Received: 28 July 2020 | Revised: 28 October 2020 | Accepted: 7 December 2020

DOI: 10.1002/psp4.12594

ARTICLE

Early predictions of response and survival from a tumor dynamics model in patients with recurrent, metastatic head and neck squamous cell carcinoma treated with immunotherapy

Ignacio González-García[1] | Vadryn Pierre[2,3] | Vincent F. S. Dubois[1] | Nassim Morsli[4] | Stuart Spencer[5] | Paul G. Baverel[1,6] | Helen Moore[7]

[1]Clinical Pharmacology & Safety Sciences, AstraZeneca, Cambridge, UK

[2]Clinical Pharmacology & Safety Sciences, AstraZeneca, Gaithersburg, Maryland, USA

[3]Clinical Pharmacology, EMD Serono, Billerica, Massachusetts, USA

[4]Clinical Development, AstraZeneca, Cambridge, UK

[5]Clinical Statistics, AstraZeneca, Cambridge, UK

[6]Clinical Pharmacology, Hoffmann-La Roche Research and Early Development, Roche Innovation Center, Basel, Switzerland

[7]Applied Mathematics, Applied BioMath, Concord, Massachusetts, USA

Correspondence
Ignacio González-García, Aaron Klug Building, Granta Park, Cambridge, CB21 6GH, UK.
Email: ignacio.gonzalez@astrazeneca.com

Funding information
The EAGLE, CONDOR, and HAWK studies were sponsored by AstraZeneca and studies 1108 and 11 were sponsored by MedImmune, a former subsidiary of AstraZeneca.

Abstract

We developed and evaluated a method for making early predictions of best overall response (BOR) and overall survival at 6 months (OS6) in patients with cancer treated with immunotherapy. This method combines machine learning with modeling of longitudinal tumor size data. We applied our method to data from durvalumab-exposed patients with recurrent/metastatic head and neck cancer. A fivefold cross-validation was used for model selection. Independent trial data, with various degrees of data truncation, were used for model validation. Mean classification error rates (90% confidence intervals [CIs]) from cross-validation were 5.99% (90% CI 2.98%–7.50%) for BOR and 19.8% (90% CI 15.8%–39.3%) for OS6. During model validation, the area under the receiver operating characteristic curves was preserved for BOR (0.97, 0.97, and 0.94) and OS6 (0.85, 0.84, and 0.82) at 24, 18, and 12 weeks, respectively. These results suggest our method predicts trial outcomes accurately from early data and could be used to aid drug development.

INTRODUCTION

Over the past decade, the increased survival and improvement on quality of life observed in some patients receiving immuno-oncology (IO) therapy have transformed the landscape of oncology care and drug development.[1] However, not all patients respond or benefit from treatment with IO therapy.[2,3] In addition, some patients who receive IO therapy experience what is termed pseudoprogression: their tumor sizes initially appear to increase, but later decrease.[4–7] The opportunity for long-term benefit could be missed if a patient experiencing pseudoprogression is removed from IO therapy. Further, some patients appear to experience hyperprogression: their tumors grow faster than expected, without any subsequent reduction throughout the remaining treatment course.[7–9] These patients may benefit from early discontinuation of IO therapy and a switch to an alternative treatment. For these and other reasons, accurate prediction of patient response to IO therapy is both important and challenging.

Various tumor dynamic models have been used to characterize drug effects on tumor size and to identify prognostic and predictive factors for overall survival for chemotherapy, targeted agents and recently IO therapy.[10–13] The relationship between early tumor dynamics and survival has been explored[11,14,15] with 8-week tumor shrinkage associated with longer survival for chemotherapy or targeted therapies. For IO therapy, tumor size change at 12 weeks demonstrated predictive value of survival.[15] However, in both cases, these tumor size-derived metrics do not provide additional benefit over the traditional Response Evaluation Criteria in Solid Tumors (RECIST)-based criteria for immunotherapies.[16] Other research evaluated the entire longitudinal time course of tumor size data and the use of joint modeling to determine the best predictors of survival.[17] Whereas providing good accuracy in predicting an independent external clinical trial, the complexity of the method and the use of long-term data present significant hurdles for its scalability and implementation in clinical practice or to guide decision making in drug development.

In this work, we propose a simple mathematical framework for early prediction of patients' best overall response (BOR) and overall survival at 6 months (OS6). Our method uses nonlinear mixed-effects (NLMEs) modeling of longitudinal tumor size data from patients on IO therapy, coupled with a machine learning classification algorithm. This method provides an early prediction of outcomes, both at individual and at study levels. Machine learning approaches have already demonstrated value in the field of pharmacometrics[18,19] but our method specifically addresses a problem of early response prediction based on clinical data.

In this proof-of-concept work, we assessed the performance of our method with data from patients with recurrent and metastatic (r/m) head and neck squamous cell carcinoma (HNSCC) treated with durvalumab (Imfinzi), an anti-PD-L1 monoclonal antibody, alone or in combination with tremelimumab, an anti-CTLA4 monoclonal antibody. Durvalumab monotherapy is currently approved for multiple indications, including urothelial carcinoma, non-small cell lung carcinoma, and extensive-stage small-cell lung carcinoma. We performed cross-validation using data from four clinical trials to select the model. We then used data from a separate randomized, blinded, and controlled phase III clinical trial to independently assess the method's predictive performance.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Previous tumor dynamics models have demonstrated that early tumor size metrics can be correlated with clinical outcomes in patients treated with chemotherapy. Predicting response to immune-oncology (IO) therapy has been challenging due to complexities, such as pseudoprogression and hyperprogression.

WHAT QUESTION DID THE STUDY ADDRESS?

Is it possible to predict the response of patients receiving IO therapies using only early data?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

A novel approach combining mixed effects modeling of tumor longitudinal data and supervised machine learning are able to predict clinical outcomes, such as best overall response and survival of an independent trial with good accuracy based on only 12 weeks of tumor size assessments. The accuracy of the method in this challenging setting is promising for its predictive potential for other cancer types and therapies.

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

Early prediction of responses of patients with cancer to various therapies could guide clinical development decision and help optimize therapy for individual patients.

METHODS

Study design, data, and patient population

Study design

This analysis is based on data from five clinical studies investigating durvalumab (Imfinzi) alone or in combination with tremelimumab (see Table 1) in patients with a diagnosis of r/m HNSCC. The training dataset was comprised of 401 patients enrolled in one of the following 4 phase I or phase II studies: study 1108 (NCT01693562), study 11 (NCT02262741), HAWK (NCT02207530), and CONDOR (NCT02319044). All studies were conducted in compliance with the Declaration of Helsinki and the US Food and Drug Administration Guidelines for Good Clinical Practice.

TABLE 1 Covariate distribution at baseline, split by study, and data set (training and validation)

	Training				Validation
Variable, mean (SD)	HAWK	CONDOR	Study_11	Study_1108	EAGLE
n	111	195	41	54	482
Age, years	57.5 (12.2)	60.3 (9.53)	60.7 (11.5)	58.9 (11.2)	59.5 (9.59)
Albumin, [g/L	38.1 (4.97)	38.6 (4.91)	39.2 (6.24)	37.6 (4.67)	39.1 (4.95)
ALP, U/L	96.0 (52.3)	99.9 (101)	102 (78.8)	105 (62.2)	123 (100)
ALT, U/L	19.5 (10.8)	18.9 (11.3)	19.1 (9.11)	22.5 (12.0)	19.1 (15.5)
GGT, U/L	49.8 (55.7)	60.4 (114)	54.9 (64.6)	61.5 (92.1)	65 (108)
HB, g/L	117 (14.0)	117 (17.5)	121 (18.6)	117 (18.4)	119 (16.6)
IC, %	15.2 (18.2)	15.8 (19.3)	25.6 (21.3)	21.1 (20.3)	17.4 (21.6)
NEUT, 10^9/L	6.28 (4.05)	6.08 (3.36)	6.43 (3.9)	5.56 (3.81)	6.24 (3.88)
NLR	8.33 (7.97)	7.32 (5.52)	9.28 (6.79)	8.33 (7.35)	7.38 (8.49)
TC, %	62.6 (26.6)	3.98 (6.24)	34.2 (34.4)	22.5 (29.9)	19.6 (28.2)
SLD, mm	68.8 (38.9)	73.3 (40.2)	75.2 (51.7)	76.3 (45.0)	64.8 (42.2)

	Training					Validation
Variable, *N* (%)	Label	HAWK	CONDOR	Study 11	Study 1108	EAGLE
HPV status	Negative	64 (58%)	126 (65%)	6 (15%)	21 (39%)	255 (53%)
	Positive	34 (31%)	58 (30%)	16 (39%)	24 (44%)	78 (16%)
	Unknown	13 (12%)	11 (6%)	19 (46%)	9 (17%)	149 (31%)
Disease stage	Locally advanced	39 (35%)	69 (35%)	39 (95%)	37 (69%)	190 (39%)
	Metastatic	72 (65%)	126 (65%)	1 (2%)	11 (20%)	279 (58%)
	Unknown	0 (0%)	0 (0%)	4 (2%)	6 (11%)	13 (3%)
Prior platinum-based therapy	No	0 (0%)	0 (0%)	39 (95%)	9 (17%)	26 (5%)
	Yes	111 (100%)	195 (100%)	2 (5%)	45 (83%)	456 (95%)
ECOG	0	33 (30%)	60 (31%)	13 (32%)	17 (31%)	125 (26%)
	1	77 (69%)	135 (69%)	28 (68%)	36 (67%)	357 (74%)
	Unknown	1 (1%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)
Sex	Male	32 (29%)	32 (16%)	9 (22%)	7 (13%)	75 (16%)
	Female	79 (71%)	163 (84%)	32 (78%)	47 (87%)	407 (84%)
Smoking status	Current	9 (8%)	29 (15%)	1 (2%)	5 (9%)	86 (18%)
	Former	59 (53%)	138 (71%)	22 (54%)	31 (57%)	295 (61%)
	Never	43 (39%)	28 (14%)	18 (44%)	18 (33%)	101 (21%)

Abbreviations: ALP, alkaline phosphatase; ALT, alanine transferase; ECOG, Eastern Cooperative Oncology Group; GGT, gamma glutamyl transferase; HB, hemoglobin; HPV, human papillomavirus; IC, immune cell PD-L1 expression, expressed in percentage staining; NEUT, neutrophil count; NLR, neutrophil to lymphocyte ratio; SLD, sum of longest diameters of the tumor size at baseline; TC, tumor cell PD-L1 expression, expressed in percentage staining.

The test dataset was comprised of data from 482 patients in a phase III study (EAGLE trial, NCT02369874) and was used to validate the proposed method. The similarity in observed clinical outcomes and distributions of patients' characteristics across trials and within trial arms allowed us to pool the data from all durvalumab-exposed patients to simplify the analysis. Consequently, patients treated with durvalumab monotherapy or with durvalumab in combination with tremelimumab were included in the analysis. Similar dose levels of durvalumab were used in the 5 studies: doses of 10 mg/kg q2w i.v. or 20 mg/kg q4w i.v. When used in combination with durvalumab, tremelimumab was dosed at 1 mg/kg q4w for 4 cycles. Two sensitivity analyses were conducted to test the impact of the assumptions made in the pooling strategy (see the Supplementary Materials).

The model predictive performance was evaluated using the area under the curve (AUC) of the receiver-operator curve (ROC) as well as the AUC of the precision and recall curve (PRC) to assess the potential impact of imbalance between the two classes, responders and nonresponders, on the classification.[20]

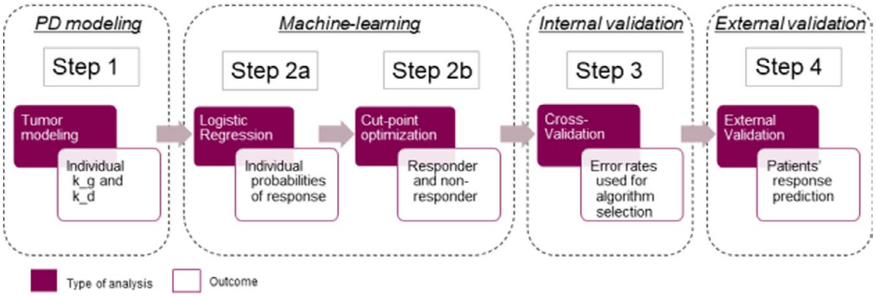

FIGURE 1 Schematic of the classification method

Tumor size data

Tumor assessments were performed using computed tomography or magnetic resonance imaging with lesion size determined according to RECIST criteria.[16] The sum of longest diameters (SLDs) of up to five target lesions was computed from each scan. All patients with at least one baseline scan with measurable disease were included; 78% of these patients also had at least one postbaseline assessment with measurable disease. The median number of measurements postbaseline was 3 (range 1–15) in the training dataset. The intended duration of treatment as defined by protocol was typically 1 year with the possibility for re-treatment for some patients. The longest follow-up was almost 30 months on treatment.

Covariate data and missingness

Available covariates in the datasets included; baseline demographics (age, sex, and performance status), kidney function (albumin level [ALB], creatinine clearance, and serum creatinine), metabolic marker (lactate dehydrogenase), cell counts (neutrophils [NEUTs], neutrophil-lymphocyte ratio [NLR]), as well as postbaseline presence of antibody-drug antibodies. The pharmacokinetics of durvalumab were not considered in the covariate analysis because most patients were dosed to achieve close to complete target suppression of PD-L1 in the periphery. Additionally, association of durvalumab exposure levels with longitudinal data of ALB and tumor size can result in confounding bias, as demonstrated by Baverel et al.[10] Potential covariates with missingness not exceeding 10% were considered, and missing values were imputed by the median and the most prevalent values for the continuous and categorical features, respectively.

Clinical outcomes and responders' definition

BOR was defined as the best response a patient had during treatment, but prior to starting any subsequent cancer therapy and up to and including RECIST progression or the last evaluable assessment in the absence of RECIST progression. BOR responders were defined as patients with a BOR of complete response or partial response, whereas BOR nonresponders were defined as those with stable disease, progressive disease, or not evaluable.

For OS6, a responder was any patient who was alive at or beyond 6 months, a nonresponder was any patient who died or was censored prior to 6 months. The 6-month landmark time ensured that a sufficient number of responders and nonresponders were in each category to test the predictive performance of the classification algorithm. Hence, dropout was factored into the OS6 responder metric to minimize informative missingness bias and avoid the need for model-based imputation methods.[21,22]

Method development and validation

A simple description of the method in four steps is provided below and a visual workflow is shown in Figure 1 with more details in the Supplementary Materials. We started with all available individual patient tumor size data in the training dataset. In step 1, we fit the following mathematical model to these SLD values over time:

$$Y = Y_0 \bullet e^{-k\_d \bullet t} + Y_1 \bullet (1 - e^{-k\_g \bullet t}) \tag{1}$$

where Y is the model value for SLD, and Y_0, Y_1, k_d (decrease), and k_g (growth) are fixed effects.[23]

This base model was fit using NLME modeling (NONMEM version 7.3; ICON, Ellicott City, MD), with an additive residual error model, and with random effects on the fixed effects parameters k_d and k_g. The resulting estimated individual rate constants k_d and k_g were used as inputs for the next step. This flexible mathematical model is able to capture tumor size dynamics specific to patients receiving IO therapy. In particular, this model can capture pseudoprogression as well as tumor sizes that approach a steady-state value that is finite and non-zero. Many previously published tumor dynamic models of responses to chemotherapy cannot exhibit these behaviors.[11,13,23–25]

In step 2a, we applied logistic regression (R version 3.5.1 or higher; R Software Foundation) to the step 1 values of k_d and k_g to obtain individual probabilities of response. We note that in exploratory analyses of other data sets (not shown), applying logistic regression to the natural logarithms (*ln*) of k_d and k_g worked well. Covariates that showed a statistically significant improvement to the model and were then tested during the multivariable analysis were NLR, ALB, Eastern Cooperative Oncology Group (ECOG), and NEUT count.

In step 2b, we chose the probability threshold that separated predicted responders from predicted nonresponders with the smallest total error when comparing these predictions to the known patient responses. In step 3, clinical markers were tested as covariates in the logistic regression model by applying fivefold cross-validation to the full time-course data. Covariate selection was performed to identify the algorithm resulting in the highest accuracy in response classification with the most parsimonious number of parameters.

In step 4, we used the test dataset to externally validate the model's ability to predict patient response from early data. Specifically, we tested the predictive performance of the model on the test data set at 24, 18, and 12 weeks post-treatment initiation as well as on the full treatment duration. NLME modeling was used with fixed-effect ("population") values frozen to those found in steps 1 and 3, and individual values estimated on the truncated data. Predictions based on those k_d and k_g values were compared with known patient outcomes to calculate the error rates.

RESULTS

Training and test datasets

The longitudinal tumor size data of the training data set used in the primary analysis are graphically illustrated in Figure 2a for all 4 trials with color-coded BOR. The corresponding Kaplan-Meier curves of overall survival are similar across the four studies (Figure 2c). Corresponding plots for the test dataset (EAGLE) are provided in Figure 2b and d. The percentages of responders in the training and test datasets were, respectively, 10% and 18% for BOR and 23% and 29% for OS6. Covariate summary statistics at baseline for each trial are shown in Table 1. No significant differences in the demographics and relevant covariates data were noted between the training and test data sets. The population those enrolled was typical of a r/m HNSCC second-line patient pool except for study 11, which had 95% of patients with no prior platinum-based therapy and a locally advanced stage of disease. Most patients were men, former or current smokers, and the patient population had a mean age approximating 60 years. Baseline ALB levels ranged from 37.6 to 39.2 g/L across trials with NEUT and NLR at baseline ~ 6 billion cells/L and 7–9, respectively. Approximately two-thirds of the patients had a baseline ECOG performance status score of 1. Around 30%–40% of patients in the training dataset had a known positive human papillomavirus (HPV) status in the training data set trials compared with 16% in the EAGLE dataset, but a generally higher proportion had unknown HPV status in the latter trial. Training and test data sets used in the sensitivity analyses of the methodology did not lead to significant imbalance in structure or data content.

Population tumor modeling of training data set

The tumor dynamics model adequately described the training and test data sets, and the model parameters were estimated with reasonable precision (Table S2). Typical values (relative standard error) of k_d, k_g, and Y1 were 0.0226 week^{-1} (46.6%), 0.00551 week^{-1} (32.1%), and 37.9 mm (27.0%), respectively, with a 17.4% residual error estimate. The final model consisted of random effects on all parameters with no covariance structure. Variability estimates were large (54.1% to 183%), reflective of the heterogeneity of tumor size patterns observed following immunotherapy. No formal goodness of fit and model qualification was undertaken besides visual assessment of the individual prediction versus observed data because residual- and population-based diagnostics were expected to be skewed by informative censoring or dropout.[21,26] Eight representative patient tumor profiles and their associated model fits (including k_g and k_d) are shown in Figure 3a. These show a well-fitting model to patterns of initial growth followed by decrease (ID = 183 and 405), tumor size growth (ID = 185), tumor size decrease (ID = 401, 431, and 440), and tumor size decrease followed by regrowth (ID = 414 and 428).

Machine learning classification

Individual parameter values obtained by fitting the tumor dynamics model to the training data sets were used as inputs of the machine learning algorithm for response classification. No clinical markers were identified as statistically significant besides k_d and k_g for the prediction of BOR, whereas baseline ALB was associated with OS6 response ($p < 0.01$). Following 5-fold cross-validation of the training dataset, ALB remained a significant predictor for the OS6 classification. Indeed, ALB, k_d, and k_g were found to be the best predictors of OS6 response, giving the lowest classification error rate (mean of 19.8% with 90% confidence interval 15.8%–39.3%) with fewest degrees of freedom.

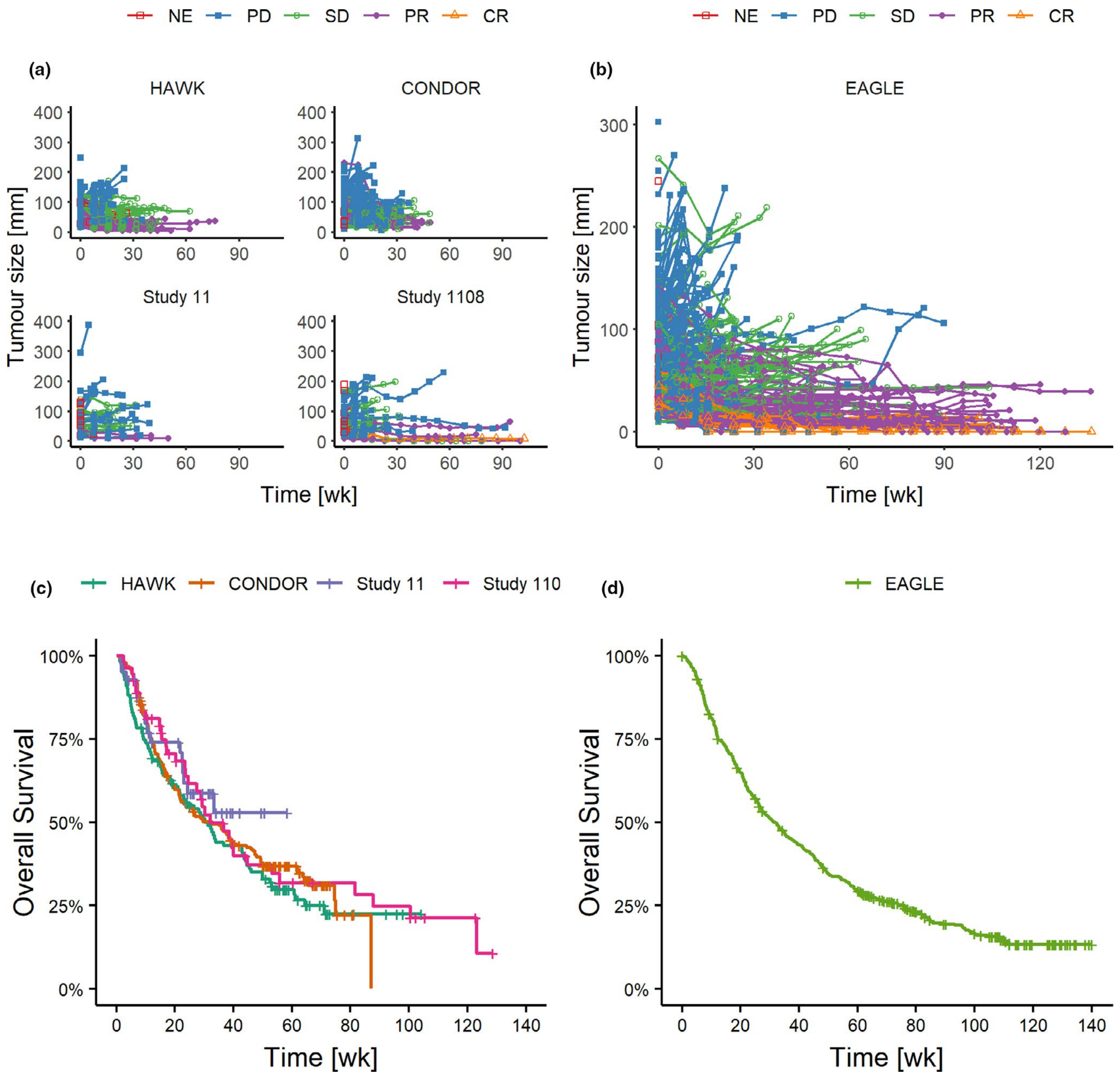

FIGURE 2 Top panels. Longitudinal tumor data of durvalumab-treated patients with recurrent and metastatic head and neck squamous cell carcinoma used for training the model (panel a, 4 studies) and for validation (panel b, one confirmatory study). Best overall response outcome classifications used are as follows: responders are patients with complete response (CR) or partial response (PR); nonresponders are patients with stable disease (SD), progressive disease (PD), or not evaluable (NE). Bottom panels. Kaplan-Meier plots of overall survival (OS) for all four clinical trials used in the training dataset stratified by study (panel c), and validation dataset OS time profile (panel d). A 6-month landmark time was used in the analysis to dichotomize OS (responder = alive and still in the trial at 6 months; nonresponder = not in the trial at 6 months), denoted as overall survival at 6 months (OS6)

A scatterplot of k_d and k_g in logarithmic scale is provided in Figure 3b for BOR observed responses, and in Figure 3c for OS6 observed responses. The classification error rates and contingency tables comparing predictions and observed responses were computed during the fivefold cross-validation of the training dataset. Error rates of the cross-validation are displayed in Figure 4 alongside AUC ROC for both the primary analysis and for the 2 sensitivity analyses. Overall, no major inflation of the error rates or AUC ROC was noted, suggesting that neither the inclusion of patients on combination therapy (sensitivity analysis 1) nor a larger training dataset sample size (sensitivity analysis 2) have significant impact on the classification accuracy (see the Supplementary Materials).

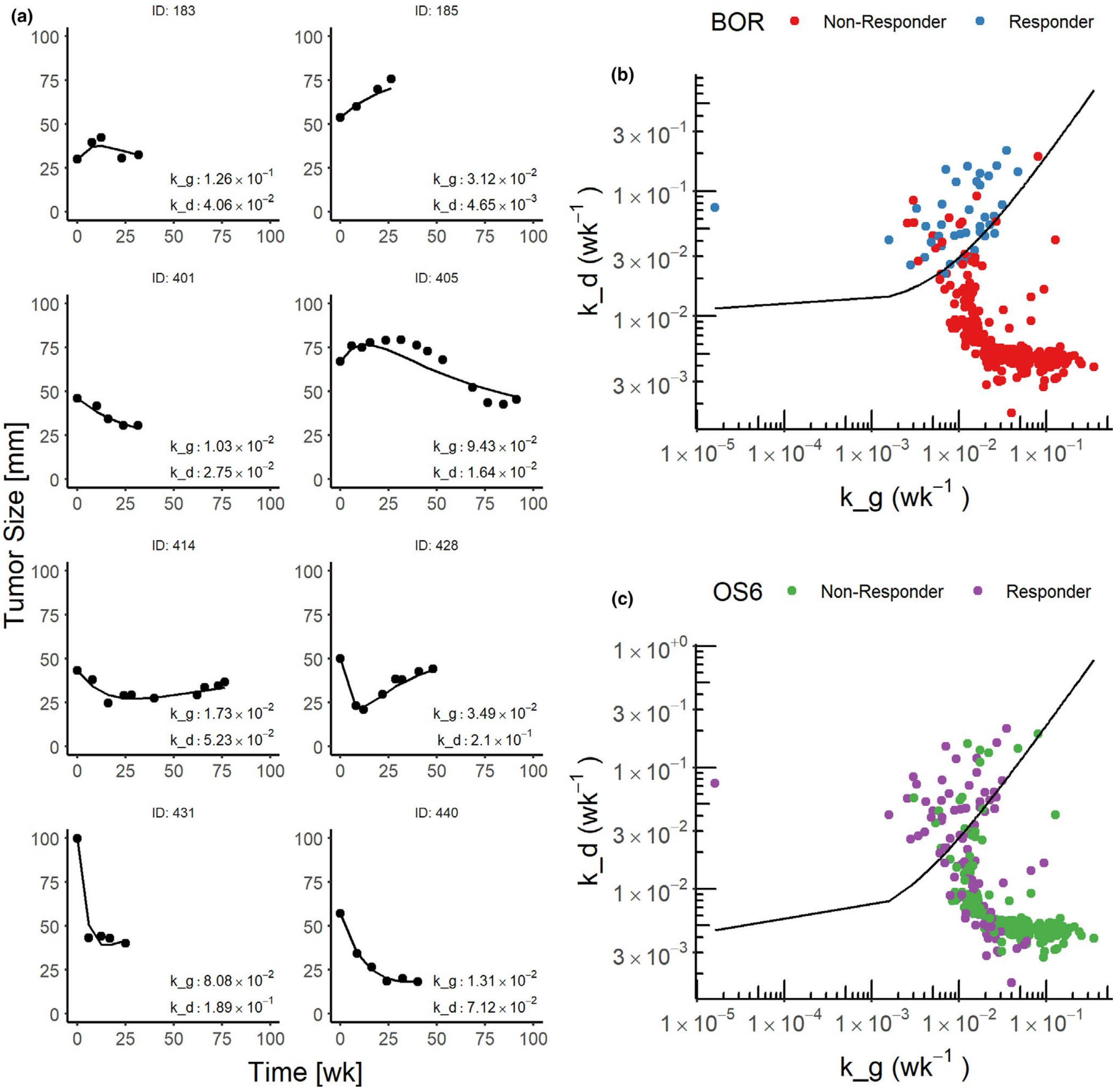

FIGURE 3 (a) Scatterplot of sum of longest diameter of target lesions time-course data (plain circles) and nonlinear mixed-effect tumor model fit (solid line) for eight representative individuals selected from the training data. Individual parameter estimates k_d and k_g, expressed in week^{-1}, are provided. (b) Scatterplot of the individual parameter estimates (k_d vs. k_g) and categorization by best overall response (BOR) in the training dataset. (c) Scatterplot of (k_d vs. k_g) and categorization by overall survival at 6 months (OS6) response in the training dataset. In (b) and (c), the solid black demarcation curve represents the machine learning output separating responders (above the curve) from nonresponders (below the curve). Observed clinical outcomes of BOR and OS6 are color-coded as indicated in the legends

External validation

During the external validation of EAGLE trial data, we used NLME modeling to fit the tumor dynamics model to the test data set, whereas the individual parameters for the training data and the overall population estimates were frozen. The longitudinal tumor size profiles of patients were well-characterized by the NLME fit, and this provided a set of individual parameters (k_d and k_g) for each patient in the test data set. Individual plots of tumor size longitudinal data and model predictions are shown in Figure 5a for data truncation at 24, 18, and 12 weeks. ROC curves showed good predictive performance irrespective of data truncation (Figure 5b). The BOR classification yielded very small decreases of AUC ROC from 0.968 (24 weeks) to 0.971 (18 weeks) to 0.937 (12 weeks). For OS6, the AUC ROC values were 0.847,

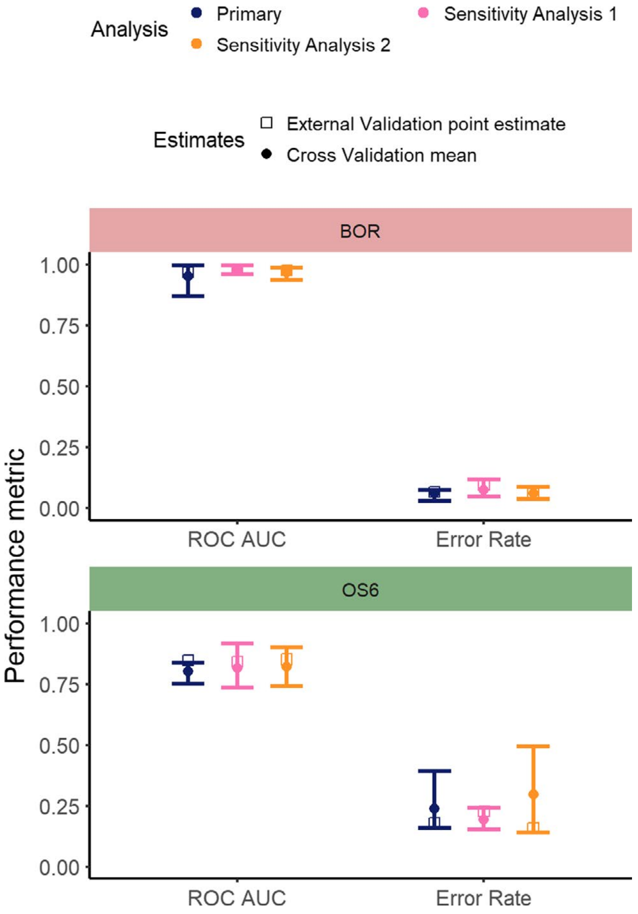

FIGURE 4 Area under the curve (AUC) estimates of receiver-operator curve (ROC) and error rate estimates of the classification algorithm in various settings. Primary analysis: Fivefold cross-validation (mean and 90% prediction interval), and external validation (12-week truncation) for best overall response (BOR) and overall survival at 6 months (OS6) trial outcomes. Sensitivity analysis 1: comparison of results for monotherapy compared with combination therapy. Sensitivity analysis 2: analysis of the effect of sample size (see the Supplementary Materials for details on these analyses)

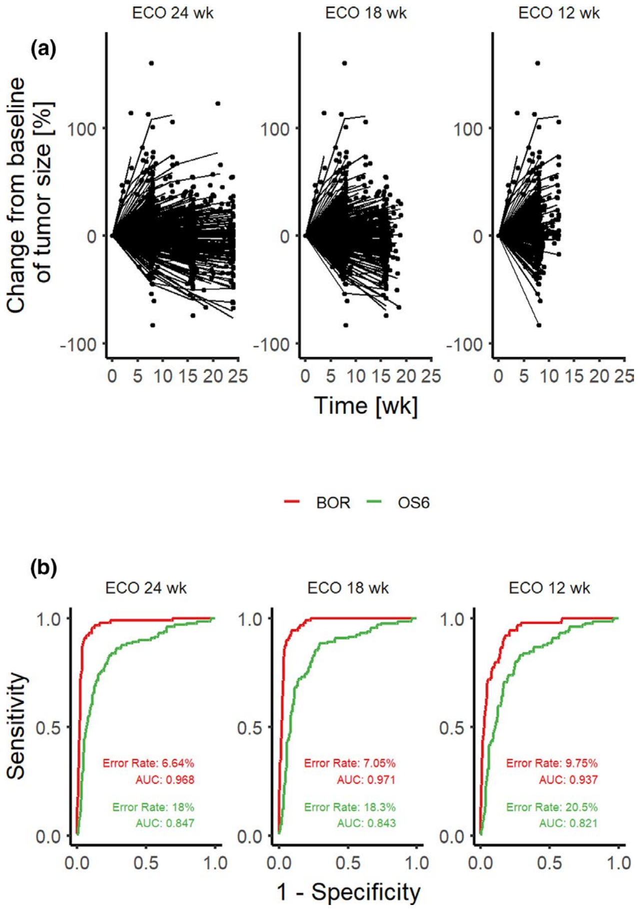

FIGURE 5 (a) Spider plots of observed (dots) and individually predicted (lines) tumor size change from baseline over time at various early cutoff times of the test data set: left (24 weeks), middle (18 weeks), and right (12 weeks). (b) Receiver operating characteristic curve for prediction of best overall response (BOR; red) and overall survival at 6 months (OS6; green) corresponding to each data truncation. Classification error rates and areas under the curve (AUC) of the ROC curve for BOR (red) and OS6 (green) corresponding to each data truncation

0.843, and 0.821, for 24, 18, and 12 weeks, respectively. Classification error rates were 6.64%, 7.05%, and 9.75% for BOR and 18.0%, 18.3%, and 20.5% for OS6 when truncating data based on 24-, 18-, and 12-week periods, respectively, from treatment initiation.

As illustrated in Figure 4, the classification error rates and AUC ROC results of the external validation at week 12 are within the distribution of values obtained during the 5-fold cross-validation both for BOR and OS6. This is suggestive of good predictive performance of the classification method with limited data, because no data truncation was performed during the cross-validation, and hence the distribution obtained represents the most conservative scenarios where all data are available for trial outcome predictions. Sensitivity analyses did not provide noticeable differences in performance compared with the primary analysis (Figure 4).

A more complete summary of the performance of the method is provided in Table 2. All specificity estimates exceeded 0.9 for the full data, 24-, 18-, or 12-week external validation. For sensitivity, the method had 0.773 probability for correct BOR responder classification for the full test data set, and 0.574 for OS6. The method was sensitive to truncation of data with higher false negative rates at earlier times. Hence, error rates appear to be affected more by the increased incidence of false negative classification rather than false positive misclassification when truncating the data. Finally, positive and negative predictive values (PPV and NPV) were also evaluated for predicting responders and nonresponders, respectively. The PPV for BOR was 0.829 when using the full data and 0.773 for 12-week truncation, whereas for OS6, PPV dropped from 0.705 to 0.660 in the same settings. NPV for BOR was 0.95 and 0.926 for full and 12-week data, respectively. For OS6, NPV was

TABLE 2 Performance metrics of the classification based on external validation results from the analysis of EAGLE trial outcomes (BOR and OS6) with the entire dataset or 24-, 18-, or 12-week data truncation

	BOR				OS6			
Data used	All data	24 wk	18 wk	12 wk	All data	24 wk	18 wk	12 wk
AUC ROC	0.957	0.968	0.971	0.937	0.838	0.847	0.843	0.821
Sensitivity	0.773	0.795	0.761	0.659	0.574	0.574	0.566	0.481
Specificity	0.964	0.964	0.967	0.957	0.912	0.909	0.909	0.909
Error rate (%)	7.05	6.64	7.05	9.75	17.8	18.0	18.3	20.5
PPV	0.829	0.833	0.838	0.773	0.705	0.698	0.695	0.66
NPV	0.950	0.955	0.948	0.926	0.854	0.854	0.851	0.827
AUC PRC	0.836	0.866	0.861	0.793	0.611	0.641	0.626	0.591
Youden Index	0.737	0.759	0.728	0.616	0.486	0.483	0.475	0.39
MCC	0.757	0.774	0.756	0.656	0.521	0.516	0.510	0.436

Abbreviations: AUC, area under the curve; BOR, best overall response; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; PRC, precision and recall curve; 0S6, overall survival at 6 months; ROC, receiver-operator curve.

0.854 and 0.827 for full and 12-week data, respectively. However, AUC PRC based on full EAGLE data was 0.836 for BOR (0.793 for 12-week truncation), and 0.611 for OS6 (0.591 for 12-week truncation).

DISCUSSION

In this proof-of-concept work, we introduced a method for early prediction of clinical response to IO therapy. Our method uses NLME model fitting to early tumor size data, coupled with machine learning to classify patients' responses, minimizing total classification error. We evaluated the predictive performance of this method using clinical trial data from durvalumab-treated patients (either monotherapy or in combination with tremelimumab) with HNSCC. Our analysis included cross-validation for model development and an external validation for predictive performance evaluation. To our knowledge, this is the first paper describing a machine learning method that incorporates pharmacometrics-based tumor dynamics model predictions for classification of responders and nonresponders in oncology.

Two clinical end points used in oncology, BOR and OS6, were examined, with overall survival (OS) being assessed at a landmark time of 6 months. Overall, our method demonstrated good predictive performance when prospectively validated with data from an external clinical trial, even for predictions made from only 12 weeks of tumor size data. Minimum gains in predictive power were obtained by incorporating data beyond 12 weeks or by increasing sample size in the sensitivity analyses conducted. The method correctly classified OS6 responders and nonresponders in ~ 80% of cases (AUC of ROC = 0.821 and error rate of 20.5% at week 12) when using at most 3 postbaseline observations of tumor size per patient (average: 1.1 postbaseline observations). For BOR, the performance metrics for external validation were even better, with AUC ROC and specificity reaching 95% in most instances of truncation and a total error rate of 9.75% using only 12 weeks of data.

Because NLME model parameter distributions can vary greatly by cancer type and treatment, the accuracy of our method depends on having adequate prior data for patients with the same indication and same treatment. Nonidentifiability of NLME model parameters for patients with limited longitudinal tumor data reduces the informativeness of individual k_d and k_g estimates, which can result in classification inaccuracy. We hypothesize that a richer dataset, with more frequent early tumor size measurements, would improve the predictive performance of the method.

In addition to its simplicity, another advantage of the proposed method is its modular structure and its adaptability for different objectives. In this work, we minimized the total classification error rate, based on a particular tumor dynamics model, using a logistic regression classification. The true negative rate of the method is consistently above 90% accuracy for BOR and OS6 with this choice of optimization. In other settings, the method could be optimized by penalizing more for false negatives, to reduce the risk of stopping treatment for patients who could benefit from the therapy.

Our method can be adapted for cancers other than HNSCC and for therapies other than IO. Equation 1 is flexible enough to provide good fits to data from patients on therapies other than IO.[23] Additionally, the type of data and features (incorporating prognostic, predictive biomarkers at baseline or with time-varying components), as well as the machine learning algorithm can also be selected or optimized to match the underlying objective of the research question.

In this work, OS6 response category was defined by a mixture of survival and censored data. Informative dropout was not handled with traditional statistical methods proposed for NLMEs analyses.[21,22] Although more than 20% of the patients had only baseline tumor size information, they were included in the predictions because the missingness is likely driven by noncompletely missing data at random effect due to disease progression, which would be informative for both OS6 and BOR outcomes. A dropout or joint model, as developed by Tardivon et al. for atezolimumab,[17] based on urothelial carcinoma SLD and OS data, could be implemented to minimize expected bias in tumor model parameter estimates due to informative dropout. However, even in its current form, our method performs with reasonable predictive accuracy, comparable to more complex methodologies.[17]

Feature selection provided only minor improvement in predictive performance. Only ALB at baseline was informative for predicting OS6 compared with prediction based solely on tumor dynamics, whereas no covariate was found to improve BOR predictions. ALB is a known prognostic marker of OS,[27] and low values of albumin at baseline improved the classification of nonresponders in cases of uninformative tumor dynamic parameters.

Last, in this paper, we only show predictions for a single indication, with all patients treated with durvalumab (either monotherapy or in combination with tremelimumab). Available data from patients with r/m HNSCC treated with durvalumab offered a good sample size for training and an independent clinical trial for a robust external validation of our method. However, patients with r/m HNSCC are a fast-progressing and difficult to treat population with low IO response rates, which did not permit full appraisal of the sensitivity and specificity of our method due to the large number of nonresponders compared with responders. To mitigate inherent limitations in our data set, we calculated precision-recall AUC as an additional performance indicator. For OS6, we noted a disparity between the ROC AUC and PRC AUC estimates (0.8 vs. 0.6) at 12 weeks. The low sensitivity and PRC AUC observed for the OS6 data may be due to a lack of predictive power of target tumor size data alone for survival beyond 6 months in the HNSCC population. Although a model-based approach accounting for dropout could improve our tumor dynamics model parameter estimates, our method performed adequately for BOR with a PRC AUC $\geq$ 0.79 at 12 weeks or higher.

We optimized the method to minimize the total error rates, but we could instead have optimized a metric that weighted PPV more heavily, for example. The choice of metric is one of the features that can be switched out in this modular approach. Youden's J statistic and the Matthews correlation coefficient were not used for the model selection here, but their values indicate that our method is well-informed at week 12 to predict OS6 with the current minimization setting. The method's performance might be further improved by implementing alternative methodologies designed to address the imbalance between responders and nonresponders in the training datasets for both OS6 and BOR.[28] Furthermore, we note that the predictive performance of machine learning methods could improve with sufficiently increased data quality and size, which would then be expected to improve predictive performance of the overall approach.

Although Gong et al.[18] used simulated data, we developed our method based on patient clinical outcomes data set. We designed an external validation strategy that ensured a robust evaluation of the method. Moreover, relying on an actual clinical dataset demonstrates its direct applicability in the drug development or clinical practice setting. Of course, the high variability in the assessment of tumor burden (SLD of target lesions, recording of new lesions or nontarget lesions), and other differences between studies or trial centers would need to be considered accordingly.

In summary, our proposed method combines traditional pharmacometrics and machine learning to make early predictions of clinical outcome in durvalumab-treated patients with HNSCC. Despite the limitations noted above, the method accurately predicted BOR and OS6 and was externally validated based on an independent phase III trial. An accurate method for predicting the response of a patient with cancer and survival using only early data could positively impact drug development as well as clinical practice. For example, such an early prediction method could inform platform trials by optimally switching patients to the most appropriate trial arms at earlier times. It could also aid in the optimization of therapy for individual patients with cancer.

ACKNOWLEDGMENTS

A poster and oral presentation based on this paper were accepted for the ASCPT Annual Meeting 2020. The authors gratefully acknowledge the expert perspective and comments provided by Alejandro Yovine during the internal review process for this manuscript.

CONFLICT OF INTEREST

Ignacio González-García, Vincent F. S. Dubois, Nassim Morsli, and Stuart Spencer are employees of AstraZeneca. Vadryn Pierre is a former employee of AstraZeneca and a current employee of EMD Serono. Paul G. Baverel is a former employee of AstraZeneca and is current employee of Roche Pharma. Helen Moore is a former employee of AstraZeneca, holds stock in Bristol-Myers Squibb, and is a current employee of Applied BioMath.

AUTHOR CONTRIBUTIONS

I.G-G., V.P., V.F.S.D., N.M., S.S., P.G.B., and H.M. wrote the manuscript. H.M. designed the research. I.G-G., V.P., V.F.S.D., P.G.B., and H.M. performed the research. I.G-G., V.P., and P.G.B. analyzed the data.

REFERENCES

1. Camacho LH. CTLA-4 blockade with ipilimumab: biology, safety, efficacy, and future considerations. *Cancer Med*. 2015;4(5):661-672.
2. Haslam A, Prasad V. Estimation of the percentage of US patients with cancer who are eligible for and respond to checkpoint inhibitor immunotherapy drugs. *JAMA Netw Open*. 2019;2(5):e192535.
3. Carretero-González A, Lora D, Ghanem I, et al. Analysis of response rate with ANTI-PD1/PDL-1 monoclonal antibodies in advanced solid tumors: a meta-analysis of randomized clinical trials. *Oncotarget*. 2018;9(9):8706-8715.
4. Kurra V, Sullivan RJ, Gainor JF, et al. Pseudoprogression in cancer immunotherapy: rates, time course and patient outcomes. *J Clin Oncol*. 2016;34(15_suppl):6580.
5. Lee JH, Long GV, Menzies AM, et al. Association between circulating tumor DNA and pseudoprogression in patients with metastatic melanoma treated with anti–programmed cell death 1 antibodies. *JAMA Oncol*. 2018;4(5):717-721.
6. Hodi FS, Ballinger M, Lyons B, et al. Immune-modified response evaluation criteria in solid tumors (IMRECIST): refining guidelines to assess the clinical benefit of cancer immunotherapy. *J Clin Oncol*. 2018;36(9):850–858.
7. Ferrara R, Mezquita L, Texier M, et al. Hyperprogressive disease in patients with advanced non-small cell lung cancer treated with PD-1/PD-L1 inhibitors or with single-agent chemotherapy. *JAMA Oncol*. 2018;4(11):1543-1552.
8. Champiat S, Dercle L, Ammari S, et al. Hyperprogressive disease is a new pattern of progression in cancer patients treated by anti-PD-1/PD-L1. *Clin Cancer Res*. 2017;23(8):1920-1928.
9. Kato S, Goodman A, Walavalkar V, Barkauskas DA, Sharabi A, Kurzrock R. Hyperprogressors after immunotherapy: analysis of genomic alterations associated with accelerated growth rate. *Clin Cancer Res*. 2017;23(15):4242-4250.
10. Baverel PG, Dubois VFS, Jin CY, et al. Population pharmacokinetics of durvalumab in cancer patients and association with longitudinal biomarkers of disease status. *Clin Pharmacol Ther*. 2018;103(4):631-642.
11. Wang Y, Sung C, Dartois C, et al. Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development. *Clin Pharmacol Ther*. 2009;86(2):167-174.
12. Yin A, Moes DJAR, van Hasselt JGC, Swen JJ, Guchelaar HJ. A Review of mathematical models for tumor dynamics and treatment resistance evolution of solid tumors. *CPT Pharmacometrics Syst Pharmacol*. 2019;8(10):720-737.
13. Ribba B, Holford Nh, Magni P, et al. A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis. *CPT Pharmacometrics Syst Pharmacol*. 2014;3(5):113.
14. Sostelly A, Mercier F. Tumor size and overall survival in patients with platinum-resistant ovarian cancer treated with chemotherapy and bevacizumab. *Clin Med Insights Oncol*. 2019;13:117955491985207.
15. Wang M, Chen C, Jemielita T, et al. Are tumor size changes predictive of survival for checkpoint blockade based immunotherapy in metastatic melanoma? *J Immunother Cancer*. 2019;7(1):1-10.
16. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228-247.
17. Tardivon C, Desmée S, Kerioui M, et al. Association between tumor size kinetics and survival in patients with urothelial carcinoma treated with atezolizumab: implication for patient follow-up. *Clin Pharmacol Ther*. 2019;106(4):810-820.
18. 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(3):305-311.
19. Koch G, Pfister M, Daunhawer I, Wilbaux M, Wellmann S, Vogt JE. Pharmacometrics and machine learning partner to advance clinical data analysis. *Clin Pharmacol Ther*. 2020;107(4):926-933.
20. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
21. Hu C, Sale ME. A joint model for nonlinear longitudinal data with informative dropout. *J Pharmacokinet Pharmacodyn*. 2003;30(1):83-103.
22. Björnsson MA, Friberg LE, Simonsson USH. Performance of nonlinear mixed effects models in the presence of informative dropout. *AAPS J*. 2015;17(1):245-255.
23. Moore H. A new tumor dynamics mathematical model. *Am Conf. Pharmacometrics*. 2016;43:S11-S122.
24. Stein WD, Gulley JL, Schlom J, et al. Tumor regression and growth rates determined in five intramural NCI prostate cancer trials: the growth rate constant as an indicator of therapeutic efficacy. *Clin Cancer Res*. 2011;17(4):907-917.
25. Bonate PL, Suttle AB. Modeling tumor growth kinetics after treatment with pazopanib or placebo in patients with renal cell carcinoma. *Cancer Chemother Pharmacol*. 2013;72(1):231-240.
26. Diggle PJ, Kenward M. Informative dropout in longitudinal data analysis. *J R Stat Soc Ser C (Applied Stat)*. 1994;43(1):49-93.
27. Gupta D, Lis CG. Pretreatment serum albumin as a predictor of cancer survival: a systematic review of the epidemiological literature. *Nutr J*. 2010;9(1):69.
28. Krawczyk B. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell*. 2016;5(4):221-232.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: González-García I, Pierre V, Dubois VF, et al. Early predictions of response and survival from a tumor dynamics model in patients with recurrent, metastatic head and neck squamous cell carcinoma treated with immunotherapy. *CPT Pharmacometrics Syst. Pharmacol.* 2021;10:230–240. https://doi.org/10.1002/psp4.12594