Citation: Clin Transl Sci (2016) 9, 43–50; doi:10.1111/cts.12384

ARTICLE

Comparative Effects of CT Imaging Measurement on RECIST End Points and Tumor Growth Kinetics Modeling

CH Li[1,2], RR Bies[1,2,3], Y Wang[4], MR Sharma[3,5], S Karovic[5], L Werk[3,6], MJ Edelman[3,7], AA Miller[3,8], EE Vokes[3,5], A Oto[3,5], MJ Ratain[3,5], LH Schwartz[3,9] and ML Maitland[3,5,*]

Quantitative assessments of tumor burden and modeling of longitudinal growth could improve phase II oncology trials. To identify obstacles to wider use of quantitative measures we obtained recorded linear tumor measurements from three published lung cancer trials. Model-based parameters of tumor burden change were estimated and compared with similarly sized samples from separate trials. Time-to-tumor growth (TTG) was computed from measurements recorded on case report forms and a second radiologist blinded to the form data. Response Evaluation Criteria in Solid Tumors (RECIST)-based progression-free survival (PFS) measures were perfectly concordant between the original forms data and the blinded radiologist re-evaluation (intraclass correlation coefficient = 1), but these routine interrater differences in the identification and measurement of target lesions were associated with an average 18-week delay (range, −20 to 55 weeks) in TTG (intraclass correlation coefficient = 0.32). To exploit computational metrics for improving statistical power in small clinical trials will require increased precision of tumor burden assessments.

Clin Transl Sci (2016) **9**, 43–50; doi:10.1111/cts.12384; published online on 21 January 2016.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

✔ Improvements in digital CT imaging offer the potential to improve the efficiency of phase II oncology clinical trials. However, in retrospective comparisons, quantitative assessments of tumor burden have not consistently proved superior to more conventional categorical time-to-event methods.

WHAT QUESTION DID THIS STUDY ADDRESS?

✔ Theoretically, quantitative assessments should be clearly superior. This study explored potential explanations for why prior comparative analyses have supported continued use of categorical end points like RECIST-based PFS.

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE

✔ We found that the PFS end point is robust to routine interrater differences in tumor burden measurement, but the measurement imprecision tolerated by RECIST caused significant discordance in the model-based end point of TTG.

HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY AND THERAPEUTICS

✔ Pharmacometrics methods offer innovative strategies to improve conduct of oncology clinical trials. To exploit these methods will require changes in how tumor burden measurements are acquired and transmitted in the course of early phase clinical trials.

One goal for computational modeling methods in cancer drug development is to enable evaluation of new therapeutics with available technology, in fewer patients, observed on treatment for shorter periods of time. One strategy to achieve this goal has been to apply computational modeling to the longitudinal growth of solid tumors in populations of patients and *in silico* simulation of clinical trials.[1–5] The ultimate goal of this effort is to improve the efficiency of cancer drug clinical development.[6–9]

Non-small cell lung cancer (NSCLC) is the leading cause of cancer-related death in the United States and an increasingly common cause of death globally.[10] Because NSCLC remains an important area of unmet need in cancer therapeutics, one of the first major investigations of computational modeling of longitudinal tumor growth to determine the relationship between early changes in tumor size and overall survival was conducted in NSCLC.[3] Clinical trial simulations used a model of overall survival in metastatic disease based on a longitudinal tumor growth model developed with data from 3,400 patients from four phase III clinical trials submitted to the US Food and Drug Administration (FDA).[3] These studies of bevacizumab, docetaxel, erlotinib, and

[1]Indiana University School of Medicine, Indianapolis, Indiana, USA; [2]Indiana Clinical and Translational Sciences Institute (CTSI), Indianapolis, Indiana, USA; [3]Alliance for Clinical Trials in Oncology, Boston, Massachusetts, USA; [4]Office of Clinical Pharmacology, US Food and Drug Administration, Silver Spring, Maryland, USA; [5]University of Chicago Medicine and Biological Sciences, Chicago, Illinois, USA; [6]Duke University, Durham, North Carolina, USA; [7]University of Maryland Greenebaum Cancer Center, School of Medicine, Baltimore, Maryland, USA; [8]Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA; [9]Columbia University College of Physicians and Surgeons, New York, New York, USA. *Correspondence: ML Maitland (mmaitlan@medicine.bsd.uchicago.edu)
Received 10 November 2015; accepted 16 December 2015; published online on 21 January 2016. doi:10.1111/cts.12384

pemetrexed led to the development of the model, derived from the sum of the longest dimension measurements of tumors by computed tomography (CT) imaging as recorded in study case report forms (CRFs). Estimations of the change in tumor size from baseline to 8 weeks of treatment (the tumor size ratio) proved an important predictor of overall survival. We undertook an independent investigation of available archived NSCLC tumor measurement data to expand on this initial study and to assess the robustness with which modeling and simulation with these data could support decision-making at the phase II to phase III transition in drug development.[1]

Another potential benefit of quantitative analysis of NSCLC tumor burden would be to redesign phase II trials to randomize fewer patients and have shorter observation periods than required for determining progression-free survival (PFS).[11–17] Previously suggested simple strategies in NSCLC have entailed measuring the median tumor size at 8 weeks for randomly assigned treatment arms,[7,18,19] or calculating the fraction of patients without progressive disease at landmark timepoints.[20] Model-based strategies have had limited testing and require validation. In studies of colorectal cancer therapy and survival outcomes, some have found advantages to continuous tumor measurement metrics, while others have not.[21–23]

We sought to assess and refine the published FDA longitudinal tumor size model for NSCLC using archived tumor measurement data so that modeling and simulation might lead to smaller, quicker early phase trials for testing new treatments for NSCLC. We intended to evaluate the power of smaller clinical trials with novel end points to detect evidence of anticancer drug treatment effects with archived CRFs from three randomized clinical trials sponsored by the US National Cancer Institute. The largest data set was sufficient for evaluation of qualitative, time-to-event end points but obviously useless for quantitative metrics. The other two data sets had inconsistencies between the measurements of tumor burden recorded on CRFs and re-measurements of tumor burden from the original CT images performed by an independent radiologist. These findings are likely to be common to historical and current solid tumor trial data sets. This study demonstrated that features of historical data on tumor burden measurement could bias comparisons between continuous measurement and categorical strategies for improving treatment evaluations. Our findings suggest that comparing conventional and computational methods on historical data is a key obstacle to progress. The simple, prospective incorporation of more precise measurement of tumor burden on CT imaging should enable computational modeling methods to clearly surpass Response Evaluation Criteria in Solid Tumors (RECIST)-based methods in assessment of treatment effects.

METHODS

Patients

Archived CRFs were available from 857 patients enrolled in three National Cancer Institute-supported studies by Cancer and Leukemia Group B (CALGB), now called the Alliance for Clinical Trials in Oncology (**Table 1**). CALGB 9730[23] was a phase III randomized trial that compared single-agent paclitaxel with combination carboplatin/paclitaxel, CALGB 30203[25] was a randomized phase II trial that evaluated eicosanoid modulation in standard first-line cytotoxic therapy regimens, and CALGB 30303[26] was a phase II randomized study of dose-dense docetaxel and cisplatin administered every 2 weeks with growth factor supportive therapy. The inclusion and exclusion criteria of the trials were previously published.[24–26]

Table 1 Three US National Cancer Institute-sponsored studies conducted by the Cancer and Leukemia Group B

CALGB study	Treatment	No. of subjects enrolled	No. of subjects treated & eligible	Dates of accrual
9730	P vs. PCB	561	561	10/1997–12/2000
30203	GCb + ZiCe/both	140	134	12/2003–9/2004
30303	DC +/- BNP	160	151	8/2004–3/2006

BNP, BNP7787; CALGB, Cancer and Leukemia Group B; DC, docetaxel and cisplatin; GCb, carboplatin and gemcitabine; P, paclitaxel; PCB, paclitaxel, carboplatin, and bevacizumab; ZiCe, zileuton celecoxib.

Original clinical trial data collection

Data relevant to reporting of the clinical trial results were captured on CRFs and entered into the CALGB digital databases. The coded, patient-level data were stored at the Core Statistical Facility for CALGB (Durham, NC, USA). Treatment response assessments were conducted according to the study protocols. The CALGB 9730 trial incorporated standard World Health Organization response criteria[27] based on imaging studies conducted every two cycles (6 weeks) as described.[24] For CALGB studies 30203 and 30303, the RECIST was used, and categorical responses were based on the sum of the longest unidimensional measurements of criteria-defined "target lesions."[28] CT imaging evaluations were conducted in all patients pretreatment, and at 6 and 12 weeks after treatment. Patients were removed from the studies for unacceptable toxicity or progression of disease. Patients who completed all study therapy were followed at minimum every 12 weeks thereafter. The target lesion and sum of the longest dimensions of target lesions measurements were captured on CRFs but not in the study database.

Tumor measurement collection

The retrospective access and analysis of these data was approved by the University of Chicago and Duke University Institutional Review Boards as consistent with the intentions of the original clinical trial consent documents.

Archived paper CRFs were obtained from storage, scanned, and saved as portable document format files. Tumor measurements from the portable document format files for CALGB 30203 and 30303 were manually extracted by a research assistant and entered into a tracking file and into the study databases simultaneously. The transcriptions were independently reviewed by two of the study authors (S.K. and C.L.) and inconsistencies were manually corrected. Individual patient tumor growth plots were inspected for atypical growth and response patterns. Aberrant plots were cross-verified with the original case report form portable document format and any additional data entry errors captured by this

review were corrected before modeling analyses were performed.

Tumor size and time-to-tumor growth modeling

Longitudinal tumor size trajectories (sum of longest tumor diameter) were analyzed with nonlinear mixed effect modeling software, NONMEM, version VII (GloboMax_LLC, Ellicott City, MD, USA) using Wings for NONMEM, version 7[29] and the model structure as described by Wang *et al.*[3] (see **Supplementary Materials** for details). This model used a combination of a linear growth function and an exponential shrinkage function to describe the tumor change with respect to baseline size (Eq. 1).

$$TS_i(t) = BASE_i e^{-SR_i t} + PR_i t \quad (1)$$

Where $TS_i(t)$ is the tumor size at time t for the i^{th} individual, $Base_i$ is the baseline tumor size, $SR_i(t)$ is the exponent tumor shrinkage rate constant, and $PR_i(t)$is the linear tumor growth rate constant. Tumor size changes were modeled using the first-order conditional estimation method with interaction. Between subject variability was assumed to be log-normally distributed and evaluated on baseline tumor size, tumor shrinkage rate, and tumor progression rate using an exponential model $P_i = P_{TV}\ x\ e^{\eta p}$ where P_i is the parameter estimate for the i^{th}individual and P_{TV} is the typical value for the parameter at the population level. Residual variability was also estimated using a proportional residual error model ($y_{ij} = \hat{y}_{ij}(1 + \varepsilon_{ij})$) where y_{ij} and $\hat{y}_{ij}$ represents the j^{th} observed tumor trajectory, and its corresponding model predicted tumor size.

The final model was examined using goodness-of-fit plots generating using R (version 2.13) based on the conditional weighted residuals distribution and the predicted vs. observed tumor size measurements at both the population and individual levels. The tumor size model was developed to evaluate data from both treatment arms individually as well as simultaneously on the combined data set. In addition to change in tumor size at 8 weeks, treatment effects on serial tumor measurements were also evaluated with time-to-tumor growth (TTG), as described by Claret *et al.*[23] More specifically, the rate of tumor growth (the differential equation dTSi/dt) was set to zero and the equation solved for time (see **Supplementary Materials** for details).

Modeling tumor burden measures from CALGB 30203 and the FDA sample

The parameter estimates for the linear growth rate and the treatment-related shrinkage rate in the CALGB trials differed from the originally published FDA sample. To determine whether the deviation of the parameter estimates was specific to the CALGB data collection, we extracted longitudinal tumor measurement data from patients with NSCLC treated with first-line platinum doublet therapy in the original FDA sample. One hundred three individual patients were selected from the platinum doublet treated patients on the FDA registration trials to match the baseline tumor size distribution of the 103 patients in CALGB 30203 based on Mahalanobis metric matching method.[30]

Blinded reevaluation of imaging data

To identify sources of variance between patient outcomes and the modeled tumor burden over time, we obtained the original sets of images from patients enrolled at one of the CALGB sites (University of Chicago) in studies 30203 and 30303. One radiologist, blinded to the original CRFs and radiology reports (coauthor A.O.) reviewed all of the baseline images and identified and measured all target lesions and measured them subsequently on all follow-up scans. PFS was determined by the time from initiation on-study until the date of the CT imaging at which, consistent with RECIST, is: the sum of the longest dimensions of target lesions increased by at least 20%; or the patient withdrew for clinical progression. One patient in this analysis had disease progression defined by development of a new lesion, and none had progression of nontarget lesions. To describe agreement between CRF and blinded evaluator-based measures for PFS and TTG in this sample, the intraclass correlation coefficient was calculated.

RESULTS

Data quality control

CRFs were reviewed for three randomized, controlled clinical trials of first-line therapy in NSCLC conducted by the CALGB (**Table 1**). CALGB 9730[23] was a phase III randomized trial that compared single-agent paclitaxel with combination carboplatin/paclitaxel. We discovered that CRFs from this study frequently included text notations of “no change” or “not available” rather than actual tumor size measurements on subsequent CT scans (**Supplementary Figure S1**). The data as entered were sufficient to determine the time of disease progression, but had too much missing data to be useful for validating the longitudinal tumor growth model and data from all 561 subjects were excluded.

CALGB 30203[24] was a randomized phase II trial that evaluated eicosanoid modulation in standard first-line cytotoxic therapy regimens, and CALGB 30303[25] was a phase II randomized study of dose-dense docetaxel and cisplatin administered every 2 weeks with growth factor supportive therapy. For the CALGB 30203 and 30303 trials, we applied the same standard for data inclusion as in the FDA model (at least a baseline measurement and measurements recorded at some subsequent timepoint). From 140 original cases in the CALGB 30203 trial and 160 in CALGB 30303, a total of 227 patients had data suitable for the analyses (**Figure 1**).

Longitudinal modeling of tumor growth in the CALGB 30203 and 30303 studies

Parameter estimates for sum of longest tumor dimensions at baseline (M_BASE), the treatment-effect/shrinkage rate (M_SR), and the linear tumor growth rate (M_PR) were determined and compared with the results of similar study arms from the original study (**Table 2**). Variance in parameter estimates increased as sample size was reduced from typical phase III to typical phase II size study arms. With a combination of both 30203 and 30303 trials, the model estimates of baseline tumor size, shrinkage rate, and progression rate were 8.1 cm, 0.025/week, and 0.059 cm/week, respectively. For example, a patient with an average baseline tumor size of 8.1 cm will, after 1 month, have the typical tumor burden

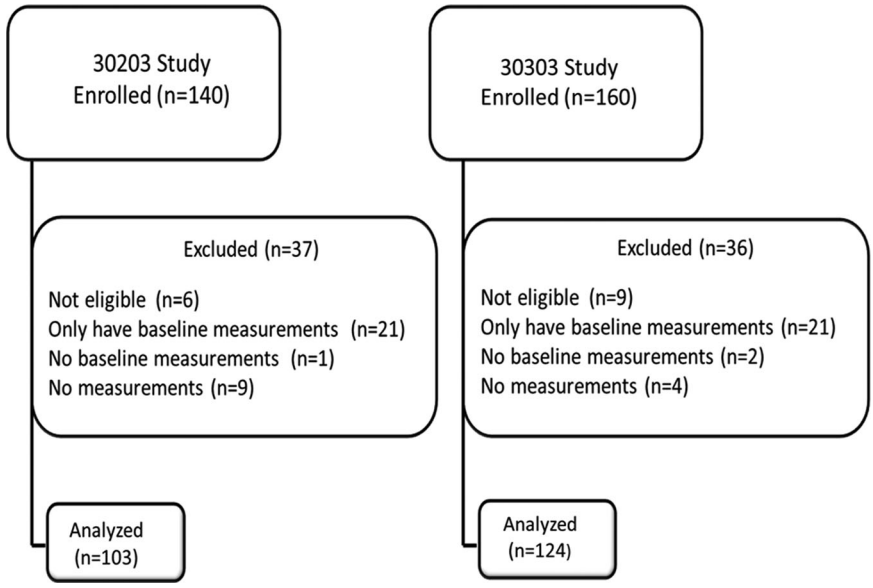

Figure 1 Selection of patients contributing tumor measurements from Cancer and Leukemia Group B (CALGB) 30203 and 30303.

Table 2 Tumor model parameter estimates and precision standard error of baseline (M_BASE), shrinkage rate (M_SR), and progression rate (M_PR) for the FDA registrational trials and CALGB 30203 and 30303 trials

Study	Treatment	No. of patients	M_BASE (cm)	M_SR (1/wk)	M_PR (cm/wk)
FDA trial treatment arms					
E4599	PCB	434	9.1 (0.33)	0.06 (0.004)	0.13 (0.02)
	PC	444	8.0 (0.30)	0.038 (0.01)	0.14 (0.04)
TAX 326	DC	408	8.7 (0.31)	0.052 (0.01)	0.16 (0.02)
	DCb	406	9.2 (0.38)	0.047 (0.005)	0.16 (0.02)
	VC	404	8.5 (0.28)	0.063 (0.01)	0.17 (0.02)
NCI trial treatment arms					
CALGB 30203	GCb +/- Zi or Ce	103	7.85 (0.45)	0.012 (0.002)	0.031 (0.002)
CALGB 30303	DC +/- BNP	124	8.28 (0.40)	0.035 (0.004)	0.072 (0.013)
	Total combined	227	8.10 (0.30)	0.025 (0.003)	0.059 (0.008)

BNP, BNP7787; CALCB, Cancer and Leukemia Group B; Ce, celecoxib; DC, docetaxel and cisplatin; DCb, docetaxel and carboplatin; FDA, US Food and Drug Administration; GCb, gemcitabine and carboplatin; M_BASE, precision standard error of baseline; M_PR, progression rate; M_SR, shrinkage rate; NCI, National Cancer Institute; PC, paclitaxel and carboplatin; PCB, paclitaxel, carboplatin, and bevacizumab; VC, vinorelbine and cisplatin; Zi, zileuton.

decrease to 8.1 cm $(-0.0251 \times 4) + (0.0594 \times 4) = 7.56$ cm. This 6.7% decrease reflects the average drug effect on tumor size. **Table 2** depicts the parameter estimates determined for patients with first-line metastatic NSCLC enrolled in five treatment arms for two multicenter phase III trials (>400 patients per study arm). Compared with these previously published findings, the CALGB results were lower for M_BASE, M_SR, and M_PR by 7%, 52%, and 61%, respectively.

Evaluation of deviations in parameter estimates

We expected these estimates to be more robust with smaller data sets and explored modifiable sources of noise in the data. First, we hypothesized that data from small cooperative group trials might be of lower quality than data perhaps more meticulously curated for submission to FDA review. We therefore identified 103 patients from the data set used to generate the FDA model, by matching their baseline tumor sizes to those of the 103 CALGB 30203 cases (who received carboplatin/gemcitabine). For the 103 patients identified from the FDA study, the observed mean and median baseline tumor sizes (**Table 3**) were comparable to those of the 103 CALGB 30203 cases, which suggested the matching method was able to identify a subset of patients from the larger FDA database to be comparable to the 103 patients in CALGB 30203. As a result, the parameter estimates for M_SR and M_PR were more similar to CALGB 30203 (**Table 3**) than to the results for any of the larger platinum doublet study arms in ECOG 4599 or TAX 326 (**Table 2**) even though the estimates for M_BASE still showed some difference. This implied that the deviation of parameter estimates between similar treatment arms in the CALGB and FDA data sets were unlikely to be due to significant differences in data quality and instead reflected effects of decreasing the size of the analyzed subject pool.

A less testable hypothesis is that the CALGB 30203 and the subset of 103 patients from the FDA data set are genuinely different from the larger population of patients on which the FDA model was based. Our experience with the multistep process of CT-imaging measurement and

Table 3 Observed baseline tumor size and tumor parameter estimates for first line platinum doublet therapy in CALGB 30203 and similarly treated patients from the FDA trials database

Study	Treatment	No. of patients	Baseline (mean) (cm)	Baseline (median) (cm)	M_BASE (cm)	M_SR (1/wk)	M_PR (cm/wk)
Subset of FDA trials database	Platinum doublets	103	9.74	8.70	9.26	0.0138	0.0346
NCI trial treatment arm (CALGB 30203)	GCb +/- Zi or Ce	103	9.71	8.70	7.85	0.0121	0.0312

CALCB, Cancer and Leukemia Group B; Ce, celecoxib; FDA, US Food and Drug Administration; GCb, gemcitabine and carboplatin; M_BASE, precision standard error of baseline; M_PR, progression rate; M_SR, shrinkage rate; NCI, National Cancer Institute; Zi, zileuton.

transmission of measurements into clinical trial databases offers an alternative hypothesis – the current RECIST-oriented clinical trial methods introduce variance in the recorded tumor burden that affects computational models of continuous tumor growth with minimum impact on RECIST-based time-to-event end points.

We therefore performed an exploratory hybrid investigation of data quality and modeling effects. We explored specific modifiable factors in the collection and reporting of tumor measurements that might contribute to the altered parameter estimates in the longitudinal growth model when the size of the population was decreased. To evaluate the reproducibility of the tumor measurements, an independent radiologist in blinded fashion measured the baseline target lesions and subsequent follow-ups from the original CT scans from 15 patients enrolled in CALGB 30203 and 30303 at one institution (**Figure 2**). For 4 of the 15 patients, at least one additional target lesion was identified (**Figure 2a**). Of the 15 subjects, 3 did not have an on-treatment assessment and therefore were not included in subsequent modeling analyses. For the 12 cases with serial measurements (**Figure 2b**), 4 (subjects 7, 8, 9, and 12) had trajectories of the measured sums of longest dimensions that were nearly superimposable between the CRFs and the blinded evaluator (BE) re-assessment. Four cases (subjects 1, 3, 4, and 5) had obvious divergence between the CRF and blinded evaluations in terms of the magnitude of change in tumor burden and timepoints at which these changes are registered. The remaining four cases had differences of unclear significance (subjects 2, 6, 10, and 11).

Estimated impact of continuous measurement variance on modeled end points

RECIST was developed to be robust to interrater variance in measurements by setting categories for tumor size changes (progressive disease, partial response, and complete response) based on thresholds for magnitudes of change that would be unlikely to be due to the greatest degree of interrater variance.[28] A patient's category of response would then likely only be due to a significant effect of treatment.[5,27] It is therefore not surprising that in settings where interrater variance is not actively controlled, assessments of continuous measurements of tumor growth will not improve upon our current RECIST-based categorical and time-to-event strategies.

We hypothesized that this interrater variance in tumor burden assessments would have a significant effect on more quantitative end points, such as TTG, with less effect on a RECIST-based time-to-event end point, such as PFS. For the 12 subjects with serial CRFs and blinded radiologist measurements (**Table 4**), we identified an average 18-week delay in TTG (range, 20–55 weeks) calculated from the re-evaluated scans compared with the CRF data, but no absolute differences in PFS assessments, corresponding with intraclass correlation coefficients of 0.32 and 1, respectively. The negative TTG values result from individuals for whom the tumor continues to progress from the baseline measurement and therefore the TTG actually occurred before the baseline measure. Despite differences in target lesion assessment and measurement, subjects met criteria for progressive disease at the same imaging session in both data sets.

DISCUSSION

This evaluation of NSCLC tumor measurements and end points in published cooperative group studies revealed limitations to using continuous measurements of tumor burden in phase II clinical trials. Modeling of typical phase III clinical trials has reproducibly demonstrated tumor burden metrics as predictors of survival.[1,16,23,31,32] These findings suggest that more quantitative evaluation of tumor growth trajectories early in the course of therapy might improve the efficiency of phase II clinical trials.[3,18,19,33] However, effective implementation of this strategy in phase II trials will require changes in the conduct and collection of data in such trials.

The primary advantage of the use of quantitative measures of tumor burden in early phase trials is to improve statistical power for detecting treatment effects. During this investigation, newly published analyses suggested that quantitative assessments of tumor burden were no more useful than RECIST-based categorical assessments or PFS.[4,21,22] Our findings are consistent with the hypothesis that the RECIST-based methods by which tumor measurement data are collected biases these evaluations. We found the modeled treatment effect and growth parameters in the 227 CALGB patients with NSCLC to diverge significantly from published results of a larger population. We then interrogated a smaller sample from the original data set from which the model was developed and obtained similar results. The large and consistent effect on computed parameters of longitudinal tumor growth models led us to scrutinize the original images and the recorded data. We identified "noise" in the process by which tumor burden is assessed and recorded to meet RECIST standards. This imprecision has no apparent effect on RECIST categories or time-to-event end points, but does affect tumor burden metrics.

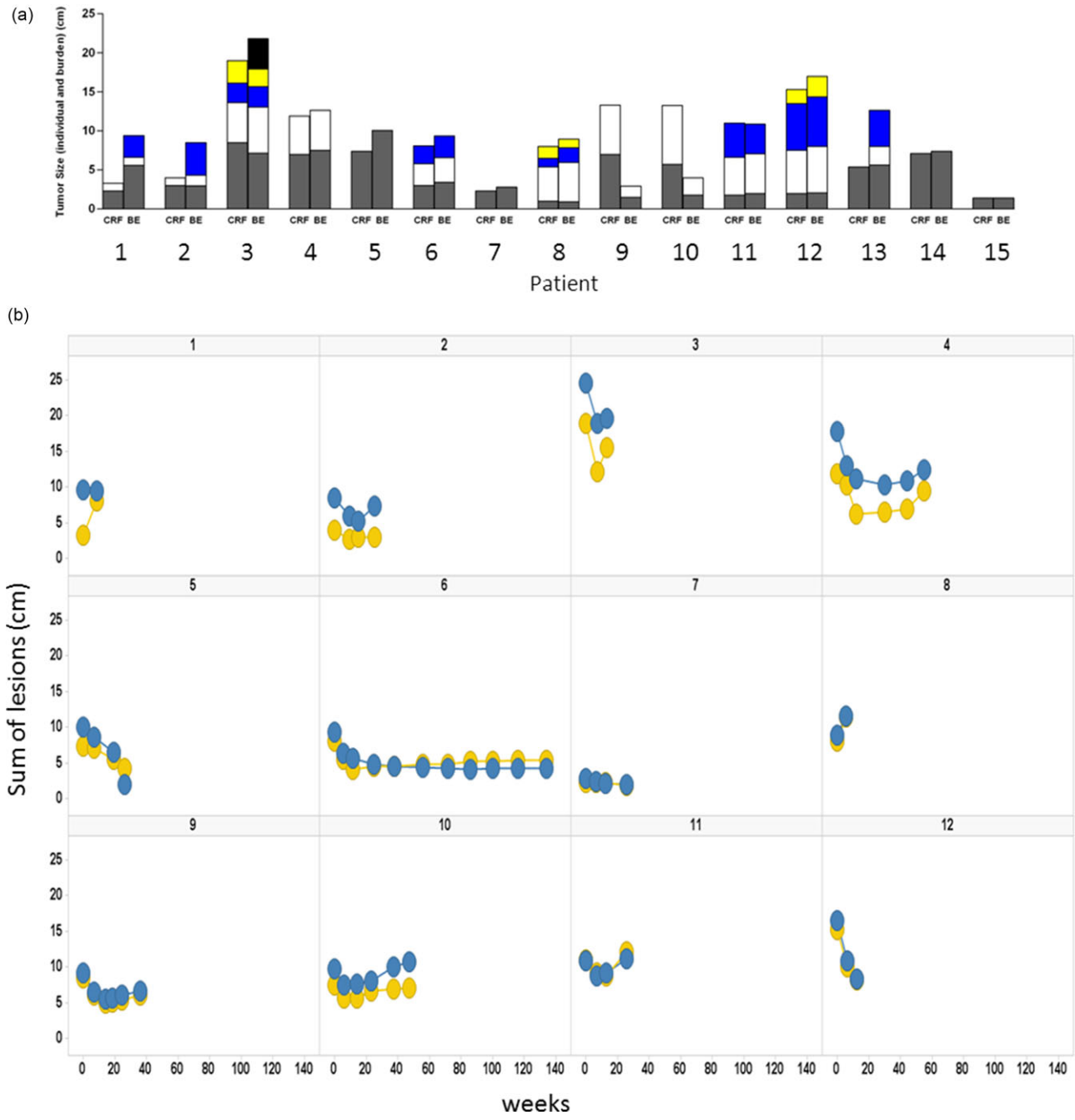

Figure 2 Baseline tumor burden represented by the sum of target lesion measurements from Cancer and Leukemia Group B (CALGB) 30203 and 30303. (**a**) Each pair of bars represents an individual patient's tumor burden, with each color representing the size of an independent target lesion, the first in gray, second in white, third in blue, fourth in yellow, fifth in black; left bar tumor measurements per case report forms (CRFs); right bar tumor measurements by independent, blinded evaluation (BE). (**b**) Tumor burden over time for subjects in CALGB 30203 and 30303. Horizontal axis reflects time in weeks; the vertical axis reflects the tumor burden by sum of the longest dimensions (cm) at each assessment timepoint for first 12 subjects in (**a**) by computed tomography (CT) imaging at each timepoint over the course of the trial. Circles represent tumor burden reported on case report forms (yellow) or on BE (blue).

There is no superior alternative approach to RECIST for the standardized assessment of anatomic tumor burden and its change over time.[34–36] This categorical system provides low interrater variance (progressive disease will be determined with high uniformity across sites in a multicenter trial and among trials) at the expense of efficiency (requires more patients to be observed over long periods of time). Our findings are consistent with investigators collecting and curating the quantitative tumor burden data with sufficient precision to support use of RECIST but not to support more computationally intensive methods of evaluating effects of treatments in small clinical trials. As long as this remains the process by

Table 4 Comparisons of PFS and calculated TTG from the target lesion measurements on original CRF and by blinded BE

Patient ID		1	2	3	4	5	6	7	8	9	10	11	12
PFS	CRF	8	26	12	48	28	128	40	5	36	48	24	18
(wk)	BE	8	26	12	48	28	128	40	5	36	48	24	18
TTG	CRF	−10	23	57	29	46	53	21	−85	32	23	30	50
(wk)	BE	65	48	69	80	53	72	24	−111	44	3	53	54

BE, blinded evaluator; CRF, case report form; PFS, progression-free survival; TTG, time-to-tumor growth.

which tumor burden data are collected, we would expect to find no consistent advantages to use of quantitative methods (such as tumor size ratio) in small phase II trials over more qualitative time-to-event strategies (such as PFS) for predicting impact on overall survival.[4,21,22]

This study had a limited data sample for analysis, but it required significant effort to obtain these data because these need to be retrospectively collected and analyzed. The primary databases maintained the RECIST-based categories in data fields, but obtaining the quantitative tumor measurements required manual retrieval and processing of archived paper forms. The small cohort of patients for whom images were available and reviewed might have been a biased sample, but this patient-recruitment site had been a major contributor to enrollment across thoracic oncology trials in CALGB with the stringent audit and quality control processes applied for member sites. The data are therefore likely representative of the overall quality of data in the larger clinical trials. Furthermore, data that included patients from independent trials submitted to the FDA yielded similar results. We cannot exclude the possibility that this particular subset of patients from the CALGB and FDA data sets represents a unique group of patients with NSCLC whose tumor growth patterns are distinct from the typical patient population. Therefore, our findings will require confirmation in other data sets.

Efforts to improve cancer therapeutics development are critical because, despite recent celebrated successes, the overall success rate of oncology drugs in phase III trials has been the lowest among fields of medicine.[37,38] The process of measuring, transmitting, analyzing, and interpreting CT imaging-based measures of tumor burden contributes significant but potentially modifiable variance to evaluations of treatment effects. This study demonstrates that this variance has greater effects on the ultimate performance of more computationally intensive metrics of tumor burden than conventional RECIST end points.

If quantitative strategies in assessing solid tumor burden are to improve the power of early phase trials to detect treatment effects, this will require changes in our methods for obtaining and recording the measurements. Centralized collection and measurement of CT images with semiautomated and digitally enhanced procedures may significantly reduce this variance. Advances in computing and digital data management in the past several years have made possible paperless systems with fewer opportunities for manual error.[39] Our findings suggest that establishing methods with less interrater variance could be a worthwhile investment in the future of cancer therapeutics assessment.

Acknowledgments. Support for this study was provided by NIH K23CA124802 (Career Development Award to M.L.M.), NIH U10CA031946 (Cancer and Leukemia Group B Chair's Development Project to M.L.M. and L.H.S.), NIH R01CA194783 (to M.L.M. and L.H.S.), and The Indiana CTSI through a gift from Eli Lilly and Company (C.H.L. and R.R.B.). The authors are grateful to Thomas Yaeger for technical assistance and Dr. Valerie Andre for expert guidance on tumor growth modeling.

Conflict of Interest/Disclosure. The views expressed in this article are those of the authors and do not necessarily reflect the official views of The FDA. L.H.S. is co-inventor on patents related to volumetric measurement of tumors on CT images.

Author Contributions. C.H.L., R.R.B., Y.W., S.K., M.J.E., M.J.R., L.H.S., and M.L.M. wrote the manuscript. C.H.L., R.R.B., Y.W., S.K., M.J.R., L.H.S., and M.L.M. designed the research. C.H.L., Y.W., M.R.S., S.K., L.W., M.J.E., A.A.M., E.E.V., A.O., and M.L.M. performed the research. C.H.L., R.R.B., Y.W., M.R.S., S.K., M.J.R., L.H.S., and M.L.M. analyzed the data. C.H.L., R.R.B., M.R.S., S.K., L.W., M.J.E., A.A.M., E.E.V., A.O., M.J.R., L.H.S., and M.L.M. contributed new reagents/analytical tools.

1. Claret, L., Lu, J.F., Bruno, R., Hsu, C.P., Hei, Y.J. & Sun, Y.N. Simulations using a drug-disease modeling framework and phase II data predict phase III survival outcome in first-line non-small-cell lung cancer. *Clin. Pharmacol. Ther.* **92**, 631–634 (2012).
2. Houk, B. Predictive capability of a Food and Drug Administration (FDA) overall survival model in non-small cell lung cancer in two phase II trials utilizing different anti-cancer agents. *Clin. Pharmacol. Ther.* **85**, abstract. (2009).
3. Wang, Y. *et al.* Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development. *Clin. Pharmacol. Ther.* **86**, 167–174 (2009).
4. Fridlyand, J., Kaiser, L.D. & Fyfe, G. Analysis of tumor burden versus progression-free survival for phase II decision making. *Contemp. Clin. Trials.* **32**, 446–452 (2011).
5. Moertel, C.G. & Hanley, J.A. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer* **38**, 388–394 (1976).
6. Barrett, J.S., Gupta, M. & Mondick, J.T. Model-based drug development applied to oncology. *Expert. Opin. Drug. Discov.* **2**, 185–209 (2007).
7. Bruno, R. & Claret, L. On the use of change in tumor size to predict survival in clinical oncology studies: toward a new paradigm to design and evaluate phase II studies. *Clin. Pharmacol. Ther.* **86**, 136–138 (2009).
8. Bruno, R., Lu, J.F., Sun, Y.N. & Claret, L. A modeling and simulation framework to support early clinical drug development decisions in oncology. *J. Clin. Pharmacol.* **51**, 6–8 (2010).
9. Ribba, B. *et al.* A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis. *CPT Pharmacom. Syst. Pharmacol.* **3**, e113 (2014).
10. Siegel, R.L., Miller, K.D. & Jemal, A. *CA Cancer J. Clin.* Jan-Feb;**65**,5-29 (2015). doi: 10.3322/caac.21254. Epub 2015 Jan 5. PMID:25559415
11. Adjei, A.A., Christian, M. & Ivy, P. Novel designs and end points for phase II clinical trials. *Clin. Cancer Res.* **15**, 1866–1872 (2009).
12. Dhani, N., Tu, D., Sargent, D.J., Seymour, L. & Moore, M.J. Alternate endpoints for screening phase II studies. *Clin. Cancer Res.* **15**, 1873–1882 (2009).
13. Maitland, M.L., Bies, R.R. & Barrett, J.S. A time to keep and a time to cast away categories of tumor response. *J. Clin. Oncol.* **29**, 3109–3111 (2011).
14. Maitland, M.L. & Schilsky, R.L. Clinical trials in the era of personalized oncology. *CA Cancer J. Clin.* **61**, 365–381 (2011).
15. Mandrekar, S.J. & Sargent, D.J. Randomized phase II trials: time for a new era in clinical trial design. *J. Thorac. Oncol.* **5**, 932–934 (2010).
16. Stein, W.D. *et al.* Tumor regression and growth rates determined in five intramural NCI prostate cancer trials: the growth rate constant as an indicator of therapeutic efficacy. *Clin. Cancer Res.* **17**, 907–917 (2011).
17. Yap, T.A., Sandhu, S.K., Workman, P. & de Bono, J.S. Envisioning the future of early anticancer drug development. *Nat. Rev. Cancer.* **10**, 514–523 (2010).
18. Lavin, P.T. An alternative model for the evaluation of antitumor activity. *Cancer Clin. Trials.* **4**, 451–457 (1981).
19. Karrison, T.G., Maitland, M.L., Stadler, W.M. & Ratain, M.J. 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).
20. Lara, P.N Jr *et al.* Southwest Oncology Group. 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).

21. An, M.W. *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).
22. Kaiser, L.D. Tumor burden modeling versus progression-free survival for phase II decision making. *Clin. Cancer Res.* **19**, 314–319 (2013).
23. Claret, L. *et al.* Evaluation of tumor-size response metrics to predict overall survival in Western and Chinese patients with first-line metastatic colorectal cancer. *J. Clin. Oncol.* **31**, 2110–2114 (2013).
24. Lilenbaum, R.C. *et al.* Single-agent versus combination chemotherapy in advanced non-small-cell lung cancer: the cancer and leukemia group B (study 9730). *J. Clin. Oncol.* **23**, 190–196 (2005).
25. Edelman, M.J. *et al.* Eicosanoid modulation in advanced lung cancer: cyclooxygenase-2 expression is a positive predictive factor for celecoxib + chemotherapy–Cancer and Leukemia Group B Trial 30203. *J. Clin. Oncol.* **26**, 848–855 (2008).
26. Miller, A.A. *et al.* Phase II randomized study of dose-dense docetaxel and cisplatin every 2 weeks with pegfilgrastim and darbepoetin alfa with and without the chemoprotector BNP7787 in patients with advanced non-small cell lung cancer (CALGB 30303). *J. Thorac. Oncol.* **3**, 1159–1165 (2008).
27. Miller, A.B., Hoogstraten, B., Staquet, M. & Winkler, A. Reporting results of cancer treatment. *Cancer* **47**, 207–214 (1981).
28. Therasse, P. *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).
29. Holford, N.H. Wings for NONMEM. (2012).
30. D'Agostino, R.B Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **17**, 2265–2281 (1998).
31. Stein, W.D. *et al.* Analyzing the pivotal trial that compared sunitinib and IFN-α in renal cell carcinoma, using a method that assesses tumor regression and growth. *Clin. Cancer Res.* **18**, 2374–2381 (2012).
32. Claret, L., Bruno, R., Lu, J.F., Sun, Y.N. & Hsu, C.P. Exploratory modeling and simulation to support development of motesanib in Asian patients with non-small cell lung cancer based on MONET1 study results. *Clin. Pharmacol. Ther.* **95**, 446-451 (2014).
33. Maitland, M.L. *et al.* Estimation of renal cell carcinoma treatment effects from disease progression modeling. *Clin. Pharmacol. Ther.* **93**, 345–351 (2013).
34. Eisenhauer, E.A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
35. Sharma, M.R., Maitland, M.L. & Ratain, M.J. RECIST: no longer the sharpest tool in the oncology clinical trials toolbox—point. *Cancer Res.* **72**, 5145–5149; discussion 5150 (2012).
36. Fojo, A.T. & Noonan, A. Why RECIST works and why it should stay–counterpoint. *Cancer Res.* **72**, 5151–5157; discussion 5158 (2012).
37. Hay, M., Thomas, D.W., Craighead, J.L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32**, 40–51 (2014).
38. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug. Discov.* **3**, 711–715 (2004).
39. Pietanza, M.C. *et al.* Harnessing technology to improve clinical trials: study of real-time informatics to collect data, toxicities, image response assessments, and patient-reported outcomes in a phase II clinical trial. *J. Clin. Oncol.* **31**, 2004–2009 (2013).

Supplementary information accompanies this paper on the *Clinical and Translational Science* website. (http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1752-8062)