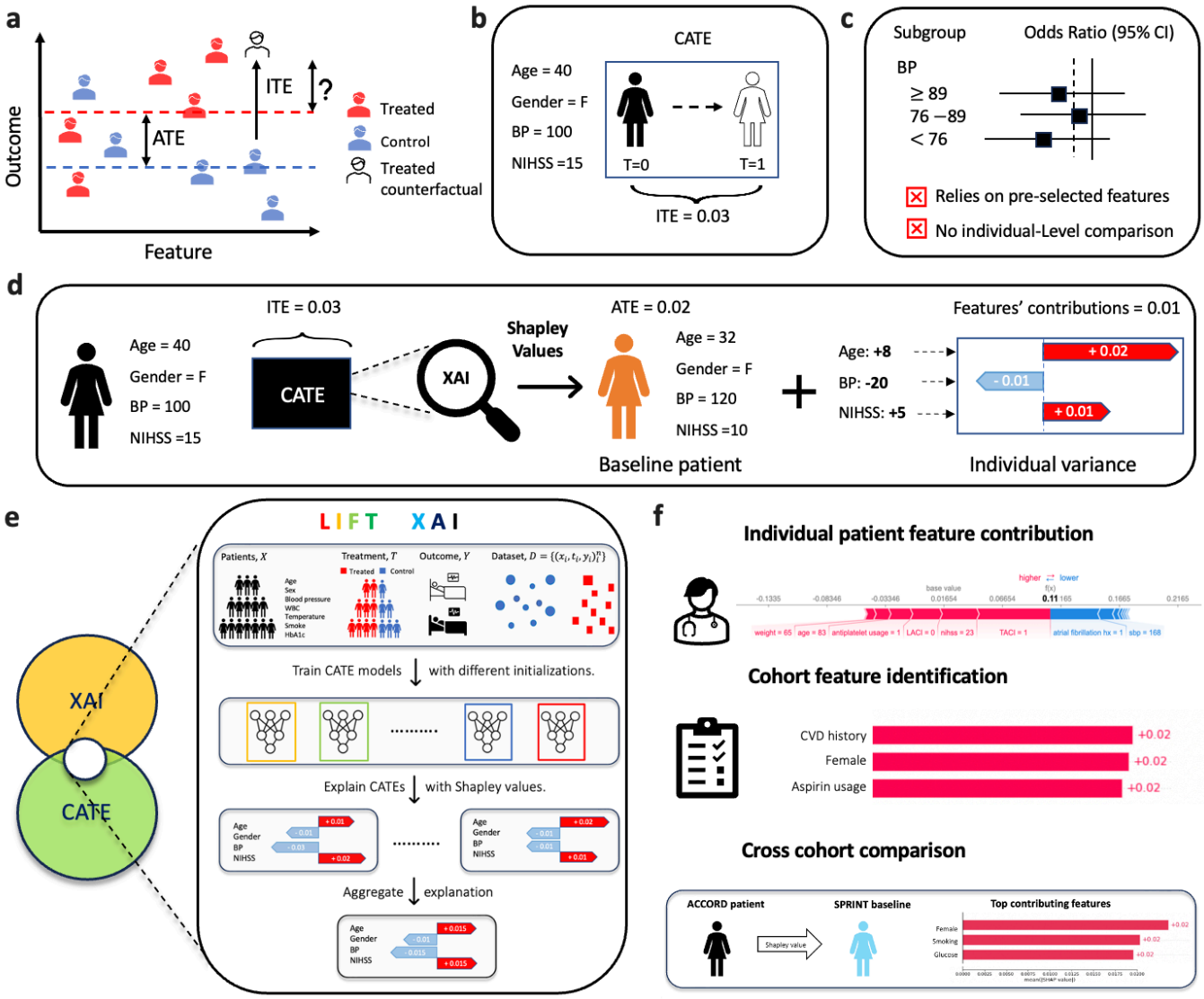

Fig. 1 | Conceptual background and the LIFT-XAI framework. (a) Treatment effect estimation distinguishes individual treatment effects (ITE) from the average treatment effect (ATE). (b) Conditional average treatment effects (CATEs) capture heterogeneity by conditioning on patient covariates. (c) Conventional subgroup analyses rely on pre-specified features and do not provide patient-level comparisons. (d) Explainable AI (XAI) enables explaining black-box models; Shapley values quantify feature contributions relative to a baseline patient. (e) **LIFT-XAI:** proposed framework that explains ensemble CATE models with Shapley values, aggregating stable attributions across learners to reveal robust, clinically meaningful signals that support decision-making and exploratory analysis. (f) Use cases for LIFT-XAI. Top: Individual-level feature contributions derived from LIFT-XAI for a representative patient. Middle: LIFT-XAI provides cohort-level feature attribution of treatment effects. Bottom: LIFT-XAI enables cross-cohort comparison, identifying features that drive differences in treatment effects between similar patient populations.

Table 1 | Model performance on IST-3, CRASH-2, SPRINT, and ACCORD. Metrics include pseudo-surrogate outcomes ($\mathcal{E}_{\text{if-pehe}}$, $\mathcal{E}_{\text{R}}$, $\mathcal{E}_{\text{DR}}$) and incremental gain scores (Qini and Uplift). Results are reported as mean ± standard deviation across random seeds. Higher values indicate better performance for Qini and Uplift, whereas lower values indicate better performance for $\mathcal{E}_{\text{R}}$, and $\mathcal{E}_{\text{DR}}$.

	IST-3[28]		CRASH-2[29]		SPRINT[30]		ACCORD[31]	
	DR-Learner	DR-Learner (Ens.)	DR-Learner	DR-Learner (Ens.)	DR-Learner	DR-Learner (Ens.)	RA-Learner	RA-Learner (Ens.)
Qini	0.284 (0.034)	**0.358 (0.016)**	0.136 (0.004)	**0.186 (0.004)**	0.066 (0.003)	**0.117 (0.003)**	0.096 (0.007)	**0.212 (0.008)**
Uplift	0.300 (0.025)	**0.395 (0.019)**	0.248 (0.005)	**0.316 (0.008)**	0.125 (0.005)	**0.212 (0.005)**	0.183 (0.012)	**0.358 (0.011)**
$\mathcal{E}_{\text{R}}$	0.521 (0.008)	**0.520 (0.008)**	**0.373 (0.003)**	0.373 (0.003)	0.241 (0.004)	**0.236 (0.006)**	0.305 (0.005)	**0.300 (0.005)**
$\mathcal{E}_{\text{DR}}$	0.921 (0.016)	**0.919 (0.015)**	0.684 (0.005)	**0.678 (0.004)**	**0.473 (0.007)**	0.472 (0.010)	0.592 (0.009)	**0.580 (0.009)**

Results

Benchmarking Explainee CATE on Real-World Clinical Data

We evaluated LIFT-XAI, our method for explaining ensemble CATE models using Shapley values. Because explanation quality depends on both the model and the explanation method, we assess the ensemble CATE and Shapley values separately, then show how LIFT-XAI improves transparency and interpretability in clinical trial feature discovery.

We first benchmarked different CATE learners across four large clinical trial datasets[28–31] and selected the best-performing learner to compare single-model and ensemble models. Performance was assessed using their prediction error (pseudo-outcome surrogates) and incremental gain (Qini/Uplift score) on held-out test sets (see Section B.3). The doubly robust learner (DR-Learner)[8] outperformed other learners on IST-3, CRASH-2, and SPRINT, with the lowest pseudo–outcome errors (0.921, 0.684, and 0.473) and highest Qini scores (0.217, 0.072, and 0.066), respectively (Table S.3). In contrast, the regression-adjustment learner (RA-Learner)[32] achieved the best performance on ACCORD. Self-interpretable models such as Causal Forest and Linear DR underperformed relative to neural network-based learners. Moreover, building ensembles from the best-performing learner consistently outperforms single models (Table 1). In addition, we computed the cohort-level average treatment effects (ATEs) based on ensemble CATEs and found that they aligned well with reported trial results for IST-3 and CRASH-2, while estimates for SPRINT and ACCORD were slightly higher. (Table S.4).

Together, these results demonstrate that ensemble CATE learners yield reliable treatment effect estimates consistent with real-world trial findings, making them the most suitable explainees.

Benchmarking Explanation Methods in Real-World Clinical Data

We next benchmarked our choice of explanation methods, Shapley values, against other explanation approaches by assessing their local fidelity, global sufficiency, and robustness (see Methods).

Local fidelity Using insertion–deletion tests[33], we evaluated whether features ranked as important by each method meaningfully affected predictions. Figure 2 (a) shows the deletion and insertion curves for the SPRINT dataset. In the deletion test (Figure 2 (a-top), pseudo-surrogate error increased as top contributing features were removed, with Shapley values outperforming Integrated Gradients (IG), LIME, and Saliency. In the insertion test (Figure 2 (a-bottom), IG performed comparably to Shapley values. Overall, when comparing Area Under the Curve (AUC), our results show that the Shapley value method achieves the highest normalized AUC in the ACCORD, SPRINT, and CRASH-2 datasets, Section D.3. In contrast, for the IST-3 dataset, LIME performs best in insertion, while the Shapley value performs best in deletion. These results indicate that Shapley explanations faithfully capture the features driving individual-level treatment effect heterogeneity.

Global sufficiency We then examined global (cohort-level) feature importance using our proposed distillation benchmark (Methods), which measures whether student models with top-ranked features can reconstruct ensemble outputs.[1] In ACCORD, Shapley values achieved the lowest distillation loss under restricted feature budgets, highlighting their efficiency in capturing features most relevant for CATE estimation, Figure 2 (b). In SPRINT and IST-3, Shapley value and LIME perform similarly to global feature importance approaches such as PermuCATE[22] and outperform LOCO[21]. We also qualitatively examined the top features identified by each method (Table S.9). In CRASH-2, both IG and Shapley values highlighted clinically meaningful factors such as GCS score and injury type, whereas Saliency tended to emphasize less informative features, such as capillary refill time.

[1] For local attribution methods, features were ranked by their average absolute attribution across individuals.

Consistency Finally, we evaluated the stability of Shapley explanations by comparing global attributions from single versus ensemble models using cosine similarity. Single-model explanation exhibited low similarity and high variance, with scores of 0.13, 0.15, 0.15, and 0.21, Figure 2 (c-middle). By contrast, ensemble-based Shapley explanations were far more consistent: the average similarity increased from 0.6 with 10 models to 0.8 with 20 models (Figure 2 c-left), highlighting the enhanced reliability and stability.

Taken together, these findings show that Shapley values yield high-fidelity explanations at both local and global levels, while ensembling further enhances robustness. This motivates our proposed **LIFT-XAI**, which combines ensemble CATEs with Shapley values to provide reliable feature attributions for treatment effect estimation.

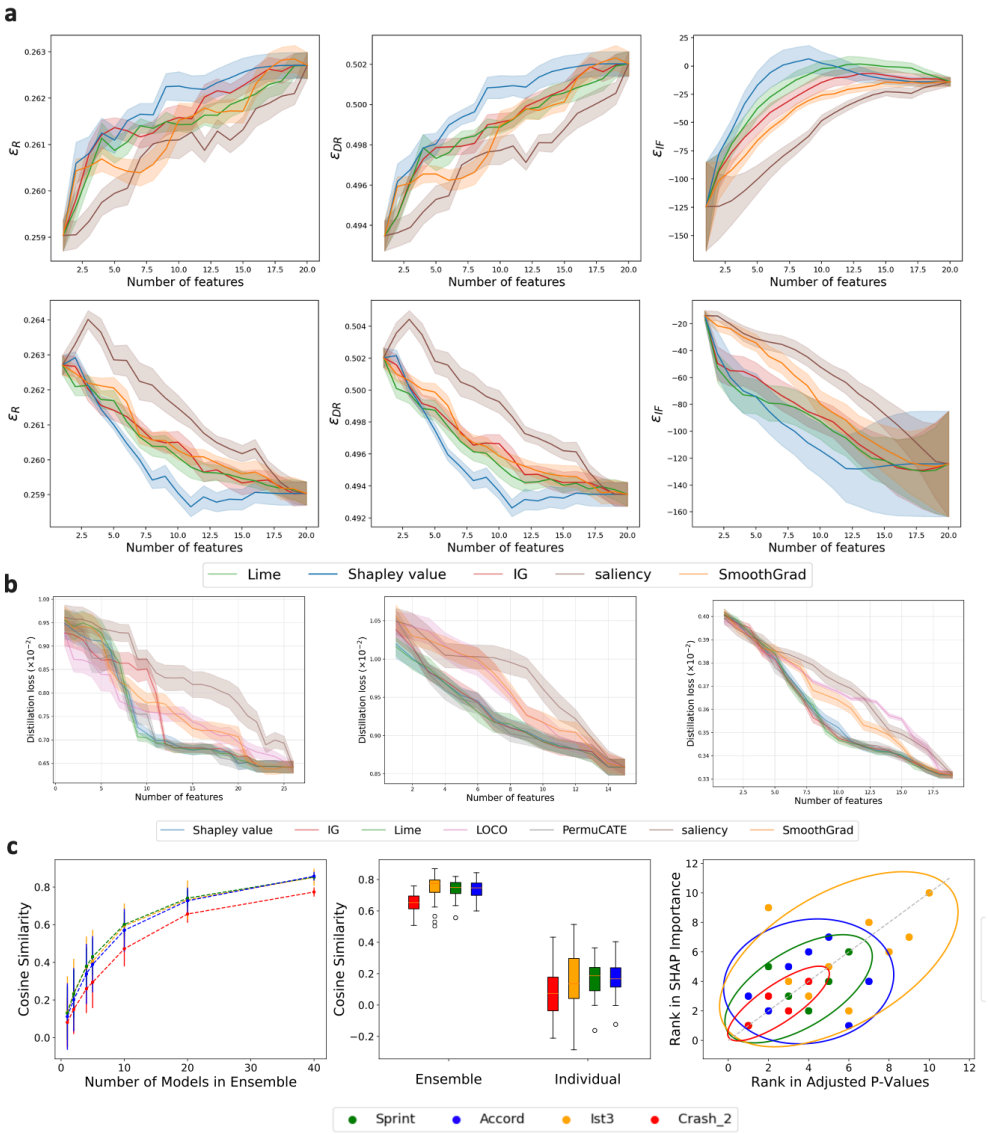

Fig. 2 | Evaluation of explanation methods for heterogeneous treatment effect models. (a) Local fidelity: deletion (top) and insertion (bottom) experiments on the SPRINT dataset using pseudo-outcome–based metrics. The x-axis denotes the number of features removed or included, and the y-axis reports surrogate error. (left to right). **(b)** Global sufficiency: Knowledge distillation performance across datasets, ACCORD, IST-3, and CRASH-2 (left to right). The x-axis denotes the number of features retained in student models, and the y-axis shows the distillation loss (MSE) relative to the ensemble teacher. **(c)** Consistency evaluation of explanation methods. (Left) Cosine similarity of feature attributions as a function of the number of models in the ensemble. (Middle) Distribution of cosine similarities for ensemble-averaged versus individual model explanations. (Right) Concordance between Shapley importance ranks and interaction p-value ranks across clinical datasets, with 95% confidence ellipses.

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

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

Validating LIFT-XAI for Cohort-Level Feature Discovery

We next evaluated whether LIFT-XAI could recover known treatment effect modifiers at the cohort level. To do this, we compared its global feature rankings with those reported in the original studies using Spearman's rank correlation[34]. For these studies, reported interaction p-values[2] were used as proxies for feature ranking[35]. 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; Figure 2 (c-right). In contrast, the ACCORD study shows a low correlation (0.05), which is expected, as no significant features were reported[31].

In the following sections, we will demonstrate how LIFT-XAI can be used to analyze clinical trial data. In what follows, we use the terms Shapley values and LIFT-XAI interchangeably.

Use Case 1: Analyzing Features' Contribution to rt-PA Treatment Effect

We begin with 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, LIFT-XAI allows direct analysis of feature impact at both individual and cohort levels. It provide individual-level explanations[25,36] 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 National Institutes of Health Stroke Score (NIHSS), Total Anterior Circulation Infarct (TACI), and usage of anti-platelet within 48 hours; the blue bars indicate 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 Partial Anterior Circulation Infarct (PACI) syndrome, who experienced a negative treatment effect.

Figure 3 (c) 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[28,37], which also identified a significant interaction between NIHSS scores and rt-PA treatment effectiveness.

Additionally, the second most impactful feature is the type or syndrome of the stroke. As shown in Figure 3 (c), rt-PA 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[38]. 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-PA, which previous studies have also discovered[38,39].

With LIFT-XAI, we can perform subgroup analysis by grouping patients based on feature values and examining their Shapley contributions. This also enables us to explore feature interactions—how one variable affects another's influence on treatment effect. In Figure 3 (d-top), we analyze the interaction between gender and NIHSS score. While NIHSS is a primary driver of rt-PA effect, LIFT-XAI reveals a gender-specific pattern: in females (red), high NIHSS scores contribute more positively, whereas low scores contribute more negatively. In contrast, the effect is less pronounced in males. A similar pattern is observed among patients without atrial fibrillation, where rt-PA appears significantly more effective in those with higher NIHSS scores, Figure 3 (d-bot).

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

Use Case 2: Attribution-based Clustering Analysis

We further explored explanation-based patient clustering in a subset of individuals from the IST-3 trial. Specifically, we focused on patients with NIHSS scores between 6 and 14—a group with uncertain treatment effects for or against rTPA treatment according to trial subgroup analysis, and no clear clinical guidelines. To address this, we applied LIFT-XAI to examine this subgroup through their Shapley values and performed hierarchical clustering directly in the explanation space.

This approach provides deeper insights and highlights substantial heterogeneity in feature contributions—even for dominant variables like the NIHSS score—indicating that similar clinical profiles based on a single feature do not necessarily imply similar treatment effect mechanisms, Figure 3 (e). Feature clustering reveals distinct patient clusters where features such as diastolic blood pressure and body weight contribute differently to the rT-PA effect, thus offering unique patient groups and potentially modifiable features that can be optimized to improve treatment effect where previous RCT's suggest none exist, Figure 3 (f). (For details of feature distribution for each cluster, please refer to Table S.11.) By moving beyond feature-based subgrouping to explanation-based subgrouping, LIFT-XAI addresses the limitations of traditional subgroup analysis and uncovers deeper patterns of treatment effect heterogeneity.

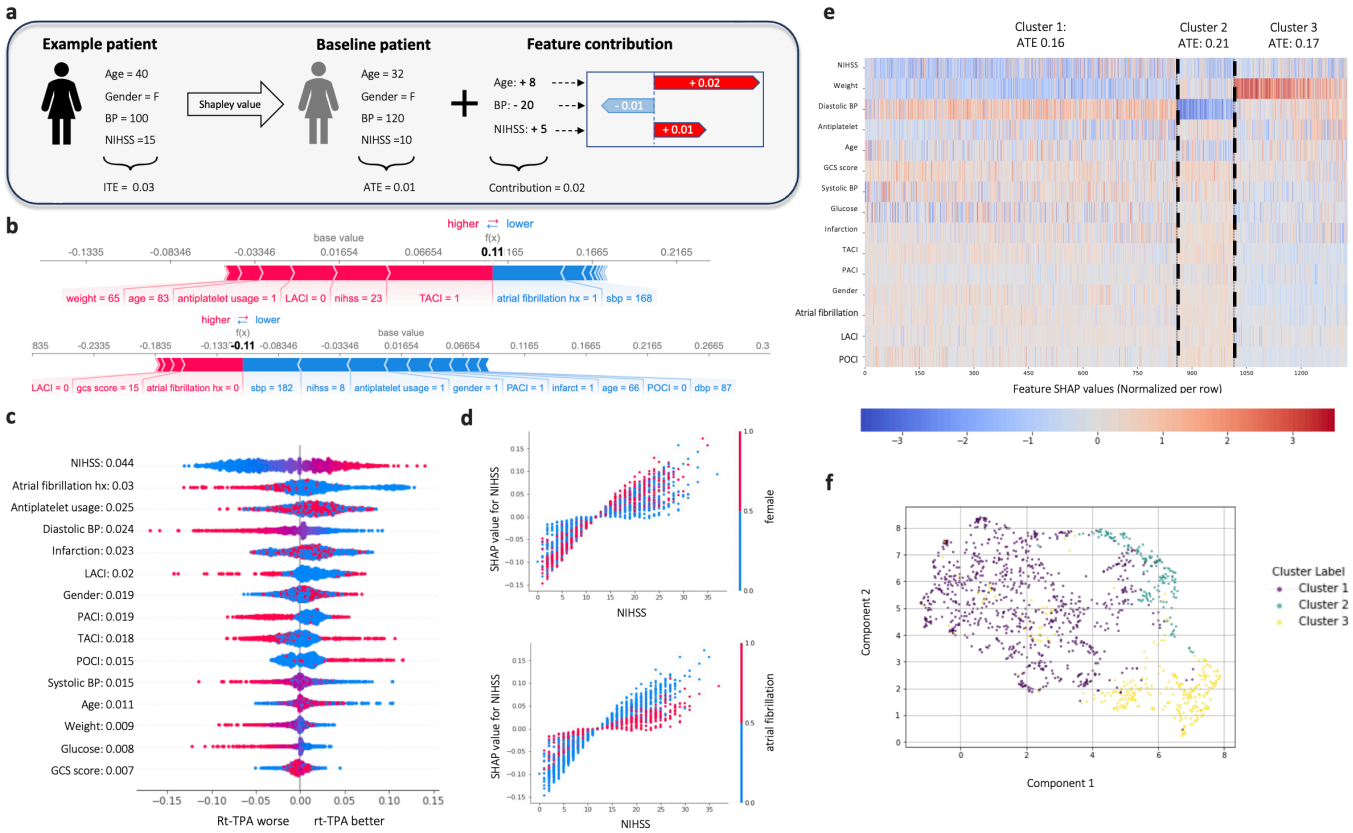

Fig. 3 | Analyzing the IST-3 Study with LIFT-XAI: (a) Decomposing feature contributions for an example patient with Shapley value. (b) Shapley values for example individuals, where red indicates positive attributions and blue represents negative attributions. (c) IST-3 summary plot showing features on 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. (d) 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. (e) Hierarchical clustering of IST-3 patients with NIHSS scores between 6 and 14 based on their corresponding SHAP values. (f) Hierarchical clustering based on Shapley value and corresponding heatmap. t-SNE visualizations of patient clusters.

Use Case 3: Deciphering Treatment Effects Across RCTs 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 LIFT-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[30,31].

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, group-level differences in Shapley values were more evident in the SPRINT trial, particularly for variables like age and CVD history, or age and gender (Figure 4c-bottom, d-bottom). In contrast, such subgroup variation was largely absent in the ACCORD trial. In some cases, the direction of feature importance even appeared reversed—for instance, in the case of glucose levels and aspirin use (Figure 4b-top).

To enable further cross-cohort comparisons of ACCORD vs. SPRINT, we substituted the ACCORD baseline of diabetic individuals with representative non-diabetic individuals from the SPRINT cohort (Section C.2). This reframes the original question—*which features are important for ACCORD individuals in compared to the average ACCORD population?*—to instead ask *which features are important for ACCORD individuals compared to the average SPRINT population?* By adjusting the reference group accordingly, this approach supports tailored, context-specific interpretations across studies.

Upon reassessing the top features from both cohorts and reanalyzing the feature rankings, we observed that *fasting glucose (fpg) emerged as a prominent feature for diabetic patients in ACCORD, but it ranked 14th among the 18 clinical features for non-diabetic patients in SPRINT*; see Figure S.8 (a). By identifying fasting glucose as a key treatment effect, LIFT-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, LIFT-XAI independently provided a clear and usable treatment metric (fasting glucose) for clinicians seeking to manage blood pressure in diabetic patients, suggesting that intensive blood pressure management may actually be useful in diabetic patients if blood glucose is tightly controlled

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 (Section B.3.1) among patients with varying glucose levels. As we show in Figure 4 (f-left) and Table S.12, 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 LIFT-XAI, we thus unravel these conflicting results in trials. Our analysis highlights variances in glucose levels as a potential explanatory and modifiable risk factor for the observed disparities in treatment outcomes between the two studies.

Use Case 4: Applying LIFT-XAI across Clinical Practice Settings.

Here, we test the ability of LIFT-XAI to identify important features contributing to 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[29]. Time from injury has emerged as having an important effect on TXA efficacy. So clinical practice has steadily crept towards using this drug faster, either 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 LIFT-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[29] and our local trauma center registry, we asked LIFT-XAI to identify features that determine TXA efficacy when administered in these different clinical practice settings (Section A.3). We then validated the feature selected by LIFT-XAI in the new pre-hospital setting by computing the treatment effect gain. To validate our results, we then compared them to features identified during a more recent randomized controlled trial of TXA when given

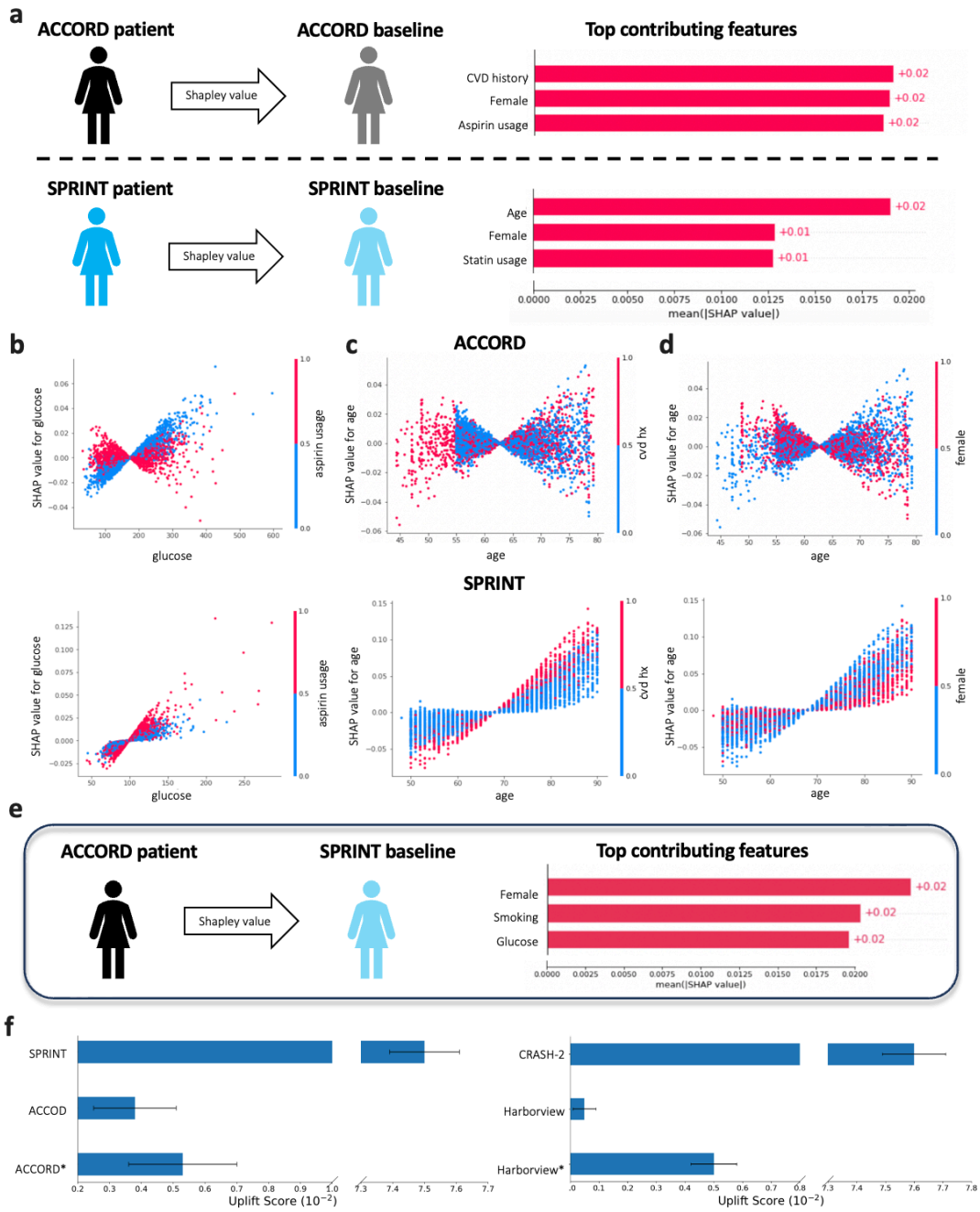

Fig. 4 | (a) Top 3 features based on mean absolute Shapley values in the ACCORD and SPRINT studies. The upper plots show results from the model trained with overlapped subsets, while the lower plots are from separately trained models.; Shapley scatter plots for feature pairs in separately trained models: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

specifically in the pre-hospital setting[40].

We first compared the top features based on their Shapley values. As shown in Figure S.10 (a-left), in the pre-hospital settings, the top features were time-to-injury, Glasgow Coma Score (GCS), trauma type, and a new effect, *age* that was not previously identified in CRASH-2 subgroup analysis. We then examined the treatment effects among different age groups. As shown in Figure 4 (f-right) and Table S.12, 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 $\times 10^{-3}$ and 8 $\times 10^{-4}$, respectively. This finding indicates that, in the pre-hospital setting, LIFT-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 [40]. This result highlights the ability of LIFT-XAI to identify important treatment effects when randomized clinical trial data are applied towards different clinical practice settings.

Discussion

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 treatment decisions. In doing so, we present a framework, LIFT-XAI, that rigorously explains CATE ensemble models with Shapley values, improving our ability to realize how randomized controlled trial data can be applied to real-world treatment decisions in a robust manner.

We demonstrated that ensemble CATE models can effectively estimate treatment effects using real-world clinical data in a controlled environment with proper model selection criteria. Furthermore, Shapley values consistently outperform alternative attribution methods in terms of local fidelity, global sufficiency, and explanation stability, making them particularly well-suited for interpreting the CATE estimates. When combined, these techniques strengthen the utility of feature attribution methods and CATE models, facilitating the integration of machine learning and XAI into clinical translation [9] and enabling a deeper understanding of the potential risks and benefits of new treatments for the individual patient.

LIFT-XAI enabled identification of key clinical features that could be used to modify treatment decisions across a range of both acute and chronic health conditions. [28–31] At the cohort level, LIFT-XAI was capable of independently explaining conflicting results between trials and identified key covariates. For example, the effect of intensive blood pressure management on health outcomes was studied in two RCTs, SPRINT and ACCORD. The trial results did not agree. The SPRINT RCT demonstrated benefit [30], however, the ACCORD RCT showed no benefit for a patient cohort that included those with diabetes mellitus [31], suggesting the presence of diabetes as having a confounding effect upon blood pressure treatment efficacy. We used LIFT-XAI to explain the treatment effects for these RCTs and it independently identified blood glucose concentration as having a primary influence on blood pressure treatment efficacy. This result is important because it both supports the evidence that intensive blood pressure treatment does not benefit those patients with diabetes mellitus, thus improving patient selection for this therapy. In addition, this result suggests that blood glucose management is a potential therapeutic maneuver that may be used to allow patients with diabetes to derive benefit from intensive blood pressure treatment once their blood glucose concentration is controlled. Importantly, we found that LIFT-XAI treatment effect explanations could be extended to individual patients, highlighting its potential to inform personalized treatment decisions.(Figure 3).

However, LIFT-XAI is constrained by inherent limitations in both CATE models and Shapley values. Although CATE estimators perform well in controlled settings such as randomized controlled trials (RCTs), their reliability is reduced in observational studies where unobserved confounders may bias effect estimates. Such confounding can violate key identification assumptions, compromising model validity [4,9] and propagating bias into the resulting explanations [14]. Beyond model limitations, while feature attribution methods are valuable for highlighting candidate features of treatment effect heterogeneity, they cannot establish causal relationships or guarantee clinically actionable conclusions [41], and are sensitive to input and model perturbation [42]. Misinterpreting attribution scores as causal drivers may therefore lead to misleading inferences about treatment effect mechanisms. We therefore recommend using LIFT-XAI primarily as an exploratory tool for feature discovery, while emphasizing that rigorous clinical trials remain essential to validate clinically actionable insights.

A promising future research direction involves developing methods to impute robust attribution scores to mitigate the impact of these confounders [43]. Additionally, integrating causal knowledge with domain expertise has been demonstrated to enhance the accuracy of feature attributions [44].

To conclude, we present a new approach to explaining treatment effects using LIFT-XAI. We propose evaluation methods to assess CATE models with XAI using the gold standard for determining treatment efficacy in medicine, randomized clinical trials. LIFT-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

Datasets

Most prior works of interpretability methods for conditional average treatment effects (CATEs) have focused on semi-synthetic or non-clinical benchmarks, such as IHDP and ACIC[14,16]. Only a few studies have examined clinical data, typically restricted to a single trial with modest sample sizes[17,21,45], leaving evaluation in real-world clinical settings underdeveloped.

To address this gap, we analyzed four publicly available randomized controlled trials (RCTs) with binary outcomes. These trials were chosen for their clinical significance and their large sample sizes, making them a reliable testbed for CATE models. Randomization in RCTs not only reduces confounding but also provides the conditions needed for valid estimation of treatment effects[2]. Each dataset was curated by removing variables with substantial missingness and applying one-hot encoding to categorical variables. The resulting datasets are:

IST-3. This trial evaluated intravenous alteplase (rt-PA) versus control in patients with acute ischemic stroke ($n = 3{,}035$)[28]. The dataset includes 7 continuous, 4 binary, and 1 categorical covariate (stroke type), with 6-month survival as the binary outcome.

CRASH-2. This trial tested tranexamic acid (TXA) against placebo in bleeding trauma patients ($n = 20{,}211$)[29]. The dataset includes 7 continuous, 1 binary, and 1 categorical (injury type) variables, with in-hospital survival defined as the binary outcome.

SPRINT. SPRINT[30]($n = 9{,}361$) compared intensive blood pressure control with a standard target . The dataset contains 12 continuous and 7 binary covariates, with the primary binary outcome defined as freedom from composite cardiovascular events.

ACCORD. ACCORD[31] evaluated intensive versus standard systolic blood pressure targets in patients with type 2 diabetes ($n = 10{,}251$). It includes 18 continuous and 7 binary covariates, with a protocol-defined primary cardiovascular composite outcome.

In particular, SPRINT and ACCORD enrolled overlapping patient populations but reported contrasting outcomes for intensive blood pressure control, offering a natural case study for whether interpretability methods capture clinically meaningful differences.

To complement these RCTs, we incorporated a retrospective cohort to assess performance in a real-world clinical setting. This retrospective cohort comes from the Harborview pre-hospital TXA registry (2007-2020) Section A.3, serving as a reference population for CRASH-2. We applied propensity-score matching[46] to construct a balanced sample of 240 patients (120 TXA, 120 controls) from trauma admissions. It contains 6 continuous and 1 categorical variable, with survival as the outcome. Detailed variable definitions of the above datasets are provided in Section A.

CATE Models

Under the potential–outcomes framework[4], the CATE for a binary treatment $T \in \{0, 1\}$ and covariates X is

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]\,, \tag{1}$$

where $Y_i(1)$ and $Y_i(0)$ denote individual i's potential outcomes under treatment and control. Because only one potential outcome is observed for each individual, $\tau(x)$ must be inferred from data. Several works have proposed neural network-based *meta-learners*[2,14] that estimate $\tau(\cdot)$ by regressing learner-specific pseudo-outcomes:

$$\hat{\tau} = \arg\min_{\tau}\ \mathbb{E}\Big[\big\{\tau(X) - \hat{Y}_{\tilde{\eta}}(X, Y, T)\big\}^2\Big], \tag{2}$$

where $\hat{Y}_{\tilde{\eta}}$ is constructed from nuisance functions $\tilde{\eta}$. The specific form of $\hat{Y}_{\tilde{\eta}}$ depends on the learner. We considered a broad family of learners, including S-, T-, X-, R-, RA-, and DR-learners, as well as TARNet, DragonNet, and CFRNet, and self-interpretable approaches, such as Causal Forests[18] and linear DR-Learners[8] (Section B.1). We implemented CATE estimators using the `CATENets` package[3]

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

Model training. Data were partitioned into training (70%), validation (20%), and test (10%) sets, with stratification by treatment (and outcome when binary) to preserve balance. Within the training set, we further applied stratified 5-fold cross-validation to obtain out-of-fold (OOF) predictions for model selection. Models were trained with early stopping, monitoring validation loss with a fixed patience parameter. Hyperparameters (learning rate, batch size, number of layers, hidden units, etc.) were tuned via grid search (Section B.2). For ensemble models, we trained 40 independent replicates with different random seeds and averaged predictions.

Evaluation. Because the true CATE $\tau^*(x)$ is unknown, we evaluate models using *pseudo–outcome surrogates*[47,48] (Section B.3.1) that approximate the prediction error, as well as Qini and uplift curves[49] (Section B.3.1), which measure treatment effect gains based on the model's estimates. Model selection uses 5-fold cross-validation, and the best-performing learner type is then designated as the *explainee* for subsequent interpretability analyses.

LIFT-XAI: Explaining Ensemble CATE with Shapley Values

Reliable interpretability of CATE models is essential in clinical contexts, where explanations can yield patient-level insights in clinical decision-making. However, explanations are often based on a single model, which may not adequately approximate the underlying treatment effect[50]. Prior works have also found that explanations from a single model are often unstable due to stochastic training factors such as random initialization and data shuffling[26], limiting their reliability as indicators of the true signal.

To address this, we propose explaining *ensemble models*[51]. Concretely, we train N instances $\{\tau_i\}_{i=1}^N$ with distinct random seeds and data shuffling. Predictions and attributions are then averaged across models:

$$\phi_j^{\mathrm{LIFT}}(x) = \frac{1}{N}\sum_{i=1}^{N}\phi_j(\tau_i, x) = \phi_j(\tau_e, x) \quad \text{where} \quad \tau_e(x) = \frac{1}{N}\sum_{i=1}^{N}\tau_i(x) \tag{3}$$

This equality holds for attribution methods that satisfy the *linearity* axiom (e.g., Shapley values, IG) (Section C.1). By reducing variance from stochastic training, this framework yields feature attributions that more robustly capture true treatment-effect signals.

Choice of attribution method. While multiple attribution methods could be employed to ensemble models, we adopt Shapley values[24,25] because they uniquely satisfy the axioms of *efficiency*, *symmetry*, and *linearity*, providing a theoretically grounded attribution. Prior studies have also shown their promise in identifying relevant features for treatment effect in semi-synthetic environments[14,17]. Moreover, Shapley values extend naturally from individual-level to cohort-level explanations, aligning with both personalized and population-level clinical analyses. Formally, for features F and a model $\tau(x)$, the Shapley value of feature j for individual x is:

$$\phi_j(x) = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\big[\tau(x_{S\cup\{j\}}) - \tau(x_S)\big], \tag{4}$$

where S is a subset of features after excluding j. In practice, we can approximate Shapley values via sampling, using the empirical feature means as baselines to estimate marginal expectations[52]. This approximation directly addresses the clinically relevant question: *Which patient features most strongly drive deviations in treatment effect from what is expected for an average patient?* We therefore introduce **LIFT-XAI** as a principled framework for explaining ensemble CATE models using Shapley values, which we employ throughout this work.

Examining Interpretability Methods

Because ground-truth feature contributions to treatment effects are rarely observable, we evaluate explanation methods using indirect but complementary criteria: (i) *local fidelity*, which tests whether attributions reflect model behavior under perturbations; (ii) *global sufficiency*, which examines whether the most important features identified by the explanation are sufficient to recover model predictions; and (iii) *consistency*, which measures the stability of explanations across independently trained models. When external ground-truth signals are available, we additionally evaluate (iv) *feature rank concordance*.

Local fidelity. Local fidelity[53] measures how well feature attributions capture the model's prediction behavior at the individual level. We adopt standard insertion-deletion benchmarks[33] with the pseudo–outcome surrogate (Section B.3) as the fidelity metric. This assesses whether removing or inserting features with high attribution values leads to the expected changes in predicted treatment effects.

Global sufficiency. Global sufficiency assesses whether features identified by an explanation capture the model's predictive signal[54]. We rank features by their attribution values,[4] then train a student model $\hat{\tau}s$ using only the top-k features to reproduce the ensemble's predictions $\tilde{y}_i$.[5] The distillation error over M test points is

$$\mathcal{E}_{\text{KD}} = \frac{1}{M}\sum_{i=1}^{M}\big(\hat{\tau}_s(X_i^k) - \tilde{y}_i\big)^2, \qquad \hat{\tau}_s = \arg\min_{\theta}\ \mathcal{L}\big(\tau(X^k;\theta), \tilde{y}\big). \tag{5}$$

Lower $\mathcal{E}_{\text{KD}}$ indicates that the explanation-selected features capture the teacher's signal with fewer inputs.

Consistency. For explanations to be useful in scientific discovery, explanations must be stable across repeated training runs. We therefore assess consistency across L independently trained ensembles, $\tau_{e,l}^k$ (with $L = 1$ corresponding to a single model), each comprising k base learners. We compute the mean pairwise cosine similarity of their *global* attribution vectors (average absolute attribution across individuals):

$$\cos(\theta_k) = \frac{1}{L(L-1)}\sum_{l=1}^{L}\sum_{\substack{j=1\\ j\neq l}}^{L}\frac{\phi(\tau_{e,l}^k)\cdot\phi(\tau_{e,j}^k)}{\|\phi(\tau_{e,l}^k)\|_2\,\|\phi(\tau_{e,j}^k)\|_2}.$$

An increase in $\cos(\theta_k)$ with k indicates reduced sensitivity to random initialization and greater reflection of underlying signal rather than noise.

Feature rank concordance. When external ground-truth signals are available, we also assess rank concordance. Specifically, we compute *Spearman's rank correlation*[34] between the global feature ranking from an explanation method, $g(\hat{\tau}, \phi)$, and oracle rankings $g(p)$ derived from interaction p-values[35], where lower p-values denote stronger evidence of treatment effect heterogeneity. The resulting correlation

$$\rho(g(\hat{\tau}, \phi),\ g(p))\,, \tag{6}$$

indicates whether explanation methods recover features consistent with ground truth features.

Baseline Explanation Methods

We benchmark LIFT-XAI against widely used explanation methods at both the individual (local) and cohort (global) levels. At the individual level, we compare Shapley values to established attribution techniques, including Integrated Gradients (IG)[55], gradient-based Saliency maps[56], SmoothGrad[57], and LIME[58], which provide feature-level attributions for each patient. At the cohort level, we additionally benchmark treatment-effect–specific methods such as Leave-One-Covariate-Out (LOCO)[21] and PermuCATE[22], which yield global importance scores but do not produce individual-level explanations. Implementation details and hyperparameter choices are provided in Section C.2.

[4]For local attribution methods, global rankings are obtained by averaging absolute attributions across individuals.
[5]This procedure is conceptually similar to RemOve And Retrain (ROAR)[54], but evaluates explanations at the global level.

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 is available at `https://github.com/q8888620002/Explain_CATE`. 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 Trial Collaborators at the London School of Hygiene & Tropical Medicine Clinical Trials Unit (University of London) for sharing CRASH-2 Trial data, and to the researchers in the Lee lab for their valuable discussions.

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] Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66**, 688 (1974).

[5] 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).

[6] Johansson, F., Shalit, U. & Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029 (PMLR, 2016).

[7] Funk, M. J. *et al.* Doubly robust estimation of causal effects. *American journal of epidemiology* **173**, 761–767 (2011).

[8] Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497* (2020).

[9] Feuerriegel, S. *et al.* Causal machine learning for predicting treatment outcomes. *Nature Medicine* **30**, 958–968 (2024).

[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] Martínez, J. A. Interpretability for conditional average treatment effect estimation (2021).

[16] Wu, Y. *et al.* Discret: Synthesizing faithful explanations for treatment effect estimation. *Proceedings of machine learning research* **235**, 53597 (2024).

[17] Svensson, D., Hermansson, E., Nikolaou, N., Sechidis, K. & Lipkovich, I. Overview and practical recommendations on using shapley values for identifying predictive biomarkers via cate modeling. *arXiv preprint arXiv:2505.01145* (2025).

[18] Athey, S. & Wager, S. Estimating treatment effects with causal forests: An application. *Observational studies* **5**, 37–51 (2019).

[19] Chen, K., Yin, Q. & Long, Q. Covariate-balancing-aware interpretable deep learning models for treatment effect estimation. *arXiv preprint arXiv:2203.03185* (2022).

[20] Retzlaff, C. O. *et al.* Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research* **86**, 101243 (2024).

[21] Hines, O., Diaz-Ordaz, K. & Vansteelandt, S. Variable importance measures for heterogeneous causal effects. *arXiv preprint arXiv:2204.06030* (2022).

[22] Paillard, J., Lobo, A. R., Kolodyazhniy, V., Thirion, B. & Engemann, D. A. Measuring variable importance in heterogeneous treatment effects with confidence. *arXiv preprint arXiv:2408.13002* (2024).

[23] Lundberg, S. M. *et al.* From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence* **2**, 56–67 (2020).

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

[25] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

[26] Mehrer, J., Spoerer, C. J., Kriegeskorte, N. & Kietzmann, T. C. Individual differences among deep neural network models. *Nature communications* **11**, 5725 (2020).

[27] Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **23**, 18 (2020).

[28] Group, I.-. 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).

[29] 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).

[30] Group, S. R. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine* **373**, 2103–2116 (2015).

[31] Group, A. S. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *New England Journal of Medicine* **362**, 1575–1585 (2010).

[32] 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).

[33] Petsiuk, V., Das, A. & Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* (2018).

[34] Spearman, C. The proof and measurement of association between two things. (1961).

[35] 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).

[36] 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).

[37] de Havenon, A. *et al.* Effect of alteplase on ischemic stroke mortality is dependent on stroke severity. *Annals of Neurology* **93**, 1106–1116 (2023).

[38] 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).

[39] 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).

[40] Investigators, P.-T. & the ANZICS Clinical Trials Group. Prehospital tranexamic acid for severe trauma. *New England Journal of Medicine* (2023).

[41] Bastani, O., Kim, C. & Bastani, H. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504* (2017).

[42] Lin, C., Covert, I. & Lee, S.-I. On the robustness of removal-based feature attributions. *Advances in Neural Information Processing Systems* **36**, 79613–79666 (2023).

[43] Sun, Q., Xia, H. & Liu, J. Data-faithful feature attribution: Mitigating unobservable confounders via instrumental variables. *Advances in Neural Information Processing Systems* (2024).

[44] 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).

[45] Wu, Y., Liu, H., Ren, K. & Chang, X. Causal rule learning: Enhancing the understanding of heterogeneous treatment effect via weighted causal rules. *arXiv preprint arXiv:2310.06746* (2023).

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

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

[48] 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).

[49] Gutierrez, P. & Gérardy, J.-Y. Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and APIs*, 1–13 (PMLR, 2017).

[50] Chen, H., Janizek, J. D., Lundberg, S. & Lee, S.-I. True to the model or true to the data? *arXiv preprint arXiv:2006.16234* (2020).

[51] Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15 (Springer, 2000).

[52] Sundararajan, M. & Najmi, A. The many shapley values for model explanation. In *International conference on machine learning*, 9269–9278 (PMLR, 2020).

[53] Molnar, C. *Interpretable machine learning* (Lulu. com, 2020).

[54] 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).

[55] Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328 (PMLR, 2017).

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

[57] Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[58] Ribeiro, M. T., Singh, S. & Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144 (2016).

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

[60] Asuncion, A. & Newman, D. Uci machine learning repository (2007).

[61] 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).

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

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

[64] Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299–319 (2021).

[65] Shalit, U., Johansson, F. D. & Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, 3076–3085 (PMLR, 2017).

[66] Shi, C., Blei, D. & Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems* **32** (2019).

[67] 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).

[68] Schulam, P. & Saria, S. Reliable decision support using counterfactual models. *Advances in neural information processing systems* **30** (2017).

[69] Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **41**, 647–665 (2014).

[70] Chen, H., Lundberg, S. M. & Lee, S.-I. Explaining a series of models by propagating shapley values. *Nature communications* **13**, 4512 (2022).

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