Tumor kinetics modeling outperforms best overall response for prediction of survival in head and neck squamous cell carcinoma

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

Tumor kinetics modeling outperforms best overall response for prediction of survival 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.

The authors declare no conflict of interest.

Social media handles:

Authors: @SBenzekry, @KevinAtsou

Institutions : @aphm_actu, @inria_sophia, @crcm_marseille

ABSTRACT

Background

In the realm of Head and Neck Squamous Cell Carcinoma (HNSCC) treatment, accurately predicting post-progression survival (PPS) and overall survival (OS) from response metrics remains a pivotal challenge. The current mainstay for response evaluation is based on the response evaluation criteria in solid tumors (RECIST) that do not account for the entire tumor response kinetics (TK). Model-derived TK parameters have been reported to carry significant OS predictive power but a rigorous comparison of such metrics versus RECIST remains to be established, and TK-OS often suffer from time-dependent covariate bias. Moreover, the comparative efficacy of TK-OS machine learning (ML) versus conventional survival models remains to be assessed.

Methods

Data on 526 advanced HNSCC patients treated with chemotherapy and cetuximab were collected from the TPExtreme clinical trial. A double-exponential model was selected to describe both first line and maintenance treatment periods. Model-based TK parameters derived from data available after first line and maintenance (TKL1) or after 4 treatment cycles (TK4) were considered to respectively predict post-progression survival (PPS) and post-4 cycles survival (OS4). They were combined with 13 baseline (BSL) clinical parameters and integrated into nine survival ML algorithms that were benchmarked for their predictive performances. A training set and a test set with 70/30 proportions were defined for model calibration and evaluation, respectively.

Results

Using TKL1 for PPS prediction, the ML algorithms felt short the classical Cox proportional hazards model. For OS4 prediction using TK4, a random survival forest (RSF) emerged as the best model. Besides the performance status that emerged as the most important, the next four top features were TK4 metrics. Furthermore, TK4 metrics outperformed RECIST-based best overall response (BOR) (c-index of $0.55\ (0.52 - 0.57)$ vs $0.58\ (0.57 - 0.6)$ on the training set, $p < 0.0001$ and $0.5\ (0.47 - 0.52)$ vs $0.53\ (0.50 - 0.55)$ on the test set, $p < 0.0001$, BOR + BSL vs TK4 + BSL, respectively).

Conclusion

The integration of TK4 model-based parameters with a random survival algorithm provides unbiased and superior OS4 predictions than RECIST in HNSCC treated with chemotherapy.

Keywords: Survival analysis, Machine learning, Tumor kinetics, Head and neck squamous cell carcinoma

INTRODUCTION

Head and Neck Squamous Cell Carcinomas (HNSCCs) arise from the mucosal epithelium of the oral cavity, pharynx and the larynx, presenting a significant clinical challenge due to their aggressive nature and the associated morbidity. Regarding resectable HNSCCs, the standard of care is based on surgery followed by and radiotherapy +/- chemotherapy [1,2]. However, for unresectable loco-regional HNSCCs, chemo-radiotherapy is the preferred modality. For patients with local or metastatic recurrence unable to benefit from local treatments such as surgery or radiotherapy, conventional treatment includes a triad of cisplatin and 5-fluorouracil (5-FU) combined with cetuximab (EXTREME regimen) [3] or a triad of cisplatin/carboplatin and 5-fluorouracil (5-FU) combined with pembrolizumab depending on the PD-L1 status [4]. These regimens are typically dispensed over six cycles, each spanning 21 days, as a first line therapeutic approach. Moreover, a GORTEC 2014-01 TPExtreme clinical trial whose objective was to compare the efficacy and safety of the TPEx regimen with EXTREME regimen, could provide an alternative to standard of care in the first-line treatment of patients with recurrent or metastatic HNSCC, especially for those who might not be good candidates for up-front pembrolizumab treatment [5]. Despite these therapeutic strides, the grim reality of a median survival oscillating between 12 to 15 months post-treatment[5] underlines a pressing need for more precise prognostic models to guide clinical decisions and tailor therapeutic regimens.

Prediction of overall survival (OS) from early surrogate markers is of critical importance, both for clinical care and drug development (for instance to forecast the final issue of a phase 3 trial from early interim data). Current surrogate markers to predict OS are based on the response evaluation criteria in solid tumors (RECIST)[6] that establish best overall response (BOR) using the relative change to baseline of the sum of the largest diameters (SLD) of target lesions. This also allows to define progression-free survival (PFS). However, such criteria have limitations. For instance, it has been observed that OS results could be uncorrelated to PFS or best overall response (BOR)[7,8].

Furthermore, in the quest for enhanced prognostic prediction models, on-treatment longitudinal SLD measurements were used to model TK using semi-mechanistic “tumor growth inhibition” (TGI) models[9,10]. These models offer a glimpse into the dynamics of tumor response to treatment over time, thereby providing TK parameters for prognostic modeling. Multiple studies combining estimated TK parameters to survival analysis (TK–OS) revealed the importance of TK parameters in predicting OS, e.g., in metastatic colorectal cancers or non-small cell lung cancer (NSCLC)[9,11,12]. However, no direct comparison between the predictive abilities of TK parameters versus RECIST BOR has yet been performed.

In addition, a slew of biases threatens the improvements afforded by the TK parameters on predicting OS. One such bias is the time-dependent covariate bias [13,14], which can skew the derived insights [15]. Amidst these backdrops, post-progression survival (PPS) arises as an adequate prognostic outcome. The pivotal rationale favoring PPS over OS is mainly anchored in the attempt to mitigate the confounding impact that subsequent lines of therapy can have on OS [7]. Unlike PFS, PPS is strongly correlated to OS ($r = 0.97,\ p < 0.05,\ R^2 = 0.94$) [8]. For instance, in cases of NSCLC, an increment in PFS does not unequivocally translate to an enhancement in OS, however, PPS exhibits a strong association with OS following the first-line treatment [8]. Moreover, PPS has been acknowledged for its potential in alleviating the immortal time bias associated with the employment of model-derived metrics as covariates for survival analysis [15].

This is mostly because such approaches where initially developed for drug development, in which case such bias does not apply. This is mainly because the output of interest was at the study or treatment arm level, rather than the individual one. For individual survival predictions, one needs to account for the time-dependent covariate bias that arises when *on-treatment* longitudinal data (i.e., TK) is used to predict at a *pre-treatment* time point, which is what is done when the outcome is OS [15]. Such approach is mathematically ill-defined as it uses data from the future to predict in the present. In the context of TK-OS, using all the available longitudinal data for predicting OS is flawed because already the total duration of the "on-treatment" phase is a strong predictor of OS. To circumvent this, our approach entails the estimation of TK parameters from SLD measurements restricted to a specific duration and thereafter predicting the subsequent survival outcome. Specifically, we examined two well-posed operational prediction problems: 1) forecasting post-cycle 4 OS (OS4) from tumor kinetics data up to that point (TK4 – OS4) and 2) predicting post-progression survival (PPS) from first line tumor kinetics (TK1L – PPS).

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

MATERIALS AND METHODS

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 loco-regional 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 6 cycles of the standard of care (fluorouracil, cisplatin and cetuximab) (Figure 1.**A**).

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

Data

Data on 526 patients were gathered from raw patients' 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 largest diameters (SLDs) of target lesions and 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, 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, country) were also retrieved from the raw records. Over 40 baseline clinical and demographical features gathered from the raw data, 15 were retained for analysis considering their clinical relevance and their imbalance ratio (IR) (**Error! Reference source not found.**). The total fraction of missingness in the dataset was 1.24% (1.45% on the test set). Therefore, due to their small amount, missing values were discarded from the training and the test set.

Kinetics-Machine Learning

The predictive model was based on a specific two-step modeling approach called the kinetics-Machine Learning [16–18] (Figure 1.**B**). 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 [19,9,10]. 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.

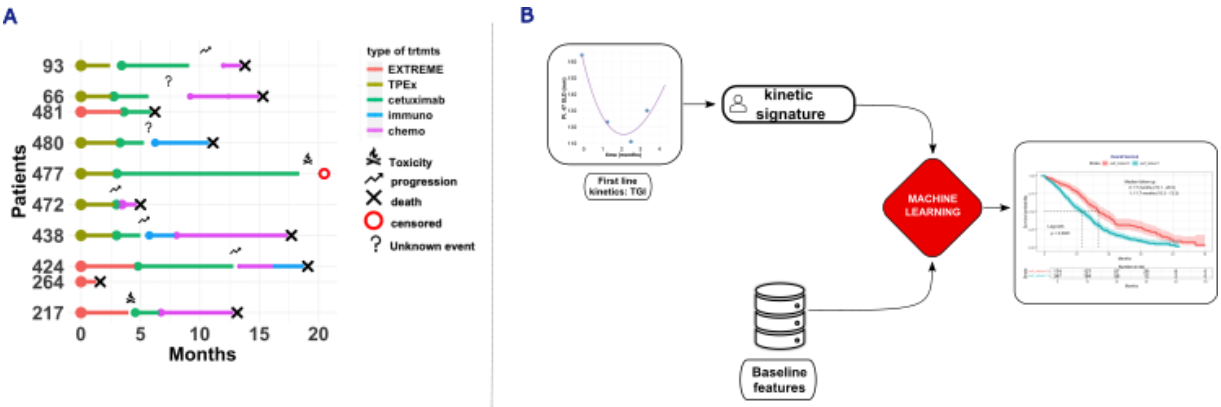

Figure 1. Treatment history and kinetics-ML prediction of survival. (A) Treatment history of a sample of patients. Cetuximab was the treatment given during maintenance. (B) Steps describing mechanistic learning. It combines modeling tumor kinetics and ML for survival prediction.

Tumor kinetics models

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 τ $(months)$, 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 $(months^{-1})$ and tumor growth is described by an exponential growth with a growth rate KG $(months^{-1})$ (*Table 1*, model 1) [19]. The two other variants of this model introduce one more parameter [9,10]. In the first one, the growth and shrinkage processes are respectively modulated by the factors α 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 λ (*Table 1*, model 3) [9,10]. 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 [10].

Table 1. Different TK models

Name	Model	Parameters	TTG[1]
Model 1	$TS(t,\Theta) = TS_0(e^{KG(t-\tau)} + e^{-KS(t-\tau)} - 1)$	TS_0, KG, KS	$\frac{\log\left(\frac{KS}{KG}\right)}{KS+KG} + \tau$
Model 2	$TS(t,\Theta) = TS_0(\alpha e^{KG(t-\tau)} + (1-\alpha)e^{-KS(t-\tau)})$	TS_0, KG, KS, α	$\frac{\log\left(\frac{KS(1-\alpha)}{\alpha KG}\right)}{KS+KG} + \tau$
Model 3	$TS(t,\Theta) = TS_0 e^{-\left(\frac{KS}{\lambda}\right)\left(1-e^{-\lambda(t-\tau)}\right)+KG(t-\tau)}$	TS_0, KG, KS, λ	$\frac{\log\left(\frac{KS}{KG}\right)}{\lambda} + \tau$

[1] Time To re-Growth

Statistical model

The framework of the Nonlinear Mixed Effects Modelling (NLME) was used as a statistical tool to describe the inter-individual variability in the observed tumor kinetics. Let $Y^i = \{y_{i1}, \cdots, y_{in^i}\}$, for $i \in \{1, \cdots, N\}$, the vector of longitudinal SLD measurements for the ith 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

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

where $TS^{(k)}(t_{ij}, \theta_i)$ 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 $\Omega^{(k)} = diag(\omega_1, \cdots, \omega_{l^{(k)}})$ (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,

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

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:

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

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.

Model calibration and selection

The parameters TS_0 and τ 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 [20]. 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 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 and the parameters were considered identifiable when their RSE did not exceed $50\ \%$.

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 were extracted. These parameters were then

utilized to standardize the test set. 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.

Benchmarking procedure

Eight predictive ML models – CoxTime [21], DeepHit [22], DeepSurv [23], LogHaz [24,25], PC-Hazard [25], Survival SVM [26,27], XGBoost and Random Survival Forest (RSF) [28] – and five classical parametric and semi-parametric survival models – Cox Proportional Hazard (PH), Accelerated Failure Time (AFT), Cox Lasso, CoxBoost and GLMBoost [29] and [30] – were assessed for their ability to predict PPS. The aforementioned models were compared by computing Harrell's Concordance index (c-index) [31] using nested and repeated k-fold cross-validation. Each inner loop contained 30 repeats of 2-fold cross-validations, used to tune the ML models. Each outer loop consisted of 10 repeats of 5-fold 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 $mlr3$ 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 were 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.

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 100 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.

Comparison of different predictor models

The best model selected from the benchmarking step was used to assess the predictive quality of different set of predictors: baseline (BSL) features, BSL + BOR, BSL + TK and TK. These models were assessed on the training set using 10 repeats of 5-fold cross validations. The generalization quality of the models was assessed on the test set using 100 bootstrap sets. Furthermore, a Kruskal-Wallis test was used to compare the c-index of the models.

Different Unbiased Two-step modeling approaches: TK4-OS4 and TKL1-PPS

To address the time-dependent covariate biases, one cannot rely on the TKL1 parameters to predict OS. Therefore, different unbiased two-step modeling approaches were investigated. On one hand, TK4-OS4, uses 4 cycles SLD data to estimate TK parameters (TK4) and predicts post cycle 4 OS. On the other hand, TKL1-PPS, uses TK parameters estimated from longitudinal SLD measurements restricted to first line + maintenance and predicts PPS.

RESULTS

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. 60.1% of the patients received a second line treatment while 33.2% got a third line treatment and finally 13.3% received a fourth line treatment. 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 1.A**. 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.

Tumor kinetics modeling

The SLD measurements on 368 (70%) patients – restricted to 4 cycles (total number of time points = 672 and median number of time points per patient = 2) – were used to calibrate the different TK models. All tested models showed a good agreement with the data. However, according to the corrected Bayesian Information Criterion (BICc), 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 $8.73\ \%$ (Table S2)

Diagnostic plots did not reveal any misspecification of the model at the individual level (Figure 2**. A-B**). Recalibration of the model using the treatment arm as a covariate to the kinetic parameters (KS, KG) revealed a discernible effect of the TPEx treatment on the tumor growth rate (KG) in comparison to its effect on the tumor shrinkage rate (KS). A Wald test demonstrated a statistically significant association between the treatment arm and KG, ($p = 6.35 \times 10^{-3}$). No association between the treatment arm and KS was found. Additionally, simulations conducted using the estimated population parameters suggested a slightly longer mean time to progression in the TPEx group compared to the EXTREME group (Figure 2**. C**). However, adding any covariate (including treatment arm) to the model resulted in a decreased goodness-of-fit and triggered identifiability issues. Therefore, a model without covariates was kept for the predictive analysis, on the entire patient cohort (i.e., both treatment arms).

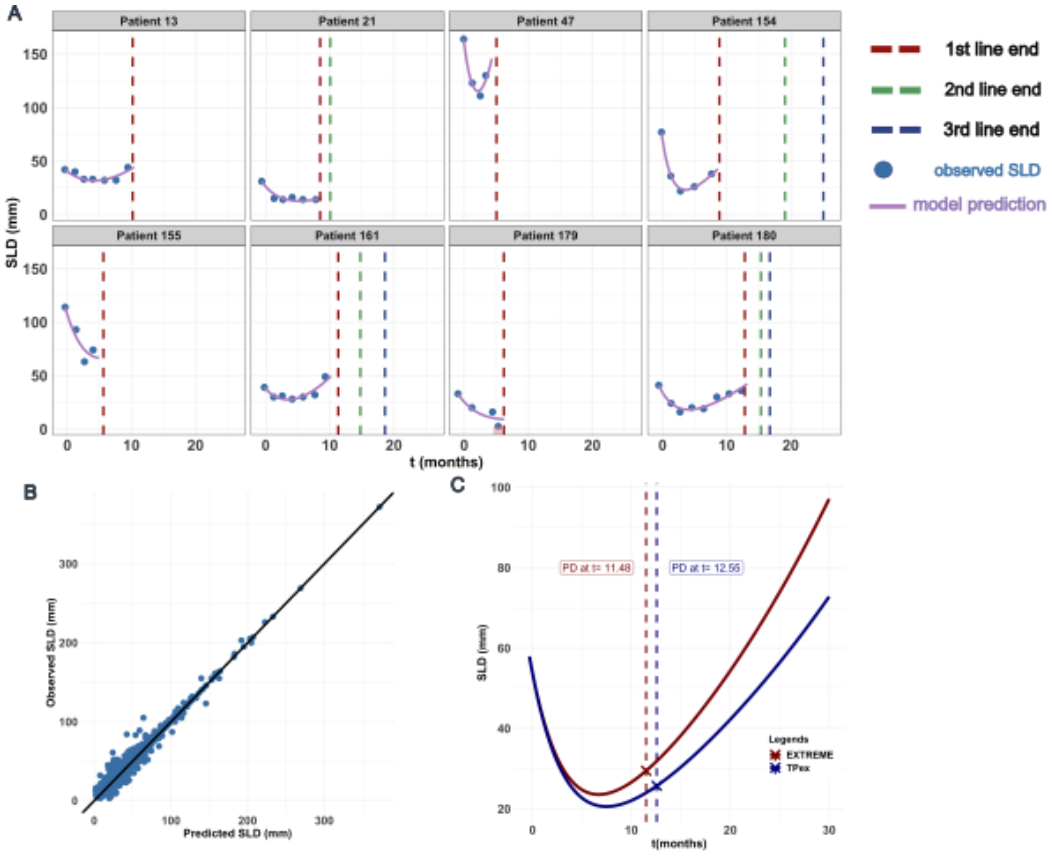

Figure 2. Diagnostic plots and treatment arms simulation. (**A**) Representative individual fits. The purple line represents the model prediction and the blue dots the observed SLDs. The red, green and dark-blue dotted vertical lines represent the end of the first, second line and third lines, respectively. (**B**) Observed SLDs vs predicted SLDs (**C**) Simulations of tumor kinetics of each treatment arm using the typical population parameters (fixed-effects). SLD: sum of largest diameters

ML for survival analysis

ML models benchmarking: unbiased predictions using TK4–OS4 et TK1L–PPS.

To avoid time-dependent covariate bias, we considered either predicting post-cycle 4 survival (OS4) from data restricted to the first 4 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–PPS, 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–PPS, 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**).

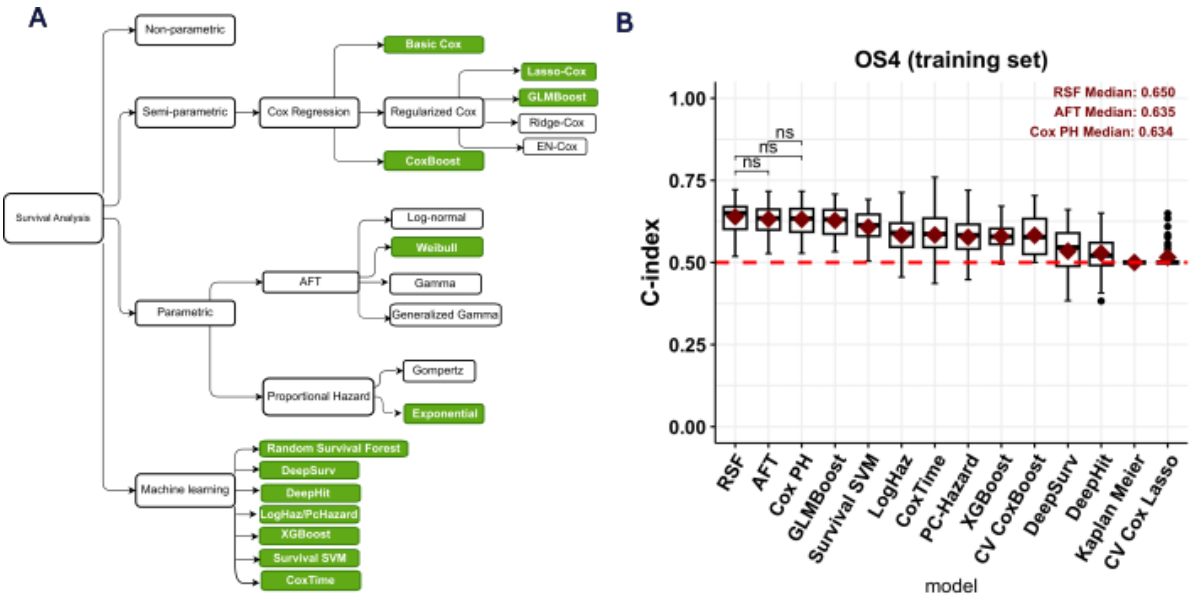

Figure 3. Benchmark of ML methods, TK4 – OS4. (**A**) Graphical representation of the different ML and statistical methods. The green boxes are the tested models. (**B**) Boxplots of the results of the benchmarking process on OS4 prediction. The red Diamond is the median c-index and the middle line the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

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 4**. **A**). In addition, utilizing the maximally selected rank statistic, TS_0, $\log KG$ and $logKS$ provided a significant separation of the survival curves (log-rank $p < 0.01$ and $p < 0.05$ respectively) (Figure 4. **B-D**).

However, for TKL1-PPS, the top-3 parameters were loco-regional relapse, followed by TS_0 , the performance status (**Figure S3.A**). Surprisingly, Unlike TK4 – OS4, $logKS$, triggered crossed survival curves **(Figure S3.D**) potentially indicating a time-dependent nature of the growth rate parameter.

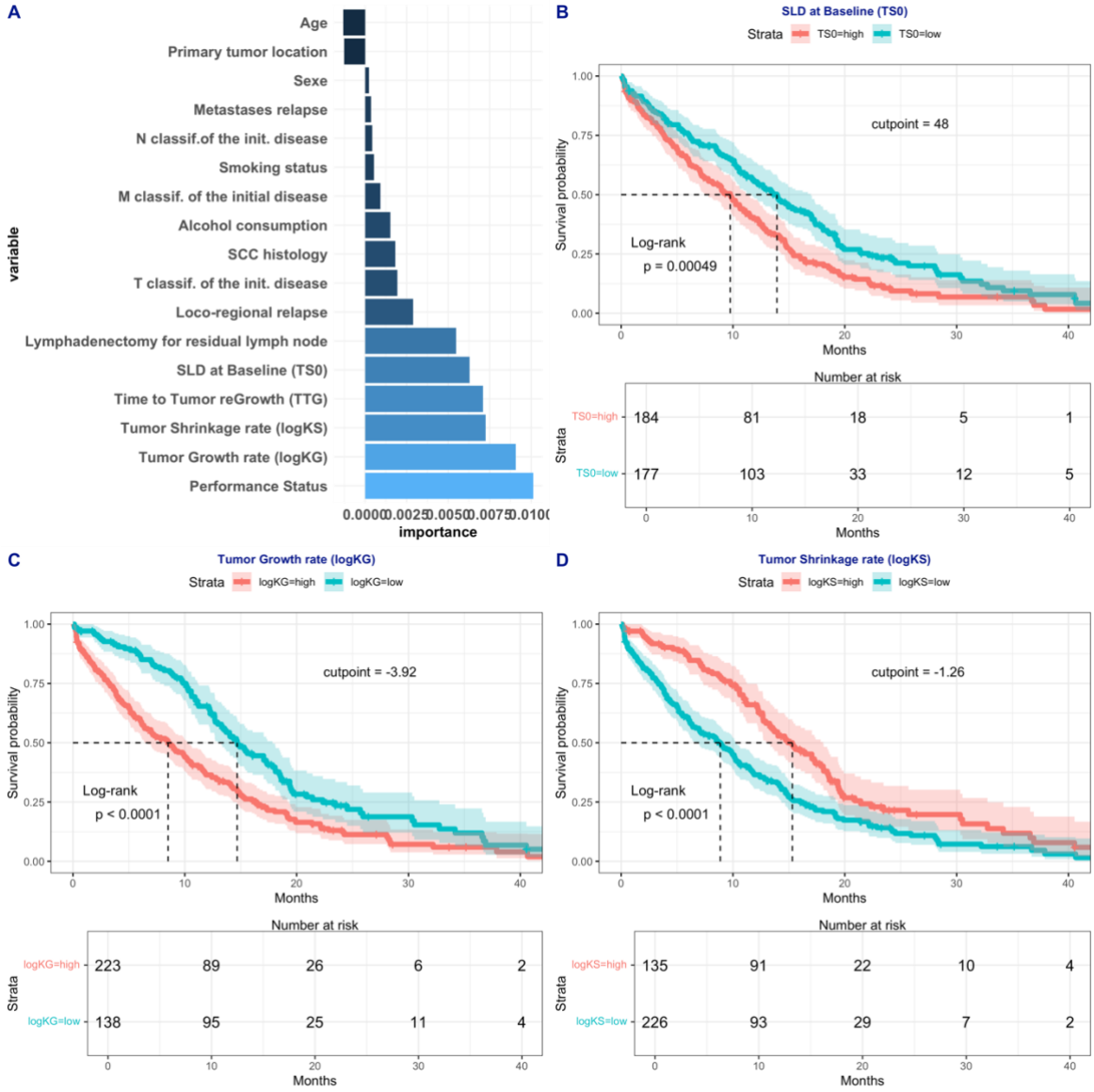

Figure 4. Features Importance and stratified survival curves for TK4-OS4. **(A)** Feature importance based on the random survival forest model and a permutation importance algorithm. **(B)** Survival curves with optimal cut-off on TS_0 (SLD at diagnosis) using the maximally selected rank statistics from 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).

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 S1**) alone, BSL + BOR, BSL + TK4 and TK4 (Figure 5). In the training set, the BSL + TK4 outperformed BOR + BSL $(0.63\ (95\%\ CI\colon 0.61 - 0.65)$ versus 0.612

$(95\%\ CI\colon 0.58-0.64)$, respectively, $p < 0.001$, Kruskal-Wallis, **Figure 5.A**). This trend persisted in the test dataset, with respective c-index of $0.63\ (95\%\ CI\colon 0.61-0.65)$ and $0.52\ (95\%\ CI\colon 0.50-0.54)$ ($p < 0.0001$, **Figure 5.B**). Intriguingly, the integration of BSL with BOR adversely impacted the model accuracy on test set as compared to BSL alone (**Figure 5.B**). However, this was not observed in TKL1-PPS, such behaviour was absent (**Figure S4.A-B**).

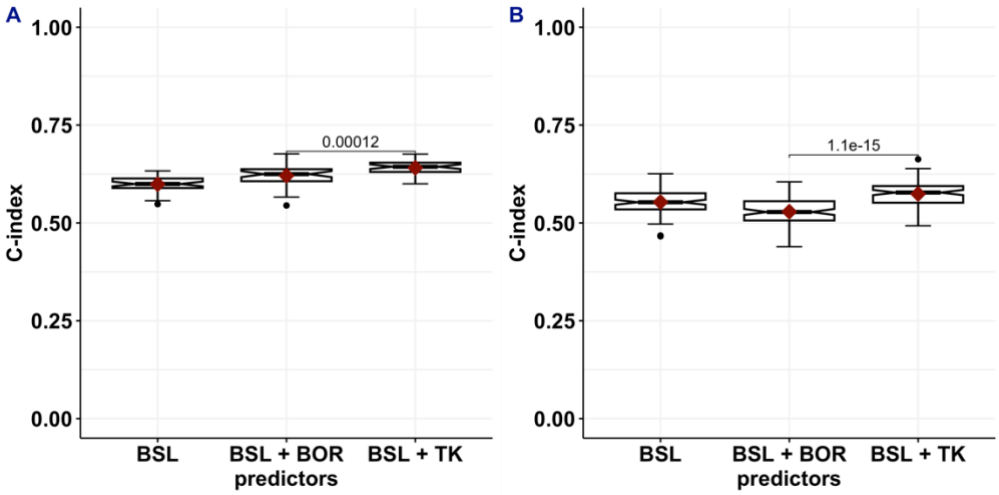

Figure 5. Comparison of different predictive models for TK4-OS4.

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.

DISCUSSION

Tumor kinetics modeling to predict overall survival (TK–OS) has been extensively developed in the last 15 years [32]. Multiple cancer types, often in the advanced or metastatic stage have been studied, including prostate [33], breast [34], colorectal [9], non-small cell lung cancer [35] or renal cell carcinoma [36]. 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 [37].

As often observed for the other cancer types, on-treatment TK could be accurately described by a simple pattern mathematically: the sum of two exponentials. A decreasing phase governed by a shrinkage parameter KS and a regrowth phase led by a parameter KG. Using the treatment arm as a covariate showed a significantly slower regrowth (parameter KG) and deeper SLD decrease in the investigational (TPEx) arm versus the control (EXTREME). However, even though the decrease in SLD is deeper in the TPEx arm, this decrease is not necessarily faster in this arm compared to the EXTREME arm, at least within the first 3 months.

This result thus prompts discussion on the necessity of undergoing TPEx for patients who present with a high tumor burden or symptoms that require a rapid response [5]. 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 [9,12,34,35,38], here we turned our interest on the ability of TK metrics to predict earlier individual survival. 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 [39,40]. One approach to do so is to rely on joint modeling [41–43]. 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 four 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 lines 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.

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 models in our setting. We therefore compared traditional (Cox proportional hazard) modeling to tree ensemble models (e.g., gradient boosting or random survival forest) and neural networks (CoxTime, DeepSurv, DeepHit). Our results indicate that the more complex (and less interpretable) models only rarely and marginally surpassed the linear standard (Cox) model. These findings 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 [44]. Rather, advanced neural network engineering performs better for computer vision or natural language processing tasks.

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. 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. Indeed, RECIST is only based on the mere static difference between the current time point and the baseline or nadir.

Unfortunately, our results indicated a limited value of TKL1 model-based metrics for PPS prediction. However, several refinements could be further investigated to improve. For example, modeling individual lesion sizes rather than the SLD unraveled substantial inter-lesion and inter-organ variability within a specific patient, and that this variability was linked to response [45–47]. In addition, considering the total volume rather one-dimensional measurements can reveal inter-lesion variability that could not be observed with diameters [48] and provide better assessment of the treatment effect [49]. Spatiotemporal heterogeneity across metastases has also been found relevant in a large study comprising more than 11,000 lesions in 2,802 colorectal cancer patients [50]. 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 [45].

Beyond invasive or irradiating tumor markers, circulating markers could also enhance predictions [51–54]. In a study modeling 4 simple blood markers kinetics (BK) from routine lab tests in addition to TK in a large dataset (862 patients in the training set for model development, 553 in the test set for model evaluation), and combining mixed-effects modeling with ML, we

recently found significant added OS predictive value of BK parameters compared with baseline + TK parameters [16,18]. One of the main advantages of such markers 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 [55,56].

In conclusion, our work suggests that early analysis of first-line tumor kinetics by ML modeling, likely reflecting tumor biology, could eventually enable the individual assessment of post-treatment survival. This could also serve as a tool to guide the selection of different therapeutic options and thus aid in studying various treatment sequences.

ACKNOWLEDGEMENTS

This work is part of the QUANTIC project funded by ITMO Cancer AVIESAN and the French Institute National du Cancer (grant #19CM148-00)

REFERENCES

1. Johnson, D. E. *et al.* Head and neck squamous cell carcinoma. *Nat Rev Dis Primers* **6**, 92 (2020).
2. Okano, S. *et al.* Induction chemotherapy in locally advanced squamous cell carcinoma of the head and neck. *Japanese Journal of Clinical Oncology* **51**, 173–179 (2021).
3. Vermorken, J. B. *et al.* Platinum-Based Chemotherapy plus Cetuximab in Head and Neck Cancer. *N Engl J Med* **359**, 1116–1127 (2008).
4. Burtness, B. *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. *The Lancet* **394**, 1915–1928 (2019).
5. Guigay, J. *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. *The Lancet Oncology* **22**, 463–475 (2021).
6. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* **45**, 228–247 (2009).
7. Saad, E., Katz, A., Machado, K. & Buyse, M. 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**, 5116–5116 (2009).
8. Imai, H. *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).
9. Claret, L. *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**, 4103–4108 (2009).

10. Claret, L. *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**, 2110–2114 (2013).
11. Claret, L. *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**, 2110–2114 (2013).
12. Claret, L. *et al.* A Model of Overall Survival Predicts Treatment Outcomes with Atezolizumab versus Chemotherapy in Non–Small Cell Lung Cancer Based on Early Tumor Kinetics. *Clinical Cancer Research* **24**, 3292–3298 (2018).
13. Beyersmann, J., Gastmeier, P., Wolkewitz, M. & Schumacher, M. An easy mathematical proof showed that time-dependent bias inevitably leads to biased effect estimation. *J Clin Epidemiol* **61**, 1216–1221 (2008).
14. Beyersmann, J., Wolkewitz, M. & Schumacher, M. The impact of time-dependent bias in proportional hazards modelling. *Stat Med* **27**, 6439–6454 (2008).
15. Mistry, H. On the relationship between tumour growth rate and survival in non-small cell lung cancer. (2017).
16. Benzekry, S. *et al.* Predicting survival and trial outcome in non-small cell lung cancer integrating tumor and blood markers kinetics with machine learning. 2023.09.26.23296135 Preprint at https://doi.org/10.1101/2023.09.26.23296135 (2024).
17. Benzekry, S. Artificial Intelligence and Mechanistic Modeling for Clinical Decision Making in Oncology. *Clin. Pharmacol. Ther.* **108**, 471–486 (2020).
18. Benzekry, S. *et al.* Supporting decision making and early prediction of survival for oncology drug development using a pharmacometrics-machine learning based model. in *PAGE* vol. 30 10276 (2022).
19. Stein, W. D., Yang, J., Bates, S. E. & Fojo, T. Bevacizumab Reduces the Growth Rate Constants of Renal Carcinomas: A Novel Algorithm Suggests Early Discontinuation of Bevacizumab Resulted in a Lack of Survival Advantage. *The Oncologist* **13**, 1055–1062 (2008).

20. Samson, A., Lavielle, M. & Mentré, F. Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model. *Computational Statistics & Data Analysis* **51**, 1562–1574 (2006).
21. Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-Event Prediction with Neural Networks and Cox Regression. Preprint at (2019).
22. Lee, C., Zame, W. & Yoon, J. DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. 8.
23. Katzman, J. L. *et al.* DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* **18**, 24 (2018).
24. Gensheimer, M. F. & Narasimhan, B. A scalable discrete-time survival model for neural networks. *PeerJ* **7**, e6257 (2019).
25. Kvamme, H. & Borgan, Ø. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis* **27**, 710–736 (2021).
26. Pölsterl, S., Navab, N. & Katouzian, A. Fast Training of Support Vector Machines for Survival Analysis. in *Machine Learning and Knowledge Discovery in Databases* (eds. Appice, A. et al.) vol. 9285 243–259 (Springer International Publishing, Cham, 2015).
27. Polsterl, S., Navab, N. & Katouzian, A. An Efficient Training Algorithm for Kernel Survival Support Vector Machines. 13.
28. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics* **2**, (2008).
29. Tutz, G. & Binder, H. Boosting ridge regression. *Computational Statistics & Data Analysis* **51**, 6044–6059 (2007).
30. Wei, L. J. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine* **11**, 1871–1879 (1992).
31. Harrell, F. E. & Califf, R. M. Evaluating the Yield of Medical Tests. 4.
32. Bruno, R. *et al.* Progress and Opportunities to Advance Clinical Cancer Therapeutics Using Tumor Dynamic Models. *Clinical Cancer Research* **26**, 1787–1795 (2020).

33. Stein, W. D. *et al.* Tumor growth rates derived from data for patients in a clinical trial correlate strongly with patient survival: a novel strategy for evaluation of clinical trial data. *The Oncologist* **13**, 1046–1054 (2008).
34. Claret, L. *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. *J Clin Oncol* (2006) doi:10.1200/jco.2006.24.18_suppl.6025;issue:issue:10.1200/jco.2006.24.issue-18_suppl.
35. Wang, Y. *et al.* Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development. *Clin Pharmacol Ther* **86**, 167–174 (2009).
36. Claret, L., Mercier, F., Houk, B. E., Milligan, P. A. & Bruno, R. Modeling and simulations relating overall survival to tumor growth inhibition in renal cell carcinoma patients. *Cancer Chemother Pharmacol* **76**, 567–573 (2015).
37. Chan, P. *et al.* Application of Machine Learning for Tumor Growth Inhibition - Overall Survival Modeling Platform. *CPT Pharmacometrics Syst Pharmacol* **10**, 59–66 (2021).
38. Bruno, R. *et al.* Tumor Dynamic Model-Based Decision Support for Phase Ib/II Combination Studies: A Retrospective Assessment Based on Resampling of the Phase III Study IMpower150. *Clinical Cancer Research* OF1–OF9 (2023) doi:10.1158/1078-0432.CCR-22-2323.
39. Mistry, H. Time-Dependent bias of tumor growth rate and time to tumor regrowth. *CPT: Pharmacometrics & Systems Pharmacology* **5**, 587–587 (2016).
40. van Walraven, C., Davis, D., Forster, A. J. & Wells, G. A. Time-dependent bias was common in survival analyses published in leading clinical journals. *J Clin Epidemiol* **57**, 672–682 (2004).
41. Ibrahim, J. G., Chu, H. & Chen, L. M. Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *JCO* **28**, 2796–2801 (2010).
42. Król, A. *et al.* Tutorial in Joint Modeling and Prediction: A Statistical Software for Correlated Longitudinal Outcomes, Recurrent Events and a Terminal Event. *Journal of Statistical Software* **81**, 1–52 (2017).

43. Tardivon, C. *et al.* Association Between Tumor Size Kinetics and Survival in Patients With Urothelial Carcinoma Treated With Atezolizumab: Implication for Patient Follow-Up. *Clin Pharmacol Ther* **106**, 810–820 (2019).
44. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* **110**, 12–22 (2019).
45. Schindler, E., Amantea, M. A., Karlsson, M. O. & Friberg, L. E. 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**, 173–181 (2016).
46. Mercier, F. *et al.* Longitudinal analysis of organ-specific tumor lesion sizes in metastatic colorectal cancer patients receiving first line standard chemotherapy in combination with anti-angiogenic treatment. *J Pharmacokinet Pharmacodyn* **47**, 613–625 (2020).
47. Kerioui, M. *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**, 100346 (2022).
48. Schindler, E. *et al.* 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**, 449–457 (2017).
49. Maitland, M. L. *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**, 6464–6474 (2020).
50. Zhou, J., Li, Q. & Cao, Y. Spatiotemporal Heterogeneity across Metastases and Organ-Specific Response Informs Drug Efficacy and Patient Survival in Colorectal Cancer. *Cancer Research* **81**, 2522–2533 (2021).
51. Almufti, R. *et al.* A critical review of the analytical approaches for circulating tumor biomarker kinetics during treatment. *Ann Oncol* **25**, 41–56 (2014).

52. Kurtz, D. M. *et al.* Dynamic Risk Profiling Using Serial Tumor Biomarkers for Personalized Outcome Prediction. *Cell* **178**, 699-713.e19 (2019).
53. Netterberg, I. *et al.* A PK/PDAnalysis of Circulating Biomarkers and Their Relationship to Tumor Response in Atezolizumab-Treated non-small Cell Lung Cancer Patients. *Clin Pharmacol Ther* **105**, 486–495 (2018).
54. Irurzun-Arana, I., Asín-Prieto, E., Martín-Algarra, S. & Trocóniz, I. F. Predicting circulating biomarker response and its impact on the survival of advanced melanoma patients treated with adjuvant therapy. *Sci Rep* **10**, 7478 (2020).
55. Khan, K. H. *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**, 1270–1285 (2018).
56. Phuong, L. N., Salas, S. & Benzekry, S. Computational modeling approaches for circulating cell-free DNA in oncology. (2024).

SUPPLEMENTARY

Table S1 Goodness-of-fit of the different TGI models on TK4

Models	AIC	BIC	BICc	OFV
Model 1	**11896.57**	**11916.11**	**11920.61**	**11886.57**
Model 2	12337.5	12364.86	12370.86	12323.5
Model 3	11919.28	11946.64	11952.64	11905.28

Table S2 Population parameter estimates

Parameter	Value	S.E.	R.S.E. (%)
KS_{pop}	0.27	0.0163	6.03
KG_{pop}	0.0309	0.0027	8.73
ω_{KS}	0.821	0.0544	6.63
ω_{KG}	1	0.0629	6.26
a	7.43	0.303	4.08

Table S3 Patients and disease characteristics

	n (%) or mean ± SD (min-max)
Sex	
Female	66 (12.54%)
Male	460 (87.46%)
Age (years)	$59.6 \pm 6.5\ (23.3 - 71.4)$
Alcohol consumption	
No	133 (25.29%)
Yes (past)	206 (39.16%)
Yes (currently)	182 (34.6%)
Tobacco status	
No	41 (7.8%)
Yes (past)	327 (62.17%)
Yes (currently)	158 (30.04%)
Primary tumor location	
Oropharynx	235 (44.68%)
Others	291 (55.32%)
Disease evolution at inclusion	
Metastasis	337 (64.07%)
Locoregional relapse	305 (57.98%)
Metastasis and locoregional relapse	116 (22.05%)

Previous cancer treatment	
No	90 (17.12%)
Yes	435 (82.7%)
Further treatment lines	
Line 2	316 (60.07 %)
Line 3	175 (33.27 %)
Line 4	70 (13.31%)
ECOG Performance status	
1	351 (66.7 %)
0	175 (33.3 %)
lymphadenectomy for residual lymph node	
yes	378 (71.8 %)
no	146 (27.8 %)
TNM classification at inclusion	
T, size of the primary tumor	
T0	2 (0.380 %)
T1	56 (10.6 %)
T2	154 (29.3 %)
T3	143 (27.2 %)
T4	151 (28.7 %)
Tx	14 (2.66 %)
N, extent to regional lymph nodes	
N0	129 (24.5 %)
N1	69 (13.1 %)
N2	260 (49.4 %)
N3	48 (9.13 %)
Nx	14 (2.66 %)
M, presence of distant metastasis	
M0	395 (75.1 %)
M1	96 (18.3 %)
Mx	28 (5.32 %)
Squamous cell carcinoma (SCC) histology	
moderately differentiated	211 (40.1%)
poorly or undifferentiated	106 (20.2 %)
Well-differentiated	154 (29.3 %)

Baseline features (BSL)	- TS_0 - Age - Sex - Smoking status - Alcohol consumption - Primary tumor location - Metastases relapse - TNM classification at inclusion - SCC histology - Loco-regional relapse - Lymphadenectomy of the residual lymph node - Performance status
Tumor Kinetic parameters (TK)	- $logKG$ - $logKS$ - TTG

Table S4 BSL and TK features

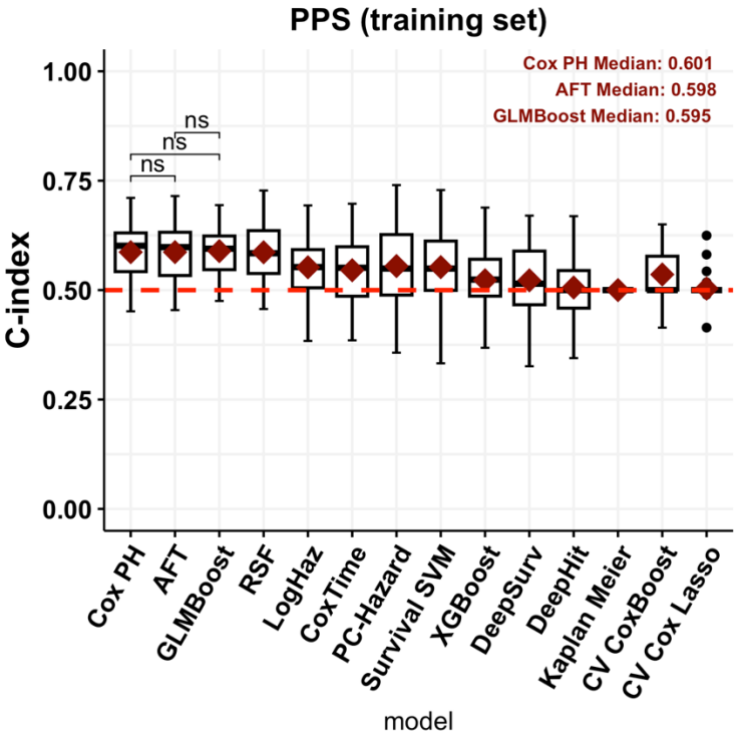

Figure S1 Benchmark of ML methods on the training set, TKL1-PPS model. Boxplots of the results of the benchmarking process on PPS prediction. The red Diamond is the median c-index and the middle line the mean c-index.

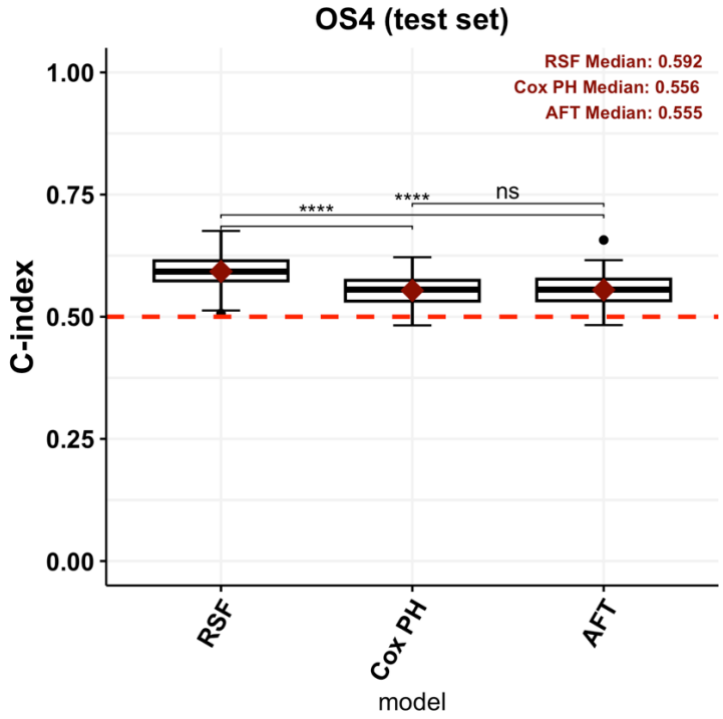

Figure S2 Best ML models on the test set, TK4 - OS4

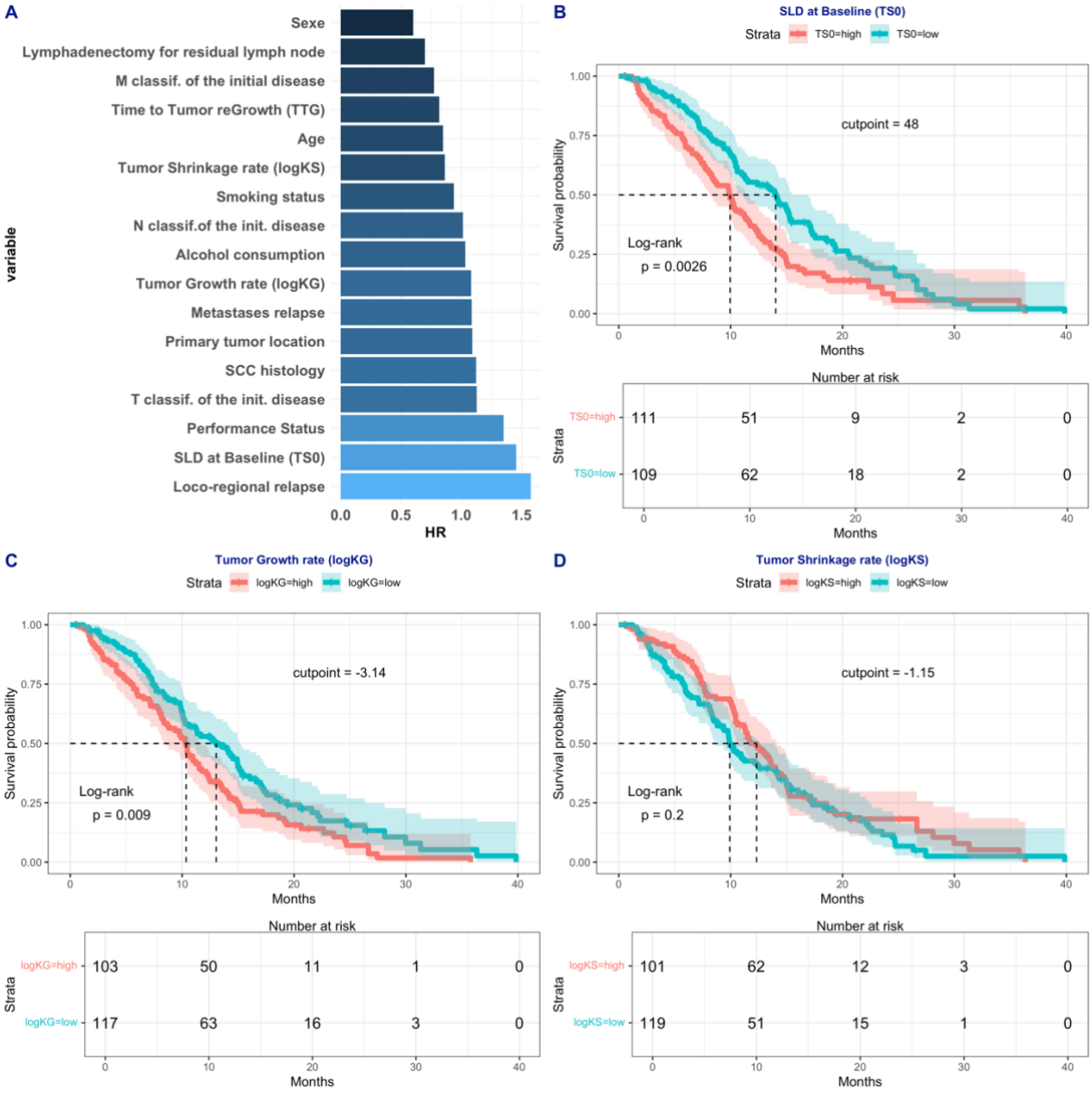

Figure S3 Parameters Importance and survival curves, TKL1-PPS model. (A) Feature importance based on the random survival forest model and a permutation importance algorithm. (B) Survival curves with optimal cut-off on TS_0 (SLD at diagnosis) using the maximally selected rank statistics from the 'maxstat' R package. (C) Survival curves with optimal cut-off on KG (tumor growth rate). (D) Survival curves with optimal cut-off on KS (tumor shrinkage rate)

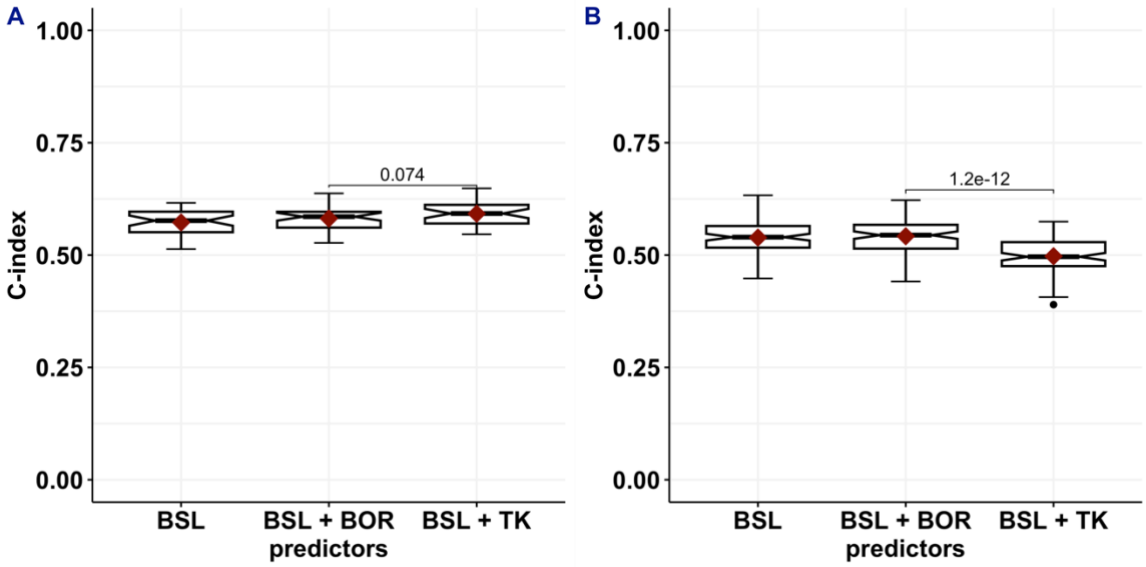

Figure S4 Comparison of different predictors models, TKL1-PPS model. The red Diamond is the median c-index and the middle line the mean c-index. p refers to the p-value of the Kruskal-Wallis test (A) Boxplots of the results of the comparison of the predictor models on the train set. (B) Boxplots of the results of the comparison of the predictor models on the test set.