Received: 18 May 2022 | Revised: 15 September 2022 | Accepted: 18 September 2022

DOI: 10.1002/psp4.12871

ARTICLE

Applying interpretable machine learning workflow to evaluate exposure–response relationships for large-molecule oncology drugs

Gengbo Liu | James Lu | Hong Seo Lim | Jin Yan Jin | Dan Lu

Department of Clinical Pharmacology, Genentech, South San Francisco, California, USA

Correspondence
Dan Lu, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080, USA.
Email: lu.dan@gene.com

Abstract

The application of logistic regression (LR) and Cox Proportional Hazard (CoxPH) models are well-established for evaluating exposure–response (E–R) relationship in large molecule oncology drugs. However, applying machine learning (ML) models on evaluating E–R relationships has not been widely explored. We developed a workflow to train regularized LR/CoxPH and tree-based XGboost (XGB) models, and derive the odds ratios for best overall response and hazard ratios for overall survival, across exposure quantiles to evaluate the E–R relationship using clinical trial datasets. The E–R conclusions between LR/CoxPH and XGB models are overall consistent, and largely aligned with historical pharmacometric analyses findings. Overall, applying this interpretable ML workflow provides a promising alternative method to assess E–R relationships for impacting key dosing decisions in drug development.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Currently, novel machine learning (ML) models focus on predictions but have not been widely applied to evaluate exposure–response (E–R) relationships yet.

WHAT QUESTION DID THIS STUDY ADDRESS?

We aim to apply an interpretable ML workflow for E–R analysis on real clinical trial datasets from large molecule oncology drugs.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

This study demonstrated that our interpretable ML workflow provides a plausible alternative method for odds ratio and hazard ratio calculations and E–R relationship evaluations. The ML approaches are robust for the analysis, including all available covariates without rigorous covariate selection, and we recommend including the control arm when available to better inform the ML models.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

The interpretable ML workflow with regularization provides a robust approach adding to our traditional pharmacometrics toolbox for E–R analysis and can be applied to impact key dose decisions in drug development.

INTRODUCTION

Exposure–response (E–R) analysis is an indispensable part in clinical drug research and development regulatory decision making over the last decades.[1] According to the US Food and Drug Administration (FDA) issued Guidance for Industry on E–R Relationship,[2] E–R analysis aims at examining the relationships between drug exposure variables, such as plasma concentration and drug efficacy/safety. The main purpose of E–R analysis is to support and justify dose selection at each phase of new drug clinical research and development.[3] Thus, the key question to be addressed by the E–R analysis is: does higher dose and consequently the higher pharmacokinetic (PK) exposure lead to higher efficacy and/or higher safety risk? However, E–R analysis is often based on the efficacy and safety data at a single dose level for large molecule oncology drugs (e.g., monoclonal antibodies [mAbs]), which can be confounded by some baseline covariates that impact both efficacy and PK exposure, leading to false-positive E–R relationship, when some of these confounders are left out from the traditional pharmacometric approaches.[4,5] In addition, the E–R relationship can be nonlinear given the potential saturation of pharmacological effects at high dose levels of mAbs. Although a dataset from a single-dose level may pose challenges in accurately recovering the E–R ground truth, it is worth exploring novel analysis approaches to overcome the methodological limitations of the traditional statistical approaches, such as logistic regression (LR) and Cox proportional hazard (CoxPH) models. These triggered the research ideas of applying novel machine learning (ML) approaches as an alternative method to address confounders by including a more comprehensive list of covariates and accounting for nonlinear relationship between covariates and hazard rate as well as the complex interactions among covariates. There were publications in this area using simulated data,[6] and we are aiming for exploring the ML methods for E–R analysis in real clinical trial datasets here.

Multivariable LR, CoxPH models, and case-matching analysis are well-established as standard pharmacometric approaches for E–R analysis in oncology trials.[7–9] The LR models are typically used to estimate the binary/categorical end point, such as the best objective response (BOR). The analysis with time-to-event end points, such as time to overall survival (OS), are done by CoxPH models. The traditional multivariable LR/CoxPH approach does not apply regularization and to avoid multicollinearity, only the most influential covariates are selected by forward addition and backward elimination. The linear and tree-based ML methods contain regularization terms that allow including many more covariates without the need of rigorous covariate selection and the issue of multicollinearity,[10] and potentially reduce the risk of missing important confounders and better recover the E–R ground truth. Further, the traditional LR and CoxPH models for E–R analysis assume linear or log-linear relationship between exposure, other baseline covariates and the efficacy end point. The tree-based ML models, such as eXtreme Gradient Boosting method (XGB),[11] can potentially be flexible to model linear, log-linear, or other nonlinear covariate-target relationships, and the complex interactions. It was reported that the explainable ML approaches, such as XGB in combination with Shapley Additive Explanation (SHAP) values outperformed the traditional CoxPH regression predictions and provided insights on survival data.[12]

The key objectives for this paper are to apply interpretable ML workflow for E–R analysis based on the efficacy end points in real oncology clinical trials for large molecule drugs, which includes the binary end point (e.g., BOR) and time-to-event end point (e.g., OS), derive the covariate-adjusted odds ratios (ORs) and hazard ratios (HRs) from ML models to evaluate the E–R relationship. We compared the results among both linear and tree-based ML methods with regularization, and with the traditional pharmacometric findings from historical analysis.

METHODS

Data sources

Data from four randomized, open-label, phase III oncology trials in patients with HER2-positive cancer were included to evaluate E–R relationship for large molecules by our ML approaches in the analysis: EMILIA,[13] TH3RESA,[14] and MARIANNE[15] trials for ado-trastuzumab emtansine (T-DM1) in patients with locally advanced or metastatic breast cancer, HELOISE[16] trial for Herceptin in gastric cancer (Tables S1–S10). For EMILIA and TH3RESA trials, the favorable risk–benefit profile of T-DM1 led to the FDA approval, and historical analysis showed positive E–R relationships.[7,8] The disease status is relatively similar among EMILIA and TH3RESA studies, and all patients in the T-DM1 arms received single-agent treatment at 3.6 mg/kg every-3-week (Q3W) regimen. Thus, we combined EMILIA and TH3RESA (referred to as EandT) studies for ML analysis. MARIANNE is a trial designed to evaluate the safety and efficacy of T-DM1 single-agent and T-DM1 + pertuzumab for the treatment of patients with HER2-positive, progressive, or recurrent locally advanced or metastatic breast cancer who had not received prior chemotherapy for their metastatic disease. All patients in the T-DM1 arms received 3.6 mg/kg Q3W regimen. In the HELOISE trial for patients with HER2 positive advanced gastric or gastro-esophageal junction cancer,[16] participants started with 8 mg/kg loading dose of Herceptin at first cycle followed by standard-of-care (6 mg/kg) or higher-dose (10 mg/kg) Q3W regimens, and the results showed a lack of clinically meaningful E–R relationship, as the higher-dose regimen does not increase OS.

The ML prediction target is the binary end point of BOR for the EandT study and the MARIANNE study, and time-to-event end point of OS for the EandT study and the HELOISE study: two classifications and two regressions on censor data tasks (Tables S1–S10). All available covariates from these four trials are applied in our ML based E–R analyses (Tables S2–S4).

Primary machine learning workflow for E–R analysis

We applied both linear and tree-based ML models: regularized LR with elastic net (L1 and L2) regularization,[17] CoxPH model with elastic net penalty (COX-NET),[18] XGB (Objective = binary:logistic) and XGB (Objective = survi val:cox),[11,19] to assess the E–R relationship between the trough concentration (C_{trough} or C_{min}) of the therapeutic agent and BOR or OS.

The analyses were performed in four steps (Figure 1a) with technical details in Figure 1b. All analyses were performed in Python 3.7 and R 4.1.0. Bayesian-based method is used for hyperparameter tuning[20] for each method, by repeated five-fold cross validation (CV) for 100 times. The median value of each hyperparameter is used for the final model. The model performance on the 20% test dataset generated by resampling without replacement was evaluated by the area under receiver operating characteristic curve (ROC-AUC) and concordance index (C-index) for classification and survival models. Details of step 1 data preprocessing and step 2 ML model building and hyperparameter tuning are provided in the Appendix S1.

In step 3, we interpret the ML model by covariate importance assessment, including the exposure covariates (e.g., C_{trough}). With the final optimized hyperparameters, we refitted the whole data on each model. To rank-order, each covariate's importance in predicting outcome, covariates coefficient, or the model-agnostic SHAP analysis were applied on explaining LR/COX-NET model or XGB model, respectively. For LR and COX-NET models, the absolute value of coefficients for each covariate was ranked for their importance. For XGB models, SHAP values measure the impact of covariates taking into account the interaction with other covariates and calculates the covariate importance by comparing what XGB model predicts with and without this covariate,[21] and the covariate importance was determined by averaging the absolute SHAP values across all patients for each covariate.

In step 4, we conducted model evaluation of E–R relationship by ORs and HRs related to C_{trough}. For the LR and XGB models with the binary BOR end point, the following values were derived: unadjusted OR between groups A and B based on unadjusted odds (formula 1a for LR and 1b for XGB), which is based on the model prediction in the absence of conditioning on other covariates[22] that some of them may be the confounders of an E–R relationship; and adjusted OR (formula 2a for LR and 2b for XGB), which is estimated as the ratio of the means of adjusted odds between group A and group B, holding the effect of all other covariates constant except the effect of C_{trough}. Similarly, the unadjusted and adjusted HR can be computed by formulas 3 and 4, respectively. It was found that among various methods to aggregate individual probabilities, both geometric mean and arithmetic mean can be used,[23,24] and here we consistently used geometric mean for the aggregation of odds or hazard rate for each group (i.e., each exposure quantile), for computing ORs and HRs.

The primary analyses of this study built the ML models using all patients, including the control arm (if available) and all available covariates. For evaluating the E–R relationship of T-DM1 in the MARIANNE study, two T-DM1 containing arms were pooled. When the control arm is included (EandT and MARIANNE), the patients in the T-DM1 treated arms were divided to four quantiles based on C_{trough} values (Q1, Q2, Q3, and Q4) and the OR or HR of each quantile to the control group were derived. In these cases, group A in the formulas refers to Q1, Q2, Q3, or Q4 and group B refers to the control group. When the control arm is not present (HELOISE study) or not used (sensitivity analysis, see section below), group A refers to Q2, Q3, or Q4 and group B refers to the Q1 group of patients with the lowest exposures.

We applied bootstrap for calculating confidence intervals (CIs) for ORs and HRs. Specifically, in each iteration of resampling, we train our ML model by resampling the

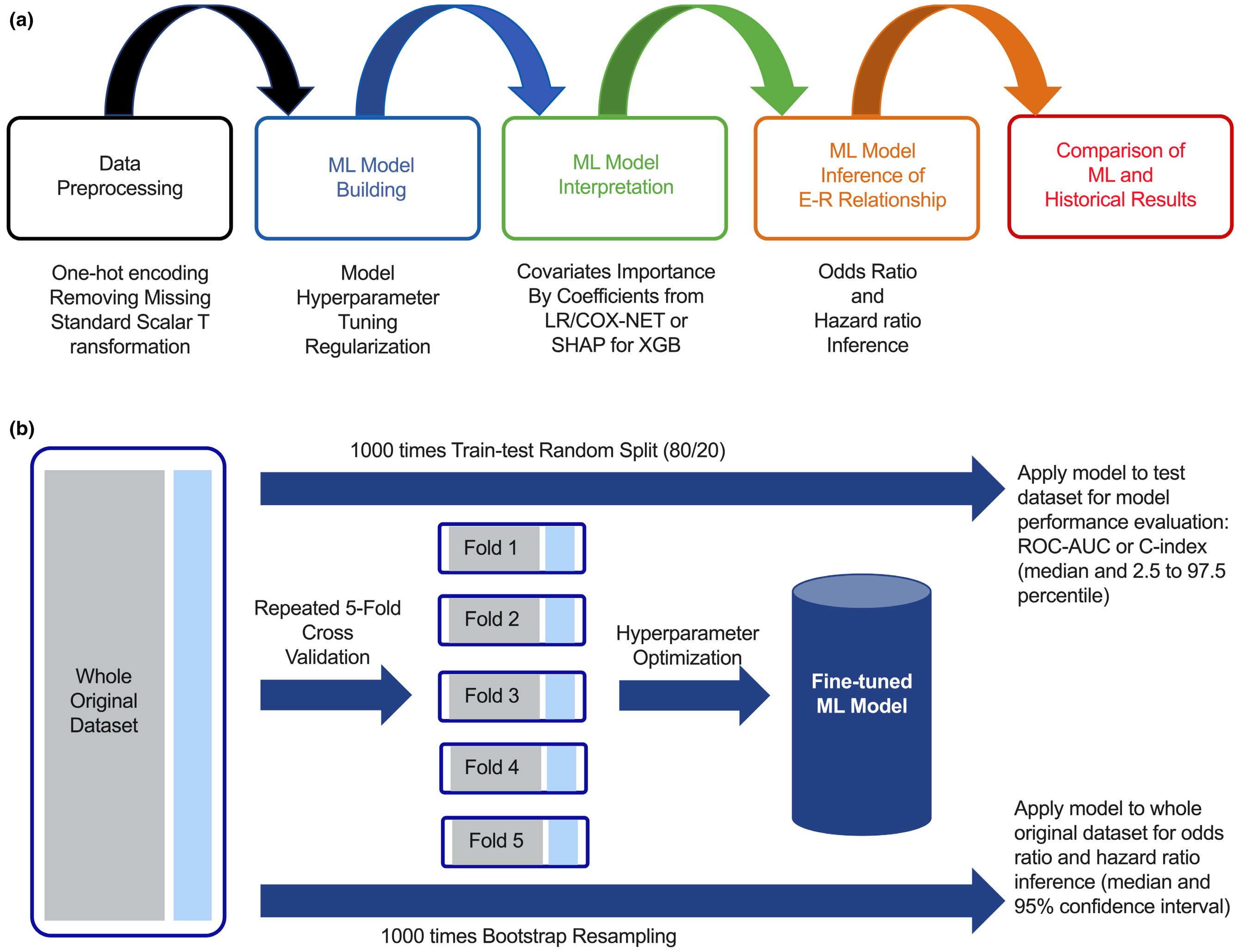
(a)
Data Preprocessing
ML Model Building
ML Model Interpretation
ML Model Inference of E-R Relationship
Comparison of ML and Historical Results
One-hot encoding Removing Missing Standard Scalar T ransformation
Model Hyperparameter Tuning Regularization
Covariates Importance By Coefficients from LR/COX-NET or SHAP for XGB
Odds Ratio and Hazard ratio Inference
(b)
1000 times Train-test Random Split (80/20)
Apply model to test dataset for model performance evaluation: ROC-AUC or C-index (median and 2.5 to 97.5 percentile)
Whole Original Dataset
Repeated 5-Fold Cross Validation
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Hyperparameter Optimization
Fine-tuned ML Model
Apply model to whole original dataset for odds ratio and hazard ratio inference (median and 95% confidence interval)
1000 times Bootstrap Resampling

FIGURE 1 (a) Scheme of ML workflow. The color arrows indicate operations within each step. (b) Detailed workflow for ML model building and evaluation of E–R relationship. AUC, area under the curve; C-index, concordance index; E–R, exposure–response; LR, logistic regression; ML, machine learning; ROC, receiver operating characteristic; SHAP, Shapley Additive Explanation; XGB, XGboost

whole original dataset with replacement and apply the ML model to the whole original dataset to evaluate the OR and HR, and this is repeated for 1000 times (Figure 1b). Based on the 95% CIs, we examine whether the intervals contain the value of one, to quantify whether there is a meaningful difference between group A and group B.

Formula 1a: Unadjusted log odds and OR for LR model

$$(\text{Unadjusted log odd})_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

$$\text{OR} = \exp^{\left(\text{mean}_{i\in \text{Group}A}(\text{logit}(p_i)) - \text{mean}_{i\in \text{Group}B}(\text{logit}(p_i))\right)} \quad (1a)$$

Here, p_i is the model predicted probability for each patient.

Formula 1b: Unadjusted log odds of each patient explained by SHAP values from the XGB model[21] and OR for XGB model.

$$(\text{Unadjusted log odd})_i = f(\vec{x}_i) = \Phi_0(f) + \sum_{j=1}^{M} \Phi_j(f,\vec{x}_i)$$
$$= \Phi_0(f) + \Phi_1(f,\vec{x}_i) + \ldots + \Phi_j(f,\vec{x}_i) + \ldots + \Phi_M(f,\vec{x}_i)$$

where $\vec{x}_i = \ < x_{1i}, \ldots, x_{ji}, \ldots, x_{Mi} >$

$$\text{OR} = \exp^{\left(\text{mean}_{i\in \text{GroupA}}(\text{unadjusted log odds})_i - \text{mean}_{i\in \text{GroupB}}(\text{unadjusted log odds})_i\right)} \quad (1b)$$

Here, $\vec{x}_i$: covariate vector with length M for patient i; f refers the XGB model function to link $\vec{x}_i$ with log odds of patient i; $\Phi_j(f,\vec{x}_i)$: SHAP value of explanatory variable j for patient i. Here, the SHAP values[21,25] explain the contribution of each covariate to the model prediction in the log odds domain. The unadjusted log odds for each patient i predicted by function f (XGB model) based on individual covariate vectors is decomposed by formula 1b using SHAP analysis,[21,25] which is the sum of all covariate attributions plus the expected value (e.g., model prediction without any covariates).

Formula 2a: Adjusted ORs from LR model

$$\mathrm{OR}^{\mathrm{LR}}\left(\frac{Q_{\mathrm{A}}}{Q_{\mathrm{B}}}\right) = \exp^{\int\left(\beta_0+\Sigma\left(\beta_j \bullet X_j\right)+\beta_{\mathrm{C}} \bullet \mathrm{C}\right) \bullet \mathrm{PDF}(\mathrm{C}|\mathrm{C}\in Q_{\mathrm{A}})\mathrm{dC}-\int\left(\beta_0+\Sigma\left(\beta_j \bullet X_j\right)+\beta_{\mathrm{C}} \bullet \mathrm{C}\right) \bullet \mathrm{PDF}(\mathrm{C}|\mathrm{C}\in Q_{\mathrm{B}})\mathrm{dC}} = \exp^{\beta_{\mathrm{C}} \bullet (\mathrm{mean}(\mathrm{C}|\mathrm{C}\in Q_{\mathrm{A}})-\mathrm{mean}(\mathrm{C}|\mathrm{C}\in Q_{\mathrm{B}}))} \tag{2a}$$

Here, the COX-NET model and XGB model with objective = "survival:cox" predict either the log of the HR of each patient to the baseline hazard in a proportional hazard function[18] or the HR,[20] respectively, both as p_i in the formula.

Formula 4: Adjusted HR of group A to group B based on adjusted prediction from $\mathrm{C}_{\mathrm{trough}}$, conditioned on other covariates

Formula 4a: Adjusted HR for COX-NET model

$$\mathrm{HR}^{\mathrm{COX-NET}}\left(\frac{Q_{\mathrm{A}}}{Q_{\mathrm{B}}}\right) = \exp^{\int\left(\log(h_0(t))+\Sigma\left(\beta_j \cdot X_j\right)+\beta_{\mathrm{C}} \cdot \mathrm{C}\right) \cdot \mathrm{PDF}(\mathrm{C}|\mathrm{C}\in Q_{\mathrm{A}})\mathrm{dC}-\int\left(\log(h_0(t))+\Sigma\left(\beta_j \cdot X_j\right)+\beta_{\mathrm{C}} \cdot \mathrm{C}\right) \cdot \mathrm{PDF}(\mathrm{C}|\mathrm{C}\in Q_{\mathrm{B}})\mathrm{dC}} = \exp^{\beta_{\mathrm{C}} \cdot (\mathrm{mean}(\mathrm{C}|\mathrm{C}\in Q_{\mathrm{A}})-\mathrm{mean}(\mathrm{C}|\mathrm{C}\in Q_{\mathrm{B}}))} \tag{4a}$$

Here, C: $\mathrm{C}_{\mathrm{trough}}$; Q_{A}, Q_{B}: group A and group B based on patient $\mathrm{C}_{\mathrm{trough}}$ quantiles; PDF: probability density function of $\mathrm{C}_{\mathrm{trough}}$ in each group. The formula assumes all other covariates (X_j) are constant. It is written based on the g-formula principle from Robins et al.[26,27] For an LR model, the adjusted OR is estimated by the ratio of the geometric means of adjusted odds of group A to group B[22,23] (formula 2a), using the covariate coefficient related to $\mathrm{C}_{\mathrm{trough}}$ estimated from the model (β_{C}). Here, the OR of group A to group B is adjusted (i.e., conditioned) on the effect of all other covariates except $\mathrm{C}_{\mathrm{trough}}$.

Formula 2b: Adjusted OR of group A to group B based on $\mathrm{C}_{\mathrm{trough}}$, using XGB model, conditioned on other model covariates

$$\mathrm{OR}^{\mathrm{XGB}} = \exp^{\left(\mathrm{mean}_{i\in\mathrm{GroupA}}(\phi_{\mathrm{C}}(f,x_i))-\mathrm{mean}_{i\in\mathrm{GroupB}}(\phi_{\mathrm{C}}(f,x_i))\right)} \tag{2b}$$

Here, $\phi_{\mathrm{C}}\left(f, x_i\right)$ is the SHAP value of explanatory variable $\mathrm{C}_{\mathrm{trough}}$ for patient i, which can be viewed as the adjusted log odds of $\mathrm{C}_{\mathrm{trough}}$, conditioned on other covariates. Thus, the adjusted ORs associated can be computed by the ratio of geometric mean of $\exp(\phi_{\mathrm{C}}\left(f, x_i\right))$ from group A and group B (formula 2b), applying a similar concept as deriving the adjusted ORs from LR models (formula 2a).

Formula 3: Unadjusted HR of group A to group B based on model prediction from all covariates for COX-NET or XGB model.

Formula 3a: Unadjusted HR for COX-NET

$$\mathrm{HR}^{\mathrm{COX-NET}} = \exp^{\left(\mathrm{mean}_{i\in\mathrm{GroupA}}(p_i)-\mathrm{mean}_{i\in\mathrm{GroupB}}(p_i)\right)} \tag{3a}$$

Formula 3b: Unadjusted HR for XGB

$$\mathrm{HR}^{\mathrm{XGB}} = \exp^{\left(\mathrm{mean}_{i\in\mathrm{GroupA}}(\log(p_i))-\mathrm{mean}_{i\in\mathrm{GroupB}}(\log(p_i))\right)} \tag{3b}$$

Here, C: $\mathrm{C}_{\mathrm{trough}}$; Q_{A}, Q_{B}: group A and group B based on patient $\mathrm{C}_{\mathrm{trough}}$ quantiles; PDF: probability density function of $\mathrm{C}_{\mathrm{trough}}$ in each group. The formula assumes all other covariates (X_j) are constant. It is similarly written based on the g-formula principle from Robins et al.[26,27] The adjusted HR is derived using the ratio of geometric mean of adjusted hazard rates for each group, based on the model estimated coefficient of $\mathrm{C}_{\mathrm{trough}}$ (β_{C}).

Formula 4b: Adjusted HR for XGB model

$$\mathrm{HR}^{\mathrm{XGB}} = \exp^{\left(\mathrm{mean}_{i\in\mathrm{GroupA}}(\phi_{\mathrm{C}}(f,x_i))-\mathrm{mean}_{i\in\mathrm{GroupB}}(\phi_{\mathrm{C}}(f,x_i))\right)} \tag{4b}$$

Here, $\phi_{\mathrm{C}}\left(f, x_i\right)$ is the SHAP value of explanatory variable $\mathrm{C}_{\mathrm{trough}}$ for patient i, which can be viewed as the adjusted log hazard rate of $\mathrm{C}_{\mathrm{trough}}$, conditioned on other covariates. For the XGB model, the adjusted HR of $\mathrm{C}_{\mathrm{trough}}$ is derived based on the research work by Sundrani et al.,[28] except that geometric mean is used instead of arithmetic mean when computing the aggregated hazard rate of group A or group B, for a consistency with other formulas in this research.

Sensitivity analysis to assess impact of methodology variations on E–R relationship estimation

There are multiple nuances in the methodology details when implementing the ML workflow. To estimate the sensitivity of our ML approach on evaluating the E–R relationship, we assessed the impacts of the following two methodology variations on evaluating the CIs of ORs and HRs for E–R relationship, using data from T-DM1 trials: (1) select only the clinically important covariates matched those used in historical analysis[29] (covariates in bold font in Table S2) versus using all covariates (primary analysis); and (2) excluding the control arm versus including in the model building. For a consistent comparison of the E–R relationship in the presence and absence of the control arm in the sensitivity analysis, we derived ORs and HRs by comparing high exposure quantiles (Q2, Q3, and Q4) to the lowest exposure quantile (Q1) as Q2/Q1, Q3/Q1, and Q4/Q1.

RESULTS

Final datasets for primary analysis and sensitivity analysis

For E–R analysis with BOR end point, there are 1286 patients from the EandT dataset, with 722 patients in the control arm and 564 patients in T-DM1 treatment arm, and 35 covariates (Table S2) are included in primary analysis; for the MARIANNE dataset, there are 610 patients with 46 covariates (Table S3). For the dataset with OS end points, in the EandT dataset, there are 1358 patients with 35 covariates; for the HELOISE dataset, there are 224 patients and 17 covariates (Table S4). The statistics of ML prediction target variables are summarized in Table 1.

Final ML models and performance

Table 2 listed the median values of each hyperparameter from four ML models tuning with 100 times repeated five-fold CV for all analyses in this study. Within all of the hyperparameters, l1_ratio, alpha_min_ratio, reg_alpha, reg_lambda determine the L1 and L2 regularization penalty terms.[18,19] Figure 2 shows model performance on test datasets with mean and SD of ROC-AUC or C-index for both classification and regression models, from 1000 datasets with train-test split datasets. For all four datasets, the performance of LR/COX-NET and XGB are comparable as assessed by ROC-AUC or C-index. Values of mean ROC-AUC of 0.61 to 0.67 and mean C-index of 0.66 to 0.72 are obtained (Table S5), suggesting moderate model performances.

Covariates importance assessment

Figure 3 shows the covariates' importance explained by LR and XGB models in predicting BOR and OS end points, with the most important covariate on the top. In the EandT trials, we found that the exposure covariate (T-DM1 concentration at the end of cycle 1 as predicted by the population PK model of T-DM1, C_{trough}/PCMIN)[30] exhibited high impact in both the BOR end point (most important for both LR and XGB models) and the OS end point (second important in COX-NET and fourth important in XGB; Figure 3a,c). On the other hand, for the T-DM1 trial in the previous patients with untreated metastatic breast cancer (MARIANNE study), C_{trough} did not appear in the top 10 important covariates for both LR and XGB methods (Figure 3b). For the HELOISE study, the clearance (CL) of large molecule drugs is generally considered to be highly correlated with the patient's health status[4,31–33] and ranked number one for both methods. $C_{\text{min,ss}}$, which is the steady-state trastuzumab C_{min} after receiving 10 or 6 mg/kg Q3W dose starting cycle 2 and thus accounted for the dose difference of the two arms, ranked number four in COX-NET and number three in XGB, suggesting a potential presence of weak E–R relationship which may not be clinically meaningful as reflected by the dose response relationship.[16] The Herceptin trough concentrations at the end of cycle 1 (C_{min}) after giving the same cycle 1 loading dose of 8 mg/kg for all patients is less important than $C_{\text{min,ss}}$, which does not appear as the top 10 important covariates in COX-NET and ranks number four in XGB.

TABLE 1 Statistics of the machine learning prediction target variables of final datasets

	Control arm patient number		Treatment arm patient number		Total number
BOR end points	**Responder**	**Nonresponder**	**Responder**	**Nonresponder**	
EandT	204 (28.25%)	518 (71.75%)	224 (39.72%)	340 (60.28%)	1286
MARIANNE	194 (67.83%)	92 (32.17%)	205 (63.27%)	119 (36.73%)	610
	Control arm patient number		**Treatment arm patient number**		
OS end points	**Censored**	**With event**	**Censored**	**With event**	
EandT	375 (55.72%)	298 (44.28%)	393 (57.37%)	292 (42.63%)	1358
	Arm A 6 mg/kg		**Arm B 10 mg/kg**		
	Censored	**With event**	**Censored**	**With event**	
HELOISE	63 (55.26%)	51 (44.74%)	51 (46.36%)	59 (53.64%)	224

Abbreviations: BOR, best objective response; EandT, the EMILIA and TH3RESA studies; OS, overall survival.

TABLE 2 Median hyperparameters from model tuning with 100 times of repeated train-test split for each of the eight models

BOR end points	**Logistic regression classifier (loss = ROC-AUC)**	**XGB classifier (loss = ROC-AUC, eval_metric = "auc")**
EandT with BOR end points	"l1_ratio": 0.507, "max_iter": 164.64, "penalty": "elasticnet," "solver": "saga"	"eta": 0.004, "max_depth": 3, "min_child_weight": 0.083, "reg_alpha": 0.608, "reg_lambda": 0.178, "subsample": 0.857,
MARIANNE with BOR endpoints	"l1_ratio": 0.44, "max_iter": 186.223, "penalty": "elasticnet," "solver": "saga"	"eta": 0.008, "max_depth": 6, "min_child_weight": 0.059, "reg_alpha": 0.46, "reg_lambda": 0.24, "subsample": 0.883
OS Endpoints	**COX-NET regression model (loss = C-index)**	**XGB regression model (loss = C-index, eval_ metric = "root mean square error")**
EandT with OS endpoints	"alpha_min_ratio": 0.021, "l1_ratio": 0.74, "alphas": 0.001	"eta": 0.004, "max_depth": 4, "min_child_weight": 0.307, "reg_alpha": 0.728, "reg_lambda": 0.623, "subsample": 0.839, "objective": "survival:cox"
HELOISE with OS endpoints	"alpha_min_ratio": 0.035, "l1_ratio": 0.727, "alphas": 0.004	"eta": 0.015, "max_depth": 4, "min_child_weight": 0.314, "reg_alpha": 0.045, "reg_lambda": 0.084, "subsample": 0.841, "objective": "survival:cox"

Abbreviations: AUC, area under the curve; BOR, best objective response; EandT, the EMILIA and TH3RESA studies; ROC, receiver operating characteristic; XGB, XG boost.

OR and HR for evaluating E–R relationship

Figure 4 presents the median value and 95% CIs of OR of EandT (Figure 4a) and MARIANNE (Figure 4b) with BOR end points and HR of EandT (Figure 4c) and HELOISE (Figure 4d) with OS end point with 1000 bootstrapping, with numerical values in Tables S6–S9.

For OR and HR in EandT trials (Figure 4a,c), the adjusted OR and HR values across Q1, Q2, Q3, and Q4 compared to the control group are overall shallower than the unadjusted values, suggesting that both linear and nonlinear ML methods offered correction of effects from confounding covariates. Further, based on the adjusted values, a positive E–R relationship is observed that higher drug exposures are associated with better efficacy in both BOR and OS. The patients in the lowest exposure quantile (Q1 group) showed equivalent or better efficacy compared to the control group as compared to the value of one, suggesting a non-detrimental effect for them. These key conclusions from ML models are overall consistent with the conclusion of statistically significant positive E–R relationship from the historical pharmacometrics analysis using corrected HRs for Cox model with selected covariates.[7,8]

Based on the OR for MARIANNE trial (Figure 4b), the unadjusted ORs from both LR and XGB for Q1 versus control group is around 0.4–0.5 in median values, suggesting a trend of E–R, which is likely false positive due to the confounding effects from other covariates. The adjusted OR from both LR and XGB suggested a flat E–R relationship across exposure quantiles with the 95% CIs of OR including the value of one.

The HELOISE study is the trial in which two doses were tested and there are no clinical meaningful OS differences from the high-dose group compared to the low-dose group.[16] We evaluated unadjusted and adjusted HRs of C_{min}, $C_{min,ss}$, and CL for the HELOISE trial. For the COX-NET and XGB models (Figure 4d), the median value

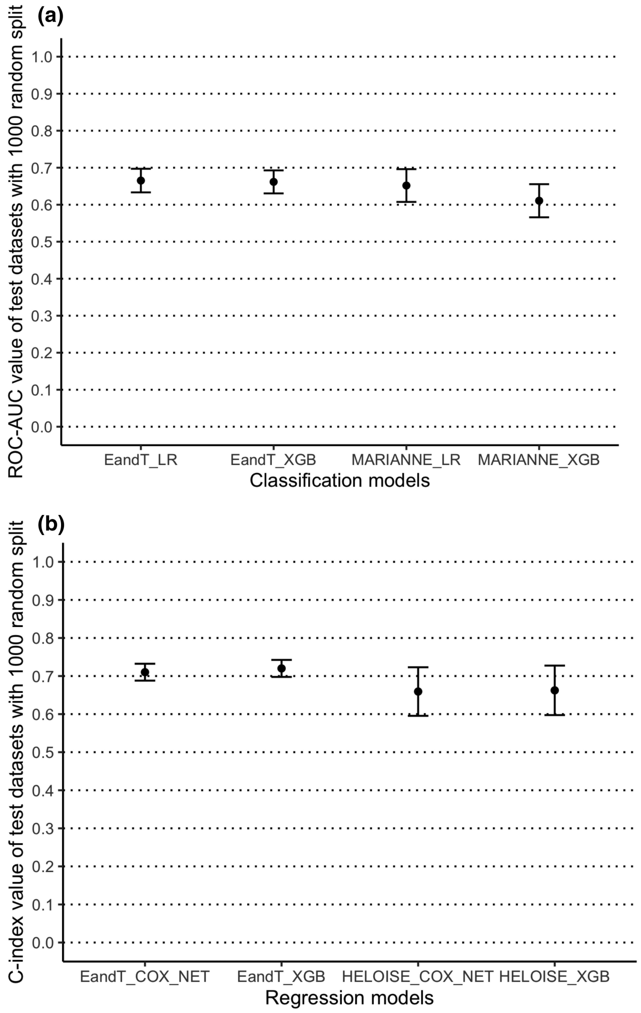

FIGURE 2 ROC-AUC and C-index (mean and SD) of test datasets from 1000 random train-test split for both classification and regression models. AUC, area under the curve; C-index, concordance index; EandT, the EMILIA and TH3RESA studies; LR, logistic regression; ROC, receiver operating characteristic; XGB, XGboost

and CIs of unadjusted HRs for Q2/Q1, Q3/Q1, and Q4/Q1 are all below one, most likely due to the effects from the confounded covariates. The adjusted HRs for cycle 1 C_{min} showed no impact on OS. The median values of adjusted HRs for $C_{min,ss}$ are less than one with 95% CIs containing one, suggesting a nonclinical meaningful E–R relationship. Patients with higher CL (Q1) showed higher HR, suggesting a higher risk for sicker patients.

In summary, the OR and HR results are consistent with the insights from the covariate importance assessment results. Across these analyses, the adjusted ORs or HRs showed a shallower E–R relationship compared with the unadjusted ones, suggesting that the ML methods can adjust for the confounding effects from other impacting covariates. The overall conclusions of either positive or flat E–R relationships are consistent between the linear and tree-based ML models, and consistent with historical analysis (when available).

Sensitivity analysis to assess impact of methodology variations on E–R relationship estimation using T-DM1 trials

First, in the sensitivity analysis when only a subset of important covariates based on disease knowledge and historical analysis are included for EandT trial data, the conclusions of positive E–R relationship based on values of ORs (Figure 5a) and HRs (Figure 5c) is consistent with the primary analysis with all covariates included, for both linear and nonlinear models. Similarly, for the MARIANNE study (Figure 5b), a flat E–R relationship is identified for both LR and XGB models given all of the 95% CI contains one, with consistent conclusions as the primary analysis. These results suggested that the regularized ML approaches in the primary analysis with all available covariates included has the advantage of skipping the step of covariate selection in the traditional analysis without a large impact on the conclusions of E–R relationship.

Second, we tested the impact on evaluating the E–R relationship with or without the control arm included. In the primary analysis, when the data contain a control arm, it is included in the model development. Here, the model was built with only the T-DM1 treatment arm of EandT or the MARIANNE study. We found that the key conclusions of E–R relationship based on ORs and HRs are largely consistent as the primary analysis, whereas the CIs appear larger in most cases based on the models built without the control arm (Figure 5). This indicates that by including the control group in the model building, there are more patients to inform the relationship of baseline covariates with the prediction target, which resulted in improved robustness of the ML model.

In summary, the results of the sensitivity analysis suggest that including the control arm when available is preferred, allowing more data to inform the ML models. In addition, the ML model evaluations of E–R conclusions do not strongly rely on a strict covariate selection process, which would be a benefit from the regularization that is not present in conventional LR/CoxPH methods. Overall, our primary ML approaches leveraging all available covariates and control arm data are robust approaches.

DISCUSSION

In this research work, we explored ML methods for E–R analysis and compared the ML results with historical E–R analysis results. In historical analysis by

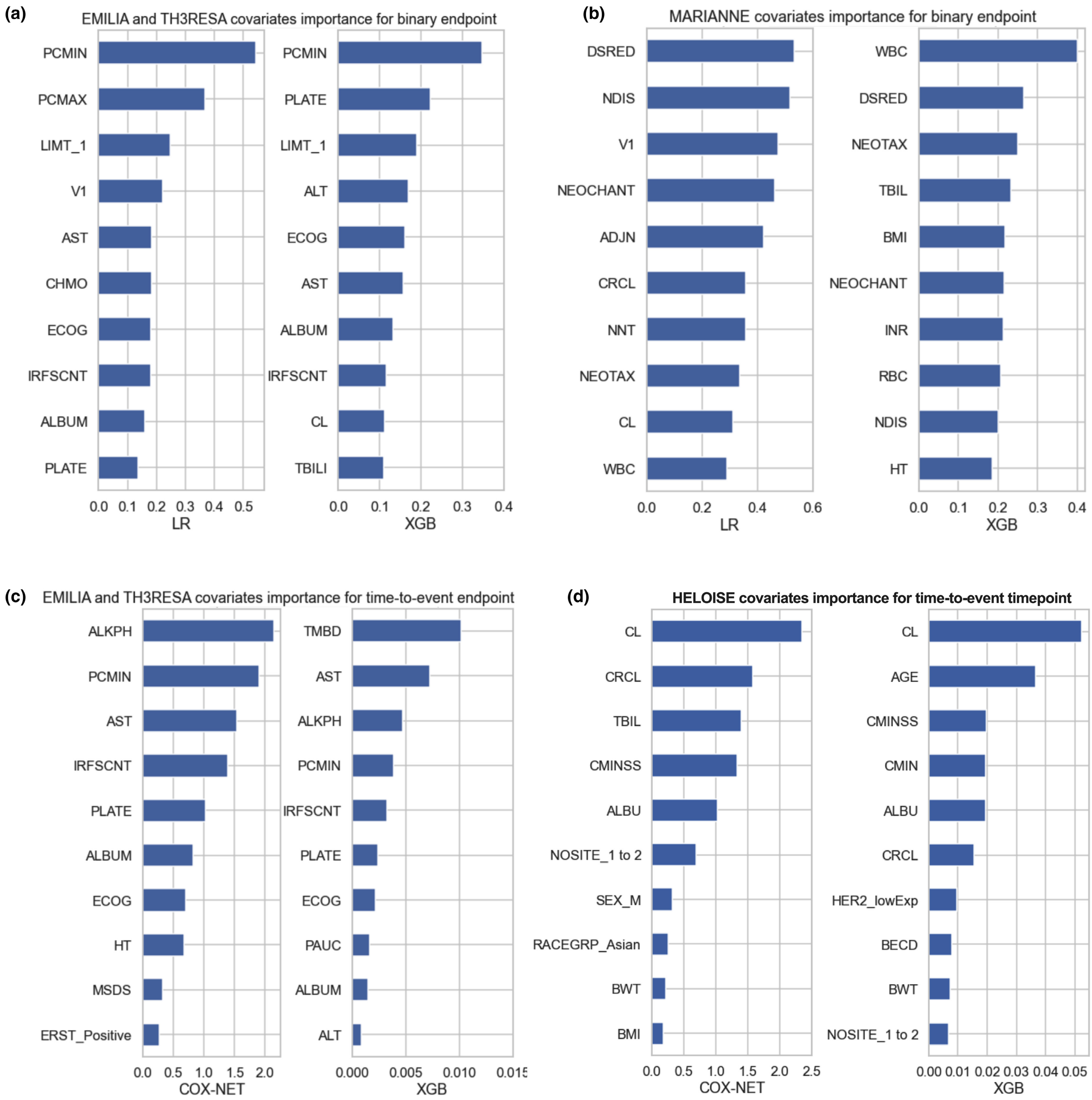

FIGURE 3 Top 10 important covariates for each analysis (see Tables S1–S3 for covariate descriptions). *X*-axes for LR or COX-NET: absolute value of coefficients for each covariate. *X*-axes for XGB: the average of absolute SHAP values across all patients for each covariate. ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; BWT, body weight; CL, clearance; CRCL, creatinine clearance; ECOG, Eastern Cooperative Oncology Group; RBC, red blood cells; SHAP, Shapley Additive Explanation; WBC, white blood cells; XGB, XGboost

case-matching approach, which is one of the causal inference methods used in historical E–R analysis, to adjust for the potential confounding from other baseline covariates,[7,8] the Q1/control HR in the EMILIA study is 0.71 (95% CI: 0.44–1.14),[7] and is 0.96 (95% CI: 0.63–1.47) in the TH3RESA study,[8] for OS end point, suggesting that the Q1 group is trending beneficial and is not worse than control given 95% CI of HR containing one. In the current analysis, the EandT study were pooled and the HR of Q1/control is 1.01 (95% CI: 0.955–1.078) and 0.852 (95% CI: 0.771–0.942) by COX-NET and XGB, respectively (Table S8), consistently suggesting that the Q1 group is not worse than the control group. It appears the 95% CI from ML models are smaller than the case-matching methods,[7,8] and more future analysis in other clinical trials may demonstrate whether this trend

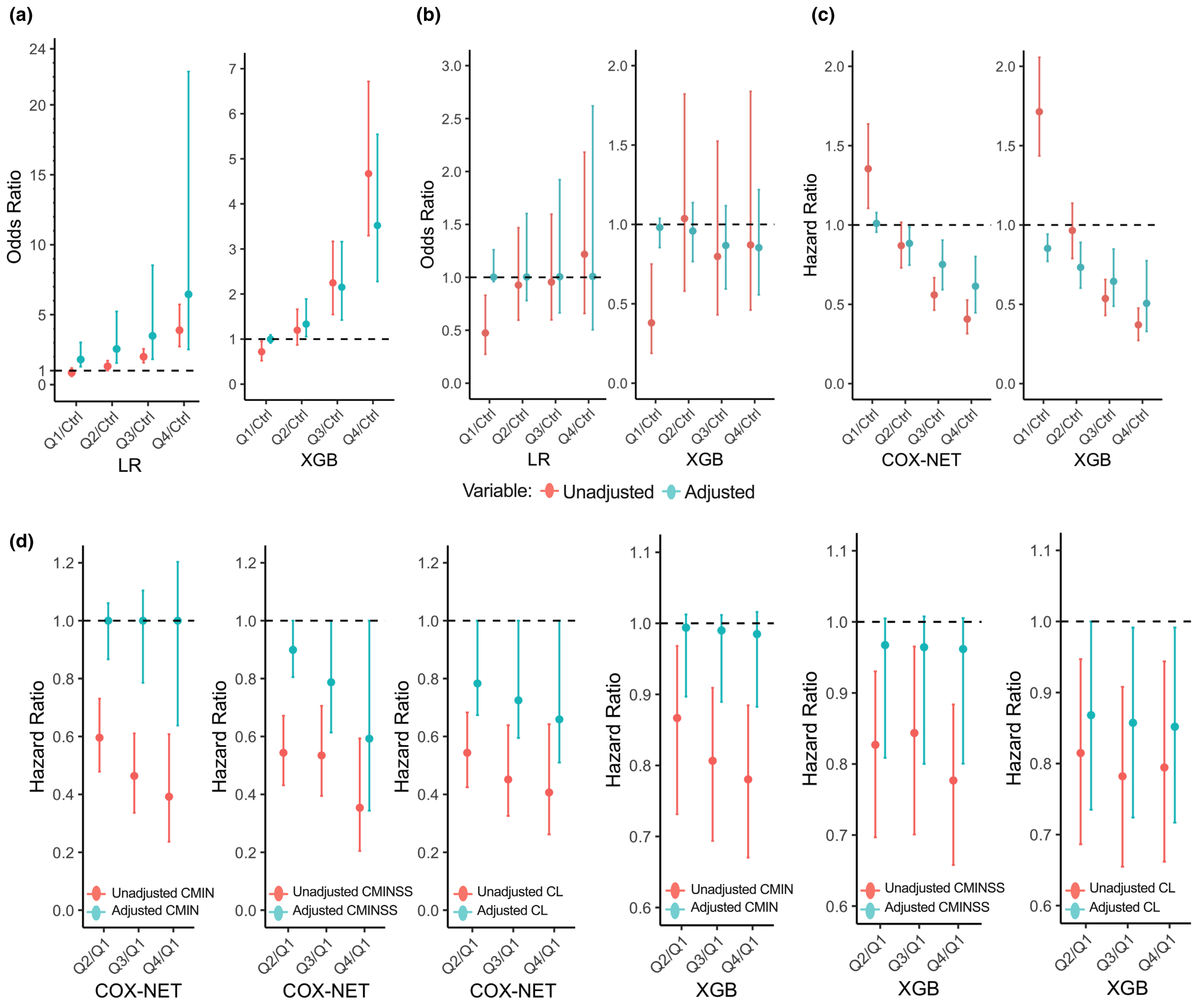

FIGURE 4 ORs of EandT and Marianne and HRs of EandT (a–c), Q1/Ctrl, Q2/Ctrl, Q3/Ctrl, and Q4/Ctrl represent the ratio of different quantiles versus the control group. HRs of HELOISE from XGB model (d), among the small three images, the first two images are HRs of C_{min} and $C_{min,ss}$, Q2/Q1, Q3/Q1, and Q4/Q1 represent the ratio of higher quantiles Q2, Q3, and Q4 versus the lowest quantile Q1, respectively. The last image shows the HRs for CL, Q2/Q1, Q3/Q1, and Q4/Q1 represent the ratio of lower quantiles Q2, Q3, and Q4 versus the highest quantile Q1. (a) EandT BOR end point; (b) MARIANNE BOR end point; (c) EandT OS end point; (d) HELOISE OS end point. BOR, best objective response; CL, clearance; C_{min}, minimum concentration; $C_{min,ss}$, minimum clearance at stead-state; EandT, the EMILIA and TH3RESA studies; HR, hazard ratio; LR, logistic regression; OR, odds ratio; XGB, XGboost

is consistent. Overall, the E–R relationship evaluated by ML methods are largely consistent with historical findings for the EMILIA[7] and TH3RESA studies.[8] For the MARIANNE study, historical analysis by LR suggested a statistically significant coefficient of log C_{trough} on BOR (Genentech data in file). The ML analysis here suggested a flat E–R relationship for BOR. Separately, it was reported that the progression free survival HR for Q1/control is trending higher than the other higher exposure quantiles with overlapping CIs, by Cox adjusted, case matching, or double adjusted methods.[34] We believe in real-world applications there will be cases where the E–R conclusions from pharmacometrics and ML approaches are consistent, or cases that are inconsistent from statistical significance point-of-view (i.e., one is significant but not another), and broader application of the ML methods to additional large-molecule oncology drug clinical trials will gain more experiences in the clinical relevance of these analyses.

Similar to the traditional pharmacometric approaches for E–R analysis, static drug exposure parameters (e.g., PC_{min}) rather than the longitudinal PK data were used in the current ML analysis, and the E–R relationship identified may represent association/correlation or causation.

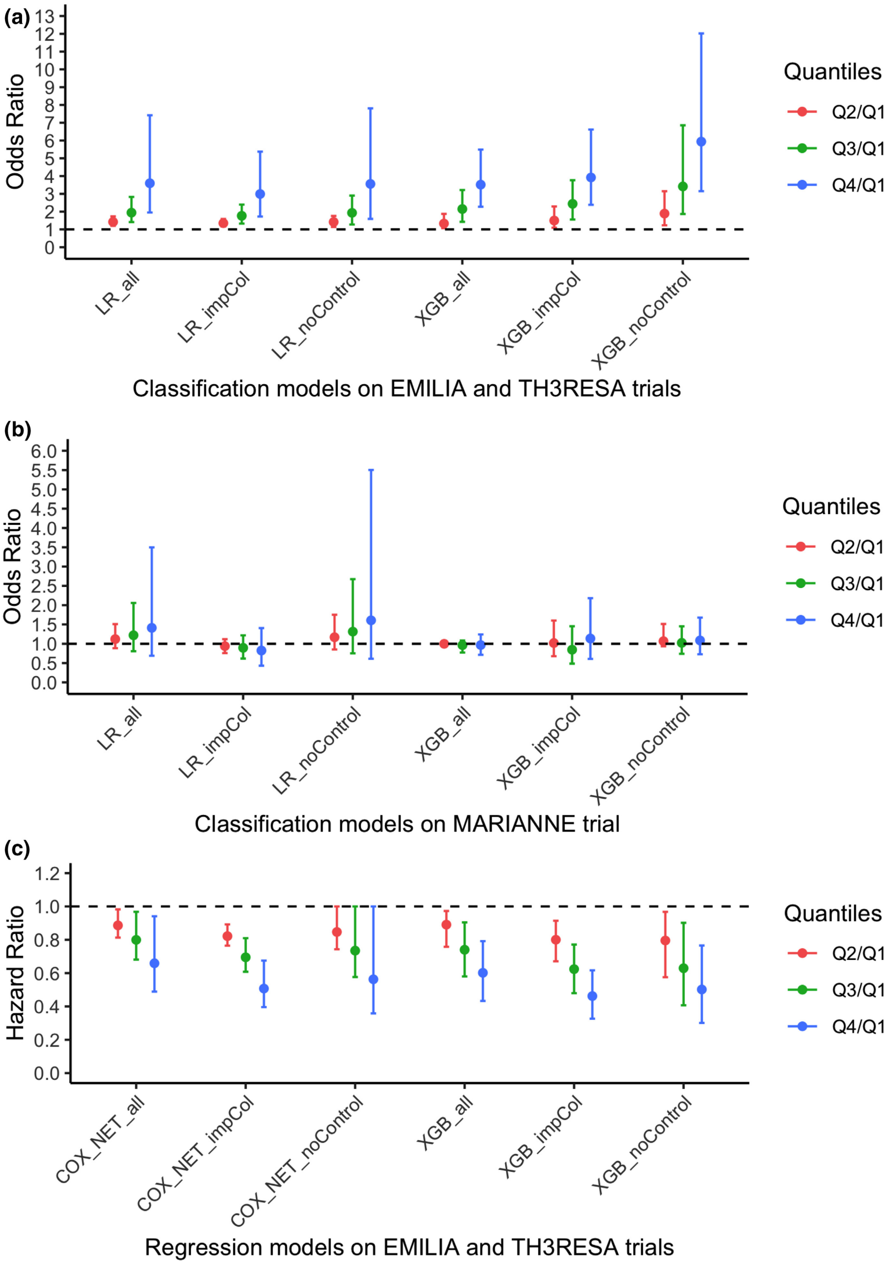

FIGURE 5 Median and 95% CI for ORs and HRs for Q2, Q3, Q4 to Q1 from 1000 bootstrap analyses. For *x*-axes, LR_all and XGB_all: primary analysis approach; LR_impCol and XGB_impCol: analysis with important covariates; LR_noControl and XGB_noControl: analysis without control arm. (a) Sensitivity analysis of EandT dataset with BOR end points; (b) sensitivity analysis of MARIANNE dataset with BOR end points; (c) sensitivity analysis of EandT dataset with OS end points. BOR, best objective response; CI, confidence interval; EandT, the EMILIA and TH3RESA studies; HR, hazard ratio; LR, logistic regression; OR, odds ratio; OS, overall survival; XGB, XGboost

The ML analysis conducted here is based on the causal assumptions of E–R relationship, as illustrated by the conceptual causal diagram for the confounded E–R relationship shown in the literature for large-molecule oncology drugs.[35] Under this causal assumption, and the key assumption that all the baseline confounders affecting both PK and efficacy are included in the ML model, the effect of drug exposure (PC_{min}) on efficacy identified by the ML model and SHAP analysis can be causal instead of merely association/correlation. However, it is often unknown whether all confounders are available in the dataset to be included and fully adjusted by the ML model, and missing key confounders may bias the E–R relationship estimation as shown in our recent simulation study.[35] A future area of study is to utilize longitudinal PK data in ML models for E–R analysis, or apply other causal-inference ML models, to further assess the causal conclusions.

Linear or log-linear models with rigorous covariate selection but without regularization are typically applied in the traditional E–R analysis with assumption of linear/log-linear relationship between covariates (including baseline and exposure covariates) and response (the efficacy end points). However, the ground truth of the E–R relationship could be nonlinear or linear, depending on the range of exposures studied. The basic pharmacological principle is that a Hill model is commonly used to describe the saturation of pharmacological effects at high exposures. Tree-based models, such as XGB, would offer the flexibility of modeling linear and nonlinear relationships. This motivated us to evaluate both linear and nonlinear (tree-based) ML models in each analysis. The ROC-AUC and C-index values are largely consistent, suggesting a similar model predictive performance for linear and nonlinear models across analyses here. It was previously reported that tree-based ML models outperform the linear Cox models as assessed by C-index in high dimensional data, and datasets with complex nonlinear relationship.[5] We foresee that non-linear tree-based methods and neural networks would outperform linear ML models and pharmacometrics models when analyzing datasets with high dimensional input covariates, such as multi-layer "omics" data (i.e., gene expression, proteomics, metabolomics, and imaging data), to more accurately evaluate E–R relationship. However, in this report, the datasets used are considered with relatively low dimension for the input covariates, and the performance differences between linear and tree-based methods are not large. Whether there is a low limit of sample size to generate a robust E–R relationship evaluation needs further assessment. The HELOISE study has the lowest number of patients (224 patients) among three analyses and XGB methods appear to have smaller differences than COX-NET between the unadjusted and adjusted HRs.

In the ML field, it is important to ensure the model performs similarly well on both training and validation/test data, known as good generalization or overfitting reduction. In this study, we aim to infer the E–R relationship from the entire dataset (instead of test data only) and also applied various techniques to ensure generalization for the models built from our ML workflow: CV for hyperparameter optimization[36,37] and regularization.[38] First, we selected the fine-tuned hyperparameters by using the median value from 100 times of repeated five-fold CV for the entire original dataset. This technique makes sure our selected hyperparameter performs stably in different bootstrap re-sampled datasets for model training and inferring the CI of ORs and HRs. CV is a common technique for overfitting reduction. Considering the relatively small data size and sometimes the imbalanced target in most E–R datasets from clinical trials, we chose to apply CV on the entire dataset. The methods of nested CV or train/test dataset split may be better in preventing data leaking, but can cause a small sample size and/or the minority label becoming even smaller after the inner and outer split of nested CV or the train/test split, for many clinical trial datasets. Second, regularization techniques are also used to increase generalization. Compared with the traditional E–R analysis method with no regularization and hence requiring forward addition and backward elimination to select the key covariates, we used both L1 and L2 regularization in all ML models to reduce the model overfitting and removed the need of the covariate selection process. It is well known that L1 regularization can lead to zero coefficient and thus can perform feature selection implicitly, whereas L2 regularization does not perform feature selection as coefficients are only reduced to values near zero instead of zero.[38] We chose elastic net regularization (both L1 and L2 regularization) instead of L1 or L2 regularization alone for the following reason. Elastic net uses a weighted linear combination of the L1 and L2 penalties in the loss function. As reported, it often outperforms the L1 method, while enjoying a similar sparsity of representation.[39] In addition, if the data contains a group of covariates that are highly correlated, instead of picking one covariate from a pair of correlated covariates randomly in ML models with L1 regularization only, elastic net inherently encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together.[39] We consider that the elastic net provides a robust regularization approach for the E–R analysis of clinical trial data which often has limited sample size, so that the model can include the effect of as many covariates as possible. This could hypothetically reduce the risk of inferring a false positive E–R relationship due to missing important confounders. We acknowledge that in the real-world analysis, there are nuances in ML methods and some other ML workflow might be more rigorous in the aspect of preventing data leaking during hyperparameter tuning, including nested CV and train/test split with hyperparameters tuned by CV in the training data. The impact of these variations on the quantitative evaluation of E–R relationships and subsequent decision making in drug development remains to be studied.

This analysis suggested that both linear methods and tree-based nonlinear approaches combining with SHAP can be good approaches for E–R analysis. One question to discuss is what are the pros and cons for ML methods, and in which scenarios ML methods would be advantageous. One major advantage of linear ML methods, such as LR and COX-NET, is the ease of covariate importance interpretation based on the absolute values of coefficients. These methods with regularization can also avoid the issue of multicollinearity. However, when the decision boundary is nonlinear, or there are complex interactions among covariates, tree-based approaches such as XGB can offer advantages. Furthermore, the SHAP approach works efficiently with the XGB model and allows assessment of covariate importance, making the XGB model no longer a black box but explainable in generating evaluations. The ORs and HRs, which are essential for quantifying E–R relationships, can be derived from SHAP values. In addition, unlike linear models, the XGB model is able to handle unnormalized and missing values internally in the algorithm thus not requiring the modeler to normalize and impute the input covariates, and include certain covariates with a relatively high percentage of missing data. Regarding the results of covariate importance, we observed discrepancy in the exact rank-order across different ML models. This may be due to the inductive bias from various models. There are different sets of explicit or implicit assumptions made by each learning algorithm (linear or tree-based) in order to perform induction, that is, to generalize a finite set of observation (training data) into a general model of the domain. Without this bias, induction would not be possible, because the observations can normally be generalized in many ways.[40] Thus, it is not surprising that the rank-order of covariate importance is different. However, the key findings related to the relative importance of PCMIN are largely similar across linear and tree-based ML methods (Figure 3), and covariate-adjusted ORs and HRs related to PC_{min} lead to the largely consistent E–R conclusions (Figure 4).

In summary, our ML methodology provides an alternative toolbox for the traditional pharmacometrics approaches to assess E–R relationships for large molecule oncology drugs. The combination of predictive ML models like XGB with interpretable tools such as SHAP is powerful to identify the most informative covariates impacting the prediction outcome. Compared with the traditional pharmacometrics approaches, our ML approaches showed the robust ability to include a large and comprehensive list of covariates and their complex interactions without rigorous step-by-step covariates selection, which may potentially reduce the risk of obtaining a false-positive E–R relationship due to missing confounders in analysis. The XGB models can handle missing covariates without manual data imputation, and can build the model without assuming a certain explicit functional form between a covariate and the prediction target and accounting for the complex interactions among covariates, thus being more flexible compared to the linear models. We consider that the workflow proposed here can be broadly applied to middle to large size data, to address the key E–R questions in various stages of drug development. In the real-world application, the best model(s) to use is determined by how the dataset properties (e.g., the underlying relationship between covariates and the prediction target) are aligned with the underlying assumptions of the linear or nonlinear models, and should be evaluated case-by-case.

AUTHOR CONTRIBUTIONS

G.L. and D.L. wrote the manuscript. D.L. designed the research. G.L., H.L., and D.L. performed research and analyzed data. G.L., J.L., J.J., and D.L. reviewed and edited the manuscript.

ACKNOWLEDGMENT

The authors acknowledge Matts Kaagedal, Pascal Chanu, Kenta Yoshida, Bing Chen, Junyi Li, Kate Sillitoe, Rashed Harun, and Yannan Tang from Genentech/Roche, and Jonathan French and Daniel Polhamus from Metrum Research Group for their review, input and support for this work.

FUNDING INFORMATION

This work is supported by Roche/Genentech.

CONFLICT OF INTEREST

Gengbo Liu, James Lu, Jin Yan Jin, and Dan Lu are employees of Genentech and own Roche stock. Hong Seo Lim was an employee of Genentech when the work was conducted.

ORCID

Gengbo Liu https://orcid.org/0000-0002-5103-7191
James Lu https://orcid.org/0000-0002-9687-5607
Jin Yan Jin https://orcid.org/0000-0002-3627-0323
Dan Lu https://orcid.org/0000-0003-4531-3599

REFERENCES

1. Pinheiro J, Duffull S. Exposure response–getting the dose right. *Pharmaceut Stat: J Appl Statist Pharmaceut Ind*. 2009;8(3):173-175.
2. https://www.fda.gov/media/71277/download#:~:text=As%20noted%20above%2C%20exposure%2Dresponse,plasma%20concentration%20and%20PD%20response.
3. Overgaard RV, Ingwersen SH, Tornøe CW. Establishing good practices for exposure–response analysis of clinical endpoints in drug development. *CPT Pharmacometrics Syst Pharmacol*. 2015;4(10):565-575.
4. Dai HI, Vugmeyster Y, Mangal N. Characterizing exposure–response relationship for therapeutic monoclonal antibodies in immuno-oncology and beyond: challenges, perspectives, and prospects. *Clin Pharmacol Therapeut*. 2020;108(6):1156-1170.
5. Kawakatsu S, Bruno R, Kågedal M, et al. Confounding factors in exposure–response analyses and mitigation strategies for monoclonal antibodies in oncology. *Br J Clin Pharmacol*. 2021;87(6):2493-2501.
6. Liu C, Xu Y, Liu Q, Zhu H, Wang Y. Application of machine learning based methods in exposure–response analysis. *J Pharmacokinet Pharmacodyn*. 2022;49:401-410.
7. Li C, Wang B, Chen SC, et al. Exposure–response analyses of trastuzumab emtansine in patients with HER2-positive advanced breast cancer previously treated with trastuzumab and a taxane. *Cancer Chemother Pharmacol*. 2017;80(6):1079-1090.
8. Chen SC, Quartino A, Polhamus D, et al. Population pharmacokinetics and exposure–response of trastuzumab emtansine in advanced breast cancer previously treated with ≥2 HER2-targeted regimens. *Br J Clin Pharmacol*. 2017;83(12):2767-2777.
9. Yang J, Zhao H, Garnett C, et al. The combination of exposure–response and case-control analyses in regulatory decision making. *J Clin Pharmacol*. 2013;53(2):160-166.
10. Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Qual Quant*. 2018;52(4):1957-1976.

11. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. and Cho, H. Xgboost: extreme gradient boosting. R package version 04–2, 2015;1(4), 1–4.
12. Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci Rep*. 2021;11(1):1-3.
13. Verma S, Miles D, Gianni L, et al. Trastuzumab emtansine for HER2-positive advanced breast cancer. *New Engl J Med*. 2012;367(19):1783-1791.
14. Krop IE, Kim SB, Martin AG, et al. Trastuzumab emtansine versus treatment of physician's choice in patients with previously treated HER2-positive metastatic breast cancer (TH3RESA): final overall survival results from a randomised open-label phase 3 trial. *Lancet Oncol*. 2017;18(6):743-754.
15. Ellis PA, Barrios CH, Eiermann W, et al. Phase III, randomized study of trastuzumab emtansine (T-DM1)±pertuzumab (P) vs trastuzumab+ taxane (HT) for first-line treatment of HER2-positive MBC: Primary results from the MARIANNE study. *J Clin Oncol*. 2015;33:507.
16. Shah MA, Xu RH, Bang YJ, et al. HELOISE: Phase IIIb randomized multicenter study comparing standard-of-care and higher-dose trastuzumab regimens combined with chemotherapy as first-line therapy in patients with human epidermal growth factor receptor 2–positive metastatic gastric or gastroesophageal junction adenocarcinoma. *J Clin Oncol*. 2017;35(22):2558-2567.
17. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
18. Pölsterl S. Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J Mach Learn Res*. 2020;21(212):1-6.
19. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. SIGKDD Conference; 2016:785-794.
20. Bergstra, J., Yamins, D., Cox, D. D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. TProc. of the 30th International Conference on Machine Learning (ICML 2013), 2013;115-123.
21. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67.
22. Karlson KB, Popham F, Holm A. Marginal and conditional confounding using logits. *Sociological Methods Res*. 2021.
23. Satopää VA, Baron J, Foster DP, Mellers BA, Tetlock PE, Ungar LH. Combining multiple probability predictions using a simple logit model. *Int J Forecast*. 2014;30(2):344-356.
24. Allard D, Comunian A, Renard P. Probability aggregation methods in geoscience. *Mathemat Geosci*. 2012;44(5):545-581.
25. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Neural Information Processing Systems*. Curran Associates, Inc.; 2017:4765-4774.
26. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol*. 2017;46(2):756-762.
27. Hernán MA, Robins JM. *Causal Inference: What If*. FLChapman & Hall/CRC; 2020.
28. Sundrani S, Lu J. Computing the hazard ratios associated with explanatory variables using machine learning models of survival data. *JCO Clin Cancer Inform*. 2021;5:364-378.
29. https://xgboost.readthedocs.io/en/latest/parameter.html?highlight=cox.
30. Wang J, Song P, Schrieber S, et al. Exposure–response relationship of T-DM1: insight into dose optimization for patients with HER2-positive metastatic breast cancer. *Clin Pharmacol Therapeut*. 2014;95(5):558-564.
31. Agrawal S, Feng Y, Roy A, Kollia G, Lestini B. Nivolumab dose selection: challenges, opportunities, and lessons learned for cancer immunotherapy. *J Immunother Cancer*. 2016;4(1):1.
32. Wang R, Shao X, Zheng J, et al. A machine-learning approach to identify a prognostic cytokine signature that is associated with nivolumab clearance in patients with advanced melanoma. *Clin Pharmacol Therapeut*. 2020;107(4):978-987.
33. Turner DC, Kondic AG, Anderson KM, et al. Pembrolizumab exposure–response assessments challenged by association of cancer cachexia and catabolic clearance. *Clin Cancer Res*. 2018;24(23):5841-5849.
34. Pharmacokinetics (PK) and exposure–response (E–R) analysis of trastuzumab emtansine (T-DM1) as a single agent or in combination With pertuzumab in patients with human epidermal growth factor receptor 2 positive (HER2+) metastatic breast cancer (MBC) who have not received prior chemotherapy for their metastatic disease, Lu D., Li C., et al., ACOP 2016 Poster. American Conference of Pharmacometrics.
35. Poon V, Lu D. Performance of Cox proportional hazard models on recovering the ground truth of confounded exposure–response relationships for large-molecule oncology drugs. *CPT Pharmacometrics Syst Pharmacol*. 2022;1-16. doi:10.1002/psp4.12859
36. Berrar, D. Cross-validation. *Encyclopedia Bioinform Comput Biol*. 2018;1:542-545.
37. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079-2107.
38. Melkumova LE, Shatskikh SY. Comparing Ridge and LASSO estimators for data analysis. *Procedia Eng*. 2017;201:746-755.
39. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodology*. 2005;67(2):301-320.
40. Hüllermeier E, Fober T, Mernberger M. Inductive Bias. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, eds. *Encyclopedia of Systems Biology*. Springer; 2013. doi:10.1007/978-1-4419-9863-7_927

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu G, Lu J, Lim HS, Jin JY, Lu D. Applying interpretable machine learning workflow to evaluate exposure–response relationships for large-molecule oncology drugs. *CPT Pharmacometrics Syst Pharmacol*. 2022;11: 1614-1627. doi:10.1002/psp4.12871