VOLUME 32 · NUMBER 8 · MARCH 10 2014

JOURNAL OF CLINICAL ONCOLOGY

STATISTICS IN ONCOLOGY

Evaluation of Alternate Categorical Tumor Metrics and Cut Points for Response Categorization Using the RECIST 1.1 Data Warehouse

Sumithra J. Mandrekar, Ming-Wen An, Jeffrey Meyers, Axel Grothey, Jan Bogaerts, and Daniel J. Sargent

Sumithra J. Mandrekar, Jeffrey Meyers, Axel Grothey, and Daniel J. Sargent, Mayo Clinic, Rochester, MN; Ming-Wen An, Vassar College, Poughkeepsie, NY; and Jan Bogaerts, European Organisation for Research and Treatment of Cancer, Brussels, Belgium.

Published online ahead of print at www.jco.org on February 10, 2014.

Written on behalf of the Response Evaluation Criteria in Solid Tumors (RECIST) Working Group.

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

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Sumithra J. Mandrekar, MD, Division of Biomedical Statistics and Informatics, Harwick 8, Mayo Clinic, 200 First St SW, Rochester, MN 55905; e-mail: mandrekar.sumithra@mayo.edu.

0732-183X/14/3208w-841w/$20.00

DOI: 10.1200/JCO.2013.52.3019

ABSTRACT

Purpose

We sought to test and validate the predictive utility of trichotomous tumor response (TriTR; complete response [CR] or partial response [PR] *v* stable disease [SD] *v* progressive disease [PD]), disease control rate (DCR; CR/PR/SD *v* PD), and dichotomous tumor response (DiTR; CR/PR *v* others) metrics using alternate cut points for PR and PD. The data warehouse assembled to guide the Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1 was used.

Methods

Data from 13 trials (5,480 patients with metastatic breast cancer, non–small-cell lung cancer, or colorectal cancer) were randomly split (60:40) into training and validation data sets. In all, 27 pairs of cut points for PR and PD were considered: PR (10% to 50% decrease by 5% increments) and PD (10% to 20% increase by 5% increments), for which 30% and 20% correspond to the RECIST categorization. Cox proportional hazards models with landmark analyses at 12 and 24 weeks stratified by study and number of lesions (fewer than three *v* three or more) and adjusted for average baseline tumor size were used to assess the impact of each metric on overall survival (OS). Model discrimination was assessed by using the concordance index (c-index).

Results

Standard RECIST cut points demonstrated predictive ability similar to the alternate PR and PD cut points. Regardless of tumor type, the TriTR, DiTR, and DCR metrics had similar predictive performance. The 24-week metrics (albeit with higher c-index point estimate) were not meaningfully better than the 12-week metrics. None of the metrics did particularly well for breast cancer.

Conclusion

Alternative cut points to RECIST standards provided no meaningful improvement in OS prediction. Metrics assessed at 12 weeks have good predictive performance.

J Clin Oncol 32:841-850. © 2014 by American Society of Clinical Oncology

INTRODUCTION

The high failure rate of phase III trials in oncology is potentially attributable to inaccurate efficacy predictions from the hypothesis-generating prior phase II trials.[1] Historically, phase II trials have used tumor response rate as the primary end point (assessed as early as 7 or 8 weeks after treatment initiation), in which response is assessed via the Response Evaluation Criteria in Solid Tumors (RECIST) criteria.[2,3] Per RECIST, the patient-level objective status is determined on the basis of unidimensional tumor measurements of the target lesions, nontarget lesions, and new lesions. A primary concern regarding the use of tumor response as a phase II trial end point is the demonstrated lack of concordance between response rates in phase II trials and the typical time-to-event outcomes (progression-free survival [PFS] and overall survival [OS]) in subsequent phase III studies.[4,5] This may be attributed to two main limitations of response: first, the assignment into "response" and "no response" categories on the basis of cut points derived from historic measurement error considerations as opposed to associations with outcome.[2,3] Specifically, a partial response (PR) is defined according to RECIST 1.1 criteria as at least a 30% decrease in the sum of the longest diameter of target lesions, taking as a reference the baseline sum of longest diameters; progressive disease (PD) is defined as at least a 20% increase, taking as a reference the smallest recorded sum or appearance of a new lesion (and at least 5 mm absolute increase in version 1.1), or new lesion recorded (with additional [^{18}F]fluorodeoxyglucose positron emission tomography assessment

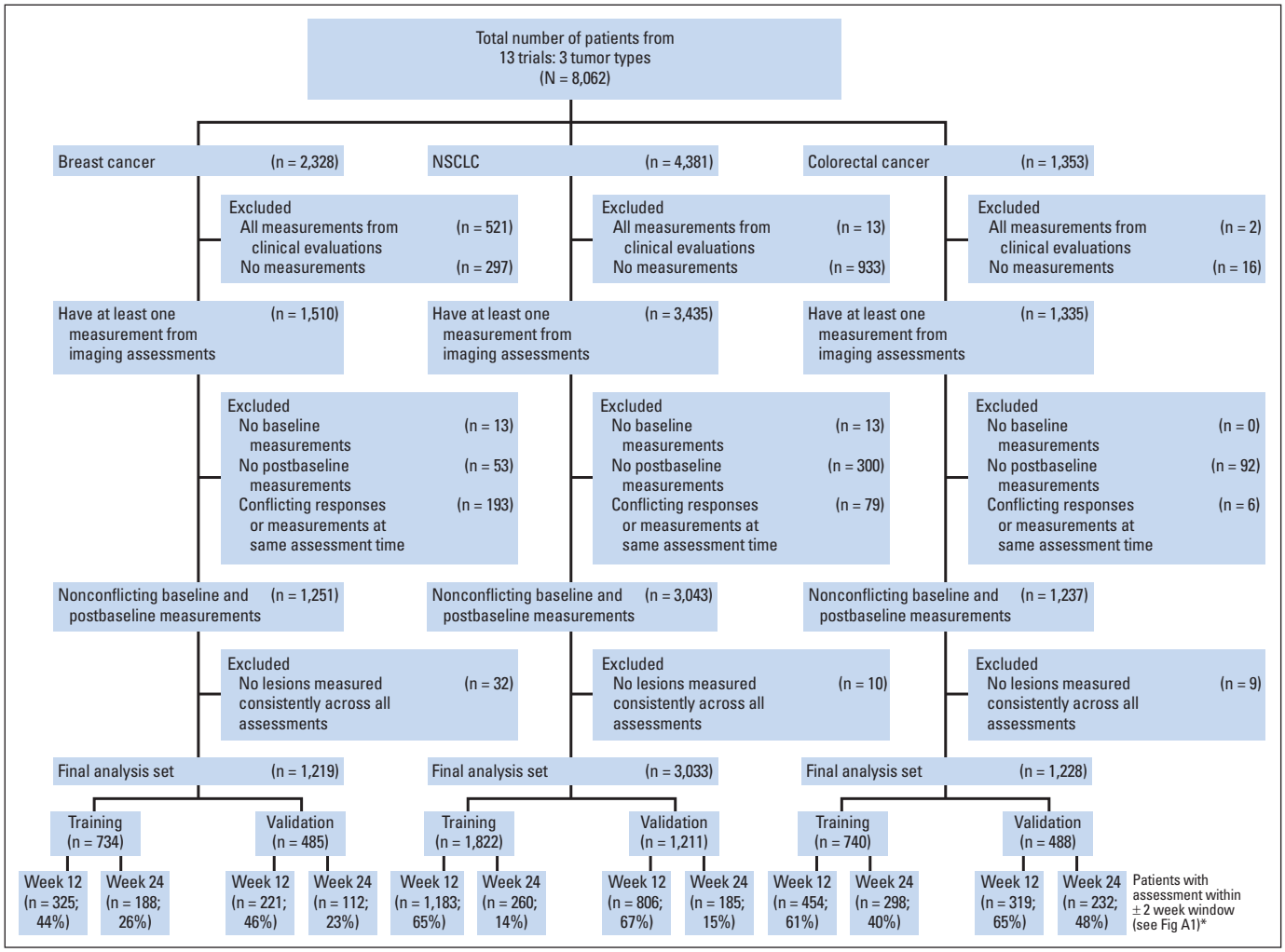

Fig 1. CONSORT diagram. (*) Twelve- and 24-week subsets are unique and created according to scenarios 1, 2, 3, and 6 in Appendix Figure A1. NSCLC, non–small-cell lung cancer.

in version 1.1). Second, the lack of distinction between stable disease (SD) and minor PD: the inability of the RECIST definition for SD to distinguish among patients whose tumors increase although not enough to be classified as progression, patients whose tumor measurements decrease although not enough to be classified as response, and patients whose tumor measurements are truly stable (neither increase nor decrease).

Alternate categorical end points have been explored and proposed to address some of these concerns.[6-11] For example, nonprogression rate or the disease control rate (DCR) classifies patients who achieve SD for an extended period of time as a success, in addition to those who achieve complete response (CR) or PR. DCR was shown to be superior to response rate in predicting survival in the setting of non–small-cell lung cancer (NSCLC).[8,9] A trichotomous tumor response (TriTR) has also been considered, in which response is categorized into CR/PR versus SD versus PD.[7,11] With the advent of targeted therapies that prolong disease stabilization, patients may experience SD rather than tumor shrinkage (CR/PR). Ignoring SD when assessing treatment efficacy, as is the case with the RECIST dichotomous tumor response (DiTR) metric, is therefore not appropriate. The TriTR metric recognizes the survival benefit associated with SD by placing such patients into their own category rather than combining them with the CR/PR (as with the dichotomous DCR metric) or with the PD (as with the DiTR metric) categories. We previously reported that confirmation of response had no impact on concordance with survival and, in particular, that response status at earlier time points performed as well as confirmed response.[11] In addition, we found that TriTR may improve prediction of subsequent survival compared with RECIST-based response metrics.[11] Although these findings suggest the potential for additional improvement on RECIST-based response metrics, they were not validated by using data from additional trials or by using a large database. Moreover, the cut point definitions used in RECIST have never been systematically examined regarding their optimality for predicting long-term survival outcomes, especially if other categorizations into the four groups (CR, PR, SD, PD) enhance prediction of survival outcomes.

In this study, we sought to systematically evaluate potential alternate cut points for PR and PD besides the RECIST cut points and then examine alternate classifications of tumor metrics for predicting OS by using the database that was assembled to guide the development of the RECIST 1.1 criteria.[3]

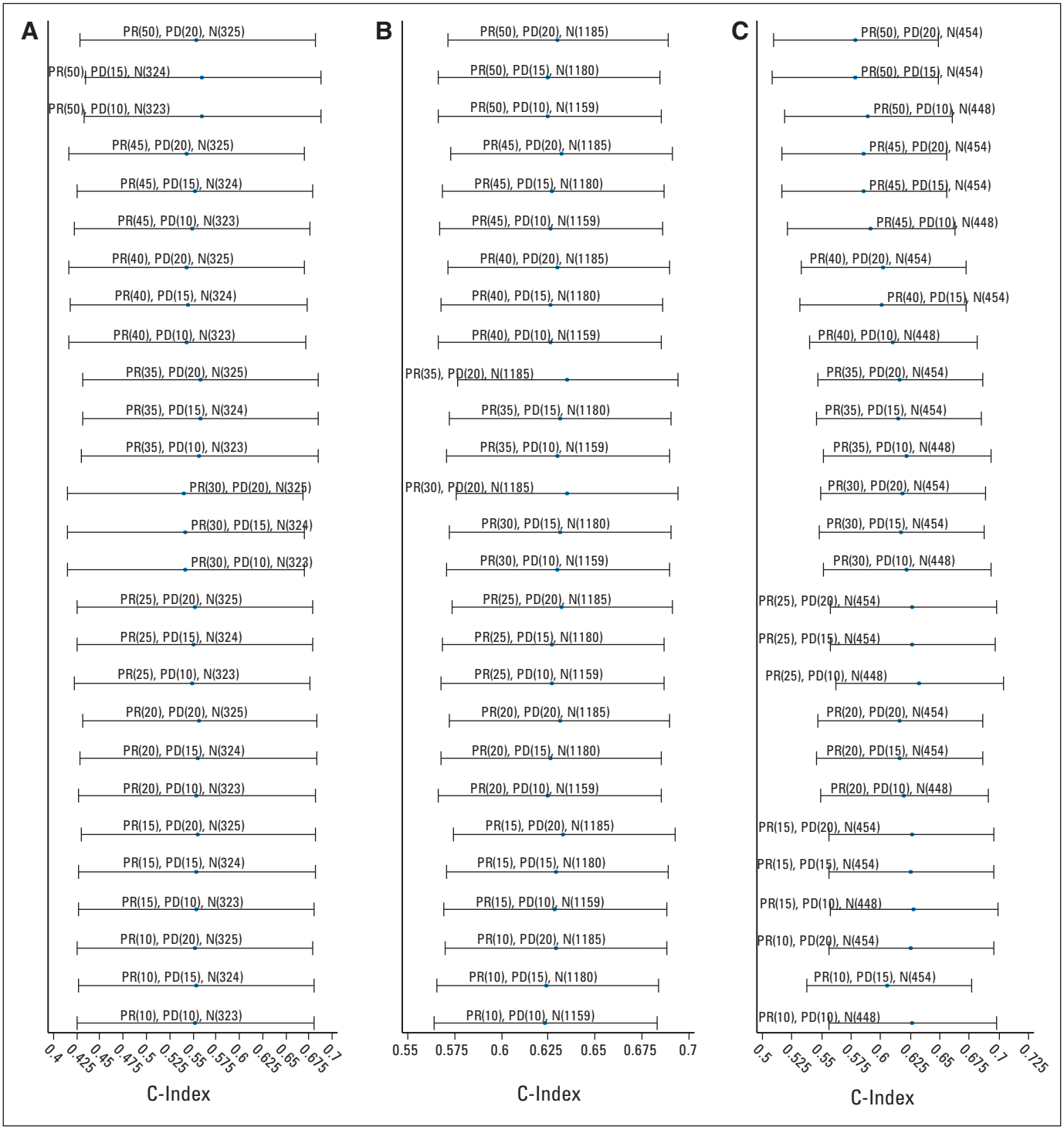

Fig 2. Forest plots of the concordance indices (c-indices) and associated 95% CIs for (A) breast cancer, (B) non–small-cell lung cancer, and (C) colorectal cancer using alternate partial response (PR) and progressive disease (PD) cut points for overall survival with the trichotomous tumor response metric on the training data sets. N(x), number of patients evaluable for the respective analysis.

METHODS

RECIST 1.1 Data Warehouse

The RECIST database includes cycle-by-cycle and lesion-by-lesion tumor measurements from 8,062 patients who were enrolled between 1993 and 2005 into 13 phase III trials in metastatic breast cancer, NSCLC, and colorectal cancer. The details of the trials and the disease assessment schedules can be found in the original report.[3] The data used in the analysis included 5,480 patients as shown in the CONSORT diagram (Fig 1). A total of 1,219 patients with breast cancer, 3,033 with NSCLC, and 1,228 with colorectal cancer were included in the final analysis, with 3,409 deaths (817

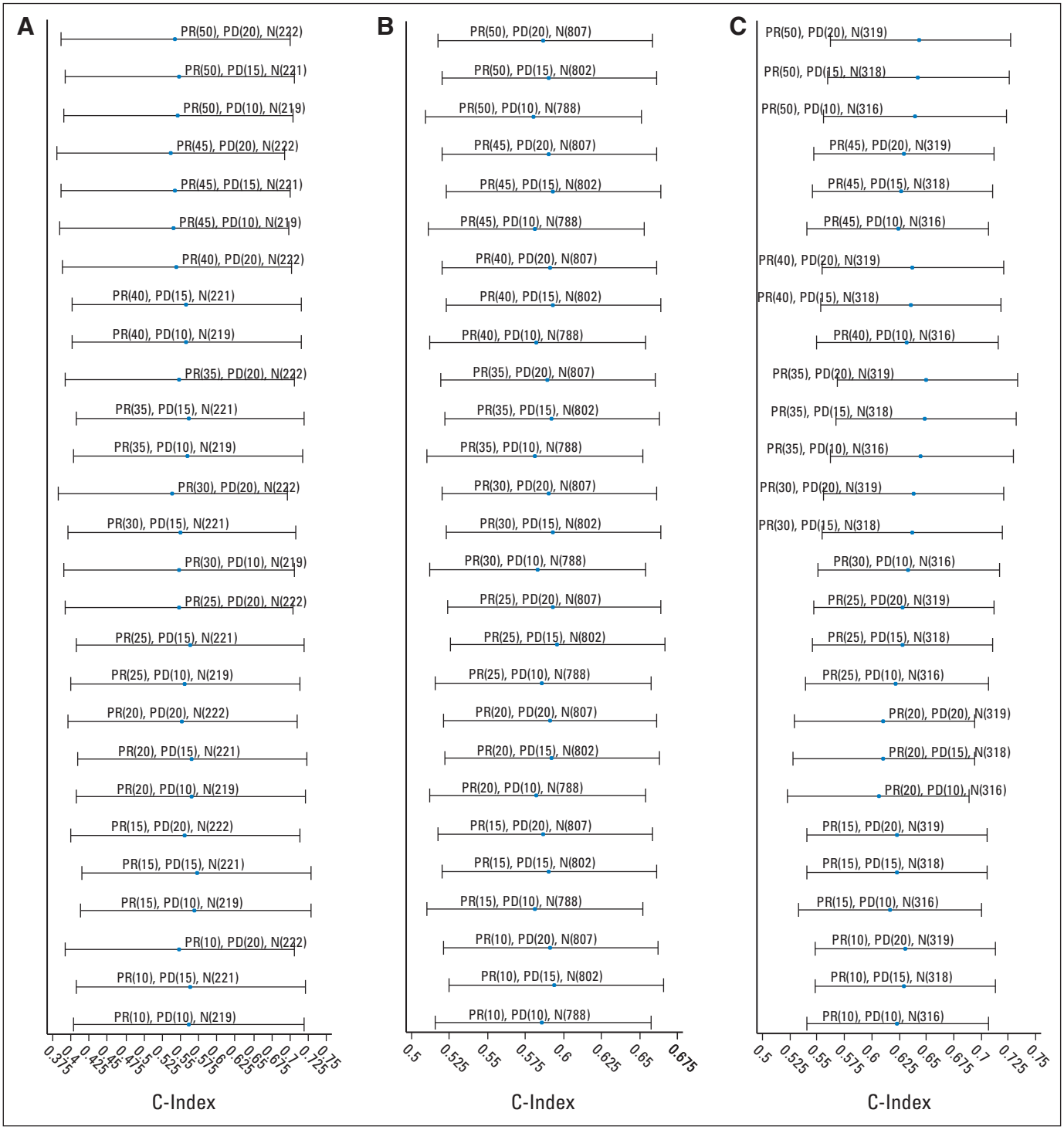

Fig 3. Forest plots of the concordance indices (c-indices) and the associated 95% CIs for (A) breast cancer, (B) non–small-cell lung cancer, and (C) colorectal cancer using alternate partial response (PR) and progressive disease (PD) cut points for overall survival with the trichotomous tumor response metric on the validation data sets. N(x), number of patients evaluable for the respective analysis.

breast cancer, 1,928 NSCLC, 664 colorectal cancer) and 3,428 progression events (762 breast cancer, 1,911 NSCLC, 755 colorectal cancer), with six patients having PD at the time of death (four breast cancer; two NSCLC). Of the progression events, 1,015 were the result of the growth of target lesions (ie, measurements available at the time of disease progression), 246 were the result of the occurrence of new lesions (ie, no measurements on target lesions recorded), 122 were the result of progression from nontarget lesions (ie, no measurements on target lesions recorded), and 67 were the result of both growth of new lesions and progression from nontarget lesions (ie, no measurements on target lesions recorded). In addition, 1,978 patients had PD documented by more than one criterion (measurements available at the time of PD).

Table 1. Cox Proportional Hazards Model Results for OS from the 12- and 24-Week Landmark Analyses for the Three Categorical Metrics Using the Training and Validation Data Sets, by Tumor Type

	12-Week								24-Week							
	Training				Validation				Training				Validation			
Metric	C-Index	95% CI	HR	Model *P*	C-Index	95% CI	HR	Model *P*	C-Index	95% CI	HR	Model *P*	C-Index	95% CI	HR	Model *P*
Breast cancer		n = 325				n = 221				n = 188				n = 112		
DiTR	0.54	0.42 to 0.67		.11	0.53	0.37 to 0.69		**.04**	0.63	0.47 to 0.79		.11	0.61	0.39 to 0.83		**.04**
SD/PD *v* CR/PR*			1.27				1.06				2.33				1.86	
TriTR	0.54	0.42 to 0.67		.09	0.54	0.38 to 0.69		**.007**	0.65	0.49 to 0.81		.09	0.60	0.38 to 0.82		**.007**
SD *v* CR/PR*			1.16				0.92				1.55				1.16	
PD *v* CR/PR*			1.63				2.15				3.98				3.15	
DCR	0.54	0.41 to 0.67		.06	0.55	0.39 to 0.70		**.003**	0.62	0.46 to 0.78		.06	0.57	0.35 to 0.79		**.003**
PD *v* CR/PR/SD*			1.55				2.23				3.59				3.04	
NSCLC		n = 1,183				n = 806				n = 260				n = 185		
DiTR	0.60	0.54 to 0.66		**< .001**	0.57	0.50 to 0.64		**< .001**	0.64	0.52 to 0.77		**< .001**	0.61	0.46 to 0.75		**< .001**
SD/PD *v* CR/PR*			1.86				1.43				2.09				2.72	
TriTR	0.63	0.57 to 0.69		**< .001**	0.59	0.52 to 0.66		**< .001**	0.68	0.55 to 0.80		**< .001**	0.63	0.49 to 0.78		**< .001**
SD *v* CR/PR*			1.44				1.21				1.04				1.48	
PD *v* CR/PR*			3.68				2.54				2.71				3.35	
DCR	0.62	0.56 to 0.68		**< .001**	0.58	0.51 to 0.66		**< .001**	0.67	0.55 to 0.80		**< .001**	0.64	0.49 to 0.78		**< .001**
PD *v* CR/PR/SD*			3.14				2.35				2.67				2.98	
Colorectal cancer		n = 454				n = 319				n = 298				n = 232		
DiTR	0.58	0.51 to 0.65		**< .001**	0.64	0.55 to 0.72		**< .001**	0.62	0.52 to 0.72		**< .001**	0.64	0.54 to 0.74		**< .001**
SD/PD *v* CR/PR*			1.82				1.64				2.32				1.88	
TriTR	0.62	0.55 to 0.69		**< .001**	0.64	0.56 to 0.72		**< .001**	0.66	0.56 to 0.76		**< .001**	0.65	0.55 to 0.74		**< .001**
SD *v* CR/PR*			1.56				1.46				1.68				1.43	
PD *v* CR/PR*			3.99				3.16				3.95				3.31	
DCR	0.58	0.51 to 0.65		**< .001**	0.63	0.55 to 0.71		**< .001**	0.63	0.53 to 0.73		**< .001**	0.65	0.55 to 0.75		**< .001**
PD *v* CR/PR/SD*			3.10				2.56				3.29				2.83	

NOTE. Bold font indicates *P* values ≤ .05. 95% CIs refer to CIs for C-index values.

Abbreviations: C-index, concordance index; CR, complete response; DCR, disease control rate; DiTR, dichotomous tumor response; HR, hazard ratio; NSCLC, non–small-cell lung cancer; OS, overall survival; PD, progressive disease; PR, partial response; SD, stable disease; TriTR, trichotomous tumor response.

*Reference category.

Data Analyses

Within each tumor type, data were randomly split 60:40 into training and validation data sets, stratified by survival and progression status and whether the observed assessments were within 2 weeks of protocol expected assessments based on a sliding window (closest available assessment within a ± 2-week window). A landmark analysis approach was used at the 12- and 24-week time points. Lesions that were consistently measured at all assessments up until 12 or 24 weeks or the closest available assessment within a ± 2-week window were used, as was done in studies by Hillman et al[12] and An et al.[11] In particular, the 12- and 24-week landmark analyses included patients who were alive and progression-free at 12 or 24 weeks with at least one postbaseline measurement within the first 12 or 24 weeks or who had PD within the landmark time window (Appendix Fig A1, online only). Although prior work suggests that earlier time points (eg, at 8 weeks) may be reasonable phase II end points,[8] the studies in this data warehouse have an assessment schedule of every 3 or 4 weeks, which is consistent with a 6-, 12-, or 24-week

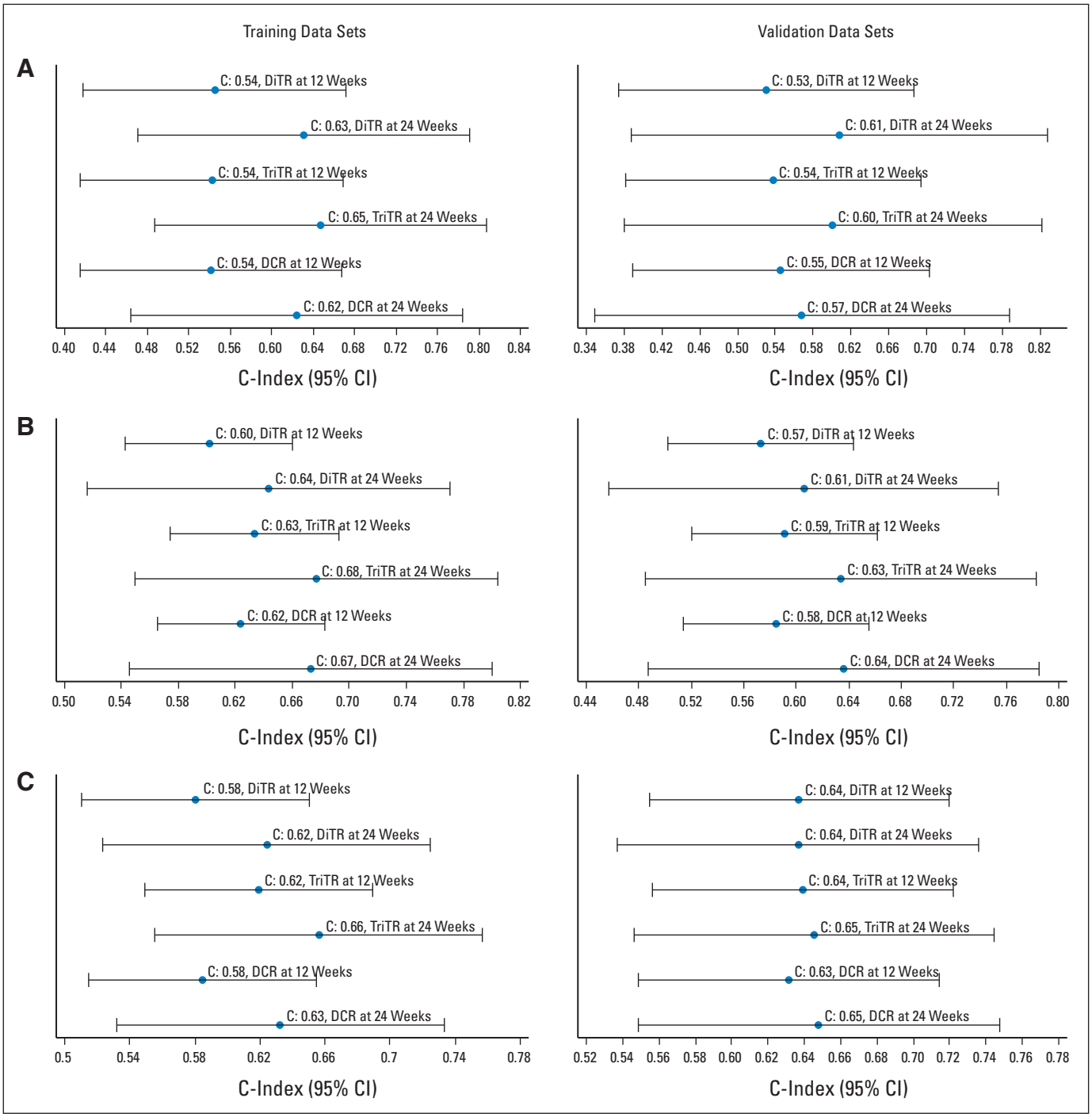

Fig 4. Forest plots of the concordance indices (c-indices) and the associated 95% CIs for (A) breast cancer, (B) non–small-cell lung cancer, and (C) colorectal cancer from the 12- and 24-week landmark analyses for the three categorical metrics using the training and validation data sets. DCR, disease control rate; DiTR, dichotomous tumor response; TriTR, trichotomous tumor response.

common assessment time point, thus making the 12- and 24-week analyses logistically relevant. Appendix Table A1 (online only) lists the numbers of patients with postbaseline assessments until 24 weeks, and Figure 1 lists the numbers of patients evaluable for the 12- and 24-week landmark analyses.

In all, 27 pairs of unique cut points for PR and PD categorization were considered: PR (10% to 50% decrease by 5% increments) and PD (10% to 20% increase by 5% increments), in which the pair (30%, 20%) corresponds to the RECIST cut points. Determination of PD was made on the basis of progression of target lesions (by using these alternate cut points) and included progression based on nontarget lesions and/or occurrence of new lesions. For each possible cut point, three metric classifications were considered: DiTR (CR/PR *v* SD/PD), TriTR (CR/PR *v* SD *v* PD), and DCR (CR/PR/SD *v* PD). Forest plots were used to visually present these results.

At each landmark time point of 12 and 24 weeks for each tumor type, by using each of the three metric classifications and for each of the 27 cut point pairs, a Cox proportional hazards model for OS was fit on the training data set. The models were stratified by study and number of consistent lesions (fewer than three and three or more), defined as the lesions that were consistently assessed at baseline and at all follow-up assessments. All models were adjusted for the average baseline tumor size defined as the sum of one-dimensional baseline tumor measurements of consistent lesions divided by the number of consistent lesions. The concordance index (c-index) along with 95% CI was used as the criterion to select the optimal cut point pair for each tumor type, for each time point, and for each metric classification.[13-15] Specifically, the 95% CIs for the difference in the c-indices were computed to determine whether they included zero.[16] The c-indices in the context of time-to-event data are a measure of model discrimination computed as the fraction of all evaluable pairs that are concordant. They range from 0.0 to 1.0, where 0.5 indicates a completely random prediction,1.0 indicates perfect prediction, and values less than 0.5 indicate prediction in the opposite direction (ie, higher risk scores indicate better survival). These findings were then confirmed in the validation data set; specifically, the c-indices were obtained by comparing the observed OS in the validation data set with the predicted OS obtained by fitting the model estimates from the training data set to the validation data set. *P* values $\leq .05$ were considered statistically significant.

RESULTS

The median number of consistent lesions by tumor type were one for breast cancer (range, one to seven), two for colorectal cancer (range, one to nine), and two for lung cancer (range, one to 10). The distribution of the average baseline tumor size (measured in millimeters) is given in Appendix Figure A2 (online only). NSCLC and colorectal cancer tumors had higher mean and median average baseline tumor size values compared with breast cancer tumors. The distribution of the numbers of patients that fall into the categories of PR, SD, and PD based on the alternate cut points for PR and PD is depicted in Appendix Figure A3 (online only) for the 12-week data set for the training and validation sets for each tumor type. As expected, a larger percentage of patients are categorized under SD as the cut point for PR progressively increases from 10% to 50%. The alternate cut points for PR and PD provided no meaningful improvement in prediction for outcome (OS) for any of the three metric classifications considered for the three tumor types at either landmark time point because all of the 95% CIs for the differences in c-indices contained zero. These results were confirmed in the respective external validation of the c-indices. Figures 2 and 3 depict example forest plots of c-indices (and the associated 95% CIs) obtained from the 12-week landmark analysis for the TriTR metric by using alternate PR and PD cut points in the training and validation sets for each tumor type (the plots for the DiTR and the DCR metrics were similar). Thus, the published RECIST cut points for PR and PD of 30% and 20% demonstrated predictive ability for OS similar to that of the alternate PR and PD cut points for all metrics, all tumor types, and at both time points (12 and 24 weeks).

Next, we assessed the predictive ability of the three metric classifications by using the published RECIST cut points. The Cox proportional hazards model results and forest plots of c-indices (95% CIs) for the three metrics and two time points that use RECIST cut points on the training sets for each tumor type are given in Table 1 and Figure 4. All of the metrics were statistically significant for OS prediction. Although the point estimates for c-index for the TriTR metrics at 24 weeks were marginally better for all tumor types, the 95% CIs for the differences included zero. The c-indices for breast cancer models were much lower than those for colorectal cancer and NSCLC. The point estimates for the externally validated c-indices were lower than those obtained from the training models for all three tumor types across the three metrics for the 12- and 24-week time points: 0.53 to 0.58 for breast cancer (training, 0.54 to 0.65), 0.56 to 0.64 for colorectal cancer (training, 0.58 to 0.66), and 0.58 to 0.63 for NSCLC (training, 0.60 to 0.68; Table 2).

For comparison, we also fit a Cox model with progression status as a time-dependent covariate by using data available over the entire follow-up (ie, no landmark analysis and including all patients who had a baseline and postbaseline assessment) for each tumor type (without splitting into training or validation data sets) to serve as a benchmark for the theoretically "best" model. It is important to note that this approach does not suggest an obvious metric nor does it allow for early assessment. The time-dependent c-indices[15] for breast cancer, colorectal cancer, and NSCLC are 0.60 (95% CI, 0.57 to 0.62), 0.66 (95% CI, 0.63 to 0.68), and 0.66 (95% CI, 0.64 to 0.67), respectively, which is within the range of the c-indices for the other metrics considered for colorectal cancer and NSCLC but is higher for breast cancer compared with the categorical metrics.

Kaplan and Meier curves of subsequent OS based on the response status at 12 or 24 weeks (without splitting into training or validation data sets) indicate that patients with PD did worse compared with those with non-PD, with some separation between the SD and the PR categories for colorectal cancer and NSCLC, respectively (Fig 5). Comparing the response status categories of CR/PR to SD, SD to PD, and CR/PR to PD revealed no statistically significant differences in OS between CR/PR and SD except for colorectal cancer at 12 and 24 weeks, and NSCLC at 12 weeks. Survival for patients with CR/PR (and SD) was statistically significantly different from PD for all three tumor types for both time points. All three metrics except the DiTR metric at 12 weeks for breast cancer were statistically significantly associated with subsequent OS.

Table 2. Validation Data Set C-Indices Generated by Using the Model Estimates From the Training Data Set

	Breast Cancer		Colorectal Cancer		NSCLC	
Metric	12-Week	24-Week	12-Week	24-Week	12-Week	24-Week
DiTR (CR/PR *v* SD/PD)	0.53	0.57	0.56	0.62	0.58	0.60
TriTR CR/PR *v* SD *v* PD)	0.54	0.58	0.63	0.64	0.60	0.63
DCR (CR/PR/SD *v* PD)	0.54	0.57	0.62	0.64	0.60	0.63

Abbreviations: C-index, concordance index; CR, complete response; DCR, disease control rate; DiTR, dichotomous tumor response; NSCLC, non–small-cell lung cancer; PD, progressive disease; PR, partial response; SD, stable disease; TriTR, trichotomous tumor response.

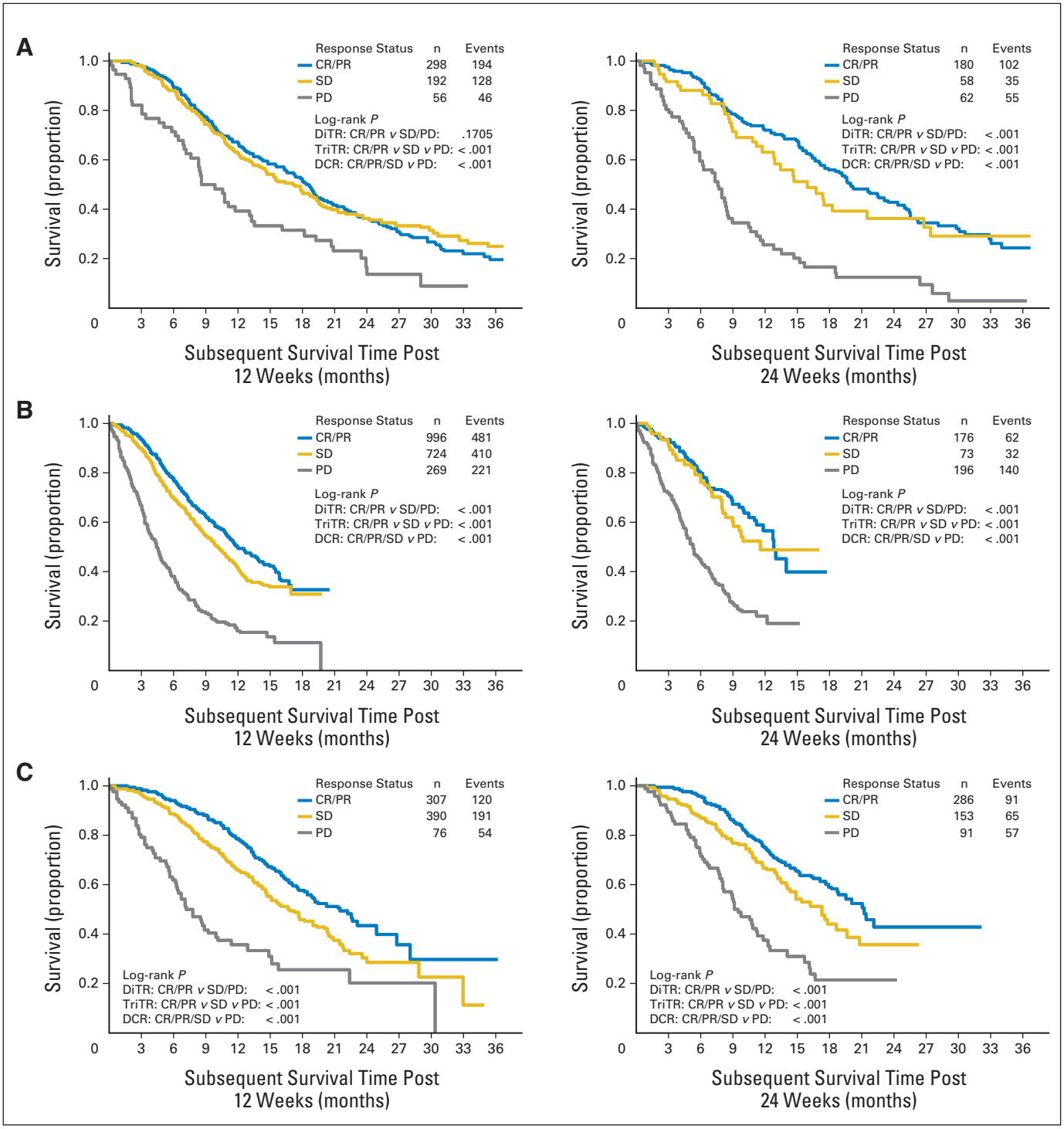

Fig 5. Kaplan-Meier curves for subsequent overall survival for (A) breast cancer, (B) non–small-cell lung cancer, and (C) colorectal cancer using the response status at 12 (left panels) and 24 (right panels) weeks. CR, complete response; DCR, disease control rate; DiTR, dichotomous tumor response; PD, progressive disease; PR, partial response; SD, stable disease; TriTR, trichotomous tumor response.

DISCUSSION

The RECIST criteria are simple, clinically relevant, and easy to implement. Although alternate categorical end points based on published RECIST cut points have been explored, this is, to the best of our knowledge, the first systematic assessment of potential alternate cut points for PR and PD besides the RECIST categorization. Similarly, although several prior studies have suggested the potential for additional improvement on RECIST-based response metrics, such findings have not been validated by using data from additional trials or by

using a large database. In addition, much of the previous work has focused on statistical significance (through the use of hazard ratios and *P* values) for selection of metrics compared with discriminatory ability (through the use of c-index) in this analysis.[17] Our study sought to assess both of these aims by using the RECIST 1.1 data warehouse, the largest assembled database on record containing lesion-by-lesion measurements over time across tumor types. A total of 27 pairs of unique cut points for PR and PD categorization were considered, in which the pair (30%, 20%) corresponds to the RECIST categorization. Our analysis demonstrates that alternate cut points have OS predictive ability similar to that of the published RECIST cut points for the three metric classifications considered and across all tumor types and at both landmark analysis time points (12 and 24 weeks). There was a statistically significant difference between CR/PR and SD status for patients with colorectal cancer based on the 12- and 24-week landmark analyses, thus lending support to the use of the TriTR metric as reported previously.[7,11] The RECIST definition for SD does not distinguish patients with true SD from those with a minor increase or minor decrease. The TriTR metric puts SD patients into their own category unlike the DiTR or the DCR metrics. However, regardless of tumor type, the three investigated categorical metrics (TriTR, DiTR, and DCR) demonstrated similar predictive performance for OS. Although the 24-week metrics had slightly improved point estimates for c-indices compared with the 12-week metrics, the 95% CIs for the difference in the c-indices for the different metrics included zero, indicating that assessing these metrics at 12 weeks might be sufficient. None of the metrics did particularly well for breast cancer. The CIs for some of the c-index comparisons were wide, suggesting that the results be interpreted with caution; however, most were symmetrical about zero, indicating sufficient overlap.

Several limitations of this work need special attention. First, these analyses used data from only three tumor types from trials that did not use modern regimens, including targeted agents, which clearly influence the biology of cancers. Moreover, these trials were conducted a decade ago or more, and imaging technology has clearly improved in the interim. Second, a complete case analysis was performed, excluding those lesions that were not measured consistently over time. Third, the analysis was limited by the fact that RECIST progression definitions were used in trials as criteria to end therapy; specifically, alternative cut points examined were constrained in that there were no measurements beyond RECIST progression. Thirteen percent of patients had no measurements documented at the time of PD resulting from progression from new lesions, clinical deterioration, or inability to document necrotic tumors, which present with the same dimensions but are likely responding to therapy. In addition, patients with breast cancer for whom the determination of response was done by clinical evaluation (physical examination) alone (ie, palpable lymph nodes) were excluded because these were not as reliable as response determined by standard imaging criteria.

The choice of the end point and the use of landmark analysis also needs to be considered. The use of OS as the end point for evaluating the utility of these metrics in the first-line setting is obscured by subsequent therapies, as demonstrated by the modest c-indices for all of the metrics considered. This leads to the question of whether alternate end points such as PFS should have been considered. The use of PFS to assess the utility of these metrics would alleviate the concerns of influence from later-line therapies; however, this analysis used a landmark approach whereby only patients alive and progression-free by 12 or 24 weeks or who had PD within the landmark time window were included. Given the relatively short median PFS in this advanced disease population (median PFS for colorectal cancer, 8.2 months; breast cancer, 8.0 months; and NSCLC, 5.7 months), all PFS events before the landmark time points are excluded, thus limiting the scope of using PFS as an end point. An alternative strategy of using a combination of landmark analysis and a time-dependent component for the first 12 or 24 weeks to account for early deaths, progressions, dropouts, and so on, or using continuous tumor size metrics based on longitudinal tumor size models is worth exploring.[18]

In summary, this analysis was undertaken to better understand the utility of the RECIST-based response metrics. No alternative cutoffs or alternative categorical metrics that were investigated in this work were better than the published RECIST standards. Ongoing work is examining imputation methods for the missing lesion measurements, as well as exploring continuous tumor measurement based metrics.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The author(s) indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Financial support: Sumithra J. Mandrekar, Ming-Wen An
Administrative support: Sumithra J. Mandrekar, Daniel J. Sargent
Manuscript writing: All authors
Final approval of manuscript: All authors

REFERENCES

1. Kola I, Landis J: Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3:711-715, 2004

2. Therasse P, Arbuck SG, Eisenhauer EA, 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 92:205-216, 2000

3. Bogaerts J, Ford R, Sargent D, et al: Individual patient data analysis to assess modifications to the RECIST criteria. Eur J Cancer 45:248-260, 2009

4. Lavin PT: An alternative model for the evaluation of antitumor activity. Cancer Clin Trials 4:451-457, 1981

5. Dhani N, Tu D, Sargent DJ, et al: Alternate endpoints for screening phase II studies. Clin Cancer Res 15:1873-1882, 2009

6. Pivot X, Thierry-Vuillemin A, Villanueva C, et al: Response rates: A valuable signal of promising activity? Cancer J 15:361-365, 2009

7. Heun JM, Grothey A, Branda ME, et al: Tumor status at 12 weeks predicts survival in advanced colorectal cancer: Findings from NCCTG N9741. Oncologist 16:859-867, 2011

8. Lara PN Jr, Redman MW, Kelly K, 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 26:463-467, 2008

9. Mandrekar SJ, Qi Y, Hillman SL, et al: Endpoints in phase II trials for advanced non-small cell lung cancer. J Thorac Oncol 5:3-9, 2010

10. Foster NR, Qi Y, Shi Q, et al: Tumor response and progression-free survival as potential surrogate endpoints for overall survival in extensive stage small-cell lung

cancer: Findings on the basis of North Central Cancer Treatment Group trials. Cancer 117:1262-1271, 2011

11. 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

12. Hillman SL, An MW, O'Connell MJ, et al: Evaluation of the optimal number of lesions needed for tumor evaluation using the Response Evaluation Criteria in Solid Tumors: A North Central Cancer Treatment Group investigation. J Clin Oncol 27:3205-3210, 2009

13. Harrell FE Jr, Lee KL, Mark DB: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361-387, 1996

14. Harrell FE Jr, Califf RM, Pryor DB, et al: Evaluating the yield of medical tests. JAMA 247: 2543-2546, 1982

15. Antolini L, Boracchi P, Biganzoli E: A time-dependent discrimination index for survival data. Stat Med 24:3927-3944, 2005

16. Newson RB: Comparing the predictive powers of survival models using Harrell's C or Somers' D. Stata J 10:339-358, 2010

17. 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

18. Claret L, Gupta M, Han K, 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 31:2110-2114, 2013

Acknowledgment

We thank the other members of the Response Evaluation Criteria in Solid Tumors (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. 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 Organisation for Research and Treatment of Cancer Breast Cancer Group and Gastrointestinal Group; Erasmus University Medical Center, Rotterdam, Netherlands; Genentech; Roche; and sanofi-aventis.

Appendix

Table A1. No. of Patients With Postbaseline Assessments Within First 24 Weeks

Disease	Maximum Postbaseline Assessment								
	1	2	3	4	5	6	7	8	9
	No. of Patients								
Breast cancer (n = 1,218)*	119	252	495	174	58	47	49	22	2
NSCLC (n = 3,033)	587	562	1,295	467	76	41	5	0	0
Colorectal cancer (n = 1,227)*	169	176	320	548	14	0	0	0	0

Abbreviation: NSCLC, non–small-cell lung cancer.
*One patient with postbaseline assessment occurring after 24 weeks.

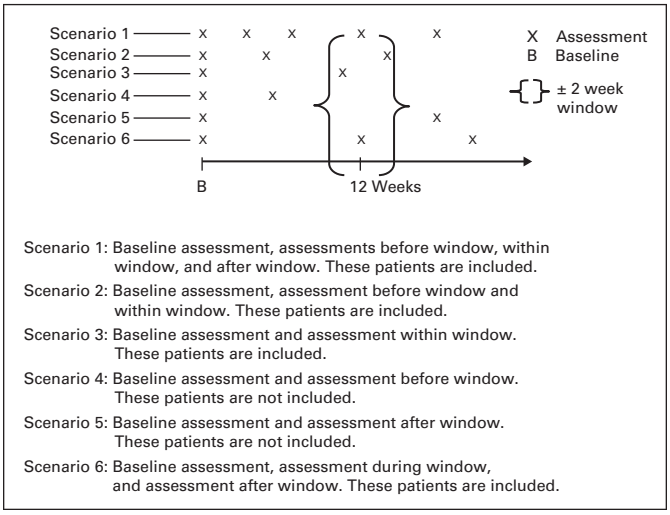

Fig A1. Criteria for selection of patients for the 12-week landmark analysis.

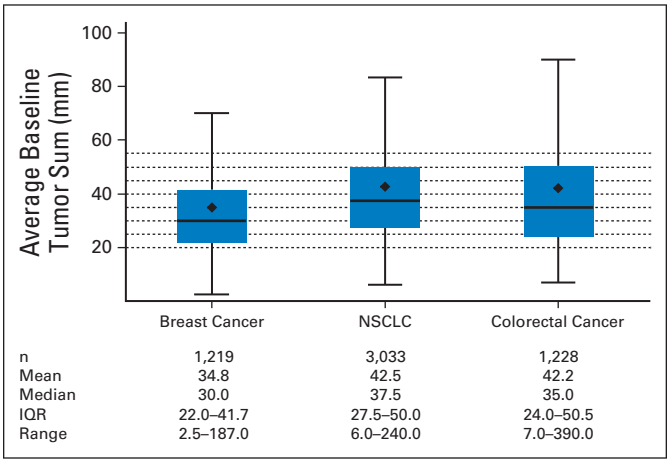

Fig A2. Distribution of average baseline tumor size in millimeters (excluding outliers*), by tumor type. (*) Values outside Q3 + 1.5 interquartile range (IQR), and Q1 − 1.5 IQR are not shown. NSCLC, non–small-cell lung cancer.

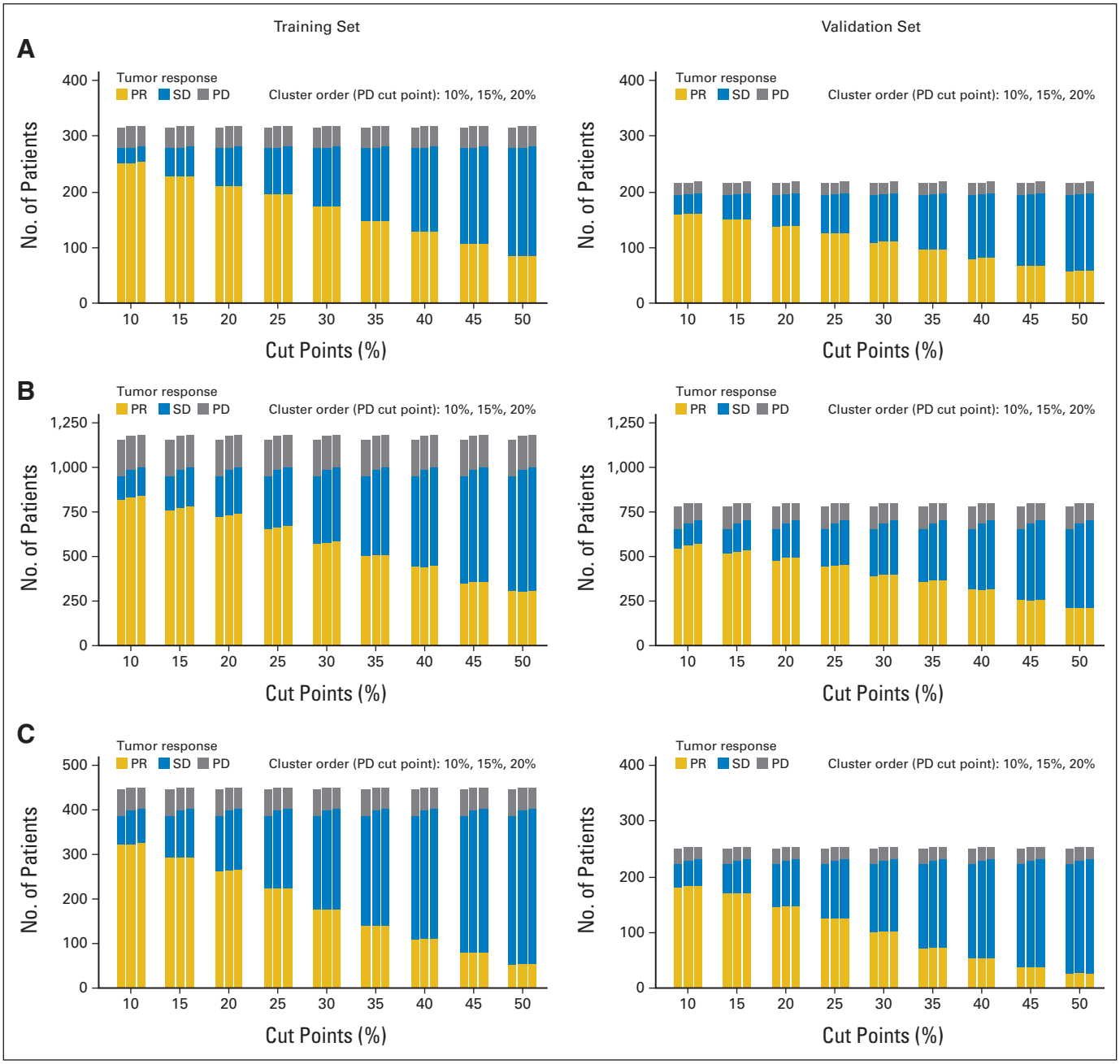

Fig A3. Frequency distribution of the numbers of patients with (A) breast cancer, (B) non–small-cell lung cancer, and (C) colorectal cancer, by tumor type, on the basis of the 12-week data set that fall under the categories of partial response (PR), stable disease (SD), and progressive disease (PD) based on the different alternative cut points for PR and PD.