Published in final edited form as:
Clin Cancer Res. 2011 October 15; 17(20): 6592–6599. doi:10.1158/1078-0432.CCR-11-0822.

Comparison of continuous versus categorical tumor measurement-based metrics to predict overall survival in cancer treatment trials

Ming-Wen An[1,2,3], **Sumithra J. Mandrekar**[1,3], **Megan E. Branda**[1], **Shauna L. Hillman**[3], **Alex A. Adjei**[4], **Henry Pitot**[5], **Richard M. Goldberg**[6], and **Daniel J. Sargent**[3]

[2]Department of Mathematics, Vassar College, Poughkeepsie NY

[3]Department of Health Sciences Research, Mayo Clinic, Rochester MN

[4]Department of Medicine, Roswell Park Cancer Institute, Buffalo NY

[5]Department of Medical Oncology, Mayo Clinic, Rochester MN

[6]Division of Hematology and Oncology, University of North Carolina, Chapel Hill, NC

Abstract

Purpose—The categorical definition of response assessed via the Response Evaluation Criteria in Solid Tumors has documented limitations. We sought to identify alternative metrics for tumor response that improve prediction of overall survival.

Experimental Design—Individual patient data from three North Central Cancer Treatment Group trials (N0026, n=117; N9741, n=1109; N9841, n=332) were used. Continuous metrics of tumor size based on longitudinal tumor measurements were considered in addition to a trichotomized response (TriTR: Response vs. Stable vs. Progression). Cox proportional hazards models, adjusted for treatment arm and baseline tumor burden, were used to assess the impact of the metrics on subsequent overall survival, using a landmark analysis approach at 12-, 16- and 24-weeks post baseline. Model discrimination was evaluated using the concordance (c) index.

Results—The overall best response rates for the three trials were 26%, 45%, and 25% respectively. While nearly all metrics were statistically significantly associated with overall survival at the different landmark time points, the c-indices for the traditional response metrics ranged from 0.59-0.65; for the continuous metrics from 0.60-0.66 and for the TriTR metrics from 0.64-0.69. The c-indices for TriTR at 12-weeks were comparable to those at 16- and 24-weeks.

Conclusions—Continuous tumor-measurement-based metrics provided no predictive improvement over traditional response based metrics or TriTR; TriTR had better predictive ability than best TriTR or confirmed response. If confirmed, TriTR represents a promising endpoint for future Phase II trials.

Keywords

continuous; tumor measurement; RECIST; prediction; survival

CORRESPONDING AUTHOR: Sumithra Mandrekar, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester MN 55905. Telephone: (507) 266-6724. Fax: (507) 266-2477. mandrekar.sumithra@mayo.edu.

[1]These authors, listed alphabetically, contributed equally to this work.

CONFLICTS OF INTERST: None.

This work was presented in part at the 2007 and 2008 Annual Meetings of the American Society of Clinical Oncology. This paper was written while M. An was on leave from Vassar College in the Department of Health Sciences Research at the Mayo Clinic.

INTRODUCTION

The high failure rate (e.g. 50-60%) of Phase III trials in oncology presents a major obstacle to the drug discovery process (1). Understanding and addressing the potential reasons for these high failure rates is crucial to making progress. Possible reasons for the high failure rate in phase III trials include 1) sub-optimal choice of patient population and 2) inaccurate predictions of effectiveness from the hypothesis generating prior Phase II trials. The focus of the present work is on the second of these reasons, and focuses on the choice of the endpoints used for Phase II trials in patients with measurable disease to identify agents worthy of further evaluation. Our goal is to consider alternative Phase II endpoints to the standardly used endpoint of tumor response, and in particular the use of endpoints based on continuous tumor measurements.

Historically, Phase II trials have used tumor response rate as the primary endpoint, where response is assessed via the Response Evaluation Criteria in Solid Tumors (RECIST; 2). RECIST was implemented in an effort to standardize assessment of tumor response and has been widely used in cancer clinical trials since 2000. Per RECIST, measurable target lesions representative of all involved organs are identified, recorded and measured at baseline using uni-dimensional tumor measurements. The overall patient-level objective status is then determined based on the assessment of the target lesions, non-target lesions and new lesions. *Best response* is defined as the best objective status (i.e., complete or partial response, stable disease, or progression; each of which is based on relative change in tumor size) on treatment. *Confirmed response* is defined as two consecutive assessments of complete or partial response (CR or PR) assessed at least 4 weeks apart. Two observations are worth noting about the RECIST criteria. First, by definition confirmed response, in contrast to best response, requires that the response status of the patient be sustained for at least a period of 4 weeks, thus avoiding to some extent possible overestimation of the observed response rate due to one-time measurement error. This is particularly important in non-randomized trials where tumor response is the primary endpoint. Second, these definitions of response are categorical and specifically, dichotomous (CR/PR or not). Modified RECIST guidelines, RECIST 1.1, were introduced in 2009. These changes do not materially impact the subject matter of the present paper.

The concerns over the response rate as a primary endpoint are well documented. First, there is a demonstrated lack of concordance between response rates in single-center Phase II trials and subsequent multicenter Phase III studies (3). More fundamentally, tumor measurements are continuous and their categorization may result in loss of information (4). A related concern is the use of an arbitrary cutoff to determine “response” and “no response” (5), and timing of assessments. With the advent of targeted therapies that prolong disease stabilization, patients may experience stable disease (SD) rather than tumor shrinkage. It has been shown that patients with SD also achieve clinical benefit (6), and hence it is not appropriate to ignore SD when assessing treatment efficacy.

Non-progression rate (also known as disease control rate (DCR)) has become one accepted alternate endpoint in assessing treatment efficacy as it includes patients who achieve SD for an extended period of time as a success, in addition to those who achieve complete or partial response. DCR was shown to be better than response rate in predicting survival in the setting of Non-Small Cell Lung Cancer (7). A trichotomous response has also been considered, where response is categorized into CR/PR vs. SD vs. progression (6). Bradbury et al (8) and Dhani et al (9) provide a recent review of the many proposed phase II trial alternate endpoints.

Actual tumor measurements are relatively simple to obtain and have been previously explored by others to be used in a Phase II endpoint. Karrison et al (10) considered log change in the sum of tumor measurements from baseline to 8 weeks as a phase II trial endpoint. Wang et al (11) developed a model for tumor size and for survival; the primary goal of the tumor model was to account for missing tumor measurement data. Claret et al (12) developed a mathematical model to predict overall survival from baseline and 7-week predicted tumor size, using a simulation study to compare observed and predicted data.

In this paper, we propose several continuous metrics based on the tumor measurements recorded over course of treatment. We hypothesized that these continuous metrics would more fully capture a patient’s tumor lesion experience over the course of the treatment, compared to traditional dichotomous or trichotomous response categorization. The goal is to identify an appropriate metric that can be assessed relatively early during treatment, which is predictive of longer term clinical outcomes such as overall survival.

METHODS

Data

We obtained individual patient tumor measurement and survival data from three North Central Cancer Treatment Group (NCCTG) cancer clinical trials: a Phase II first line Pemetrexed plus Gemcitabine study in advanced non small cell lung cancer (N0026, n=157; 13), a Phase III randomized study of IFL, FOLFOX4, and IROX as first line therapy for advanced colorectal cancer (N9741, n=1691; 14), and a Phase III randomized study of CPT-11 vs. OXAL/5-FU/CF as second line therapy for advanced colorectal cancer (N9841, n=491; 15). N0026 had three treatment arms in which none was found superior (overall response rate RR = 19%). N9741 randomized patients to three treatment arms IFL (Arm A), FOLFOX (Arm F) and IROX (Arm G). Arm F was found to be the most effective treatment, arm A was the previous standard of care, and arm G was found to have some efficacy but high toxicity. Overall RR for N9741 was 38%. N9841 had two treatment arms between which there was no difference found in overall survival (overall RR = 22%).

All patients with measurable disease who had a baseline measurement as well as at least one post-baseline measurement were included in our analysis. Patients who progressed or went off study for any reason prior to their first post-baseline measurement were excluded. The analysis dataset therefore included 117, 1109, and 332 patients from N0026, N9741, and N9841 respectively; with associated response rates of 26%, 45%, and 25% respectively. Baseline characteristics for the patients included in the analyses are summarized in Table 1.

Both N9841 and N0026 utilized the RECIST criteria measurement for collection and assessment. N9741 opened prior to RECIST and instead collected and assessed tumor measurements according to WHO criteria. Applying the RECIST criteria to N9741 data, we used the maximum of the two measurements recorded for each lesion. The number of lesions measured at each assessment varied over time within patients, thus only lesion measurements that were available across all assessments for the patient were utilized. Each study was designed to assess the lesions at 4-6 week intervals and assessed up to 10 lesion measurements.

Metrics

The continuous metrics we consider are based on the tumor measurements recorded over course of treatment. As such, these metrics are different than previously considered work on continuous tumor measurements (e.g., of tumor size at *k*-weeks for some fixed *k*) in that these capture an overall tumor burden over the course of study. Table 1 summarizes the metrics. The total sum of measurements (TSM) was calculated simply as the sum of measurements at each assessment starting from baseline. To account for total time on study, we considered the average sum of measurements (ASM), which is calculated by dividing the TSM by the total number of assessments. As another metric, we considered the relative change from baseline (RCB), defined as follows: for each assessment, divide the sum of measurements at that assessment by the sum of the baseline measurements and subtract 1 from this ratio, and then sum this quantity over all assessments. With this definition, negative values of the RCB indicate a decrease in tumor measurements. We also considered the average relative change from baseline (ACB), which is the RCB divided by the number of assessments. In addition to these continuous metrics, we also considered trichotomized response, defined as CR/PR vs. Stable vs. Progression, using the RECIST criteria. In particular, we considered trichotomized response (TriTR) status at a pre-defined timepoint and best trichotomized response (Best TriTR), as well as the traditional confirmed response.

Statistical Analysis

Both the total sum of measurements (TSM) and the average sum of measurements (ASM) were log-transformed to normalize the distribution in order to satisfy model assumptions. Distributions of other continuous metrics appeared approximately symmetric and unimodal, and therefore were not transformed. A Cox proportional hazards model adjusting for the metric, treatment arm, and sum of baseline measurements (i.e., baseline tumor burden) was fit, where the metrics were calculated from tumor measurements available at randomization (baseline) until (1) 12-weeks post-randomization (12-week landmark analysis); (2) 16-weeks post-randomization (16-week landmark analysis); and (3) 24-weeks post-randomization (24-week landmark analysis). In each analysis, we considered the continuous metrics, confirmed response, trichotomized best response, and trichotomized response status at the pre-defined timepoints. For this last metric, the objective status at the assessment closest to that time point, i.e. within 3 weeks from the expected assessment time, was utilized. When utilizing a landmark analysis approach, TSM and ASM are theoretically equivalent since they only differ by a constant factor, that is, TSM divided by the number of assessments yields the ASM. However not all patients have the same number of assessments in a given time period so that ASM and TSM do not necessarily differ by a constant factor across all patients. We have therefore included both metrics for consideration. A similar comment applies to RCB and ARCB.

To understand the associations between each metric and survival, we considered hazard ratios for each metric. As a measure of model fit and a means to compare non-nested models, we considered the Akaike Information Criteria (AIC).

Discrimination and calibration were considered to better understand predictive utility of our response metrics, our primary goal. To determine how well a model (and by extension, a metric) discriminates among patients with different outcomes we utilized the concordance index (or c-index; 16-17). The c-index in the context of survival considers all pairs of individuals. If one can determine which individual in the pair first died, then this pair is *evaluable*. If the patient with a *higher* estimated hazard ratio dies *prior* to the one with a *lower* estimated hazard ratio then the pair is deemed *concordant*, otherwise they are considered discordant. The index is the fraction of all evaluable pairs that are concordant. It ranges from 0.5 to 1.0 where 0.5 indicates no association and 1 indicates perfect association. We assessed calibration by comparing expected and observed survival probabilities at 1-year as follows: patients were grouped into deciles of their predicted probabilities from a Cox model. Within each decile, we calculated the average predicted probability (“expected”) and the Kaplan-Meier estimate (“observed”), and compared these in plots and via an informal measure calculated as the sum of the squared difference between expected and observed probabilities. In comparing metrics, our focus was on discrimination.

RESULTS

Table 3 provides results from the 12-week landmark analysis, adjusting for metric, treatment arm and sum of baseline measurements. Figure 1 presents the c-indices for all 7 metrics (represented by different symbols) for each study. In this section, unless otherwise noted, we present results from the 12-week landmark analysis. Results for the 16- and 24-week landmark analyses were similar and are included as supplemental tables.

All of the categorical and continuous response metrics were found to be statistically significantly associated with overall survival (hazard ratios, p<0.05; Table 3). The AIC values ranged from 799 (for the model with ASM) to 815 (for Confirmed Response) in N0026; from 11872 (for TriTR) to 12699 (for Confirmed Response) in N9741; and from 2987 (for ARCB) to 3058 (for RCB) in N9841.

Based on the plots and measure of observed versus expected 1-year survival probabilities, calibration across models was reasonable and comparable (Figure 2; Supplemental Figures 1-2). Plots for 2- and 3-year survival probabilities (not shown) revealed similar findings. The c-indices for confirmed response ranged from 0.59-0.65; for the continuous metrics, from 0.60-0.66; and for TriTR metrics, from 0.64-0.69 (Table 3). Three observations regarding the c-indices hold across all three studies. First, while continuous metrics were all statistically significantly associated with overall survival (via hazard ratios; Table 3), they provided minimal (if any) improvement in prediction for overall survival compared to the various categorical response metrics based on the c-index. In Figure 1, the c-indices for categorical response metrics (open circles, open point-down triangles, and solid circles; range: 0.59-0.69 from Table 3) are as high as or higher than those for continuous metrics (other symbols; range: 0.60-0.66 from Table 3). This suggests that in general the categorical response metrics may be better predictors of overall survival than the continuous metrics. Second, the c-indices for trichotomized response status at 12-weeks (solid circles; range: 0.64-0.69 from Table 3) are at least as high as or higher than those for best trichotomized response (open circles; range: 0.64-0.67 from Table 3) *and* confirmed response (open point-down triangles; range: 0.59-0.65 from Table 3) within the same time frame. Third, the c-indices for trichotomized response at 12-weeks (solid circles; range: 0.64-0.69) are comparable to those for trichotomized response at 16- and 24-weeks (range: 0.65-0.70 and 0.65-0.66 respectively from Supplemental Tables). As reference, we also calculated the c-indices for models that only included treatment and baseline sum of measurements based on measurement data available at 12-weeks. These were 0.59, 0.61, and 0.64 for N0026, N9741, and N9841 respectively.

All of these analyses were also repeated within each arm for N9741, the large randomized trial, and the results were similar (data not shown). Of note is that the c-indices for response at 12-weeks were consistently as high as or higher than those for other metrics. Figure 3 shows Kaplan-Meier survival curves by trichotomized response status at 12-weeks and by ACB (dichotomized at the median) for each study.

DISCUSSION

Contrary to our hypothesis, we found that the continuous metrics we assessed provide no predictive advantage over the categorical response metrics. However, we do recommend further study of the trichotomized response at early time points (e.g. 12-, 16-, and 24-weeks) with particular attention to 12-week status. This metric has at least two advantages. First, it addresses the concern over ignoring stable disease by including SD as a separate category; second it can be assessed earlier since it does not require confirmation and does not require data from the entire study period.

It is interesting to note that N0026 and N9841 had relatively low response rates (25-26%), yet the trichotomized response still performed as well as in N9741 which had a higher response rate (45%). Likely the trichotomized response appropriately recognizes the survival benefit associated with stable disease by placing such patients into their own category rather than combined in the same category as progression. A natural extension to the trichotomized response would be a 5-level metric (CR vs. PR vs. Stable vs. Increasing vs. Prog). However, this also has some inherent limitations, specifically, 1) the need to specify a cut-point to distinguish between Increasing and Progression, where the choice for this is not obvious, and 2) the complete response (CR) rate is often small in oncology studies, for example, in our data, the CR rates for N0026, N9741, and N9841 were 0%, 4.2%, and 3.3% respectively.

The inability for the continuous metrics we assessed to improve survival prediction may be due to several factors. First, when considered over an entire study population, tumor growth may be sufficiently ‘regular’ that measurements at a fixed time point post-baseline adequately characterize tumor activity. Second, the imaging frequency could be too infrequent to capture the tumor size changes. Alternatively, unidimensional tumor size may not be the most accurate measure of disease aggressiveness; functional imaging, volumetric assessment, or other advanced imaging methods may offer improvements. Finally, it may be too much to expect any early tumor measurement related endpoints to predict overall survival in settings where second and later line therapies are used (18). An important assumption for the validity of endpoints based on our continuous metrics is that patient tumors are measured at regular intervals which do not differ by arm. This is to eliminate the possible bias that could arise in the following situation: two patients have similar tumor growth trajectories, but one has a tumor measurement at j weeks and the other has a tumor measurement as j+i weeks (for i>0) by which time the tumor is a different size than at week j. As a result, these patients may have different tumor response profiles based on our continuous metrics

An additional limitation to the current data is the inability to effectively assess the impact of the missing measurement data due to clinical progression, new lesions and missing assessments. Moreover, the number of lesions measured at each assessment was variable, and the current analysis used only the lesion measurements that were available across all assessments for the patient. Since not all lesion measurements at each assessment were used, the measurement data from each cycle used to compute the metrics could be biased. Future work should consider further exploration of trichotomized response as well as alternative continuous metrics since simple scalar summaries such as those we considered may not likely capture “the” key features of the tumor growth curve. For example, it is possible to have one patient for whom the tumor decreases over time and another patient for whom tumor increases over time, but for these two patients to have identical sums of measurements. Further, tumor growth curves often exhibit non-linearity, e.g. initial tumor shrinkage followed by progression. In order to capture key features of the tumor growth curve and thereby to improve prediction, a metric will likely need to be composite, for example, a linear combination of multiple scalar summaries such as those considered in this paper. Longitudinal modeling, e.g. mixed models, is another option others have previously considered (e.g. 12).

In conclusion, our data suggest that categorical response metrics predict survival as well as or better than the continuous tumor-measurement-based metrics considered in this work. Furthermore, trichotomized response at early timepoints, possibly as early as 12-weeks, are worthy of further study as an alternative endpoint in Phase II trials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

FINANCIAL SUPPORT: National Cancer Institute Grants No CA-15083 (Mayo Clinic Cancer Center) and CA-25224 (North Central Cancer Treatment Group)

References

1. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2004; 3:711–715. [PubMed: 15286737]
2. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst. 2000; 92:205–216. [PubMed: 10655437]
3. Zia MI, Siu LL, Pond GR, Chen EX. Comparison of outcomes of phase II studies and subsequent randomized control studies using identical chemotherapeutic regimens. J Clin Oncol. 2005; 23:6982–91. [PubMed: 16192585]
4. Lavin PT. An alternative model for the evaluation of antitumor activity. Cancer Clin Trials. 1981; 4:451–457. [PubMed: 7318127]
5. Pivot X, Thierry-Vuillemin A, Villanueva C, Bazan F. Response Rates: A Valuable Signal of Promising Activity? Cancer J. 2009; 15:361–365. [PubMed: 19826354]
6. Sargent D, Campbell M, Grothey A, Goldberg R. Overall and 12-week tumor response versus actual tumor measurements as predictors of overall survival in advanced colorectal cancer – findings from NCCTG N9741. ASCO. 2008 Abstract.
7. Lara P, Redman M, Kelly K, Edelman M, Williamson S, Crowley J, et al. Disease Control Rate at 8 Weeks Predicts Clinical Benefit in Advanced Non-Small Cell Lung Cancer: Results From Southwest Oncology Group Randomized Trials. J Clin Oncol. 2008; 26:463–467. [PubMed: 18202421]
8. Bradbury P, Seymour L. Tumor Shrinkage and Objective Response Rates: Gold Standard for Oncology Efficacy Screening Trials, or an Outdated Endpoint? Cancer J. 2009; 15(5):354–360. [PubMed: 19826353]
9. Dhani N, Tu D, Sargent DJ, Seymour L, Moore MJ. Alternate Endpoints for Screening Phase II Studies. Clin Caner Res. 2009; 15(6):1873–1882.
10. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of Phase II Cancer Trials Using a Continuous Endpoint of Change in Tumor Size: Application to a Study of Srrafenib and Erlotinib in Non-Small-Cell Lung Cancer. J Natl Cancer Inst. 2007; 99(19):1455–1461. [PubMed: 17895472]
11. Wang Y, Sung C, Dartois C, Ramchandani R, Booth BP, Rock E, 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. 2009; 86(2):167–174. [PubMed: 19440187]
12. Claret L, Girard P, Hoff PM, Van Cutsem E, Zuideveld KP, Jorga K, et al. Model-Based Prediction of Phase III Overall Survival in Colorectal Cancer on the Basis of Phase II Tumor Dynamics. J Clin Oncol. 2009; 27:4103–4108. [PubMed: 19636014]
13. Ma CX, Nair S, Thomas S, Mandrekar SJ, Nikcevich DA, Rowland KM, 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. 2005; 23:5929–5937. [PubMed: 16135464]
14. Goldberg RM, Sargent DJ, Morton RF, Fuchs CS, Ramanathan RK, Williamson SK, 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. 2004; 22:23–30. [PubMed: 14665611]
15. Kim GP, Sargent DJ, Mahoney MR, Rowland KM Jr, Philip PA, Mitchell E, 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. 2009; 27:2848–2854. [PubMed: 19380443]

16. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. J Am Med Assoc. 1982; 247:2543–2546.
17. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996; 15:361–387. [PubMed: 8668867]
18. Broglio KR, Berry DA. Detecting an Overall Survival Benefit that is Derived From Progression-Free Survival. J Natl Cancer Inst. 2000; 101:1642–1649. [PubMed: 19903805]

STATEMENT OF TRANSLATIONAL RELEVANCE

The high failure rate (e.g. 50-60%) of Phase III trials in oncology, attributable in part to less than optimal predictions of effectiveness from hypothesis generating Phase II trials, presents a major obstacle to the drug discovery process. Historically, Phase II trials have used (categorical) tumor response rate as the primary endpoint, an approach that has documented concerns. In this paper, we explore longitudinal tumor measurement-based continuous metrics and alternative categorical response metrics. Our results suggest that an unconfirmed trichotomized objective status assessed as early as 12-weeks post treatment initiation predicts for subsequent survival as well as or better than traditional response based or our continuous metrics. This trichotomized objective status metric, if validated, may positively impact the drug development process by: a) accurately reflecting the intended goals of current therapies, b) shortening assessment time, and c) improving prediction of subsequent survival.

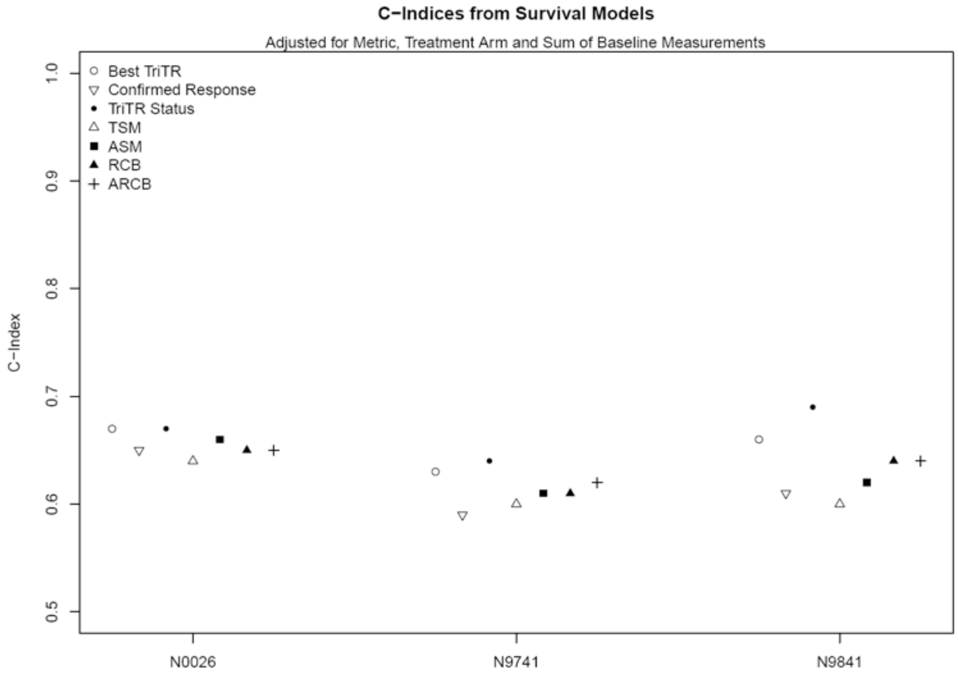

Figure 1.

C-Indices from survival models adjusted for metric, treatment arm and sum of baseline measurements

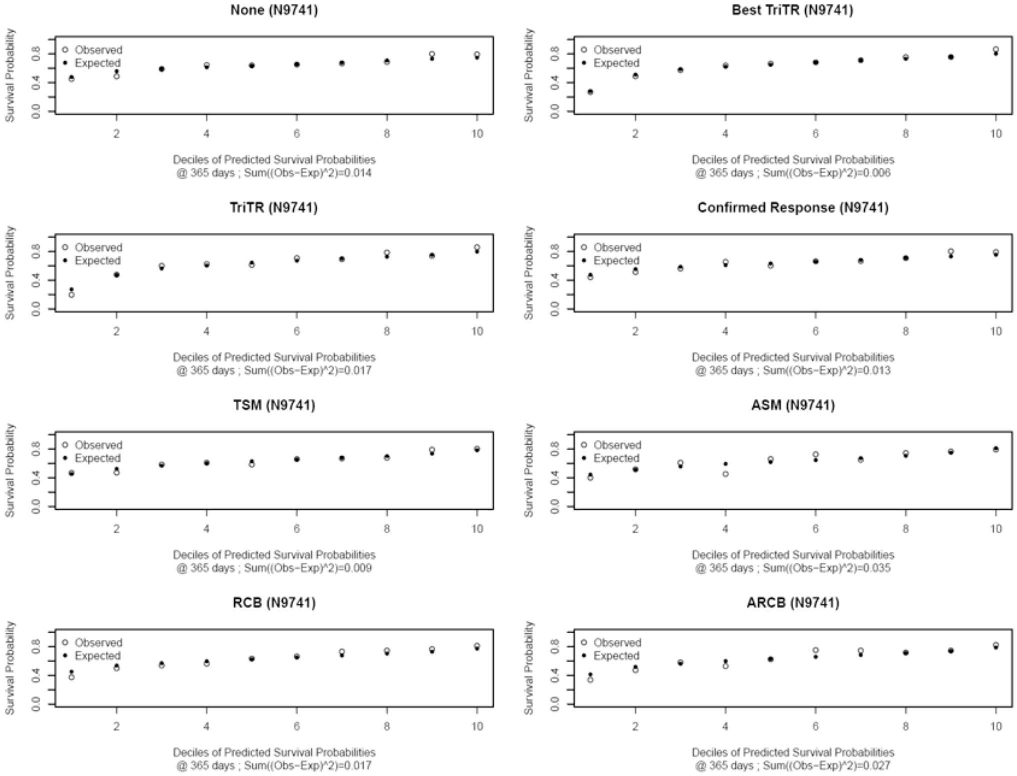

Figure 2.

Plots of observed versus expected 1-year survival probabilities for N9741 based on models for each metric and a model with no metric, adjusting for treatment arm and sum of baseline tumor measurements

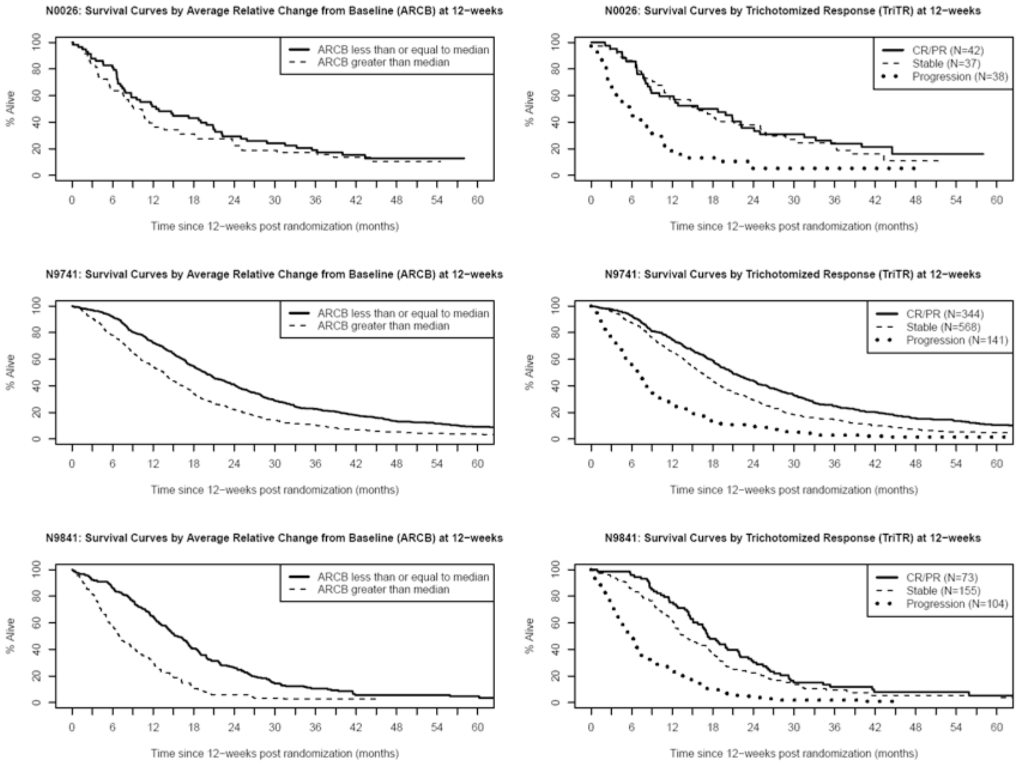

Figure 3.

Kaplan-Meier Survival Curves by Dichotomized Average Change since Baseline (ACB) (left column) and Trichotomized Response (TriTR) (right column) at 12-weeks, for each study (row)

Table 1

Definition of Metrics.

Metric	**Definition**
Trichotomized Best Response (Best TriTR)	Best objective status (CR/PR vs. Stable vs. Progression)
Confirmed Response	Two consecutive assessments of complete or partial response assessed at least 4 weeks apart (CR/PR vs. no CR/PR)
Trichotomized Response Status (TriTR) at a pre-defined timepoint	Objective status at the assessment closest to the pre-defined timepoint (CR/PR vs. Stable vs. Progression)
Total Sum of Measurements (TSM)	Sum of measurements at each assessment
Average Sum of Measurements (ASM)	Sum of measurements divided by the number of assessments
Relative change from baseline (RCB)	Sum of the relative changes since baseline, calculated as follows: divide the sum of measurements at each assessment by the sum of the baseline measurements and subtract 1; then sum this quantity over all assessments
Average relative change from baseline (ARCB)	RCB divided by the number of assessments

Table 2

Baseline characteristics of patients included in analysis.

Characteristics	N0026 (N=117)	N9741 (N=1109)	N9841 (N=332)
Age: Median (Range)	65 (39, 81)	61 (19, 85)	63 (28, 86)
PS: 0-1[1]	117 (100%)	1062 (96%)	321 (97%)
2	0 (0%)	47 (4%)	11 (3%)
Male	70 (60%)	692 (62%)	195 (59%)
Stage: IIIB	16 (14%)	0 (0%)	0 (0%)
IV[2]	101 (86%)	1109 (100%)	332 (100%)
Overall Survival (Years): Median (95% CI)	1.1 (1.0, 1.4)	1.6 (1.5, 1.7)	1.0 (1.0, 1.2)
% Alive[3]: 2 Years	13 (11%)	198 (18%)	42 (13%)
3 Years	8 (7%)	88 (8%)	11 (3%)
4 Years	9 (8%)	52 (5%)	4 (1%)
5 Years	5 (4%)	62 (6%)	4 (1%)
Overall Confirmed Response Rate (RR)	30 (26%)	498 (45%)	83 (25%)

[1] Per eligibility criteria for N0026.

[2] Per eligibility criteria for N9741 and N9841.

[3] Percent alive at year and up to following year.

Table 3

Up to 12 Week Assessments adjusting for treatment arm and sum of baseline tumor measurements.

Study	Metric	HR[2]	AIC[3]	Chi-Square	P-Value	C-Index
N0026	Best TriTR @ 12 wks[1]:					
	Stable	1.29	805.3	31.2	<.0001	0.67
	Progression	3.71				
	TriTR @ 12 weks[1]					
	Stable	1.12	807.0	29.5	<.0001	0.67
	Progression	2.58				
	Confirmed Response	3.67	815.4	19.1	0.0003	0.65
	TSM	1.44	809.0	15.9	0.0012	0.64
	ASM	3.91	799.0	25.9	<.0001	0.66
	RCB	1.92	807.0	17.9	0.0005	0.65
	ARCB	9.40	805.1	19.8	0.0002	0.65
N9741	Best TriTR @ 12 wks[1]:					
	Stable	1.40	12590.5	201.9	<.0001	0.63
	Progression	3.71				
	TriTR @ 12 weks[1] :					
	Stable	1.41	11872.0	198.4	<.0001	0.64
	Progression	3.37				
	Confirmed Response	1.37	12699.0	91.3	<.0001	0.59
	TSM	1.39	12055.7	112.1	<.0001	0.60
	ASM	1.76	12026.6	141.2	<.0001	0.61
	RCB	1.24	12055.4	112.4	<.0001	0.61

Study	Metric	HR[2]	AIC[3]	Chi-Square	P-Value	C-Index
	ARCB	3.99	12029.2	138.6	<.0001	0.62
N9841	Best TriTR @ 12 wks[1]:					
	Stable	1.26	3017.7	81.7	<.0001	0.66
	Progression	3.42				
	TriTR @ 12 weks[1] :					
	Stable	1.30	3005.2	105.8	<.0001	0.69
	Progression	3.77				
	Confirmed Response	1.15	3028.6	34.0	<.0001	0.61
	TSM	1.18	3004.8	34.7	<.0001	0.60
	ASM	1.70	2995.8	43.6	<.0001	0.62
	RCB	1.55	3058.4	50.6	<.0001	0.64
	ARCB	4.28	2987.3	52.2	<.0001	0.64

[1] Reference CR/PR

[2] HR: Hazard Ratio for 1 unit change in metric

[3] Akaike Information Criteria