Predicting Survival in Patients with Advanced NSCLC Treated with Atezolizumab Using Pre- and on-Treatment Prognostic Biomarkers

Sébastien Benzekry, Mélanie Karlsen, Célestin Bigarré, Abdessamad El Kaoutari, Bruno Gomes, Martin Stern, Ales Neubert, René Bruno, François Mercier, Suresh Vatakuti, et al.

Predicting Survival in Patients with Advanced NSCLC Treated with Atezolizumab Using Pre- and on-Treatment Prognostic Biomarkers

Sébastien Benzekry[1,*], Mélanie Karlsen[1], Célestin Bigarré[1], Abdessamad El Kaoutari[1], Bruno Gomes[2], Martin Stern[3], Ales Neubert[4], Rene Bruno[5], François Mercier[6], Suresh Vatakuti[7], Peter Curle[8] and Candice Jamois[9]

Existing survival prediction models rely only on baseline or tumor kinetics data and lack machine learning integration. We introduce a novel kinetics-machine learning (kML) model that integrates baseline markers, tumor kinetics, and four on-treatment simple blood markers (albumin, C-reactive protein, lactate dehydrogenase, and neutrophils). Developed for immune-checkpoint inhibition (ICI) in non-small cell lung cancer on three phase II trials (533 patients), kML was validated on the two arms of a phase III trial (ICI and chemotherapy, 377 and 354 patients). It outperformed the current state-of-the-art for individual predictions with a test set C-index of 0.790, 12-months survival accuracy of 78.7% and hazard ratio of 25.2 (95% CI: 10.4–61.3, $P < 0.0001$) to identify long-term survivors. Critically, kML predicted the success of the phase III trial using only 25 weeks of on-study data (predicted HR = 0.814 (0.64–0.994) vs. final study HR = 0.778 (0.65–0.931)). Modeling on-treatment blood markers combined with predictive machine learning constitutes a valuable approach to support personalized medicine and drug development. The code is publicly available at https://gitlab.inria.fr/benzekry/nlml_onco.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Current overall survival (OS) prediction models for anticancer treatments rely primarily on pre-treatment or tumor kinetics (TK) data and use classical parametric or semiparametric survival models.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ This study aimed to predict OS in advanced non-small cell lung cancer (NSCLC) patients undergoing immune-checkpoint inhibition (ICI), by introducing a novel kinetics-machine learning (kML) model. Specifically, we sought to integrate baseline markers, tumor kinetics, and, critically, simple on-treatment blood markers (albumin, C-reactive protein, lactate dehydrogenase, and neutrophils) using a combination of nonlinear mixed-effects modeling and machine learning. The primary questions were whether this new approach could improve individual OS prediction and anticipate the outcome of a phase III trial based on previous phase II and early on-study data.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ The integration of on-treatment blood markers kinetics (BK) with predictive machine learning outperformed the current state-of-the-art survival prediction models for ICI in NSCLC. Predictive capabilities with the BKs were superior to models with baseline or baseline and TK data.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ The improved predictive performance of the kML model suggests its potential application in clinical decision-making for individual predictions of overall survival, possibly leading to treatment adaptations. Additionally, the ability to predict the success of a phase III trial based on previous phase II and early on-study data could significantly impact drug development, potentially reducing attrition rates and improving decision-making in immuno-oncology.

[1]COMPutational Pharmacology and Clinical Oncology Department, Centre Inria de l'Université Côte d'Azur, Cancer Research Center of Marseille, Inserm UMR1068, CNRS UMR7258, Aix Marseille University UM105, Marseille, France; [2]Pharma Research and Early Development, Early Development Oncology, Roche Innovation Center Basel, Basel, Switzerland; [3]Pharma Research and Early Development, Early Development Oncology, Roche Innovation Center Zurich, Zurich, Switzerland; [4]Pharma Research and Early Development, Data & Analytics, Roche Innovation Center Basel, Basel, Switzerland; [5]Modeling and Simulation, Clinical Pharmacology, Genentech Research and Early Development, Marseille, France; [6]Modeling and Simulation, Clinical Pharmacology, Genentech Research and Early Development, Roche Innovation Center Basel, Basel, Switzerland; [7]Pharma Research and Early Development, Predictive Modeling and Data Analytics, Roche Innovation Center Basel, Basel, Switzerland; [8]Inovigate, Basel, Switzerland; [9]Pharma Research and Early Development, Translational PKPD and Clinical Pharmacology, Roche Innovation Center Basel, Basel, Switzerland.
*Correspondence: Sébastien Benzekry (sebastien.benzekry@inria.fr)

Received March 14, 2024; accepted June 19, 2024. doi:10.1002/cpt.3371

Lung cancer is the leading cause of cancer death worldwide,[1] with non-small cell lung cancer (NSCLC) being the most prevalent type, representing 80–85% of cases. Immune-checkpoint inhibitors (ICI) (e.g., atezolizumab (ATZ)) have led to significant improvements in survival rates for patients with advanced cancers.[2] However, there is still a large variability in clinical response and progression eventually occurs in a majority of patients.[3] Additionally, drug development in immuno-oncology is highly challenging, with a 95% attrition rate.[4] Current approaches for go/no-go decisions are based on interim endpoints (e.g., progression-free survival, overall response rate) that have often been found to be poor predictors of the primary endpoint of most clinical trials in oncology, namely, overall survival (OS).[5] This calls for better surrogate markers at interim analyses. Altogether, there is a need for better and validated predictive models of OS for both personalized health care (individual predictions) and drug development (trial predictions).

Currently, programmed-death ligand 1 (PDL1) expression is the only routine predictive biomarker used for NSCLC patients,[3,6] despite being controversial.[7] Tumor mutational burden[6] and transcriptomic data[3] have also been investigated but did not reach clinical practice. Here we posit that such static and single marker approaches are intrinsically limited and that improved predictive performances could be achieved by: (i) using multimodal integrative analyses relying on a combination of markers and machine learning algorithms[3,8] and (ii) including dynamic markers obtained from early on-treatment data.[9,10] The nonlinear mixed-effects (NLME) modeling approach is well suited for the latter, and tumor kinetics (TK) model-based metrics have already been shown to carry significant predictive value for OS in oncology, including ATZ monotherapy in advanced NSCLC.[10,11] The first main novelty of the current study is to investigate the predictive value of model-based parameters of simple blood markers kinetics (BK), in addition to TK.

The second main novelty is to apply ML algorithms, increasingly used in biology and medicine,[12] but only rarely for TK-OS modeling,[13] instead of classical survival models. Extensions of classical ML models to survival data have been proposed (e.g., random survival forests[14]) that are appealing to build improved predictive models. However, the actual superiority of ML algorithms over standard approaches for clinical prediction models remains controversial.[15] In addition, most ML studies to date have low sample sizes in both training and test sets, which affects their reliability due to overfitting, as well as generalizability.

Here, we coupled the strengths of NLME modeling with ML to derive a predictive model of OS from baseline and on-treatment data, called kinetics-machine learning (kML, **Figure 1a**). We leveraged large training and test datasets to achieve robust results (**Figure 1b**). Subsequently, we tested the operational predictive capabilities of kML in two relevant scenarios: (i) individual prediction of OS and (ii) prediction of the outcome of a phase III trial from early on-study data.

METHODS

Data

The data consisted of advanced NSCLC patients enrolled in ATZ trials ($N = 1936$, **Figure 1b**; **Figure S1**). The training set comprised the FIR (NCT01846416),[16] POPLAR (NCT01903993)[2] and BIRCH (NCT02031458)[17] phase II clinical trials. The test sets were the atezolizumab arm of the OAK phase III trial (NCT02008227)[18] for individual predictions and additionally the docetaxel (DTX) arm for trial predictions (**Figure S1**). These studies were conducted in accordance with the Declaration of Helsinki after approval by institutional review boards or independent ethics committees. All patients provided written informed consent. Patients from French centers were excluded for legal reasons ($N = 118$).

The outcome considered was OS, defined as the time between treatment start and death or last follow-up, in which case the data were right-censored. The median follow-up was 35.2 months (95% CI: 34.5–35.7) in the training set and 26.8 months (95% CI: 26.3–27.5) in the test set.

Preprocessing

Baseline data. The baseline data (BSL) consisted of 63 variables (43 numeric and 20 categorical) variables spanning demographic and biological data, clinical information and disease status (see **Figure S3** for a description of the main variables).

Tumor and blood markers kinetics. Longitudinal TK data consisted of the investigator-assessed sum of largest diameters (SLD) of lesions as per the RECIST criteria.[19] Patients with only one baseline SLD measurement and no SLD measurement during the treatment period were excluded ($N = 110$). This resulted in 5,473, 3,015, and 2,020 data points in the training, ATZ and DTX test sets, respectively, with a median of 5, 4, and 4 data points per patient.

On-treatment BK data on albumin, C-reactive protein (CRP), lactate dehydrogenase LDH, and neutrophils were analyzed. Data points exclusion criteria were set: (i) physiological bounds, (ii) duplicates, (iii) extreme out-of-range shifts. In order to have enough data from Bayesian estimations, patients with less than three pre-cycle five observations were removed, mostly due to missing CRP data ($N = 282$ (33%) in training, $N = 155$ (28%) and 176 (34%) in test sets). We refer to the **Supplementary Methods** and **Figure S2** for details. After preprocessing, 60,779, 38,460, and 11,799 data points remained in the training, ATZ and DTX test sets, respectively.

Nonlinear mixed-effects modeling

Population approach. Statistical hierarchical nonlinear mixed-effects modeling (NLME) was used to implement a population approach[20] for the kinetic data and parameter estimation was conducted using the Monolix software.[21] Mathematical details are given in the supplement.

Structural models. Following previous work, the TK structural model was assumed to be the sum of two exponentials[10,22]:

$$y_j^i = \begin{cases} y_0^i e^{\mathrm{KG}^i t} & t \le 0 \\ y_0^i \left(e^{-KS^i t} + e^{\mathrm{KG}^i t} - 1 \right) & t > 0 \end{cases},$$

where $t = 0$ corresponds to treatment initiation and y_0, KG and KS are three parameters, representing respectively the baseline value, growth and shrinkage rates. This model was also considered for the BKs, together with three other models: constant ($y_j^i = \alpha^i, \forall j$), linear ($y_j^i = \alpha^i + \beta^i t_j^i, \forall j$) and hyperbolic ($y^i = p^i + \frac{\varepsilon^p (q^i - p^i)}{t_j^i + \varepsilon^p}$)[23] (**Figure S4**). In the

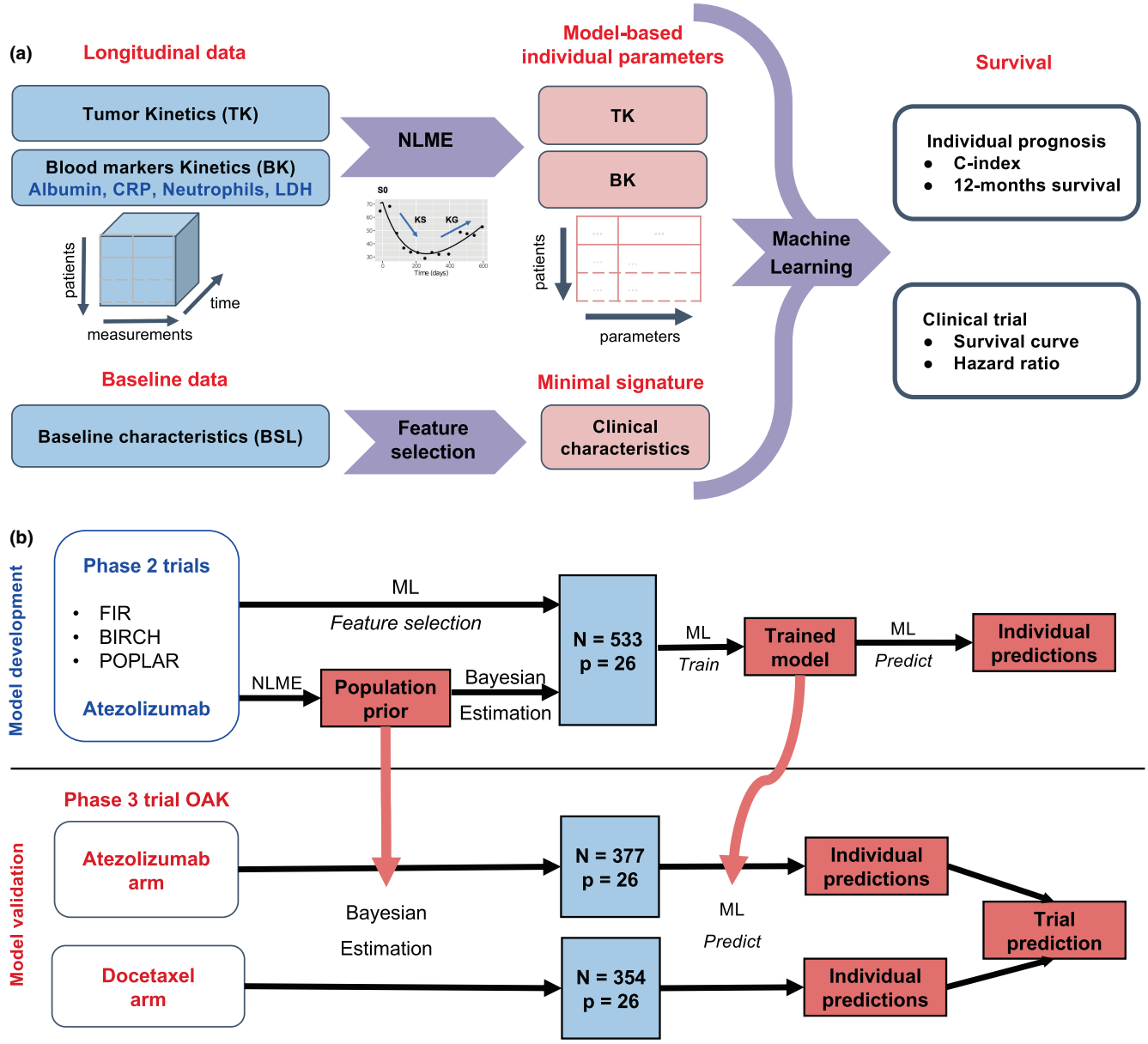


Figure 1 Study schematic. (**a**) Baseline and longitudinal data were combined into a machine learning algorithm in order to predict individual survival prognosis. Longitudinal data were modeled using nonlinear mixed-effects modeling, whereas machine learning-based feature selection was applied to the baseline data to derive a minimal signature. Tumor kinetics and biological kinetics parameters were combined with the minimal signature to predict survival. Predictive performances were assessed using survival metrics (C-index and survival at horizon times). (**b**) Algorithm used to develop the model on the train data and carry it to the test set for external validation. Each step—preprocess, learning of the Bayesian priors, dimensionality reduction, feature selection, choice, tuning, and training of the machine learning algorithm—was calibrated on the training set and then applied to the test set. TK, tumor kinetics; BK, blood markers kinetics; ML, machine learning; NLME, nonlinear mixed-effects modeling.

latter, parameter q is the model-based initial condition, parameter p is the asymptotic value as $t \to +\infty$ and parameter l is the time-point at which $y = \frac{p+q}{2}$. Quantitative comparison of goodness-of-fit between models was assessed using the corrected Bayesian information criterion (BICc).[24]

Identification of individual model-based parameters. For TK and each BK, the population parameters θ_{pop} and ω were identified using the training set. The resulting distribution was used as the prior in subsequent Bayesian estimations of the individual parameters $\hat{\theta}^i$, in both the training and test sets.

Model-estimated baseline parameters were not kept. We additionally considered the ratio of the model-predicted value at cycle 3 day 1 to the model-estimated baseline parameter.

Truncated data: individual level. Data was truncated, for each patient, at cycle 3 day 1 (C3D1), C5D1, and C10D1 from longitudinal TK and BK data. Training priors were recalculated from each CXD1 set.

Truncated data: study-level. For trial predictions, the data was truncated on the basis of landmark times denoted lt after study initiation: lt = 10, 25, and 60 weeks. Only the patients enrolled before this time and their data collected up to lt were kept. Note that in these datasets, patients have varying follow-up duration (from 0 to lt).

Machine learning

See the **Supplementary Methods** for details on missing values and scaling.

Models. Model elaboration and development were performed on the training set, using 10-folds cross-validation for evaluation of performances. The following survival models were used: proportional hazards Cox regression,[25] extreme gradient boosting (XGB) with either Cox or accelerated failure time (AFT) models[26] and random survival forests (RSF).[14] Nested cross-validation with inner bagging in each 10-fold cross-validation outer loop was used to evaluate the benefit of tuning the hyperparameters (see **Supplementary Methods** for details).[27] Improvement of the performances was negligible with hyperparameter tuning (**Figure S4**). Therefore, we used the default values of the hyperparameters. For the final RSF model: number of trees = 500, number of variables to possibly split at each node = 5, minimum size of terminal node = 15, number of random splits for splitting a variable = 10.

Evaluation. For each patient, the RSF model gives two prediction outputs: a scalar value termed "mortality" that we will refer to as "ML score," and a time-dependent survival curve.[14] We refer to the supplement for detailed specifications of all the evaluation metrics.

For patient stratification (dichotomized KM curves), the cut-points were defined at the 20th percentiles of the ML score to identify the top 20% long-term survivors. This approach ensured fair comparison across variables/scores. Differences between KM strata were assessed with the log-rank test and hazard ratios were computed using Cox regression.

Variable selection and minimal signature. Variable selection was performed only for the BSL data. The method was based on two steps: (i) sorting the variables using least absolute shrinkage and selection operator (Cox-LASSO)[28] and (ii) building RSF incremental models including increasing numbers of Cox-LASSO sorted variables.

See Supplement for details on the feature selection and computation of predicted HRs.

Code availability

The code designed for conducting analyses, constructing, and validating models can be accessed at https://gitlab.inria.fr/benzekry/nlml_onco and at a software heritage permalink.[29]

RESULTS

Data

The training dataset comprised 862 patients[2,16,17] and the external validation (test) sets 553 and 521 patients, respectively.[18]

Variables comprised baseline (pretreatment, 63 features) and longitudinal (on-treatment) data (**Figure 1a**). The latter included TK and on-treatment measurements of four BKs: albumin, CRP, LDH, and neutrophils. These markers were selected during data exploration where we observed a significant association of OS with the C3D1 to baseline ratio (P = 0.0046, 0.08, 0.0001, and 0.0026, respectively univariable Cox regression).

Following preprocessing (see Methods section), there remained 533 patients in the training dataset and 377 and 354 patients in the test datasets (ATZ and DTX arms, respectively). The major determinant of patients' withdrawal was the lack of CRP measurements during the first three treatment cycles (**Figure S2**).

Nonlinear mixed-effects modeling (NLME) of on-treatment biomarkers

The TK structural model was the sum of an increasing and a decreasing exponential function (double-exponential model).[22] It was able to accurately describe the training data with no goodness-of-fit misspecification (**Figure 2a**; **Figure S6**). Population parameters were estimated with good accuracy (all relative standard errors smaller than 9%, **Table 1**).

For the BKs, we first investigated whether significant on-treatment kinetics could be observed beyond random noise (see raw data in **Figures S7–S10**). The null hypothesis was set to a constant model and tested against three alternative empiric models: linear, hyperbolic (monotonous but nonlinear and saturating), and double-exponential (nonlinear and non-monotonous). We found significant kinetics for each BK, demonstrated by a lower BICc and proportional error parameter b (**Figure S11**). The best models were hyperbolic for albumin and double-exponential for the other BKs. Representative fits are shown in **Figure 2a** and goodness-of-fit metrics are in **Figures S12–S15**. Parametric identifiability was excellent for all models (**Table 1**).

We further assessed the stratification value of the individual model-based kinetic marker for OS prognosis (**Figure 2b**). The TK parameter KG (growth rate) exhibited good stratifying ability (HR = 4.39 (2.8–6.89)), which was similar to the CRP_KG parameter (HR = 4.37, (2.76–6.91)). Ranked by HR importance (controlled by the 20th percentile definition of the cut-point, see Methods section), the next four best parameters were: albumin_p (HR = 3.17 (2.11–4.78)), neutrophils_KG (HR = 3.07 (2.04–4.63)), neutrophils_KS (HR = 2.33 (1.6–3.39)) and TK_KS (HR = 2.02 (1.42–2.89)). All kinetic parameters carried substantial prognostic power (P < 0.0001, log-rank test).

Survival prediction using kinetics-machine learning (kML): model development

Three feature sets resulted from the analysis above: BSL, TK, and BK (**Figure 1a**). The development of kML comprised two main steps: (i) choice of the algorithm and (ii) derivation of a minimal signature (**Figure 1b**). The first was achieved by benchmarking four models that used all variables (P = 119, N = 553). The RSF model was found to exhibit the best performances (**Figure S6**) and was thus selected. Notably, we found significantly better predictive performances of RSF over a classical Cox proportional hazard regression model (significant differences in C-index, P = 0.0006, Student's t-test).

Feature selection on the BSL variables was performed building incremental RSF models based on LASSO importance-sorted variables (**Figure 3a**). They were evaluated using 10-fold cross-validated AUCs and C-indices performed on the training set. The model using all of them achieved the best score. Nevertheless, keeping in mind the objective to ultimately support decision-making and patient stratification, a minimal (11 features), near-optimal, set of BSL variables was selected and denoted mBSL. It was defined as the first seven variables reaching the plateau (baseline

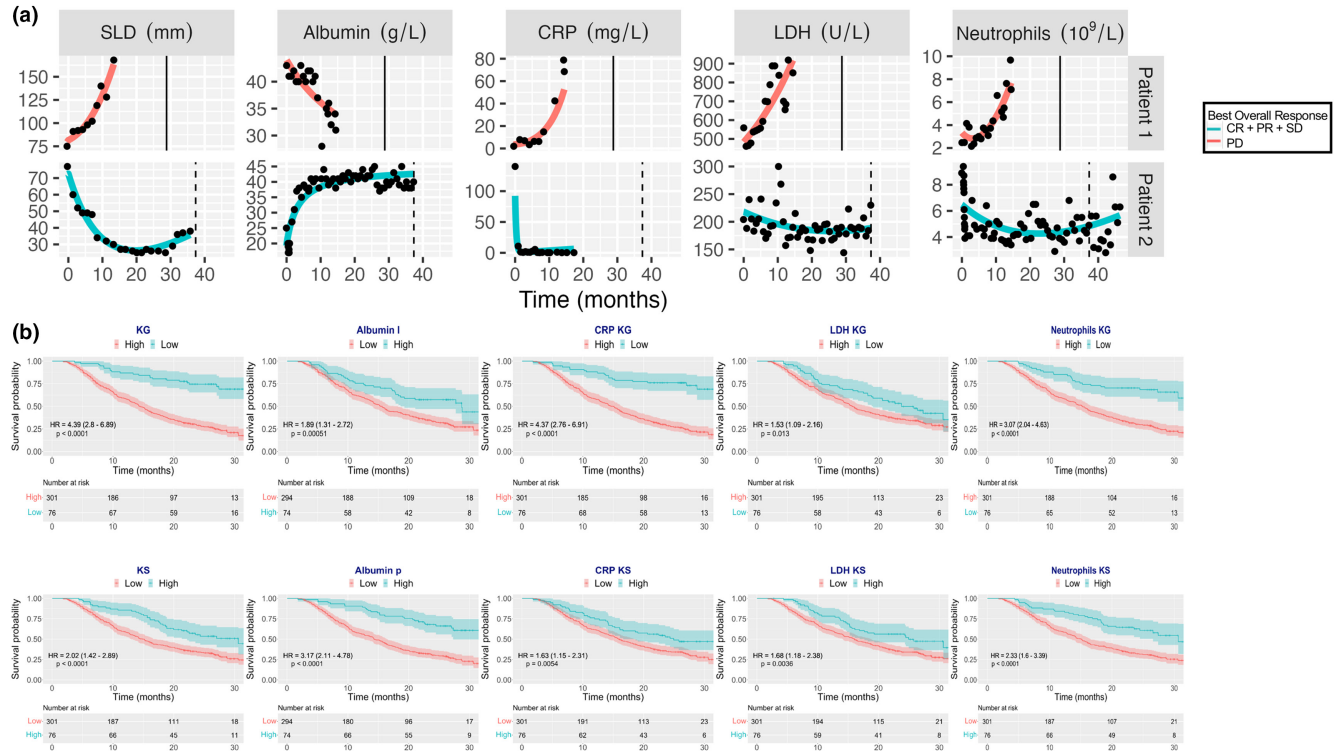


Figure 2 Goodness-of-fit metrics and plots of dynamic BK models. (**a**) Representative individual fits for the TK and BK best empirical models showing nontrivial kinetic parameters well captured by the dynamic models. Survival is indicated by a vertical line (solid = death, dashed = censored). (**b**) Stratified Kaplan–Meier curves at the 20th percentile level on the test set, for TK and BK model-based parameters. Missing values were removed in this univariable analysis, explaining the difference of initial number of patients for albumin that had 9 patients in this case. CR, complete response; PD, progressive disease; PR, partial response; SD, stable disease.

Table 1 Parameters from nonlinear mixed-effects modeling of tumor and blood marker kinetics

	TK	CRP	LDH	Neutrophils	Albumin	
KG_{pop} ($week^{-1}$)	0.00492 (6.80)	0.00814 (9.38)	0.00238 (10.48)	0.00436 (8.69)	p_{pop} (g/L)	29.4 (3.82)
KS_{pop} ($week^{-1}$)	0.00778 (8.22)	0.0137 (14.14)	0.00184 (13.73)	0.000987 (21.16)	l_{pop} (log(day))	8.09 (2.74)
ω_{KG}	1.36 (3.80)	1.61 (4.25)	1.55 (5.36)	1.41 (4.48)	ω_p	0.476 (7.48)
ω_{KS}	1.41 (4.66)	1.81 (6.29)	1.92 (5.34)	2.46 (5.82)	ω_l	0.359 (6.42)
error	6.82 (1.15)	0.559 (1.23)	0.138 (0.79)	0.207 (0.82)	error	0.0549 (0.774)

Parameter value (relative standard error (%)). CRP, C-reactive protein; LDH, lactate deshydrogenase; TK, constant error; others, proportional error.

CRP, heart rate, neutrophils-to-lymphocytes ratio, neutrophils, lymphocytes-to-leukocytes ratio, liver metastases, and ECOG score), complemented with four variables with established prognostic or predictive value and available in routine care: PD-L1 expression (50% cutoff),[2] hemoglobin,[30] SLD[13] and LDH.[31,32]

We then compared the cross-validated C-indices of each feature set (**Figure 3b**). mBSL exhibited moderate discrimination performances (C-index = 0.710 ± 0.038), which was slightly outperformed by the TK set (C-index = 0.723 ± 0.025). Interestingly, the BK set significantly outperformed both mBSL and TK (C-index = 0.793 ± 0.038, $P = 0.0004$ and 0.0005, respectively, Student's *t*-test). Jointly, mBSL, TK, and BK performed significantly better than any feature set alone (C-index = 0.824 ± 0.050, $P = 0.00007$, 0.0002 and 0.055), as well as any combination of two sets among the three (mBSL + TK: 0.77 ± 0.026, mBSL + BK: 0.81 ± 0.027, TK + BK: 0.80 ± 0.049). The resulting model combining mBSL, TK, and BK was denoted kML (kinetics-machine learning).

During cross-validation on the training set, kML exhibited excellent predictive performances across multiple metrics, with minimal between-folds variability (e.g., AUC = 0.919 ± 0.056, accuracy = 0.873 ± 0.052, **Figure 3c**).

External validation

The predictive performance of kML (mBSL, TK, and BK) was assessed on the ATZ test set (377 patients). At the population level, the model-predicted survival curve was in excellent agreement with the observed data (**Figure 4a**). Notably, the prediction interval from the model was narrow, indicating high precision. At the individual level, consistent with the cross-validation results, substantial discrimination performances were observed (C-index = 0.790,

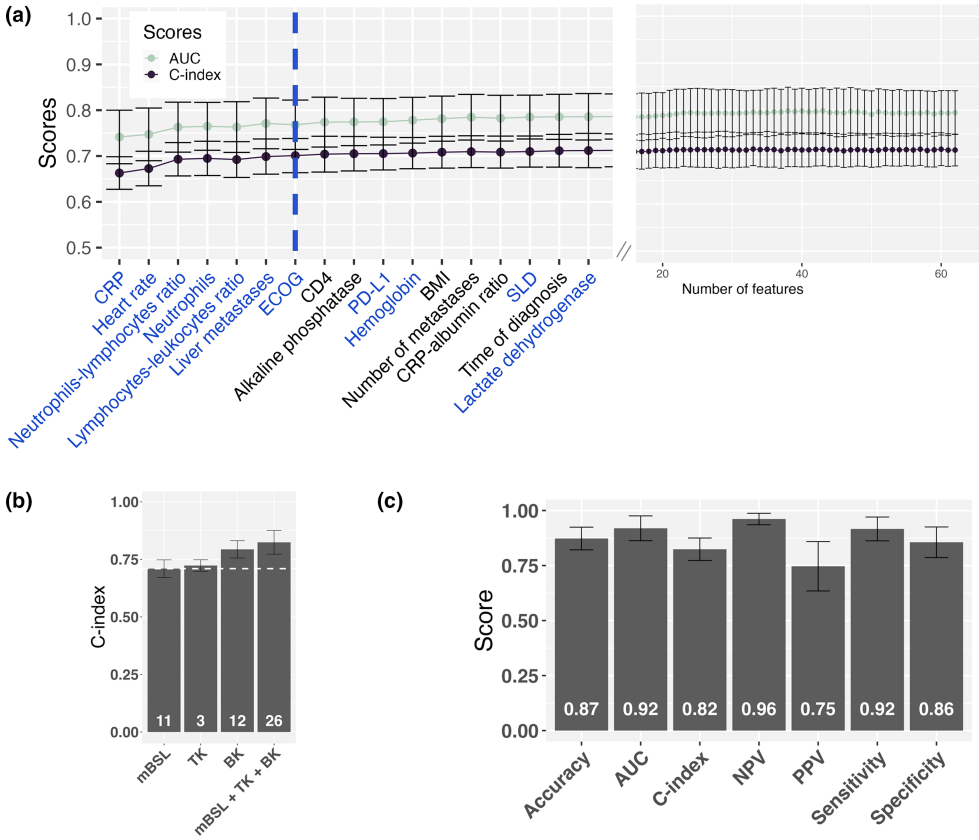


Figure 3 Minimal baseline (mBSL) signature and kinetics-ML (kML) model. (**a**) Cross-validated (CV) performance scores on the training set (C-index and AUC, mean ± standard deviation) for incremental random survival forest (RSF) models using an increasing number of baseline clinical and biological variables sorted by LASSO importance. The dashed blue line shows the minimal number of variables reaching the plateau. Blue-colored variables correspond to the minimal clinical signature (mBSL). (**b**) Comparative CV c-indices of RSF models based either on mBSL, TK, BK, and mBSL + TK + BK (final model, kML) variables showing increased predictive performances over baseline when using model-based parameters of kinetic markers. Numbers on the bars indicate the number of variables. The dashed horizontal line represents the mBSL mean C-index. (**c**) CV performances of the kML model for discrimination (C-index) and classification (survival prediction at 12 months OS).

accuracy and AUC for 12-months survival probability = 0.787 and 0.874, respectively, **Figure 4b**). All classification metrics for prediction of survival at 12 months were high (≥0.78), except PPV, indicating a worse ability to predict death than survival. Although smaller, they were similar to the cross-validation results.

In addition, calibration curves revealed good performance, at multiple horizon times (**Figure 4c**). Model-predicted probabilities were concordant with the observed KM estimates of the survival probabilities, over the entire range of the binned predicted probabilities. This is further illustrated by the contingency **Table 2**. For instance, among 212 patients predicted to be alive at 12 months, 182 (85.8%) were actually alive. Predictive AUC was good at other horizon times (0.846 and 0.910 at 6 and 24 months, respectively, **Figure S16**). However, PPV and sensitivity were very low at 6 months.

Notably, the kML mortality score derived from the model and learned on the training set was able to accurately stratify OS in the test set (HR = 25.2 (10.4–61.3), $P < 0.0001$, **Figure 4d**), indicating excellent ability to identify the 20% of long-term survivors. It outperformed all single kinetic markers (**Figure 2c**).

Variable's importance was assessed by running a *post hoc* multivariable Cox regression (**Figure 4f**). Interestingly, the top two variables were BKs model-derived parameters (CRP_KG and CRP ratio C3). In addition, TK and BKs covered six out of the seven top important features and were found more important than PDL1.

Given the large sample size of our data, we further assessed the model performances when trained on smaller data sets (**Figure S17**). The learning curve revealed that approximately 200 patients were necessary to reach similar performances to the ones obtained with the full training set ($N = 533$), for both cross-validation and external validation on the test set (C-index = 0.82 ± 0.056 vs. 0.82 ± 0.050 in cross-validation, 0.78 vs 0.79 on the test set, models trained with 200 vs. 533 patients, respectively). Trained with only 60 patients, kML reached already good performances (C-index = 0.76 ± 0.15 and 0.74 in cross-validation and test, respectively).

Together, these results demonstrate important predictive performances of overall survival following ATZ treatment using the kML model.

Application to individual survival prognosis from early on-treatment data

Results above required full on-treatment time-course data to compute TK and BK markers, thus cannot be used to make early predictions. To investigate the operational applicability of our

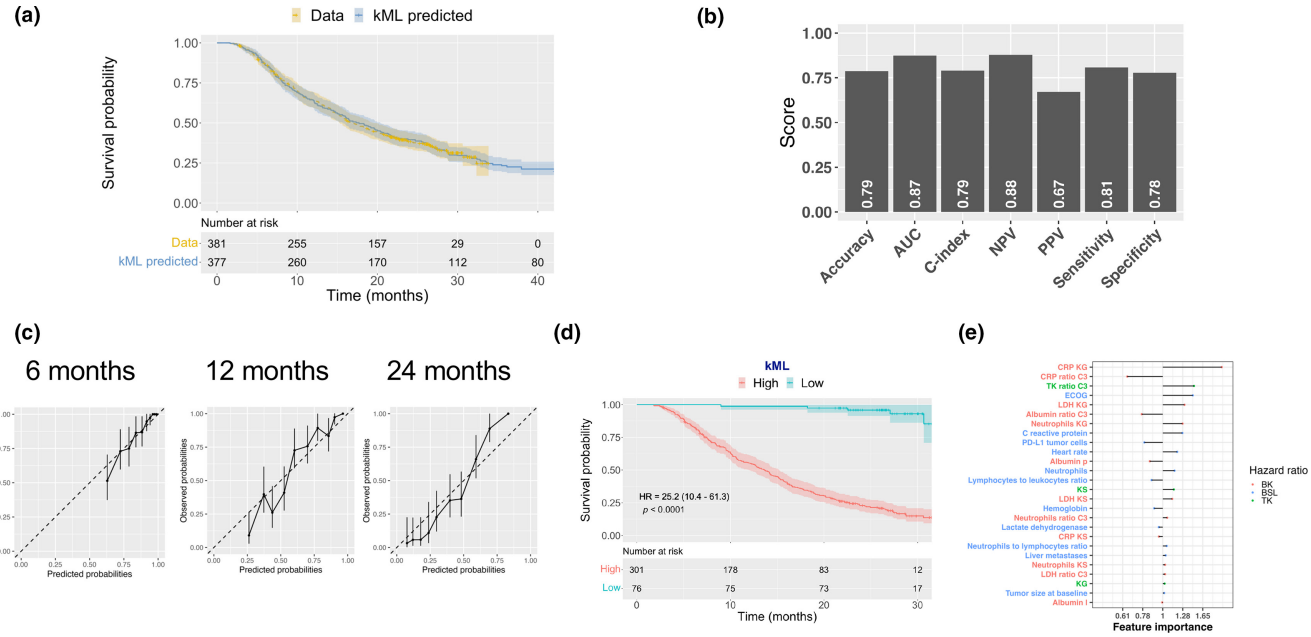


Figure 4 Predictive performances of kML on the ATZ test set. (**a**) Comparison of the population-level survival curves between the data (KM estimator) and the kML model prediction in the ATZ arm of OAK. (**b**) Scores of discrimination metrics. Classification metrics were computed for the prediction of OS at 12 months. (**c**) Calibration curves at 6, 12, and 24 months, showing the observed survival probabilities (with KM 95% confidence interval) versus the predicted ones in 10 bins corresponding to the model-predicted survival probability deciles. Dashed line is the identity. (**d**) Dichotomized KM survival curves based on the ML model-predicted score (high vs. low), at the 20th percentile cutoff. (**e**) Variables importance (multivariable hazard ratios) in the full time-course kML model.

methodology, data from the test set were truncated at day 1 of the treatment cycles 3, 5, and 10, respectively corresponding to 1.5, 3, and 6.75 months. We found that integrating longer on-treatment data in kML, the predictive performances steadily increased (**Figure 5a**; **Figure S18**). Using the baseline variables only (mBSL), the stratification ability was significant but moderate (HR = 1.74 (1.24–2.46), $P = 0.0014$, **Figure 5b**). In contrast, kML exhibited increasing stratification ability from data at 1.5 months (HR = 2.19 (1.53–3.12), $P < 0.0001$), 3 months (HR = 3.51 (2.33–5.3), $P < 0.0001$) and 6.8 months (HR = 5.01 (3.16–7.95), $P < 0.0001$), see **Figure 5c**.

Further investigation of the predictive performances of individual kinetic markers revealed that TK parameters were the most informative at 6 weeks (1.5 months, first imaging assessment). Adding BKs to TKs brought additional predictive value starting at 3 months, and BKs outperformed TK from 6.75 months on (**Figure S19A**). Among BKs, neutrophils kinetics appeared to be the most predictive, followed by CRP, albumin, and LDH. However, the combined BK signature outperformed each individual BK, indicating that their collective predictive capabilities were not driven by any single BK alone.

Interestingly, the most important variable at 1.5 months was a kinetic one, TK ratio C3 with the following variables being from mBSL (liver metastases, PDL1 and ECOG). When more on-treatment variables become available, this shifted to TK and BK (TK ratio C3, KS, KG, CRP_KG, LDH_KG), see **Figure S19B**.

Table 2 Contingency table for OS prediction at 12 months

	Truth		
	Alive (0)	Dead (1)	Total
Model			
Alive (–)	182	30	212 (58.7%)
Dead (+)	48	101	149 (41.3%)
Total	230 (63.7%)	131 (36.3%)	361

16/377 censored patients with survival time ≤12 months removed for computation of accuracy, sensitivity, specificity, PPV and NPV don't correspond exactly to the numbers because they are computed from KM estimate, thus adjusting for censoring bias.

Application to clinical trial outcome prediction from early on-study data

The kML model can also be applied for the prediction of the outcome of a clinical trial (survival curves and associated hazard ratio), from early on-study data. We performed on-study truncations on the two arms of the test set based on a number of weeks after the date of the first patient recruited (see Methods section). Here, we applied the model to predict not only patients receiving ATZ, but also DTX (**Figure 1b**). Predictions of the kML model applied to each arm yielded very accurate results when using data from the entire study (predicted HR: 0.784 (0.7–0.842), vs. data HR: 0.778 (0.65–0.931), **Figure 6a,b**). Notably, the model prediction intervals were narrower than the data Kaplan–Meier confidence intervals, probably because the kML-trained model incorporates the information from the three phase II trials. Using only early data, the model was already able to detect a (non-significant) tendency at 10 weeks, with only 23 and 30 patients in each arm, and a very short follow-up. Starting from data available at 25 weeks

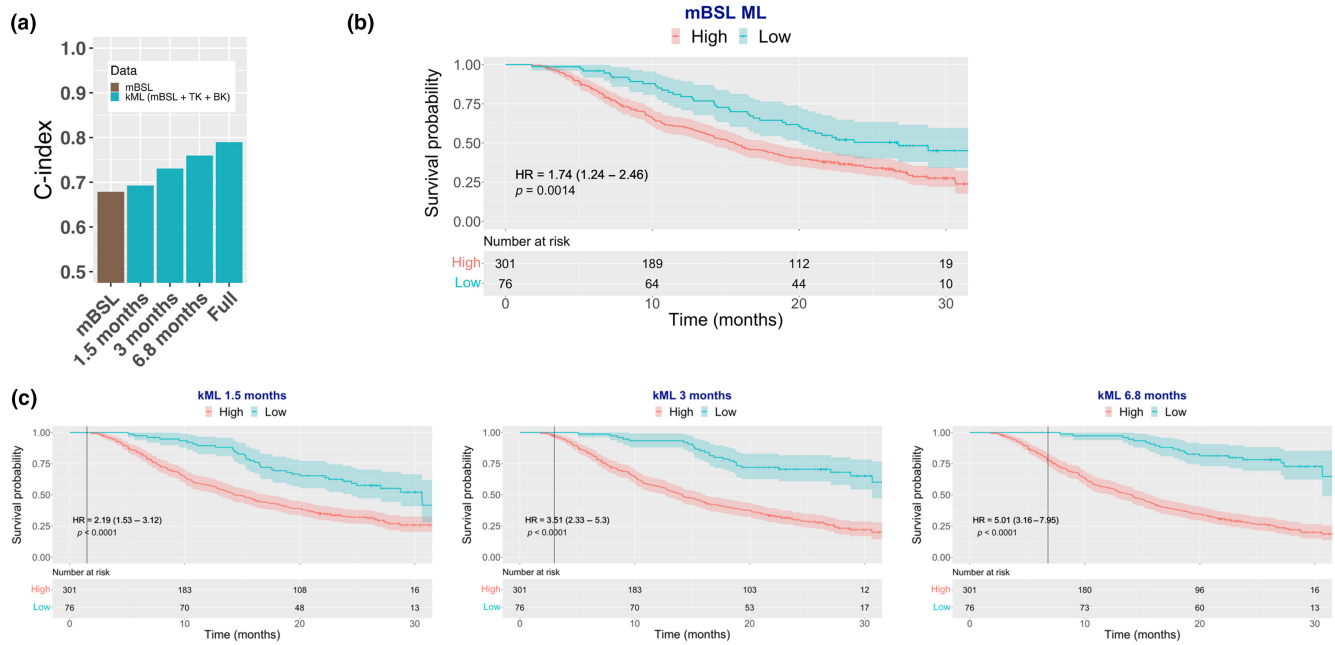


Figure 5 Predictive value of kML from cycle-truncated data. (**a**) Predictive power (C-index) of ML models using baseline (BSL) or truncated data at 1.5, 3, and 6.8 months as well as the full time-course. (**b**) Stratified KM survival curves using a RSF model trained on the minimal baseline (mBSL) variables. (**c**) Stratified KM survival curves using kML from 1.5 months (two cycles), 3 months (four cycles) and 6.8 months (nine cycles) truncated data. Truncation time is indicated by the vertical line. BK, biological kinetics; CRP, C-reactive protein; LDH, lactate dehydrogenase; TK, tumor kinetics.

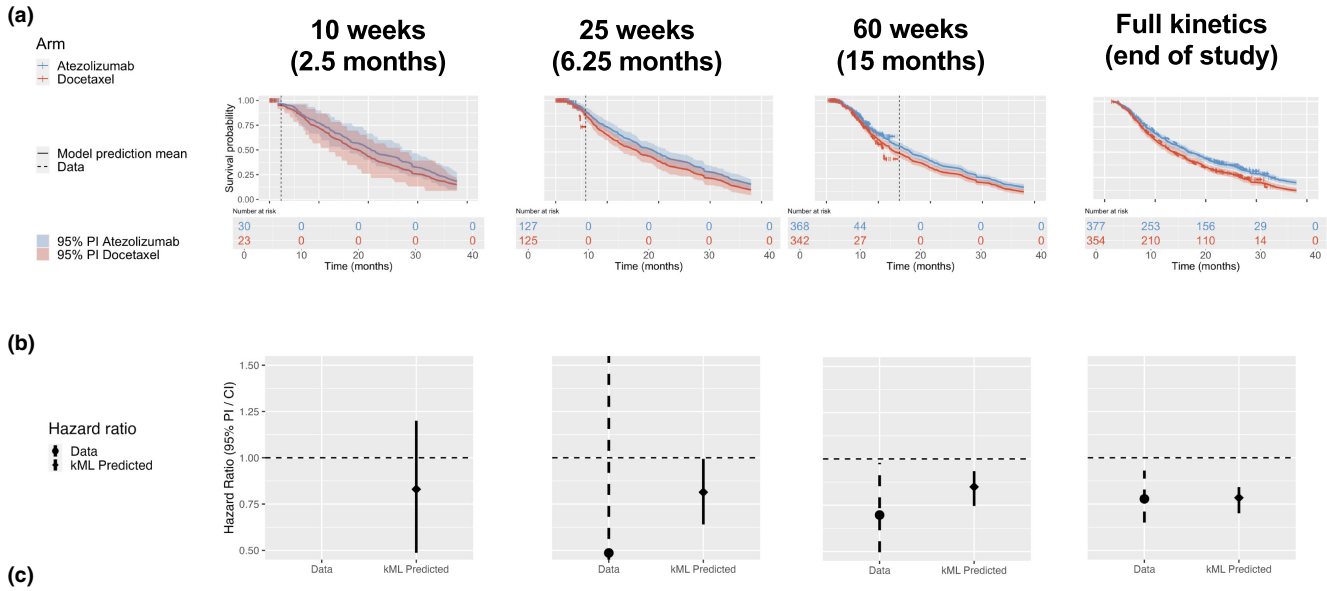
(6.25 months), the model correctly predicted a positive outcome of the study, with a 95% prediction interval of the HR below 1. Of note, the available OS data at this time (dashed lines, **Figure 6a,b**) was far from being conclusive. The model prediction was stable from 25 weeks on whereas the OS data only exhibited significant HR starting from 60 weeks and required more than 300 patients in each arm to be conclusive.

DISCUSSION

Blood markers from hematology and biochemistry are routinely collected during clinical care or drug trials. They are cost-effective and easily obtained both before and during treatment. There is to date limited exploration regarding the predictive capabilities of the kinetics of such data. Combining BSL variables with on-treatment data (TK and BK), we investigated this using a novel hybrid NLME-ML methodology. The resulting kML model demonstrated excellent performances in two aspects: (i) patient-level and (ii) trial-level OS predictions. The kML model outperformed current state-of-the-art methods based on either baseline or only TK data. The latter include PD-L1 (AUC = 0.601 for durable response),[7,33] tumor mutational burden (AUC = 0.646),[34,35] baseline blood counts (AUC = 0.74)[31,36–38] or the ROPRO score, derived from a large pan-cancer cohort and incorporating baseline clinical and biological data (27 variables). ROPRO achieved a C-index of 0.69 and a 3-months AUC of 0.743 for a similar prediction as our case (OAK clinical trial).[39]

TK—OS modeling value is now well-established[10,11] and further confirmed here. BK and BK—OS modeling are less common in the literature. In 2020, Irurzun-Arana et al. proposed a semi-mechanistic model for the time-course of three circulating biomarkers and their association with OS and toxicity, for melanoma patients. They found that LDH increase was negatively associated with OS, a finding confirmed in our results for NSCLC. In 2021, Gavrilov et al.[23] proposed a hyperbolic model of NLR kinetics that exhibited improved OS predictions over TK alone. We focused here on four BKs: albumin, CRP, LDH and neutrophils. Albumin is associated with nutritional status (cachexic state) and is known to evolve with time in responders. CRP is a marker of systemic inflammation.[32] Increased CRP, decreased albumin level, and increased CRP/albumin ratio have been reported to be associated with poor survival.[40] Neutrophils play a role in inflammation by promoting a favorable microenvironment for cancer cell growth and spread, and activation of carcinogenic signaling pathways.[41] Elevated LDH levels are a marker of cancer cell turnover rate, and LDH has a potential role in prediction of potentially invisible metastases.[32] We found that all these markers had nontrivial on-treatment kinetics and that all of these kinetics were significantly associated with OS. In addition, when combined together, they outperformed TK alone.

While significant and predictive, the model fits with empiric models were far from perfect. This leaves room for improvement with more complex, mechanistic models of the joint TK—BKs—OS. Further, mechanistic models using multimodal data collected during, for example, the PIONeeR clinical study (NCT03833440) will bring additional information for the understanding and prediction of mechanisms of resistance to ICI.[42–44] This data includes pharmacokinetics, quantitative image analysis, genomics and transcriptomics, circulating DNA,[45–48] immune- and vasculo-monitoring as well as a large number of BKs and soluble factors.[49]



(c)

Data HR (95% CI)	NA (NA – NA)	0.486 (0.0863 – 2.73)	0.697 (0.496 – 0.979)	0.778 (0.65 – 0.931)
Predicted HR (95% PI)	0.836 (0.547 – 1.4)	0.814 (0.64 – 0.994)	0.849 (0.746 – 0.934)	0.784 (0.7 – 0.842)
Number of patients (DTX - ATZ)	23 – 30	125 – 127	342 – 368	354 – 377
Data points (TK/Alb/CRP/LDH/ Neutrophils)	1/3/3/2/2.5 – 1/2/2/2/2	2/4/4/4/4 – 2/4/4/4/4	3/6/5/6/7 – 3/7/6/7/7	4/8/6/7/8 – 4/9.5/9/9/10

Figure 6 Use of kML for early-prediction of the outcome of a clinical trial. (**a**) Survival curves model-based predictions and prediction intervals vs. actual data from on-study data at multiple landmark times after study initiation. Note that the model is able to predict full survival curves even if based on early kinetics. The vertical dashed line corresponds to the landmark time, pointing to the data available at this moment. (**b**) Compared data and kML-predicted hazard ratios. (**c**) Description of hazard ratios, number of patients and number of data points available in each arm, at the landmark on-study time points. ATZ, atezolizumab arm; CI, confidence interval; DTX, DTX arm; PI, prediction interval.

A strength of our study is that we relied on well-curated data with high number of patients from clinical trials. However, in other settings (e.g., earlier trial phases or real-world data), a smaller number of patients and larger heterogeneity (e.g., dosing delays or posology adaptations) are to be expected. For lower sample sizes, we found that 60 patients were sufficient to reach near-optimal performances, indicating that very large sample size might not be required as long as the number of variables remains small (**Figure S17**). Heterogeneity could be addressed with further (e.g., K-PD) modeling for dosing differences. A much larger rate of missing values might also be a limitation. Advanced missing value imputation methods should be considered and developed, especially for longitudinal data. When some markers (e.g., CRP) would not be available, kML could easily be retrained using only the available data.

Unexpectedly, kML, trained on ATZ data, yielded good predictions in the DTX (control) arm, despite a different mechanism of action (ICI vs. chemotherapy). This suggests that the relationships between TK/BK and OS were drug-independent and highlights that the empirical models were generic enough to be predictive in both cases. More importantly, kML could be used to anticipate the outcome of the ATZ vs. DTX phase III trial using phase II trials data for training and early on-study data for predictions. The positive outcome was correctly predicted at a 6.25 months landmark time after study initiation. This outperformed previous results using TK only (10 months)[10] or the OS data itself (15 months). These findings could have important implications for go/no-go decisions during drug development. They could also help to better detect futility and reassign patients, funds, and energy to other researches. Of note, a recent evaluation of this strategy has been published in the context of earlier phases, where it might be even more relevant. Bruno et al. resampled the first-line NSCLC ATZ study IMpower150 to mimic small and short follow-up of a phase 1b study. They demonstrated that TK model-based metrics had better predictive operating characteristics compared with the objective response rate or progression-free survival RECIST standard endpoints.[50] Extension of such results with the addition of BKs is a perspective of our work.

In conclusion, our study shows that integrating model-based, on-treatment longitudinal data from routine biological markers shows great promise for both personalized health care and drug development.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

FUNDING

This work was sponsored by the Roche Pharma Research and Early Development (pRED) One-D Modeling and Simulation Digital Initiative. It also benefited from funding from ITMO Cancer AVIESAN and French Institut National du Cancer (grant #19CM148-00).

CONFLICT OF INTEREST

The authors declare the following competing interests: B. Gomes, M. Stern, A. Neubert, R. Bruno, F. Mercier, S. Vatakuti and C. Jamois are Roche employees and stock holder. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

S.B., R.B., F.M., and C.J. wrote the manuscript. S.B., B.G., M.S., and C.J. designed the research. S.B., M.K., and A.E.K. performed the research. S.B., A.N., R.B., F.M., P.C., and C.J. analyzed the data. C.B. and S.V. contributed analytical tools.

DATA AVAILABILITY

Qualified researchers may request access to individual patient-level data through the clinical study data request platform (https://vivli.org/). For further details on how to request access, see https://www.roche.com/innovation/process/clinical-trials/data-sharing/.

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. & Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
2. 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).
3. Camidge, D.R., Doebele, R.C. & Kerr, K.M. Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC. *Nat. Rev. Clin. Oncol.* **16**, 341–355 (2019).
4. Hutchinson, L. & Kirk, R. High drug attrition rates—where are we going wrong? *Nat. Rev. Clin. Oncol.* **8**, 189–190 (2011).
5. Hua, T., Gao, Y., Zhang, R., Wei, Y. & Chen, F. Validating ORR and PFS as surrogate endpoints in phase II and III clinical trials for NSCLC patients: difference exists in the strength of surrogacy in various trial settings. *BMC Cancer* **22**, 1022 (2022).
6. Rizvi, H. *et al*. Molecular determinants of response to anti-programmed cell death (PD)-1 and anti-programmed death-ligand 1 (PD-L1) blockade in patients with non-small-cell lung cancer profiled with targeted next-generation sequencing. *J. Clin. Oncol.* **36**, 633–641 (2018).
7. So, W.V., Dejardin, D., Rossmann, E. & Charo, J. Predictive biomarkers for PD-1/PD-L1 checkpoint inhibitor response in NSCLC: an analysis of clinical trial and real-world data. *J. Immunother. Cancer* **11**, e006464 (2023).
8. Acosta, J.N., Falcone, G.J., Rajpurkar, P. & Topol, E.J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
9. Kurtz, D.M. *et al*. Dynamic risk profiling using serial tumor biomarkers for personalized outcome prediction. *Cell* **178**, 699–713.e19 (2019).
10. 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).
11. Chan, P. *et al*. Prediction of overall survival in patients across solid tumors following atezolizumab treatments: a tumor growth inhibition-overall survival modeling framework. *CPT Pharmacometrics Syst. Pharmacol.* **10**, 1171–1182 (2021).
12. Benzekry, S. Artificial intelligence and mechanistic modeling for clinical decision making in oncology. *Clin. Pharmacol. Ther.* **108**, 471–486 (2020).
13. Chan, P. *et al*. Application of machine learning for tumor growth inhibition—overall survival modeling platform. *CPT Pharmacometrics Syst. Pharmacol.* **10**, 59–66 (2021).
14. Ishwaran, H., Kogalur, U.B., Blackstone, E.H. & Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
15. Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y. & van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019).
16. Spigel, D.R. *et al*. FIR: efficacy, safety, and biomarker analysis of a phase II open-label study of Atezolizumab in PD-L1–selected patients with NSCLC. *J. Thorac. Oncol.* **13**, 1733–1742 (2018).
17. Peters, S. *et al*. Phase II trial of Atezolizumab As first-line or subsequent therapy for patients with programmed death-ligand 1-selected advanced non-small-cell lung cancer (BIRCH). *J. Clin. Oncol.* **35**, 2781–2789 (2017).
18. 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).
19. Eisenhauer, E.A. *et al*. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
20. Lavielle, M. *Mixed Effects Models for the Population Approach* (CRC Press, New York, NY, USA, 2014).
21. Lixoft. *Monolix Version 2020R1* (Lixoft SAS, Antony, France, 2020).
22. Stein, W.D. *et al*. Tumor growth rates derived from data for patients in a clinical trial correlate strongly with patient survival: a novel strategy for evaluation of clinical trial data. *Oncologist* **13**, 1046–1054 (2008).
23. Gavrilov, S., Zhudenkov, K., Helmlinger, G., Dunyak, J., Peskov, K. & Aksenov, S. Longitudinal tumor size and neutrophil-to-lymphocyte ratio are prognostic biomarkers for overall survival in patients with advanced non-small cell lung cancer treated with Durvalumab. *CPT Pharmacometrics Syst. Pharmacol.* **10**, 67–74 (2021).
24. Delattre, M., Lavielle, M. & Poursat, M.-A. A note on BIC in mixed-effects models. *Electron J. Stat.* **8**, 456–475 (2014).
25. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. B. Methodol.* **34**, 187–220 (1972).
26. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, New York, NY, USA, 2016).
27. Cawley, G.C. & Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
28. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B. Methodol.* **58**, 267–288 (1996).
29. Benzekry, S. nlml_onco—Software Heritage archive <https://archive.softwareheritage.org/browse/snapshot/f97ab20eb6cb44405add6531b5fa0f2d89e8d526/directory/?origin_url=https://gitlab.inria.fr/benzekry/nlml_onco>.
30. Benzekry, S. *et al*. Machine learning for prediction of immunotherapy efficacy in non-small cell lung cancer from simple clinical and biological data. *Cancer* **13**, 6210 (2021).
31. Havel, J.J., Chowell, D. & Chan, T.A. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat. Rev. Cancer* **19**, 133–150 (2019).

32. Blank, C.U., Haanen, J.B., Ribas, A. & Schumacher, T.N. The "cancer immunogram". *Science* **352**, 658–660 (2016).
33. Doroshow, D.B. *et al*. PD-L1 as a biomarker of response to immune-checkpoint inhibitors. *Nat. Rev. Clin. Oncol.* **18**, 345–362 (2021).
34. Hellmann, M.D. *et al*. Nivolumab plus Ipilimumab in lung cancer with a high tumor mutational burden. *N. Engl. J. Med.* **378**, 2093–2104 (2018).
35. Peters, S. *et al*. Atezolizumab versus chemotherapy in advanced or metastatic NSCLC with high blood-based tumor mutational burden: primary analysis of BFAST cohort C randomized phase 3 trial. *Nat. Med.* **28**, 1831–1839 (2022).
36. Soyano, A.E. *et al*. Peripheral blood biomarkers correlate with outcomes in advanced non-small cell lung cancer patients treated with anti-PD-1 antibodies. *J. Immunother. Cancer* **6**, 129 (2018).
37. Diem, S. *et al*. Neutrophil-to-lymphocyte ratio (NLR) and platelet-to-lymphocyte ratio (PLR) as prognostic markers in patients with non-small cell lung cancer (NSCLC) treated with nivolumab. *Lung Cancer* **111**, 176–181 (2017).
38. Peng, L. *et al*. Peripheral blood markers predictive of outcome and immune-related adverse events in advanced non-small cell lung cancer treated with PD-1 inhibitors. *Cancer Immunol. Immunother.* **69**, 1813–1822 (2020).
39. Becker, T. *et al*. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. *Ann. Oncol.* **31**, 1561–1568 (2020).
40. Yang, J.-R. *et al*. Post-diagnostic C-reactive protein and albumin predict survival in Chinese patients with non-small cell lung cancer: a prospective cohort study. *Sci. Rep.* **9**, 8143 (2019).
41. Bruni, D., Angell, H.K. & Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat. Rev. Cancer* **20**, 662–680 (2020).
42. Ciccolini, J., Benzekry, S. & Barlesi, F. Deciphering the response and resistance to immune-checkpoint inhibitors in lung cancer with artificial intelligence-based analysis: when PIONeeR meets QUANTIC. *Br. J. Cancer* **123**, 337–338 (2020).
43. Greillier, L. *et al*. Abstract LB120: comprehensive biomarkers analysis to explain resistances to PD1-L1 ICIs: the precision immuno-oncology for advanced non-small cell lung cancer (PIONeeR) trial. *Cancer Res.* **82**, LB120 (2022).
44. Barlesi, F. *et al*. Comprehensive biomarkers (BMs) analysis to predict efficacy of PD1/L1 immune checkpoint inhibitors (ICIs) in combination with chemotherapy: a subgroup analysis of the precision Immuno-oncology for advanced non-small CEll lung CancER (PIONeeR) trial. *Ann. Oncol.* **16**, 100108 (2022).
45. Gandara, D.R. *et al*. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nat. Med.* **24**, 1441–1448 (2018).
46. Assaf, Z.J.F. *et al*. A longitudinal circulating tumor DNA-based model associated with survival in metastatic non-small-cell lung cancer. *Nat. Med.* **29**, 859–868 (2023).
47. Nabet, B.Y. *et al*. Noninvasive early identification of therapeutic benefit from immune checkpoint inhibition. *Cell* **183**, 363–376. e13 (2020).
48. Cabel, L. *et al*. Clinical potential of circulating tumour DNA in patients receiving anticancer immunotherapy. *Nat. Rev. Clin. Oncol.* **15**, 639–650 (2018).
49. Barrera, L. *et al*. Cytokine profile determined by data-mining analysis set into clusters of non-small-cell lung cancer patients according to prognosis. *Ann. Oncol.* **26**, 428–435 (2015).
50. Bruno, R. *et al*. Tumor dynamic model-based decision support for phase Ib/II combination studies: a retrospective assessment based on resampling of the phase III study IMpower150. *Clin. Cancer Res.* **29**, 1047–1055 (2023).