Predicting survival and trial outcome for anti-PDL1 in non-small cell lung cancer using tumor and blood markers kinetics

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], **Candice Jamois**[9]

[1]COMPutational pharmacology and clinical Oncology Department, Inria Sophia Antipolis–Méditerranée, 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, Switzerland; [3]Pharma Research and Early Development, Early Development Oncology, Roche Innovation Center Zurich, Switzerland; [4]Pharma Research and Early Development, Data & Analytics, Roche Innovation Center 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; [7]Pharma Research and Early Development, Predictive Modeling and Data Analytics, Roche Innovation Center Basel, Switzerland; [8]Inovigate, Basel, Switzerland; [9]Pharma Research and Early Development, Translational PKPD and Clinical Pharmacology, Roche Innovation Center Basel, Switzerland

✉ **For correspondence:** sebastien.benzekry@inria.fr

Present address: COMPO team, Pharmacy faculty 27 Bd Jean Moulin 13385 Marseille, FRANCE

Data availability: Qualified researchers may request access to individual patient level data through the clinical study data request platform (https://vivli.org/). Further details on Roche's criteria for eligible studies are available here (https://vivli.org/members/ourmembers/). For further details on Roche's Global Policy on the Sharing of Clinical Information and how to request access to related clinical study documents, see here https://www.roche.com/innovation/process/clinical-trials/data-sharing/.

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)

Competing interests: The authors declare the existence of a financial competing interest

Abstract

Current predictive models of survival typically use baseline or tumor kinetics data. We propose a novel kinetics-machine learning (kML) integrative model of overall survival following anti-PDL1 treatment in non-small cell lung cancer. It incorporates eleven baseline markers and four on-treatment blood markers: albumin, C-reactive protein, lactate dehydrogenase and neutrophils. The kML model was developed on three phase 2 trials (862 patients) and validated on a phase 3 trial (553 patients). It outperformed the current state-of-the-art for individual predictions with a test set c-index of 0.79, a 12-months AUC of 0.86 and a hazard ratio of 25.2 (95% CI: $10.4-61.3$, $p < 0.0001$) for identification of long-term survivors. kML was also able to anticipate the success of the phase 3 trial by utilizing only 25 weeks of on-study data. It constitutes a valuable approach to support personalized medicine and drug development.

Introduction

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 case[2]. Immune-checkpoint inhibitors (ICI) (e.g., atezolizumab (ATZ)) have led to significant improvements in survival rates for pa-

tients with advanced cancers such as NSCLC[3,4]. However, there is still a large variability in clinical response and progression eventually occurs in a majority of patients[5]. Additionally, drug development in immuno-oncology is highly challenging, with a 95% attrition rate[6]. 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, overall survival (OS)[7]. 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, PDL1 expression is the only routine biomarker used for NSCLC patients[5,8] despite being controversial[9,10]. Tumor mutational burden[8,11,12] and transcriptomic data[5,13,14] have also been investigated but did not reach clinical practice. Here we posit that such static and single marker approach is intrinsically limited and that substantial additional predictive performances could be gained by: 1) using multi-modal integrative analyses relying on a combination of markers and machine learning algorithms[5,12,14,15] and 2) including dynamic markers obtained from early on-treatment data[15,16]. The nonlinear mixed-effects (NLME) modeling approach is well suited for the latter[17], and tumor kinetics (TK) model-based metrics have been shown to carry significant predictive value for OS in oncology, including ATZ monotherapy in advanced NSCLC[18–20]. The first main novelty of the current study is to establish the predictive value of model-based parameters of simple blood markers kinetics (BK), in addition to TK.

The second main novelty is to apply machine learning (ML) algorithms, increasingly used in biology and medicine[21] but only rarely for TK-OS modeling[22], instead of classical survival models. Extensions of classical ML models to survival data have been proposed (e.g., random survival forests[23]), but their actual superiority over standard approaches remains controversial[24].

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). We leveraged extensive training and testing datasets to achieve robust results. Subsequently, we tested the operational predictive capabilities of kML in two relevant scenarios: 1) individual prediction of OS and 2) prediction of the outcome of a phase 3 trial from early on-study data.

Methods

Data

For both training and external validation (testing) sets, patients from French centers were excluded for legal reasons ($N = 118$, not included in the numbers above). The training set comprised the FIR (NCT01846416)[25], POPLAR (NCT01903993)[3] and BIRCH (NCT02031458)[26] phase 2 clinical trials. The testing set was the atezolizumab arm of the OAK phase 3 trial (NCT02008227)[27]. 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.

The outcome considered was overall survival (OS), defined as the time between treatment start and death or last follow-up, in which case the data was 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 consisted of 63 variables spanning demographic and biological data, clinical information and disease status (see Appendix 1—figure 1–4 for a description of the main variables). PD-L1 expression on tumor cells was measured by immunohistochemistry or quantitative polymerase chain reaction, with four possible levels (0: $< 1\%$; 1: $\geq 1\%$; 2: $\geq 5\%$ and 3: $\geq 50\%$)[3]. We refer to the above-mentioned identifiers and references for further details on the other variables. Data were measured in accordance to the studies principles.

Missing values (1.6% total, maximum 12% in one variable) were imputed with median for numeric variables and mode for categorical variables, learned on the training set, even when applying the model to the testing set. Following preprocessing, all numeric variables were centered and scaled. Means and standard deviations was learned on the train and carried to the test set.

Dimensionality reduction for RNAseq

Initial expression data from RNAseq consisted of 715 patients and $58,311$ transcripts. The first step of data filtering removed all transcripts with less than 10 read counts for all patients, then selected genes with highest variability between patients (top $15,000$ transcripts most variable). Then, data were normalized using upper quartile normalization which consisted in dividing each read count by the 75th percentile of the read counts of the corresponding sample and the final expression values were $\log_2$ transformed. Subsequently, a univariable Cox regression model was employed to statistically assess the correlations between the expression levels of the transcripts and overall survival. Bonferroni correction was used to adjust p-values from multiple univariate tests. This step was performed using the `RegParallel` R package. We selected transcripts with high predictive values using following criteria: adjusted log rank < 0.01 and $HR < 0.85$ or $HR > 1.2$. The remaining transcripts were used to perform a bootstrap Lasso Cox regression[28] with cross-validation using mainly the `glmnet` R package. Finally, the smallest number of transcripts with best predictive model (highest c-index) was selected for further analysis.

Tumor kinetics (TK)

Patients with only one baseline SLD measurement and no SLD measurement during the treatment period were excluded ($N = 110$).

Blood markers kinetics (BK)

Only patients with at least two observations on-treatment, pre-cycle 5 (3 months) were considered. Time points prior to treatment start were discarded. Five rules were established to exclude irrelevant data points or patients from the BK dataset. First, lower (LB) and upper (UB) physiological bounds were established after discussion with a clinical oncologist. Outliers (observations outside these bounds) were discarded using the following bounds: albumin, LB = 10g/l, UB = 100g/l; CRP, no LB, UB = 300mg/l; LDH, LB = 50U/l, UB = 2000U/l; neutrophils, no LB, UB = 20G/l. Second, duplicate time points were removed, keeping the first one recorded. Third, aberrant outliers were identified. Denoting BK_n^k the value of the k-th BK at time t_n for a give patient, we excluded values such that: $\mathrm{BK}_n^k \notin \left(BK_{n-1}^k,\, BK_{n+1}^k\right)$, and $\left|\mathrm{BK}_n^k - \mathrm{BK}_{n-1}^k\right| > 3 \times sd_{BK^k}$, and $\left|\mathrm{BK}_n^k - \mathrm{BK}_{n+1}^k\right| > 3 \times \mathrm{sd}_{\mathrm{BK}^k}$, where $\mathrm{sd}_{\mathrm{BK}^k}$ is the standard deviation of $\{\mathrm{BK}_n^k\}$. Fourth, for each patient, only the BK value at the closest time point to treatment initiation was kept, provided this time point was no more than 40 days before or 10 days after treatment initiation (otherwise, patient was disregarded). Fifth, in order to have sufficient data for Bayesian parameter estimation with early data, patients with less than three data points before cycle five were removed.

Nonlinear mixed-effects modeling

Population approach

Statistical hierarchical nonlinear mixed-effects modeling (NLME) was used to implement a population approach for the kinetic data[29]. Briefly, denoting by $M\,(t;\theta)$ a structural dynamic model that depends on time t and a set of parameters θ, longitudinal observations y_j^i in patient i at time t_j^i were assumed to follow the observation model

$$y_j^i = M\left(t_j^i;\, \theta^i\right) + \varepsilon_j^i,$$

where $\varepsilon_j^i \sim \mathcal{N}\left(0,\, \sigma_j^i\right)$ is the gaussian-distributed error model. The latter was either constant ($\sigma_j^i = a,\, \forall i, j$) for TK or proportional ($\sigma_j^i = bM\,(t;\theta),\, \forall i, j$) for BK. To describe inter-individual variability,

individual parameters θ^i were assumed to follow log-normal distributions:

$$\ln(\theta^i) = \ln\left(\theta_{\text{pop}}\right) + \eta^i, \quad \eta^i \sim \mathcal{N}\left(0, \omega^2\right)$$

with population-level parameters θ_{pop} and ω. Estimation of these was performed using the stochastic approximation of expectation maximization algorithm implemented in the Monolix software[30,31].

Structural models

Following previous work, the TK structural model was assumed to be the sum of two exponentials[19,32]:

$$y_j^i = \begin{cases} y_0^i e^{\text{KG}^i t} & t \leq 0 \\ y_0^i (e^{-\text{KS}^i t} + e^{\text{KG}^i t} - 1) & t > 0 \end{cases}$$

where $t = 0$ corresponds to treatment initiation and y_0, KG and KS are three parameters. This model was also considered for BK, 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 ($y_j^i = p^i + \frac{e^{l^i}(q^i - p^i)}{t_j^i + e^{l^i}}$). Quantitative comparison of goodness-of-fit between models was assessed using the corrected Bayesian information criterion[33].

Identification of individual model-based parameters

The population parameters identified on the training set were used to define prior distributions of the TK and BK model parameters. These "training" priors were used for Bayesian estimation (maximum a posteriori estimate) of the individual TK and BK model parameters, in both the training and testing sets. To avoid biased comparison with baseline variables, the model-estimated baseline parameters (y_X^0 for X = TK, CRP, LDH and neutrophils, and albumin_q for albumin) were not kept in the TK and BKs feature sets. We additionally considered the ratio of the model-predicted value at cycle 3 day 1 to the model-estimated baseline parameter, denoted X_{ratio} for marker X. Altogether, there were three individual parameters for each marker: X_{KG}, X_{KS} and X_{ratio} for X = TK, CRP, LDH and neutrophils; and albumin_p, albumin_l and albumin_{ratio} for albumin. Once centered and scaled (from means and standard deviations derived from the training set), these individual model-based parameters constituted the TK and BKs feature sets.

Truncated data: individual-level

Individual-level truncated datasets were derived from the longitudinal TK and BK data at the following treatment cycle horizons: cycle 3 day 1 (C3D1, 1.5 months), C5D1 (3 months) and C10D1 (6.75 months). That is, for each patient, post-CXD1 values were discarded, for both the training and testing sets. New training priors were estimated from each CXD1 training set and used for Bayesian estimation of the individual parameters in the CXD1 training and testing sets. The resulting TK and BK truncated model parameter Y for marker X at cycle horizon i were denoted by $X_{\text{Y},i}$ (e.g., $ldh_{\text{KG},5}$).

Truncated data: study-level

To assess the ability of kML to early predict the final outcome of a clinical trial, we considered the two arms of the OAK phase 3 study. That is, not only the atezolizumab arm (testing set) but also the docetaxel arm (unused previously). For each arm, the data was truncated at multiple on-study landmark times lt ($lt = 10, 25$ and 60 weeks) after study initiation (first patient recruited). That is, only the patients enrolled before this time and only the data collected up to lt was used. For both arms the NLME population priors estimated from the training set (atezolizumab monotherapy) was used for Bayesian estimation of the individual model parameters, for each on-study-truncated dataset. For example, for a patient i recruited at 12 weeks, it was absent from the $lt = 10$ weeks data set, had model parameter values $\theta^i_{lt=25}$ derived from 13 weeks of on-study data in the $lt = 25$ weeks data set and different model parameter values $\theta^i_{lt=60}$ derived from 48 weeks of on-study data in the $lt = 60$ data weeks set.

Machine learning

Models

Model elaboration and development was performed exclusively on the training set, using 10 folds cross-validation for predictive performances evaluation. Due to censoring in the data, survival models were used: proportional hazards Cox regression[34], extreme gradient boosting (XGB) with either Cox or accelerated failure time (AFT) models[35] and random survival forests[23]. Nested cross-validation with inner bagging in each 10-fold cross-validation outer loop was used to evaluate the benefit of tuning the hyperparameters[36]. Improvement of the performances was negligible with hyperparameter tuning, that has higher computational cost (Appendix 1—figure 15). Therefore, we used the default values of the hyperparameters (that is, number of trees *ntree* $= 500$, number of variables to possibly split at each node *mtry* $= 5$ (rounded up square root of the number of variables), minumum size of terminal node *nodesize* $= 15$, non-negative integer specifying number of random splits for splitting a variable *nsplit* $= 10$ for the RSF model).

Evaluation

Predictive performances were assessed for either discrimination (c-index and classification metrics at horizon times τ), calibration (calibration curves) or stratification (dichotomized KM survival curves). For each individual, the RSF model gives two prediction outputs: a scalar value termed "mortality" that we will refer to as "ML score", and time-dependent predicted survival curves[23]. The former was used to compute the c-index using the `rcorr.cens` function of the `hmisc` R package[37,38].

For classification (prediction of survival at a horizon time τ), we used the latter to compute model-predicted probabilities of death at τ. Unless otherwise specified, $\tau = 12$ months. Survival-adapted metrics of predictive performance were used for sensitivity, specificity, area under the receiver-operator curve (ROC AUC) and negative and positive predictive value (NPV and PPV) to account for censoring[39,40]. For computation of accuracy, censored patients before τ were discarded ($N = 17/396$ in the test set at 12 months). Event was defined as death. Therefore, e.g., PPV corresponds to the ability of the model to correctly predict death. The optimal cut-points used for individual OS predictions on the test set were defined as the Kaplan-Meier estimated OS in the training set at τ (0.257 at 6 months, 0.437 at 12 months, 0.634 at 24 months).

For patient stratification (dichotomized KM curves), the ML score was used, with models trained on the training set and predicted on the test set. In order to assess stratification abilities to capture the 20% of long-term survivors, cut-points were set at the 20th percentiles for each variable/score evaluated. For fair comparisons, the population was also restricted to the patients with enough data to be predicted by the ML model (see preprocessing above). Significance of differences in KM curves was established using the logrank test, and hazard ratios were computed using proportional hazards Cox regression.

Variable selection and minimal signature

Three strategies were investigated to account for the multi-modal nature of the data: 1) variable selection on all the variables together, 2) variable selection per feature set (clinical, RNAseq, TK and BK) or 3) variable selection on the pooled sets resulting from 2). The general method for variable selection in a feature set was based on two steps: i) sorting the variables by importance and ii) building incremental models including increasing numbers of variables. For i), multiple algorithms were tested: Cox-sorting based on either univariate p-values or absolute hazard ratios, backward/forward stepwise selection, variable importance from RSF, or least absolute shrinkage and selection operator (LASSO)-based importance[41]. The latter was the one ultimately selected. It was defined as the sorting resulting from coefficients gradually becoming non-zero during likelihood maximization, when the regularization parameter λ decreases. For ii), the algorithm used for incremental models was RSF. Resulting c-indices and AUCs were plotted against the number of variables. Selected variables were defined as the minimal subset of variables able to reach the maximum c-index. Minimal signatures were defined as the minimal set of variables able to achieve

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

	K		CRP		LDH		Neutrophils		Albumin		
KG_{pop} ($week^{-1}$)	0.004 92	(6.80)	0.008 14	(9.38)	0.002 38	(10.48)	0.004 36	(8.69)	p_{pop} (g/l)	29.4	(3.82)
KG_{pop} ($week^{-1}$)	0.007 78	(8.22)	0.0137	(14.14)	0.001 84	(13.73)	0.000 987	(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[1]	6.82	(1.15)	0.559	(1.23)	0.138	(0.79)	0.207	(0.82)	error[1]	0.0549	(0.77)

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

a c-index larger than 0.75 and an AUC larger than 0.8.

Survival simulations and computation of predicted HRs

For each patient i, one output of the kML model is a survival curve $S^i(t)$. This gives the cumulative distribution function $1 - S^i(t)$ of the random variable T^i of the time to death for patient i, which was used to simulate 100 replicates of T^i. Pooling all patients together, we thus obtained 100 replicates of $\{T^{i,ATZ}, T^{j,DTX}\}$ for i and j being the patient indices within the ATZ and docetaxel arms, respectively. Each replicate then led to 1) a predicted survival curve in each arm and 2) a Cox proportional hazard HR between the two arms. Taking the mean and the 5th and 95th percentiles over all replicates yielded the reported point estimate and corresponding 95% prediction interval. The same procedure was used for study-truncated data.

Results

data

The data consisted of individual measurements of NSCLC patients treated with ATZ monotherapy. Three phase 2 trials were pooled into a training dataset[3,25,26] ($N = 862$ patients, Appendix 1—figure 1). The external validation (test) set comprised data from the OAK phase 3 trial ($N = 553$)[27].

Variables comprised baseline (pre-treatment) and longitudinal (on-treatment) data (Figure 1A). The former included: patients and disease characteristics ($p = 63$ variables, 43 numeric and 20 categorical, denoted BSL) and transcriptomic ("RNAseq", $p = 58,311$ transcripts) data. The latter included: longitudinal investigator-assessed sum of largest diameters (SLD) of lesions as per the RECIST criteria[42], denoted by tumor kinetics (TK, $k = 5,473/3,015$ time points in the train/test sets, respectively, median $5/4$ data points per patient, range $2/2$ —$24/20$); and longitudinal measurements of four blood markers (albumin, C-reactive protein (CRP), lactate dehydrogenase (LDH) and neutrophils), denoted together as blood markers kinetics (BK, $k = 60,779/38,460$ data points, median 11–7–11–$11/9$–9–9–10 data points per patient, range 3–3–3–$3/3$–3–3–3 —60–63–63–$78/82$–47–77–89 for albumin–CRP–LDH-neutrophils in the train/test sets, respectively). See Figure 1B and Appendix 1—figure 2–4 for details of the data and overall algorithmic procedures.

Nonlinear mixed-effects modeling (NLME) of longitudinal markers

The TK structural model was the sum of an increasing and a decreasing exponential function (double exponential model)[32]. It was able to accurately describe the data with no goodness-of-fit misspecification (Figure 2A and Appendix 1—figure 5). Population parameters were estimated with good accuracy (all relative standard errors smaller than 9%, Table 1).

To analyze the BK data, we first investigated whether significant kinetic patterns could be observed beyond random noise (due to, e.g., measurement errors, see raw data in Appendix 1—figure 6–9). The latter was considered as the null hypothesis, described by a constant model. It was tested against three alternative empiric models: linear, hyperbolic (monotonous but nonlinear and saturating) and double-exponential (nonlinear and non-monotonous). For all four BKs, we found significant kinetics compared with the constant model, as shown by lower corrected

Figure 1: Study schematic

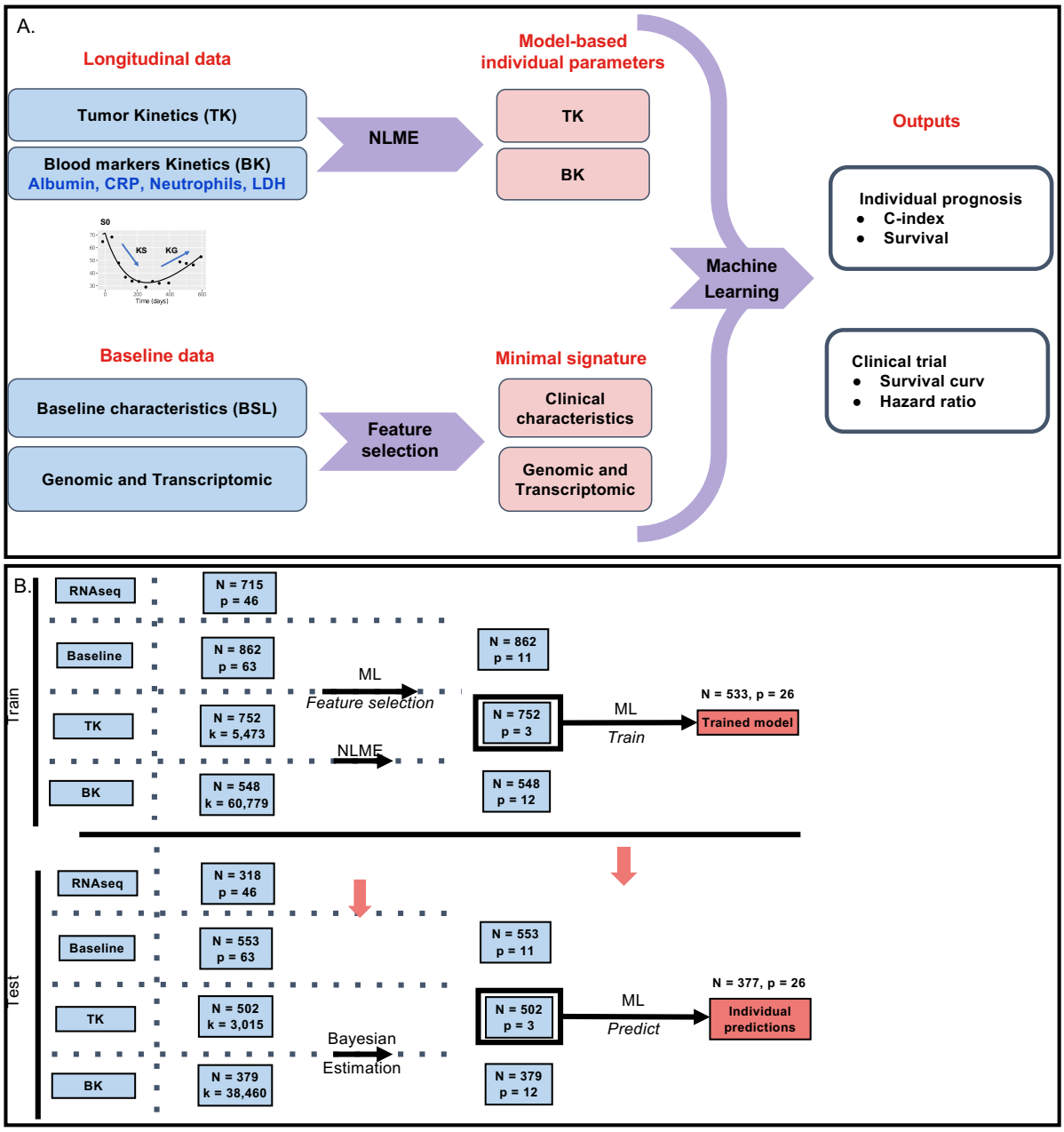


Figure 1. Study schematic A. Baseline and longitudinal data were combined into a machine learning algorithm in order to predict individual survival prognosis. Longitudinal data were modelled using nonlinear mixed-effects modelling, whereas machine learning-based feature selection was applied to the baseline 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 — were 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 modelling

Figure 2: Goodness-of-fit metrics and plots of dynamic BK models

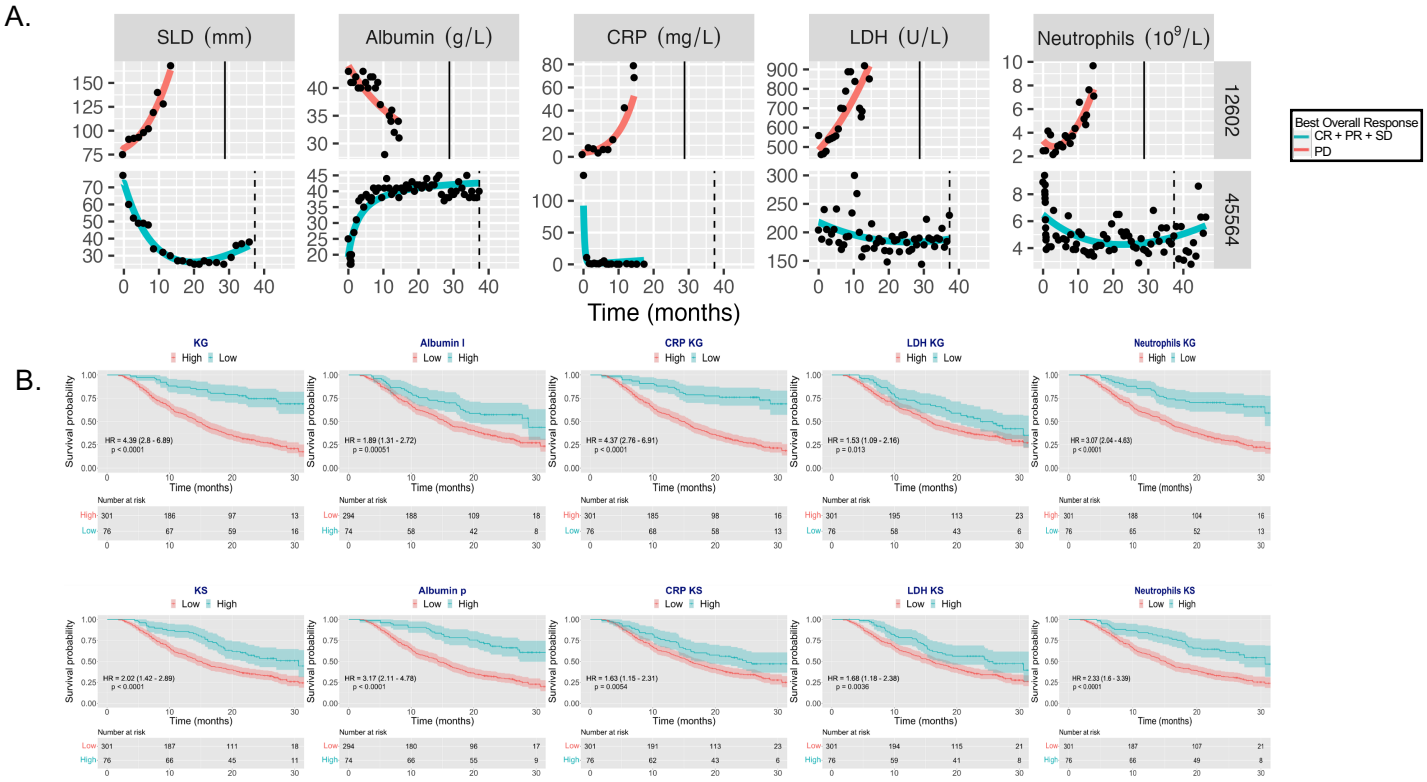


Figure 2. Goodness-of-fit metrics and plots of dynamic BK models A. Representative individual fits for the TK and BK best empirical models showing non-trivial kinetic parameters well captured by the dynamic models. Survival is indicated by a vertical line (solid = death, dashed = censored). **B.** Stratified 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; PR: partial response; SD: stable disease; PD: progressive disease.

Bayesian information criterion and relative error between model fits and data (Appendix 1—figure 10). Best descriptive models were hyperbolic for albumin and double-exponential for the other BKs. Individual fits to patient kinetics with the best models showed substantial descriptive power (Figure 2A), which was confirmed by data versus model fits plots (11–14). Parametric identifiability of population parameters 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, the following 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).

For TK and BKs we complemented the initial model parameters with an additional metric that was considered valuable for early prediction: the model-predicted ratio of change over baseline at cycle 3 day 1.

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

Four feature sets resulted from the analysis above: BSL, RNAseq, TK and BK (Figure 1B). The development of a kinetics-machine learning (kML) comprised two main steps: choice of the algorithm

and derivation of a minimal signature. They were performed using cross-validation on the training set. The first was achieved by benchmarking four models that used all variables ($p = 119$, $N = 553$). The random survival forest (RSF) model was selected as it exhibited the best performances (Appendix 1—figure 15). Notably, we found significantly better predictive performances of RSF over a Cox proportional hazard regression model ($p = 0.0006$).

Feature selection on BSL variables was performed building incremental RSF models based on LASSO importance-sorted variables (Figure 3A). 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 (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% cut-off)[3], hemoglobin[43], SLD[22] and LDH[44,45].

Applying stringent criteria to the RNAseq data (see methods), we selected 167 transcripts as candidates for final variable selection using Bolasso regression model to identify the optimal set of predictors[28]. Finally, we ended up with 52 RNAseq variables that corresponded to the highest average c-index of 0.64.

Performing incremental models with BKs, the first four LASSO-based most important features were LDH_{KG}, $neutrophils_{KG}$, $albumin_p$ and CRP_{KG} parameters were the four most important features (not shown). This indicated that the combination of all BKs was required to achieve significant predictive performances. Nevertheless, we kept all sets of three model-based parameters in the TK and BK signatures (15 parameters in total) because each set depends only on one marker (per time point).

We then compared the cross-validated c-index of each feature set on the train data (Figure 3B). Because of negligible discrimination performances (*c-index* $= 0.62 \pm 0.050$) and non-systematic availability of those data, the RNAseq set was removed from the model. The selected set of clinical data at baseline (mBSL) exhibited moderate discrimination performances (*c-index* $= 0.710 \pm 0.038$), which was slightly outperformed by the TK set (*c-index* $= 0.723 \pm 0.025$). Interestingly, the BK set significantly outperformed both baseline clinical and TK (*c-index* $= 0.793 \pm 0.038$, $p = 0.0004$ and 0.0005 respectively, Student's t-test). Jointly, mBSL, TK and BK performed significantly better than any feature set alone (*c-index* $= 0.824 \pm 0.050$, $p = 0.00007$, 0.0002 and 0.055), as well as any combination of two sets among the three (mBSL + TK: *c-index* $= 0.77 \pm 0.026$, mBSL + BK: *c-index* $= 0.81 \pm 0.027$, TK + BK: *c-index* $= 0.80 \pm 0.049$). 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 (*AUC* $= 0.919 \pm 0.056$, *accuracy* $= 0.873 \pm 0.052$, Figure 3C).

External validation

The predictive performance of the final kML model (mBSL, TK and BK) was assessed on the 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* $= 0.789$, Figure 4B, AUC for 12-months survival probability $= 0.874$). All classification metrics for prediction of survival at 12 months were high (≥ 0.78), except PPV. 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 149 patients predicted to be dead at 12 months,

Figure 3: Minimal baseline (mBSL) signature and kinetics-ML (kML) model

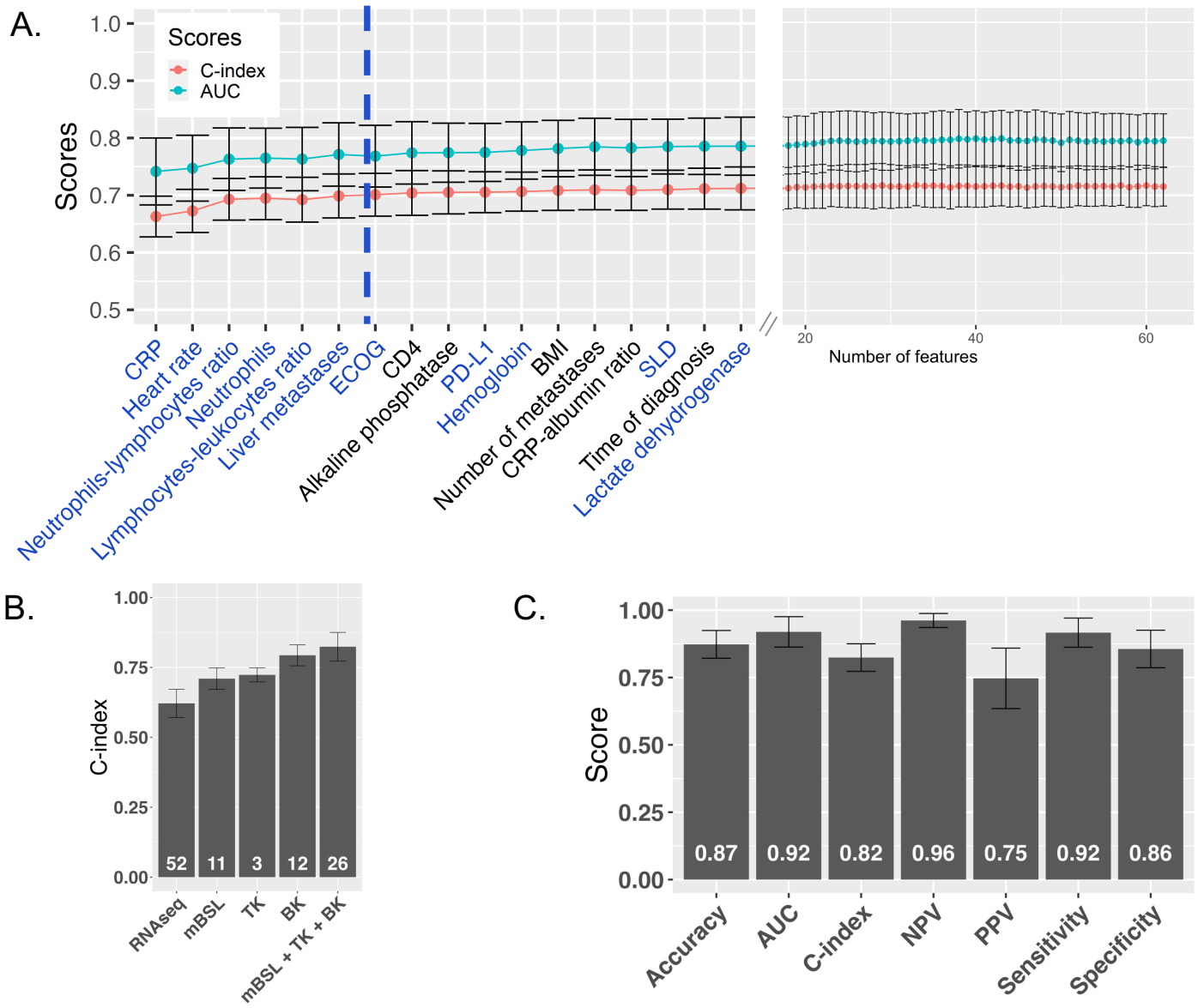


Figure 3. Minimal baseline (mBSL) signature and kinetics-ML (kML) model A. Cross-validated (CV) performance scores on the training set (c-index and AUC, mean ± standard deviation) for incremental random survival forest (RSF) models using an increasing number of baseline clinical and biological variables 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 RNAseq, 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. **C.** CV performances of the kML model for discrimination (c-index) and classification (survival prediction at 12-months OS).

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	

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

101 (67.8%) were actually deceased. Predictive AUC was good at other horizon times (0.846 and 0.910 at 6 and 24 months, respectively, Appendix 1—figure 16). 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).

Variables importance was assessed by running a post-hoc multivariable Cox regression (Figure 4F). Interestingly, the top two variables were BKs (CRP_{KG} and CRP ratio C3). In addition, TK and BK made up for six out of the seven top important features and were found more important than PD-L1.

Given the large sample size of our data, we further assessed the model performances when trained on smaller data sets (Appendix 1—figure 17). The learning curve revealed that approximately 200 patients were necessary to reach similar performance to the ones obtained with the full training set ($N = 533$), for both cross-validation and external validation on the test set (*c-index* $= 0.82 \pm 0.056$ vs *c-index* $= 0.82 \pm 0.050$ 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* $= 0.76 \pm 0.15$ 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 of future survival. To investigate the operational applicability of our methodology, data from the test set were truncated at the beginning of treatment cycles number three, five or ten, 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 and Appendix 1—figure 18). 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 (Appendix 1—figure 19A). 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

Figure 4: Predictive performances on the test set

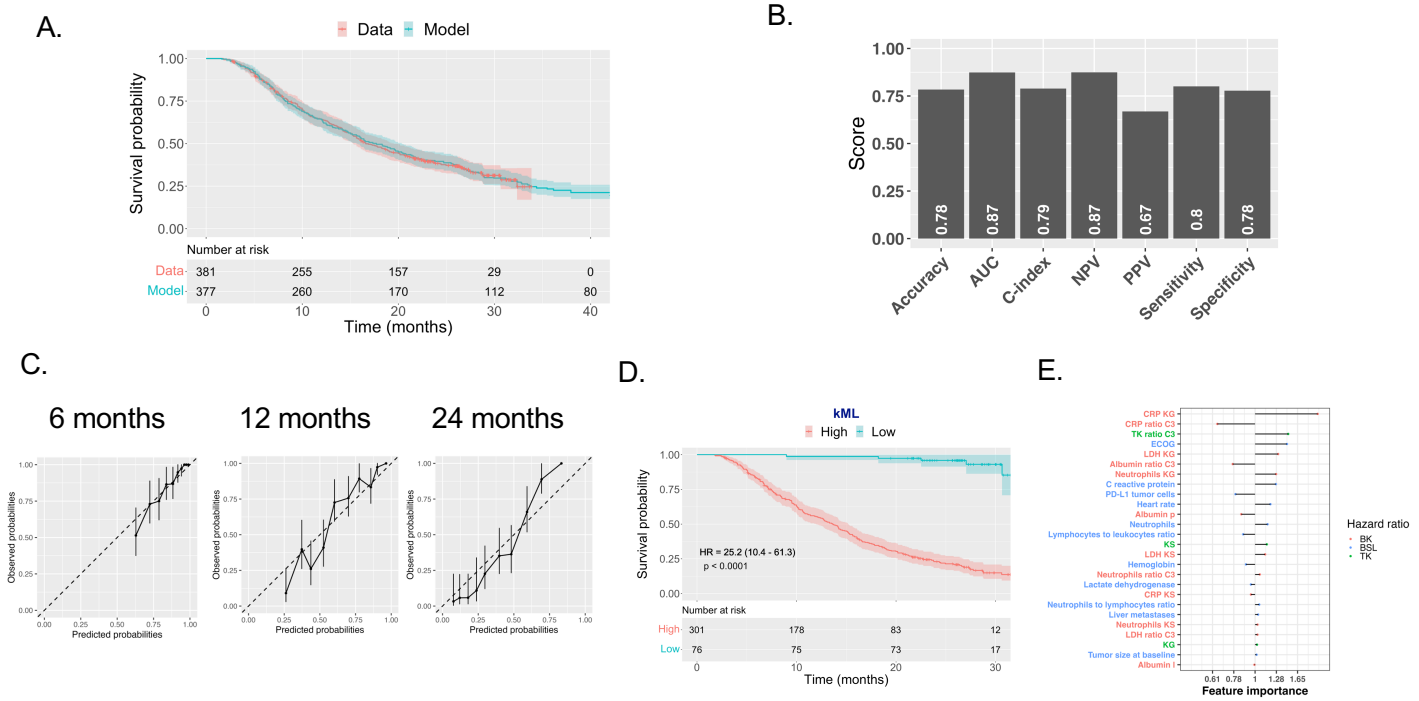


Figure 4. Predictive performances on the test set A. Comparison of the population-level survival curves between the data (KM estimator) and the model prediction. **B.** Scores of discrimination metrics. Classification metrics were computed for prediction of OS at 12 months. **C.** Calibration curves 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 versus low), at the 20th percentile cut-off. **E.** Variables importance (multivariable hazard ratios) in the full time-course kML model.

Figure 5: Individual-level predictions from cycle-truncated data

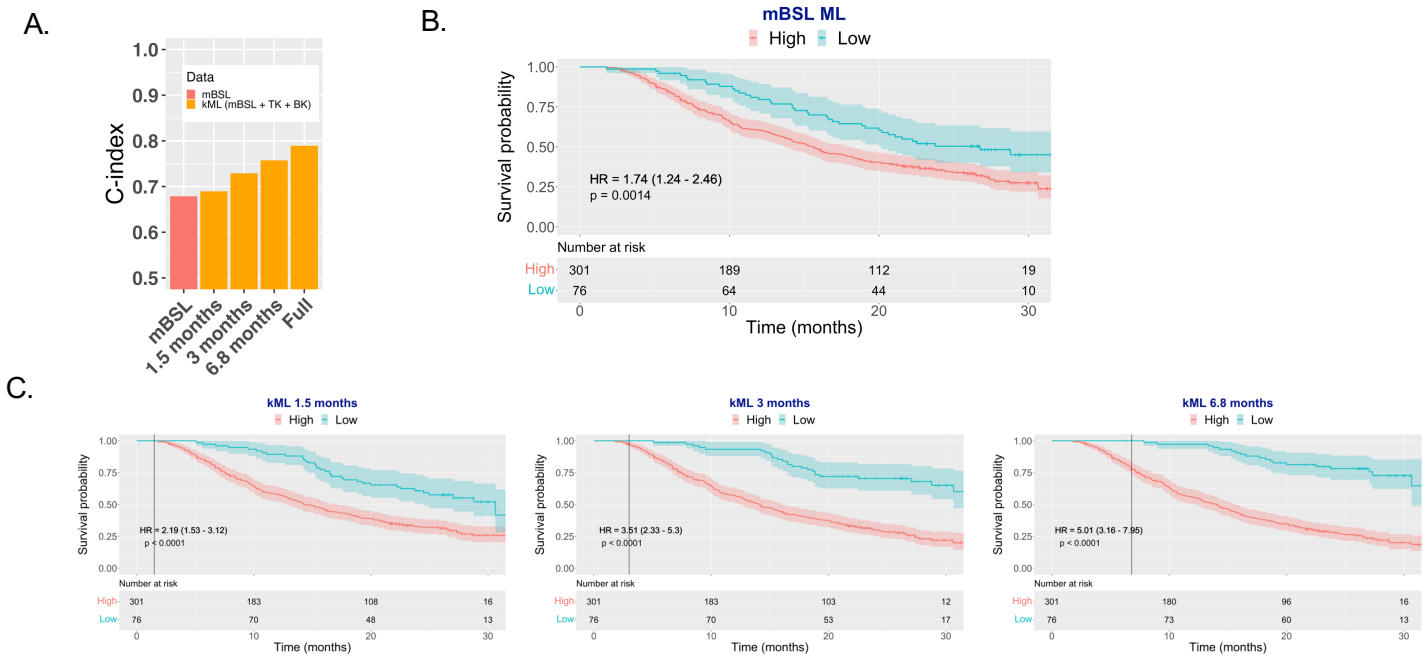


Figure 5. Predictive value from cycle-truncated data A. Predictive power (c-index) of ML models using baseline (BSL) or truncated data at 1.5, 3 and 6.8 months as well as the full time-course. **B.** Stratified KM survival curves using a RSF model trained on the minimal baseline (mBSL) variables. **C.** Stratified KM survival curves using kML from 1.5 months (2 cycles), 3 months (4 cycles) and 6.8 months (9 cycles) truncated data. Truncation time is indicated by the vertical line.
TK: tumor kinetics; BK: biological kinetics; LDH: lactate dehydrogenase; CRP: C-reactive protein.

driven by any single biomarker alone.

Interestingly, the most important variable at 1.5 months was a kinetic one, TK ratio C3 with following variables being from mBSL (e.g., liver metastases, PDL1 and ECOG). When more on-treatment variables become available, this shifted to TK and BK (TK ratio C3, TK_{KS}, TK_{KG}, CRP_{KG}, LDH_{KG}), see Appendix 1—figure 19B.

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. To this respect, a different truncation needs to be performed. Indeed, the quantity of data available is not determined by the time a given patient has received the treatment, but rather by the recruitment rate and the number of patients and associated data available at a given on-study landmark time. We thus performed such truncations on the test set based on a number of weeks after the date of the first patient recruited. Predictions of the kML model applied to each arm (atezolizumab and docetaxel) yielded very accurate results when using data from the entire study (predicted HR $= 0.784$ $(0.7-0.842)$), versus 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. 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 very small follow-up. Starting from data available at 25 weeks (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 data at this time (dashed lines, Figure 6A and red HR CIs in Figure 6B) was far from being conclusive. The model prediction was stable from 25 weeks on whereas the data only exhibited significant HR from 60 weeks and required more than 300 patients in each arm to

be conclusive.

Discussion

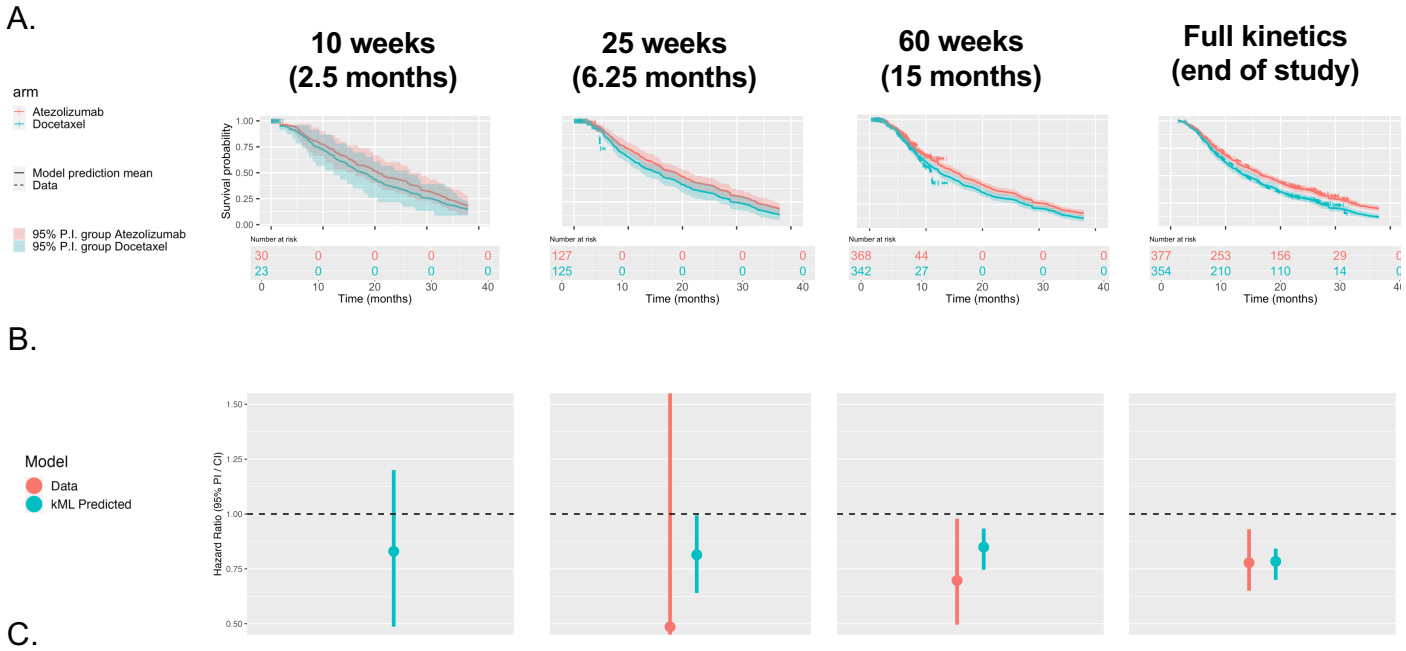
Blood markers from hematology and biochemistry are routinely collected during clinical care or drug trials. They are cost-effective and easily obtained both before and during treatment. There is limited exploration regarding the predictive capabilities of the kinetics of such data. Combining BSL variables with on-treatment data (TK and BK), we addressed this question using a novel hybrid NLME–ML methodology. The resulted kML model demonstrated excellent predictive performances for OS in two aspects: 1) patient-level predictions (discrimination, calibration and patient stratification) and 2) trial-level predictions. The kML model outperformed current state-of-the-art methods based on either baseline or on-treatment data alone, utilizing only routine clinical information, with a c-index of 0.79 and an area under the curve (AUC) of 0.87 on the test dataset. Overall, kML incorporates 26 features, out of which 15 features require monitoring five quantities over time (tumor size, albumin, CRP, LDH and neutrophils).

Regarding baseline markers, the predictive value of PD-L1 expression, commonly used in clinical care, is controversial[9,10]. Previous studies reported an AUC for durable response of 0.601 and a PFS HR of 1.90 (PD-L1 $\geq 1\%$ vs 0%)[8]. Baseline tumor mutational burden showed similar predictive value initially ($AUC = 0.646$)[11], but led to disappointing results in a recent prospective study[46] and others found it to be more prognostic than predictive[47]. Baseline blood counts were previously reported to predict overall survival[44,48–50] and treatment response ($AUC = 0.74$)[43]. The ROPRO score, derived from a large pan-cancer cohort and incorporating baseline clinical and biological data (27 variables) achieved a c-index of 0.69 and a 3-months AUC of 0.743 for prediction of survival in the OAK clinical trial[51]. Here, we confirmed these findings and established a minimal signature of such data composed of only 11 variables (CRP, heart rate, neutrophils to lymphocytes ratio, neutrophils, lymphocytes to leukocytes ratio, liver metastases, ECOG, PD-L1 $\geq 50\%$, hemoglobin, SLD and LDH), yet with similar predictive performances ($c\text{-}index = 0.678$) and significant stratification ability ($HR = 1.74$, $p = 0.0014$). Altogether, our kML model demonstrated substantially better predictive performances than these baseline models.

The main novelty of our work lies in the use of on-treatment blood markers kinetics (BK). We first further confirmed the established predictive value of TK model-based parameters[19,20]. Blood- or serum-derived longitudinal markers kinetics have to date rarely been modeled. Gavrilov et al. proposed to model NLR kinetics and demonstrated improved OS predictions over TK alone[52]. Here we extended to four BKs: albumin, CRP, LDH and neutrophils. This choice was not only motivated by observed statistical associations, but also from biological considerations. Albumin is associated with nutritional status (cachexic state) and is known to evolve with time in responders. CRP is a marker of systemic inflammation[45]. Increased CRP, decreased albumin level, and increased CRP/albumin ratio have been reported to be associated with poor survival[53]. Neutrophils play a role in inflammation by promoting a favorable microenvironment for cancer cell growth and spread, and activation of carcinogenic signaling pathways[54]. Elevated LDH levels are a marker of cancer cells turnover rate, and LDH has a potential role for prediction of potential invisible metastases[45]. We found that all these markers had non-trivial on-treatment kinetics. However, data fits were not perfect, possibly due to the simplicity and empiric nature of the models we used. Further mechanistic modeling of the joint kinetics of BKs and TK could bring relevant biological information and yield more accurate predictive parameters. We found that all four BKs were contributive to the model and that, combined, they outperformed TK performances.

We analyzed the RNAseq data using standard methods and found only negligible predictive performances. Such result could be explained by the fact that the tissue of origin that was used was heterogeneous across the patients (primary tumor or metastasis), was limited to a local area of the tumor, and could come from tissue sampled long before treatment initiation. Given that our main objective was to derive a predictive model from markers available in routine practice, we

Figure 6: Use of kML for early prediction of the outcome of a clinical trial



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 versus actual data from on-study data at multiple horizon times after study initiation. Note that the model is able to predict full survival curves even if based on early kinetics. **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.
PI: prediction interval, CI: confidence interval, DTX: docetaxel arm, ATZ: atezolizumab arm.

excluded it from our minimal signature. A refined analysis, especially focusing on immune-based signatures, could improve our results[5].

Machine learning models, although increasingly used in pharmacological studies —including recently for TK-OS modeling and variable selection[22,55] —have yet rarely been rigorously compared to classical statistical models[24]. Here, such comparison revealed significantly better performance of the nonlinear random survival forest RSF model compared to the linear proportional hazards Cox model. In our approach, we did not use the propagation of standard statistical quantification of the parameters' estimates uncertainty to evaluate the accuracy of the model predictions. Rather, we relied on the RSF-outputted individual survival curves to sample virtual individuals and compute prediction intervals.

A drawback of classical TK–OS studies is that they make use of the full observed kinetics to predict overall survival, which can lead to time-dependent covariate bias[56] and limit their practical applicability at bedside. We used individual-truncated data sets and found that kML was already improving predictions over mBSL using data at 1.5 months, which corresponds to the first imaging assessment of the treatment effect. At later times, stratification abilities increased to highly significant levels (e.g., $HR = 5$ at 6.8 months).

A strength of our study is that we relied on well-curated data with high number of patients from clinical trials. However, when extrapolating to other settings — earlier trial phases, real-world data — limited number of patients might be available. Yet, we found that using only 60 patients to train kML was sufficient to reach near-optimal performances.

Not only kML has value for personalized health care, but it revealed also useful for prediction of a phase 3 trial using early on-study data. Our predictions compared favorably with previous work, being able to predict the study's positive outcome from 25 weeks onwards, versus 40 weeks using TK only[19] and 60 weeks if relying on the observed data alone. Of note, in a recent evaluation based on resampling the first-line NSCLC ATZ study IMpower150 to mimic small, short follow up early Phase Ib studies, TK model-based metrics had better operating characteristics to predict Phase III success compared with RECIST endpoints ORR and PFS[55]. Extension of such results with the addition of BKs is thus a promising line of research. In addition, kML, trained on ATZ data, yielded excellent predictive abilities for the docetaxel (control) arm. This suggests that the relationships between TK / BK and OS might be drug-independent. In turn, this opens future perspectives in terms of testing kML on drugs with different mechanism of action, or combinations.

Further avenues of research comprise the development of integrative models from advanced multi-modal data such as the one collected during the PIONeeR clinical study (NCT03833440)[57,58] that include quantitative image analysis from multiplex immune–histochemistry staining of tumor tissue, genomic and transcriptomic data, biological and clinical markers. In addition, mechanistic modeling of quantitative and physiologically meaningful longitudinal data (immune-monitoring, vasculo-monitoring, circulating DNA[12,59–61], soluble factors[62], pharmacokinetics, TK and a large number of BKs from either hematology or biochemistry) paves the way to an improved understanding and prediction of mechanisms of relapse to ICI[63]. Furthermore, the predictive abilities of kML — at the both individual and study levels — should be evaluated in model-based prospective trials[64].

In conclusion, our study shows that integrating model-based on-treatment dynamic data from routine biological markers shows great promise for both personalized health care and early prediction of the outcome of late-phase trials during drug development.

References

1. Bray, F., Ferlay, J., *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians,* 394–424. ISSN: 1542-4863. doi:10.3322/caac.21492 (2018).

2. Duma, N., Santana-Davila, R. & Molina, J. R. Non–Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment. *Mayo Clinic Proceedings,* 1623–1640. ISSN: 0025-6196, 1942-5546. doi:10.1016/j.mayocp.2019.01.013 (2019).

3. Fehrenbacher, L., Spira, A., *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. *The Lancet,* 1837–1846. ISSN: 0140-6736, 1474-547X. doi:10.1016/S0140-6736(16)00587-0 (2016).

4. Grant, M. J., Herbst, R. S. & Goldberg, S. B. Selecting the optimal immunotherapy regimen in driver-negative metastatic NSCLC. *Nature Reviews Clinical Oncology,* 625–644. ISSN: 1759-4782. doi:10.1038/s41571-021-00520-1 (2021).

5. Camidge, D. R., Doebele, R. C. & Kerr, K. M. Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC. *Nature Reviews Clinical Oncology,* 341–355. ISSN: 1759-4782. doi:10.1038/s41571-019-0173-9 (2019).

6. Hutchinson, L. & Kirk, R. High drug attrition rates—where are we going wrong? *Nature Reviews Clinical Oncology,* 189–190. ISSN: 1759-4782. doi:10.1038/nrclinonc.2011.34 (2011).

7. 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,* 1022. ISSN: 1471-2407. doi:10.1186/s12885-022-10046-z (2022).

8. Rizvi, H., Sanchez-Vega, F., *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. *Journal of Clinical Oncology.* doi:10.1200/JCO.2017.75.3384 (2018).

9. Doroshow, D. B., Bhalla, S., *et al.* PD-L1 as a biomarker of response to immune-checkpoint inhibitors. *Nature Reviews Clinical Oncology,* 345–362. ISSN: 1759-4782. doi:10.1038/s41571-021-00473-5 (2021).

10. 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. *Journal for Immunotherapy of Cancer,* e006464. ISSN: 2051-1426. doi:10.1136/jitc-2022-006464 (2023).

11. Hellmann, M. D., Ciuleanu, T.-E., *et al.* Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *New England Journal of Medicine,* 2093–2104. doi:10.1056/NEJMoa1801946 (2018).

12. Gandara, D. R., Paul, S. M., *et al.* Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nature Medicine,* 1441–1448. ISSN: 1546-170X. doi:10.1038/s41591-018-0134-3 (2018).

13. Cristescu, R., Mogg, R., *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade–based immunotherapy. *Science,* eaar3593. doi:10.1126/science.aar3593 (2018).

14. Sankar, K., Ye, J. C., *et al.* The role of biomarkers in personalized immunotherapy. *Biomarker Research,* 32. ISSN: 2050-7771. doi:10.1186/s40364-022-00378-0 (2022).

15. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nature Medicine,* 1773–1784. ISSN: 1546-170X. doi:10.1038/s41591-022-01981-2 (2022).

16. Kurtz, D. M., Esfahani, M. S., *et al.* Dynamic Risk Profiling Using Serial Tumor Biomarkers for Personalized Outcome Prediction. *Cell,* 699–713.e19. ISSN: 1097-4172. doi:10.1016/j.cell.2019.06.011 (2019).

17. Bonate, P. L. *Pharmacokinetic-Pharmacodynamic Modeling and Simulation* 2nd ed. 2011. ISBN: 978-1-4419-9484-4 (Springer-Verlag New York Inc., New York, 2011).

18. Claret, L., Girard, P., *et al.* Model-based prediction of phase III overall survival in colorectal cancer on the basis of phase II tumor dynamics. *J Clin Oncol,* 4103–4108. doi:10.1200/JCO.2008.21.0807 (2009).

19. Claret, L., Jin, J. Y., *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,* 3292–3298. doi:10.1158/1078-0432.CCR-17-3662 (2018).

20. Chan, P., Marchand, M., *et al.* Prediction of overall survival in patients across solid tumors following atezolizumab treatments: A tumor growth inhibition-overall survival modeling framework. *CPT: pharmacometrics & systems pharmacology,* 1171–1182. ISSN: 2163-8306. doi:10.1002/psp4.12686 (2021).

21. Benzekry, S. Artificial intelligence and mechanistic modeling for clinical decision making in oncology. *Clinical Pharmacology & Therapeutics,* 471–486. ISSN: 1532-6535. doi:10.1002/cpt.1951 (2020).

22. Chan, P., Zhou, X., *et al.* Application of Machine Learning for Tumor Growth Inhibition - Overall Survival Modeling Platform. *CPT: pharmacometrics & systems pharmacology,* 59–66. ISSN: 2163-8306. doi:10.1002/psp4.12576 (2021).

23. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics,* 841–860. ISSN: 1932-6157, 1941-7330. doi:10.1214/08-AOAS169 (2008).

24. Christodoulou, E., Ma, J., *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology,* 12–22. ISSN: 0895-4356. doi:10.1016/j.jclinepi.2019.02.004 (2019).

25. Spigel, D. R., Chaft, J. E., *et al.* FIR: Efficacy, Safety, and Biomarker Analysis of a Phase II Open-Label Study of Atezolizumab in PD-L1–Selected Patients With NSCLC. *Journal of Thoracic Oncology,* 1733–1742. ISSN: 1556-0864. doi:10.1016/j.jtho.2018.05.004 (2018).

26. Peters, S., Gettinger, 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). *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology,* 2781–2789. ISSN: 1527-7755. doi:10.1200/JCO.2016.71.9476 (2017).

27. Rittmeyer, A., Barlesi, F., *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. *The Lancet,* 255–265. ISSN: 0140-6736, 1474-547X. doi:10.1016/S0140-6736(16)32517-X (2017).

28. Bach, F. R. *Bolasso: model consistent Lasso estimation through the bootstrap* in (Association for Computing Machinery, New York, NY, USA, 2008), 33–40. ISBN: 978-1-60558-205-4. doi:10.1145/1390156.1390161.

29. Lavielle, M. *Mixed Effects Models for the Population Approach* ISBN: 1-4822-2650-2 (CRC Press, 2014).

30. Delyon, B., Lavielle, M. & Moulines, E. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics,* 94–128. ISSN: 0090-5364, 2168-8966. doi:10.1214/aos/1018031103 (1999).

31. Lixoft. *Monolix version 2020R1.* Antony, France, 2020.

32. Stein, W. D., Figg, 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. *The Oncologist,* 1046–1054. doi:10.1634/theoncologist.2008-0075 (2008).

33. Delattre, M., Lavielle, M. & Poursat, M.-A. A note on BIC in mixed-effects models. *Electronic Journal of Statistics,* 456–475. ISSN: 1935-7524, 1935-7524. doi:10.1214/14-EJS890 (2014).

34. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological),* 187–220. ISSN: 0035-9246 (1972).

35. 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* (Association for Computing Machinery, San Francisco, California, USA, 2016), 785–794. ISBN: 9781450342322. doi:10.1145/2939672.2939785.

36. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research,* 2079–2107 (2010).

37. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA,* 2543–2546. ISSN: 0098-7484 (1982).

38. Harrell, F. E. Hmisc: Harrell miscellaneous. *R package.* https://CRAN.R-project.org/package=Hmisc (2022).

39. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics,* 337–344. ISSN: 0006-341X. doi:10.1111/j.0006-341x.2000.00337.x (2000).

40. Heagerty, P. J. & Saha-Chaudhuri, p. b. P. *survivalROC: Time-dependent ROC curve estimation from censored survival data* tech. rep. (2013). https://CRAN.R-project.org/package=survivalROC.

41. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 267–288. ISSN: 0035-9246 (1996).

42. Eisenhauer, E. A., Therasse, P., *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer. Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers* 228–247. ISSN: 0959-8049. doi:10.1016/j.ejca.2008.10.026 (2009).

43. Benzekry, S., Grangeon, M., *et al.* Machine Learning for Prediction of Immunotherapy Efficacy in Non-Small Cell Lung Cancer from Simple Clinical and Biological Data. *Cancers,* 6210. doi:10.3390/cancers13246210 (2021).

44. Havel, J. J., Chowell, D. & Chan, T. A. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nature Reviews Cancer,* 133–150. ISSN: 1474-1768. doi:10.1038/s41568-019-0116-x (2019).

45. Blank, C. U., Haanen, J. B., Ribas, A. & Schumacher, T. N. The "cancer immunogram". *Science,* 658–660. doi:10.1126/science.aaf2834 (2016).

46. Peters, S., Dziadziuszko, R., *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. *Nature Medicine,* 1831–1839. ISSN: 1546-170X. doi:10.1038/s41591-022-01933-w (2022).

47. Liu, Y., Wang, B., Tian, H. & Hsu, J. C. Rejoinder for discussions on correct and logical causal inference for binary and time-to-event outcomes in randomized controlled trials. *Biometrical Journal,* 246–255. ISSN: 1521-4036. doi:10.1002/bimj.202100089 (2022).

48. Soyano, A. E., Dholaria, B., *et al.* Peripheral blood biomarkers correlate with outcomes in advanced non-small cell lung Cancer patients treated with anti-PD-1 antibodies. *Journal for Immunotherapy of Cancer,* 129. ISSN: 2051-1426. doi:10.1186/s40425-018-0447-2 (2018).

49. Diem, S., Schmid, 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 (Amsterdam, Netherlands),* 176–181. ISSN: 1872-8332. doi:10.1016/j.lungcan.2017.07.024 (2017).

50. Peng, L., Wang, Y., *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 immunology, immunotherapy: CII,* 1813–1822. ISSN: 1432-0851. doi:10.1007/s00262-020-02585-w (2020).

51. Becker, T., Weberpals, J., *et al.* An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. *Annals of Oncology,* 1561–1568. ISSN: 0923-7534. doi:10.1016/j.annonc.2020.07.013 (2020).

52. Gavrilov, S., Zhudenkov, K., *et al.* 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 & systems pharmacology,* 67–74. ISSN: 2163-8306. doi:10.1002/psp4.12578 (2021).

53. Yang, J.-R., Xu, J.-Y., *et al.* Post-diagnostic C-reactive protein and albumin predict survival in Chinese patients with non-small cell lung cancer: a prospective cohort study. *Scientific Reports,* 8143. ISSN: 2045-2322. doi:10.1038/s41598-019-44653-x (2019).

54. Bruni, D., Angell, H. K. & Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nature Reviews Cancer,* 662–680. ISSN: 1474-1768. doi:10.1038/s41568-020-0285-7 (2020).

55. Bruno, R., Marchand, M., *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. *Clinical Cancer Research,* OF1–OF9. ISSN: 1078-0432. doi:10.1158/1078-0432.CCR-22-2323 (2023).

56. Desmée, S., Mentré, F., Veyrat-Follet, C. & Guedj, J. Nonlinear Mixed-Effect Models for Prostate-Specific Antigen Kinetics and Link with Survival in the Context of Metastatic Prostate Cancer: a Comparison by Simulation of Two-Stage and Joint Approaches. *The AAPS Journal,* 691–699. ISSN: 1550-7416. doi:10.1208/s12248-015-9745-5 (2015).

57. Greillier, L., Monville, F., *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 Research,* LB120. ISSN: 0008-5472. doi:10.1158/1538-7445.AM2022-LB120 (2022).

58. Barlesi, F., Monville, 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. *Annals of Oncology.* doi:10.1016/iotech/iotech100100 (2022).

59. Assaf, Z. J. F., Zou, W., *et al.* A longitudinal circulating tumor DNA-based model associated with survival in metastatic non-small-cell lung cancer. *Nature Medicine,* 859–868. ISSN: 1546-170X. doi:10.1038/s41591-023-02226-6 (2023).

60. Nabet, B. Y., Esfahani, M. S., *et al.* Noninvasive Early Identification of Therapeutic Benefit from Immune Checkpoint Inhibition. *Cell,* 363–376.e13. ISSN: 1097-4172. doi:10.1016/j.cell.2020.09.001 (2020).

61. Cabel, L., Proudhon, C., *et al.* Clinical potential of circulating tumour DNA in patients receiving anticancer immunotherapy. *Nature Reviews. Clinical Oncology,* 639–650. ISSN: 1759-4782. doi:10.1038/s41571-018-0074-3 (2018).

62. Barrera, L., Montes-Servín, E., *et al.* Cytokine profile determined by data-mining analysis set into clusters of non-small-cell lung cancer patients according to prognosis. *Annals of Oncology: Official Journal of the European Society for Medical Oncology,* 428–435. ISSN: 1569-8041. doi:10.1093/annonc/mdu549 (2015).

63. 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. *British Journal of Cancer,* 1–2. ISSN: 1532-1827. doi:10.1038/s41416-020-0918-3 (2020).

64. Ciccolini, J., Barbolosi, D., André, N., Barlesi, F. & Benzekry, S. Mechanistic Learning for Combinatorial Strategies With Immuno-oncology Drugs: Can Model-Informed Designs Help Investigators? *JCO Precision Oncology,* 486–491. doi:10.1200/PO.19.00381 (2020).

Appendix 1

Supplementary figures

Study	Description	Population	N
FIR GO28625	Phase 2 study for the efficacy and safety of ATZ in advanced NSCLC	PD-L1 positive locally advanced or metastatic NSCLC (lines 1 and 2+)	133
POPLAR GO28753	Phase 2 randomised controlled trial of ATZ versus docetaxel in NSCLC	Locally advanced or metastatic NSCLC who failed platinum therapy (line 2)	134
BIRCH GO28754	Phase 2 study of ATZ in advanced or metastatic NSCLC	Locally advanced or metastatic NSCLC (lines 1, 2 or 3)	595
Train			**862**
Test - OAK GO28915	Phase 3 RCT of ATZ versus docetaxel (DTX) in patients with previously treated NSCLC	Stage IIIB or IV, previously chemo treated	**553**
Train + Test			**1415**

Four monotherapy studies of atezolizumab in advanced NSCLC. NSCLC: Non-Small Cell Lung Cancer; p = number of parameters, N: number of patients treated with atezolizumab (patients from French centers were excluded for legal reasons (N=118); In total, data from 1074 patients from OAK were used as Test set (553 from the ATZ arm, 521 from the DTX arm); PD: Pharmacodynamic; SLD: Sum of the Largest Diameters. CRP: C Reactive Protein; LDH: Lactate Dehydrogenase.

1

Appendix 1—figure 1. Train and test data sets

Characteristic	Total, N = 1415[1]	FIR, N = 133[1]	POPLAR, N = 134[1]	BIRCH, N = 595[1]	OAK, N = 553[1]	p-value[2]
Age	64 (57, 70)	67 (60, 73)	62 (55, 69)	65 (57, 71)	64 (57, 70)	<0.001
Sex						0.3
Female	568 (40%)	57 (43%)	47 (35%)	251 (42%)	213 (39%)	
Male	847 (60%)	76 (57%)	87 (65%)	344 (58%)	340 (61%)	
Weight	72 (61, 82)	70 (60, 83)	73 (63, 84)	72 (61, 82)	71 (60, 82)	0.3
BMI	24.9 (22.1, 28.1)	24.8 (21.9, 27.6)	25.2 (22.8, 28.7)	25.0 (22.1, 28.2)	24.7 (22.0, 28.1)	0.4
Unknown	65	8	5	30	22	
Race						<0.001
Asian	228 (16%)	6 (4.5%)	23 (17%)	77 (13%)	122 (22%)	
Others, unknown or missing	73 (5.2%)	9 (6.8%)	9 (6.7%)	23 (3.9%)	32 (5.8%)	
White	1,114 (79%)	118 (89%)	102 (76%)	495 (83%)	399 (72%)	
Smoking history						0.13
Current	171 (12%)	18 (14%)	22 (16%)	60 (10%)	71 (13%)	
Never	248 (18%)	16 (12%)	27 (20%)	102 (17%)	103 (19%)	
Previous	996 (70%)	99 (74%)	85 (63%)	433 (73%)	379 (69%)	

Characteristic	Total, N = 1415[1]	FIR, N = 133[1]	POPLAR, N = 134[1]	BIRCH, N = 595[1]	OAK, N = 553[1]	p-value[2]
Heart rate	81 (71, 92)	80 (70, 94)	84 (74, 96)	80 (70, 90)	81 (72, 93)	0.049
Systolic blood pressure	122 (111, 133)	122 (113, 132)	122 (112, 131)	120 (110, 133)	123 (113, 135)	0.13

[1]Median (IQR); n (%)
[2]Kruskal-Wallis rank sum test; Pearson's Chi-squared test

Appendix 1—figure 2. Patient characteristics: demographics and clinics

Characteristic	Total, N = 1415[1]	FIR, N = 133[1]	POPLAR, N = 134[1]	BIRCH, N = 595[1]	OAK, N = 553[1]	p-value[2]
Disease type						0.4
Locally advanced	77 (5.4%)	3 (2.3%)	8 (6.0%)	32 (5.4%)	34 (6.1%)	
Metastatic	1,338 (95%)	130 (98%)	126 (94%)	563 (95%)	519 (94%)	
Line						<0.001
≥2	1,255 (89%)	102 (77%)	134 (100%)	466 (78%)	553 (100%)	
1	160 (11%)	31 (23%)	0 (0%)	129 (22%)	0 (0%)	
Histology						<0.001
Non-squamous	1,016 (74%)	95 (100%)	87 (65%)	427 (72%)	407 (74%)	
Squamous	361 (26%)	0 (0%)	47 (35%)	168 (28%)	146 (26%)	
Unknown	38	38	0	0	0	
Stage						<0.001
I	123 (8.9%)	11 (8.7%)	4 (3.0%)	73 (12%)	35 (6.5%)	
II	140 (10%)	11 (8.7%)	9 (6.8%)	73 (12%)	47 (8.7%)	
III	367 (27%)	30 (24%)	37 (28%)	165 (28%)	135 (25%)	
IV	753 (54%)	75 (59%)	82 (62%)	273 (47%)	323 (60%)	
Unknown	32	6	2	11	13	
Number of metastases						0.4
1	386 (28%)	35 (27%)	34 (26%)	158 (27%)	159 (29%)	
2	655 (48%)	55 (42%)	61 (47%)	289 (50%)	250 (46%)	
3	334 (24%)	41 (31%)	34 (26%)	128 (22%)	131 (24%)	
Unknown	40	2	5	20	13	
Liver metastases	272 (19%)	29 (22%)	33 (25%)	105 (18%)	105 (19%)	0.3
Number of met. loc.						0.3
Four sites	70 (4.9%)	13 (9.8%)	6 (4.5%)	27 (4.5%)	24 (4.3%)	
One site	426 (30%)	37 (28%)	39 (29%)	178 (30%)	172 (31%)	
Three sites	264 (19%)	28 (21%)	28 (21%)	101 (17%)	107 (19%)	
Two sites	655 (46%)	55 (41%)	61 (46%)	289 (49%)	250 (45%)	
Tumor size	59 (43, 95)	59 (57, 59)	70 (43, 107)	60 (37, 96)	70 (43, 102)	0.001

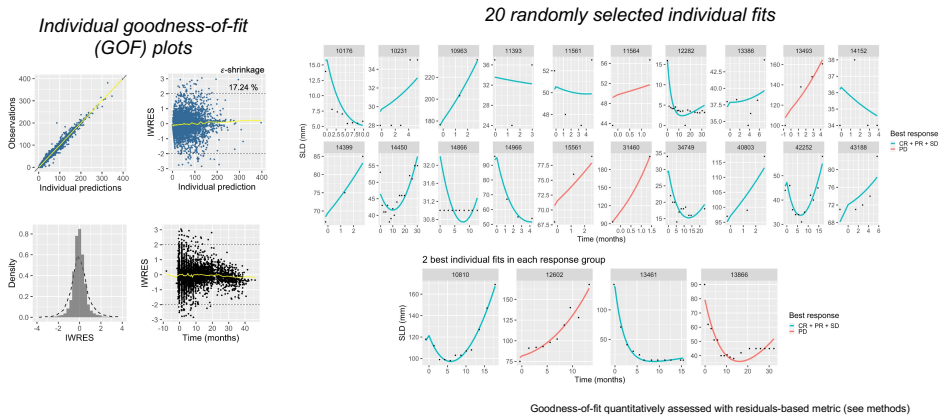
[1]n (%); Median (IQR)
[2]Pearson's Chi-squared test; Kruskal-Wallis rank sum test

Appendix 1—figure 3. Patient characteristics: disease

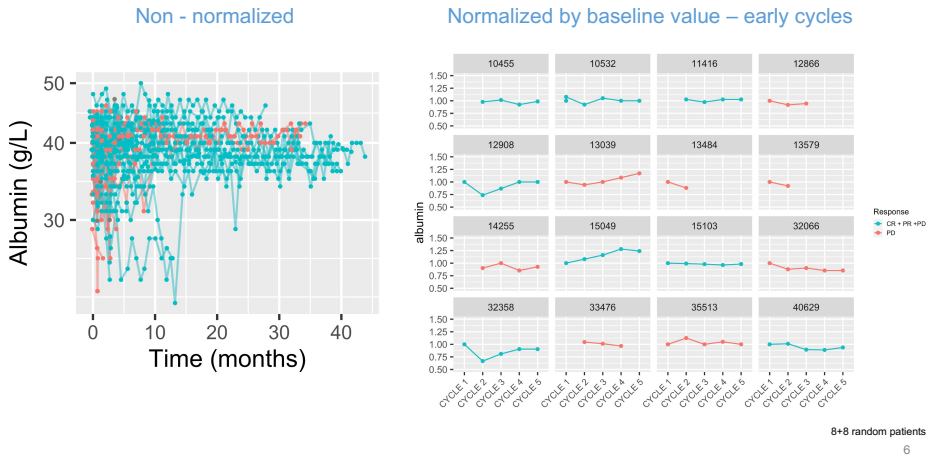
Characteristic	Total, N = 1415[1]	FIR, N = 133[1]	POPLAR, N = 134[1]	BIRCH, N = 595[1]	OAK, N = 553[1]	p-value[2]
PD-L1 tumor cells						<0.001
0	606 (43%)	4 (3.0%)	40 (30%)	212 (36%)	350 (64%)	
1	209 (15%)	21 (16%)	56 (42%)	69 (12%)	63 (11%)	
2	381 (27%)	105 (79%)	37 (28%)	160 (27%)	79 (14%)	
3	217 (15%)	3 (2.3%)	1 (0.7%)	154 (26%)	59 (11%)	
Unknown	2	0	0	0	2	
PD-L1 immune cells						<0.001
0	252 (18%)	3 (2.3%)	40 (30%)	16 (2.7%)	193 (35%)	
1	398 (28%)	5 (3.8%)	56 (42%)	122 (21%)	215 (39%)	
2	439 (31%)	25 (19%)	19 (14%)	297 (50%)	98 (18%)	
3	322 (23%)	99 (75%)	19 (14%)	159 (27%)	45 (8.2%)	
Unknown	4	1	0	1	2	
ECOG						0.6
Status 0	505 (36%)	42 (32%)	44 (33%)	215 (36%)	204 (37%)	
Status 1 or 2	909 (64%)	90 (68%)	90 (67%)	380 (64%)	349 (63%)	
Unknown	1	1	0	0	0	

[1]n (%)
[2]Pearson's Chi-squared test

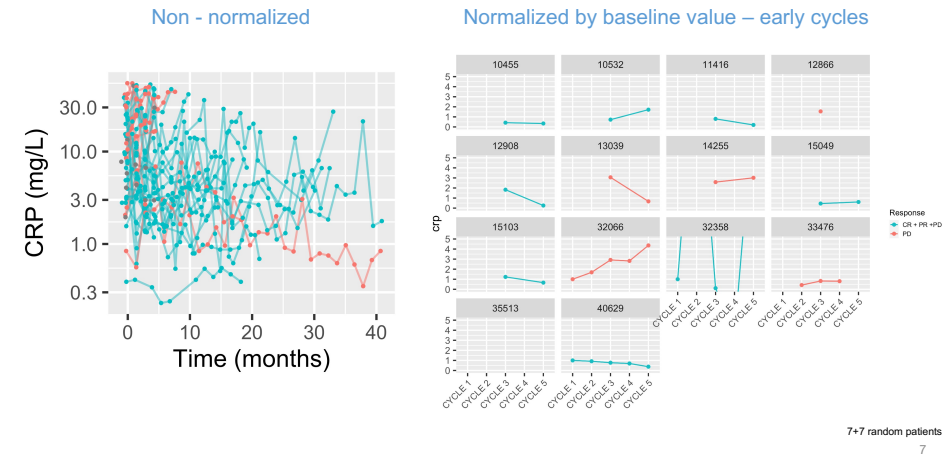
Appendix 1—figure 4. Patient characteristics: PD-L1 and ECOG



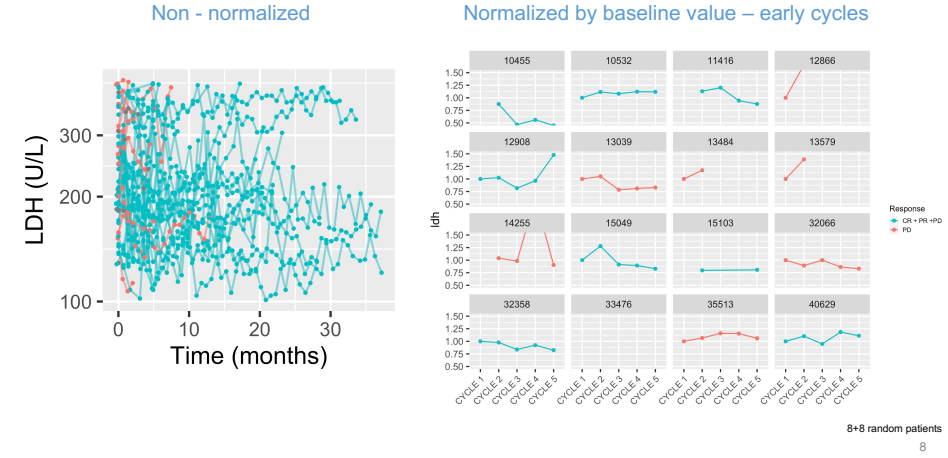
Appendix 1—figure 5. TK modeling goodness-of-fit



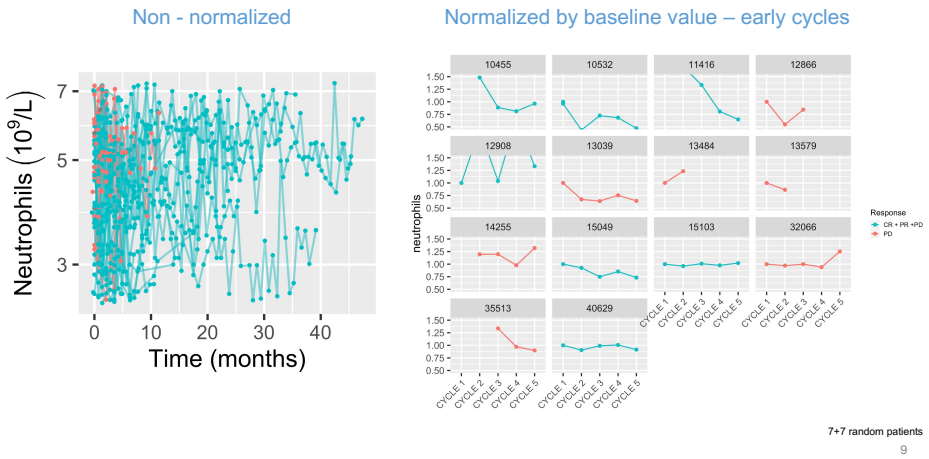
Appendix 1—figure 6. Examples of longitudinal kinetics: Albumin



Appendix 1—figure 7. Examples of longitudinal kinetics: CRP



Appendix 1—figure 8. Examples of longitudinal kinetics: LDH



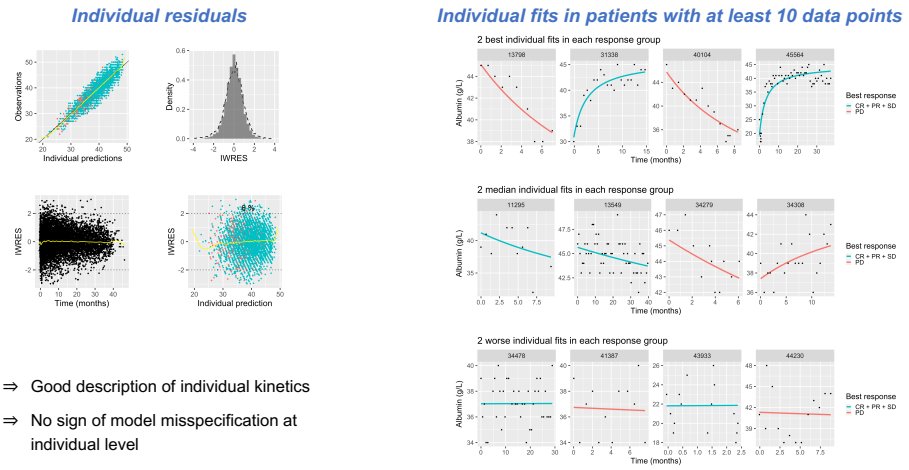
Appendix 1—figure 9. Examples of longitudinal kinetics: Neutrophils

	Albumin		CRP		LDH		Neutrophils	
Model	BICc	b	BICc	b	BICc	b	BICc	b
Double-exponential	48,395	0.058	**28,764**	**0.21**	**39,886**	**0.56**	**102,449**	**0.14**
Hyperbolic	**48,007**	**0.056**	29,712	0.22	40,915	0.62	102,943	0.14
Linear	49,436	0.063	30,020	0.23	42,462	0.70	105,193	0.17
Constant	49,724	0.065	31,332	0.25	42,982	0.74	106,249	0.18

Corrected Bayesian Information Criterion (BICc) for four empirical kinetic models of BK. b : standard deviation of the proportional error model

10

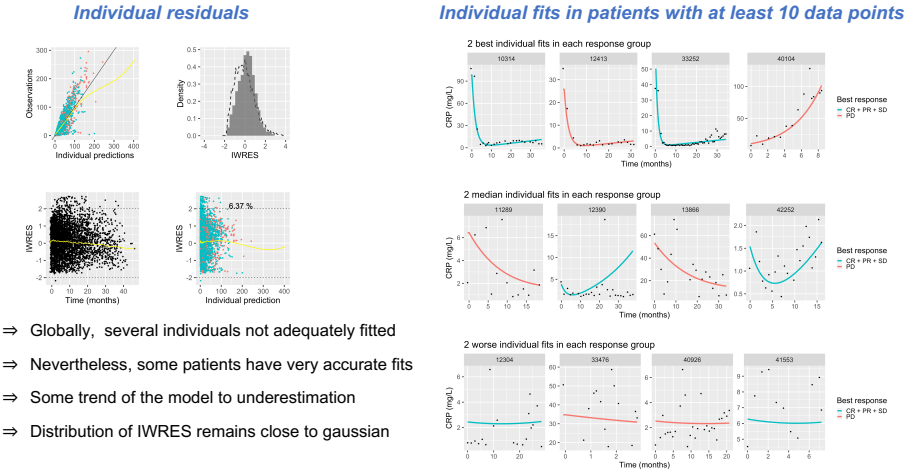
Appendix 1—figure 10. Goodness-of-fit metrics of dynamic BK models



⇒ Good description of individual kinetics

⇒ No sign of model misspecification at individual level

Appendix 1—figure 11. Albumin: hyperbolic individual fits



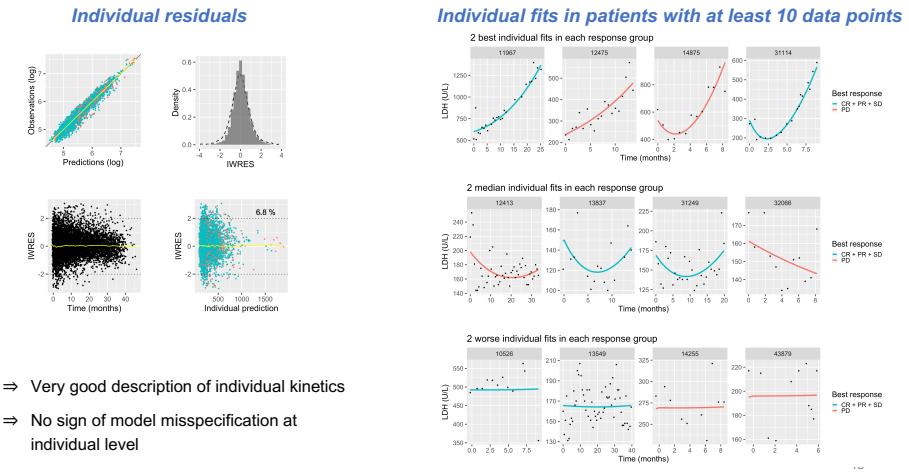
⇒ Globally, several individuals not adequately fitted

⇒ Nevertheless, some patients have very accurate fits

⇒ Some trend of the model to underestimation

⇒ Distribution of IWRES remains close to gaussian

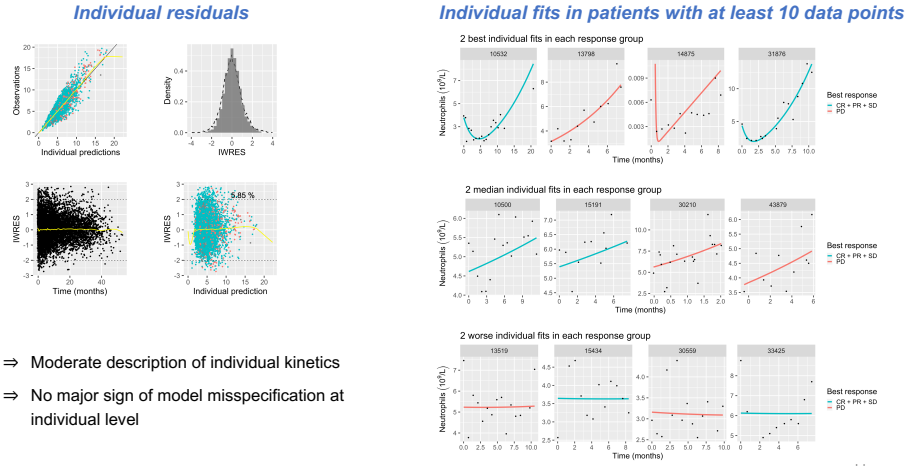
Appendix 1—figure 12. CRP: dexp individual fits



⇒ Very good description of individual kinetics

⇒ No sign of model misspecification at individual level

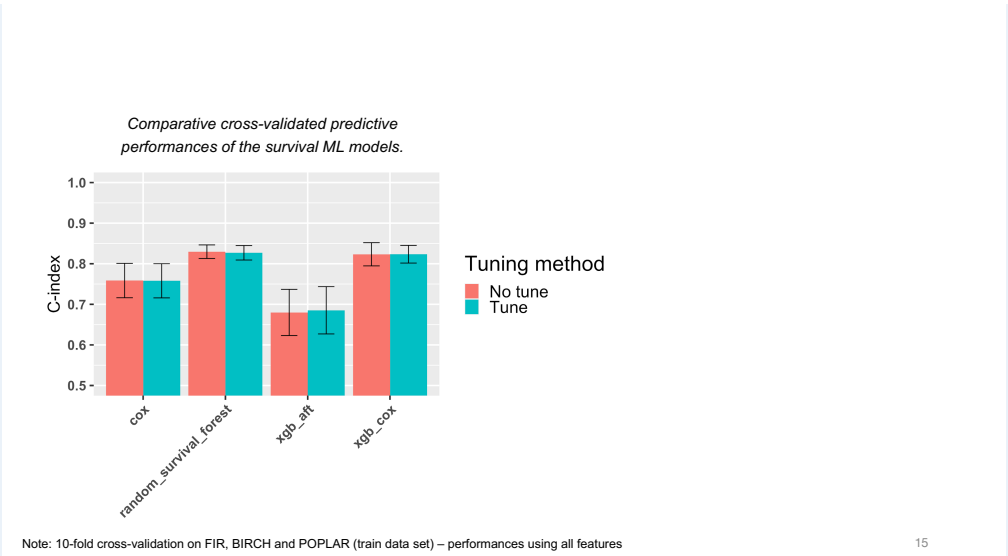
Appendix 1—figure 13. LDH: dexp individual fits



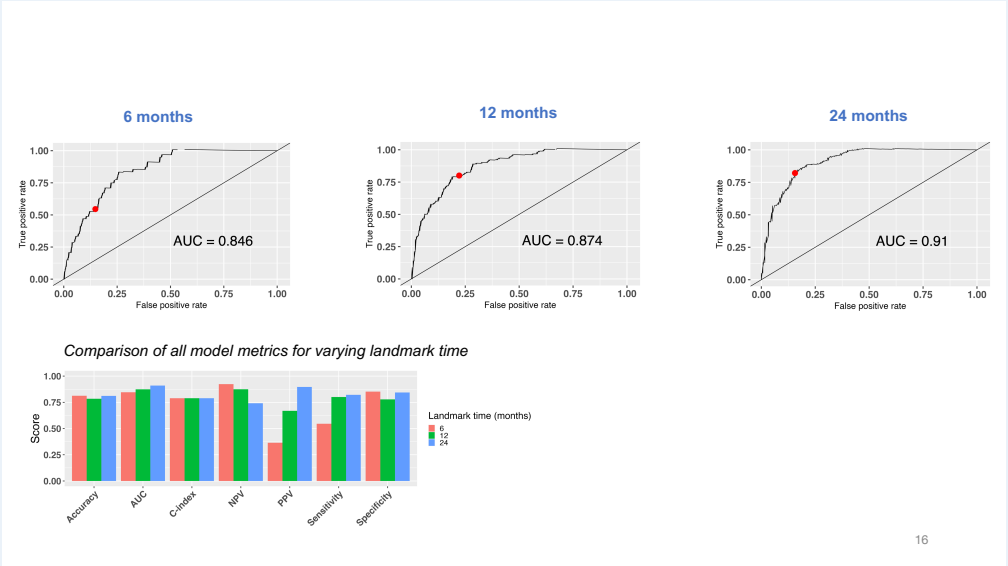
⇒ Moderate description of individual kinetics

⇒ No major sign of model misspecification at individual level

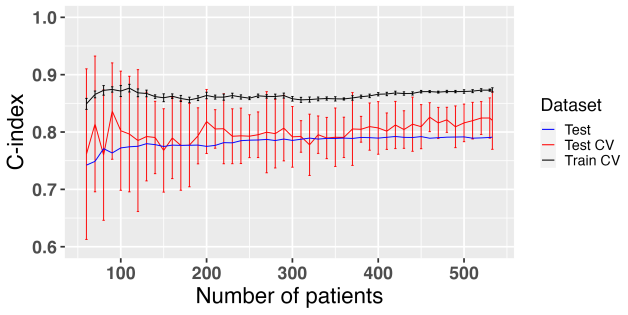
Appendix 1—figure 14. Neutrophils: dexp individual fits



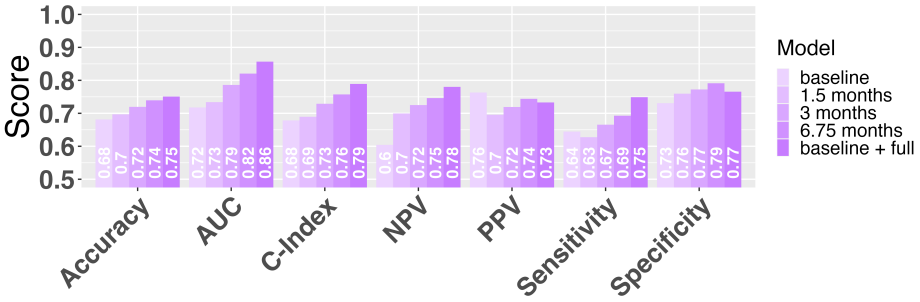
Appendix 1—figure 15. Comparison of ML algorithms and tuning methods



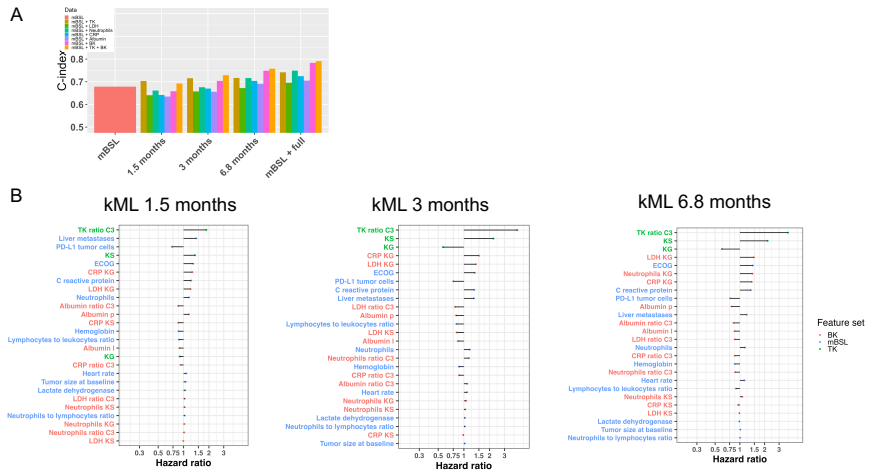
Appendix 1—figure 16. ROC curves for variables landmark times (test set - OAK)



Appendix 1—figure 17. Learning curve



Appendix 1—figure 18. Additional value of NLME to baseline for multiple metrics



Appendix 1—figure 19. kML models using only single kinetic markers