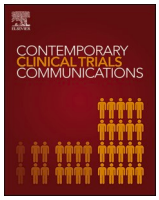
Contemporary Clinical Trials Communications 23 (2021) 100827

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc



Modeling tumor measurement data to predict overall survival (OS) in cancer clinical trials

Fang-Shu Ou [a,1,*], Jun Tang [b,1], Ming-Wen An [c], Sumithra J. Mandrekar [a]

[a] Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA
[b] Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA, USA
[c] Department of Mathematics and Statistics, Vassar College, Poughkeepsie, NY, USA

ARTICLE INFO

Keywords:
Tumor measurement data
RECIST
Prediction
Cancer trial

ABSTRACT

Introduction: Longitudinal tumor measurements (TM) are commonly recorded in cancer clinical trials of solid tumors. To define patient response to treatment, the Response Evaluation Criteria in Solid Tumors (RECIST) categorizes the otherwise continuous measurements, which results in substantial information loss. We investigated two modeling approaches to incorporate all available cycle-by-cycle (continuous) TM to predict overall survival (OS) and compare the predictive accuracy of these two approaches to RECIST.

Material and methods: Joint modeling (JM) for longitudinal TM and OS and two-stage modeling with potential time-varying coefficients were utilized to predict OS using data from three trials with cycle-by-cycle TM. The JM approach incorporates TM data collected throughout the course of the clinical trial. The two-stage modeling approach incorporates information from early assessments (before 12 weeks) to predict subsequent OS outcome. The predictive accuracy was quantified by c-indices.

Results: Data from 577, 337, and 126 patients were included for the analysis (from two stage IV colorectal cancer trials (N9741, N9841) and an advanced non-small cell lung cancer trial (N0026), respectively). Both the JM and two-stage modeling reached a similar conclusion, i.e. the baseline covariates (age, gender, and race) were mostly not predictive of OS (p-value > 0.05). Quantities derived from TM were strong predictors of OS in the two colorectal cancer trials ($p < 0.001$ for both association in JM and two-stage modeling parameters); but less so in the lung cancer trial ($p = 0.053$ for association in JM and $p = 0.024$ and 0.160 for two-stage modeling parameters). The c-indices from the two-stage modeling were higher than those from a model using RECIST (range: 0.611–0.633 versus 0.586–0.590). The dynamic c-indices from the JM were in the range of 0.627–0.683 indicating good predictive accuracy.

Conclusion: Both modeling approaches provide highly interpretable and clinical meaningful results; the improved predictive performance compared with RECIST indicates the possibility of deriving better trial endpoints from these approaches.

1. Introduction

The Response Evaluation Criteria in Solid Tumors (RECIST), first defined in 2000 [1] and subsequently updated in 2009 (RECIST 1.1) [2], uses measurement-based tumor response to define patient response to treatment and has been widely utilized in cancer clinical trials. Objective response rate and complete response, endpoints which are defined based on RECIST, are suitable for regulatory approval of new therapeutics by both the Food and Drug Administration (FDA) [3] and European Medicines Agency (EMA) [4]. Even though RECIST-based tumor response has been established as a convincing measure of anti-tumor activity, multiple studies have suggested that it may not be the best predictor of overall survival (OS) [5–10]. Research efforts have been devoted to finding alternative endpoints [7,9,11,12] and to investigating other features that may impact the prediction of OS, e.g. missing data [13] and mixed responses [14,15].

It is understandable that categorizing otherwise continuous tumor measurements, e.g. as is done in the binary RECIST-based tumor response, may result in substantial information loss. To address this, continuous endpoints have been developed to utilize the absolute/

* Corresponding author. Department of Quantitative Health Sciences, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA.
E-mail address: ou.fang-shu@mayo.edu (F.-S. Ou).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.conctc.2021.100827
Received 17 November 2020; Received in revised form 7 June 2021; Accepted 26 July 2021
Available online 9 August 2021

relative slope between baseline and 6- and 12-week assessments. Unfortunately, no improvements have been observed in either predictive ability (i.e. c-index, Brier score, and Hosmer-Lemeshow statistic) or clinical utility (e.g. positive and negative predictive values) compared to the RECIST-based binary response metric [10,12]. There are a couple of possible reasons behind the inability of improving prediction using continuous tumor measurements. The tumor measurements are taken by imaging at scheduled assessment timepoints which may not be sufficiently frequent to capture the continuous tumor size changes. Multiple lines of therapies are available for most advanced cancers which weaken the association of early tumor measurements to downstream OS.

We aim to investigate other flexible statistical models which either use all the tumor measurements as a longitudinal biomarker or early tumor measurements-based metrics while relaxing the proportional hazards assumption. This analysis explores the use of two different analysis approaches, namely joint modeling [16] and two-stage modeling with potential time-varying coefficients [17], applied to tumor measurements data to predict OS. The strength of the two-stage modeling approach includes the ease of obtaining metrics from the first stage modeling, which involves fitting a simple linear regression model on individual tumor measurements collected prior to a pre-defined landmark time point, i.e. 12 weeks, for each patient. The first-stage model estimates, which represent tumor measurement-based metrics and subsequently enter the second-stage as covariates, have natural and clinically relevant interpretations, i.e. the average tumor size at baseline and the trajectory of tumor growth within the first 12 weeks. Incorporating time-varying coefficients further improves the predictive ability and is easily achieved given modern statistical software. The strength of the joint modeling approach is that it takes into account all tumor measurements throughout the course of a clinical trial and can be used to perform risk prediction for individual patients. Both models circumvent the missing tumor measurement issue by using all recorded data instead of limiting to the complete cases or imputing missing values, as was done in previous analyses [9,10,12,13]. Both models can also handle tumor measurements taken at varying assessment times allowing for flexibility in protocol adherence, which make the methods more applicable to real life data.

2. Patients and methods

2.1. Analysis population

This analysis included patients enrolled in one of three Alliance/NCCTG cancer clinical trials: a phase III randomized study between May 1999 and April 2001 of IFL (bolus 5-fluorouracil [5-FU], leucovorin [LV], irinotecan), FOLFOX (oxaliplatin and infused fluorouracil plus leucovorin), and IROX (irinotecan and oxaliplatin) as first line therapy for advanced colorectal cancer (N9741 [18], n = 795); a phase III randomized trial of CPT-11 (irinotecan) versus OXAL (oxaliplatin)/5-FU (5-Fluorouracil)/CF (Leucovorin) as second line therapy for patients with advanced colorectal carcinoma (N9841 [19], n = 491); and a phase II first line pemetrexed plus gemcitabine study in advanced non-small cell lung cancer (N0026 [20], n = 157). N9741 found FOLFOX to be an active regimen for treatment of patients with previously untreated advanced colorectal cancer and led to a statistically significantly improved response rate and time to disease progression compare to IFL. N9841 demonstrated noninferiority of FOLFOX4 to irinotecan in OS as second-line therapy in patients with FU-refractory disease. N0026 showed that pemetrexed followed by gemcitabine on day 1 and gemcitabine on day 8 was less toxic compared with the other treatment schedules. These trials were chosen for illustrative purposes only; all 3 trials collected cycle-by-cycle tumor measurements and individual patient data were available.

Only patients with both a baseline measurement and at least one post-baseline image-assessed tumor measurement were included in this analysis. For the two-stage modeling, patients who did not have a measurement in the first 12 weeks since randomization or died within 12 weeks were further excluded. This additional exclusion was needed because a landmark analysis approach was adopted for the second stage modeling with the landmark time chosen to be 12 weeks, and the baseline and early tumor measurements (prior to 12 weeks) were used for the first stage modeling.

2.2. Tumor measurements

Each trial collected cycle-by-cycle, lesion-by-lesion tumor measurements which were assessed by site investigators. While both N9841 and N0026 used RECIST 1.0[2] for collection and assessment, N9741 was activated prior to RECIST and therefore collected and assessed tumor measurements according to WHO criteria [21]. To facilitate fair comparisons across trials, as well as to be compliant with current standards, we applied RECIST 1.1[3] to all three trials. Specifically, we chose the maximum of the bi-dimensional measurements recorded as the single dimensional measurement for each lesion and included data from up to 5 target lesions. We did not consider non-measurable lesions or new lesion information in this analysis.

2.3. Endpoint

Overall survival (OS), defined as time from trial registration to death due to any cause, was the primary outcome of this analysis.

2.4. Statistical analyses

Notation

Let $D_i = \{T_i, \delta_i, y_{ijk}, t_{ik}, w_i\}$ be the observed data for the ith patient($i = 1, \ldots, n$), where $T_i = \min(T_i^*, C_i)$ is the observed failure time, T_i^* is the underlying (possibly unobserved) failure time, and C_i is the censoring time, and $\delta_i = 1(T_i^* < C_i)$ is the failure indicator. All time-to-event variables are measured in weeks. The observed longitudinal tumor measurement y_{ijk} represents the size (in millimeters) of tumor j from patient i at time t_{ik}, where $j = 1, \ldots, n_i \leq 5$. The sum of the longest diameter (as in RECIST 1.1) for a patient at a given time, t_{ik}, is denoted by $z_{ik} = \sum_{j=1}^{n_i} y_{ijk}$. w_i is the vector of static demographic covariates, age, sex (male vs. female), and race (white vs. non-white) for patient i at baseline. To investigate the performance of using observed longitudinal tumor measurements in predicting OS, we considered two modeling approaches as illustrated below.

2.4.1. Joint modeling

One approach to modeling a longitudinal outcome and a time-to-event outcome simultaneously is through joint modeling [16]. A fundamental assumption of joint modeling is that b_i (a vector of time-independent random effects) underlies both the longitudinal and survival processes, which implies that the random effects account for the association between both processes as well as the correlation between repeated measures in the longitudinal process, i.e. the longitudinal and survival processes are conditionally independent given b_i. Let $m_i(t)$ denote the true and possibly unobserved tumor measurements on the same scale at time t. We considered a linear mixed effects subject-specific longitudinal sub-model for tumor measurements and a Cox proportional hazards survival sub-model for OS.

The longitudinal sub-model is defined as follows:

$$z_i(t) = m_i(t) + \varepsilon_i(t) = \beta_0 + \beta_1 t + \beta_2 x_i(t) + b_{i0} + b_{i1} t + \varepsilon_i(t), \varepsilon_i(t) \sim N\left(0,\ \sigma^2\right), \quad (1)$$

where $\{\beta_0, \beta_1, \beta_2\}$ are fixed effects, $x_i(t)$ is the number of lesions being measured at time t, and b_{i0} and b_{i1} are random intercept and random slope per individual, respectively. We further assume a multivariate

normal distribution for b_i and independence between $\varepsilon_i(t)$and b_i. Note that a non-linear trajectory of the tumor measurements can be incorporated into the model by including polynomial or spline terms in the longitudinal sub-model.

The hazard function of the survival sub-model is defined as follows:

$$h_i(t|M_i(t), w_i) = h_0(t)\exp\{\gamma^T w_i + \alpha m_i(t)\} \quad (2)$$

where $M_i(t) = \{m_i(u) : 0 \le u < t\}$ denotes the longitudinal history of true unobserved tumor measurements until time$t$, $h_0(t)$ is the baseline hazard function, and parameter α quantifies the association between true tumor measurements and the risk of death. Due to computational complexity, we only considered parametric models for $h_0(\cdot)$. Specifically, the baseline risk function was assumed to be piecewise constant with six internal knots placed at equally spaced percentiles of the observed event times. The number of knots was set to be six to allow a certain degree of flexibility in the baseline risk function yet, at the same time, to include a sufficient number of events in each segment.

Given the data are observed at discrete times t_{ik}, when fitting the model, $z_i(t)$ in Equation (1) is replaced by the observed sum of lesions, z_{ik}.

The joint distribution of $\{T_i, \delta_i, z_{ik}\}$ can then be derived based on the assumption that random effects b_i explain all interdependencies within a patient. Thus, the longitudinal and survival processes are conditionally independent given b_i and the joint likelihood can be solved using a two-step approach. For further details of the mathematical derivation of these results, we refer to Rizopoulos 2012 [16].

2.4.2. Two-stage modeling

Though joint modeling is capable of utilizing all observed tumor measurement data, thereby likely to have a better prediction accuracy, it does not suggest an obvious metric (such as in RECIST 1.1) and therefore may be difficult to implement in clinical practice. To address this limitation, we propose a landmark analysis combined with a two-stage modeling strategy with potential time-varying coefficients which relaxes the proportional hazard assumption in the survival process.

In the first stage, we fit a simple linear regression model to all available individual tumor measurements (y_{ijk}) through 12-weeks for each patient and estimate subject-specific baseline tumor size, β_{0i}, and tumor size changing rate, β_{1i}. In the second stage, we fit a Cox model, landmarked at 12 weeks, with potential time-varying coefficients for the estimated, subject-specific, tumor measurement-based metrics, i.e. $\widehat{\beta}_{0i}$ and $\widehat{\beta}_{1i}$, from the first stage. Specifically, the hazard model is defined as:

$$h_i(t|y_i, \quad w_i, \quad t_M) = h_0(t)\exp\left\{\gamma^T w_i + \gamma_0(t)\widehat{\beta}_{0i} + \gamma_1(t)\widehat{\beta}_{1i}\right\}$$

for $t > t_M$, where t_M is the landmark time. The landmark time point of 12 weeks was chosen given the fact that the protocol required trials N9741, N9841 and N0026 to have 6- and 12-week assessments, and 12-week is a commonly utilized landmark time point for tumor measurement-based metrics in other literatures [9,10,22,23].

Unlike in the joint modeling approach, we did not impose any restrictions on $h_0(\cdot)$ in the model. We conducted the Grambsch-Therneau test [24] for non-proportionality. The functional forms of the potential time-varying coefficients $\gamma_0(t)$ and $\gamma_1(t)$ were chosen based on visual inspection of the lowess fit of the scaled Schoenfeld residuals and model selection criterion Akaike Information Criteria (AIC).

2.5. Model performance and predictive accuracy

To provide a benchmark for comparison, a proportional hazards model with OS as the response variable was used. RECIST-based best response (complete response vs. partial response vs. stable disease vs. progressive disease) by 12 weeks was the independent variable in the model while adjusting for the same baseline characteristics of age, sex, and race. Model performance was summarized by AIC and BIC. It is important to note that the AIC and BIC from the joint modeling and two-stage modeling should not be compared because their underlying likelihood functions are substantially different. It is, however, appropriate to compare the AIC and BIC between the proportional hazard model with RECIST-based best response and two-stage modeling. For predictive accuracy, Harrell's C [25] was used for both the proportional hazard model and two-stage modeling. For joint modeling, Rizopoulos developed a time-dependent ROC curve [26] which corresponds to a cumulative sensitivity and dynamic specificity (C/D, see the general definition of Heagerty and Zheng [27]). The area under the curve (AUC) for this time-dependent ROC curve provides a summary measure of the discriminatory power of the joint model and is termed the dynamic c-index [26]. It should be noted that Harrell's C is a weighted area under the incident sensitivity and dynamic specificity (I/D, see the general definition of Heagerty and Zheng [27]) ROC curve [27]. Given the different definitions of Harrell's C (i.e. using the I/D definition) and dynamic c-index (i.e. using the C/D definition), these c-indices are not comparable.

A 2-sided p-value of <0.05 was considered statistically significant for all tests. No adjustments were made for multiple comparisons. All analyses were performed using R (version 3.6.1) with the `JM` package for joint modeling and the `survival` package for two-stage modeling.

3. Results

Patients included in this analysis are shown in the CONSORT diagram (Fig. 1). Among the 1443 patients enrolled into one of the three clinical trials, 403 patients were excluded due to a lack of requisite measurements to perform the analysis. Patients were excluded from this analysis for the following reasons: (1) all measurements were based on clinical evaluations only (n = 17); (2) no lesion measurements available in the trial dataset (n = 236); (3) no baseline measurements recorded (n = 7); and (4) having baseline but no post-baseline measurements available (n = 143). Additional exclusions for the two-stage analysis include patients with no post-baseline assessments before 12 weeks (n = 63) or died before 12 weeks (n = 26). Baseline characteristics for patients included in the joint modelling analysis are summarized in Table 1. In total, 1273, 695, 226 lesions were recorded at baseline from patients enrolled in N9741, N9841, and N0026, respectively. Baseline tumor burden, defined as the sum of the longest diameter (cm) of all target lesions within a patient, are similar in terms of median but data from N9741 are more skewed than the others. All three trials have long-term survival follow-up with mature OS outcomes (number of events/number of patients are 559/577, 320/337, and 119/126 for N9741, N9841, and N0026, respectively).

3.1. Joint modeling

Results from the longitudinal sub-model and the random effects estimates are shown in Supplemental Table 1. The estimates of random intercepts and slopes are larger in comparison to the fixed effects estimates indicating that there is a lot of variability across individuals. Furthermore, the baseline sum of the tumor measurement ($\widehat{\beta_0}$) tends to be larger in the two colorectal trials (N9741 and N9841) than in the lung trial (N0026). A closer to zero slope in the change of tumor measurement ($\widehat{\beta_1}$) in N9841, a second-line study, indicates that the tumor shrinkage is less pronounced in N9841 compared to the first-line studies, N9741 and N0026. Results from the survival sub-model are given in Table 2. It is clear that OS does not associate with age, gender and race in any of the trials considered. The association between the risk of death and the biomarker values (i.e. sum of tumor measurements) was found to be statistically significant for N9741 and N9841 but not for N0026 (potentially due to the smaller sample size). For all 3 studies, the estimates of the hazard ratio of association variable are greater than one,

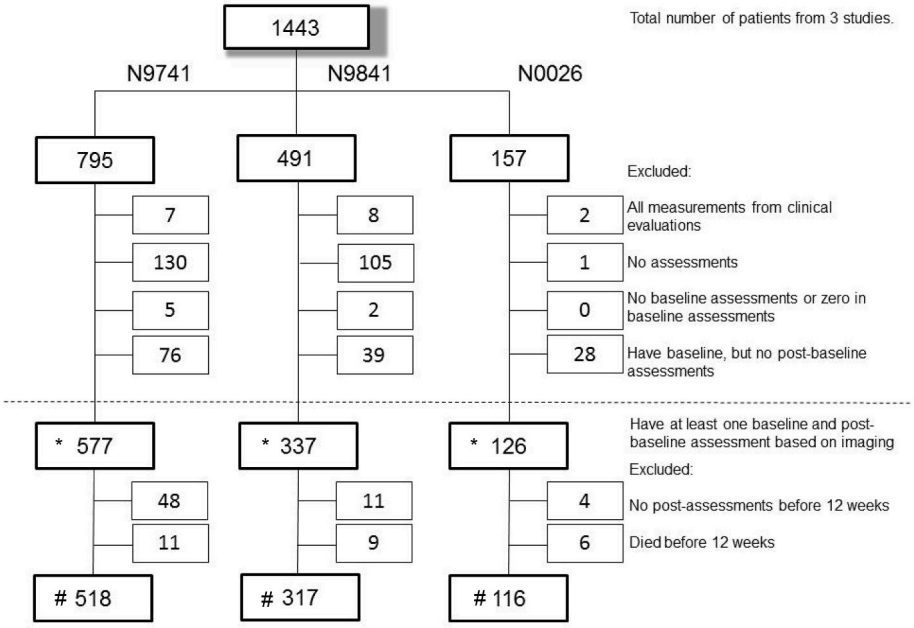


Fig. 1. CONSORT diagram depicting exclusions from each trial to reach the analysis dataset. * Data used for joint modeling portion of the analysis. # Data used for two-stage and RECIST related analysis. Number of OS events available from each trial are 559 (out of 577), 320 (out of 337), and 119 (out of 126) for N9741, N9841, and N0026, respectively.

Table 1
Baseline characteristics of patients included in the joint modelling analysis.

Characteristics	N9741 (n = 577)	N9841 (n = 337)	N0026 (n = 126)
Age, median (range)	61 (27–88)	63 (28–86)	65 (39–82)
Gender, n (%)			
Male	357 (61.87)	198 (58.75)	76 (60.32)
Female	220 (38.13)	139 (41.25)	50 (39.68)
Race, n (%)			
White	504 (87.35)	291 (86.35)	113 (89.68)
None-white	73 (12.65)	46 (13.65)	13 (10.32)
Number of TL per patient at baseline, median (range)	2 (1, 5)	2 (1, 5)	1.5 (1, 5)
Tumor burden (cm)[a] per patient at baseline, median (range)	8.2 (1.0, 42.6)	7.0 (1.0, 28.7)	6.3 (2.0,22.5)

TL: Target lesions.
[a] Tumor burden is defined as the sum of the longest diameter of all target lesions within a patient.

indicating that a larger tumor measurement (higher tumor burden) is associated with higher risk of death, as expected.

3.2. Two-stage modeling

The non-proportionality test was not statistically significant (smallest p-value = 0.48 for first-stage slope) for N0026, so no time-varying coefficient was considered for that particular trial. For models involving data from N9741 and N9841, the non-proportionality test indicates a time-dependent effect for the first-stage slope (p-value = 0.007 and 0.013, respectively). Either a piecewise linear function or a linear function with a constant after a change point was considered after examining the scaled Schoenfeld residuals with loess fit (Supplemental Fig. 1). A rough grid search (per 10 weeks increase) was conducted for the change point and the final model was chosen by AIC value (a smaller AIC was desired). A linear function with a constant after a change point was selected for N9741 and N9841 with change points at 150 and 130 weeks post-landmark time point (i.e. 12 weeks), respectively. The two-stage modeling results are shown in Table 3. Age and race were not

Table 2
Joint modeling results.

Parameter	N9741		N9841		N0026	
	Hazard Ratio (95% CI)	p-value	Hazard Ratio (95% CI)	p-value	Hazard Ratio (95% CI)	p-value
Age	1.003 (0.995, 1.011)	0.480	0.998 (0.988, 1.008)	0.726	1.013 (0.993, 1.034)	0.233
Race (White vs. Non-white)	1.135 (0.882, 1.461)	0.325	0.886 (0.629, 1.249)	0.490	0.943 (0.515, 1.726)	0.846
Gender (male vs. female)	1.100 (0.923, 1.310)	0.287	0.912 (0.720, 1.156)	0.448	1.154 (0.769, 1.732)	0.469
Association[a]	1.002 (1.001, 1.003)	<0.0001	1.006 (1.004, 1.008)	<0.0001	1.028 (0.994, 1.063)	0.053
Baseline hazard [b]	**Estimate**	**Std Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**
$\log(\xi_1)$	−5.798	0.278	−5.564	0.366	−6.397	0.741
$\log(\xi_2)$	−4.917	0.278	−4.697	0.363	−5.462	0.716
$\log(\xi_3)$	−4.927	0.278	−4.573	0.361	−5.136	0.705
$\log(\xi_4)$	−4.440	0.278	−4.318	0.360	−5.084	0.698
$\log(\xi_5)$	−4.400	0.279	−4.008	0.358	−5.588	0.680
$\log(\xi_6)$	−4.432	0.275	−3.900	0.354	−5.452	0.671
$\log(\xi_7)$	−5.078	0.287	−3.993	0.386	−5.518	0.694

Std Err: Standard Error.
[a] The estimate for α in Equation (2) which measures the association between longitudinal biomarker (i.e. tumor measurements) and the risk of death.
[b] Baseline hazard is assumed to be piecewise constant with seven knots placed at equally spaced percentiles of the observed event times.

predictive of OS, consistent with findings from the joint modeling approach, with the exception of N9841, where male gender was associated with longer OS. For N0026, the subject-specific first-stage intercept was statistically significant for predicting OS; however, the subject-specific first-stage slope was not. Since there were no time-varying coefficients incorporated into the model for N0026, the effects of subject-specific first-stage intercept and slope on OS remained constant throughout the course of the study. For N9741 and N9841, on the other hand, the subject-specific first-stage intercept and slope were highly significant and were strong predictors of OS. Given that the coefficients of the subject-specific slope from the first-stage may change over time, for illustrative purposes, we calculated the hazard ratios (HR) at 12 weeks (landmark), 1 year after landmark time, 2 years after landmark time, and the change point (Table 3). Although the association decreased over time (i.e. HR of 1.357 and 1.274 at 12 weeks and decreased to 1.127 and 0.924 at 2 years after landmark for N9741 and N9841, respectively), the subject-specific slope from the first-stage remained a strong predictor of OS within one year post landmark.

3.2.1. Model performance and predictive accuracy

AIC and BIC are summarized in Table 4. One should note that the AIC and BIC presented in Table 4 should not be compared across the joint modeling and two-stage modeling because the underlying likelihood functions are different for the two models. Predictive accuracy of the two approaches is presented in Table 5, along with results from the model using RECIST response by 12 weeks, which is intended to be used as a benchmark. The Harrell's C-index from the two-stage modeling is consistently higher (range from 0.611 to 0.633) than that obtained using RECIST response (range from 0.586 to 0.590). The dynamic c-index for joint modeling ranges from 0.627 to 0.683. Even though the dynamic c-index cannot be compared to Harrell's C-index directly, it does display a similar pattern as Harrell's C for the different trials (i.e. N9841 has the highest dynamic c-index [0.683] and Harrell's C [0.633]; while N0026 has the lowest dynamic c-index [0.627] and Harrell's C [0.611]).

Table 4
Model performance.

	N9741	N9841	N0026
Joint Modeling			
AIC	35490.060	16858.160	5631.858
BIC	35568.500	16926.920	5682.911
Two-stage Model			
AIC	5316.500	2910.281	859.102
BIC	5341.800	2932.543	872.559
RECIST[a]			
AIC	5082.823	2855.010	834.287
BIC	5107.879	2877.152	850.268

[a] Proportional hazard model with RECIST-based best response by 12 weeks treated as a 4-level categorical variable (complete response vs. partial response vs. stable disease vs. progressive disease) in the model.

Table 5
Predictive accuracy.

Model	Harrell's C-index (95% CI)	Dynamic C-index
N9741		
RECIST[a]	0.586 (0.557, 0.616)	–
Two-stage modeling	0.613 (0.584, 0.642)	–
Joint modeling[b]	–	0.646
N9841		
RECIST[a]	0.587 (0.552, 0.622)	–
Two-stage modeling	0.633 (0.598, 0.668)	–
Joint modeling[b]	–	0.683
N0026		
RECIST[a]	0.590 (0.523, 0.657)	–
Two-stage modeling	0.611 (0.554, 0.668)	–
Joint modeling[b]	–	0.627

[a] Proportional hazard model with RECIST-based best response by 12 weeks treated as a 4-level categorical variable (complete response vs. partial response vs. stable disease vs. progressive disease) in the model.
[b] Δt was set to 12 weeks.

4. Discussion

In this analysis, we adopted two different approaches, joint modeling and two-stage modeling, to predict OS using tumor measurements. The associations in the joint model were highly statistically significant for N9741 and N9841 (both colorectal cancer trials) indicating that tumor measurement data is predictive of OS for those two studies. Despite the fact that the association in the joint model for N0026 was not statistically significant, it was the greatest in magnitude (i.e. HR of 1.028) compared to N9741 and N9841 (HR of 1.002 and 1.006, respectively). The statistically non-significant result may be due to the fact that N0026 has the smallest sample size among all three trials considered.

Similar conclusions can be drawn from the two-stage modeling approach as well. A larger tumor measurement at baseline and a larger slope between baseline and landmark time both associated with a shorter OS.

A primary difference between the two approaches is in how much tumor measurement data were used to model OS. The joint modeling approach uses all tumor measurement data throughout the course of the clinical trial; whereas the two-stage modeling approach only incorporates information from early assessments (i.e. prior to 12 weeks). Our models did not account for patients developing new lesions or progression in non-target lesions. In joint modeling, the new lesion information (a binary variable) and non-target lesion status (a categorical, qualitative variable) can be easily incorporated into the longitudinal

Table 3
Two-stage modeling results.

Parameter	N9741		N9841		N0026	
	Hazard Ratio (95% CI)	p-value	Hazard Ratio (95% CI)	p-value	Hazard Ratio (95% CI)	p-value
Age	1.004 (0.996, 1.012)	0.377	0.999 (0.989, 1.010)	0.911	1.009 (0.988, 1.031)	0.394
Race (White vs. Non-white)	1.179 (0.899, 1.548)	0.234	1.072 (0.758, 1.516)	0.694	0.809 (0.417, 1.569)	0.530
Gender (Male vs. Female)	1.014 (0.844, 1.219)	0.884	0.782 (0.620, 0.987)	0.038	1.203 (0.794, 1.824)	0.384
First-stage intercept (mm)	1.008 (1.004, 1.011)	<0.0001	1.008 (1.003, 1.012)	0.0008	1.010 (1.001, 1.019)	0.024
First-stage slope (mm/week)	$e^{[0.305-0.002 a \min(\text{time in weeks},150)]}$	<0.0001	$e^{[0.242-0.003 a \min(\text{time in weeks},130)]}$	<0.0001	1.119 (0.957, 1.309)	0.160
e.g. At 12 weeks (landmark)	1.357 (1.232, 1.496)	–	1.274 (1.155, 1.405)	–	1.119 (0.957, 1.309)	–
e.g. At 1 year post landmark	1.237 (1.167, 1.311)	–	1.085 (1.007, 1.169)	–	1.119 (0.957, 1.309)	–
e.g. At 2 years post landmark	1.127 (1.044, 1.216)	–	0.924 (0.783, 1.090)	–	1.119 (0.957, 1.309)	–
e.g. At change point[a]	1.038 (0.919, 1.173)	–	0.854 (0.686, 1.062)	–	1.119 (0.957, 1.309)	–

[a] Change point is 150 weeks post landmark time point (i.e. 12 weeks) for N9741 and 130 weeks for N9841.

sub-model; however, this information remains challenging to incorporate into the two-stage modeling approach. Incorporating new lesions and non-target lesion information can potentially enhance the predictive ability of modeling and more closely adhere to the current clinical practice." Advantages of the two-stage modeling include the ease of obtaining a metric from the first stage model and ease of computation. It allows time-varying coefficients and offers better predictive accuracy than classical Cox models using RECIST. It also provides an opportunity to derive a better alternative tumor measurement-based endpoint for trial design. Advantages of the joint modeling include taking into account all tumor measurements throughout the entire course of the clinical trial and performing risk prediction for an individual patient. However, the joint modeling approach is computationally intensive. For it to be successfully implemented in practice, it will require a centralized database to house all tumor measurements, a centralized computing system to update the prediction algorithm as more data are collected, and a user-friendly interface so results can be easily understood. Since this will require a joint effort from multiple stakeholders with infrastructure, sophisticated software, and computing support, we acknowledge that this approach would be difficult to implement in practice. Further, unlike two-stage modeling, joint modeling does not readily provide an obvious metric for clinical decision making.

In conclusion, these two approaches provide easily interpretable and clinical meaningful results while using tumor measurement data with potentially missing lesion information. Alternative endpoints based on early tumor measurement data may be better at predicting OS. The approaches presented in this paper allow for utilizing serial continuous tumor measurements for predicting OS. With the advancement of computer technology, statisticians should strive to bring approaches that can help enhance risk prediction and facilitate individualized medicine decision making.

Author declaration

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of competing interest

None.

Acknowledgments

This work was partially supported by the National Institutes of Health Grants: CA167326 and P30CA15083 (Mayo Comprehensive Cancer Center Grant). The data from N9741, N9841, and N0026 were obtained directly from the Alliance for Clinical Trials in Oncology, a National Clinical Trials Network cooperative group.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2021.100827.

References

[1] P. Therasse, S.G. Arbuck, E.A. Eisenhauer, et al., New guidelines to evaluate the response to treatment in solid tumors. European organization for research and treatment of cancer, national cancer institute of the United States, national cancer institute of Canada, J. Natl. Cancer Inst. 92 (2000) 205–216.
[2] E.A. Eisenhauer, P. Therasse, J. Bogaerts, et al., New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1), Eur. J. Canc. 45 (2009) 228–247.
[3] Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics, Guidance for Industry.
[4] Guideline on the Evaluation of Anticancer Medicinal Products in Man.
[5] T.G. Karrison, M.L. Maitland, W.M. Stadler, et al., Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non small-cell lung cancer, J. Natl. Cancer Inst. 99 (2007) 1455–1461.
[6] T. Jaki, V. Andre, T.L. Su, et al., Designing exploratory cancer trials using change in tumour size as primary endpoint, Stat. Med. 32 (2013) 2544–2554.
[7] C. Suzuki, L. Blomqvist, A. Sundin, et al., The initial change in tumor size predicts response and survival in patients with metastatic colorectal cancer treated with combination chemotherapy, Ann. Oncol. 23 (2012) 948–954.
[8] H. Piessevaux, M. Buyse, M. Schlichting, et al., Use of early tumor shrinkage to predict long-term outcome in metastatic colorectal cancer treated with cetuximab, J. Clin. Oncol. 31 (2013) 3764–3775.
[9] S.J. Mandrekar, M.W. An, J. Meyers, et al., Evaluation of alternate categorical tumor metrics and cut points for response categorization using the RECIST 1.1 data warehouse, J. Clin. Oncol. 32 (2014) 841–850.
[10] M.W. An, X. Dong, J. Meyers, et al., Evaluating continuous tumor measurement-based metrics as phase II endpoints for predicting overall survival, J. Natl. Cancer Inst. 107 (2015).
[11] 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, J. Clin. Oncol. 31 (2013) 2110–2114.
[12] M.W. An, S.J. Mandrekar, M.E. Branda, et al., Comparison of continuous versus categorical tumor measurement-based metrics to predict overall survival in cancer treatment trials, Clin. Canc. Res. 17 (2011) 6592–6599.
[13] M.W. An, J. Tang, A. Grothey, et al., Missing tumor measurement (TM) data in the search for alternative TM-based endpoints in cancer clinical trials, Contemp Clin Trials Commun 17 (2020) 100492.
[14] F.-S. Ou, J.M. Hubbard, P.M. Kasi, et al., Heterogeneity in early lesion changes on treatment as a marker of poor prognosis in patients (pts) with metastatic colorectal cancer (mCRC) treated with first line systemic chemotherapy ± biologic: findings from 9,092 pts in the ARCAD database, J. Clin. Oncol. 35 (2017), 3535-3535.
[15] F.-S. Ou, Y. Lou, E.V. Cutsem, et al., Evaluation of lesion-based response at 12 weeks (LBR12) of treatment (Rx) in metastatic colorectal cancer (mCRC): findings from 9,092 patients (pts) in the ARCAD database, J. Clin. Oncol. 36 (2018), 612-612.
[16] D. Rizopoulos, Joint Models for Longitudinal and Time-To-Event Data : with Applications in R, CRC Press, Boca Raton, 2012.
[17] T.M. Therneau, P.M. Grambsch, Modeling Survival Data : Extending the Cox Model, Springer, New York, 2000.
[18] R.M. Goldberg, D.J. Sargent, R.F. Morton, et al., A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer, J. Clin. Oncol. 22 (2004) 23–30.
[19] G.P. Kim, D.J. Sargent, M.R. Mahoney, et al., Phase III noninferiority trial comparing irinotecan with oxaliplatin, fluorouracil, and leucovorin in patients with advanced colorectal carcinoma previously treated with fluorouracil: N9841, J. Clin. Oncol. 27 (2009) 2848–2854.
[20] C.X. Ma, S. Nair, S. Thomas, et al., Randomized phase II trial of three schedules of pemetrexed and gemcitabine as front-line therapy for advanced non-small-cell lung cancer, J. Clin. Oncol. 23 (2005) 5929–5937.
[21] A.B. Miller, B. Hoogstraten, M. Staquet, et al., Reporting results of cancer treatment, Cancer 47 (1981) 207–214.
[22] M.R. Sharma, E. Gray, R.M. Goldberg, et al., Resampling the N9741 trial to compare tumor dynamic versus conventional end points in randomized phase II trials, J. Clin. Oncol. 33 (2015) 36–41.
[23] M.W. An, Y. Han, J.P. Meyers, et al., Clinical utility of metrics based on tumor measurements in phase II trials to predict overall survival outcomes in phase III trials by using resampling methods, J. Clin. Oncol. 33 (2015) 4048–4057.
[24] P.M. Grambsch, T.M. Therneau, Proportional hazards tests and diagnostics based on weighted residuals, Biometrika 81 (1994) 515–526.
[25] F.E. Harrell Jr., K.L. Lee, D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Stat. Med. 15 (1996) 361–387.
[26] D. Rizopoulos, Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data, Biometrics 67 (2011) 819–829.
[27] P.J. Heagerty, Y. Zheng, Survival model predictive accuracy and ROC curves, Biometrics 61 (2005) 92–105.