Foundation Models for Heterogeneous Treatment Effect Estimation: Adapting Pretrained EHR Embeddings in Small Clinical Cohorts

Xuange Liang[0009-0008-7897-280X], Bin Yang[0000-0003-1320-7273], and Yuanjia Wang[0000-0002-1510-331]

Abstract Estimating heterogeneous treatment effects (HTE) from electronic health records (EHR) is critical for precision medicine, yet challenging with small clinical cohorts, confounding bias, and high-dimensional feature space. Pretrained EHR foundation models encode rich patient medical history into feature embeddings that capture patients' characteristics; however, whether these embeddings can be effectively utilized for HTE estimation remains unknown. We leverage embeddings generated by a large pretrained EHR foundation model, CLMBR, and apply the R-learner framework for HTE estimation. We develop an outcome-specific dimensionality reduction strategy that adapts to small sample sizes through principal component analysis (PCA) and strong regularization. We compare the foundation model embedding with five baseline methods. Evaluated on comparing GLP-1 receptor agonists versus insulin/sulfonylureas for type 2 diabetes patients across four clinical outcomes on Stanford University EHRs dataset, our approach achieves rank #1 on all four outcomes and averages 7.9% improvement over the best baseline per outcome. Our results demonstrate that pretrained EHR foundation model embedding can significantly enhance HTE estimation with limited data when properly adapted. Furthermore, we develop interpretable tools that connect clinical event codes to HTE predictions by mapping dimension-level importance to token-level importance, providing actionable insights for personalized treatment decisions.

Xuange Liang
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA e-mail: xl3493@cumc.columbia.edu

Bin Yang
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA e-mail: by2303@cumc.columbia.edu

Yuanjia Wang
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA e-mail: yw2016@cumc.columbia.edu

1 Introduction

Precision medicine requires understanding patient-specific treatment responses rather than relying on population-level average effects. Conditional average treatment effects (CATE) estimates account for heterogeneity among patients by estimating the effects of a treatment for a patient subgroup defined by covariates. Suppose we have access to n independent and identically distributed subjects (X_i, T_i, Y_i) for $i = 1, 2..., n$, where X_i denote a patients' feature variables before treatment (e.g., medical history), $T_i \in \{0, 1\}$ denotes the treatment indicator, and Y_i is the outcome of interest after treatment. Following the potential outcomes framework of Rubin [4], we posit $Y^{(1)}$ and $Y^{(0)}$ as potential outcomes a patient would have experienced given treatment $T = 1$ or $T = 0$, and define CATE as

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

There is an emerging statistical and machine learning literature that focuses on the estimation and application of complex form heterogeneous treatment effect (HTE) under the causal inference framework. Early approaches to accounting for confounding include propensity score stratification [**?**], matching methods [**?**], and inverse probability weighting [**?**]. However, these methods often rely on parametric assumptions and may not handle high-dimensional covariates.

Modern machine learning approaches have significantly advanced HTE estimation by leveraging flexible nonparametric methods. Meta-learners, a class of model-agnostic frameworks, transform the CATE estimation problem into standard supervised learning tasks. The T-learner [**?**] trains separate outcome models for treated and control groups, while the X-learner [**?**] uses cross-estimation to improve efficiency. The S-learner [**?**] treats treatment as a feature in a single model. However, these methods can suffer from regularization bias when the treatment effect is small relative to the main effect [**?**].

The R-learner [2] addresses these limitations through the Robinson transformation [3], which orthogonalizes the treatment effect from nuisance components (propensity score and outcome models). This approach achieves desirable error bounds comparable to oracle methods and provides doubly-robust estimation, meaning the estimator is consistent if either the propensity score or outcome model is correctly specified. The R-learner has been extended to various settings, including the optimal R-learner with complete cross-fitting [1], which we adopt in this work.

Alternative approaches to HTE estimation include causal forests [**?**], which adapt random forests for treatment effect estimation, and neural network-based methods [**?**] that learn balanced representations to control for confounding. However, these methods often require large sample sizes and may not perform well with high-dimensional EHR data in small clinical cohorts.

Electronic health records (EHR) provide rich longitudinal patient histories for personalized treatment decisions. Recent advances in EHR foundation models [5] have shown promise in learning rich patient representations from large-scale pretraining. Benchmarks like EHRSHOT [7] and recent evaluations [6] demonstrate that pretrained embeddings can improve clinical prediction tasks. CLMBR (Clinical Language Model for Biomedical Records) [5] is a transformer-based foundation model pretrained on 2.5 million patient timelines via next-event prediction. For each patient, CLMBR pro-

cesses the sequential medical history (diagnoses, procedures, medications, lab values) up to the treatment date and outputs a 768-dimensional contextualized embedding from the final transformer layer. This embedding encodes rich temporal patterns and medical relationships learned during pretraining. However, adapting these high-dimensional representations (typically 768 dimensions) for estimating CATE and HTE presents unique challenges, as causal inference and HTE estimation require removing confounding rather than maximizing prediction accuracy. In addition, adapting a high-dimensional embedding directly to a limited sample size and achieving clinical interpretability requires careful regularization and dimensionality reduction strategies.

We perform a case study comparing HTE of GLP-1 receptor agonists versus insulin/sulfonylureas for type 2 diabetes (T2D) patients on four outcomes: hypertension, hyperlipidemia, systolic blood pressure, and heart rate (46-151 patients received treatments and had available outcomes).

Key contributions: (1) We demonstrate that pretrained EHR foundation models (CLMBR [5]) can be successfully adapted for HTE estimation through PCA reduction across all outcomes. (2) We implement on the R-learner framework with complete cross-fitting [1] and ultra-extreme regularization to prevent overfitting. (3) We achieve rank #1 on all four outcomes, outperforming all baseline methods, including Binary PCA variants and traditional count-based methods. (4) We introduce interpretability methods that combine attention-based temporal relevance with dimension-based treatment effect sensitivity, enabling clinicians to identify key drivers of treatment heterogeneity.

2 Methods

2.1 R-learner Framework

R-learner [2] addresses challenges in estimating CATE in an observational study by allowing fully automatic specification for heterogeneous treatment effect estimators while achieving desirable error bounds comparable to oracle methods where the true data generation mechanism is known.

We adopt the efficient local polynomial R-learner variant by [1], and specifically, we estimate CATE via a two-stage procedure that constructs pseudo-outcomes using the Robinson transformation [3]. This procedure explicitly accounts for confounding by orthogonalizing the target parameter from nuisance components and achieves doubly-robustness, which means the estimation is consistent if either the propensity score model $e(X) = \mathbb{P}(T = 1|X)$ or outcome model $m(X) = \mathbb{E}[Y|X]$ is correctly specified. Specifically, the R-learner minimizes the following objective:

$$\hat{\tau} = \arg\min_{\tau} \mathbb{E}\left[w(X) \cdot \left(\tilde{R}(X) - \tau(X)\right)^2\right], \tag{2}$$

where $\tilde{R}(X) = (Y - m(X)) \cdot (T - e(X))$ is the Robinson-transformed pseudo-outcome, and $w(X) = 1/[e(X)(1 - e(X))]$ is the inverse variance weight.

Complete cross-fitting procedure: We use R-learner with complete cross-fitting [1] for doubly-robust CATE estimation. For each of K cross-validation folds, we compute the following steps on disjoint training and validation sets:

1. **Propensity score estimation:** Fit logistic regression with L_2 penalty on training data:

$$\hat{e}(X) = \sigma(\hat{\beta}^T X), \quad \text{where } \hat{\beta} = \arg\min_{\beta} \sum_i \log(1 + e^{-y_i \beta^T X_i}) + \tfrac{1}{2C}\|\beta\|^2.$$

2. **Outcome model:** Fit Ridge regression with cross-fitting:

$$\hat{m}(X) = \hat{\theta}^T X, \quad \text{where } \hat{\theta} = \arg\min_{\theta} \sum_i (Y_i - \theta^T X_i)^2 + \alpha_m \|\theta\|^2.$$

3. **Pseudo-outcomes:** Construct Robinson-transformed targets with inverse variance weights:

$$\tilde{R}_i = (Y_i - \hat{m}(X_i)) \cdot (T_i - \hat{e}(X_i)), \quad w_i = \frac{1}{\hat{e}(X_i)(1 - \hat{e}(X_i)) + \epsilon},$$

   where $\epsilon = 10^{-4}$ prevents division by zero.
4. **CATE estimation:** Fit weighted Ridge regression on pseudo-outcomes:

$$\hat{\tau}(X) = \hat{\gamma}^T X, \quad \text{where } \hat{\gamma} = \arg\min_{\gamma} \sum_i w_i \cdot (\tilde{R}_i - \gamma^T X_i)^2 + \alpha_\tau \|\gamma\|^2.$$

This three-layer cross-fitting procedure ensures Neyman orthogonalization: $e(X)$, $m(X)$, and $\tau(X)$ are estimated on disjoint data subsets, which reduces overfitting bias and maintains the doubly-robust property in finite samples.

2.2 Dimensionality Reduction and Regularization

A large pretrained model includes rich information about the patient's medical history. We utilize the high-dimensional embeddings from the pretrained model for CATE estimation. For each patient i, we generate a high-dimensional embedding $X_i \in \mathbb{R}^d$ with $d = 768$ from the EHR foundation model CLMBR [5], and we compare the utility of this embedding with count-based features ($d = 1000 - 5000$), and binary PCA ($d = 1000$). Due to the high dimensionality of the feature embedding, there is a severe risk of overfitting with a limited sample size ($46 - 151$ patients). We select the reduced PCA dimension via hyperparameter search: for each candidate d_{PCA} and for each feature family, we run 5,000 Latin hypercube sampling (LHS) trials; we then choose d_{PCA} based on the average of the best validation R-loss across trials for that dimension, yielding $d_{\text{PCA}} = 14$ for binary outcomes and $d_{\text{PCA}} \in \{32, 64\}$ for continuous outcomes.
Implementation: (1) Apply standardization: $X' = (X - \mu)/\sigma$. (2) Fit PCA on training data only within each CV fold to prevent leakage. (3) Verify 85-96% variance retention.

Ultra-extreme regularization: Small samples require aggressive regularization beyond standard practices. For a specific dimension, we select the regularization hyperparameters for each feature via LHS by minimizing the empirical R-loss (2) via 3-fold cross-validation. For CLMBR on binary outcomes (optimal $d_{\text{PCA}} = 14$): $C = 10^{-7}$ (propensity), $\alpha_m = 2000$ (outcome), $\alpha_\tau = 2500$ (treatment). For CLMBR on continuous outcomes (optimal $d_{\text{PCA}} = 32$ or 64): $C = 5 \times 10^{-5}$ (propensity), $\alpha_m = 500$ (outcome), $\alpha_\tau = 600$ (treatment). Binary PCA methods also use ultra-strong regularization: $C = 10^{-7}$, $\alpha_m = 500$, $\alpha_\tau = 1000$ for binary outcomes.

3 Experimental Setup

3.1 Dataset and Cohorts

We use Stanford Medicine EHR data following the EHRSHOT benchmark framework [7] with 1,331 T2D patients, of whom 81 (6.6%) received GLP-1 receptor agonists and 71 (5.3%) received insulin or sulfonylureas. We construct treatment cohorts (GLP-1 receptor agonists vs insulin/sulfonylureas) using 180-day baseline windows for four clinical outcomes after treatment: hypertension ($n = 46$), hyperlipidemia ($n = 55$), systolic BP ($n = 151$), and heart rate ($n = 151$).

3.2 Baselines Methods and Evaluation

Baselines: To ensure fair comparison, all methods apply dimensionality reduction to match the target dimension. We evaluate: (1) CLMBR (pretrained embeddings $d = 768$ with PCA), (2) Binary-Bernoulli (binary event codes ICD10/SNOMED etc. $d = 1000$ with Bernoulli PCA), (3) Binary-Logistic (binary codes $d = 1000$ with Logistic PCA), (4) Counts-SVD (count codes $d = 5000$ with TruncatedSVD), and (5) Counts-PCA (count codes $d = 1000$ with PCA).
Evaluation: Binary outcomes use 5-fold CV with stratified splitting; continuous use 3-fold CV. We use the empirical evaluation of the R-loss (2) as our primary metric, which measures the mean squared residuals after accounting for prognostic effect and treatment main effect. Lower R-loss indicates more accurate CATE estimation. We use standardized features and outcome scaling for numerical stability. All experiments are performed using a random seed of 42.

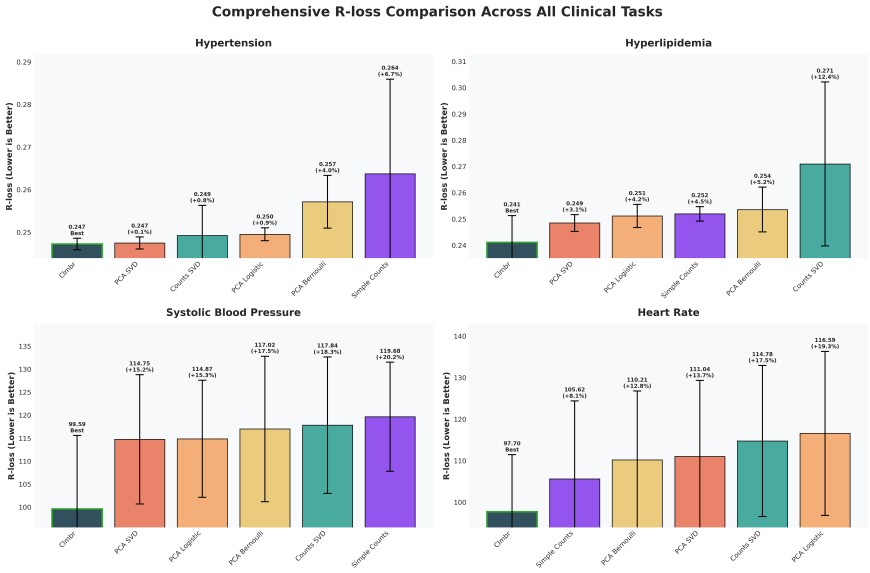


Fig. 1 Comprehensive cross-outcome performance comparison, where a smaller R-loss indicates better performance. CLMBR achieves rank #1 on all four outcomes with consistent performance. Count-based methods show high variance and occasional catastrophic failures.

4 Results

4.1 Main Results

Figure 1 presents R-loss across all methods and outcomes. CLMBR with outcome-specific dimensionality reduction achieves the lowest R-loss on all four clinical outcomes, demonstrating consistent superiority over baseline methods.

Performance summary. CLMBR achieves 100% success rate (rank #1 on 4/4 outcomes) with improvements of 4.3%, 4.6%, 10.8%, and 12.0% over the best baseline for each outcome (average 7.9%). For continuous outcomes (systolic BP and heart rate), CLMBR shows substantial improvements (10.8-12.0%) over all baselines. For binary outcomes (hypertension and hyperlipidemia), CLMBR maintains consistent superiority with 4-5% improvements.

4.2 Ablation: Effect of Model Architecture and Dimension Reduction

Figure 2 shows how model architecture and dimension reduction affect performance for systolic blood pressure (151 patients). The foundation model (transformer-based), “CLMBR Best” (parameters 141M, context length 512), achieves the best performance compared to the generic “GPT-2 based CLMBR” and “DeepSeek-V3 based CLMBR”,

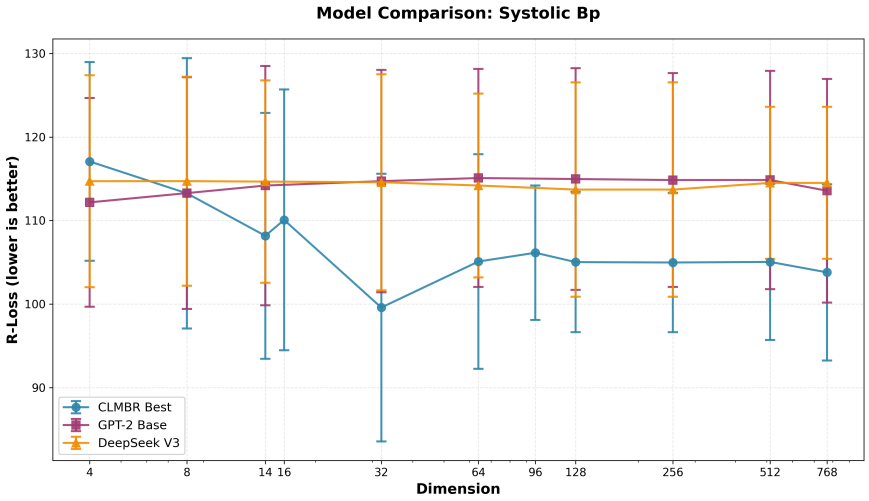


Fig. 2 Effect of architecture and dimensionality on R-loss for systolic blood pressure (151 patients). "CLMBR Best" (transformer-based foundation model pretrained on the full EHRSHOT dataset) consistently outperforms generic GPT-2 and DeepSeek-V3 across most dimensionalities, with optimal performance at $d_{\text{PCA}} = 32$. The curves show R-loss changes with the PCA dimension, revealing the importance of both model architecture and compression level.

both using 117M parameters and 256 as context length. The optimal performance for CLMBR is at $d_{\text{PCA}} = 32$ (R-loss = 99.6), with performance degrading at very low dimensions ($d_{\text{PCA}} = 4$, R-loss = 117.0) due to information loss and no further improvement with a larger dimension, even at $d_{\text{PCA}} = 768$ (R-loss = 103.8). This demonstrates the importance of maintaining model size and longer context length, and finding the right balance between retaining information and preventing overfitting through regularization is crucial for small-sample HTE estimation.

4.3 Interpretability: Connecting Event Codes to CATE Predictions

To address the black-box nature of CLMBR embeddings, we developed interpretability tools that directly connect clinical event codes to CATE predictions by mapping dimension-level importance to token-level importance. Our approach bridges the 768-dimensional embedding space with clinically actionable insights through attention-weighted CATE relevance scoring.

CATE-based event importance. We calculate token-level importance scores using:

$$s_i = \alpha_i \times (h_i \cdot w), \tag{3}$$

where h_i is the hidden state for token i (768-dim vector), w is the dimension importance weight derived from HTE feature analysis (768-dim vector indicating which embed-

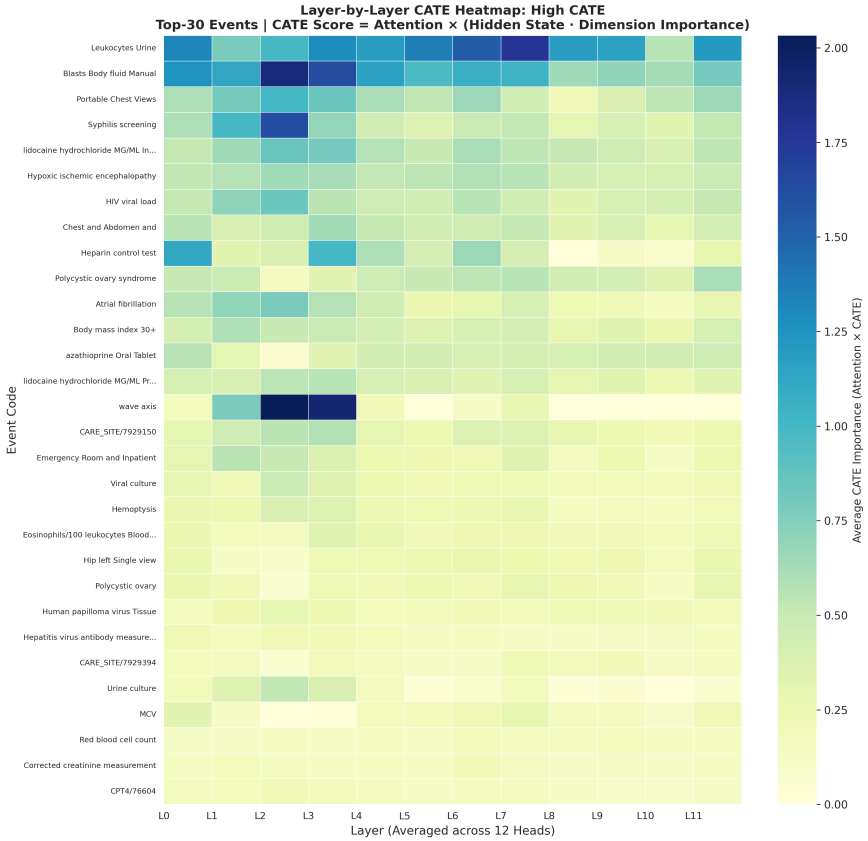


Fig. 3 Attention weight heatmap for high CATE (CATE > Median) systolic blood pressure patients aggregated across all patients. Darker colors indicate higher CATE importance.

ding dimensions contribute most to CATE prediction), and α_i is the attention weight from CLMBR's final layer. This formulation combines two complementary signals: (1) *model-level importance* (α_i) capturing which events the model attends to in the feature space when predicting EHR events, and (2) *task-level relevance* ($h_i \cdot w$) identifying how each event activates CATE-relevant embedding dimensions. The dimension importance weights w are obtained from R-learner's treatment effect model coefficients after PCA projection, providing a direct link between the 768D embedding space and heterogeneous treatment effects.

Figure 3 shows the aggregated attention weight distribution across CLMBR's transformer layers for patients with high conditional treatment effects. The model exhibits progressive refinement through layers, with later layers showing stronger focus on cardiovascular risk factors and medication history that drive treatment effect predictions.

Temporal attention layer pattern. To understand how the model's attention varies across both time and transformer layers for different patient subgroups, we generate layer × time heatmaps by aggregating attention weights into temporal bins. For each of

Table 1 Sample size robustness analysis for heart rate outcome across 30 independent runs per sample fraction.

Sample Size	R-loss Mean	Std Dev	CV (%)
50% (70)	127.47	25.74	20.2
60% (84)	121.38	20.84	17.2
70% (98)	118.24	17.48	14.8
80% (112)	118.62	13.40	11.3
90% (126)	106.34	2.58	2.4
100% (141)	**97.70**	**3.23**	**3.3**

the 15 High CATE and 15 Low CATE hypertension patients, we divide the 180-day pre-treatment window into 60 time bins (3-day intervals), then aggregate attention weights received by events in each bin across all patients and all attention heads within each layer. The resulting heatmap in Figure 4 shows average attention per patient for each (layer, time bin) combination, revealing how different CATE groups exhibit distinct temporal attention patterns.). High CATE patients show attention on mid-range historical events, while Low CATE patients focus on more recent events.

Figure 5 demonstrates how our CATE-based scoring directly links individual clinical codes to treatment effect predictions. For this patient with a large positive CATE (5.35 mmHg higher with GLP-1 vs insulin/sulfonylureas), the model identifies the dynamic nature of CATE estimation, where recent clinical measures have a larger treatment effect. Unlike pure attention-based approaches that only show what the model attends to in the feature space, our method reveals which events matter for CATE prediction through their activation of treatment-effect-relevant embedding dimensions. This provides clinicians with interpretable, event-level explanations for predicted treatment heterogeneity. These interpretable scores thus transform CLMBR from a black-box predictor into a more transparent clinical decision support tool.

4.4 Sample Size Robustness Analysis

To assess the reliability of our HTE estimation across varying sample sizes, we conduct a sample size analysis on the heart rate outcome (141 patients). We evaluate model performance across 10 sample fractions (10%–100%) with 30 independent random runs per fraction, using unique random seeds for each (sample fraction, run, CV fold) combination to ensure statistical independence. Table 1 presents R-loss values (lower is better) across different sample sizes. R-loss and variance decrease substantially (CV: 20.2% to 3.3%) as sample size increases, suggesting foundation model embeddings would be more effective in larger studies.

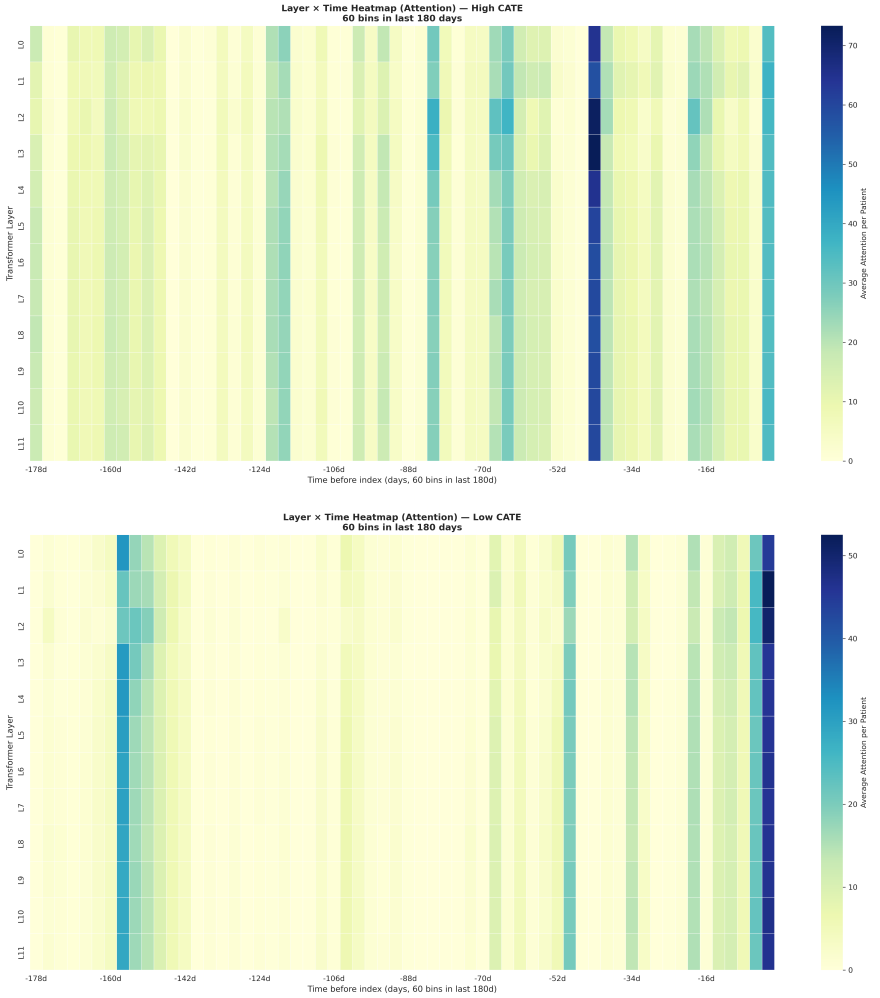
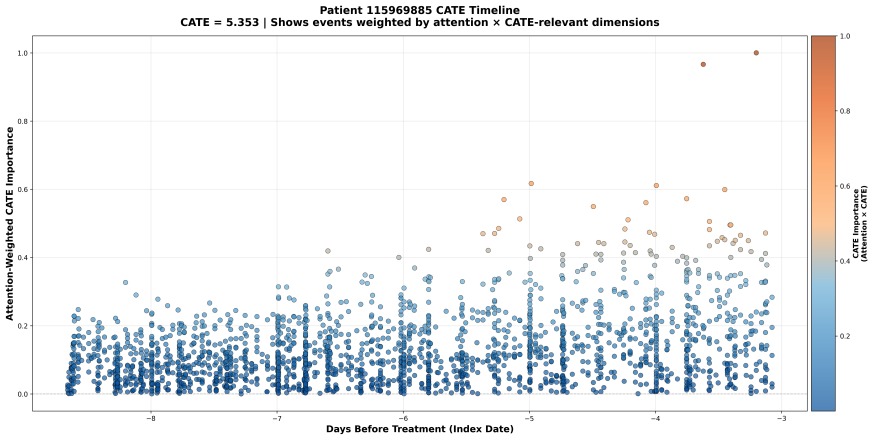


Fig. 4 Layer × time attention heatmaps for hypertension patients comparing High CATE (top) and Low CATE (bottom) groups. Each heatmap aggregates attention patterns from 15 patients over a 180-day pre-treatment window with 60 time bins. The X-axis shows time relative to treatment date (negative days indicate days before treatment), Y-axis shows transformer layers (L0–L11), and color intensity represents average attention per patient. High CATE patients show concentrated attention on mid-range historical events (-43d to -82d), suggesting the model leverages longitudinal disease progression patterns. Low CATE patients show stronger focus on recent events (closer to 0d), indicating reliance on current clinical state. These distinct temporal attention patterns reflect differences in baseline patient characteristics between treatment response groups.

5 Discussion

Foundation models enable small-sample HTE estimation. Our results demonstrate that large-scale pretrained EHR foundation models are effective for estimating HTE in



Rank	Event Code	Description	Days Relative	Importance
1	LOINC/8480-6	Systolic blood pressure (#1)	-3.2	1.0000
2	SNOMED/61928009	Platelet count	-3.6	0.9663
3	SNOMED/115267000	Specimen description	-5.0	0.6170
4	LOINC/8480-6	Systolic blood pressure (#2)	-3.5	0.5991
5	LOINC/789-8	Erythrocytes [#/volume] in Blood by Automated count	-3.8	0.5723
6	LOINC/LP21258-6	Oxygen saturation (#1)	-4.1	0.5607
7	LOINC/LP21258-6	Oxygen saturation (#2)	-4.5	0.5494
8	LOINC/LP21258-6	Oxygen saturation (#3)	-5.1	0.5132
9	LOINC/LP173418-7	Note	-4.2	0.5106
10	LOINC/2947-0	Sodium [Moles/volume] in Blood	-3.6	0.5057
11	LOINC/19254-2	Oxygen [Partial pressure] adjusted to patient's actual tempe	-3.4	0.4955
12	LOINC/8867-4	Heart rate	-3.4	0.4953

Fig. 5 CATE-based event importance timeline for a high-CATE systolic blood pressure patient (CATE = +5.35 mmHg). The scatter plot shows attention-weighted CATE importance scores in equation (3) for each clinical event over time, where events closer to the treatment date contribute more to predicting the HTE. The table lists the top 12 most important events with their temporal positions and importance scores, revealing which specific medical history factors drove the model's HTE prediction for this patient.

cohorts with limited sample sizes as small as 46–151 patients. CLMBR captures rich clinical information through pretraining on 2.5 million patients, generating a patient embedding that remains highly predictive even after dimensionality reduction by up to 50-fold.

Interpretability through dimension-attention fusion. To derive clinical insight from the pretrained black-box foundation model, our framework provides a token-level importance score in equation (3) by computing the attention weights from the foundation model and leveraging the learned coefficients for HTE prediction, thereby linking clinical code with HTE estimates. Our framework establishes a direct mapping from embeddings to CATE estimates that links the patient feature patterns, such as metabolic markers and medication histories, to treatment decisions. As a limitation, we note that the importance scores should be interpreted as model attributions rather than causal explanations of HTE. In addition, the codes that have high importance scores may reflect site-specific documentation patterns or care pathways, not patient biology. The interpretations of these scores, therefore, should be corroborated through sensitivity analysis and external validation in independent cohorts before being used to draw clinical or mechanistic conclusions.

Generalizability limitations. The small cohort sizes (46–151 patients per outcome) from a single institution limit generalizability. While our approach utilizes longitudinal EHR history, the data is irregularly sampled and lacks the high-frequency granularity of continuous physiological monitoring. However, our sample size robustness analysis (Table 1) reveals R-loss and variance decrease with increasing sample size. This scaling pattern suggests that foundation model embeddings would be more effective and reliable in larger cohorts, motivating multi-institutional validation studies across diverse treatment classes and patient populations before clinical deployment. Additionally, CLMBR embeddings may inherit biases from the pretraining data, and future work should validate performance across diverse datasets to assess generalizability and potential biases.

Our work bridges pretrained foundation models and precision medicine, demonstrating that large-scale pretraining yields robust representations that remain informative even after aggressive dimensionality reduction for small sample settings. For future studies, these findings should be validated in larger cohorts from multiple institutions and diverse treatment classes.

References

[1] Kennedy, E.H.: Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics **17**(2), 3008–3049 (2023)

[2] Nie, X., Wager, S.: Quasi-oracle estimation of heterogeneous treatment effects. Biometrika **108**(2), 299–319 (2021)

[3] Robinson, P.M.: Root-n-consistent semiparametric regression. Econometrica **56**(4), 931–954 (1988)

[4] Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**(5), 688–701 (1974)

[5] Steinberg, E., Fleming, S.L., Xu, Y., Fries, J.A., Chen, B.S., Chen, C., Shao, Y., Golbus, J.R., Cipriano, L.E., Pageler, N., et al.: Language models are an effective representation learning technique for electronic health record data. Journal of Biomedical Informatics **113**, 103637 (2021)

[6] Wornow, M., Thapa, R., Steinberg, E., Fries, J.A., Callahan, A., Patel, B., Fleming, S., Lemmon, M.A., Lehman, E., Shah, N.H.: Context clues: Evaluating long context models for clinical prediction tasks on ehrs. arXiv preprint arXiv:2412.16178 (2024)

[7] Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Lemmon, M.A., Lehman, E., Kalpathy-Cramer, J., Fries, J.A., et al.: Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)