

Predicting survival and trial outcome in non-small cell lung cancer integrating tumor and blood markers kinetics with machine learning

Sébastien Benzekry¹✉, Mélanie Karlsen¹, Célestin Bigarré¹, Abdessamad El Kaoutari¹, Bruno Gomes², Martin Stern³, Ales Neubert⁴, Rene Bruno⁵, François Mercier⁶, Suresh Vatakuti⁷, Peter Curle⁸, Candice Jamois⁹

✉ 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

¹COMPUTational pharmacology and clinical Oncology Department, Inria Sophia Antipolis-Méditerranée, Cancer Research Center of Marseille, Inserm UMR1068, CNRS UMR7258, Aix Marseille University UM105, Marseille, France; ²Pharma Research and Early Development, Early Development Oncology, Roche Innovation Center Basel, Switzerland; ³Pharma Research and Early Development, Early Development Oncology, Roche Innovation Center Zurich, Switzerland; ⁴Pharma Research and Early Development, Data & Analytics, Roche Innovation Center Basel, Switzerland; ⁵Modeling and Simulation, Clinical Pharmacology, Genentech Research and Early Development, Marseille France; ⁶Modeling and Simulation, Clinical Pharmacology, Genentech Research and Early Development, Roche Innovation Center Basel; ⁷Pharma Research and Early Development, Predictive Modeling and Data Analytics, Roche Innovation Center Basel, Switzerland; ⁸Inovigate, Basel, Switzerland; ⁹Pharma Research and Early Development, Translational PKPD and Clinical Pharmacology, Roche Innovation Center Basel, Switzerland

Abstract

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, CRP, lactate dehydrogenase and neutrophils). Developed for immune-checkpoint inhibition (ICI) in non-small cell lung cancer on three phase 2 trials (533 patients), kML was validated on the two arms of a phase 3 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, a 12-months survival accuracy of 78.7% and a 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 3 trial using only 25 weeks of on-study data (predicted HR = 0.814 (0.64 – 0.994) versus final study HR = 0.778 (0.65 – 0.931)). Our model constitutes a valuable approach to support personalized medicine and drug development.

37 Introduction

38 Lung cancer is the leading cause of cancer death worldwide¹, with non-small cell lung cancer
39 (NSCLC) being the most prevalent type, representing 80%–85% of case². Immune-checkpoint in-
40 hibitors (ICI) (e.g., atezolizumab (ATZ)) have led to significant improvements in survival rates for pa-
41 tients with advanced cancers such as NSCLC^{3,4}. However, there is still a large variability in clinical
42 response and progression eventually occurs in a majority of patients⁵. Additionally, drug devel-
43 opment in immuno-oncology is highly challenging, with a 95% attrition rate⁶. Current approaches
44 for go/no-go decisions are based on interim endpoints (e.g., progression-free survival, overall re-
45 sponse rate) that have often been found to be poor predictors of the primary endpoint of most
46 clinical trials in oncology, overall survival (OS)⁷. This calls for better surrogate markers at interim
47 analyses. Altogether, there is a need for better and validated predictive models of OS for both
48 personalized health care (individual predictions) and drug development (trial predictions).

49 Currently, PDL1 expression is the only routine biomarker used for NSCLC patients^{5,8} despite
50 being controversial^{9,10}. Tumor mutational burden^{8,11,12} and transcriptomic data^{5,13,14} have also
51 been investigated but did not reach clinical practice. Here we posit that such static and single
52 marker approach is intrinsically limited and that substantial additional predictive performances
53 could be gained by: 1) using multi-modal integrative analyses relying on a combination of markers
54 and machine learning algorithms^{5,12,14,15} and 2) including dynamic markers obtained from early
55 on-treatment data^{15,16}. The nonlinear mixed-effects (NLME) modeling approach is well suited for
56 the latter¹⁷, and tumor kinetics (TK) model-based metrics have been shown to carry significant
57 predictive value for OS in oncology, including ATZ monotherapy in advanced NSCLC^{18–20}. The first
58 main novelty of the current study is to establish the predictive value of model-based parameters
59 of simple blood markers kinetics (BK), in addition to TK.

60 The second main novelty is to apply machine learning (ML) algorithms, increasingly used in
61 biology and medicine²¹ but only rarely for TK-OS modeling²², instead of classical survival mod-
62 els. Extensions of classical ML models to survival data have been proposed (e.g., random survival
63 forests²³), but their actual superiority over standard approaches remains controversial²⁴. In addi-
64 tion, most ML studies to date are underpowered due to low sample sizes in both training and test
65 sets.

66 Here, we coupled the strengths of NLME modeling with ML to derive a predictive model of OS
67 from baseline and on-treatment data, called kinetics-machine learning (kML, Figure 1A). We lever-
68 aged large training and test datasets to achieve robust results (Figure 1B). Subsequently, we tested
69 the operational predictive capabilities of kML in two relevant scenarios: 1) individual prediction of
70 OS and 2) prediction of the outcome of a phase 3 trial from early on-study data.

71 Methods

72 Data

73 For both training and external validation (testing) sets, patients from French centers were excluded
74 for legal reasons ($N = 118$, not included in the numbers above). The training set comprised the
75 FIR (NCT01846416)²⁵, POPLAR (NCT01903993)³ and BIRCH (NCT02031458)²⁶ phase 2 clinical trials.
76 The test set was the atezolizumab arm of the OAK phase 3 trial (NCT02008227)²⁷ for individual pre-
77 dictions and additionally the docetaxel arm for trial predictions (Supplementary Figure 1). These
78 studies were conducted in accordance with the Declaration of Helsinki after approval by institu-
79 tional review boards or independent ethics committees. All patients provided written informed
80 consent.

81 The outcome considered was overall survival (OS), defined as the time between treatment start
82 and death or last follow-up, in which case the data was right-censored. The median follow-up was
83 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.

84 Preprocessing

85 Baseline data

86 The baseline data consisted of 63 variables spanning demographic and biological data, clinical infor-
87 mation and disease status (see Supplementary Figure 2–4 for a description of the main variables).
88 PD-L1 expression on tumor cells was measured by immunohistochemistry or quantitative poly-
89 merase chain reaction, with four possible levels (0: < 1%; 1: $\geq 1\%$; 2: $\geq 5\%$ and 3: $\geq 50\%$)³. We refer
90 to the above-mentioned identifiers and references for further details on the other variables. Data
91 were measured in accordance to the studies principles.

92 Tumor and blood markers kinetics (TK and BK)

93 Patients with only one baseline SLD measurement and no SLD measurement during the treatment
94 period were excluded ($N = 110$). For BK, first time points prior to treatment start were discarded.
95 Then, four exclusion rules were established to identify anomalous data points: 1) values outside
96 physiologically possible bounds, 2) duplicates, 3) values that abruptly went to an extreme out-of-
97 range value between two measurements, 4) only the BK value at the closest time point to treatment
98 initiation was kept. Eventually, in order to have sufficient data for Bayesian estimation with early
99 data, patients with less than three observations before cycle 5 were removed. We refer to the
100 supplementary methods for details.

101 Nonlinear mixed-effects modeling

102 Population approach

103 Statistical hierarchical nonlinear mixed-effects modeling (NLME) was used to implement a popula-
104 tion approach²⁸ for the kinetic data and parameter estimation was conducted using the Monolix
105 software²⁹. Mathematical details are given in the supplementary methods.

106 Structural models

107 Following previous work, the TK structural model was assumed to be the sum of two exponen-
108 tials^{19,30}:

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

109 where $t = 0$ corresponds to treatment initiation and y_0 , KG and KS are three parameters, represent-
110 ing respectively the baseline value, growth and shrinkage rates. This model was also considered
111 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 hy-
112 perbolic ($y_j^i = p^i + \frac{e^{q^i (t_j^i - p^i)}}{t_j^i + e^{q^i}}$)³¹. Quantitative comparison of goodness-of-fit between models was
113 assessed using the corrected Bayesian information criterion³².

114 Identification of individual model-based parameters

115 The population parameters identified on the training set were used to define prior distributions of
116 the TK and BK model parameters. These “training” priors were used for Bayesian estimation (maxi-
117 mum a posteriori estimate) of the individual TK and BK model parameters, not only for the training
118 set but also for the test sets, in order to avoid leakage. To focus on the pure kinetic parameters,
119 the model-estimated baseline parameters were not kept. We additionally considered the ratio of
120 the model-predicted value at cycle 3 day 1 to the model-estimated baseline parameter. Altogether,
121 there were three individual parameters for each marker: X_{KG} , X_{KS} and X_{ratio} for $X = \text{TK, CRP, LDH}$
122 and neutrophils; and albumin_p , albumin_l and albumin_{ratio} for albumin.

123 Truncated data: individual-level

124 Individual-level truncated datasets were derived from the longitudinal TK and BK data by keeping
125 data only up to: cycle 3 day 1 (C3D1, 1.5 months), C5D1 (3 months) and C10D1 (6.75 months). New
126 training priors were estimated from each CXD1 training set. The resulting TK and BK truncated
127 model parameter Y for marker X at cycle i were denoted by $X_{Y,i}$ (e.g., $ldh_{KG,5}$).

128 Truncated data: study-level for trial prediction

129 Study-level truncated datasets were defined at the following on-study landmark times t after study
130 initiation (first patient recruited): $t = 10, 25$ and 60 weeks.

131 Only the patients enrolled before this time and their data collected up to t was used. Note
132 that here $t = 0$ corresponds to study initiation and thus patients in these datasets have varying
133 follow-up duration (from 0 to t), in contrast to individual-level truncated datasets.

134 Machine learning

135 Data preparation

136 Missing values (1.6% total, maximum 12% in one variable) were imputed with the median for nu-
137 meric variables and mode for categorical variables, learned on the training set, even when applied
138 to the test set. All numeric variables were centered and scaled. Means and standard deviations
139 were learned on the train and carried to the test set.

140 Models

141 Model elaboration and development was performed exclusively on the training set, using 10 folds
142 cross-validation for predictive performances evaluation. Due to censoring in the data, survival
143 models were used: proportional hazards Cox regression³³, extreme gradient boosting (XGB) with
144 either Cox or accelerated failure time (AFT) models³⁴ and random survival forests (RSF)²³. Nested
145 cross-validation with inner bagging in each 10-fold cross-validation outer loop was used to evaluate
146 the benefit of tuning the hyperparameters³⁵. Improvement of the performances was negligible
147 with hyperparameter tuning (Supplementary Figure 5). Therefore, we used the default values of
148 the hyperparameters. For the final RSF model: number of trees $n_{tree} = 500$, number of variables
149 to possibly split at each node $m_{try} = 5$, minimum size of terminal node $n_{odesize} = 15$, number of
150 random splits for splitting a variable $n_{split} = 10$.

151 Evaluation

152 Predictive performances were assessed for either discrimination (c-index and classification met-
153 rics at horizon times τ), calibration (calibration curves) or stratification (dichotomized KM survival
154 curves). For each individual, the RSF model gives two prediction outputs: a scalar value termed
155 “mortality” that we will refer to as “ML score”, and time-dependent predicted survival curves²³. The
156 former was used to compute the c-index using the `rcorr.cens` function of the `hmisc` R package^{36,37}.
157 For prediction of survival at a horizon time τ , we used the latter to compute model-predicted prob-
158 abilities of death at τ . Unless otherwise specified, $\tau = 12$ months. Survival-adapted metrics of pre-
159 dictive performance were used for sensitivity, specificity, area under the receiver-operator curve
160 (ROC AUC) and negative and positive predictive value (NPV and PPV) to account for censoring^{38,39}.
161 For computation of accuracy, censored patients before τ were discarded ($N = 17/396$ in the test
162 set at 12 months). The optimal cut-points used for individual OS predictions on the test set were
163 defined as the Kaplan-Meier estimated survival probability in the training set at τ (0.257 at 6 months,
164 0.437 at 12 months, 0.634 at 24 months).

165 For patient stratification (dichotomized KM curves), the ML score was used, with models trained
166 on the training set and predicted on the test set. In order to assess stratification abilities to capture
167 the 20% of long-term survivors, cut-points were set at the 20th percentiles for each variable/score
168 evaluated. This cut-point arbitrary definition was also motivated by the aim to ensure fair compar-
169 ison between multiple parameters on the same data. Significance of differences in KM curves was
170 established using the logrank test, and hazard ratios were computed using proportional hazards
171 Cox regression.

172 Variable selection and minimal signature

173 Variable selection was performed only for the BSL data. The method was based on two steps: 1)
174 sorting the variables using least absolute shrinkage and selection operator (LASSO)⁴⁰ and 2) build-
175 ing RSF incremental models including increasing numbers of variables. LASSO sorting was defined

176 as taking the coefficients gradually becoming non-zero during likelihood maximization when the
177 regularization parameter decreases. The minimal signature was defined as the minimal set of vari-
178 ables able to achieve a c-index larger than 0.75 and an AUC larger than 0.8, with the addition of 4
179 well-established prognosis.

180 Survival simulations and computation of predicted HRs

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

189 Data Availability

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

195 Code availability

196 Algorithms used for data analysis are all publicly available from the indicated libraries and refer-
197 ences in the Methods section.

198 Results

199 Data

200 The data consisted of advanced NSCLC patients enrolled in ATZ trials ($N = 1936$, Figure 1B and
201 Supplementary Figure 1). Three ATZ phase 2 trials were pooled into a training dataset^{3,25,26} ($N =$
202 862). The external validation (test) set comprised data from the ATZ arm ($N = 553$) of the OAK phase
203 3 trial²⁷. For trial outcome prediction the docetaxel arm ($N = 521$) was added as an additional test
204 set.

205 Variables comprised baseline (pre-treatment) and longitudinal (on-treatment) data (Figure 1A).
206 The former included: patients and disease characteristics ($p = 63$ variables, 43 numeric and 20
207 categorical, denoted BSL) and transcriptomic ("RNAseq", $p = 58,311$ transcripts) data. The latter
208 included: longitudinal investigator-assessed sum of largest diameters (SLD) of lesions as per the
209 RECIST criteria⁴¹, denoted by tumor kinetics (TK, $k = 5, 473/3, 015$ time points in the train/test sets, re-
210 spectively, median 5/4 data points per patient, range 2/2 —24/20); and longitudinal measurements
211 of four blood markers (albumin, C-reactive protein (CRP), lactate dehydrogenase (LDH) and neu-
212 trophils), denoted together as blood markers kinetics (BK, $k = 60, 779/38, 460$ data points, median
213 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
214 albumin-CRP-LDH-neutrophils in the train/test sets, respectively).

215 Nonlinear mixed-effects modeling (NLME) of longitudinal markers

216 We first developed NLME models for the longitudinal data (Figure 1B). The TK structural model was
217 the sum of an increasing and a decreasing exponential function (double exponential model)³⁰. It
218 was able to accurately describe the training data with no goodness-of-fit misspecification (Figure
219 2A and Supplementary Figure 6). Population parameters were estimated with good accuracy (all
220 relative standard errors smaller than 9%, Table 1).

Figure 1: Schematic of the kML framework

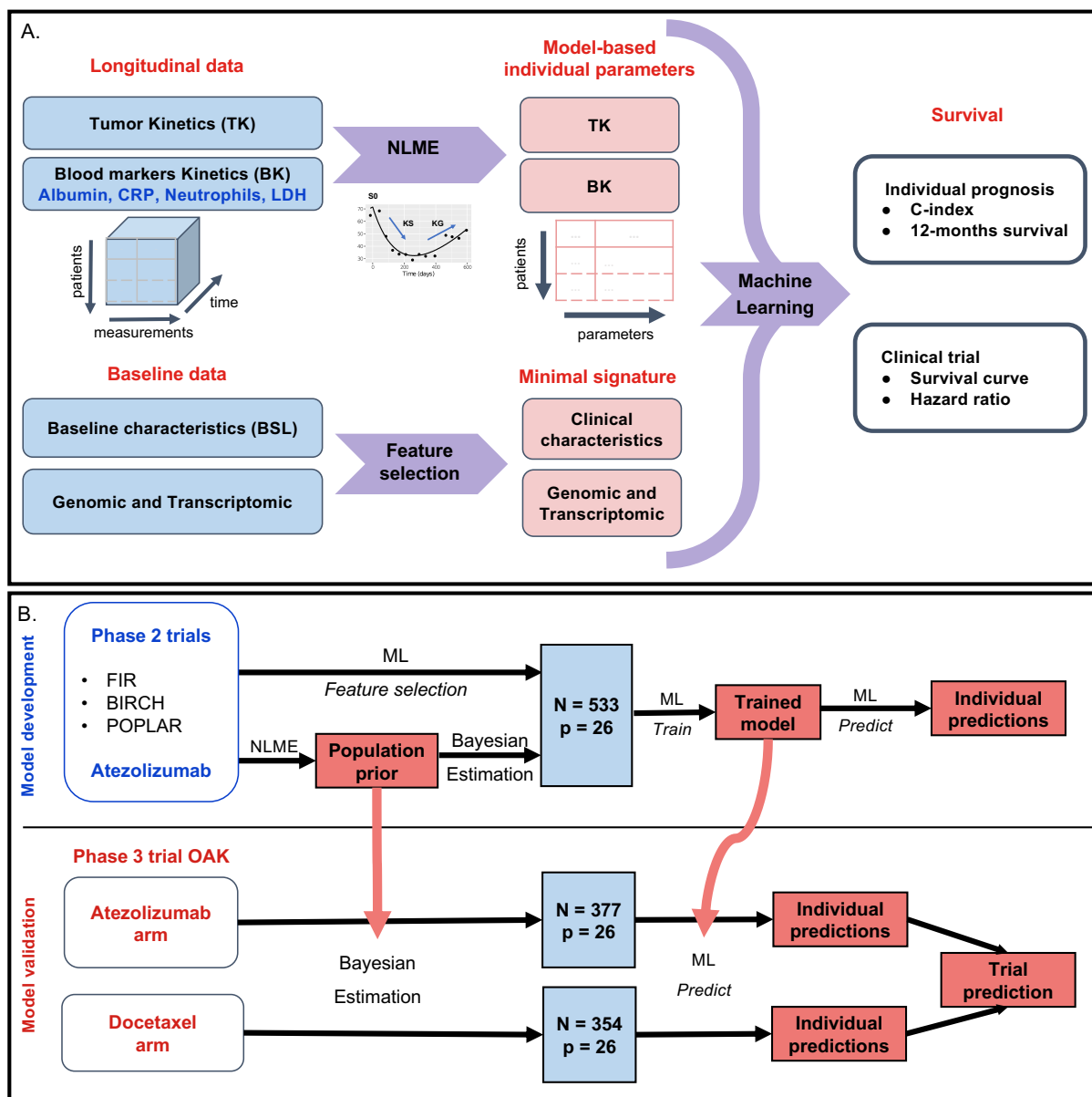


Figure 1. Study schematic **A.** Baseline and longitudinal data were combined into a machine learning algorithm in order to predict individual survival prognosis. Longitudinal data were modelled using nonlinear mixed-effects modelling, whereas machine learning-based feature selection was applied to the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

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

	K		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)
KG_{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 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 dehydrogenase.

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 Supplementary Figures 7–10). 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 non-linear 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 Bayesian information criterion and relative error between model fits and data (Supplementary Figure 11). The 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 (Supplementary Figures 12–15). 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; (controlled by the 20th percentile definition of the cut-point, see methods), 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.

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

Four feature sets resulted from the analysis above: BSL, RNAseq, TK and BK (Figure 1A). The development of a kinetics-machine learning (kML) comprised two main steps: choice of the algorithm and derivation of a minimal signature (Figure 1B). The first was achieved by benchmarking four models that used all variables ($p = 119$, $N = 553$). The random survival forest (RSF) model found to exhibit the best performances (Supplementary Figure 5) and was thus selected. Notably, we found significantly better predictive performances of RSF over a classical 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, hemoglobin⁴², SLD²² and LDH^{43,44}.

Applying stringent criteria to the RNAseq data (see supplementary methods), we selected 167 transcripts as candidates for final variable selection using Bolasso regression model to identify the

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

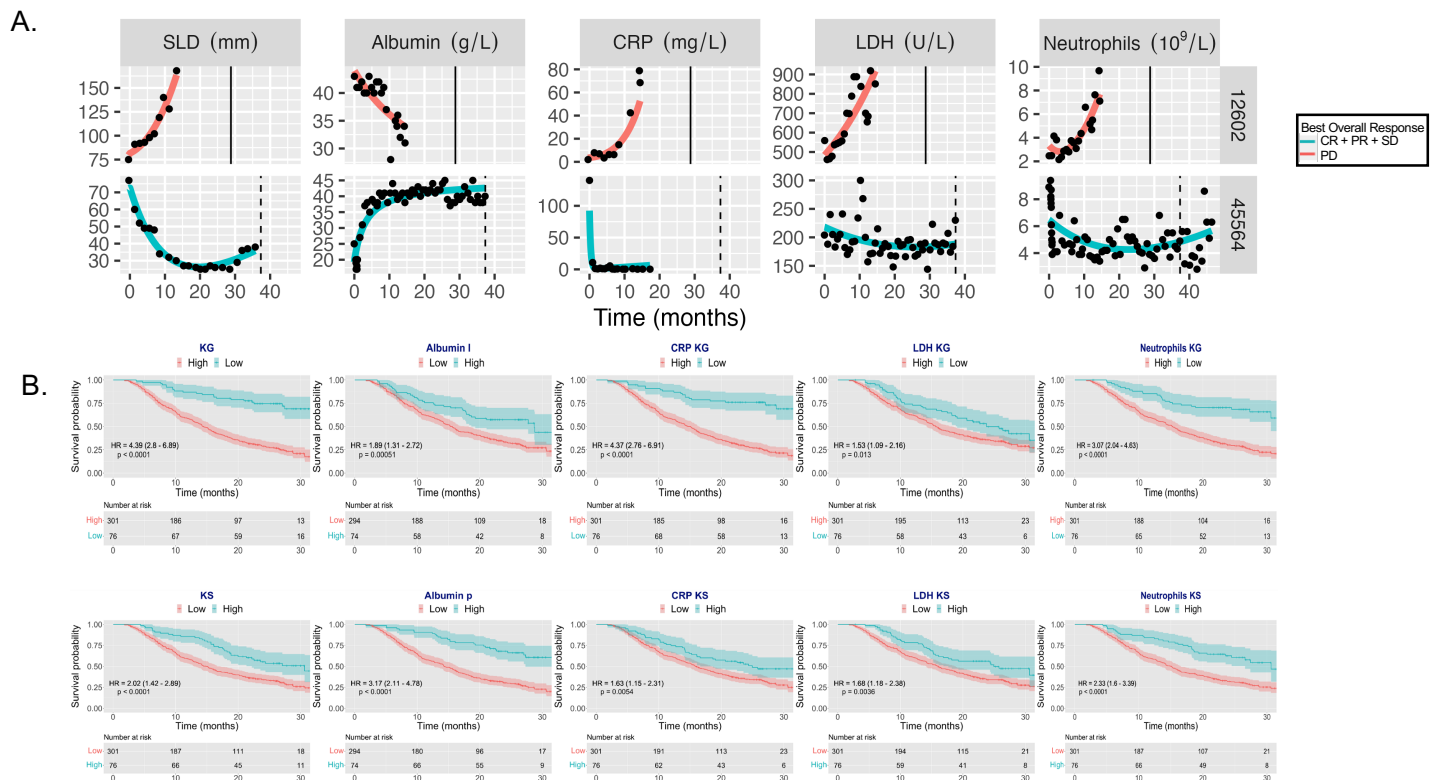


Figure 2. Goodness-of-fit metrics and plots of dynamic BK models **A.** Representative individual fits for the TK and BK best empirical models showing non-trivial kinetic parameters well captured by the dynamic models. Survival is indicated by a vertical line (solid = death, dashed = censored). **B.**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

Table 2. Contingency table for OS prediction at 12 months

MODEL		TRUTH		
		Alive (0)	dead (1)	Total
	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 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.

260 optimal set of predictors⁴⁵. Finally, we ended up with 52 RNAseq variables that corresponded to
261 the highest average c-index of 0.64.

262 We then compared the cross-validated c-index of each feature set on the train data (Figure
263 3B). Because of negligible discrimination performances ($c\text{-index} = 0.62 \pm 0.050$) and non-systematic
264 availability of those data, the RNAseq set was removed from the model. The selected set of clinical
265 data at baseline (mBSL) exhibited moderate discrimination performances ($c\text{-index} = 0.710 \pm 0.038$),
266 which was slightly outperformed by the TK set ($c\text{-index} = 0.723 \pm 0.025$). Interestingly, the BK set
267 significantly outperformed both baseline clinical and TK ($c\text{-index} = 0.793 \pm 0.038$, $p = 0.0004$ and 0.0005
268 respectively, Student's t-test). Jointly, mBSL, TK and BK performed significantly better than any
269 feature set alone ($c\text{-index} = 0.824 \pm 0.050$, $p = 0.00007$, 0.0002 and 0.055), as well as any combination
270 of two sets among the three (mBSL + TK: $c\text{-index} = 0.77 \pm 0.026$, mBSL + BK: $c\text{-index} = 0.81 \pm 0.027$,
271 TK + BK: $c\text{-index} = 0.80 \pm 0.049$). The resulting model combining mBSL, TK and BK was denoted kML
272 (kinetics-machine learning).

273 During cross-validation on the training set, kML exhibited excellent predictive performances
274 across multiple metrics, with minimal between-folds variability ($AUC = 0.919 \pm 0.056$, $accuracy =$
275 0.873 ± 0.052 , Figure 3C).

276 External validation

277 The predictive performance of the final kML model (mBSL, TK and BK) was assessed on the ATZ
278 test set (377 patients). At the population level, the model-predicted survival curve was in excellent
279 agreement with the observed data (Figure 4A). Notably, the prediction interval from the model
280 was narrow, indicating high precision. At the individual level, consistent with the cross-validation
281 results, substantial discrimination performances were observed ($c\text{-index} = 0.790$, accuracy and AUC
282 for 12-months survival probability 0.787 and 0.874, respectively, Figure 4B). All classification metrics
283 for prediction of survival at 12 months were high (≥ 0.78), except PPV, indicating worse ability to
284 predict death than survival. Although smaller, they were similar to the cross-validation results.

285 In addition, calibration curves revealed good performance, at multiple horizon times (Figure
286 4C). Model-predicted probabilities were concordant with the observed KM estimates of the survival
287 probabilities, over the entire range of the binned predicted probabilities. This is further illustrated
288 by the contingency Table 2. For instance, among 212 patients predicted to be alive at 12 months,
289 182 (85.8%) were actually alive. Predictive AUC was good at other horizon times (0.846 and 0.910 at
290 6 and 24 months, respectively, Supplementary Figure 16). However, PPV and sensitivity were very
291 low at 6 months.

292 Notably, the kML mortality score derived from the model and learned on the training set was
293 able to accurately stratify OS in the test set ($HR = 25.2$ (10.4–61.3), $p < 0.0001$, Figure 4D), indicat-
294 ing excellent ability to identify the 20% of long-term survivors. It outperformed all single kinetic
295 markers (Figure 2C).

296 Variables importance was assessed by running a post-hoc multivariable Cox regression (Figure

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

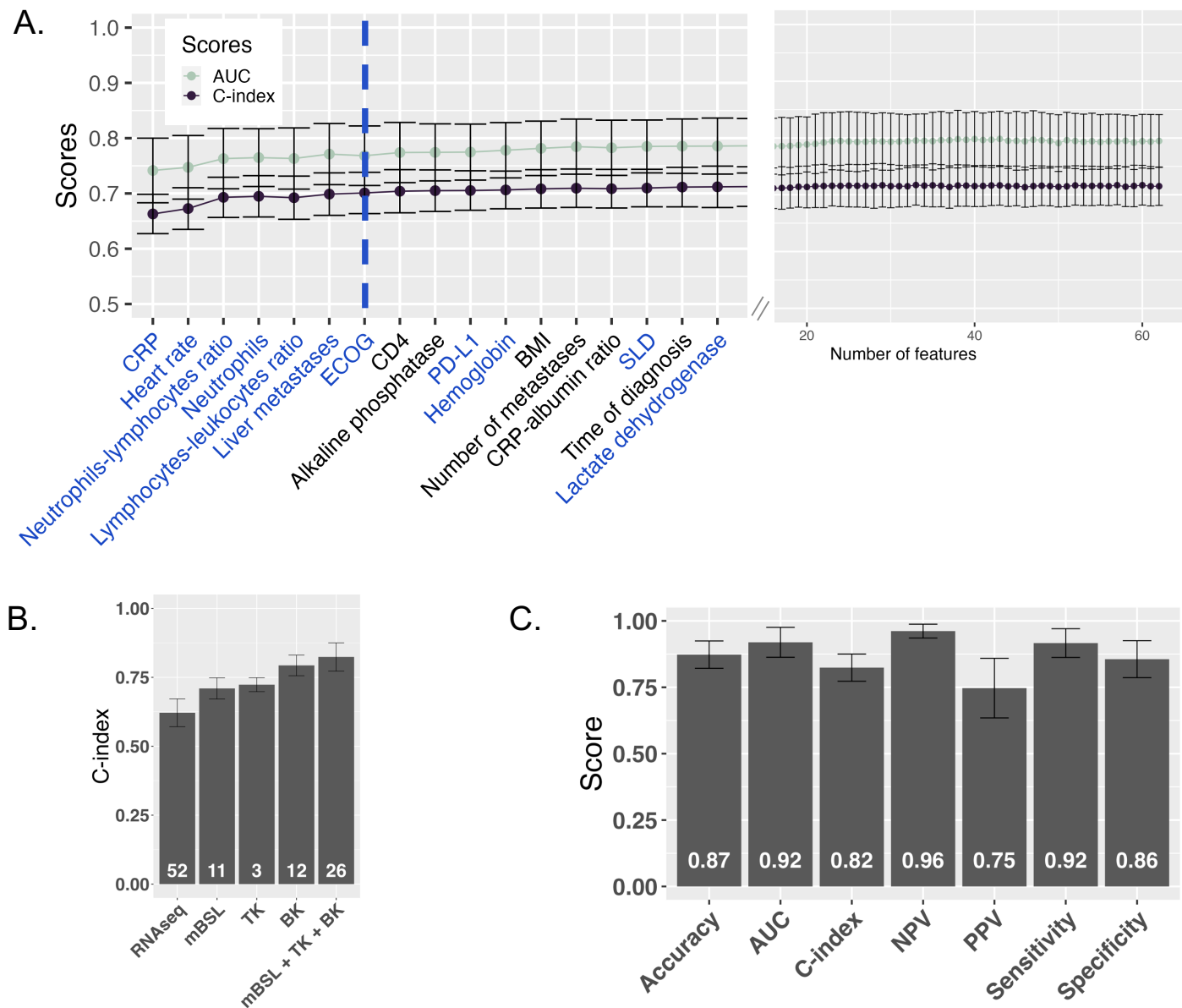


Figure 3. Minimal baseline (mBSL) signature and kinetics-ML (kML) model **A.** Cross-validated (CV) performance scores on the training set (c-index and AUC, mean \pm standard deviation) for incremental random survival forest (RSF) models using an increasing number of baseline clinical and biological 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).

Figure 4: Predictive performances of kML on the ATZ test set

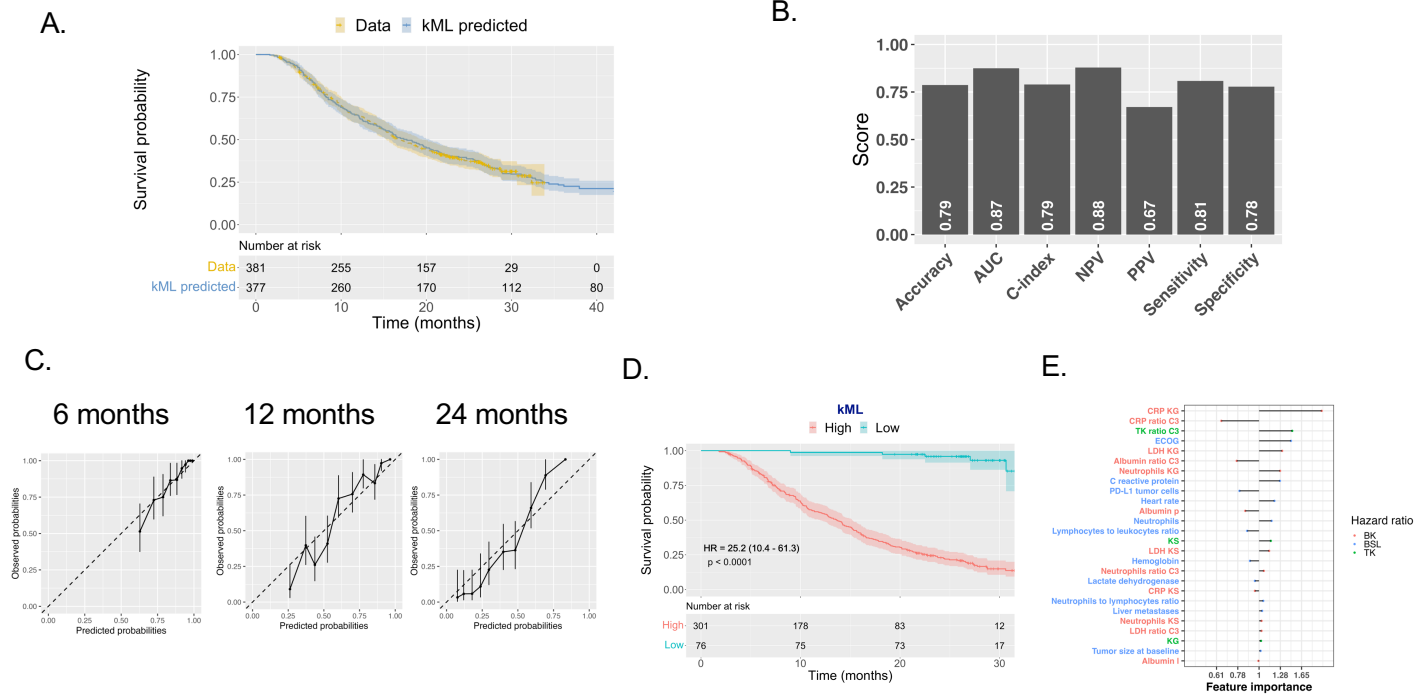


Figure 4. Predictive performances of kML on the ATZ test set **A.** Comparison of the population-level survival curves between the data (KM estimator) and the model prediction. **B.** Scores of discrimination metrics. Classification metrics were computed for prediction of OS at 12 months. **C.** Calibration 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.

297 4F). Interestingly, the top two variables were BKs (CRP_{KG} and CRP ratio C3). In addition, TK and
 298 BK made up for six out of the seven top important features and were found more important than
 299 PD-L1.

300 Given the large sample size of our data, we further assessed the model performances when
 301 trained on smaller data sets (Supplementary Figure 17). The learning curve revealed that ap-
 302 proximately 200 patients were necessary to reach similar performance to the ones obtained with
 303 the full training set ($N = 533$), for both cross-validation and external validation on the test set
 304 ($c\text{-index} = 0.82 \pm 0.056$ vs $c\text{-index} = 0.82 \pm 0.050$ in cross-validation, 0.78 vs 0.79 on the test set, models
 305 trained with 200 vs 533 patients, respectively). Trained with only 60 patients, kML reached already
 306 good performances ($c\text{-index} = 0.76 \pm 0.15$ and 0.74 in cross-validation and test, respectively).

307 Together, these results demonstrate important predictive performances of overall survival fol-
 308 lowing ATZ treatment using the kML model.

309 Application to individual survival prognosis from early on-treatment data

310 Results above required full on-treatment time-course data to compute TK and BK markers, thus
 311 cannot be used to make early predictions. To investigate the operational applicability of our method-
 312 ology, data from the test set were truncated at the beginning of treatment cycles number 3, 5 and
 313 10, respectively corresponding to 1.5, 3 and 6.75 months. We found that integrating longer on-
 314 treatment data in kML, the predictive performances steadily increased (Figure 5A and Supplemen-
 315 tary Figure 18). Using the baseline variables only (mBSL), the stratification ability was significant
 316 but moderate (HR = 1.74 (1.24 - 2.46), $p = 0.0014$, Figure 5B). In contrast, kML exhibited increasing
 317 stratification ability from data at 1.5 months (HR = 2.19 (1.53 - 3.12), $p < 0.0001$), 3 months (HR = 3.51

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

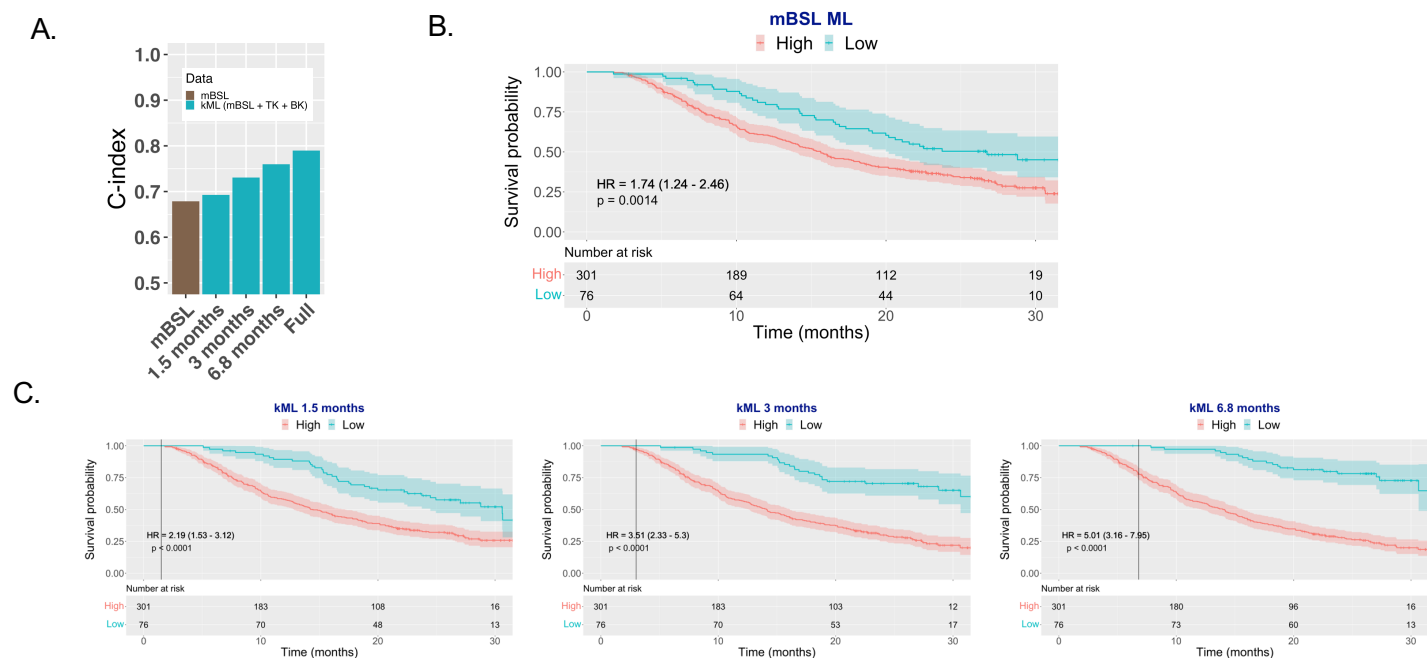


Figure 5. Predictive value of kML from cycle-truncated data **A.** Predictive power (c-index) of ML models using baseline (BSL) or truncated data at 1.5, 3 and 6.8 months as well as the full time-course. **B.** Stratified KM survival curves using a RSF model trained on the minimal baseline (mBSL) variables. **C.** Stratified KM survival curves using kML from 1.5 months (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.

318 (2.33 – 5.3), $p < 0.0001$) and 6.8 months (HR = 5.01 (3.16 – 7.95), $p < 0.0001$), see Figure 5C.

319 Further investigation of the predictive performances of individual kinetic markers revealed
 320 that TK parameters were the most informative at 6 weeks (1.5 months, first imaging assessment).
 321 Adding BKs to TKs brought additional predictive value starting at 3 months, and BKs outperformed
 322 TK from 6.75 months on (Supplementary Figure 19A). Among BKs, neutrophils kinetics appeared to
 323 be the most predictive, followed by CRP, albumin and LDH. However, the combined BK signature
 324 outperformed each individual BK, indicating that their collective predictive capabilities were not
 325 driven by any single biomarker alone.

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

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

331 The kML model can also be applied for the prediction of the outcome of a clinical trial (survival
 332 curves and associated hazard ratio), from early on-study data. We performed on-study runca-
 333 tions on the test set based on a number of weeks after the date of the first patient recruited (see
 334 methods). Here, we applied the model to predict not only patients receiving ATZ, but also doc-
 335 etaxel (Figure 1B). Predictions of the kML model applied to each arm yielded very accurate results
 336 when using data from the entire study (predicted HR = 0.784 (0.7 – 0.842)), versus data HR = 0.778
 337 (0.65 – 0.931), Figure 6A–B). Notably, the model prediction intervals were narrower than the data
 338 Kaplan-Meier confidence intervals, probably because the kML-trained model incorporates the in-
 339 formation from the three phase 2 trials. Using only early data, the model was already able to detect

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

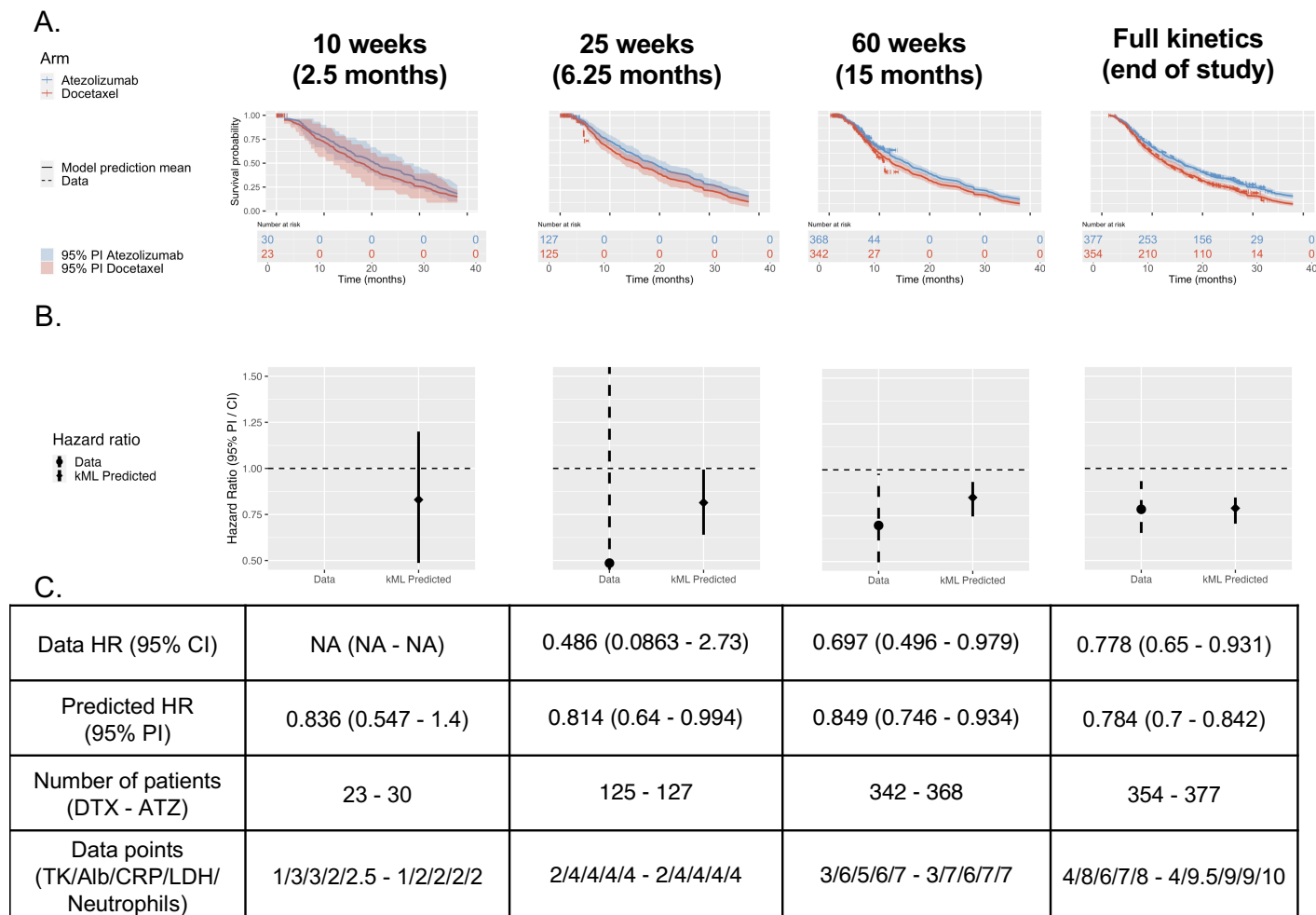


Figure 6. Use of kML for early-prediction of the outcome of a clinical trial **A.** Survival curves model-based predictions and prediction intervals versus actual data from on-study data at multiple horizon times after study initiation. Note that the model is able to predict full survival curves even 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.

340 a (non-significant) tendency at 10 weeks, with only 23 and 30 patients in each arm, and very short
 341 follow-up. Starting from data available at 25 weeks (6.25 months), the model correctly predicted
 342 a positive outcome of the study, with a 95% prediction interval of the HR below 1. Of note, the
 343 available data at this time (dashed lines, Figure 6A and red HR CIs in Figure 6B) was far from being
 344 conclusive. The model prediction was stable from 25 weeks on whereas the OS data only exhib-
 345 ited significant HR starting from 60 weeks and required more than 300 patients in each arm to be
 346 conclusive.

347 Discussion

348 Blood markers from hematology and biochemistry are routinely collected during clinical care or
 349 drug trials. They are cost-effective and easily obtained both before and during treatment. There
 350 is limited exploration regarding the predictive capabilities of the kinetics of such data. Combining
 351 BSL variables with on-treatment data (TK and BK), we addressed this question using a novel hy-
 352 brid NLME-ML methodology. The resulted kML model demonstrated excellent predictive perfor-

353 mances for OS in two aspects: 1) patient-level predictions (discrimination, calibration and patient
354 stratification) and 2) trial-level predictions. The kML model outperformed current state-of-the-art
355 methods based on either baseline or on-treatment data alone, utilizing only routine clinical infor-
356 mation, with a c-index of 0.79 and an accuracy of 78% for prediction of 12-month survival, on the
357 test dataset. Overall, kML incorporates 26 features, out of which 15 features require monitoring
358 five quantities over time (tumor size, albumin, CRP, LDH and neutrophils).

359 Regarding baseline markers, the predictive value of PD-L1 expression, commonly used in clinical
360 care, is controversial^{9,10}. Previous studies reported an AUC for durable response of 0.601 and
361 a PFS HR of 1.90 (PD-L1 $\geq 1\%$ vs 0%)⁸. Baseline tumor mutational burden showed similar predic-
362 tive value initially (AUC = 0.646)¹¹, but led to disappointing results in a recent prospective study⁴⁶.
363 Baseline blood counts were previously reported to predict overall survival^{43,47-49} and treatment
364 response (AUC = 0.74)⁴². The ROPRO score, derived from a large pan-cancer cohort and incorpo-
365 rating baseline clinical and biological data (27 variables) achieved a c-index of 0.69 and a 3-months
366 AUC of 0.743 for prediction of survival in the OAK clinical trial⁵⁰. Here, we confirmed these find-
367 ings and established a minimal signature of such data composed of only 11 variables (CRP, heart
368 rate, neutrophils to lymphocytes ratio, neutrophils, lymphocytes to leukocytes ratio, liver metas-
369 tases, ECOG, PD-L1 $\geq 50\%$, hemoglobin, SLD and LDH), yet with similar predictive performances
370 (*c-index* = 0.678) and significant stratification ability (*HR* = 1.74, *p* = 0.0014). Altogether, our kML
371 model demonstrated substantially better predictive performances than these baseline models.

372 We further confirmed the established predictive value of TK model-based parameters^{19,20}. Blood-
373 or serum-derived longitudinal markers kinetics have to date rarely been modeled. Gavrilov et al.
374 proposed to model NLR kinetics and demonstrated improved OS predictions over TK alone³¹. Here
375 we extended to four BKs: albumin, CRP, LDH and neutrophils. This choice was not only motivated
376 by observed statistical associations, but also from biological considerations. Albumin is associated
377 with nutritional status (cachexic state) and is known to evolve with time in responders. CRP is a
378 marker of systemic inflammation⁴⁴. Increased CRP, decreased albumin level, and increased CR-
379 P/albumin ratio have been reported to be associated with poor survival⁵¹. Neutrophils play a role
380 in inflammation by promoting a favorable microenvironment for cancer cell growth and spread,
381 and activation of carcinogenic signaling pathways⁵². Elevated LDH levels are a marker of cancer
382 cells turnover rate, and LDH has a potential role for prediction of potential invisible metastases⁴⁴.
383 We found that all these markers had non-trivial on-treatment kinetics. However, data fits were not
384 perfect, possibly due to the simplicity and empiric nature of the models we used. Further mecha-
385 nistic modeling of the joint kinetics of BKs and TK could bring relevant biological information and
386 yield more accurate predictive parameters. We found that all four BKs were contributive to the
387 model and that, combined, they outperformed TK performances.

388 We analyzed the RNAseq data using standard methods and found only negligible predictive
389 performances. Such result could be explained by the fact that the tissue of origin that was used
390 was heterogeneous across the patients (primary tumor or metastasis), was limited to a local area
391 of the tumor, and could come from tissue sampled long before treatment initiation. Given that
392 our main objective was to derive a predictive model from markers available in routine practice, we
393 excluded it from our minimal signature. A refined analysis, especially focusing on immune-based
394 signatures, could improve our results⁵.

395 Machine learning models, although increasingly used in pharmacological studies—including
396 recently for TK-OS modeling and variable selection^{22,53}—have yet rarely been rigorously compared
397 to classical statistical models²⁴. Here, such comparison revealed significantly better performance
398 of the nonlinear random survival forest RSF model compared to the linear proportional hazards
399 Cox model. In our approach, we did not use the propagation of standard statistical quantification
400 of the parameters' estimates uncertainty to evaluate the accuracy of the model predictions. Rather,
401 we relied on the RSF-outputted individual survival curves to sample virtual individuals and compute
402 prediction intervals.

403 A drawback of classical TK-OS studies is that they make use of the full observed kinetics to

404 predict overall survival, which can lead to time-dependent covariate bias⁵⁴ and limit their practical
405 applicability at bedside. We used individual-truncated data sets and found that kML was already
406 improving predictions over mBSL using data at 1.5 months, which corresponds to the first imag-
407 ing assessment of the treatment effect. At later times, stratification abilities increased to highly
408 significant levels (e.g., $HR = 5$ at 6.8 months).

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

413 Not only kML has value for personalized health care, but it also revealed useful for prediction
414 of a phase 3 trial using early on-study data. Our model was able to predict the study's positive
415 outcome with data at ~ 6 months, versus 10 months using TK only¹⁹. Relying on the data alone,
416 such positive outcome was only detectable at 15 months. These results could have important im-
417 plications for drug development as they could inform earlier on go/no-go decisions. Consequently,
418 this could allow to detect futility more easily and more rapidly during clinical trials, allowing to
419 avoid treating patients with an inefficient investigational treatment and to reassign funds and en-
420 ergy to other researches. Of note, in a recent evaluation based on resampling the first-line NSCLC
421 ATZ study IMpower150 to mimic small, short follow up early Phase Ib studies, TK model-based
422 metrics had better operating characteristics to predict Phase III success compared with RECIST
423 endpoints ORR and PFS⁵⁵. Extension of such results with the addition of BKs is thus a promising
424 line of research. In addition, kML, trained on ATZ data, yielded excellent predictive abilities for the
425 docetaxel (control) arm. This suggests that the relationships between TK / BK and OS might be
426 drug-independent. In turn, this opens future perspectives in terms of testing kML on drugs with
427 different mechanism of action, or combinations.

428 Further avenues of research comprise the development of integrative models from advanced
429 multi-modal data such as the one collected during the PIONeeR clinical study (NCT03833440)^{56,57}
430 that include quantitative image analysis from multiplex immune-histochemistry, genomic and tran-
431 scriptomic data, biological and clinical markers. In addition, mechanistic modeling of quantitative
432 and physiologically meaningful longitudinal data (immune-monitoring, vasculo-monitoring, circu-
433 lating DNA^{12,58-60}, soluble factors⁶¹, pharmacokinetics, TK and a large number of BKs from either
434 hematology or biochemistry) paves the way to an improved understanding and prediction of mech-
435 anisms of relapse to ICI⁶². Furthermore, the predictive abilities of kML — at the both individual and
436 study levels — should be evaluated in model-based prospective trials⁶³.

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

440 References

- 441 1. Bray, F., Ferlay, J., *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and
442 mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 394-
443 424. ISSN: 1542-4863. doi:[10.3322/caac.21492](https://doi.org/10.3322/caac.21492) (2018).
- 444 2. Duma, N., Santana-Davila, R. & Molina, J. R. Non-Small Cell Lung Cancer: Epidemiology, Screen-
445 ing, Diagnosis, and Treatment. *Mayo Clinic Proceedings*, 1623-1640. ISSN: 0025-6196, 1942-
446 5546. doi:[10.1016/j.mayocp.2019.01.013](https://doi.org/10.1016/j.mayocp.2019.01.013) (2019).
- 447 3. Fehrenbacher, L., Spira, A., *et al.* Atezolizumab versus docetaxel for patients with previously
448 treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised
449 controlled trial. *The Lancet*, 1837-1846. ISSN: 0140-6736, 1474-547X. doi:[10.1016/S0140-6736\(16\)](https://doi.org/10.1016/S0140-6736(16)00587-0)
450 [00587-0](https://doi.org/10.1016/S0140-6736(16)00587-0) (2016).

- 451 4. Grant, M. J., Herbst, R. S. & Goldberg, S. B. Selecting the optimal immunotherapy regimen in
452 driver-negative metastatic NSCLC. *Nature Reviews Clinical Oncology*, 625–644. ISSN: 1759-4782.
453 doi:[10.1038/s41571-021-00520-1](https://doi.org/10.1038/s41571-021-00520-1) (2021).
- 454 5. Camidge, D. R., Doebele, R. C. & Kerr, K. M. Comparing and contrasting predictive biomarkers
455 for immunotherapy and targeted therapy of NSCLC. *Nature Reviews Clinical Oncology*, 341–355.
456 ISSN: 1759-4782. doi:[10.1038/s41571-019-0173-9](https://doi.org/10.1038/s41571-019-0173-9) (2019).
- 457 6. Hutchinson, L. & Kirk, R. High drug attrition rates—where are we going wrong? *Nature Reviews*
458 *Clinical Oncology*, 189–190. ISSN: 1759-4782. doi:[10.1038/nrclinonc.2011.34](https://doi.org/10.1038/nrclinonc.2011.34) (2011).
- 459 7. Hua, T., Gao, Y., Zhang, R., Wei, Y. & Chen, F. Validating ORR and PFS as surrogate endpoints in
460 phase II and III clinical trials for NSCLC patients: difference exists in the strength of surrogacy
461 in various trial settings. *BMC Cancer*, 1022. ISSN: 1471-2407. doi:[10.1186/s12885-022-10046-z](https://doi.org/10.1186/s12885-022-10046-z)
462 (2022).
- 463 8. Rizvi, H., Sanchez-Vega, F., *et al.* Molecular Determinants of Response to Anti-Programmed
464 Cell Death (PD)-1 and Anti-Programmed Death-Ligand 1 (PD-L1) Blockade in Patients With
465 Non-Small-Cell Lung Cancer Profiled With Targeted Next-Generation Sequencing. *Journal of*
466 *Clinical Oncology*. doi:[10.1200/JCO.2017.75.3384](https://doi.org/10.1200/JCO.2017.75.3384) (2018).
- 467 9. Doroshow, D. B., Bhalla, S., *et al.* PD-L1 as a biomarker of response to immune-checkpoint
468 inhibitors. *Nature Reviews Clinical Oncology*, 345–362. ISSN: 1759-4782. doi:[10.1038/s41571-](https://doi.org/10.1038/s41571-021-00473-5)
469 [021-00473-5](https://doi.org/10.1038/s41571-021-00473-5) (2021).
- 470 10. So, W. V., Dejardin, D., Rossmann, E. & Charo, J. Predictive biomarkers for PD-1/PD-L1 check-
471 point inhibitor response in NSCLC: an analysis of clinical trial and real-world data. *Journal for*
472 *Immunotherapy of Cancer*, e006464. ISSN: 2051-1426. doi:[10.1136/jitc-2022-006464](https://doi.org/10.1136/jitc-2022-006464) (2023).
- 473 11. Hellmann, M. D., Ciuleanu, T.-E., *et al.* Nivolumab plus Ipilimumab in Lung Cancer with a
474 High Tumor Mutational Burden. *New England Journal of Medicine*, 2093–2104. doi:[10.1056/](https://doi.org/10.1056/NEJMoa1801946)
475 [NEJMoa1801946](https://doi.org/10.1056/NEJMoa1801946) (2018).
- 476 12. Gandara, D. R., Paul, S. M., *et al.* Blood-based tumor mutational burden as a predictor of clini-
477 cal benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nature Medicine*,
478 1441–1448. ISSN: 1546-170X. doi:[10.1038/s41591-018-0134-3](https://doi.org/10.1038/s41591-018-0134-3) (2018).
- 479 13. Cristescu, R., Mogg, R., *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-
480 based immunotherapy. *Science*, eaar3593. doi:[10.1126/science.aar3593](https://doi.org/10.1126/science.aar3593) (2018).
- 481 14. Sankar, K., Ye, J. C., *et al.* The role of biomarkers in personalized immunotherapy. *Biomarker*
482 *Research*, 32. ISSN: 2050-7771. doi:[10.1186/s40364-022-00378-0](https://doi.org/10.1186/s40364-022-00378-0) (2022).
- 483 15. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nature Medicine*,
484 1773–1784. ISSN: 1546-170X. doi:[10.1038/s41591-022-01981-2](https://doi.org/10.1038/s41591-022-01981-2) (2022).
- 485 16. Kurtz, D. M., Esfahani, M. S., *et al.* Dynamic Risk Profiling Using Serial Tumor Biomarkers for
486 Personalized Outcome Prediction. *Cell*, 699–713.e19. ISSN: 1097-4172. doi:[10.1016/j.cell.2019.](https://doi.org/10.1016/j.cell.2019.06.011)
487 [06.011](https://doi.org/10.1016/j.cell.2019.06.011) (2019).
- 488 17. Bonate, P. L. *Pharmacokinetic-Pharmacodynamic Modeling and Simulation* 2nd ed. 2011. ISBN:
489 978-1-4419-9484-4 (Springer-Verlag New York Inc., New York, 2011).
- 490 18. Claret, L., Girard, P., *et al.* Model-based prediction of phase III overall survival in colorectal
491 cancer on the basis of phase II tumor dynamics. *J Clin Oncol*, 4103–4108. doi:[10.1200/JCO.](https://doi.org/10.1200/JCO.2008.21.0807)
492 [2008.21.0807](https://doi.org/10.1200/JCO.2008.21.0807) (2009).
- 493 19. Claret, L., Jin, J. Y., *et al.* A Model of Overall Survival Predicts Treatment Outcomes with Ate-
494 zolizumab versus Chemotherapy in Non-Small Cell Lung Cancer Based on Early Tumor Kinet-
495 ics. *Clin Cancer Res*, 3292–3298. doi:[10.1158/1078-0432.CCR-17-3662](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1016/S0140-6736(16)32517-X) (2017).
28. Lavielle, M. *Mixed Effects Models for the Population Approach* ISBN: 1-4822-2650-2 (CRC Press, 2014).
29. Lixoft. *Monolix version 2020R1*. Antony, France, 2020.
30. 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](https://doi.org/10.1634/theoncologist.2008-0075) (2008).
31. Gavrillov, 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](https://doi.org/10.1002/psp4.12578) (2021).
32. 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](https://doi.org/10.1214/14-EJS890) (2014).
33. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187–220. ISSN: 0035-9246 (1972).
34. 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](https://doi.org/10.1145/2939672.2939785).
35. 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).

- 543 36. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical
544 tests. *JAMA*, 2543–2546. ISSN: 0098-7484 (1982).
- 545 37. Harrell, F. E. Hmisc: Harrell miscellaneous. *R package*. [https://CRAN.R-project.org/package=](https://CRAN.R-project.org/package=Hmisc)
546 [Hmisc](https://CRAN.R-project.org/package=Hmisc) (2022).
- 547 38. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-dependent ROC curves for censored survival
548 data and a diagnostic marker. *Biometrics*, 337–344. ISSN: 0006-341X. doi:[10.1111/j.0006-341x.](https://doi.org/10.1111/j.0006-341x.2000.00337.x)
549 [2000.00337.x](https://doi.org/10.1111/j.0006-341x.2000.00337.x) (2000).
- 550 39. Heagerty, P. J. & Saha-Chaudhuri, p. b. P. *survivalROC: Time-dependent ROC curve estimation*
551 *from censored survival data* tech. rep. (2013). <https://CRAN.R-project.org/package=survivalROC>.
- 552 40. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical*
553 *Society. Series B (Methodological)*, 267–288. ISSN: 0035-9246 (1996).
- 554 41. Eisenhauer, E. A., Therasse, P., *et al.* New response evaluation criteria in solid tumours: Revised
555 RECIST guideline (version 1.1). *European Journal of Cancer. Response assessment in solid tumours*
556 *(RECIST): Version 1.1 and supporting papers* 228–247. ISSN: 0959-8049. doi:[10.1016/j.ejca.2008.](https://doi.org/10.1016/j.ejca.2008.10.026)
557 [10.026](https://doi.org/10.1016/j.ejca.2008.10.026) (2009).
- 558 42. Benzekry, S., Grangeon, M., *et al.* Machine Learning for Prediction of Immunotherapy Efficacy
559 in Non-Small Cell Lung Cancer from Simple Clinical and Biological Data. *Cancers*, 6210. doi:[10.](https://doi.org/10.3390/cancers13246210)
560 [3390/cancers13246210](https://doi.org/10.3390/cancers13246210) (2021).
- 561 43. Havel, J. J., Chowell, D. & Chan, T. A. The evolving landscape of biomarkers for checkpoint in-
562 hibitor immunotherapy. *Nature Reviews Cancer*, 133–150. ISSN: 1474-1768. doi:[10.1038/s41568-](https://doi.org/10.1038/s41568-019-0116-x)
563 [019-0116-x](https://doi.org/10.1038/s41568-019-0116-x) (2019).
- 564 44. Blank, C. U., Haanen, J. B., Ribas, A. & Schumacher, T. N. The “cancer immunogram”. *Science*,
565 658–660. doi:[10.1126/science.aaf2834](https://doi.org/10.1126/science.aaf2834) (2016).
- 566 45. Bach, F. R. *Bolasso: model consistent Lasso estimation through the bootstrap* in (Association for
567 Computing Machinery, New York, NY, USA, 2008), 33–40. ISBN: 978-1-60558-205-4. doi:[10.](https://doi.org/10.1145/1390156.1390161)
568 [1145/1390156.1390161](https://doi.org/10.1145/1390156.1390161).
- 569 46. Peters, S., Dziadziuszko, R., *et al.* Atezolizumab versus chemotherapy in advanced or metastatic
570 NSCLC with high blood-based tumor mutational burden: primary analysis of BFAST cohort C
571 randomized phase 3 trial. *Nature Medicine*, 1831–1839. ISSN: 1546-170X. doi:[10.1038/s41591-](https://doi.org/10.1038/s41591-022-01933-w)
572 [022-01933-w](https://doi.org/10.1038/s41591-022-01933-w) (2022).
- 573 47. Soyano, A. E., Dholaria, B., *et al.* Peripheral blood biomarkers correlate with outcomes in ad-
574 vanced non-small cell lung Cancer patients treated with anti-PD-1 antibodies. *Journal for Im-*
575 *munotherapy of Cancer*, 129. ISSN: 2051-1426. doi:[10.1186/s40425-018-0447-2](https://doi.org/10.1186/s40425-018-0447-2) (2018).
- 576 48. Diem, S., Schmid, S., *et al.* Neutrophil-to-Lymphocyte ratio (NLR) and Platelet-to-Lymphocyte
577 ratio (PLR) as prognostic markers in patients with non-small cell lung cancer (NSCLC) treated
578 with nivolumab. *Lung Cancer (Amsterdam, Netherlands)*, 176–181. ISSN: 1872-8332. doi:[10.1016/](https://doi.org/10.1016/j.lungcan.2017.07.024)
579 [j.lungcan.2017.07.024](https://doi.org/10.1016/j.lungcan.2017.07.024) (2017).
- 580 49. Peng, L., Wang, Y., *et al.* Peripheral blood markers predictive of outcome and immune-related
581 adverse events in advanced non-small cell lung cancer treated with PD-1 inhibitors. *Cancer*
582 *immunology, immunotherapy: CII*, 1813–1822. ISSN: 1432-0851. doi:[10.1007/s00262-020-02585-](https://doi.org/10.1007/s00262-020-02585-w)
583 [w](https://doi.org/10.1007/s00262-020-02585-w) (2020).
- 584 50. Becker, T., Weberpals, J., *et al.* An enhanced prognostic score for overall survival of patients
585 with cancer derived from a large real-world cohort. *Annals of Oncology*, 1561–1568. ISSN: 0923-
586 7534. doi:[10.1016/j.annonc.2020.07.013](https://doi.org/10.1016/j.annonc.2020.07.013) (2020).
- 587 51. Yang, J.-R., Xu, J.-Y., *et al.* Post-diagnostic C-reactive protein and albumin predict survival in
588 Chinese patients with non-small cell lung cancer: a prospective cohort study. *Scientific Reports*,
589 8143. ISSN: 2045-2322. doi:[10.1038/s41598-019-44653-x](https://doi.org/10.1038/s41598-019-44653-x) (2019).

- 590 52. Bruni, D., Angell, H. K. & Galon, J. The immune contexture and Immunoscore in cancer progn-
591 nosis and therapeutic efficacy. *Nature Reviews Cancer*, 662–680. ISSN: 1474-1768. doi:[10.1038/
592 s41568-020-0285-7](https://doi.org/10.1038/s41568-020-0285-7) (2020).
- 593 53. Liu, G., Lu, J., Lim, H. S., Jin, J. Y. & Lu, D. Applying interpretable machine learning workflow to
594 evaluate exposure-response relationships for large-molecule oncology drugs. *CPT: pharmaco-
595 metrics & systems pharmacology*, 1614–1627. ISSN: 2163-8306. doi:[10.1002/psp4.12871](https://doi.org/10.1002/psp4.12871) (2022).
- 596 54. Desmée, S., Mentré, F., Veyrat-Follet, C. & Guedj, J. Nonlinear Mixed-Effect Models for Prostate-
597 Specific Antigen Kinetics and Link with Survival in the Context of Metastatic Prostate Cancer:
598 a Comparison by Simulation of Two-Stage and Joint Approaches. *The AAPS Journal*, 691–699.
599 ISSN: 1550-7416. doi:[10.1208/s12248-015-9745-5](https://doi.org/10.1208/s12248-015-9745-5) (2015).
- 600 55. Bruno, R., Marchand, M., *et al.* Tumor Dynamic Model-Based Decision Support for Phase Ib/II
601 Combination Studies: A Retrospective Assessment Based on Resampling of the Phase III Study
602 IMpower150. *Clinical Cancer Research*, OF1–OF9. ISSN: 1078-0432. doi:[10.1158/1078-0432.CCR-
22-2323](https://doi.org/10.1158/1078-0432.CCR-
603 22-2323) (2023).
- 604 56. Greillier, L., Monville, F., *et al.* Abstract LB120: Comprehensive biomarkers analysis to explain
605 resistances to PD1-L1 ICIs: The precision immuno-oncology for advanced non-small cell lung
606 cancer (PIONeer) trial. *Cancer Research*, LB120. ISSN: 0008-5472. doi:[10.1158/1538-7445.
607 AM2022-LB120](https://doi.org/10.1158/1538-7445.AM2022-LB120) (2022).
- 608 57. Barlesi, F., Monville, F., *et al.* Comprehensive biomarkers (BMs) analysis to predict efficacy of
609 PD1-L1 immune checkpoint inhibitors (ICIs) in combination with chemotherapy: a subgroup
610 analysis of the Precision Immuno-Oncology for advanced Non-Small Cell Lung CancER (PIO-
611 NeeR) trial. *Annals of Oncology*. doi:[10.1016/j.ionc.2022.09.001](https://doi.org/10.1016/j.ionc.2022.09.001) (2022).
- 612 58. Assaf, Z. J. F., Zou, W., *et al.* A longitudinal circulating tumor DNA-based model associated with
613 survival in metastatic non-small-cell lung cancer. *Nature Medicine*, 859–868. ISSN: 1546-170X.
614 doi:[10.1038/s41591-023-02226-6](https://doi.org/10.1038/s41591-023-02226-6) (2023).
- 615 59. Nabet, B. Y., Esfahani, M. S., *et al.* Noninvasive Early Identification of Therapeutic Benefit from
616 Immune Checkpoint Inhibition. *Cell*, 363–376.e13. ISSN: 1097-4172. doi:[10.1016/j.cell.2020.09.
617 001](https://doi.org/10.1016/j.cell.2020.09.001) (2020).
- 618 60. Cabel, L., Proudhon, C., *et al.* Clinical potential of circulating tumour DNA in patients receiv-
619 ing anticancer immunotherapy. *Nature Reviews. Clinical Oncology*, 639–650. ISSN: 1759-4782.
620 doi:[10.1038/s41571-018-0074-3](https://doi.org/10.1038/s41571-018-0074-3) (2018).
- 621 61. Barrera, L., Montes-Servín, E., *et al.* Cytokine profile determined by data-mining analysis set
622 into clusters of non-small-cell lung cancer patients according to prognosis. *Annals of Oncology:
623 Official Journal of the European Society for Medical Oncology*, 428–435. ISSN: 1569-8041. doi:[10.
624 1093/annonc/mdu549](https://doi.org/10.1093/annonc/mdu549) (2015).
- 625 62. Ciccolini, J., Benzekry, S. & Barlesi, F. Deciphering the response and resistance to immune-
626 checkpoint inhibitors in lung cancer with artificial intelligence-based analysis: when PIONeer
627 meets QUANTIC. *British Journal of Cancer*, 1–2. ISSN: 1532-1827. doi:[10.1038/s41416-020-0918-3
628](https://doi.org/10.1038/s41416-020-0918-3) (2020).
- 629 63. Ciccolini, J., Barbolosi, D., André, N., Barlesi, F. & Benzekry, S. Mechanistic Learning for Com-
630 binatorial Strategies With Immuno-oncology Drugs: Can Model-Informed Designs Help Inves-
631 tigators? *JCO Precision Oncology*, 486–491. doi:[10.1200/PO.19.00381](https://doi.org/10.1200/PO.19.00381) (2020).

632 **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

633
635

Supplementary 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

¹Median (IQR); n (%)

²Kruskal-Wallis rank sum test; Pearson's Chi-squared test

636
638

Supplementary 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 (48%)	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

¹n (%); Median (IQR)

²Pearson's Chi-squared test; Kruskal-Wallis rank sum test

639

640

Supplementary 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	

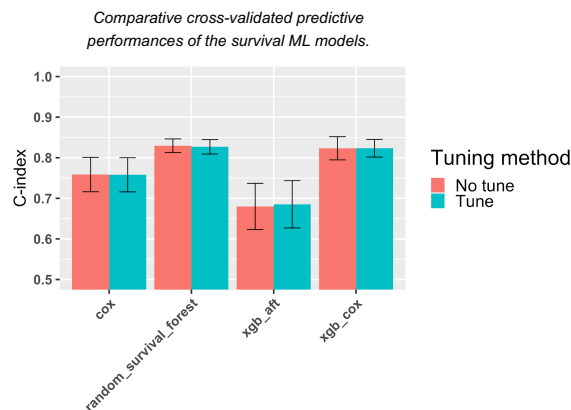
¹n (%)

²Pearson's Chi-squared test

642

643

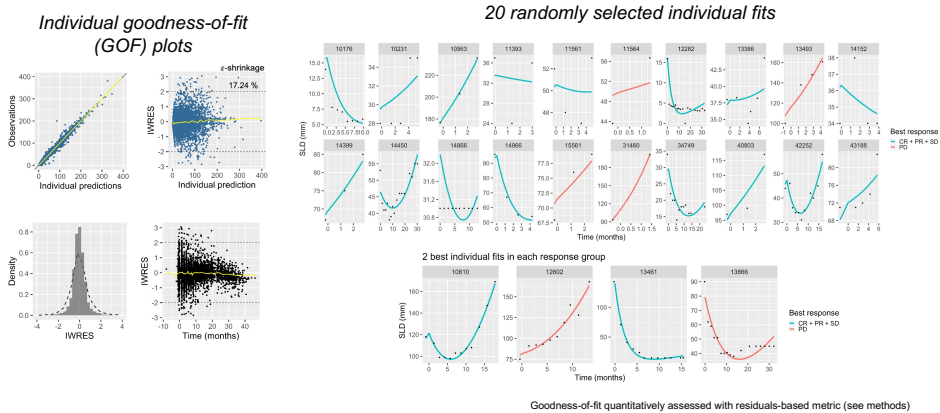
Supplementary Figure 4. Patient characteristics: PD-L1 and ECOG



645
646

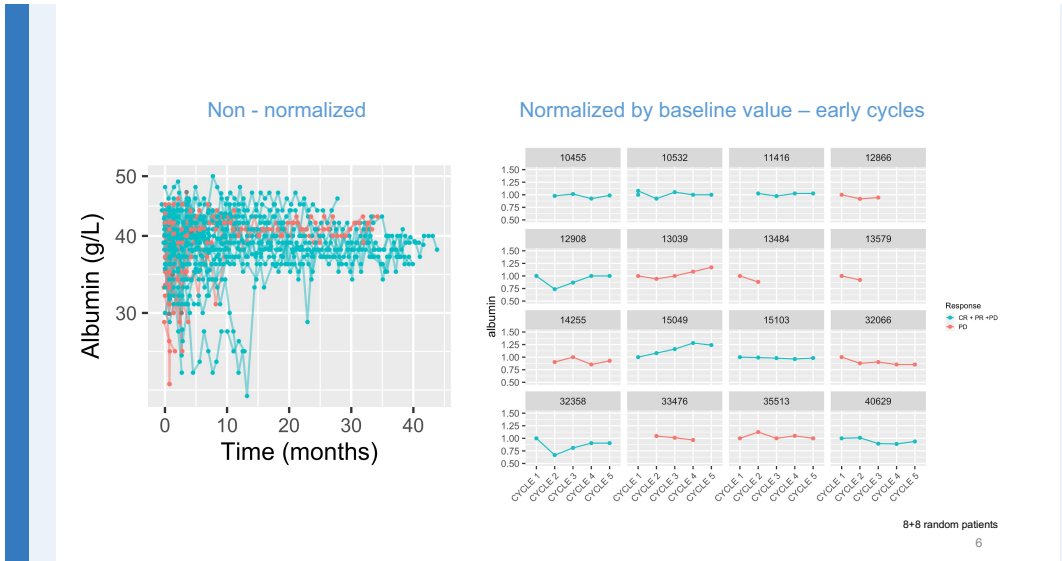
15

Supplementary Figure 5. Comparison of ML algorithms and tuning methods



648
649

Supplementary Figure 6. TK modeling goodness-of-fit

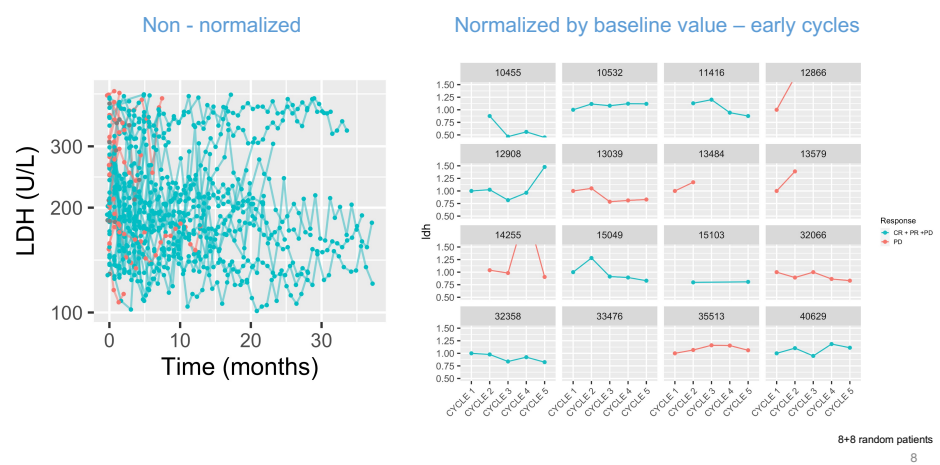


Supplementary Figure 7. Examples of longitudinal kinetics: Albumin



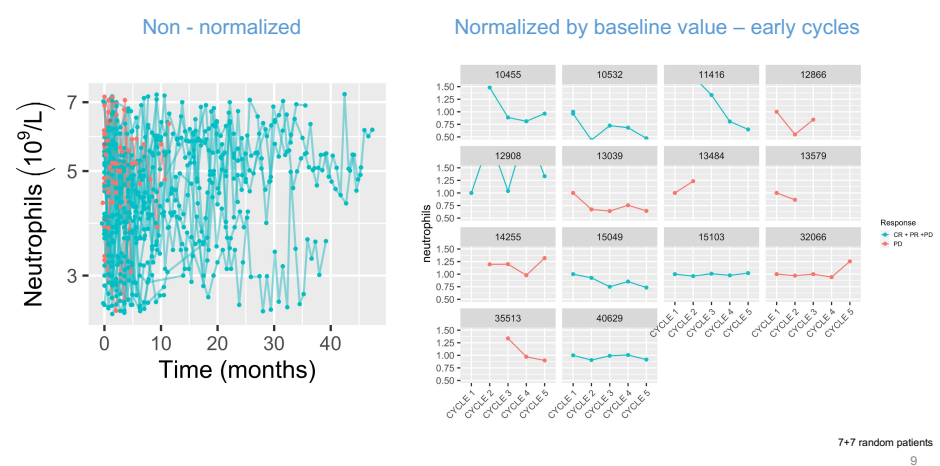
Supplementary Figure 8. Examples of longitudinal kinetics: CRP

657
659



Supplementary Figure 9. Examples of longitudinal kinetics: LDH

660
662



Supplementary Figure 10. Examples of longitudinal kinetics: Neutrophils

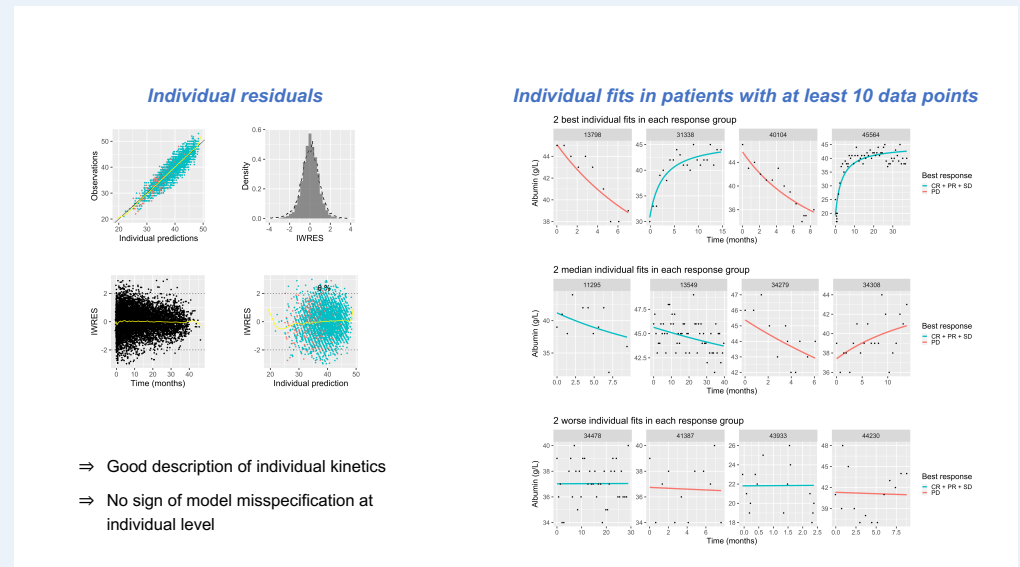
Model	Albumin		CRP		LDH		Neutrophils	
	BICc	b	BICc	b	BICc	b	BICc	b
Double-exponential	48,395	0.058	28,764	0.21	39,886	0.56	102,449	0.14
Hyperbolic	48,007	0.056	29,712	0.22	40,915	0.62	102,943	0.14
Linear	49,436	0.063	30,020	0.23	42,462	0.70	105,193	0.17
Constant	49,724	0.065	31,332	0.25	42,982	0.74	106,249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

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

10

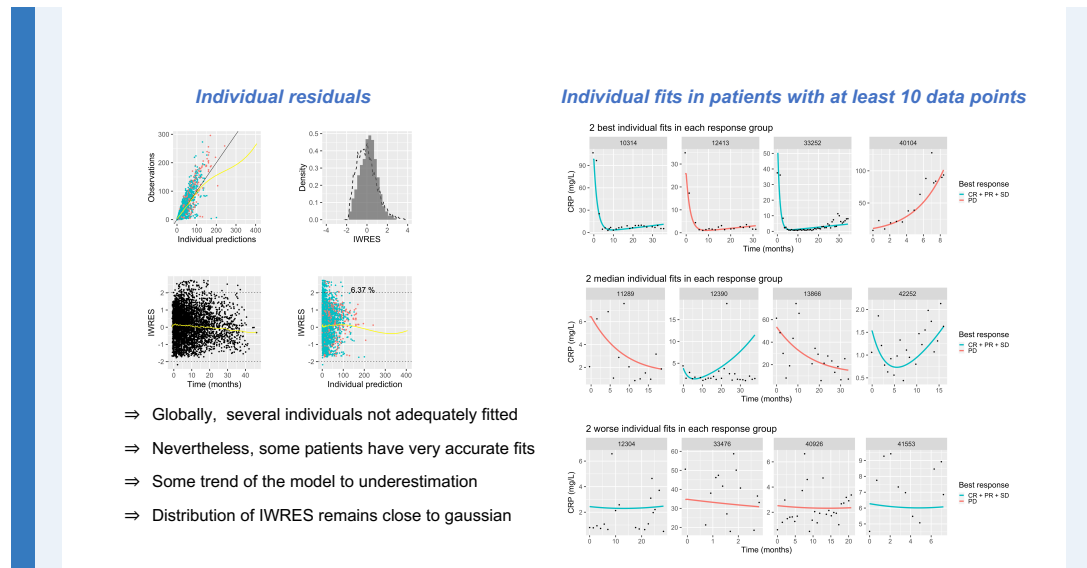
663
668

Supplementary Figure 11. Goodness-of-fit metrics of dynamic BK models



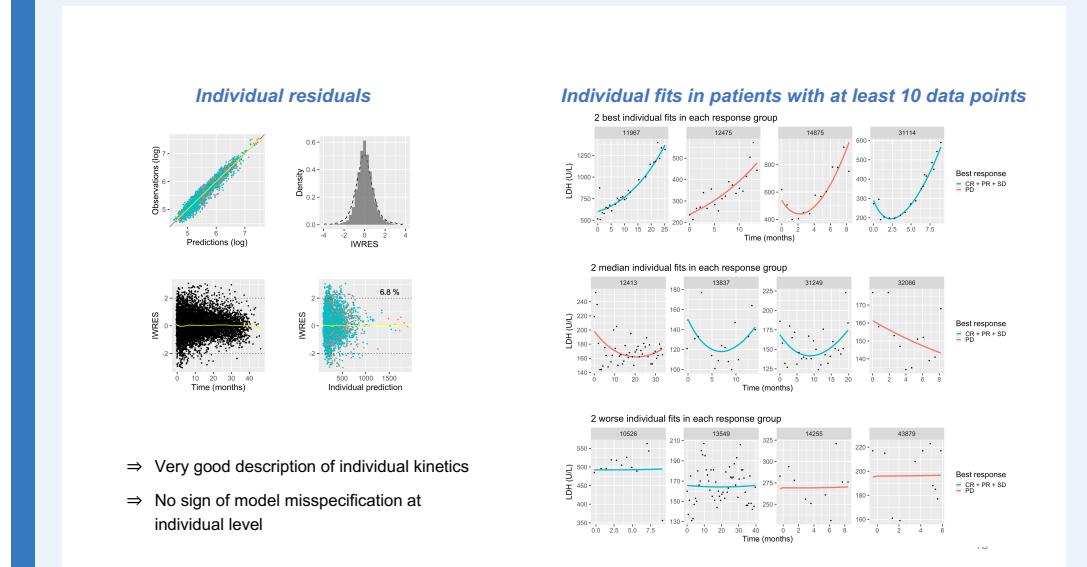
666
668

Supplementary Figure 12. Albumin: hyperbolic individual fits



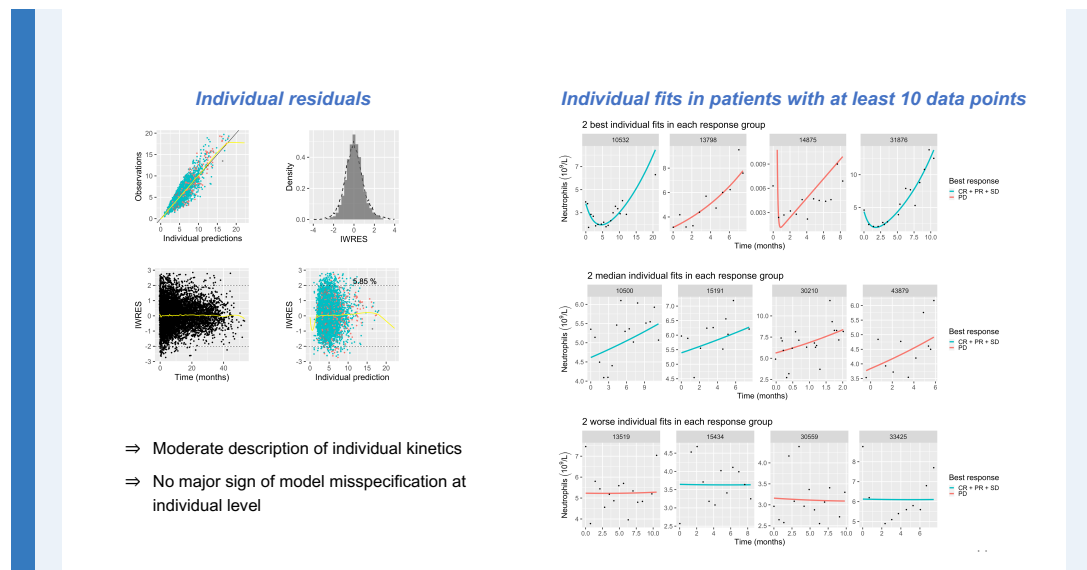
669
670

Supplementary Figure 13. CRP: dex individual fits



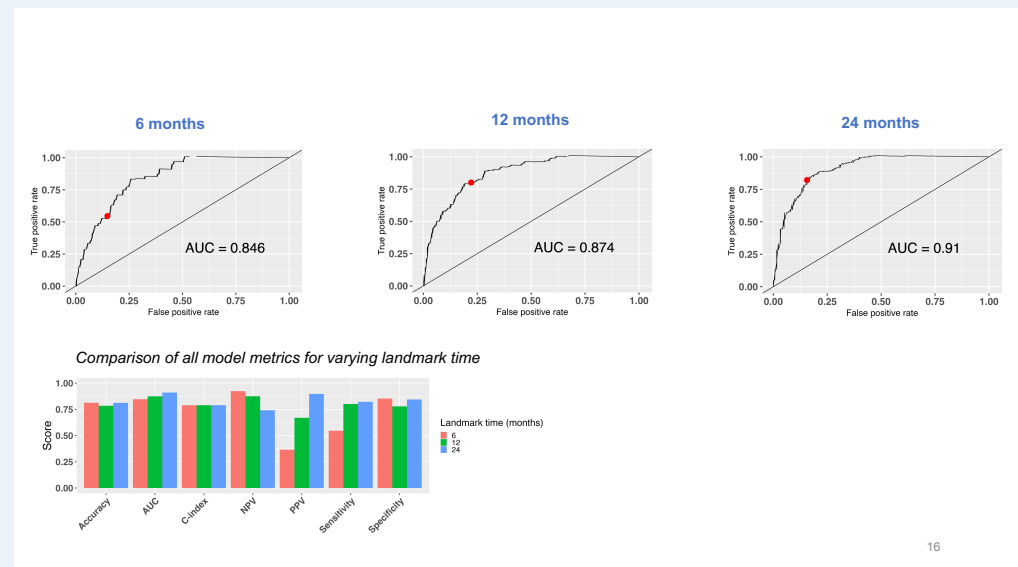
672
673

Supplementary Figure 14. LDH: dex individual fits



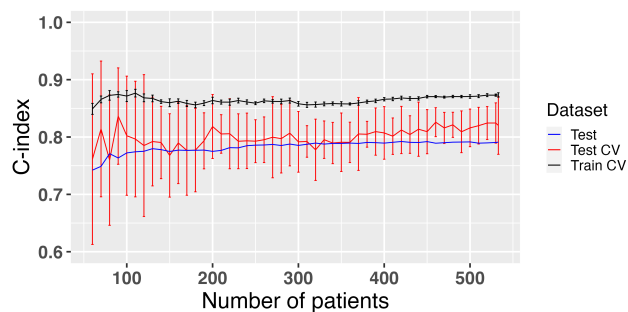
675
676

Supplementary Figure 15. Neutrophils: dexp individual fits



678
680

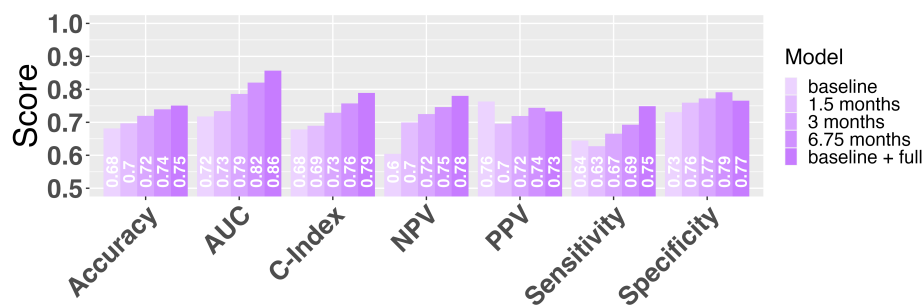
Supplementary Figure 16. ROC curves for variables landmark times (test set - OAK)



17

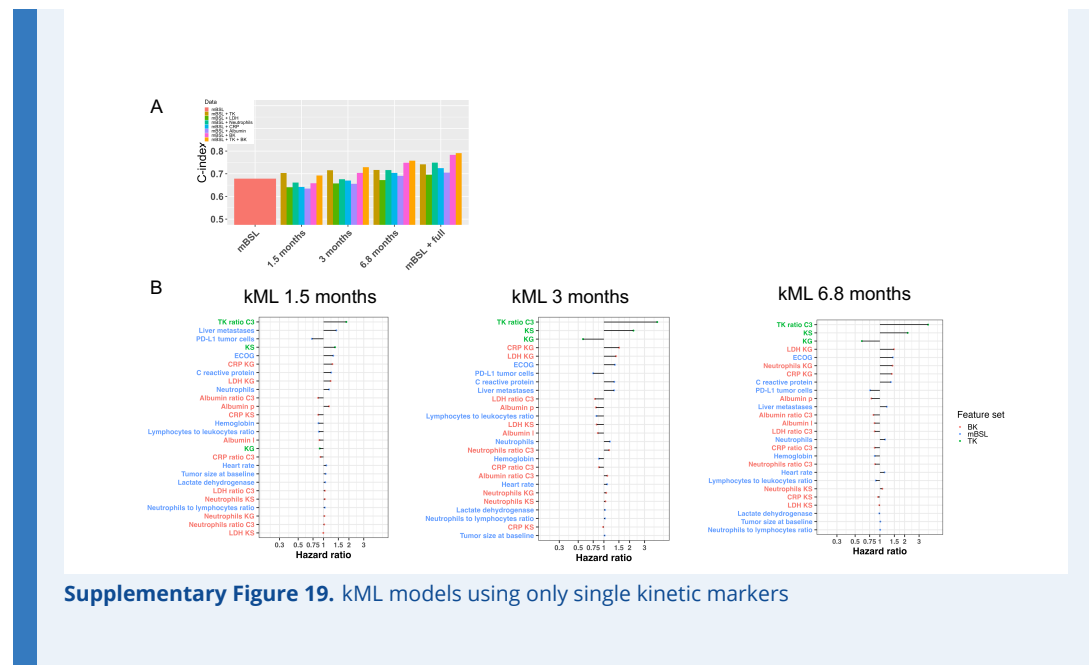
681
683

Supplementary Figure 17. Learning curve



684
686

Supplementary Figure 18. Additional value of NLME to baseline for multiple metrics



687
689

Supplementary Figure 19. kML models using only single kinetic markers

690 Supplementary Methods

691 Dimensionality reduction for RNAseq

692 Initial expression data from RNAseq consisted of 715 patients and 58,311 transcripts. The
693 first step of data filtering removed all transcripts with less than 10 read counts for all pa-
694 tients, then selected genes with highest variability between patients (top 15,000 transcripts
695 most variable). Then, data were normalized using upper quartile normalization which con-
696 sisted in dividing each read count by the 75th percentile of the read counts of the corre-
697 sponding sample and the final expression values were \log_2 transformed. Subsequently, a
698 univariable Cox regression model was employed to statistically assess the correlations be-
699 tween the expression levels of the transcripts and overall survival. Bonferroni correction
700 was used to adjust p-values from multiple univariate tests. This step was performed using
701 the `RegPara11e1` R package. We selected transcripts with high predictive values using follow-
702 ing criteria: adjusted log rank < 0.01 and HR < 0.85 or HR > 1.2 . The remaining transcripts
703 were used to perform a bootstrap Lasso Cox regression with cross-validation using mainly
704 the `glmnet` R package. Finally, the smallest number of transcripts with best predictive model
705 (highest C-index) was selected for further analysis.

706 Rules for BK processing

- 707 1. Observations outside lower (LB) and upper (UB) physiological bounds were discarded
708 using the following values, determined from discussion with a clinical oncologist: al-
709 bumin, LB = 10 g L^{-1} , UB = 100 g L^{-1} ; CRP, no LB, UB = 300 mg L^{-1} ; LDH, LB = 50 U/L ,
710 UB = 2000 U/L ; neutrophils, no LB, UB = 20.
- 711 2. For duplicates, the first one recorded was kept.
- 712 3. Denoting BK_n the value of the a "BK" at time t_n for a given patient, we excluded values
713 such that: $BK_n \notin (BK_{n-1}, BK_{n+1})$ AND $|BK_n - BK_{n-1}| > 3 \times sd_{BK}$ AND $|BK_n - BK_{n+1}| >$
714 $3 \times sd_{BK}$, where sd_{BK} is the standard deviation of $\{BK_n\}_n$, i.e. all time points for this
715 patient.
- 716 4. The BK value at the closest time point to treatment initiation was kept, provided this
717 time point was no more than 40 days before or 10 days after treatment initiation (oth-
718 erwise, patient was disregarded).

720 Nonlinear mixed-effects modeling

721 Denoting by $\mathcal{M}(t; \theta)$ a structural dynamic model that depends on time t and a set of pa-
722 rameters θ , longitudinal observations y_j^i in patient i at time t_j^i were assumed to follow the
723 observation model

$$724 y_j^i = \mathcal{M}(t_j^i; \theta^i) + \epsilon_j^i,$$

725 where $\epsilon_j^i \sim \mathcal{N}(0, \sigma_j^i)$ is the gaussian-distributed error model. The latter was either constant
726 ($\sigma_j^i = a, \forall i, j$) for TK or proportional ($\sigma_j^i = b \cdot \mathcal{M}(t_j^i; \theta^i), \forall i, j$) for BK. To describe inter-individual
727 variability, individual parameters θ^i were assumed to follow log-normal distributions:

$$728 \ln(\theta^i) = \ln(\theta_{pop}) + \eta^i, \quad \eta^i \sim \mathcal{N}(0, \omega^2)$$

729 with population-level parameters θ_{pop} and ω . Estimation of these was performed using the
730 stochastic approximation of expectation maximization algorithm implemented in the Mono-
731 lix software.
732
733
734