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

Ming-Wen An, Yu Han, Jeffrey P. Meyers, Jan Bogaerts, Daniel J. Sargent, and Sumithra J. Mandrekar

Ming-Wen An, Vassar College, Poughkeepsie, NY; Yu Han, Novartis Pharmaceuticals, East Hanover NJ; Jeffrey Meyers, Daniel J. Sargent, and Sumithra J. Mandrekar, Mayo Clinic, Rochester, MN; and Jan Bogaerts, European Organisation for Research and Treatment of Cancer, Brussels, Belgium.

Published online ahead of print at www.jco.org on October 26, 2015.

Written on behalf of the RECIST Steering Committee.

Supported in part by Grant No. CA167326 from the National Cancer Institute, National Institutes of Health.

Authors' disclosures of potential conflicts of interest are found in the article online at www.jco.org. Author contributions are found at the end of this article.

Corresponding author: Ming-Wen An, PhD, Department of Mathematics and Statistics, Vassar College, 124 Raymond Ave, Poughkeepsie NY 12604; e-mail: mian@vassar.edu.

© 2015 by American Society of Clinical Oncology

0732-183X/15/3334w-4048w/\$20.00

DOI: 10.1200/JCO.2015.60.8778

ABSTRACT

Purpose

Phase II clinical trials inform go/no-go decisions for proceeding to phase III trials, and appropriate end points in phase II trials are critical for facilitating this decision. Phase II solid tumor trials have traditionally used end points such as tumor response defined by Response Evaluation Criteria for Solid Tumors (RECIST). We previously reported that absolute and relative changes in tumor measurements demonstrated potential, but not convincing, improvement over RECIST to predict overall survival (OS). We have evaluated the metrics by using additional measures of clinical utility and data from phase III trials.

Methods

Resampling methods were used to assess the clinical utility of metrics to predict phase III outcomes from simulated phase II trials. In all, 2,000 phase II trials were simulated from four actual phase III trials (two positive for OS and two negative for OS). Cox models for three metrics landmarked at 12 weeks and adjusted for baseline tumor burden were fit for each phase II trial: absolute changes, relative changes, and RECIST. Clinical utility was assessed by positive predictive value and negative predictive value, that is, the probability of a positive or negative phase II trial predicting an effective or ineffective phase III conclusion, by prediction error, and by concordance index (c-index).

Results

Absolute and relative change metrics had higher positive predictive value and negative predictive value than RECIST in five of six treatment comparisons and lower prediction error curves in all six. However, differences were negligible. No statistically significant difference in c-index across metrics was found.

Conclusion

The absolute and relative change metrics are not meaningfully better than RECIST in predicting OS.

J Clin Oncol 33:4048-4057. © 2015 by American Society of Clinical Oncology

INTRODUCTION

Phase II trials inform go/no-go decisions for proceeding to phase III trials, and appropriate phase II end points are critical for facilitating this decision. Phase II trials for solid tumors have traditionally used tumor response (TR) as defined by the Response Evaluation Criteria for Solid Tumors (RECIST).¹ However, concerns over the appropriateness of RECIST in measuring treatment benefit and predicting long-term outcomes have led to the pursuit of alternative end points.²⁻⁶ Although many alternatives have been identified as promising, none have emerged a clear winner and, more importantly, none have replaced RECIST response in trial practice.

We previously explored various tumor measurement (TM)-based end points as RECIST alternatives. We evaluated trichotomous TR (complete response [CR] or partial response [PR] v stable disease [SD] v progressive disease [PD]), disease control rate (CR/PR/SD v PD), and dichotomous TR (CR/PR v SD/PD), by using a range of cutoff points for defining categories.⁷ These alternative categorical end points offered no meaningful improvement over RECIST in predicting overall survival (OS). We subsequently evaluated absolute (and relative) changes in TM between baseline and 6 weeks and between weeks 6 and 12.⁸ Again, no meaningful improvement in OS prediction was found with these continuous end points over RECIST. In these analyses, the primary criterion of predictive ability was

discrimination measured by the concordance index (c-index),⁹ which is commonly used when comparing phase II end points.

In our study, we applied a different set of measures for assessing predictive ability. These measures more directly address the concern of high failure rates (50% to 60%) in phase III trials.^{10,11} Specifically, we considered positive predictive value (PPV) and negative predictive value (NPV) defined as the probability for a phase III trial to be a success or failure (ie, OS benefit associated with treatment [or not]) if the phase II trial yields a go or no-go decision. PPV combines the true-positive and false-positive rates and NPV combines the true-negative and false-negative rates of phase II trials on drug efficacy. In addition to PPV/NPV, we considered two other measures that inform clinical utility: prediction error and SE for the c-index. To calculate these, we simulated phase II data by resampling from actual phase III trials. For each simulated phase II trial, we evaluate three metrics—RECIST response and absolute and relative changes in TM—by fitting separate Cox models for each, as described by Mandrekar et al.^{7,8} Then we calculated the three measures by using model-predicted risk scores and averaging across the simulated phase II trials.

The approach of resampling from actual trials has been applied previously. Tang et al¹² compared single-arm historically controlled versus randomized concurrently controlled phase II designs in terms of power and type I error rate, which differs from the objective of our work. Sharma et al^{13,14} shared a goal similar to ours of evaluating alternative end points. However, a key distinction is that they calculated power or the true-positive rate (ie, the probability of a phase II trial to yield a go decision, given that the phase III trial was a success) instead of PPV/NPV. In a commentary on the study by Sharma et al,¹³ LeBlanc and Tangen¹⁵ recommended PPV/NPV for evaluating the clinical utility of phase II metrics; Rubenstein et al¹⁶ made a similar recommendation. We therefore calculated PPV/NPV to assess predictive ability of phase II TM-based end points.

METHODS

Data

Data from four phase III trials (two on colon cancer; two on non–small-cell lung cancer [references not provided because of data confidentiality]) were used. These trials, hereafter referred to as trials 1 to 4, were selected for having sufficient sample size to make a resampling approach meaningful and for having treatment arm information available. Three trials used OS as the primary end point, and one used progression-free survival (PFS) as the primary and OS as the secondary end point. Each trial included cycle-by-cycle TM, treatment arm, and survival data.

We excluded patients from analysis for reasons indicated in the CONSORT diagram (Fig 1). Of note, for trial 1, we included patients from the control arm and the most promising treatment arm. Our final analysis data set included 292, 434, 351, and 328 patients from trials 1, 2, 3, and 4, respectively. Because trials 3 and 4 had three treatment arms, we performed pairwise comparisons (active 1 v control, active 2 v control). In total, we considered six treatment comparisons, two of which were positive (ie, OS benefit was associated with treatment; trials 1 and 2) and four of which were negative (ie, no OS treatment benefit; trial 3 active 1 v control, trial 3 active 2 v control, trial 4 active 1 v control, and trial 4 active 2 v control).

Following the protocols set forth in Mandrekar et al,^{7,8} we used a 12-week landmark analysis. Specifically, we included patients who were alive and on study at 12 weeks and used TM from baseline to 12 weeks to predict subsequent OS (ie, with baseline defined at 12 weeks). To investigate potential selection bias introduced by using landmark analyses, we analyzed OS by using the original cohort of all patients with baseline at random assignment and

subsequent OS by using landmarked patients with baseline at 12 weeks. Results from these two analyses followed similar trends, alleviating concerns about bias (not shown because of data confidentiality). Figure 2 shows Kaplan-Meier curves of subsequent OS by trial and treatment arm. Table 1 summarizes trial outcomes and TM for these phase III trials.

Simulation Framework

Figure 3 shows a schematic of our simulation framework. To simulate a single phase II trial by using data from a given (actual) phase III trial, we sampled 90 patients with replacement from among the phase III patients with 1:1 randomization to treatment and control arms. This design required 32 events and had 80% power to detect a hazard ratio of 0.55 comparing treatment versus control based on a one-sided test with $\alpha = .20$. We then fit each of the three previously validated Cox models of subsequent OS (RECIST response, absolute changes, relative changes⁸), adjusting for average baseline sum defined as the sum of baseline lesions divided by the number of baseline lesions. The absolute (relative) changes model included terms for the slope (% change) in TM between baseline and week 6 and between weeks 6 and 12 as well as linear spline terms that allowed for different relationships between risk of death and TM change, depending on whether the change was positive or negative. We explored the possibility of treatment effect heterogeneity across values of each metric by using subpopulation treatment effect pattern plots.^{17,18} A sensitivity analysis was performed to investigate the robustness of results to choice of parameter values. There was no evidence for treatment effect heterogeneity (Data Supplement). We therefore did not include treatment-by-metric interaction terms in our models.

From each model, we obtained predicted risk scores (ie, linear predictor) for each patient. Higher scores reflect higher risk for death. We further obtained three outcomes from the risk scores. First, we compared risk scores between arms by using a Wilcoxon rank sum test. The test result was recorded as positive if the risk scores on the treatment arm were statistically significantly less than those on the control arm and negative otherwise. Second, we calculated prediction error at specific time points according to Gerds and Schumacher.¹⁹ Prediction error at time t measures calibration by quantifying the discrepancy between observed and predicted probability of survival beyond time t . Prediction error can be any non-negative value, with values closer to 0 reflecting better calibration. Third, we calculated the c-index. The c-index measures discrimination by quantifying a model's ability to differentiate between patients with different survival outcomes. C-indexes range from 0 to 1, with values of 0.5 to 1.0 reflecting good discrimination, 0.5 reflecting no discrimination, and 0.0 to 0.5 reflecting good reverse discrimination. In addition to calculating c-index for each model, we calculated pairwise differences in c-indexes for any two models (ie, RECIST v absolute, RECIST v relative, and absolute v relative).

Summarizing Results Across 2,000 Simulated Phase II Trials for a Given Phase III Trial

From each phase III trial, we simulated 2,000 phase II trials according to the framework previously described. For each of the three Cox models, by using the values of the three outcomes (Wilcoxon result, prediction error, and c-index) from across these 2,000 trials, we estimated three measures: (1) PPV or NPV for each validated model if the given phase III trial was positive or negative, (2) prediction error, and (3) SE for the c-index.

Regarding (1), PPV is defined as $\Pr(\text{positive phase III} | \text{positive phase II})$ and NPV as $\Pr(\text{negative phase III} | \text{negative phase II})$, where $\Pr()$ denotes probability. To calculate PPV, we first calculated the true-positive rate (TPR) as $\Pr(\text{positive phase II} | \text{positive phase III})$ by taking the proportion of the 2,000 simulated phase II trials that were positive if the corresponding phase III trial was positive. We also need the probability of a positive phase III trial [$\Pr(\text{positive phase III})$]—a measure of drug efficacy—and the false-positive rate (FPR) [$\Pr(\text{positive phase II} | \text{negative phase III})$]. Neither could be estimated from the simulated data, so we specified a range of values: $0 \leq \Pr(\text{positive phase III}) \leq 0.90$ and $0.10 \leq \text{FPR} \leq 0.90$. Finally we combined these probabilities by using Bayes theorem to calculate PPV. Calculations for NPV were similar, except the true-negative rate (TNR) was estimated from phase II trials that were simulated from negative phase III trials, and a range of values for $\Pr(\text{negative phase III})$ and false-negative rate (FNR) was specified. Note that TPR is equivalent to

Fig 1. CONSORT diagram showing exclusions from final analysis data set, by trial.

the power that Sharma et al^{12,13} calculated and that TPR is related to but distinct from PPV. Regarding (2), we averaged prediction errors at time t across the 2,000 trials. Regarding (3), SE estimates of the c-index were taken as the standard deviation of the 2,000 c-indexes. In addition, to compare c-indexes between two models, we obtained 95% CIs for the true pairwise differences in c-indexes, by taking the 2.5th and 97.5th percentiles of the 2,000 pairwise differences.

Plots of PPV versus Pr(positive phase III) and of NPV versus Pr(negative phase III) for $\alpha = .20$ and FPR/FNR fixed at 0.10 and plots of prediction error versus time were generated.

RESULTS

Table 1 summarizes data from the observed phase III trials and offers a benchmark comparison for simulated phase II trial results. The Data

Supplement includes box plots of TM changes by RECIST response categories; as expected, the PD RECIST category showed the highest positive change in both metrics. Response rates are consistently higher in the treatment versus control arms across all four trials. In terms of continuous metrics (absolute and relative changes), there are some cases in which the control arm has, on average, slightly larger negative change (ie, improvement) compared with the treatment arm. For example, in trial 2, for the last slope term (between weeks 6 and 12), the control arm had an average decrease of 1.89 mm (v 1.85 mm for the treatment arm). Except for three marginal cases (RECIST model in trial 3 active 1 v control, absolute change model in trial 4 active 1 v control, and relative change model in trial 4 active 2 v control), there were no statistically significant differences in risk scores between arms. In trial 2, the RECIST model had the highest c-index; in trial 1, the

Fig 2. Kaplan-Meier survival curves showing subsequent survival using landmarked patients with baseline defined at 12 weeks. (A) Trial 1: 1-0. (B) Trial 2: 1-0. (C) Trial 3: 1-0. (D) Trial 3: 2-0. (E) Trial 4: 1-0. (F) Trial 4: 2-0.

absolute change model had the highest c-index; and in trials 3 and 4, the relative change model had the highest c-index.

The median observed number of events across the 2,000 phase II trials exceeded 32, the required number of events based on the sample size calculations described previously. Specifically, the median numbers of events were 83, 37, 49, 49, 53, and 53 for the six treatment comparisons.

Figure 4 shows plots of PPV versus $\Pr(\text{positive phase III})$ and NPV versus $\Pr(\text{negative phase III})$ for $\alpha = .20$ and $\text{FPR} = \text{FNR} = 0.10$. Higher curves reflect better predictive values. Values of PPV and NPV on the diagonal correspond to no additional information provided by the phase II trial beyond what we already knew of the phase III trial's probability of success or failure. In three of the six comparisons (trial 2, trial 3 active 1 v control, trial 4 active 2 v control), the curve for the

Table 1. Observed Phase III Trials: Summary of Subsequent Overall Survival (landmarked at 12 weeks), Tumor Measurement Data, and Cox Models

Statistic	Positive Phase III Trials				Negative Phase III Trials							
	Trial 1		Trial 2		Trial 3				Trial 4			
	Treatment (n = 160)	Control (n = 132)	Treatment (n = 221)	Control (n = 213)	Control (0) (n = 134)	Treatment (1) (n = 97)	Treatment (2) (n = 120)	Control (0) (n = 114)	Treatment (1) (n = 115)	Treatment (2) (n = 99)		
Subsequent overall survival data summary												
No. of events	142	127	73	104	73	52	63	67	67	58		
Median survival, days (95% CI)	655 (589 to 794)	594 (517 to 690)	694 (569 to NA)	530 (480 to 660)	433 (394 to NA)	436 (332 to NA)	423 (352 to NA)	376 (352 to 552)	428 (335 to 468)	453 (351 to 511)		
Hazard ratio (95% CI)	0.76 (0.59 to 0.96)	—	0.58 (0.43 to 0.79)	—	—	1.05 (0.74 to 1.5)	1.02 (0.73 to 1.43)	—	0.99 (0.71 to 1.39)	0.94 (0.66 to 1.34)		
Log-rank <i>P</i>	.022		< .001		.815		.909		.95	.726		
Tumor measurement data summary												
Response rate, No. (%)	92 (58)	69 (52)	92 (42)	88 (41)	84 (63)	63 (65)	78 (65)	48 (21)	55 (48)	50 (51)		
First slope, mean (SD)	−0.27 (0.57)	−0.27 (0.34)	−3.59 (3.75)	−3.08 (5.40)	−4.08 (4.14)	−4.22 (3.82)	−4.13 (4.26)	−2.46 (2.92)	−3.68 (4.09)	−3.19 (3.95)		
Last slope, mean (SD)	−0.26 (0.52)	−0.14 (0.24)	−1.85 (2.97)	−1.89 (3.05)	−1.44 (2.34)	−1.54 (2.91)	−0.83 (2.09)	−0.86 (2.44)	−0.89 (2.42)	−1.05 (2.84)		
First percent change, mean (SD)	−0.27 (0.47)	−0.29 (0.29)	−0.30 (0.31)	−0.24 (0.39)	−0.45 (0.33)	−0.49 (0.35)	−0.46 (0.34)	−0.32 (0.34)	−0.41 (0.39)	−0.40 (0.36)		
Last percent change, mean (SD)	−0.32 (0.42)	−0.24 (0.33)	−0.19 (0.33)	−0.18 (0.30)	−0.25 (0.42)	−0.27 (0.41)	−0.18 (0.36)	−0.17 (0.45)	−0.15 (0.35)	−0.23 (0.58)		
Cox model 1: RECIST response												
Median risk score	0.98	1.03	1.13	1.14	0.94	0.90	0.89	1.13	1.13	1.01		
Wilcoxon <i>P</i>	.22		.39		.04		.25		.23	.07		
C-index	0.580		0.601		0.566		0.557		0.543	0.572		
Cox model 2: slope												
Median risk score	0.91	0.92	0.95	0.95	0.94	0.94	0.88	0.93	0.90	0.90		
Wilcoxon <i>P</i>	.76		.26		.78		.57		.04	.33		
C-index	0.606		0.573		0.536		0.568		0.565	0.570		
Cox model 3: percent change												
Median risk score	0.98	0.97	1.04	1.02	0.99	0.97	0.97	0.99	0.99	0.93		
Wilcoxon <i>P</i>	.64		.42		.11		.72		.35	.04		
C-index	0.587		0.588		0.570		0.599		0.576	0.580		
NOTE. Cox model 1 (RECIST): average baseline sum + RECIST response; Cox model 2 (absolute changes): average baseline sum + first slope + last slope												

NOTE. Cox model 1 (RECIST): average baseline sum + RECIST response; Cox model 2 (absolute changes): average baseline sum + first slope + first slope, pos + last slope + last slope, pos; Cox model 3 (relative changes): average baseline sum + first % change, pos + last % change + last % change, pos. Risk score is taken to be the linear predictor from the Cox model. Here, first slope, pos represents a linear spline term, with a knot at first slope equaling 0. These terms are similarly defined for last slope, pos, first % change, pos, and last % change, pos.

Abbreviations: c-index, concordance index; NA, not attainable; SD, standard deviation.

Fig 3. Schematic of the simulation framework. From one observed phase III trial, we simulated 2,000 phase II trials and recorded three outcomes for each simulated trial. The results across the 2,000 simulated phase II trials were then combined to calculate measures of clinical utility for the phase III trial. c-index, concordance index; NPV, negative predictive value; PPV, positive predictive value.

absolute change metric is consistently higher than those for the RECIST and relative change metrics, suggesting that in terms of PPV and NPV, the absolute change metrics outperform the other two. However, in trial 1, the RECIST curve is highest and the absolute change curve is lowest and, in fact, near the diagonal. In trial 3 active 2 versus control and trial 4 active 1 versus control, the relative change curve is highest.

Figure 5 presents the prediction error curves. Lower curves reflect better calibration. For all six comparisons, the absolute change metric curves are almost always the lowest over time, suggesting that calibration of the absolute change metrics is better than or at least no worse than the other two.

Figure 6 shows box plots of the 2,000 c-indexes for the three Cox models and six comparisons. Each distribution approximates the sampling distribution of a c-index estimator, with its mean (standard deviation) serving as point (SE) estimates for the true c-index. The means of the 2,000 c-indexes are close to the c-indexes from the corresponding phase III trials, suggesting that phase II trial c-indexes may be unbiased for phase III trial c-indexes. Moreover, the 95% CIs

for the pairwise differences in c-indexes include 0 (Data Supplement), suggesting no statistically significant difference in discriminatory ability of the metrics.

DISCUSSION

Continuous TM-based metrics have increasingly been suggested to have greater promise than RECIST in predicting longer-term outcomes.²⁻⁵ The goal of this study was to assess the predictive ability of RECIST and continuous TM-based end points, specifically absolute and relative changes as phase II end points, on phase III outcomes to help understand the high failure rates of phase III trials. Results from the c-indexes in this study agree with those previously reported,⁸ suggesting no statistically significant difference in discriminatory ability across the three metrics. Similarly, results from the prediction curves in this study agree with the Hosmer-Lemeshow goodness-of-fit statistics from Mandrekar,⁸ suggesting similar calibration across the three metrics. However, it is unclear how these directly measure a

Fig 4. The positive predictive value (negative predictive value) versus the probability of negative (positive) phase III trial with assumed false-positive rate (false-negative rate) of 0.10 estimated by using simulated data for trials 1 and 2 (trials 3 and 4). Statistical significance level was $\alpha = .20$. Higher curves reflect higher predictive value and are therefore more desirable. “0” denotes the control arm in trials 3 and 4. (A) Trial 1. (B) Trial 2. (C) Trial 3: 1-0. (D) Trial 3: 2-0. (E) Trial 4: 1-0. (F) Trial 4: 2-0.

metric’s ability to improve the high failure rate of phase III trials. For example, it may be that neither the absolute nor the relative change metrics outperform RECIST in terms of discriminatory ability, but one might nevertheless have a lower phase III failure rate than RECIST. We were interested in investigating such possibilities and specifically asking, If a phase II trial is positive, what is the probability that the subsequent phase III trial will also be positive?

PPV and NPV provide direct measures for this question.^{15,16} However, unlike other measures of predictive ability such as the c-index, they require the probability that a phase III trial is a success (failure) and require the FPR (FNR), neither of which can be estimated from our data. So we treat these as parameters with a range of specified values. We therefore obtain a range of PPV/NPV values instead of a single PPV/NPV value.

Fig 5. Prediction error curves for trials 1, 2, 3, and 4. Lower curves reflect less prediction error and are therefore more desirable. “0” denotes the control arm in trials 3 and 4. (A) Trial 1. (B) Trial 2. (C) Trial 3: 1-0. (D) Trial 3: 2-0. (E) Trial 4: 1-0. (F) Trial 4: 2-0.

Sharma et al¹³ adopted a similar approach to evaluate different phase II end points by using one positive and one negative phase III trial in rectal cancer and one positive phase III trial in colorectal cancer.¹⁴ Rather than PPV/NPV, they focused on TPR/FPR. They found the mean log ratio of tumor size between baseline and 6 (or 8) weeks had a higher TPR and also a higher FPR than RECIST.¹³ They also found time to tumor growth had comparable or higher TPR than PFS and log ratio of tumor size at 18 weeks; they did not evaluate RECIST response.¹⁴ Sharma et al were interested in the power of a

metric; hence, they focused on TPR/FPR, whereas we were interested in a related but different outcome—the impact of metrics on the high phase III failure rates—and so we calculated PPV/NPV. It is also important to note that our study, similar to the work of others, included data from only two disease groups. Furthermore, the colorectal cancer trials were positive and the lung cancer trials were negative, thus limiting the generalizability of these findings. Finally, subsequent therapies may change the OS-metrics relationship, thus limiting the use of OS to evaluate the metrics. To the best of our knowledge, two of the

Fig 6. Distributions of concordance index (c-index) values. Each set of three plots corresponds to a treatment comparison. Within each set, for each of the three models, a box plot shows the distribution of 2,000 c-index values, a black square denotes the mean of the 2,000 c-index values, and an open triangle denotes the c-index value from the given phase III trial. The mean of the 2,000 values is close to the value from the phase III trial. The bootstrap percentile CIs for pairwise differences in c-indexes include 0 (not shown). Cox model 1 (RECIST): average baseline sum + RECIST response; Cox model 2 (absolute changes): average baseline sum + first slope + first slope.pos + last slope + last slope.pos; Cox model 3 (relative changes): average baseline sum + first % change + first % change.pos + last % change + last % change.pos. Here, first slope.pos represents a linear spline term, with a knot at first slope equaling 0. These terms are similarly defined for last slope.pos, first % change.pos, and last % change.pos. (A) Trial 1. (B) Trial 2. (C) Trial 3: 1-0. (D) Trial 3: 2-0. (E) Trial 4: 1-0. (F) Trial 4: 2-0.

four trials did not allow crossover or postprogression therapy, one allowed postprogression therapy but not crossover, and the last allowed both. PFS could be considered an end point instead of OS. However, in these advanced disease settings, median PFS is relatively short (8.2 months for colorectal cancer; 5.7 months for non-small-cell lung cancer), thus excluding a large number of PFS events from a landmark analysis and limiting the use of PFS. It could be worthwhile to develop an alternative strategy for combin-

ing a landmark approach with a time-dependent component for the first 12 weeks to account for early deaths, progressions, and dropouts.

On the basis of findings from the various measures of predictive ability—c-index, prediction curves, and PPV/NPV—a consistent message emerges. With a few exceptions, RECIST has slightly lower predictive ability. However, accounting for sampling variability, these differences appear negligible, suggesting that absolute

and relative change metrics are no better than RECIST at predicting phase III outcomes. This general message reiterates the message found in Mandrekar et al.^{7,8} Therefore, on the basis of our exhaustive complete-case analysis of the RECIST data warehouse, we do not recommend replacing RECIST with absolute and relative changes. However, it will still be important to repeat these analyses after addressing the important limitations encountered in previous analyses^{7,8}: accounting for missing measurements from missed assessments or lesions and incorporating progression from new lesions or nontarget lesions into the metric calculations to verify that the same message emerges. At that point, if the conclusions remain unchanged, it may be necessary to expand the scope of data beyond unidimensional TM and include other OS-associated tumor characteristics to further improve phase III predictions.

In conclusion, we adopted a resampling-based approach by using data from four real-world phase III trials to evaluate the PPV, NPV, prediction error, and discriminatory ability of three phase II end points—RECIST, and absolute and relative changes in TM—for pre-

dicting OS. We conclude that the absolute and relative change metrics are not meaningfully better than RECIST in predicting OS.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the authors are available with this article at www.jco.org.

AUTHOR CONTRIBUTIONS

Conception and design: Ming-Wen An, Jan Bogaerts, Daniel J. Sargent, Sumithra J. Mandrekar

Collection and assembly of data: Ming-Wen An, Jeffrey P. Meyers, Jan Bogaerts, Daniel J. Sargent, Sumithra J. Mandrekar

Data analysis and interpretation: All authors

Manuscript writing: All authors

Final approval of manuscript: All authors

REFERENCES

- Eisenhauer EA, Therasse P, Bogaerts J, et al: New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer* 45:228-247, 2009
- Karrison TG, Maitland ML, Stadler WM, 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* 99:1455-1461, 2007
- Jaki T, André V, Su TL, et al: Designing exploratory cancer trials using change in tumour size as primary endpoint. *Stat Med* 32:2544-2554, 2013
- Claret L, Gupta M, Han K, et al: Evaluation of tumour-size response metrics to predict overall survival in Western and Chinese patients with first-line metastatic colorectal cancer. *J Clin Oncol* 31:2110-2114, 2013
- Suzuki C, Blomqvist L, Sundin A, et al: The initial change in tumor size predicts response and survival in patients with metastatic colorectal cancer treated with combination chemotherapy. *Ann Oncol* 23:948-954, 2012
- An MW, Mandrekar SJ, Branda ME, et al: Comparison of continuous versus categorical tumor measurement-based metrics to predict overall survival in cancer treatment trials. *Clin Cancer Res* 17:6592-6599, 2011
- Mandrekar SJ, An MW, Meyers J, et al: Evaluation of alternate categorical tumor metrics and cut points for response categorization using the RECIST 1.1 data warehouse. *J Clin Oncol* 32:841-850, 2014
- Mandrekar SJ, An MW, Dong X, et al: Multi-trial evaluation of longitudinal tumor measurement (TM)-based metrics for predicting overall survival (OS) using the RECIST 1.1 data warehouse. *J Clin Oncol* 31, 2013 (suppl 15s; abstr 6520).
- Harrell FE Jr, Califf RM, Pryor DB, et al: Evaluating the yield of medical tests. *JAMA* 247:2543-2546, 1982
- Kola I, Landis J: Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 3:711-715, 2004
- Gan HK, You B, Pond GR, et al: Assumptions of expected benefits in randomized phase III trials evaluating systematic treatments in cancer. *J Natl Cancer Inst* 104:590-598, 2012
- Tang H, Foster NR, Grothey A, et al: Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. *J Clin Oncol* 28:1936-1941, 2010
- Sharma MR, Karrison TG, Jin Y, et al: Resampling phase III data to assess phase II trial designs and endpoints. *Clin Cancer Res* 18:2309-2315, 2012
- Sharma MR, Gray E, Goldberg RM, et al: Resampling the N9741 trial to compare tumor dynamic versus conventional end points in randomized phase II trials. *J Clin Oncol* 33:36-41, 2014
- Le Blanc M, Tangen C: Choosing phase II endpoints and designs: Evaluating the possibilities. *Clin Cancer Res* 18:2130-2132, 2012
- Rubenstein L, LeBlanc M, Smith MA: More randomization in phase II trials: Necessary but not sufficient. *J Natl Cancer Inst* 103:1075-1077, 2011
- Bonetti M, Gelber RD: A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. *Stat Med* 19:2595-2609, 2000
- Bonetti M, Gelber RD: Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* 5:465-481, 2004
- Gerds TA, Schumacher M: Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom J* 48:1029-1040, 2006

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

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

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or jco.ascopubs.org/site/ifc.

Ming-Wen An

No relationship to disclose

Yu Han

Employment: Novartis Pharmaceuticals

Jeffrey P. Meyers

No relationship to disclose

Jan Bogaerts

No relationship to disclose

Daniel J. Sargent

No relationship to disclose

Sumithra J. Mandrekar

No relationship to disclose

Acknowledgment

We thank the other members of the RECIST Working Group for their valuable feedback: Janet Dancey, Elisabeth de Vries, Robert Ford, Steve Gwyther, Wendy Hayes, Otto Hoekstra, Erich Huang, Saskia Litiere, Yan Liu, Margaret Mooney, Larry Rubinstein, Larry Schwartz, Lesley Seymour, Lalitha Shankar, Patrick Therasse, and Helen Young (authors Jan Bogaerts, Daniel J. Sargent, and Sumithra J. Mandrekar are also members of RECIST Working Group). We also thank the following organizations for making data available for the RECIST data warehouse: Amgen, AstraZeneca, Breast Cancer International Research Group, Bristol-Myers Squibb, European Organization for Research and Treatment of Cancer, Breast Cancer Group and GI Group, Erasmus University Medical Center, Rotterdam, Netherlands, Genentech, Roche, and sanofi-aventis.