ARTICLE

Application of Machine Learning for Tumor Growth Inhibition – Overall Survival Modeling Platform

Phyllis Chan[1,*,†], Xiaofei Zhou[1,2,4,†], Nina Wang[1], Qi Liu[1], René Bruno[3] and Jin Y. Jin[1]

Machine learning (ML) was used to leverage tumor growth inhibition (TGI) metrics to characterize the relationship with overall survival (OS) as a novel approach and to compare with traditional TGI-OS modeling methods. Historical dataset from a phase III non-small cell lung cancer study (OAK, atezolizumab vs. docetaxel, *N* = 668) was used. ML methods support the validity of TGI metrics in predicting OS. With lasso, the best model with TGI metrics outperforms the best model without TGI metrics. Boosting was the best linear ML method for this dataset with reduced estimation bias and lowest Brier score, suggesting better prediction accuracy. Random forest did not outperform linear ML methods despite hyperparameter optimization. Kernel machine was marginally the best nonlinear ML method for this dataset and uncovered nonlinear and interaction effects. Nonlinear ML may improve prediction by capturing nonlinear effects and covariate interactions, but its predictive performance and value need further evaluation with larger datasets.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Traditional modeling methods could be limited for data mining purpose, and machine learning (ML) has the potential as a new tool to improve prediction in model-informed drug development.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Four ML methods of lasso, boosting, random forest, and kernel machine were applied to investigate their predictive performance for comparison with each other and with traditional tumor growth inhibition-overall survival (TGI-OS) modeling, to explore the interactions between predictor variables, and to incorporate nonlinear relationships for the prediction of OS.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ ML methods support the validity of TGI metrics in predicting OS and can serve as an alternative tool to improve prediction by capturing nonlinear effects and covariate interactions using a dataset with small dimensionality, but their predictive performance and value need further evaluation with larger datasets.

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

☑ Leveraging model-based drug development through ML to improve treatment response prediction and identification of the best predictors has great potential to further advancements in precision medicine.

The major success of artificial intelligence (AI) in medicine so far has been in the augmentation of disease diagnosis based on imaging data, and AI is increasingly applied to improve prediction in drug development, such as candidate selection and analysis of genomics data. In terms of the use of AI in clinical pharmacology, a series of studies were published in the late 1990s and early 2000s investigating the usage of AI, in particular, neural network for predicting pharmacokinetic concentrations of mainly antibiotics and immunosuppressants, to guide medication dosage based on patient characteristics.[1–8] In more recent years, AI is experiencing a resurgence in drug development due to the availability of increased computing power, as the strength of AI over traditional analytical methodologies is mainly its ability to deal with large, not easily interpretable data, from which multiple comparisons and interactions can be made from the number of possible covariates in clinical studies. The ultimate aim of AI approaches is model prediction, and AI algorithms have the ability to incorporate many important but collinear variables simultaneously.[5] Additionally, certain AI methodologies can also investigate nonlinear and higher-order interactions without assuming a structure for covariates, which is not possible with the majority of standard approaches of analyzing data from clinical trials.[5] Therefore, not surprisingly, AI/machine learning (ML) was listed as a potential new tool that can expand the horizon of

[†] These authors contributed equally to this work.

The results of the analysis were previously presented in part at the American Society of Clinical Pharmacology and Therapeutics meeting in March 2019.

[1] Clinical Pharmacology, Roche/Genentech, South San Francisco, California, USA; [2] Formerly of Department of Statistics, The Ohio State University, Columbus, Ohio, USA; [3] Clinical Pharmacology, Roche/Genentech, Marseille, France; [4] Present address: Biostatistics, Novartis, East Hanover, New Jersey, USA. *Correspondence: Phyllis Chan (chan.hui-min@gene.com)

Received: June 23, 2020; accepted: October 28, 2020. doi:10.1002/psp4.12576

model-informed drug development by the Office of Clinical Pharmacology at the US Food and Drug Administration (FDA).[9] ML is a subfield of AI, and ML algorithms are data-mining tools and techniques used for pattern recognition based on models for the prediction of new data.[10]

A major challenge in oncology drug development is the ability to predict clinical response to anticancer drugs according to the individual characteristics and drug exposure metrics, and traditionally Cox proportional hazard (Cox PH) or parametric survival regression (also known as accelerated failure time (AFT)) model is used for overall survival (OS) prediction. Alternatively, the tumor growth inhibition-OS (TGI-OS) approach, in which multivariate parametric survival regression models with baseline patient characteristics and model-based TGI metrics, have been shown to be predictive of OS outcome in a number of solid tumor types as well as in hematologic diseases.[11,12] Recently, a TGI-OS model was developed using data from patients with non-small cell lung cancer treated with atezolizumab or docetaxel in a phase II study (POPLAR) using on-treatment tumor growth rate constant estimated using time profiles of sum of longest diameters (SLDs) as the TGI metrics. The model performance was further evaluated by leveraging early tumor kinetic data before OS maturation to predict long-term survival in a phase III study (OAK).[13]

However, traditional modeling methods could be limited for data mining purpose because, in clinical studies, many of the variables may be correlated, causing the covariate selection process to force some important variables out of the model. Furthermore, the relationship between a covariate and OS could be nonlinear, thus creating complex interactions among variables and the possibility of overfitting.[14] These characteristics may compromise the predictive power of the model and lead to less accurate survival prediction. ML can be used as an alternative methodology to circumvent these potential issues by using ML processes for pattern recognition and prediction purposes.

The aims of our study were to apply four well-established ML methods of lasso, boosting, random forest, and kernel machine to data from the OAK study to investigate the predictive performance of each method, to explore the interactions between predictor variables, and to incorporate nonlinear relationships for the prediction of OS. The predictive performance for the classification of the patient outcomes were compared among the four ML methods and to the TGI-OS results by using OAK dataset with the same prespecified covariates. Leveraging model-based drug development through ML to improve treatment response prediction and identification of the best predictors based on a large list of prognostic and predictive factors has great potential to further advancements in precision medicine.

METHODS

The OAK study protocol has been previously described.[15] Briefly, patients with previously treated non-small cell lung cancer in the randomized, open-label, phase III clinical trial were randomly assigned (1:1) to receive i.v. atezolizumab (1,200 mg) or docetaxel (75 mg/m^2) once every 3 weeks. Co-primary end points of the OAK study were OS in the intention-to-treat and programmed death-ligand 1 (PD-L1)-expression population tumor cells (TCs)1/2/3 or immune cells (ICs)1/2/3. In the primary population, 425 patients were randomly assigned to receive atezolizumab and 425 patients were assigned to receive docetaxel. The study was conducted in accordance with the Declaration of Helsinki after approval by institutional review boards or independent ethics committees. All patients provided written informed consent. The methods of the TGI-OS model have been previously published.[13]

The following baseline patient characteristics were tested to explain variability in OS: age, sex, body weight (BWT), Eastern Cooperative Oncology Group (ECOG) performance status, smoking status (never smokers vs. other), total protein, albumin, alkaline phosphatase, aspartate aminotransferase, lactate dehydrogenase (LDH), serum creatinine, estimated glomerular filtration rate, tumor size estimated as the sum of longest diameter of the target lesions (baseline SLD), number of metastatic sites (number of metastatic sites as a continuous variable, number of metastatic sites as a categorical variable with > 2 as a single category, or number of metastatic sites as a categorical variable with > 4 as a single category), histology (non-squamous vs. squamous), years since metastasis, number of prior chemotherapy regimens for advanced disease (second-line vs. third-line), and PD-L1 expression (binary variables of TC123IC123, TC23IC23, TC3, or IC3 categorized as yes/no), as well as treatment-related variables of the treatment arm, area under the concentration-time curve of atezolizumab after cycle 1 (AUC1), and TGI metrics. PD-L1, which is expressed on TCs and tumor-infiltrating ICs on a wide variety of cancer expressions and is targeted by atezolizumab, was scored by immunohistochemistry as percentage of PD-L1–expressing TC (TC3 ≥ 50%, TC2 ≥ 5%, and < 50%, TC1 ≥ 1% and < 5%, and TC0 < 1%) and as percentage of PD-L1–expressing tumor area for IC (IC3 ≥ 10%, IC2 ≥ 5%, and < 10%, IC1 ≥ 1% and < 5%, and IC0 < 1%).[16]

Four ML methods were implemented in the current study for covariate selection and log(hazard) or log(OS) predictions with censoring, and the same 27 following baseline patient characteristics from the TGI-OS analysis dataset were tested as covariates. For covariate selection using each of the ML methods, three variations were examined: excluding all covariates from the model, retaining only the important covariates based on the tuning parameter (lambda), and including all 27 prespecified covariates; this was to explore whether models with covariates would perform similarly as a model that does not use any covariate information and whether the inclusion of less significant covariates contribute to more accurate prediction. A schematic of the data used for model development is illustrated in **Figure S1**. Brier scores, which is a calculation of the mean square of error of the test set and can range from 0 to 1, compared the observed and model-predicted binary outcome of OS at various timepoints,[17] with lower scores associated with more accurate model prediction and better model performance.[18] The lasso method was applied for variable selection using the Cox PH model[19] or the AFT model,[20] and the linear effects were on the log of hazard for Cox lasso, whereas the linear effects were on the log of survival time for AFT lasso.

Like lasso, Cox boosting is another linear supervised ML approach implemented, due to the linear base learners in the analysis.[21] Similar to boosting, random forest is another decision tree-based method for building classification and regression prediction models and is able to incorporate many predictor variables simultaneously without compromising the accuracy of the risk prediction. The most important variables are identified based on their contribution to the predictive accuracy of the model and are those that most frequently result in early splitting of the decision trees.[22] Survival random forest constructs decorrelated trees to evaluate effects of the covariates on the hazard and can detect complex nonlinear relationship between predictor variables and outcome.[23,24] Several hyperparameters were tuned for random forest, including the number of variables randomly selected as candidates for splitting a node and the number of random splits to consider for each candidate splitting variable. Another ML method able to capture the nonlinear effects of the covariates on the log of survival time was kernel machine, also known as support vector machine, and can also incorporate the interaction between the vcovariates.[25] Last, graphical display of an input important measure from one-dimensional (marginal) covariate analysis, pairwise interaction matrix of multidimensional kernel machine analysis, and three-dimensional surface visualization of the kernel machine estimates were generated.

The prediction models based on the ML approaches were trained using internal validation from bootstrap sampling with replacement of the same size as the original dataset and a cross-validation step of 50 (B = 50 samples). The ML models were implemented using functions in publically available R (version 3.6.1) packages, which are listed in the **Supplementary Material**.

RESULTS

Analysis dataset

Data from 850 patients randomized to the study was used for the analysis and 751 patients (88%) were identified to have at least one post-baseline tumor size assessment for the estimation of TGI metrics, with a total of 668 patients having data for all 27 covariates. In the analysis dataset, 334 patients were treated with atezolizumab, and 334 patients were treated with docetaxel. The median survival time of the 668 patients included in the analysis was 361 days, with a minimum and maximum of 24 and 823 days, respectively.

TGI-OS

The results of the TGI-OS model were published previously.[13] In summary, the following individual TGI metrics were estimated using a previously published longitudinal TGI model[13]: TGI model parameters of tumor growth rate constant (KG) and tumor shrinkage rate constant (KS), and time to tumor regrowth (TTG). From the univariate Cox PH analysis, log(KG) was the most significant factor followed by TTG and early change in tumor size and a number of baseline prognostic factors and treatment (atezolizumab vs. docetaxel). Among parametric models, the log-normal distribution had the best likelihood of describing the OS distribution. After backward elimination under the log-normal AFT model, log(KG), number of metastatic sites and albumin level remained the only significant independent covariates in the final OS model developed based on POPLAR data and externally validated using OAK data.

Machine learning

The top predictors selected by lasso, boosting, random forest, and kernel machine are shown in **Table 1** and were chosen based on variable importance methods, such as solution path, permutation accuracy importance, and sensitivity analysis. The number of top predictors was determined from cross-validation. For the four ML methods investigated, two TGI metrics, log(KG) and TTG, are consistently the most informative predictors of OS, in keeping with results from the TGI-OS modeling approach. Of the 24 baseline covariates, baseline SLD, LDH, and albumin were the most discriminative predictors.

The solution paths using lasso, which is one of the two ML methods investigating linear effects, are displayed in **Figure S2** and **Figure S3**, and log(KG) was ranked the highest among 27 covariates in variable importance, and 5 other covariates were chosen by lasso to be the top predictors: baseline SLD, ECOG, albumin, LDH, and TTG. With Cox boosting, the same six covariates as lasso were selected as the top predictors, with two of the six top predictors being TGI metrics.

The predictive performance among various covariate models using the Cox PH method by lasso is shown in **Figure 1** and AFT method by lasso in **Figure 2**. **Table 2** shows the Brier score averaged over different times for all the subjects for each ML method, and the Brier scores were used to assess the predictive accuracy of each ML model and to compare prediction performance among different models. Although models without covariates performed the worst, there was little difference between the models that incorporated all 27 covariates (also known as a feature in ML) and those that only had the top predictors in the model, indicating that the inclusion of all 27 features was not needed for optimal prediction, as the information from the lower ranked variables were either irrelevant or redundant and reduce prediction accuracy by increasing the noise. Additional sensitivity analysis revealed that when TGI metrics were not included in the list for feature selection, AUC1 was a significant predictor by all methods. With the incorporation of TGI metrics, AUC1 was no longer significant, possibly because

Table 1 Top predictors of overall survival selected by different analysis methods

Analysis methods	Model selected predictors for OS (in order of significance)
TGI-OS	logKG, Metsites, ALBU
Lasso	logKG, BSLD, ECOG, TTG, ALBU, LDH
Boosting	logKG, BSLD, ECOG, TTG, ALBU, LDH
Random forest	logKG, TTG, BSLD, LDH, ALBU
Kernel machine	logKG, BSLD, TTG, logKS

ALBU, albumin; BSLD, baseline tumor size; ECOG, Eastern Cooperative Oncology Group performance status; LDH, lactate dehydrogenase; logKG, log of tumor growth rate constant; logKS, log of tumor shrinkage rate constant; Metsites, number of metastatic sites; OS, overall survival; TGI, tumor growth inhibition; TTG, time to tumor regrowth.

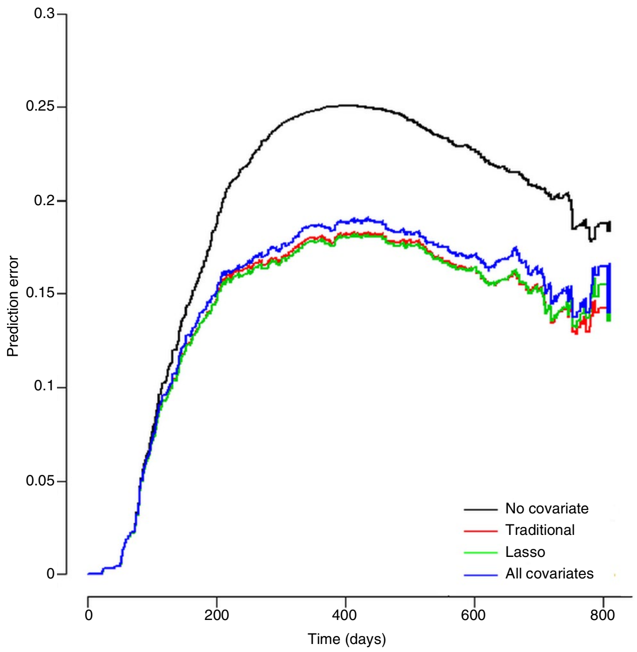

Figure 1 Brier scores of different covariate models based on lasso Cox proportional hazard (Cox PH).

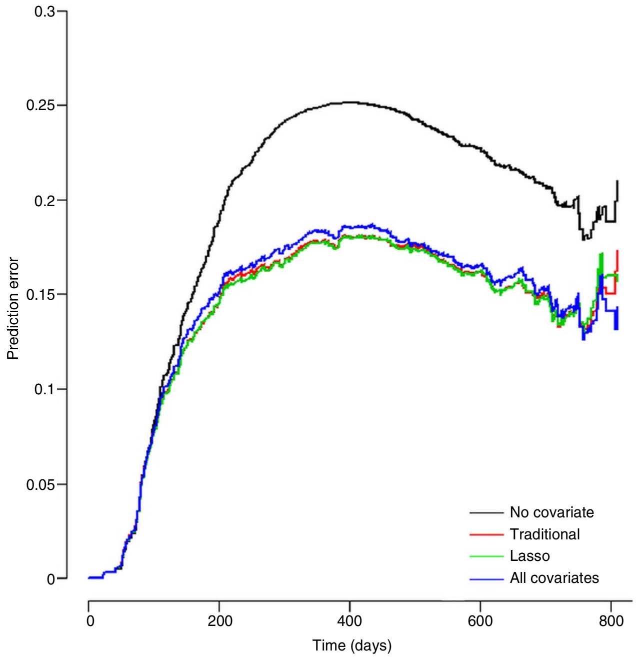

Figure 2 Brier scores of different covariate models based on lasso accelerated failure time.

exposure information is already covered by TGI metrics that are meant to capture treatment effect,[13] and the best model for each ML method with TGI metrics outperformed the best model without TGI metrics. When comparing the prediction performance of models that only included top predictors selected by lasso vs. models with statistically significant covariates selected by TGI-OS through backward elimination, the models developed using the lasso approach provided marginally more accurate prediction (Brier scores of 0.141 vs. 0.142 in the Cox PH framework, and 0.139 vs. 0.140 in the traditional AFT framework; **Table 2**).

A second linear effect ML method tested was Cox boosting, which returned slightly better prediction accuracy compared with the best lasso model (Brier score of 0.137 vs. 0.139) because its coefficient estimates are less biased. In addition to the terms of variable selection, the model with the top predictors chosen by Cox boosting performed better than the traditional AFT model with covariates selected by TGI-OS (Brier score of 0.137 vs. 0.139; **Figure S4**).

ML methods based on nonlinear effects were also investigated. The model with covariates selected using random forest had a Brier score of 0.142, which is higher than the Cox PH and AFT models with covariates chosen by lasso or Cox boosting. Using the kernel machine method, potential collinearity and nonlinearity among the predefined list of 27 covariates were examined in further detail, and nonlinear and interaction effects were detected. A two-dimensional plot shows the nonlinear and marginal effects of log(KG), log(KS), and TTG on the log of survival time (**Figure 3**), whereas the three-dimensional surface plot illustrates the complex interactive effects of log(KG) and log(KS) on the log of survival time (**Figure 4**). A pairwise comparison of the interaction effects among the 27 covariates were also conducted using kernel machine and indicates the covariate pairs that have strong influence in the prediction model as shown in the squares with greater grayscale intensity (**Figure 5**).

DISCUSSION

There has been much renewed interest in recent years in AI/ML approaches due to rapid growth in computer hardware that make parallel processing possible and numerically intensive computations faster. ML is a comprehensive model development tool that allows selection of predictors among all available parameters without subjective preselection and is able to reveal complex hierarchical relationships between covariates, which enables more flexible data modeling.[26] ML can handle datasets with high dimensionality for the identification of the best predictors from a large list of covariates, as a means of hypothesis generation without prior assumptions. Because of this capability, a large number of variables that might not reach statistical significance and would be excluded using traditional methods, but nevertheless could cumulatively predict outcome, can be incorporated into an ML model. Increasingly, ML has been explored and accepted as a complementary analytical tool in model-informed drug development.[27–30]

The reported use of ML for survival outcome in oncology has been few and typically limited to using high-dimensional imaging or gene expression data as predictors,[31–33] and recently ML was applied to identify the association between baseline biomarker signature and nivolumab clearance, which is linked to survival outcome.[34] An evaluation by the FDA of simulated data showed that ML-based methods outperformed Cox model in survival prediction performance and in identifying the preset influential variables, and the authors of that analysis concluded that ML-based methods provide a powerful tool for time-to-event analysis, due

Table 2 **Brier scores of models with top predictor selection based on different methods**

Frame work	No covariate	Covariates from TGI-OS	Lasso	All covariates
Cox	0.19	0.142	0.141	0.145
AFT	0.19	0.14	0.139	0.142

AFT, accelerated failure time; OS, overall survival; TGI, time to tumor regrowth.

to their capacity for high-dimensional data and better performance when the predictor variables assume nonlinear relationships in the hazard function.[18]

The current exploratory analysis was based on the atezolizumab study OAK, using historical patient-level data typically available from a phase III clinical trial, applying four well-established ML methods to investigate feature selection and the predictive performance of each method and to compare strengths and weaknesses among the four ML methods and the previously developed TGI-OS model applied to the same data. The dataset would not be considered high-dimensional for ML analysis, but the application of ML approaches to the data could nevertheless provide comparative insights. The TGI metrics used in the models were previously derived based on tumor size data[13] and have been shown to capture treatment effect and could be considered as a predictive biomarker for OS.[11–13,35]

The four ML methods evaluated in the analysis included covariate/feature selection and prediction, where feature selection builds a screening model by removing the redundant features. In traditional modeling approaches, the rule of parsimony dictates a goal of finding the fewest number of variables that can accurately predict outcome for model stability, even though important information may be lost if variables that are not sufficiently predictive are excluded from the model.[36] ML approaches are able to use information from more covariates even in high-dimensional datasets, because correlated variable are not forced out of the equation. Nevertheless, caution should be taken to avoid overfitting when a dataset with small dimensionality is used for model development even when ML approaches have been applied. In the current analysis, the findings from the ML models confirmed the importance of TGI metrics in predicting OS treatment outcome, and the results showed similar survival outcome prediction performance compared with the models with covariates chosen by the TGI-OS approach. One ML model in particular, namely Cox boosting, had slightly better accuracy than the traditional approach or other ML methods in terms of Brier scores. If AFT boosting also had been implemented, it is likely that this parametric method would perform better than the semiparametric Cox boosting method for linear effects because the parametric model captures the data better than semiparametric, given parametric model's distribution is specified correctly.

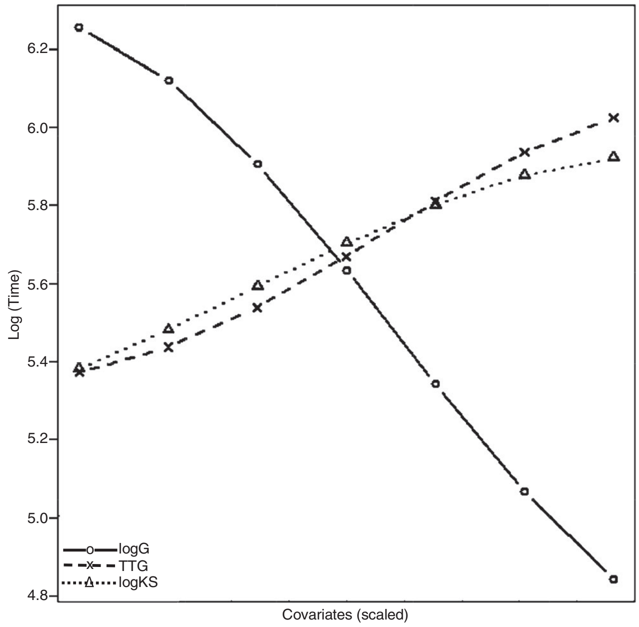

Figure 3 Marginal effects of log(KG), log(KS), and TTG on the log of survival time using kernel machine. KG = 1/week, KS = 1/week, TTG = week. KG, tumor growth rate constant; KS, tumor shrinkage rate constant; TTG, time to tumor regrowth.

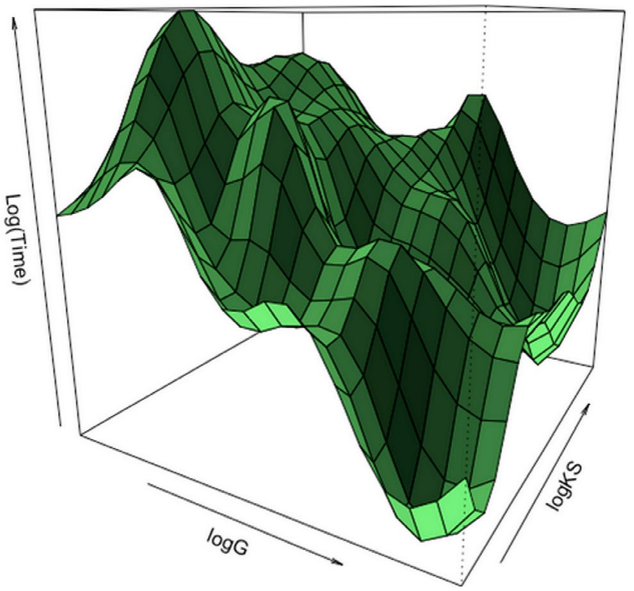

Figure 4 Interactive effects of log(KG) and log(KS) on the log of survival time using kernel machine. KG, tumor growth rate constant; KS, tumor shrinkage rate constant.

The predictive performance between the models were evaluated using Brier score, which is mainly a relative measure for comparison purposes.[18] Additional investigations regarding the performance of the models at the individual level were not conducted due to the small dimensionality nature of the analysis dataset. The predictive error from Brier score was calculated as the weighted average that corresponds to the probabilities of not being censored and depends on the covariates and is observed to be highest at ~ 200–500 days.

Because traditional approaches for OS prediction do not accommodate nonlinear relationships between predictor variables and outcomes, yet many variables are thought

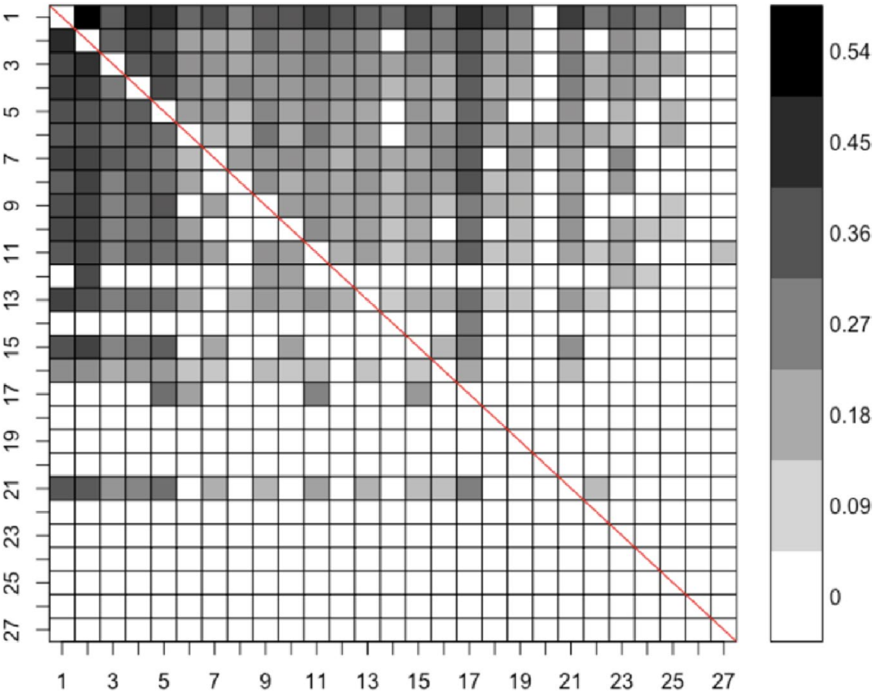

Figure 5 Pairwise effects of the 27 covariates estimated by kernel machine. 1 = log(KG), 2 = log(KS), 3 = AUC1, 4 = TTG, 5 = LDH, 6 = ALBU, 7 = Histology, 8 = AST, 9 = ALP, 10 = Age, 11 = BSLD, 12 = SCr, 13 = Smoking status, 14 = TPRO, 15 = ECOG, 16 = Sex, 17 = BWT, 18 = Metsite2, 19 = Metsite4, 20 = Metsites, 21 = YSM, 22 = eGFR, 23 = TC123IC123, 24 = TC23IC23, 25 = number of prior chemotherapy regimens for advanced disease (second-line vs. third-line) = Line3, 26 = TC3, 27 = IC3. Red line is the line of identity. The intensity of the grayscale corresponds to the strength of the interaction effect. ALBU, albumin; ALP, alkaline phosphatase; AST, aspartate aminotransferase; AUC, area under the concentration-time curve; BSLD, baseline sum of longest diameter; BWT, body weight; ECOG, Eastern Cooperative Oncology Group; eGFR, estimated glomerular filtration rate; KG, tumor growth rate constant; KS, tumor shrinkage rate constant; LDH, lactate dehydrogenase; SCr, serum creatinine; TTG, time to tumor regrowth; TPRO, total protein; YSM, years since metastasis.

to exhibit such relationships, it was hypothesized that incorporating the nonlinear relationships through nonlinear effects, the ML models might provide better predictive performance compared with the linear ML or traditional models. However, in the current analysis, random forest did not outperform the linear models, possibly due to sample size limitation. This is because the limited sample size (668 patients with complete data) is unlikely to be large enough for the random forest step functions to approximate the true nonlinear and linear effects from the covariates. However, a systematic exploration of potential interactions between predictor variable through the use of kernel machine offered interesting insights to the interactions that were not revealed previously using the TGI-OS approach. For example, even though increased KG has been previously associated with decreased survival time, the three-dimensional surface plot in **Figure 4** showed the relationship may not be linear. The pairwise comparison of the interaction effects displays the relative predictive impact of covariate pairs. The asymmetric characteristic of the matrix indicates imbalanced contribution between each variable in some covariate pairs.[37] For example, the contributions of the interaction effect of log(KG) given BWT is clearly more impactful than the interaction effect of BWT given log(KG), as shown by the different intensity in the grayscale between the square in row 1 column 17 and the square in row 17 column 1. Reasons for the unequal contributions could be investigated in future explorations.

Much more data are usually required for ML approaches, but a typical clinical trial dataset is not particularly large to train an ML model. In addition, analysis of clinical data typically initiates with a preselected list for covariate screening and therefore does not fit the data mining application of ML.[38] Hence, additional work is needed to compare the predictive performance of the ML models and the important predictors identified by the ML approaches in larger datasets, possibly by combining data from multiple clinical trials or leveraging alternative data sources, such as longitudinal tumor dynamic data instead of TGI metrics or real-world data, as well as incorporating other covariates, although a physiological understanding of the correlation between covariates and the survival outcome should still remain, as the lack of interpretability has been one of the major criticisms of using ML approaches. Another commonly cited drawback of ML approaches is generalizability, because ML models cannot extrapolate beyond feature space of the training data, whereas the traditional empirical models used in pharmacometrics and quantitative systems pharmacology in particular can be extrapolated to a certain extent.[39] Therefore, the dataset used to train ML models must be relevant to the population being studied. Due to the relatively small dimensionality of the analysis dataset by ML standards, cross validation using bootstrap[40] was performed in the current analysis instead of using k-fold cross validation, however, our investigation was designed to be exploratory, and an external validation with data from additional studies should be performed in the future to confirm the validity of the conclusions.

The current analysis demonstrated that ML methods support the validity of TGI metrics in predicting OS and that ML techniques have great potential to overcome certain limitations of traditional modeling approaches by their ability to incorporate large numbers of predictor variables without compromising the accuracy of the prediction. However, they did not provide significantly more accurate predictions than traditional methods with the current analysis dataset. ML methods can serve as an alternative tool, in addition to TGI-OS modeling, to improve prediction by capturing nonlinear effects and covariate interactions, but their predictive performance and value need further evaluation. To our knowledge, this is the first work that utilizes ML approaches to leverage TGI metrics to characterize the relationship with OS. Additional analyses and examples, especially by utilizing the newer algorithms, are required to definitively conclude that ML applications can improve model-informed drug development and have the potential to advance precision medicine in drug discovery, development, and use.

Supporting Information. Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website (www.psp-journal.com).

Acknowledgment. The authors acknowledge the editorial assistance that was provided by Anshin Biosolutions, Corp.

Funding. This work was supported by Genentech, Inc.

Conflict of Interest. P.C., N.W., Q.L., R.B., and J.Y.J. are employees and stockholders of Genentech, Inc. X.Z. was a graduate student summer intern at Genentech while the analysis was conducted and is currently an employee of Novartis.

Author Contributions. P.C., X.Z., R.B., and J.Y.J. wrote the manuscript. P.C., X.Z., N.W., Q.L., R.B., and J.Y.J. designed the research. P.C., X.Z., N.W., Q.L., R.B., and J.Y.J. performed the research. X.Z. analyzed the data.

1. Brier, M.E., Zurada, J.M. & Aronoff, G.R. Neural network predicted peak and trough gentamicin concentrations. *Pharm. Res.* **12**, 406–412 (1995).
2. Camps-Valls, G. *et al.* Prediction of cyclosporine dosage in patients after kidney transplantation using neural networks. *IEEE Trans. Biomed. Eng.* **50**, 442–448 (2003).
3. Chen, H.Y., Chen, T.C., Min, D.I., Fischer, G.W. & Wu, Y.M. Prediction of tacrolimus blood levels by using the neural network with genetic algorithm in liver transplantation patients. *Ther. Drug Monit.* **21**, 50–56 (1999).
4. Corrigan, B.W., Mayo, P.R. & Jamali, F. Application of a neural network for gentamicin concentration prediction in a general hospital population. *Ther. Drug Monit.* **19**, 25–28 (1997).
5. Hall, R.G.N., Pasipanodya, J.G., Meek, C., Leff, R.D., Swancutt, M. & Gumbo, T. Fractal geometry-based decrease in trimethoprim-sulfamethoxazole concentrations in overweight and obese people. *CPT Pharmacometrics Syst. Pharmacol.* **5**, 674–681 (2016).
6. Smith, B.P. & Brier, M.E. Statistical approach to neural network model building for gentamicin peak predictions. *J. Pharm. Sci.* **85**, 65–69 (1996).
7. Solomon, I. *et al.* Applying an artificial neural network to warfarin maintenance dose prediction. *Isr. Med. Assoc. J.* **6**, 732–735 (2004).
8. Yamamura, S. *et al.* Application of artificial neural network modelling to identify severely ill patients whose aminoglycoside concentrations are likely to fall below therapeutic concentrations. *J. Clin. Pharm. Ther.* **28**, 425–432 (2003).
9. Wang, Y., Zhu, H., Madabushi, R., Liu, Q., Huang, S.M. & Zineh, I. Model-informed drug development: current US regulatory practice and future considerations. *Clin. Pharmacol. Ther.* **105**, 899–911 (2019).
10. Ma, H., Xu, C.F., Shen, Z., Yu, C.H. & Li, Y.M. Application of machine learning techniques for clinical predictive modeling: A cross-sectional study on nonalcoholic fatty liver disease in China. *Biomed Res. Int.* **2018**, 4304376 (2018).
11. Bruno, R. *et al.* Progress and opportunities to advance clinical cancer therapeutics using tumor dynamic models. *Clin. Cancer Res.* **26**, 1787–1795 (2020).
12. Bruno, R., Mercier, F. & Claret, L. Evaluation of tumor size response metrics to predict survival in oncology clinical trials. *Clin. Pharmacol. Ther.* **95**, 386–393 (2014).
13. Claret, L. *et al.* A model of overall survival predicts treatment outcomes with atezolizumab versus chemotherapy in non-small cell lung cancer based on early tumor kinetics. *Clin. Cancer Res.* **24**, 3292–3298 (2018).
14. Ambale-Venkatesh, B. *et al.* Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ. Res.* **121**, 1092–1011 (2017).
15. Rittmeyer, A. *et al.* Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* **389**, 255–265 (2017).
16. Fehrenbacher, L. *et al.* Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *Lancet* **387**, 1837–1846 (2016).
17. Brier, G.W. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* **78**, 1–3 (1950).
18. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **63**, 938–939 (2010), author reply 939.
19. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997).
20. Khan, M.H.R. & Shaw, J.E.H. Variable selection for survival data with a class of adaptive elastic net techniques. *Stat. Comput.* **26**, 725–741 (2016).
21. Tutz, G. & Binder, H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics* **62**, 961–971 (2006).
22. Konerman, M.A., Zhang, Y., Zhu, J., Higgins, P.D., Lok, A.S. & Waljee, A.K. Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology* **61**, 1832–1841 (2015).
23. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
24. Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
25. Cortez, P. Data mining with neural networks and support vector machines using the r/rminer tool. In: Advances in data mining. Applications and theoretical aspects (ed. Perner, P.) 572-583 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010).
26. Yip, T.C. *et al.* Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. *Aliment. Pharmacol. Ther.* **46**, 447–456 (2017).
27. Chaturvedula, A., Calad-Thomson, S., Liu, C., Sale, M., Gattu, N. & Goyal, N. Artificial intelligence and pharmacometrics: time to embrace, capitalize, and advance? *CPT Pharmacometrics Syst. Pharmacol.* **8**, 440–443 (2019).
28. Hutchinson, L. *et al.* Models and machines: how deep learning will take clinical pharmacology to the next level. *CPT Pharmacometrics Syst. Pharmacol.* **8**, 131–134 (2019).
29. Koch, G., Pfister, M., Daunhawer, I., Wilbaux, M., Wellmann, S. & Vogt, J.E. Pharmacometrics and machine learning partner to advance clinical data analysis. *Clin. Pharmacol. Ther.* **107**, 926–933 (2020).
30. Liu, Q. *et al.* Application of machine learning in drug development and regulation: current status and future potential. *Clin. Pharmacol. Ther.* **107**, 726–729 (2020).
31. Gupta, P. *et al.* Prediction of colon cancer stages and survival period with machine learning approach. *Cancers (Basel)* **11**, 2007 (2019).
32. Johansen, R. *et al.* Predicting survival and early clinical response to primary chemotherapy for patients with locally advanced breast cancer using DCE-MRI. *J. Magn. Reson. Imaging* **29**, 1300–1307 (2009).
33. Shipp, M.A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**, 68–74 (2002).
34. Wang, R. *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.* **107**, 978–987 (2020).
35. Hopkins, A.M. *et al.* Early tumor shrinkage identifies long-term disease control and survival in patients with lung cancer treated with atezolizumab. *J. Immunother Cancer.* **8**, e000500 (2020).
36. Hutmacher, M.M. & Kowalski, K.G. Covariate selection in pharmacometric analyses: a review of methods. *Br. J. Clin. Pharmacol.* **79**, 132–147 (2015).
37. Cortez, P. & Embrechtz, M.J. Using sensitivity analysis and visualization techniques to open black box data mining models. *Inform. Sciences.* **225**, 1–17 (2013).
38. Vidyasagar, M. Identifying predictive features in drug response using machine learning: opportunities and challenges. *Annu. Rev. Pharmacol. Toxicol.* **55**, 15–34 (2015).

39. Benson, N. Quantitative systems pharmacology and empirical models: friends or foes? *CPT Pharmacometrics Syst. Pharmacol.* **8**, 135–137 (2019).
40. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Stat. Assoc.* **92**, 548–560 (1997).