Performance of Cox proportional hazard models on recovering the ground truth of confounded exposure–response relationships for large-molecule oncology drugs

Victor Poon | Dan Lu

Modeling and Simulation Group, Department of Clinical Pharmacology, Genentech, Inc., 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
A Cox proportional hazard (CoxPH) model is conventionally used to assess exposure–response (E–R), but its performance to uncover the ground truth when only one dose level of data is available has not been systematically evaluated. We established a simulation workflow to generate realistic E–R datasets to assess the performance of the CoxPH model in recovering the E–R ground truth in various scenarios, considering two potential reasons for the confounded E–R relationship. We found that at high doses, when the pharmacological effects are largely saturated, missing important confounders is the major reason for inferring false-positive E–R relationships. At low doses, when a positive E–R slope is the ground truth, either missing important confounders or mis-specifying the interactions can lead to inaccurate estimates of the E–R slope. This work constructed a simulation workflow generally applicable to clinical datasets to generate clinically relevant simulations and provide an in-depth interpretation on the E–R relationships with confounders inferred by the conventional CoxPH model.

Study Highlights
WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?
It is well recognized that large-molecule drugs have confounded exposure–response (E–R) relationships. The Cox proportional hazard (CoxPH) model is conventionally used to assess E–R, but its performance to uncover the ground truth is not systematically evaluated when only one dose level of data is available.

WHAT QUESTION DID THIS STUDY ADDRESS?
We aim to establish a simulation workflow to generate biologically plausible confounded E–R datasets that then were used to assess the capability of CoxPH models to recover the E–R ground truth in various scenarios.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?
At high-dose ranges, missing important confounders is the major reason for a false-positive E–R relationship. At low-dose ranges, either missing important...
INTRODUCTION

Large-molecule drugs such as monoclonal antibodies and antibody–drug conjugates have been developed for cancer treatment using time-to-event data (e.g., overall survival) as one of the primary clinical trial end points. To justify the dose selection for optimal efficacy, exposure–response (E–R) analysis using the Cox proportional hazard (CoxPH) model has been widely used. However, most large-molecule oncology clinical trials only test one dose level in their pivotal study to infer the E–R relationship, which is used to address the question of whether a higher dose is more efficacious. Unfortunately, such analyses are often confounded by various factors due to the complex interplay of patient characteristics, disease, drug exposure, clearance, and treatment response.1–4 The underlying biological reasons for these confounders may vary for different drugs and are not entirely clear. One plausible explanation for anticancer biologics is that higher drug clearance parallels disease severity markers associated with end-stage cancer anorexia–cachexia syndrome.4 It is also possible that the confounding could be due to other baseline covariates that relate to both patient health status and tumor burden.2 In this work, to construct biologically plausible simulation scenarios, we hypothesized two major confounding reasons that could lead to a false-positive E–R relationship where increasing exposures would not actually increase efficacy (i.e., a lack of direct causal relationship), as shown in the hypothetical causal diagram (Figure 1a). The Confounding Reason 1 is that sicker patients with poorer health status and higher tumor burden (TMDB) tend to have both shorter survival and faster clearance and thus lower pharmacokinetic (PK) exposures.2,4 The Confounding Reason 2 is that hypothetically sicker patients could have both lower maximum effects of PK exposures on decreasing the hazard rate and lower PK exposures. These reasons have posed challenges in uncovering the E–R ground truth in analyzing real clinical data. Therefore, statistical methods aiming for causal inference (e.g., propensity score matching and inverse probability weighting) were explored.5–7 However, there are limitations in these approaches. For example, the case matching method usually requires a control arm, which is sometimes absent from pivotal studies in recent years. As a result, the traditional multivariate CoxPH model is still widely used in E–R analysis.

To our knowledge, there is no systematic evaluation yet on the performance and limitations of the CoxPH model, the widely used statistical method, in uncovering the E–R ground truth when analyzing a clinical dataset at only one dose level where potential confounders may be present. A simulation approach with known E–R ground truth is used here. The objectives of this research are as follows. First, we designed a clinical trial simulation workflow to generate realistic exposure–survival datasets with known E–R ground truth and the presence of biologically plausible and clinically relevant confounders from baseline covariates across low- to high-dose levels. Second, we analyzed the simulated data and explored various scenarios where the analysis model deviates from the ground truth used in the simulation model to systematically assess in which situation(s) the multivariate CoxPH model analysis can largely adjust for confounders and recover the E–R ground truth and in which situation(s) the CoxPH model may not fully recover the ground truth. The simulated datasets contain doses ranging from low to high levels, whereas each E–R analysis was performed using the simulated data at each dose level to mimic the typical situation in the pivotal studies that only one dose level is studied.

METHODS

Part I: Establish the base-case simulation workflow to generate survival dataset with known E–R ground truth and presence of confounders

We designed realistic and biologically plausible simulation scenarios based on the covariate matrix augmentation of a real clinical trial dataset by using a novel clinical trial simulation method accounting for correlations among continuous covariates, categorical covariates, and post hoc PK parameters.8 The survival time simulation includes the effect of baseline covariates and the effect of confounders or mis-specifying the interactions can lead to inaccurate estimation of the E–R slope.

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

This work deepened our understanding on the performance of multivariate CoxPH models for recovering the E–R ground truth to guide better dose decisions and motivate further research of novel models to overcome certain methodology limitations of CoxPH models.
drug exposure on survival time in cancer patients based on two potential reasons for confounded E–R relationship (Figure 1a). For Confounding Reason 1, it is designed by using explicit mathematical functions to link the worse baseline health status-related covariates with higher hazard rate. In addition, there are implicit correlations of these covariates representing worse health status with higher PK clearance and consequently lower exposures, implemented by the conditional distribution modeling method for covariate simulation. For Confounding Reason 2, it is designed by adding explicit functions as the interaction term so that patient health status–related covariates have interactions with the maximum effect of drug trough concentration (PCMIN) on hazard rate of overall survival.
overall survival (i.e., the maximum effect of PCMIN on hazard rate of overall survival $E_{\text{max}}$ parameter in the Hill function; Figure 1a), thus sicker patients have lower $E_{\text{max}}$ values. There are also implicit correlations of these covariates with PK clearance. Examples of Confounding Reason 1 are included in all simulations, whereas examples of Confounding Reason 2 are added on top of Reason 1 by including interaction terms.

In this research, we used a historical study of trastuzumab emtansine (T-DM1) in human epidermal growth factor receptor 2-positive (HER2+) breast cancer patients (EMILIA study, ClinicalTrials.gov identifier NCT00829166) as a starting point of the simulation workflow. The steps taken to simulate and analyze survival datasets with confounded E–R relationships are outlined next. These steps can also be readily applied to other clinical trial datasets.

Step 1 is to augment a data matrix of 5000 virtual patients which includes baseline covariates and post hoc PK parameters, based on a subset of patients from the EMILIA study dataset. This subset are the patients in the T-DM1 treatment arm with T-DM1 PK parameter estimates ($N = 263$), after receiving 3.6 mg/kg doses of T-DM1 every 3 weeks [Q3W]. The dataset contains 18 baseline covariates (see summary statistics of baseline covariates in Table S1), 4 T-DM1 post hoc PK parameters, and their survival time. Among the baseline covariates, body weight (WT), TMBD, serum albumin level (ALBUM), aspartate aminotransferase (AST), and extracellular domain of human epidermal growth factor receptor 2 (ECD) were identified as covariates for PK clearance. Other covariates in the dataset include alanine aminotransferase (ALT), Eastern Cooperative Oncology Group performance status (ECOG), number of disease sites (IRFSCNT), visceral disease (VSCR), prior trastuzumab treatment (yes/no) (PHSTSET), prior anthracycline treatment (yes/no) (PRIANTHR), liver metastasis (yes/no), brain metastasis (yes/no), bone metastasis (yes/no), region, race and age. Hypothetically, among these covariates, some may only affect PK (Covariate Set 1; Figure 1a), some may only affect hazard rate of overall survival (Covariate Set 2; Figure 1a) and some may affect both (Covariate Set 3; Figure 1a, red arrows) or have interactions with the effect of PCMIN on hazard rate (Figure 1a, orange arrow). Covariate Set 3 includes the potential confounders in inferring the E–R relationship and should be considered in the multivariate CoxPH analysis. The post hoc PK parameters from a two-compartmental model include clearance (CL), central volume of distribution (V1), intercompartmental clearance (Q), and peripheral volume of distribution (V2). The post hoc PK parameters have the implicit correlations with baseline covariates, so patients with lower clearance of large-molecule drugs are generally healthier with better prognostic factors, that is, lower TMBD, ECOG, and AST and higher ALBUM who also have a longer survival. For Step 1, a novel clinical trial simulation approach by the multiple imputation by chained equations (MICE) method was used to generate an augmented covariate matrix with 5000 patients that maintains the original distribution and implicit correlations among continuous covariates, categorical covariates, and individual post hoc PK parameters as in the original EMILIA study. This method is based on conditional distribution modeling to generate virtual patients with a covariate distribution similar to the original observed distribution. It is performed using the MICE algorithm as implemented in R package mice Version 3.13.0 with R Version 3.6.3. In essence, each covariate vector in the covariate matrix is simulated/imputed by estimating their conditional probability density using predictive mean matching and multinomial logistic regression for continuous and categorical covariates, respectively, where all covariates were assumed to be missing, leveraging the original data to train each imputation model. The probability distribution and implicit correlations for each of the simulated baseline covariates and post hoc PK parameters were evaluated to ensure consistency with the EMILIA dataset (Figure 2). The post hoc PK parameters generated here reflect both the interindividual variability (IVV) explained by the baseline covariates and the remaining covariate-unexplained IVV. Using the augmented individual post hoc PK parameters for the simulated patients, Day 21 T-DM1 PCMIN at multiple dose levels (0.3, 0.6, 1.2, 2.4, 3.6, and 4.8 mg/kg) were simulated using the R package mrgsolve Version 0.11.0 based on the T-DM1 population PK model. The simulation used individual patient WT, which impacts both the dose amount and T-DM1 CL. It is worth noting that PCMIN is chosen here as a representative exposure metric in this simulation workflow, which is proportional to the given dose level.

Step 2 is to prepare for simulation of survival data by selecting the key covariates, their corresponding coefficient ($\beta$) values, and the variance–covariance matrix of the coefficients based on the EMILIA study E–R data. The CoxPH model specified (Equation 1) was applied to the original EMILIA data with control arm ($N = 649$) to obtain realistic estimates of baseline covariate relationships with hazard rate, with PCMIN included in the analysis:

$$
\lambda(t) = \lambda_0(t) \exp(\beta_1 \text{LOGTMBD} + \beta_2 \text{ECOG} + \beta_3 \text{IRFSCNT} + \beta_4 \text{LOGALKPH} + \beta_5 \text{LOGECD} + \beta_6 \text{LOGAST} + \beta_7 \text{LOGALBUM} + \beta_8 \text{PCMIN})
$$

where $\lambda_0(t)$ is the baseline hazard rate and $\beta_1$–$\beta_8$ are model coefficients. The covariates included in the CoxPH model were ECOG, IRFSCNT, ALKPH, ECD, ALBUM, AST, TMBD, and PCMIN where a log-linear relationship was
assumed for the continuous baseline covariates (ALKPH, ECD, ALBUM, AST, TMBD). Besides PCMIN, for the baseline covariates augmented from the MICE method, only seven of them were included in the CoxPH model to simplify the analysis and avoid multicollinearity. These baseline covariates were selected empirically based on knowledge of the most clinically relevant covariates in HER2+ breast cancer patients. The variance–covariance matrix of $\beta$ for the seven baseline covariates was obtained by using the vcov R function on the CoxPH R object. The forest plots of hazard ratios (HRs) for seven baseline covariates and PCMIN based on the original EMILIA dataset with control arm ($N = 649$) are shown in Figure S3, and forest plots of HRs for all covariates are shown in Figure S4.

Step 3 is to simulate the survival time based on the augmented baseline covariate matrix and PCMIN of 5000 patients at each dose level (Step 1), the hypothetical sigmoidal relationship between PCMIN and hazard rate that represents the E–R ground truth, and random sampling from the coefficient ($\beta$) variance–covariance matrix associated with the seven baseline covariates assuming multivariate-normal distribution (Step 2). For each vector of resampled $\beta$ ($\beta_1$–$\beta_7$), which represents the effects of baseline covariates on hazard rate and the $E_{max}$ for the...
Weibull distribution was used for simulating the survival verse probability method assuming a baseline hazard with ship of T-DM1 in the EMILIA study. Cumulative hazard in-

is worth noting that this simulation ground truth is assumed ≤

Based on basic pharmacological principles that drug effects doses

the multivariate CoxPH analysis (Equation 3b) would

ear the censoring indicator, where 41.2 months was the

covariates by dividing PCMIN to 5 quintiles, thus the β value associated with this categorical covariate of the PCMIN quintile (Equation 3a,b) represents the natural logarithm of HR between each higher PCMIN quintile with the lowest PCMIN quintile. The HRs relative to the patients in the lowest PCMIN quintile (Q1/Q1, Q2 = Q2, Q3, Q4, or Q5, where Q1, Q2, Q3, Q4, and Q5 represent PCMIN quintiles from low to high exposures within each dose level) is obtained by Equation (3c), and the median and 5th and 95th percentile values were summarized from 100 simulated trials. This approach avoided forcing an explicit mathematical functional form such as linear or log-linear relationship to link the continuous PCMIN variable with the hazard rate, which is actually described by the Hill equation in the simulation.

recover the E-R simulation ground truth (Equation 4). Given the goal of identifying E-R relationships, PCMIN, as the representative exposure metric, is always included in both univariate and multivariate analyses. To estimate the HRs between higher PCMIN quintiles with the lowest PCMIN quintile, the continuous covariate of PCMIN within each dose level was transformed to categorical co-

where x is the covariate matrix with dimension of 5000 rows and eight columns, β is the vector of eight rows, E_max is the maximum effect of PCMIN on hazard rate (E_max = -1), EC50 is the drug concentration reaching 50% of E_max (EC50 = 0.2 μg/ml), and γ is the Hill coefficient (γ = 2). Based on basic pharmacological principles that drug effects usually follow the Hill equation (Figure S1), we assumed unsaturated drug effects at low doses so that the hazard rate decreases with exposure increases while the drug effects are largely saturated at high doses. Thus, EC50 of 0.2 μg/ml is chosen so that >90% patients have PCMIN above EC50 for doses ≥2.4 mg/kg and thus having a relatively flat E–R relationship compared with low doses ≤1.2 mg/kg (Figure S1). It is worth noting that this simulation ground truth is assumed for this research and does not represent the E–R relationship of T-DM1 in the EMILIA study. Cumulative hazard inverse probability method assuming a baseline hazard with Weibull distribution was used for simulating the survival time (Equation 2b)¹²:

where T is the survival time, V is the survival probability that is assumed to follow a uniform distribution between 0 and 1, λ and ρ are the Weibull distribution parameters for baseline hazard, and x · β is the matrix multiplication of covariate values and the corresponding β coefficients for each patient (defined in Equation 2a). A uniform distribution between 0.5 and 41.2 months for censoring time was assumed to generate the censoring indicator, where 41.2 months was the maximum follow-up time in the EMILIA dataset. Survival data were simulated for 100 trials by random resampling of the β variance–covariance matrix at each dose level of 0.3, 0.6, 1.2, 2.4, 3.6, and 4.8 mg/kg.

Step 4 is to assess whether the simulated dataset replicates the confounded E–R relationship at a single dose level by univariate CoxPH analysis (Equation 3a) and whether the multivariate CoxPH analysis (Equation 3b) would
quintile $Q_1$ (Figure S1). The model-estimated HRs were compared with the simulation ground truth (Equation 4). Because the multivariate analysis (Equation 3b) used the formula similar to the base-case simulation (Equation 2a) in terms of the baseline covariates, it is expected to largely recover the E–R ground truth. This analysis was therefore used as a benchmark for the results in Part II of this article.

**Part II: Assessing the recovery of ground truth at various survival time simulation and CoxPH analysis scenarios**

The second part of this research is to evaluate the ability of the CoxPH model in recovering the E–R ground truth with confounders under the following three scenarios (the simulation and analysis schemes are summarized in Figure 1b): (1) leave-one-out analysis, (2) analysis using linear CoxPH model on the simulated dataset with non-linear relationships of a baseline covariate with hazard rate, and (3) analysis using linear CoxPH model without interactions on the simulated data with interactions of baseline covariates and PCMIN (Table 1). Scenario 1 aims to recreate situations where unmeasured/unknown confounding covariates are not included in the CoxPH model for E–R analysis, reflecting the Confounding Reason 1.

| Scenario | Model for survival data simulation | Multivariate model to analyze simulation data and derive HRs for inferring E-R relationship |
|----------|------------------------------------|----------------------------------------------------------------------------------|
| Leave-one-out analysis on simulated data | • Based on Equation (2) • Linear for baseline categorical covariates (ECOG, IRFSCNT) effects • Log-linear for continuous covariates (ECD, ALBUM, TMBD) effects • Hill function for PCMIN effect | • Based on Equation (3) but leave one of the baseline covariates (ECD, ALBUM, TMBD, ECOG, IRFSCNT) out in analysis model |
| Nonlinear relationship between baseline covariates and survival time | • Based on Equation (2) with nonlinearity between ECD or TMBD with survival time • Monotonic functions: (LOG(X)$^2$, SQRT (LOG (X)), EXP (LOG (X)$^3$), EXP (X/median (X)), LOG (X)$^4/(4^4 + LOG (X)^4)$, and nonmonotonic functions, $X = ECD$ or TMBD) | • Equation (3) |
| Simulated data with impact of interactions between baseline covariates and PCMIN on survival time | • Based on Equation (2) with impact of baseline covariates (ECOG, ECD, ALBUM, TMBD) on PCMIN’s effect on survival time (E$_{max}$) by multiplication of the following terms: ECOG$^{0.5}$, (LOG (ECD)/median (LOG (ECD))$^{-0.5}$, (ECOG$^{0.5}$) * (LOG (ALBUM)/median (LOG (ALBUM)))$^{10}$, (ECOG$^{0.5}$) * (LOG (ALBUM)/median (LOG (ALBUM)))$^{-10}$, and (ECOG$^{0.5}$) * (LOG (ALBUM)/median (LOG (ALBUM)))$^{-10}$ * (LOG (ECOGER)/median (LOG (ECOGER)))$^{-0.5}$ * (LOG (TMDB)/median (LOG (TMDB)))$^{2}$ | • Equation (3) |

Abbreviations: ALBUM, albumin; ECD, extracellular domain of human epidermal growth factor receptor 2; ECOG, Eastern Cooperative Oncology Group performance status; E$_{max}$, maximum effect of PCMIN on hazard rate of overall survival; E-R, exposure-response; EXP, exponential function; HR, hazard ratio; IRFSCNT, number of disease sites; LOG, natural log transformation; LOGECOGER, natural log transformation of ECD; LOGTMDB, natural log transformation of tumor burden; PCMIN, drug trough concentration; SQRT, square root; TMDB, tumor burden.

$^a$Nonmonotonic functions: $-0.0844 - 0.388 * LOGTMDB - 0.052 * LOGTMDB^2 + 0.245 * LOGTMDB^3; -0.0744 - 0.238 * LOGECOGER - 0.0974 * LOGECOGER^2 + 0.0453 * LOGECOGER^2$. 
covariates impact hazard log linearly or linearly on top of the impact of PCMIN’s effect on hazard (Equation 3b), except for Scenario 1 that certain covariate was not included in the formula. The HRs were derived (Equation 3c) and compared with the simulation ground truth (Equation 4). In each of the aforementioned scenarios (Table 1) and the base case (benchmark scenario), we quantitatively assessed the deviation of the analysis model’s inferred HRs from the E–R ground truth. First, the root mean squared error (RMSE) of the HRs was derived as follows (N = 100 for 100 simulated trials; Equation 5):

\[
\text{RMSE} = \sqrt{\frac{\sum (\text{Analysis}_{HR} - \text{Ground Truth}_{HR})^2}{N}} \tag{5}
\]

The differences between the HRs for \(Q_5/Q_1\) (\(Q_5 = Q_2, Q_3, Q_4, \) or \(Q_5, 100\) trials) and the ground truth HRs were used to compute the respective four RMSE values at each dose level. These four RMSE values were then averaged and plotted against the averaged RMSE from the benchmark scenario (the base case). Furthermore, to evaluate the steepness of the E–R relationship, in each of the 100 trials, a linear E–R slope metric was inferred by linear regression of HRs for \(Q_5/Q_1\), \(Q_3/Q_1\), \(Q_4/Q_1\), and \(Q_5/Q_1\), within each dose level. The E–R slopes were summarized by the median and 5th and 95th percentiles for 100 trials and compared with the E–R slope of the simulation ground truth at the corresponding dose level.

**RESULTS**

**Part I: Establishment of simulation framework and analysis of the E–R relationship for the base-case simulation**

Using the MICE method, the simulated covariate matrix has a similar distribution (Figure 2a) as the original dataset. There are a total of eight baseline continuous covariates, 10 baseline categorical covariates, and four PK parameters. In the simulated covariate matrix, the baseline covariates have low Pearson correlation coefficient values (absolute values <0.5) with PK clearance and the simulated PCMIN, similar to the original EMILIA dataset \((N = 263)\) (Figure 2b). This weak correlation is realistic and is potentially due to the presence of baseline covariate unexplained IIV for the post hoc clearance values. The simulated PCMIN has a moderate correlation (absolute value of the Pearson correlation coefficient >0.6) with the natural logarithm of CL, which is expected (Figure 2b).

For the base-case scenario, the univariate analysis with only the categorical PCMIN included in the CoxPH model (Equation 3a) generated unadjusted HRs and suggested a large deviation from the ground truth with a much steeper positive E–R relationship at all dose levels (red vs. blue symbols; Figure 3), suggesting that the simulated dataset created confounding effects between PCMIN and other baseline covariates and that the univariate analysis without adjusting for the confounders could not recover the E–R ground truth. In comparison, the multivariate analysis (Equation 3b) generated adjusted HRs and largely recovered the E–R ground truth, as shown by similar median values of HRs (green vs. blue symbols; Figure 3). Using the categorical PCMIN is a more accurate method to recover the E–R ground truth, as sensitivity analysis with continuous PCMIN, or natural logarithm of PCMIN, showed that adjusting for confounders by the multivariate analysis was not sufficient to properly recover the E–R ground truth (Figures S10 and S11). Using the natural logarithm of PCMIN (Figure S11), when compared with continuous PCMIN (Figure S10), did show an improvement in recovering the ground truth by the multivariate CoxPH model but was still outperformed by the categorical PCMIN model (Figure 3).

**Part II: Assessing the recovery of ground truth at various survival simulation and CoxPH analysis scenarios**

As shown in Figure 1b, three scenarios were assessed.

Scenario 1: Leave-one-out CoxPH analysis on the simulated dataset

It was found that if key baseline confounders were left out from CoxPH analysis, the ground truth of the E–R relationship may not be fully recovered for some covariates (Figure 4). When ECD was included in the simulation but not included in the analysis, there is a large deviation between the adjusted HRs from multivariate analysis (green) and ground truth (blue) (Figure 4a). Consequently, there is a large RMSE (Figure 4b) and E–R slope (Figure 4c) deviation from the fully specified benchmark analysis model when ECD is excluded in analysis, and consistently, the absolute values of the E–R slopes are larger than the ground truth, suggesting that a steeper E–R relationship is inferred across all dose levels. There are also slight deviations when ECOG is left out, as reflected from RMSE (Figure 4b). However, these deviations are not observed when leaving out the other covariates, as shown in the HR plots (Figure 4b,c, Figure S5).
Scenario 2: Simulation with nonlinear relationship of baseline covariates with hazard rate and analysis with linear/log-linear CoxPH model

As shown in Figure 5 and Figures S6 and S7 for the HR plots, when various hypothetical nonlinear relationships between baseline covariates (ECD or TMBD) and hazard rate were included in the simulation (Table 1), the multivariate CoxPH model, which assumes a linear/log-linear relationship of these baseline covariates, recovers the E–R ground truth reasonably well when using categorical PCMIN. For adjusted HRs, the RMSEs (black dots) from the mis-specified analysis models aligned well with the RMSE of the correctly specified benchmark analysis model (red dots) (Figure 5a for ECD and Figure S8 for TMBD). Similarly, the E–R slopes also line up well between the mis-specified analysis model and the ground truth (Figure 5b for ECD and Figure 5c for TMBD). These results suggest that although biologically plausible nonlinear relationships between baseline covariates and hazard rate may exist, the linear CoxPH model linking baseline covariate with hazard rate can recover the E–R ground truth in the current simulation analysis, with categorical PCMIN used. It should be noted that results may differ upon further increases in the simulation for the degree of nonlinearity between baseline covariates and hazard rate and the strength of correlation of baseline covariates with PCMIN, or when the continuous PCMIN is used (see the Discussion section).

Scenario 3: Simulation with impact of interactions between baseline covariates and PCMIN on hazard rate and analysis with linear/log-linear CoxPH model without interactions

Various biologically plausible models with interactions between PCMIN and other baseline covariates were included in the simulation models (Table 1). Figure 6a shows the adjusted HRs obtained from the linear CoxPH model on the simulated data that contains interactions between ECOG, ALBUM, ECD, TMBD, and PCMIN compared with the E–R ground truth (Equation 4). At the high-dose levels (e.g., 3.6 and 4.8 mg/kg), the adjusted HRs by the linear CoxPH model without interactions are largely aligned with the ground truth. Thus, in these cases when the efficacy is
largely saturated based on drug-specific pharmacological principles, the multivariate CoxPH analysis would conclude that a further dose increase would not increase efficacy, although the underlying complex interactions in the simulation are not included in the analysis model. However, at low doses there is a trend of overcorrection by the analysis model, inferring a shallower E–R trend quantitatively. Thus, in these cases a correct dose-selection decision related to efficacy may be compromised if the investigated dose is at the steep part of the E–R curve. As shown in Figure S9, HR plots for other interaction scenarios showed similar conclusions, especially when ECOG is included in the interaction term.

Consistent with these HR plots, when interactions involve ECOG, increased RMSE values are observed at low- to middle-dose levels, but similar to the benchmark scenario at high-dose levels (3.6 and 4.8 mg/kg) (Figure 6b). When additional interaction terms are added along with ECOG, the RMSE values tend to increase. Correspondingly, there is an underestimation of E–R slope at low doses (0.3, 0.6, 1.2 mg/kg) when a positive E–R slope is present (Figure 6c). At higher doses (≥2.4 mg/kg), when the ground truth E–R slope is close to zero, the 5th to 95th percentile includes the ground truth (Figure 6c).
DISCUSSION

For this research, we started from a real clinical trial dataset to design the various simulation scenarios. The simulation scenarios are designed to be realistic and biologically plausible, and the workflow can be applied to generate simulations based on other real datasets. The key findings are summarized for two dosing scenarios in drug development: (1) high-dose range when the efficacy is largely saturated and (2) low-dose range when the efficacy is not saturated and the doses are at the steep part of the E–R curve. These two dosing scenarios were analyzed in the context of two potential confounding reasons: sicker patients have higher hazard rate and lower PCMIN, and a
false-positive E–R relationship would be obtained if some of the covariates related to patient health status are left out from the analysis (Confounding Reason 1), and sicker patients have lower $E_{\text{max}}$ in the Hill equation of the E–R relationship and also lower PCMIN (Confounding Reason 2). The notions of “low” and “high” dose ranges here are often based on indirect evidence from in vitro and in vivo pharmacology data, which are drug specific and should be considered when we interpret the E–R results from multivariate CoxPH models.

The key learnings are summarized here. First, for the current large-molecule oncology clinical trials that still largely follow the convention to seek high doses to achieve maximal efficacy, the doses usually demonstrate saturation in pharmacological effects based on in vitro and in vivo evidence. The work here suggested that if
the E–R analysis from a “high”-dose pivotal study shows a positive E–R slope, the results should be interpreted with caution as there is a relatively large possibility that this is a false-positive, with the underlying reason more likely due to missing unknown key confounders in the CoxPH analysis that impact both exposure and response (Confounding Reason 1) and less likely due to missing the underlying interactions (Confounding Reason 2) or the nonlinear relationship of baseline confounders with hazard rate. Second, with a changing dosing paradigm in oncology and the effort of Project Optimus to reform the dose-optimization and dose-selection paradigm in oncology, there are efforts to seek optimum benefit–risk profiles by using low doses.13 If the knowledge from in vitro and in vivo pharmacology provides strong evidence that the drug effect is not saturated, the work here suggests that we are at risk of overestimating the absolute value of the E–R slope if unknown key confounders are missing or underestimating the slope if underlying interactions are not included. Third, the linear CoxPH model that misspecifies the nonlinear relationship between baseline covariates/confounders and hazard rate may not affect the recovery of E–R ground truth when categorical PCMIN is used, which do not force a linear/log-linear relationship between PCMIN and hazard rate.

It is important to emphasize that these learnings are from the analysis results based on the simulation workflow using the clinically relevant covariates for a specific trial. The simulation workflow is designed to start from a real clinical trial dataset, and in this report we chose the EMILIA study dataset. There are pros and cons for this approach. The pros are that this method generates realistic simulated datasets to closely mimic the real-world situation. The cons are that the learnings related to specific covariates may be limited to one study or similar studies and thus is not the focus here. For example, it is not entirely clear why in the leave-one-out analysis, ECD is the key covariate needed to properly adjust for confounders, but not other covariates, such as ALBUM, which is usually strongly related to cancer patients’ health status. The reason why ECD matters is not apparent from the low Pearson correlation coefficient (Figure 3b) or from the forest plot that shows the magnitude of impact of each covariate (Figure S3). It is also not clear why ignoring the ECOG-related interactions, but not other covariates, would lead to underestimating the E–R slope at lower doses. We acknowledge that the findings could be related to the specific real data chosen here to start the simulation workflow. The key learnings generated are not intended to be molecule and covariate specific but, rather, from methodology evaluation aspects. Furthermore, this workflow is easy to implement, and future applications on other trial datasets will shed light on the potential confounders specific for each trial/molecule and enhance the understanding of CoxPH performances in the datasets with different implicit characteristics.

Currently, for large-molecule oncology clinical trials where only one dose level is tested in most pivotal studies, the E–R ground truth is usually unknown due to the presence of unknown confounders that could be missed from the analysis.1–4 There are exceptions in which two doses were tested. One example is the postmarketing requirement trial of Herceptin in gastric cancer patients (Heloise study, ClinicalTrials.gov identifier NCT01450696) in which patients were randomly assigned to two dose groups, and the high-dose group did not offer survival benefits compared with the low-dose group in gastric cancer patients.14 The results implied a lack of clinically meaningful E–R relationship and a potentially false-positive E–R relationship identified from the ToGA study (ClinicalTrials.gov identifier NCT01414041)15 in which only one dose level of Herceptin is tested in gastric cancer patients, and the E–R conclusions from ToGA study7 subsequently led to the postmarketing commitment of the Heloise study.14 Based on in vitro pharmacology and clinical data, the dose given in the pivotal Herceptin trial in gastric cancer patients (ToGA study) reached the trough concentration well above the threshold values and likely saturated the underlying pharmacological effects.16,17 Using the learnings from the current research as identified in the leave-one-out analysis, it is reasonable to hypothesize that the false-positive E–R relationship in the ToGA study could potentially be due to certain unknown confounders impacting both hazard rate and exposures not being included in the E–R analysis. Similar situations may occur for other large-molecule drugs. Evidence to support this hypothesis can be drawn from the finding that the statistically significant baseline covariates identified in the population PK (popPK) model only explained a portion of the IIV of PK clearances for multiple large-molecule drugs.10,18–20 For T-DM1, the final popPK model identified covariates (WT, ECD, TMBD, ALBUM, AST, and baseline trastuzumab concentrations) only explained 44% of total IIV of the clearance.10 This explains a relatively weak correlation between the baseline covariates and PCMIN in both the real and simulated datasets (Figure 2b). Recently, high-dimensional data such as baseline cytokine signatures are found to be good predictors of nivolumab clearance.21 Thus, there is a high likelihood that some confounders affecting both PK clearance and patient survival are not available in traditional clinical datasets and therefore missing from the E–R and popPK analysis. Furthermore, for the traditional CoxPH and popPK approaches without regularization (i.e., not including a penalization term in model parameter values), the covariate selection is usually based on stepwise covariate modeling (SCM; i.e., select covariates by forward
addition and backward elimination and associated statistical significance of objective function value changes) and full covariate approaches. Thus, to avoid multicollinearity, highly correlated covariates cannot be all included in the model. Consequently, the traditional CoxPH model without regularization cannot include a comprehensive list of covariates if some of them are highly correlated and cannot fully use high-dimensional baseline covariates such as bioinformatics data, and thus having a relatively high risk of leaving important baseline confounders out. This is one potential limitation determined by the underlying methodology of CoxPH models.

Given the insights generated from methodology perspectives in the current analysis, we hypothesize that the machine-learning (ML)/deep-learning (DL) models for survival analysis\textsuperscript{22} may provide great advantages in better approximating the E–R ground truth. With regularization, ML/DL models can include a more comprehensive covariate list as well as high-dimensional patient-level biomarker and bioinformatics data to the E–R analysis, eliminate the SCM process, and consequently reduce the risk of missing important confounders and account for the complex interactions between the baseline covariates and exposure metrics with hazard rate in the models. Thus, ML/DL models could hypothetically be more accurate in recovering the ground truth and reducing the risk of underestimating the E–R slope due to not correctly specifying the interactions or overestimating the E–R slope due to missing important confounders. This remains to be evaluated in future work.

Another well-known advantage for tree-based ML/DL models is to accurately describe the complex nonlinear relationship between covariates and the prediction target, whereas the CoxPH models usually assume a linear or log-linear relationship. It was reported that nonlinear ML/DL models generated a higher C-index on the test dataset when the covariates assumed nonlinear relationships with the hazard rate,\textsuperscript{23} and ML models better approximated the ground truth of the E–R relationship when compared with traditional CoxPH models for highly nonlinear systems.\textsuperscript{24} Conceptually, we want to emphasize that there are two types of nonlinearitys considered in the current research: (1) various nonlinear functional forms (Table 1) between baseline confounders with hazard rate and (2) the Hill relationship between exposure (e.g., PCMIN) and hazard rate. In this research, we used categorical PCMIN without forcing a linear/log-linear functional form relating continuous PCMIN with hazard rate and explicit linear/log-linear functions between baseline covariates and hazard rate in the analysis model. This CoxPH model can reasonably recover the E–R ground truth, even when the relationship between baseline confounders and hazard rate is mis-specified compared with the simulation formula. Potential reasons for this finding are likely due to two aspects: the formulas of nonlinearity assumed for baseline confounders and strength of correlation implicitly implemented in the simulation between the baseline confounders and PCMIN. The simulation formulas (Table 1) assume a plausible trend of increasing hazard rate with baseline confounder values related to sicker health status (instead of cosine or concave/convex functions used in literature), which is likely well approximated by the log-linear relationship; this approach based on a real clinical dataset better reflects realistic situations where the confounding effects from baseline covariates are implicit through relatively weak correlations with PCMIN (Figure 3b) instead of explicitly specified in the simulation formulas as done in the literature.\textsuperscript{24} We consider this weak correlation to be more realistic given that only a portion of the IIV of clearance is explained by the traditionally available baseline covariates.\textsuperscript{10,18–20} Importantly, this analysis also showed that it is essential to use categorical PCMIN based on its quintiles (Equation 3a) to estimate HRs on the simulated data that assume a nonlinear relationship (Hill function) between PCMIN and hazard rate as the ground truth (Equation 2) while using continuous PCMIN or logarithm of PCMIN generates higher bias from the E–R ground truth compared with categorical PCMIN (Figure 3, Figures S10 and S11) due to forcing a linear/log-linear functional form that deviates from the Hill function. Thus, if the research question is to estimate HRs between exposure quintiles, it is important to convert PCMIN to categorical covariates in the CoxPH model. In this aspect, the tree-based ML/DL approaches may offer more flexibility in the functional form relating continuous PCMIN to hazard rate and thus may not require the transformation of continuous PCMIN to its categorical counterpart to capture the HRs of the E–R ground truth.

In summary, this article presents a systematic assessment of the capabilities of CoxPH models to recover the E–R ground truth for optimal dose selection in oncology. An overall finding is that the conventionally used CoxPH models are plausible in many cases to uncover the E–R ground truth with some exceptions. This research work deepened our understanding on the potential reasons for these exceptions, which is helpful for interpreting the E–R results for dose decision making in clinical trials. To enable a robust methodology assessment mimicking the analysis on the confounded E–R relationship, a rigorous simulation workflow was established to computationally augment a clinical trial dataset while ensuring the implicit characteristics/correlations of the dataset remained intact and then subsequently simulate survival times for the augmented dataset by explicitly incorporating a sigmoidal dependence of hazard rate on the representative exposure variable, which represented the E–R ground truth. This workflow can simulate realistic E–R datasets.
with the presence of biologically plausible and clinically relevant baseline confounders. Furthermore, by assessing the performance of CoxPH models in recovering the E–R ground truth, we found that multivariate CoxPH models largely capture the ground truth in many cases, however, some overestimations or underestimations of E–R slope were identified in certain low-dose range (steep E–R relationship) or high-dose range (shallow E–R relationship) scenarios. In real-world cases, the E–R ground truth is often unknown especially when only one dose level is tested. This work suggests that an interpretation of the E–R trend from one-dose level clinical trials of large-molecule oncology drugs by CoxPH models would benefit from considering the in vitro and in vivo pharmacological evidence, being aware of potential reasons that may bias the estimated E–R slope from the ground truth, for a more accurate dose decision. Lastly, this work identified the methodology limitations of the CoxPH models that are likely to cause deviations from the ground truth and trigger further research in assessing novel ML/DL models for their capability to overcome these potential drawbacks.

**AUTHOR CONTRIBUTIONS**
V.P. and D.L. wrote the manuscript. D.L. designed the research. V.P. and D.L. performed the research and analyzed the data.

**ACKNOWLEDGMENTS**
We acknowledge Tong Lu, Phyllis Chan, Kenta Yoshida, Jin Yan Jin, and Pascal Chanu from Genentech and Jonathan French and Daniel Polhamus from Metrum Research Group for providing insightful technical discussions, comprehensive review, and strong support for this research; Liang Zhao from the Food and Drug Administration for providing simulation R codes of his publication related to the inverse probability method; and Anshin Biosolutions for editing support.

**FUNDING INFORMATION**
This work is supported by Roche/Genentech.

**CONFLICT OF INTEREST**
Victor Poon and Dan Lu are Roche/Genentech employees. Victor Poon and Dan Lu own Roche/Genentech stocks.

**ORCID**
Victor Poon  
https://orcid.org/0000-0002-4215-2198  
Dan Lu  
https://orcid.org/0000-0003-4531-3599  

**REFERENCES**
1. 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;4:72.

2. 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 Ther*. 2020;108:1156-1170.

3. Kawakatsu S, Bruno R, Kagedal M, et al. Confounding factors in exposure–response analyses and mitigation strategies for monoclonal antibodies in oncology. *Br J Clin Pharmacol*. 2021;87:2493-2501.

4. 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;24:5841-5849.

5. 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;83:2767-2777.

6. 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;80:1079-1090.

7. 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;53:160-166.

8. Smania G, Jonsson EN. Conditional distribution modeling as an alternative method for covariates simulation: comparison with joint multivariate normal and bootstrap techniques. *CPT Pharmacometrics Syst Pharmacol*. 2021;10:330-339.

9. Verma S, Miles D, Gianni L, et al. Trastuzumab emtansine for HER2-positive advanced breast cancer. *N Engl J Med*. 2012;367:1783-1791.

10. Lu D, Girish S, Gao Y, et al. Population pharmacokinetics of trastuzumab emtansine (T-DM1), a HER2-targeted antibody-drug conjugate, in patients with HER2-positive metastatic breast cancer: clinical implications of the effect of covariates. *Cancer Chemother Pharmacol*. 2014;74:399-410.

11. Elmokadem A, Riggs MM, Baron KT. Quantitative systems pharmacology and physiologically-based pharmacokinetic modeling with mrgsolve: a hands-on tutorial. *CPT Pharmacometrics Syst Pharmacol*. 2019;8:883-893.

12. Bender R, Augustin T, Blettner M. Generating survival times to simulate cox proportional hazards models. *Stat Med*. 2005;24:1713-1723.

13. Shah M, Rahman A, Theoret MR, Pazdur R. The drug-dosing conundrum in oncology - when less is more. *N Engl J Med*. 2021;385:1445-1447.

14. Shah MA, Xu RH, Bang YJ, et al. HELIOSE: 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;35:2558-2567.

15. Cosson VF, Ng VW, Lehle M, Lum BL. Population pharmacokinetics and exposure–response analyses of trastuzumab in patients with advanced gastric or gastroesophageal junction cancer. *Cancer Chemother Pharmacol*. 2014;73:737-747.

16. Baselga J, Carbonell X, Castaneda-Soto NJ, et al. Phase II study of efficacy, safety, and pharmacokinetics of trastuzumab monotherapy administered on a 3-weekly schedule. *J Clin Oncol*. 2005;23:2162-2171.
17. Pegram M, Hsu S, Lewis G, et al. Inhibitory effects of combinations of HER-2/neu antibody and chemotherapeutic agents used for treatment of human breast cancers. Oncogene. 1999;18:2241-2251.

18. Bruno R, Washington CB, Lu JF, Lieberman G, Banken L, Klein P. Population pharmacokinetics of trastuzumab in patients with HER2+ metastatic breast cancer. Cancer Chemother Pharmacol. 2005;56:361-369.

19. Garg A, Quartino A, Li J, et al. Population pharmacokinetic and covariate analysis of pertuzumab, a HER2-targeted monoclonal antibody, and evaluation of a fixed, non-weight-based dose in patients with a variety of solid tumors. Cancer Chemother Pharmacol. 2014;74:819-829.

20. Lu D, Lu T, Gibiansky L, et al. Integrated two-analyte population pharmacokinetic model of Polatuzumab Vedotin in patients with non-Hodgkin lymphoma. CPT Pharmacometrics Syst Pharmacol. 2020;9:48-59.

21. 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 Ther. 2020;107:978-987.

22. Wang P, Li Y, Reddy CK. Machine learning for survival analysis: a survey. ACM Comput Surv. 2019;51:Article 110.

23. Gong X, Hu M, Zhao L. Big data toolsets to Pharmacometrics: application of machine learning for time-to-event analysis. Clin Transl Sci. 2018;11:305-311.

24. 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.

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: 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;11:1511-1526. doi:10.1002/psp4.12859