# Mechanistic Learning for Predicting Survival Outcomes in Head and Neck Squamous Cell Carcinoma

Kevin Atsou, Anne Auperin, Jôel Guigay, Sébastien Salas, Sébastien Benzekry

*CPT: Pharmacometrics & Systems Pharmacology*

ARTICLE OPEN ACCESS

# Mechanistic Learning for Predicting Survival Outcomes in Head and Neck Squamous Cell Carcinoma

Kevin Atsou[1] | Anne Auperin[2] | Jôel Guigay[3] | Sébastien Salas[1,4] | Sebastien Benzekry[1]

[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]Biostatistical and Epidemiological Division, Institut Gustave Roussy, Villejuif, France | [3]Clinical Oncology, Centre Antoine Lacassagne, Nice, France | [4]Clinical Oncology, Hôpital Timone, Aix-Marseille University, Marseille, France

**Correspondence:** Kevin Atsou (atsoukevin@gmail.com; kokoukevin.atsou@pfizer.com) | Sebastien Benzekry (sebastien.benzekry@inria.fr)

**Received:** 7 June 2024 | **Revised:** 18 November 2024 | **Accepted:** 5 December 2024

**Funding:** This work is part of the QUANTIC project funded by ITMO Cancer AVIESAN and the French Institute National du Cancer (grant #19CM148-00) and of the DIGPHAT project which was supported by a grant by a grant from the French government, managed by the National Research Agency (ANR), under the France 2030 program, reference ANR-22-PESN-0017.

**Keywords:** deep learning | head and neck squamous cell carcinoma | machine learning | survival analysis | tumor kinetics

## ABSTRACT

We employed a mechanistic learning approach, integrating on-treatment tumor kinetics (TK) modeling with various machine learning (ML) models to address the challenge of predicting post-progression survival (PPS)—the duration from the time of documented disease progression to death—and overall survival (OS) in Head and Neck Squamous Cell Carcinoma (HNSCC). We compared the predictive power of model-derived TK parameters versus RECIST and assessed the efficacy of nine TK-OS ML models against conventional survival models. Data from 526 advanced HNSCC patients treated with chemotherapy and cetuximab in the TPExtreme trial were analyzed using a double-exponential model. TK parameters from the first line and maintenance (TKL1) or after four cycles (TK4) were used to predict PPS and post-cycle 4 OS (OS4), combined with 12 baseline parameters. While ML algorithms underperformed compared to the Cox model for PPS, a random survival forest was superior for OS prediction using TK4 and surpassed RECIST-based metrics. This model demonstrated unbiased OS4 prediction, suggesting its potential for improving HNSCC treatment evaluation.

**Trial Registration:** ClinicalTrials.gov identifier: NCT02268695.

## 1 | Introduction

Head and Neck Squamous Cell Carcinomas (HNSCCs) arise from the mucosal epithelium of the oral cavity, pharynx and the larynx, posing a significant clinical challenge due to their aggressive nature and associated morbidity. The standard of care for resectable HNSCCs involves surgery followed by radiotherapy, with or without chemotherapy [1, 2]. For unresectable loco-regional HNSCCs, chemo-radiotherapy is the preferred treatment. In cases of local or metastatic recurrence where patients are not eligible for local treatments such as surgery or radiotherapy, standard treatment options includes the EXTREME regimen, which combines cisplatin and 5-fluorouracil (5-FU) with cetuximab [3] or a combination of cisplatin/carboplatin and 5-fluorouracil (5-FU) with pembrolizumab depending on the PD-L1 status [4]. These regimens are typically administered over 6 cycles of 21 days each, as a first line therapeutic approach. The GORTEC 2014-01 TPExtreme clinical trial aimed to compare the efficacy and safety of the TPEx regimen—4 cycles of docetaxel in combination with cisplatin and cetuximab followed by 500 mg / m$^2$ Cetuximab maintenance every 2 weeks—with the EXTREME

**Summary**

- What is the current knowledge on the topic?
  - Current surrogate markers for predicting overall survival (OS) in head and neck squamous cell carcinoma (HNSCC) rely on RECIST criteria. However, these do not account for the full dynamics of tumor kinetics (TK), limiting their predictive accuracy.
- What question did this study address?
  - This study evaluated the efficacy of model-derived TK parameters versus RECIST-based metrics for predicting post-progression survival (PPS) and overall survival (OS) in HNSCC patients, comparing traditional survival models with advanced machine learning (ML) and deep learning (DL) approaches.
- What does this study add to our knowledge?
  - The study demonstrated that TK parameters, especially when combined with random survival forests (RSF), provided superior OS predictions compared to RECIST metrics. While complex ML and DL models did not significantly outperform classical models, RSF was the most effective for OS prediction using TK data.
- How might this change drug discovery, development, and/or therapeutics?
  - Integrating TK-based metrics with ML models could improve early predictions of individual patient outcomes, potentially leading to more personalized and effective treatment strategies in HNSCC. This approach could refine the evaluation and monitoring of treatment responses beyond traditional RECIST guidelines.

regimen, potentially offering an alternative for the standard of care in the first-line treatment, particularly in patients who may not be good candidates for up-front pembrolizumab treatment [5]. Despite these therapeutic strides, the grim reality of a median survival oscillating between 12 and 15 months post-treatment [5] underlines a pressing need for more precise prognostic models to guide clinical decision-making and personalize therapeutic strategies.

Accurate prediction of overall survival (OS) from early surrogate markers is crucial for both for clinical care and drug development, especially to forecast the outcomes of phase 3 trials based on early interim data. Current surrogate markers to predict OS are based on the response evaluation criteria in solid tumors (RECIST) [6]. RECIST defines the best overall response (BOR) as the most favorable response recorded from treatment initiation to disease progression, based on the relative change in the sum of the largest diameters (SLD) of target lesions from baseline. Additionally, RECIST allows the definition of progression-free survival (PFS), which is the time from the start of treatment until objective tumor progression or death. However, PFS and BOR have been weakly supported as surrogate markers for OS [7–11].

To enhance prognostic models, on-treatment longitudinal SLD measurements have been employed to model TK using semi-mechanistic "tumor growth inhibition" (TGI) models [12, 13]. These models provide insights into the dynamics of tumor response to treatment over time and generate TK parameters for prognostic modeling. Several studies combining estimated TK parameters to survival analysis (TK-OS) revealed the importance of TK parameters in predicting OS, for example, in metastatic colorectal cancer or non-small cell lung cancer (NSCLC) [12–14]. However, a direct comparison between the predictive abilities of TK parameters versus RECIST-based BOR has not yet been performed.

Initially introduced to assist drug development (e.g., predicting phase 3 outcomes from phase 2 data) [12], TK parameters can also be used for individual OS prediction. However, in such context, on-treatment longitudinal data cannot be used to predict survival at a pre-treatment time point, as is the case when the outcome is OS [15, 16]. This phenomenon has been referred to as a type of time-dependent covariate bias and used as a criticism of some TK-OS studies [16]. For instance, using the entire duration of on-treatment data to predict OS is problematic because the total on-treatment duration is itself a strong predictor of OS. One possibility to alleviate this bias is to identify TK parameters from SLD measurements within a restricted timeframe to predict subsequent survival. Predicting post-progression survival (PPS), defined as the time from disease progression until death or last follow-up, is also a suitable alternative to OS. Specifically, we focused on two well-defined predictive tasks: (1) predicting post-cycle 4 OS (OS4) from tumor kinetics data up to that point (TK4-OS4) and (2) predicting post-progression survival (PPS) from the first line tumor kinetics only (TK1L-PPS). Note that point (2) is of particular interest to clinical oncologists because the time of failure of a first treatment is pivotal for therapeutic decision. Being able to accurately predict PPS could help to choose whether to give a treatment or not, and what treatment to give as second line in the first case.

To address this, we harnessed the strengths of population TK modeling, utilizing non-linear mixed effects (NLME), coupled with machine learning (ML) survival analysis. By benchmarking the currently available ML survival models, we compared the predictive power of TK parameters against BOR for survival prediction.

## 2 | Methods

### 2.1 | Patients

Data consisted of all patients enrolled in the GORTEC 2014-01 TPExtreme clinical trial [5]. Included patients had at least one measurable lesion according to the Response Evaluation Criteria in Solid Tumors (RECIST) v1.1 [6], ECOG performance status of 1 or 0, and were aged 18–70 years with histologically confirmed metastatic or recurrent squamous-cell carcinoma of the oral cavity, hypopharynx, oropharynx, or larynx, unsuitable for locoregional curative treatments. A total of 526 participants were randomly assigned (1:1) to each treatment group. On one hand, the TPEx regimen was administrated and consisted of 4 cycles of a combination of docetaxel, cisplatin, and cetuximab. On the other hand, the EXTREME regimen consisted of six cycles of the standard of care (fluorouracil, cisplatin, and cetuximab) (Figure 1A).

All enrolled patients gave written informed consent before any study procedure. The trial was carried out in respect of good clinical practice guidelines and the declaration of Helsinki. It

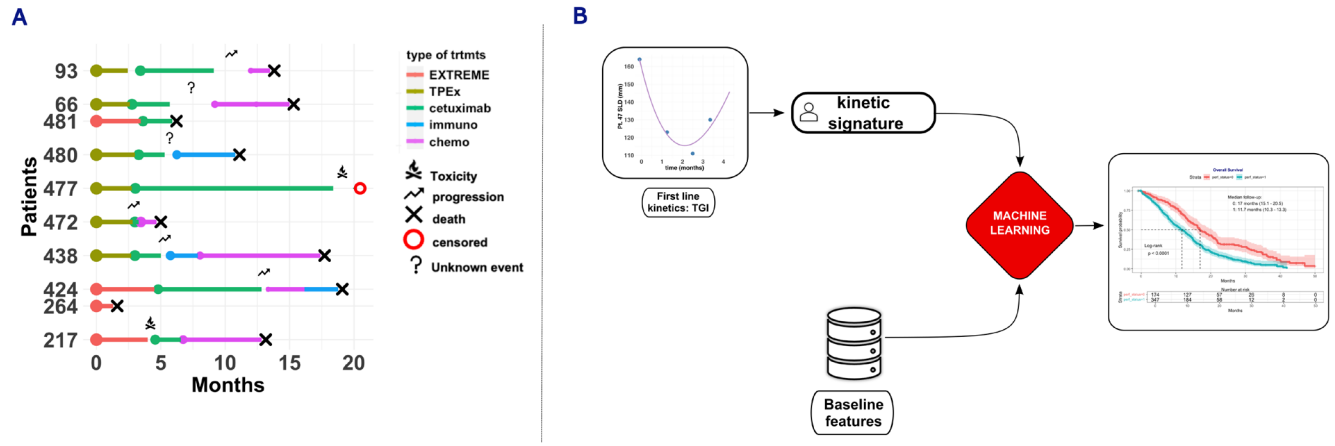


**FIGURE 1** | Treatment history and kinetics-ML prediction of survival. (A) Treatment history of a sample of patients. Cetuximab was the treatment given during maintenance. (B) Steps describing mechanistic learning. It combines modeling tumor kinetics and ML for survival prediction.

was approved by competent authorities and ethics committees in France, Spain, and Germany and registered with ClinicalTrials.gov under the reference code NCT02268695.

## 2.2 | Data

Data on 526 patients were gathered from patient records structured by a Case Report Form (CRF). Tumor response was measured every 8 weeks after the start of the treatment. Each measurement was performed by Computed Tomography (CT) scan or MRI for neck and CT scan for chest, and abdomen until disease progression. The sum of the largest diameters (SLDs) of target lesions and the best overall response (BOR) were computed according to RECIST v1.1. OS outcome was defined as the time from the start of treatment to death or to the (right censored) date of last follow-up and PPS as the duration from the time of documented disease progression after the first line treatment to death from any cause or to the (right censored) date of last follow-up. Data about treatments (type of treatment, treatment line, and toxicity events), disease characteristics (histology, metastasis status, loco-regional recurrence, tumor node metastasis (TNM) classification of the initial disease, TNM classification of the loco-regional recurrence), clinical, demographic, and epidemiological data (gender, performance status, tobacco status, alcohol status, age, and country) were also retrieved from the records. From the original set of over 40 baseline clinical and demographical features extracted from the data, 12 were retained for analysis based on their clinical relevance and their imbalance ratio, which refers to the extent of uneven distribution among different categories within these features. The dataset initially included seven patients who died before receiving the first treatment (three from the training set and four from the test set); these records were excluded from the analysis. Following this, the dataset exhibited 58 missing observations (1.32% missingness) in the training set and 23 missing observations (1.24% missingness) in the test set. To ensure robust model training and evaluation, rows containing missing values were also excluded. This step reduced the training set from 365 to 307 records and the test set from 154 to 131 records. Although the overall percentage of missingness was relatively low, these exclusions were necessary to maintain the integrity of the subsequent analyses and to avoid potential biases associated with imputation methods.

## 2.3 | Kinetics-Machine Learning

The predictive model was based on a specific two-step modeling approach called the kinetics-Machine Learning [17–19] (Figure 1B). It combines phenomenological modeling of the tumor response and ML techniques for survival analysis to build predictive models. The time evolution of the SLDs (tumor kinetics, TK) was modeled by tumor growth inhibition models [12, 13, 20]. Mixed-effects statistical learning techniques were used to estimate the individual parameters. Finally, using different ML methods for survival analysis, the estimated tumor kinetic parameters (patient-specific kinetic signature) were combined with relevant baseline features to predict individual patients' survival.

## 2.4 | Tumor Kinetics Models

### 2.4.1 | Structural Models

Multiple TK models were assessed for their ability to fit the observed SLD measurements. Let $t$, the time variable (months), TS, the tumor size (i.e., SLD, mm), $TS_0$ (mm), the tumor size at the first SLD assessment and $\tau$ (months), the time lag between first SLD measurement and the treatment start ($t = 0$). In the first model, tumor on-treatment regression or shrinkage is characterized by an exponential decrease with decay rate KS $\left(\text{months}^{-1}\right)$ and tumor growth is described by an exponential growth with a growth rate KG $\left(\text{months}^{-1}\right)$ (Table 1, model 1) [20]. The two other variants of this model introduce one more parameter [12, 13]. In the first one, the growth and shrinkage processes are respectively modulated by the factors $\alpha$ and $1-\alpha$, respectively describing proportions of resistant and non-resistant tumor cells (Table 1, model 2). In the third model, the decrease in the treatment effect is modeled by a decay rate $\lambda$ (Table 1, model 3) [12, 13]. For each model, an additional parameter, Time To re-Growth (TTG), known for its ability to predict survival in other tumor types, was also assessed and tested as a predictor of OS [13].

**TABLE 1** | Different TK models.

| Name | Model | Parameters | TTG |
|---|---|---|---|
| Model 1 | $\mathrm{TS}(t,\Theta)=\mathrm{TS}_0\left(e^{\mathrm{KG}(t-\tau)}+e^{-\mathrm{KS}(t-\tau)}-1\right)$ | $\mathrm{TS}_0, \mathrm{KG}, \mathrm{KS}$ | $\frac{\log\left(\frac{\mathrm{KS}}{\mathrm{KG}}\right)}{\mathrm{KS}+\mathrm{KG}}+\tau$ |
| Model 2 | $\mathrm{TS}(t,\Theta)=\mathrm{TS}_0\left(\alpha e^{\mathrm{KG}(t-\tau)}+(1-\alpha)e^{-\mathrm{KS}(t-\tau)}\right)$ | $\mathrm{TS}_0, \mathrm{KG}, \mathrm{KS}, \alpha$ | $\frac{\log\left(\frac{\mathrm{KS}(1-\alpha)}{\alpha\mathrm{KG}}\right)}{\mathrm{KS}+\mathrm{KG}}+\tau$ |
| Model 3 | $\mathrm{TS}(t,\Theta)=\mathrm{TS}_0 e^{-\left(\frac{\mathrm{KS}}{\lambda}\right)\left(1-e^{-\lambda(t-\tau)}\right)+\mathrm{KG}(t-\tau)}$ | $\mathrm{TS}_0, \mathrm{KG}, \mathrm{KS}, \lambda$ | $\frac{\log\left(\frac{\mathrm{KS}}{\mathrm{KG}}\right)}{\lambda}+\tau$ |

Abbreviation: TTG, Time To re-Growth.

#### 2.4.2 | Statistical Model

The framework of the Nonlinear Mixed Effects Modeling (NLME) was used as a statistical tool to describe the inter-individual variability in the observed tumor kinetics. Let $Y^i=\left\{y_{i1},\ \cdots,y_{in^i}\right\}$, for $i\in\{1,\ \cdots,N\}$, the vector of longitudinal SLD measurements for the $i$th patient, where $N$ represents the total number of patients and $n^i$, the number of measurements performed for patient $i$. The statistical model used in the fitting step was given by

$$y_{ij}=\mathrm{TS}^{(k)}\left(t_{ij};\theta_i^{(k)}\right)+e_{ij}^{(k)} \tag{1}$$

where $\mathrm{TS}^{(k)}\left(t_{ij},\theta_i\right)$ is the value of a specific structural model $k$ (for $k\in\{1,2,3\}$) (Table 1) at time $t_{ij}$, $\theta_i^{(k)}\in\mathbb{R}^p$ is the corresponding vector of parameters specific to the individual $i$ and $e_{ij}^{(k)}$ represents the residual error model. The vector of individual parameters $\theta_i^{(k)}$ was described by the combination of fixed effects $\theta_{pop}^{(k)}$, which are constant among the population and random effects $\eta_i^{(k)}$ which describe the inter-individual variability observed in the SLD kinetics. The random effects were assumed to be normally distributed with mean zero and variance–covariance matrix $\Omega^{(\mathrm{k})}=diag\left(\omega_1,\ \cdots,\omega_{l^{(k)}}\right)$ (where $l^{(k)}$ refers to the number of kinetic parameters in a specific model $k$). Furthermore, the positivity of the parameters was ensured by assuming a log-normal distribution on the individual parameters. More precisely,

$$\log\theta_i^{(k)}=\log\left(\theta_{pop}^{(k)}\right)+\eta_i^{(k)},\eta_{\mathrm{i}}^{(k)}\sim\mathcal{N}\left(0,\Omega^{(\mathrm{k})}\right) \tag{2}$$

For the different structural models, the error model was assumed to be constant and was defined as follows:

$$e_{ij}^{(k)}=\sigma^{(k)}\,\varepsilon_{ij} \tag{3}$$

where $\varepsilon_{ij}\sim\mathcal{N}(0,1)$ represents the residual error and $\sigma^{(k)}$ its standard deviation. In summary, the vector of population parameters is:

$$\psi^{(\mathrm{k})}=\left(\theta_{pop}^{(k)^T},\sigma^{(k)},\omega^{(k)^T}\right)$$

where $\theta_{pop}^{(k)^T}$ and $\omega^{(k)^T}$ represent respectively the entries of the vector of fixed effects parameters and the entries of the vector of the random effects standard deviations.

#### 2.4.3 | Model Calibration and Selection

The parameters $\mathrm{TS}_0$ and $\tau$ were retrieved from the data. Some SLD measurements which were censored and set to zero due to the maximum CT (or MRI) scan slice thickness were reset to an offset of 2.5 mm (half of the value required for a lesion to be considered as measurable according to RECIST v1.1) and marked as left-censored. The contribution of censored measurements to the observed likelihood function was handled by lower Limit Of Quantification (LOQ) censoring [21]. Furthermore, the population parameters $\psi^{(k)}$ of each statistical model were estimated by computing the maximum likelihood estimate $\hat{\psi}^{(k)}$ of the observed likelihood function with the Stochastic Approximation of the Expectation–Maximization Algorithm implemented in the Monolix R API (Version 2021R2). The Fisher information matrix was derived from the observed likelihood by a Markov Chain Monte Carlo algorithm implemented in Monolix. The relative standard errors (RSE) and corresponding confidence intervals were computed for each component of $\hat{\psi}^{(k)}$. The estimator $\hat{\theta}_i^{(k)}$ of the individual kinetic parameters was defined as the mode of the posterior conditional distribution $p\left(\theta_i^k\mid y_{ij};\hat{\psi}^{(k)}\right)$.

Goodness-of-fit of the NLME model was assessed by the Bayesian information criterion (BIC). A RSE of $\leq 50\,\%$ was established as a threshold to ensure parameter stability and model identifiability, thereby indicating that the parameter estimates were reliable and precise. Additionally, diagnostic plots, including individual curve fits and comparisons of observed versus predicted SLD measurements, were employed to illustrate and validate the model's performance.

### 2.5 | ML for Survival Analysis

The data on the 526 patients were split into a training and test set, which represented respectively 70% and 30% of the full dataset. Initial data pre-processing was performed on the training set, from which the scaling parameters—specifically the mean and standard deviation of continuous variables—were computed to standardize the features. These scaling parameters were then applied to the test set to maintain consistency. TK model selection and multiple survival ML algorithms were benchmarked on the training set. The best TK model selected on the training set was calibrated on the test set using the NLME priors obtained on the training set and the corresponding kinetic parameters were retrieved to assess the accuracy of the best ML model selected

on the training set. The estimated individual kinetic parameters—including TTG—were added to the baseline features for the assessment of different predictive ML and statistical models for survival analysis.

To avoid skewness and to be consistent with the log-normal hypothesis in the modeling process, a logarithmic transformation was applied to the parameters KG and KS.

#### 2.5.1 | Benchmarking Procedure

Eight predictive ML models—CoxTime [22], DeepHit [23], DeepSurv [24], LogHaz [25, 26], PC-Hazard [27], Survival SVM [28, 29], XGBoost [30], and Random Survival Forest (RSF) [31]—and five classical parametric and semi-parametric survival models—Cox Proportional Hazard (PH), Accelerated Failure Time (AFT) [32], Cox Lasso [33], CoxBoost, and GLMBoost [34] and [35]—were assessed for their ability to predict PPS. A brief description of each model and its methodological approach is provided in Table S1. The aforementioned models were compared by computing Harrell's Concordance index (c-index) [36] using nested and repeated k-fold cross-validation. Each inner loop contained 30 repeats of twofold cross-validations, used to tune the ML models. Each outer loop consisted of 10 repeats of fivefold cross-validations, used to assess the generalizability of the different models. The models were tuned by performing a random search on the numerical hyperparameters domain. Overall, 50 training and test sets were used to compute the predictive performance of each model. The benchmarking procedure and the different models were implemented using the *mlr*3 packages in *R* version 4.1.0.

To assess the features importance, a permutation-based method was applied to the best ML model on the training set. However, for cases where Cox PH emerged as the best model, the Hazard Ratios (HR) were used to quantify feature significance. To stratify the survival curves based on TK parameters, the maximally selected rank statistic from the *maxstat R* package was employed.

#### 2.5.2 | Test Step

First, the best TGI model obtained on the training set was calibrated on the test set using the priors obtained on the training set and the estimated kinetic parameters were added to the baseline features. Afterwards, using 1000 bootstrap sets derived from the test set, the accuracy of the best performing ML model was assessed on the test set by computing the c-index and its 95% confidence interval.

#### 2.5.3 | Comparison of Different Predictor Models

The best model selected from the benchmarking step was used to assess the predictive quality of different sets of predictors: baseline (BSL) features, BSL + BOR, BSL + TK, and TK. These models were assessed on the training set using 10 repeats of fivefold cross validations. The generalization quality of the models was assessed on the test set using 1000 bootstrap sets. Furthermore, a Kruskal–Wallis test was used to compare the c-index of the models.

### 2.6 | Unbiased Two-Step Modeling Approaches: TK4-OS4 and TKL1-PPS

It would be biased to predict the time-to-death of a patient from baseline (i.e., OS) using TK data that is not available at this time. To avoid this, we introduced two unbiased two-step modeling approaches. The first, denoted TK4-OS4, used 4 cycles SLD data to estimate TK parameters (TK4) and predicts post cycle 4 OS (OS4). The second approach, denoted TKL1-PPS, used TK parameters identified from longitudinal SLD measurements restricted to the first line of treatment + the maintenance period (e.g., the time to disease progression), to predict PPS.

## 3 | Results

### 3.1 | Patients and Disease Characteristics

Overall, 526 patients treated with either the EXTREME or TPEx regimen were retrieved from the GORTEC 2014-01 TPExtreme clinical trial [5]. Patients and disease characteristics are summarized in Table S3. In this study, 60.1% of the patients received a second-line treatment while 33.2% received a third-line treatment, and finally 13.3% received a fourth-line treatment. In this study, 44.6% of the primary tumors were in the oropharynx and 64% of the patients had metastasis at inclusion (Table S3). The typical treatment history of a patient is depicted in Figure 1A. It highlights the inter-patient heterogeneity in terms of both efficacy and toxicity, thus emphasizing the complexity of clinical management of such patients at bedside.

### 3.2 | Tumor Kinetics Modeling

The SLD measurements on 365 (70%) patients—restricted to four cycles (total number of time points = 672 and median number of time points per patient = 2)—were used to calibrate the different TK models. Among the tested models, Model 1 demonstrated the best agreement with the data, as indicated by the lowest values for all information criteria (AIC, BIC, and BICc), suggesting a superior balance between model fit and complexity. Furthermore, model 1 was found to be the best to describe the data in the different cases (TK4 and TKL1) (Table S1 for TK4). All the parameters were identifiable with a maximum RSE of 8.73%, confirming model stability and parameter identifiability (Table S2).

Diagnostic plots did not reveal any misspecification of the model at the individual level, supporting the adequacy of model 1 for describing the observed tumor dynamics (Figure 2A,B). Recalibrating the model using the treatment arm as a covariate to the kinetic parameters (KS, KG) revealed a potential effect of the TPEx treatment on the tumor growth rate (KG) in comparison to its effect on the tumor shrinkage rate (KS). Indeed, a Wald test demonstrated a statistically significant association between the treatment arm and KG, ($p = 6.35 \times 10^{-3}$) But no association was found between the treatment arm and KS. Furthermore,

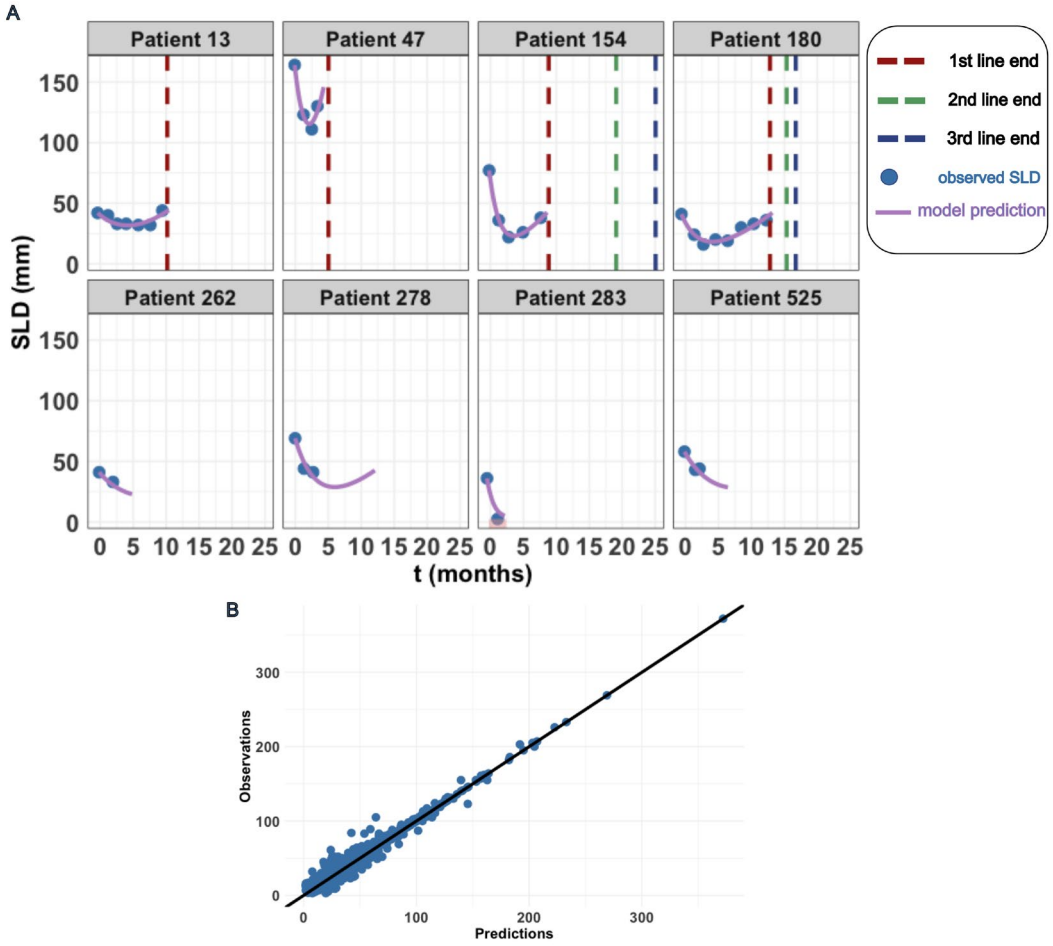


**FIGURE 2** | Diagnostic plots (A) Representative individual fits. The purple line represents the model prediction and the blue dots the observed SLDs. The red, green and dark-blue dotted vertical lines represent the end of the first, second line and third lines, respectively. (B) Observed SLDs vs. predicted SLDs.

incorporating the treatment arm or any additional covariates to the model resulted in a decreased goodness-of-fit, introduced identifiability issues and did not show any significant improvement in the survival analysis. Given the goal of predicting outcomes on the entire patient cohort without introducing covariate-related bias, a model without covariates was kept for the predictive analysis. This choice ensured a more generalizable model for the population, regardless of specific treatment arm effects. The differences in the models did not significantly alter the overall conclusions regarding TK parameters and survival predictions.

## 3.3 | ML For Survival Analysis

### 3.3.1 | ML Models Benchmarking Unbiased Predictions Using TK4-OS4 and TK1L-PPS

To avoid time-dependent covariate bias, we considered either predicting post-cycle 4 survival (OS4) from data restricted to the first four treatment cycles (TK4-OS4) or predict post-progression survival (PPS) from TK parameters derived from the data during the first line of treatment until progression (TKL1-PPS). On the training set, for both TK4-OS4 and TKL1-OS, the random survival forest (RSF) model did not significantly outperformed Cox PH or AFT (median c-indices = 0.65, 0.635, and 0.63 for TK4-OS4, Figure 3 and 0.601, 0.598, and 0.595 for TKL1-OS, Figure S1). However, an examination of the results on the test set for TK4-OS4 revealed that RSF exhibited the best generalization properties in comparison to Cox PH and AFT (Figure S2). Overall, TKL1-PPS (best training c-index = 0.6) exhibited worse performances than TK4-OS4 (best training c-index = 0.65, Figure 3B). Moreover, deep learning architectures, specifically DeepHit and DeepSurv, exhibited suboptimal performances for both TK4-OS4 and TKL1-PPS. Upon examination of the test set, for TK4-OS4, the Cox PH model achieved a mean c-index of 0.54 (95% CI: 0.52–0.56) (Figure S2).

### 3.3.2 | Feature Importance

Noteworthy, the most important feature for OS4 prediction was first the patient performance status followed by the TK parameters (logKG, logKS, and TTG) and the baseline tumor size $TS_0$ (Figure 4A). In addition, utilizing the maximally selected rank statistic, $TS_0$, logKG and logKS provided a significant separation of the survival curves (log-rank $p < 0.01$ and $p < 0.05$, respectively) (Figure 4B–D).

However, for TKL1-PPS, the top 3 parameters were loco-regional relapse, followed by $TS_0$, the performance status (Figure S3A). Surprisingly, Unlike TK4-OS4, logKS, triggered crossed survival

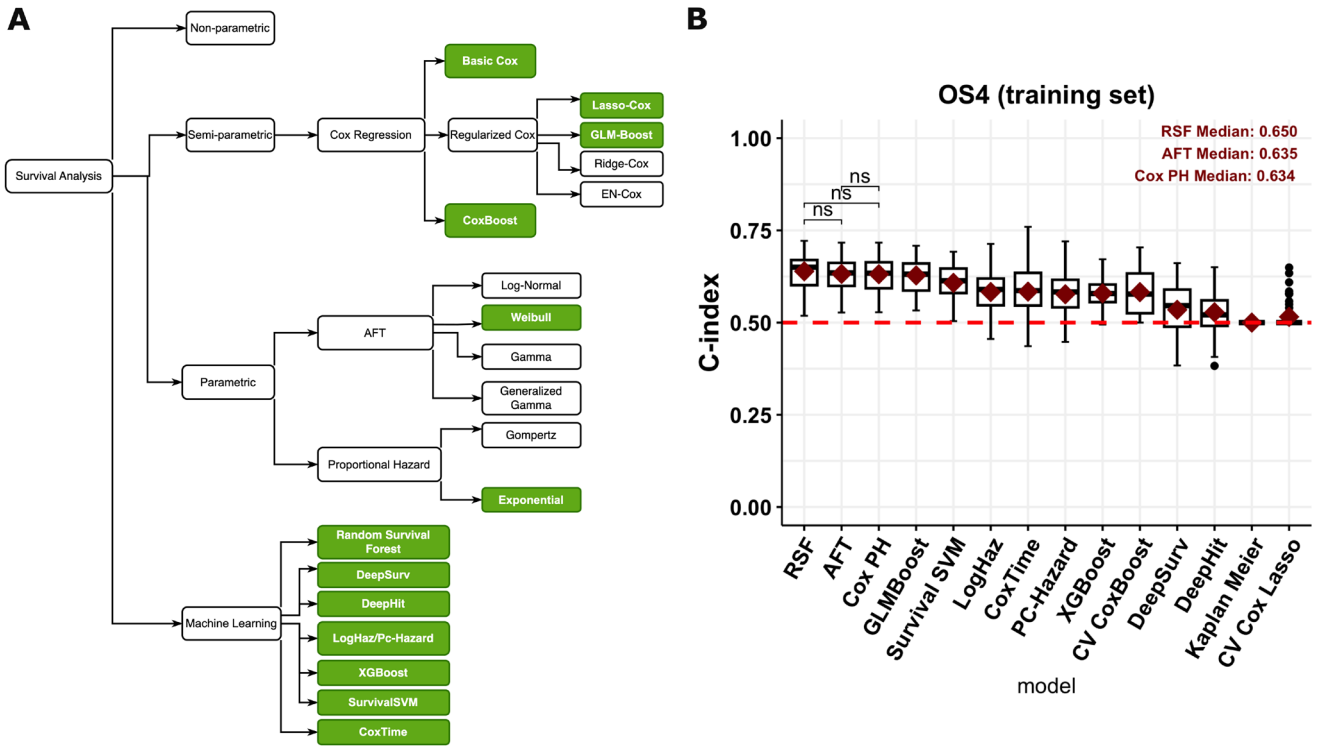


**FIGURE 3** | Benchmark of ML methods, TK4-OS4. (A) Graphical representation of the different ML and statistical methods. The green boxes are the tested models. (B) Boxplots of the results of the benchmarking process on OS4 prediction. The red Diamond is the median c-index and the middle line the mean c-index.

curves (Figure S3D) potentially indicating a time-dependent nature of the growth rate parameter.

### 3.3.3 | Comparison of Different Predictor Models: TK4 Parameters Are Better Predictors of OS4 Than BOR

To compare the relative predictive performances of BOR and TK4 as predictors of OS4, four sets of predictors were considered in combination with RSF: 12 baseline features (BSL, see Table S4) alone, BSL + BOR, BSL + TK4 and TK4 (Figure 5). In the training set, the BSL + TK4 outperformed BOR + BSL 0.63 (95% CI: 0.61 − 0.65) versus 0.612 (95% CI: 0.58 − 0.64), respectively, $p < 0.001$, Kruskal-Wallis, (Figure 5A). This trend persisted in the test dataset, with respective c-index of 0.63 (95% CI: 0.61 − 0.65) and 0.52 (95% CI: 0.50 − 0.54) ($p < 0.0001$, Figure 5B). The integration of BSL with BOR adversely impacted the model accuracy on test set as compared to BSL alone (Figure 5B). However, this was not observed in TKL1-PPS, such behavior was absent (Figure S4A,B).

## 4 | Discussion

Tumor kinetics modeling to predict overall survival (TK-OS) has been extensively developed in the last 15 years [37]. Multiple cancer types, often in the advanced or metastatic stage have been studied, including prostate [38], breast [39], colorectal [12], non-small cell lung cancer (NSCLC) [40], or renal cell carcinoma [41]. Our study is the first to apply such modeling to HNSCC, using data from the TPExtreme trial [5]. In addition, it is one of the few to consider and benchmark machine learning survival models versus classical (semi-)parametric survival analysis methods [42].

As often observed for the other cancer types, such as NSCLC [14], breast, and colorectal cancer [39], on-treatment TK could be accurately described by a simple pattern mathematically: the sum of two exponentials. A decreasing phase is governed by a shrinkage parameter KS and a regrowth phase is led by a parameter KG.

The primary objective of the TK modeling was to assess the predictive value of the TK parameter on the survival outcomes. Incorporating the treatment arm as a covariate to investigate its potential impact on tumor dynamics resulted in a reduced model performance, characterized by decreased goodness-of-fit and identifiability issues. Consequently, covariates (including treatment arms) were excluded from the final model to ensure identifiability and robustness. Despite these limitations, it is noteworthy that the exploratory analysis of treatment arms as covariates using simulated SLDs revealed a significant effect of the TPEx treatment on the tumor growth rate (KG), indicating a potential differential impact of the treatment regimens on tumor dynamics. Specifically, Using the treatment arm as a covariate demonstrated a significantly slower regrowth (parameter KG) and deeper SLD decrease in TPEx arm versus the control EXTREME, particularly within the first 3 months. Additionally, observations of simulated median TK suggested a benefit of TPEx versus EXTREME, observed at ~12 months, with a better safety profile [5].

Unlike several previous TK modeling studies mostly aimed to assist drug development [12, 14, 39, 40, 43], here we turned

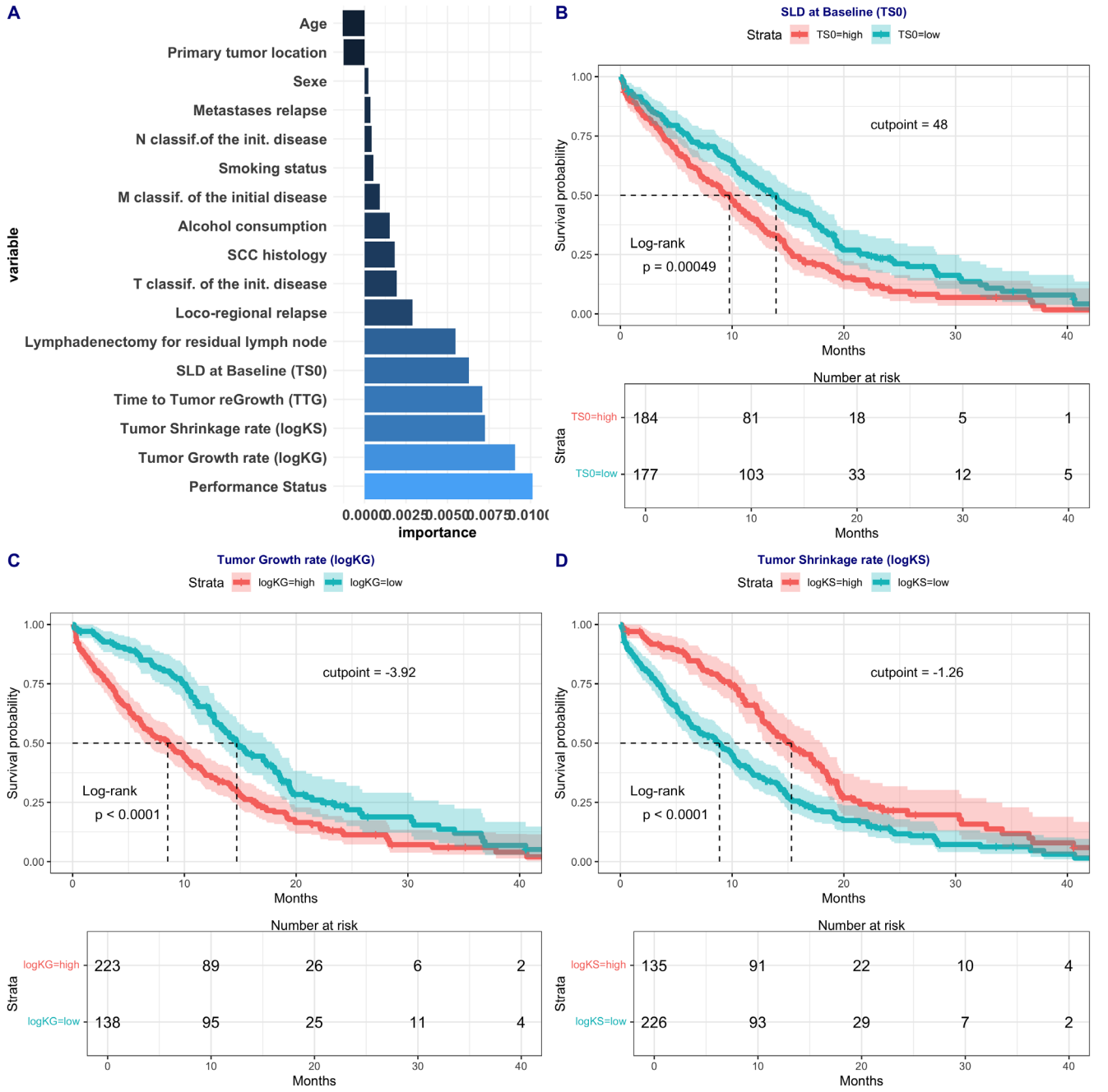


**FIGURE 4** | Features Importance and stratified survival curves for TK4-OS4. (A) Feature importance based on the random survival forest model and a permutation importance algorithm. (B) Survival curves with optimal cut-off on TS_0 (SLD at diagnosis) using the maximally selected rank statistics from the 'maxstat' R package. (C) Survival curves with optimal cut-off on logKG (tumor growth rate). (D) Survival curves with optimal cut-off on logKS (tumor shrinkage rate).

our interest on the ability of TK metrics to predict survival of patients early in the course of treatment. In such context, the value of predictive modeling does not rely on the ability to forecast study-level survival curves but rather lies in the evaluation of discrimination metrics such as the c-index. It is also crucial to avoid time-dependent covariate bias [15, 16]. One approach to do so is to rely on joint modeling [44–46]. Here, motivated by clinically relevant scenarios in routine care, we rather focused on two operational survival prediction problems. The first was to evaluate the survival predictive value of on-treatment TK metrics during the 4 cycles corresponding to TPEx chemotherapy duration (TK4-OS4). The second leveraged data from the first line of treatment plus maintenance to predict post-progression survival (TK1L-PPS). These could be of value to guide clinical decisions regarding maintenance or second-line therapy. The latter setting is of particular interest because, while clear guidelines are established from large phase 3 trials in the first line setting, the diversity of subsequent line options is wider and often left to the appreciation of the medical oncologist. Disposing of an individualized PPS predictive tool could for example help decide whether to orientate a given patient to one therapeutic option or another and thus optimize the treatment sequence. For instance, if the

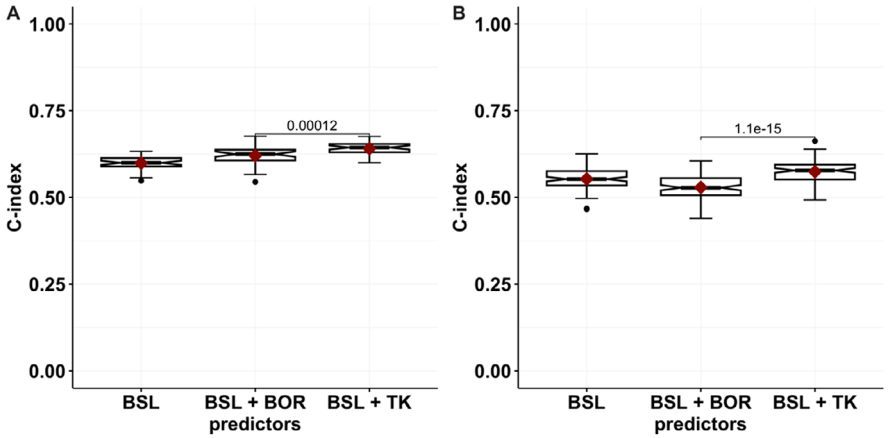


**FIGURE 5** | Comparison of different predictive models for TK4-OS4. The red diamond and middle line are the median and mean c-index, respectively. (A) Predictive models on the training set. (B) Predictive models on the test set. *p* refers to the *p*-value of the Kruskal-Wallis test.

survival chances of a patient are predicted to be very small, the clinician or the patient might decide to stop anti-cancer therapy. However, to really decipher the impact of the first line TK on the response to second-line options would require the knowledge of the treatments received post-first line. A limitation of our study is that such data was not available in the trial records. Further investigations should be conducted to address this question.

To maximize predictive performances and account for possible nonlinear interactions between predictors of survival, we benchmarked the current state-of-the art survival machine learning for the two survival prediction problems. We compared traditional (Cox proportional hazard) modeling to tree ensemble models (e.g., gradient boosting or random survival forest) and neural networks (CoxTime, DeepSurv, and DeepHit). Our results indicate that the more complex (and less interpretable) models only rarely and marginally surpassed the standard, linear (Cox) model. This contrasts with another work with more patients and variables, where a random survival forest outperformed a Cox model [17]. However, they align with other studies showing that complex ML might not overcome classical statistical models for cases relative to predictive modeling in healthcare from structured data, especially when the number of samples is small (of the order of 100–1000) [47–49]. Rather, advanced neural network engineering performs better for computer vision or natural language processing tasks [49]. Yet, the best model for TK4-OS4 was a random survival forest model. It demonstrated a modest (but significant) predictive performance of individual survival with a c-index of 0.592. It is important to note that the relatively modest c-index reflects the inherent complexity of HNSCC and the heterogeneity of the patient population. These findings underscore the necessity of applying the mechanistic learning approach to a larger and more diverse dataset, incorporating additional biomarkers.

When reducing the number of variables to only BSL + TK4, the model: (1) did not exhibit signs of overfitting (same performances in the training and test sets) but most importantly (2) indicated a substantially better prediction of OS4 than a model using BSL + BOR. This last point indicates that the current standard RECIST guidelines for evaluation and monitoring of response could be improved using dynamic TK-based metrics that account for the entire on-treatment longitudinal data.

Unfortunately, our results indicated a limited value of TKL1 model-based metrics for PPS prediction. However, several refinements could be further investigated. For example, modeling individual lesion sizes rather than the SLD unraveled substantial inter-lesion and inter-organ variability within a specific patient, and this variability was linked to response [50–52]. In addition, considering the total volume rather than one-dimensional measurements can reveal inter-lesion variability that could not be observed with diameters [53] and provide a better assessment of the treatment effect [54]. Spatiotemporal heterogeneity across metastases has also been found relevant in a large study comprising more than 11,000 lesions in 2802 colorectal cancer patients [55]. Further, modeling the kinetics of functional tumor markers could add predictive value. For instance, Schindler et al. used the per-lesion maximal standard uptake value ($SUV_{max}$) obtained from positron emission tomography and not only found significant inter-lesion variability but also that this marker outperformed tumor size for prediction of survival [50].

Beyond invasive or irradiating tumor markers, circulating markers could also enhance predictions [56–59]. We recently found improved survival predictive performances when adding four simple blood markers kinetics (BK) from routine lab tests to TK data in a large dataset of non-small cell lung cancer patients treated with an immune-checkpoint inhibitor [17, 19]. The data comprised three phase 2 trials used as the training set for model development (862 patients) and a phase 3 trial used as the test set for model evaluation (377 and 354 patients in the investigational and control (docetaxel) arms, respectively). One of the main advantages of such BKs is that they are non-invasive and can be sampled more frequently than tumor measurements. In addition, the advent of liquid biopsies provides qualitative and quantitative data on circulating free or tumor DNA that pave a promising avenue to develop more accurate predictive computational models leveraging on-treatment data [60, 61].

In conclusion, our work suggests that early analysis of first-line tumor kinetics through mechanistic learning provides valuable prognostic insights. Future studies should aim to integrate additional biomarkers and explore the utility of larger datasets to improve prediction accuracy and better guide treatment decisions.

### Author Contributions

K.A., A.A., J.G., S.S., and S.B. wrote the manuscript. K.A., S.S., and S.B. designed the research. K.A. performed the research. K.A., S.B., S.S., and A.A. analyzed the data. K.A. contributed new reagents/analytical tools.


### Acknowledgments

This work is part of the QUANTIC project funded by ITMO Cancer AVIESAN and the French Institute National du Cancer (grant #19CM148-00). This work is part of the DIGPHAT project which was supported by a grant by a grant from the French government, managed by the National Research Agency (ANR), under the France 2030 program, reference ANR-22-PESN-0017.


### Conflicts of Interest

The authors declare no conflicts of interest.

### References


1. D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis, "Head and Neck Squamous Cell Carcinoma," *Nature Reviews. Disease Primers* 6 (2020): 92.

2. S. Okano, A. Homma, N. Kiyota, et al., "Induction Chemotherapy in Locally Advanced Squamous Cell Carcinoma of the Head and Neck," *Japanese Journal of Clinical Oncology* 51 (2021): 173–179.

3. J. B. Vermorken, R. Mesia, F. Rivera, et al., "Platinum-Based Chemotherapy Plus Cetuximab in Head and Neck Cancer," *New England Journal of Medicine* 359 (2008): 1116–1127.

4. B. Burtness, K. J. Harrington, R. Greil, et al., "Pembrolizumab Alone or With Chemotherapy Versus Cetuximab With Chemotherapy for Recurrent or Metastatic Squamous Cell Carcinoma of the Head and Neck (KEYNOTE-048): A Randomised, Open-Label, Phase 3 Study," *Lancet* 394 (2019): 1915–1928.

5. J. Guigay, A. Aupérin, J. Fayette, et al., "Cetuximab, Docetaxel, and Cisplatin Versus Platinum, Fluorouracil, and Cetuximab as First-Line Treatment in Patients With Recurrent or Metastatic Head and Neck Squamous-Cell Carcinoma (GORTEC 2014-01 TPExtreme): A Multicentre, Open-Label, Randomised, Phase 2 Trial," *Lancet Oncology* 22 (2021): 463–475.

6. E. A. Eisenhauer, P. Therasse, J. Bogaerts, et al., "New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (Version 1.1)," *European Journal of Cancer* 45 (2009): 228–247.

7. Z. Zhang, C. Xie, T. Gao, Y. Yang, Y. Yang, and L. Zhao, "Identification on Surrogating Overall Survival With Progression-Free Survival of First-Line Immunochemotherapy in Advanced Esophageal Squamous Cell Carcinoma—An Exploration of Surrogate Endpoint," *BMC Cancer* 23 (2023): 145.

8. A. Shahnam, N. Hitchen, U. Nindra, et al., "Objective Response Rate and Progression-Free Survival as Surrogates for Overall Survival Treatment Effect: A Meta-Analysis Across Diverse Tumour Groups and Contemporary Therapies," *European Journal of Cancer* 198 (2024): 113503.

9. L. Belin, A. Tan, Y. De Rycke, and A. Dechartres, "Progression-Free Survival as a Surrogate for Overall Survival in Oncology Trials: A Methodological Systematic Review," *British Journal of Cancer* 122 (2020): 1707–1714.

10. E. Saad, A. Katz, K. Machado, and M. Buyse, "Post-Progression Survival (PPS) and Overall Survival (OS) According to Treatment Type in Contemporary Phase III Trials in Advanced Breast Cancer (ABC)," *Cancer Research* 69 (2009): 5116.

11. H. Imai, K. Mori, K. Wakuda, et al., "Progression-Free Survival, Post-Progression Survival, and Tumor Response as Surrogate Markers for Overall Survival in Patients With Extensive Small Cell Lung Cancer," *Annals of Thoracic Medicine* 10 (2015): 61–66.

12. L. Claret, P. Girard, P. M. Hoff, et al., "Model-Based Prediction of Phase III Overall Survival in Colorectal Cancer on the Basis of Phase II Tumor Dynamics," *Journal of Clinical Oncology* 27 (2009): 4103–4108.

13. L. Claret, M. Gupta, K. Han, et al., "Evaluation of Tumor-Size Response Metrics to Predict Overall Survival in Western and Chinese Patients With First-Line Metastatic Colorectal Cancer," *Journal of Clinical Oncology* 31 (2013): 2110–2114.

14. L. Claret, J. Y. Jin, C. Ferté, 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," *Clinical Cancer Research* 24 (2018): 3292–3298.

15. C. van Walraven, D. Davis, A. J. Forster, and G. A. Wells, "Time-Dependent Bias Was Common in Survival Analyses Published in Leading Clinical Journals," *Journal of Clinical Epidemiology* 57 (2004): 672–682.

16. H. Mistry, "Time-Dependent Bias of Tumor Growth Rate and Time to Tumor Regrowth," *CPT: Pharmacometrics & Systems Pharmacology* 5 (2016): 587.

17. S. Benzekry, M. Karlsen, C. Bigarré, et al., "Predicting Survival in Patients With Advanced NSCLC Treated With Atezolizumab Using Pre- and On-Treatment Prognostic Biomarkers," *Clinical Pharmacology & Therapeutics* 116 (2024): 1110–1120.

18. S. Benzekry, "Artificial Intelligence and Mechanistic Modeling for Clinical Decision Making in Oncology," *Clinical Pharmacology and Therapeutics* 108 (2020): 471–486.

19. S. Benzekry, M. Karlsen, A. El Kaoutari, et al., "Supporting Decision Making and Early Prediction of Survival for Oncology Drug Development Using a Pharmacometrics-Machine Learning Based Model," *PAGE* 30 (2022): 10276.

20. W. D. Stein, J. Yang, S. E. Bates, and T. Fojo, "Bevacizumab Reduces the Growth Rate Constants of Renal Carcinomas: A Novel Algorithm Suggests Early Discontinuation of Bevacizumab Resulted in a Lack of Survival Advantage," *Oncologist* 13 (2008): 1055–1062.

21. A. Samson, M. Lavielle, and F. Mentré, "Extension of the SAEM Algorithm to Left-Censored Data in Nonlinear Mixed-Effects Model: Application to HIV Dynamics Model," *Computational Statistics & Data Analysis* 51 (2006): 1562–1574.

22. H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-Event Prediction With Neural Networks and Cox Regression," *Journal of Machine Learning Research* 20 (2019): 1–30.

23. C. Lee, W. Zame, and J. Yoon, "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks," 2018.

24. J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network," *BMC Medical Research Methodology* 18 (2018): 24.

25. M. F. Gensheimer and B. Narasimhan, "A Scalable Discrete-Time Survival Model for Neural Networks," *PeerJ* 7 (2019): e6257.

26. H. Kvamme and Ø. Borgan, "Continuous and Discrete-Time Survival Prediction With Neural Networks," *Lifetime Data Analysis* 27 (2021): 710–736.

27. O. Bouaziz and G. Nuel, "L0 Regularisation for the Estimation of Piecewise Constant Hazard Rates in Survival Analysis," (2016), https://arxiv.org/abs/1609.04595.

28. S. Pölsterl, N. Navab, and A. Katouzian, "Fast Training of Support Vector Machines for Survival Analysis," in *Machine Learning and Knowledge Discovery in Databases*, vol. 9285, ed. A. Appice (Cham, Switzerland: Springer International Publishing, 2015), 243–259.

29. S. Polsterl, N. Navab, and A. Katouzian, "An Efficient Training Algorithm for Kernel Survival Support Vector Machines," 2016.

30. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA: Association for Computing Machinery, 2016), 785–794, https://doi.org/10.1145/2939672.2939785.

31. H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random Survival Forests," *Annals of Applied Statistics* 2 (2008): 841–860.

32. L. L. George, "The Statistical Analysis of Failure Time Data," *Technometrics* 45 (2003): 265–266.

33. R. Tibshirani, "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine* 16 (1997): 385–395.

34. G. Tutz and H. Binder, "Boosting Ridge Regression," *Computational Statistics & Data Analysis* 51 (2007): 6044–6059.

35. L. J. Wei, "The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis," *Statistics in Medicine* 11 (1992): 1871–1879.

36. F. E. Harrell and R. M. Califf, "Evaluating the Yield of Medical Tests," *JAMA* 247 (1982): 2543–2546.

37. R. Bruno, D. Bottino, D. P. de Alwis, et al., "Progress and Opportunities to Advance Clinical Cancer Therapeutics Using Tumor Dynamic Models," *Clinical Cancer Research* 26 (2020): 1787–1795.

38. W. D. Stein, W. D. Figg, W. Dahut, et al., "Tumor Growth Rates Derived From Data for Patients in a Clinical Trial Correlate Strongly With Patient Survival: A Novel Strategy for Evaluation of Clinical Trial Data," *Oncologist* 13 (2008): 1046–1054.

39. L. Claret, P. Girard, J. O'Shaughnessy, et al., "Model-Based Predictions of Expected Anti-Tumor Response and Survival in Phase III Studies Based on Phase II Data of an Investigational Agent," *Journal of Clinical Oncology* 24 (2006): 6025.

40. Y. Wang, C. Sung, C. Dartois, et al., "Elucidation of Relationship Between Tumor Size and Survival in Non-Small-Cell Lung Cancer Patients Can Aid Early Decision Making in Clinical Drug Development," *Clinical Pharmacology and Therapeutics* 86 (2009): 167–174.

41. L. Claret, F. Mercier, B. E. Houk, P. A. Milligan, and R. Bruno, "Modeling and Simulations Relating Overall Survival to Tumor Growth Inhibition in Renal Cell Carcinoma Patients," *Cancer Chemotherapy and Pharmacology* 76 (2015): 567–573.

42. P. Chan, X. Zhou, N. Wang, Q. Liu, R. Bruno, and J. Y. Jin, "Application of Machine Learning for Tumor Growth Inhibition - Overall Survival Modeling Platform," *CPT: Pharmacometrics & Systems Pharmacology* 10 (2021): 59–66.

43. R. Bruno, M. Marchand, K. Yoshida, 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 (2023): 1047–1055, https://doi.org/10.1158/1078-0432.CCR-22-2323.

44. J. G. Ibrahim, H. Chu, and L. M. Chen, "Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data," *JCO* 28 (2010): 2796–2801.

45. A. Król, A. Mauguen, Y. Mazroui, A. Laurent, S. Michiels, and V. Rondeau, "Tutorial in Joint Modeling and Prediction: A Statistical Software for Correlated Longitudinal Outcomes, Recurrent Events and a Terminal Event," *Journal of Statistical Software* 81 (2017): 1–52.

46. C. Tardivon, S. Desmée, M. Kerioui, et al., "Association Between Tumor Size Kinetics and Survival in Patients With Urothelial Carcinoma Treated With Atezolizumab: Implication for Patient Follow-Up," *Clinical Pharmacology and Therapeutics* 106 (2019): 810–820.

47. E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. van Calster, "A Systematic Review Shows No Performance Benefit of Machine Learning Over Logistic Regression for Clinical Prediction Models," *Journal of Clinical Epidemiology* 110 (2019): 12–22.

48. R. Knauer, M. Grimm, and E. Rodner, "PMLBmini: A Tabular Classification Benchmark Suite for Data-Scarce Applications," arXiv (2024), https://doi.org/10.48550/arXiv.2409.01635.

49. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?," in *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA: Curran Associates Inc., 2024), 507–520.

50. E. Schindler, M. A. Amantea, M. O. Karlsson, and L. E. Friberg, "PK-PD Modeling of Individual Lesion FDG-PET Response to Predict Overall Survival in Patients With Sunitinib-Treated Gastrointestinal Stromal Tumor," *CPT: Pharmacometrics & Systems Pharmacology* 5 (2016): 173–181.

51. F. Mercier, M. Kerioui, S. Desmée, J. Guedj, O. Krieter, and R. Bruno, "Longitudinal Analysis of Organ-Specific Tumor Lesion Sizes in Metastatic Colorectal Cancer Patients Receiving First Line Standard Chemotherapy in Combination With Anti-Angiogenic Treatment," *Journal of Pharmacokinetics and Pharmacodynamics* 47 (2020): 613–625.

52. M. Kerioui, S. Desmée, F. Mercier, et al., "Assessing the Impact of Organ-Specific Lesion Dynamics on Survival in Patients With Recurrent Urothelial Carcinoma Treated With Atezolizumab or Chemotherapy," *ESMO Open* 7 (2022): 100346.

53. E. Schindler, S. M. Krishnan, R. Mathijssen, A. Ruggiero, G. Schiavon, and L. E. Friberg, "Pharmacometric Modeling of Liver Metastases' Diameter, Volume, and Density and Their Relation to Clinical Outcome in Imatinib-Treated Patients With Gastrointestinal Stromal Tumors," *CPT: Pharmacometrics & Systems Pharmacology* 6 (2017): 449–457.

54. M. L. Maitland, J. Wilkerson, S. Karovic, et al., "Enhanced Detection of Treatment Effects on Metastatic Colorectal Cancer With Volumetric CT Measurements for Tumor Burden Growth Rate Evaluation," *Clinical Cancer Research* 26 (2020): 6464–6474.

55. J. Zhou, Q. Li, and Y. Cao, "Spatiotemporal Heterogeneity Across Metastases and Organ-Specific Response Informs Drug Efficacy and Patient Survival in Colorectal Cancer," *Cancer Research* 81 (2021): 2522–2533.

56. R. Almufti, M. Wilbaux, A. Oza, et al., "A Critical Review of the Analytical Approaches for Circulating Tumor Biomarker Kinetics During Treatment," *Annals of Oncology* 25 (2014): 41–56.

57. D. M. Kurtz, M. S. Esfahani, F. Scherer, et al., "Dynamic Risk Profiling Using Serial Tumor Biomarkers for Personalized Outcome Prediction," *Cell* 178 (2019): 699–713.

58. I. Netterberg, C. C. Li, L. Molinero, et al., "A PK/PDAnalysis of Circulating Biomarkers and Their Relationship to Tumor Response in Atezolizumab-Treated Non-Small Cell Lung Cancer Patients," *Clinical Pharmacology and Therapeutics* 105 (2018): 486–495.

59. I. Irurzun-Arana, E. Asín-Prieto, S. Martín-Algarra, and I. F. Trocóniz, "Predicting Circulating Biomarker Response and Its Impact on the Survival of Advanced Melanoma Patients Treated With Adjuvant Therapy," *Scientific Reports* 10 (2020): 7478.

60. K. H. Khan, D. Cunningham, B. Werner, et al., "Longitudinal Liquid Biopsy and Mathematical Modeling of Clonal Evolution Forecast Time to Treatment Failure in the PROSPECT-C Phase II Colorectal Cancer Clinical Trial," *Cancer Discovery* 8 (2018): 1270–1285.

61. L. N. Phuong, S. Salas, and S. Benzekry, "Computational Modeling Approaches for Circulating Cell-Free DNA in Oncology," 2024.

**Supporting Information**

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