A Pilot Study of Volume Measurement as a Method of Tumor Response Evaluation to Aid Biomarker Development

Binsheng Zhao1,2, Geoffrey R. Oxnard3, Chaya S. Moskowitz4, Mark G. Kris3, William Pao3,5, Pingzhen Guo2, Valerie M. Rusch6, Marc Ladanyi5,7, Naiyer A. Rizvi3, and Lawrence H. Schwartz2

Abstract

Purpose: Tissue biomarker discovery is potentially limited by conventional tumor measurement techniques, which have an uncertain ability to accurately distinguish sensitive and resistant tumors. Semiautomated volumetric measurement of computed tomography imaging has the potential to more accurately capture tumor growth dynamics, allowing for more exact separation of sensitive and resistant tumors and a more accurate comparison of tissue characteristics.

Experimental Design: Forty-eight patients with early stage non–small cell lung cancer and clinical characteristics of sensitivity to gefitinib were studied. High-resolution computed tomography was done at baseline and after 3 weeks of gefitinib. Tumors were then resected and molecularly profiled. Unidimensional and volumetric measurements were done using a semiautomated algorithm. Measurement changes were evaluated for their ability to differentiate tumors with and without sensitizing mutations.

Results: Forty-four percent of tumors had epidermal growth factor receptor–sensitizing mutations. Receiver operating characteristic curve analysis showed that volumetric measurement had a higher area under the curve than unidimensional measurement for identifying tumors harboring sensitizing mutations (P = 0.009). Tumor volume decrease of >24.9% was the imaging criteria best able to classify tumors with and without sensitizing mutations (sensitivity, 90%; specificity, 89%).

Conclusions: Volumetric tumor measurement was better than unidimensional tumor measurement at distinguishing tumors based on presence or absence of a sensitizing mutation. Use of volume-based response assessment for the development of tissue biomarkers could reduce contamination between sensitive and resistant tumor populations, improving our ability to identify meaningful predictors of sensitivity. Clin Cancer Res; 16(18): 4647–53. ©2010 AACR.

Tumor imaging for response evaluation has a fundamental role in oncology care and drug development. A reduction in tumor size or response has conventionally suggested a tumor is biologically vulnerable to a treatment; yet, it remains unclear what magnitude of size reduction is biologically meaningful. Published criteria for the classification of response are used widely in the reporting of clinical trial results (1, 2), but studies on individual patient outcomes have failed to identify a clear survival benefit associated with a partial response (3, 4). Importantly, criteria such as Response Evaluation Criteria In Solid Tumors (RECIST; ref. 1) were developed for consistency with historical definitions of response and for minimizing the impact of measurement variability (5, 6) but were not developed as biologically based criteria for tumor response.

The lack of a biological basis to conventional response assessment becomes particularly important for tissue biomarker development, especially because the efficacy of newer targeted therapies is often limited to a subset of biologically vulnerable cancers. The development of tissue biomarkers predicting sensitivity to therapy are essential, yet biomarker studies are potentially limited by their use of conventional response criteria, which divide patients into responders and nonresponders based on simple unidimensional tumor measurements that fail to accurately capture tumor growth dynamics. By treating tumors with RECIST response as sensitive and others as resistant, investigators compare the tissue characteristics of two arbitrarily defined subgroups.
The clearest example of the deficiency of RECIST is in the treatment of non–small cell lung cancer (NSCLC) with epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors like gefitinib. In advanced NSCLC, EGFR-sensitizing mutations in exons 19 and 21 have recently been validated as the strongest biomarker predicting prolonged progression-free survival on gefitinib (7); patients harboring a sensitizing mutation have a median progression-free survival of 10 months compared to 1.5 months for those lacking a sensitizing mutation (8). However, the RECIST response rate for EGFR tyrosine kinase inhibitors in patients with sensitizing mutations is only 70% (8–10), meaning nearly one of three patients is classified as a nonresponder despite gaining important clinical benefit. Published waterfall plots have shown that most EGFR mutant cancers develop tumor shrinkage on tyrosine kinase inhibitor, although criteria for a partial response may not be met (10, 11). By grouping tumors harboring a sensitizing mutation as part of the resistant pool, this arbitrary response threshold could impair our ability to identify meaningful tissue differences, potentially hampering the development of tissue biomarkers for novel therapies.

We hypothesized that more accurate imaging techniques would be better able to distinguish resistant tumors from those harboring sensitizing mutations and could thus aid biomarker development. Toward this goal, we developed a semiautomated algorithm for computed tomography–based tumor volume measurement, which we found to be feasible in lung nodules with low variability (12, 13). Using gefitinib therapy as a model and by measuring the entire tumor mass before and after treatment, we proposed to find a biologically based threshold that could optimally dichotomize lung cancers into sensitive (EGFR mutant) and resistant [EGFR wild type (WT)] groups. Such a finding would support further investigations into the application volumetric response as a tool for biomarker research in therapies where a predictive tissue biomarker has not yet been identified.

**Materials and Methods**

Imaging and tissue data were obtained prospectively as an exploratory analysis within a phase II trial of neoadjuvant gefitinib in patients with NSCLC (14), the results of which are being published separately. All patients had stage I or II NSCLC and were deemed operable and resectable. To enrich the population for EGFR-mutant cancers, patients were only included if (a) they had a smoking history of <15 pack-years or (b) their tumors had histologic features of bronchioloalveolar cancer, characteristics associated with EGFR tyrosine kinase inhibitor sensitivity (15). Between July 2004 and March 2008, 50 patients were treated. Nine patients were eligible based on histology and 41 based on smoking history. Patients received gefitinib daily for 3 weeks before surgery and discontinued it 2 days before their operation. Computed tomography imaging was done before starting gefitinib and before surgery. This study was approved by our institutional review board.

**Tumor genomic analysis**

At time of resection, tumor tissue was snap frozen in liquid nitrogen and stored in a −80°C freezer. Representative areas of these specimens were pathologically reviewed to confirm the diagnosis and presence of tumor. Genomic DNA was analyzed for the most common EGFR-sensitizing mutations (exons 19 and 21) using previously described PCR-based methods (16–18). EGFR WT tumors were also tested for KRAS mutations, which are found in a nonoverlapping subset of lung adenocarcinomas that have been found to be resistant to EGFR tyrosine kinase inhibitor therapy (18–20). If no EGFR or KRAS mutations were found, then the remaining EGFR exons (18–24) were assessed by standard dideoxynucleotide sequencing. Pretreatment tissue was analyzed by the above method for the first 18 patients on the study, but this requirement was removed after we observed 100% concordance between the pretreatment and resection results (21). Selected specimens that were found to be EGFR/KRAS WT were submitted for more detailed mutational testing using mass spectrometry.

**Tumor imaging and measurement**

Baseline computed tomography of each patient was done within 2 weeks before gefitinib initiation. A follow-up computed tomography scan was done using the same imaging acquisition technique 3 weeks later, before surgery. Non–contrast enhanced diagnostic chest computed tomographies were done with a LightSpeed 16 scanner (GE Medical Systems) during a breath hold. High-resolution
images with 1.25-mm slice thickness were reconstructed per our volumetric imaging protocol (13). Two patients were excluded because 1.25-mm slice thickness reconstructions were not done as required by protocol, leaving 48 of 50 patients qualified for our analysis.

Tumor contours were semiautomatically delineated using a three-dimensional segmentation algorithm (12, 13). To ensure appropriateness of the segmentation results, the computer-generated tumor contours were inspected by two radiologists blinded to mutation status and scan dates (P. Guo and L.H. Schwartz), and suboptimal computer segmentation results were corrected by one of the radiologists. The areas defined by these final tumor contours were then summed across all computed tomography slices to calculate a total tumor volume measurement. Tumor greatest diameter was measured from a single transverse image plane as described by the RECIST guidelines (1). Changes in tumor measurements were calculated by subtracting the baseline from the follow-up measurement, then dividing by the baseline measurement.

Statistical analysis
Tumors were divided into molecular subgroups defined by presence or absence of a sensitizing mutation. The distribution of tumor measurement changes was compared among subgroups using a Wilcoxon rank-sum test. Using estimated 95% limits of agreement for unidimensional (−7.3%, +6.2%) and volumetric (−12.1%, +13.4%) measurement from a previous study on measurement reproducibility (13), an exact McNemar's test was used to compare the proportion of measurement changes that decrease outside these ranges of variability.

Measurement methods were then evaluated to explore how well they could differentiate between tumors with and without a sensitizing mutation. Nonparametric estimates of receiver operating characteristic curves and the area under these curves were calculated for both measurement methods. For the receiver operating characteristic calculations, the relative percent change of the unidimensional and volumetric measurements were multiplied by −1, thus using the percent decrease in measurement as a positive value to estimate the receiver operating characteristic curves and areas under the curve. To test whether the areas under the curve were equal, methods suggested by DeLong et al. (22) were used.

Youden's index was calculated to determine what threshold for volumetric and unidimensional measurement change optimally dichotomized tumors with and without sensitizing mutations (23). Youden's index is equal to sensitivity + specificity − 1. By choosing its maximum value as a diagnostic threshold, we place equal importance on sensitivity and specificity. For context, we also calculated the sensitivity and specificity of a RECIST-based threshold; for unidimensional measurement, we used a 30% decrease in diameter (1), whereas for volumetric measurement, we used the mathematically equivalent volumetric decrease (65%) from which RECIST was developed (1, 5).

All tests were considered significant at the P ≤ 0.05 level. Analyses were done in Stata 10.0 for Windows (StataCorp LP).

Results
Tumor measurements grouped by mutation status
Twenty-one (44%) of 48 tumors harbored an EGFR mutation. Twenty-seven (56%) were EGFR WT, with five of these harboring a KRAS mutation. A median of 24 days passed between baseline and follow-up scans (range, 20-41 d), with a median of 22 days on gefitinib (range, 20-28 d). Unidimensional and volumetric tumor measurements are summarized in Table 1 for all 48 tumors and grouped by presence or absence of sensitizing mutation. As predicted biologically, there was a significant difference between the mean measurement changes of these subgroups of tumors, showed with unidimensional measurement (−11.1% versus −2.6%; P = 0.002) and volumetric measurement (−45.9% versus −4.4%; P < 0.001). Across all tumors, the overall magnitude of change is greater for volumetric measurement, consistent with the mathematical relationship between change in diameter and change in volume (5).

Individual tumor measurement changes were compared with the aforementioned 95% limits of agreement determined from a previous study (13). Seventy-five percent of volumetric measurement changes surpassed the expected range of variability, whereas only 46% of unidimensional changes surpassed the expected range of variability (P = 0.003). This indicates that most unidimensional measurement changes after 3 weeks of treatment were indistinguishable from the expected changes because of variability in computed tomography acquisition and measurement.

Using tumor measurements to dichotomize tumors with and without sensitizing mutations
Figure 1 shows the distribution of tumor measurement changes for both measurement methods, grouping tumors

| Table 1. Change in tumor measurements, grouped by mutation status, for each measurement method |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                | All tumors      | EGFR mutant     | EGFR WT*        | P               |
| N               | 48              | 21              | 27              |                 |
| Unidimensional percentage change (%) |                 |                 |                 |                 |
| Median          | −5.3            | −11.1           | −1.8            | 0.002           |
| IQR             | −13.1, 0.7      | −28.9, −4.4     | −7.0, 1.3       |                 |
| Volumetric percentage change (%) |                 |                 |                 |                 |
| Median          | −23.9           | −45.9           | −4.4            | <0.001          |
| IQR             | −44.8, 0.0      | −54.1, −29.1    | −16.5, 7.0      |                 |

NOTE: P values are from a Wilcoxon rank-sum test comparing EGFR mutant tumors versus EGFR WT tumors. Abbreviation: IQR, interquartile range.

*Includes all KRAS mutant tumors.
by mutation status. With unidimensional measurement, there is overlap between the interquartile ranges of the measurement changes from the two tumor subgroups, whereas there is better separation of these subgroups using volumetric measurement.

To determine which imaging method is a better diagnostic test for identifying tumors with sensitizing mutations, we calculated receiver operating characteristic curves (Fig. 2). The receiver operating characteristic analysis evaluates each tumor measurement change as a potential threshold for a binary diagnostic test predicting EGFR mutation status. The receiver operating characteristic curve plots the sensitivity by 1 - specificity for all such potential thresholds, and the area under the curve represents the strength of the test across multiple possible thresholds. We found that volumetric measurement has a significantly higher area under the curve than unidimensional measurement ($P = 0.009$). Figure 3 shows this observation in one tumor harboring a sensitizing mutation, in which volume measurement detected a change that was not detected by conventional unidimensional measurement.

**Optimal thresholds for biomarker development**

For each measurement method, we calculated which threshold had the highest combined sensitivity and specificity for dichotomizing tumors based on presence or absence of a sensitizing mutation (Table 2). For volumetric measurement, a threshold of 24.9% decrease (Youdens' index, 0.79) classified 90% of tumors with a mutation as responders and 89% of tumors without a mutation as nonresponders (positive predictive value, 86%; negative predictive value, 92%). For unidimensional measurement, the threshold with the highest sensitivity and specificity was a 7.0% decrease (Youdens' index, 0.49), which classified 71% of tumors with a mutation as responders and 78% of tumors without a mutation as nonresponders (positive predictive value, 71%; negative predictive value, 78%). Because a 7.0% decrease would be within the expected measurement variability of unidimensional measurement ($-7.3\%, +6.2\%$), we did an ad hoc calculation to determine what threshold outside of that range had the highest Youdens' index. We determined that a 9.0% unidimensional decrease (Youdens' index, 0.38) would have a sensitivity of 52% and a specificity of 85% (positive predictive value, 73%; negative predictive value, 70%) and would be less influenced by measurement variability than a threshold of 7.0% decrease. These thresholds will require validation in an independent data set as a confirmation of their accuracy.

We then evaluated the effectiveness of RECIST-based thresholds at separating tumors based on presence or absence of a sensitizing mutation using the conventional 30% decrease for unidimensional measurement and a mathematically equivalent 65% decrease for volumetric measurement (5). Both thresholds had a high specificity...
(Table 2) but classified only 3 of 21 EGFR mutant tumors as responders (sensitivity, 14%). Although these thresholds accurately classified EGFR WT tumors as nonresponders, the high false-negative rate meant that only 59% to 60% of the nonresponding tumors were actually EGFR WT (negative predictive value, ~60%). Importantly, these results were calculated from applying RECIST-based thresholds after only 3 weeks of treatment and therefore do not necessarily represent the accuracy of RECIST in biomarker analyses of full phase II trial results.

**Discussion**

The chance of a novel therapy achieving success in the clinical setting is vastly improved by the identification of a tissue biomarker predicting increased tumor sensitivity. For example, through the identification of EGFR-sensitizing mutations in NSCLC, erlotinib and gefitinib have been transformed from one of many second-line therapies to the primary therapy for a significant subset of cancers. Alternatively, cetuximab is an EGFR targeted therapy that, despite positive phase III trial results (24), has not received Food and Drug Administration approval for NSCLC and has garnered limited clinical enthusiasm in lung cancer therapy, in large part because candidate biomarkers (KRAS mutation, EGFR expression) have failed to show predictive ability (24–27). Although a recently published study in colorectal cancer has explored newer biomarkers of cetuximab response (amphiregulin, epiregulin; ref. 28), the results were notable for marginal statistical significance. We hypothesize that such biomarker analyses could be strengthened by better use of imaging and biologically based definitions of tumor response.

Our data indicate that biomarker studies like the one described above are limited by the modest ability of unidimensional measurement to distinguish tumors with and without sensitizing mutations. In our study, there was significant overlap between the unidimensional measurement changes of EGFR mutant and EGFR WT tumors; the optimal threshold for diameter measurement, a 7% decrease, had a disappointing sensitivity and specificity (71%, 78%). Volumetric measurement, however, was better able to dichotomize tumors with and without sensitizing mutations across multiple possible response thresholds, with a significantly better area under the curve by receiver operating characteristic curve analysis ($P = 0.009$). Our data indicate that, using a 24.9% volume decrease to dichotomize tumors for biomarker analysis, sensitivity and specificity would be higher (90%, 89%), potentially improving the ability to identify meaningful tissue differences between the tumors being studied. These results will need to be validated in a larger, more generalizable patient population.

One possible value of volume measurement, aside from its ability to capture changes to the whole tumor

**Table 2. Sensitivity and specificity of different possible thresholds for dichotomizing tumors into EGFR mutant and EGFR WT groups**

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidimensional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal*</td>
<td>7.0% Decrease</td>
<td>71%</td>
</tr>
<tr>
<td>Alternate†</td>
<td>9.0% Decrease</td>
<td>52%</td>
</tr>
<tr>
<td>RECIST</td>
<td>30% Decrease</td>
<td>14%</td>
</tr>
<tr>
<td>Volumetric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal*</td>
<td>24.9% Decrease</td>
<td>90%</td>
</tr>
<tr>
<td>RECIST</td>
<td>65% Decrease</td>
<td>14%</td>
</tr>
</tbody>
</table>

Abbreviations: Sens, sensitivity = percentage of EGFR mutant tumors classified as responders; Spec, specificity = percentage of EGFR WT tumors classified as nonresponders.

*Threshold calculated to have highest summed sensitivity and specificity.

†Threshold outside of the 95% limits of agreement with the highest summed sensitivity and specificity.
mass, is that measurement variability seems to have a relatively limited impact on the measurement changes that are found. In a recently completed study, we showed that, for patients with NSCLC undergoing same-day repeat computed tomography scans, the 95% limits of agreement for volumetric measurement were (−12.1% to +13.4%), whereas for unidimensional measurement, they were (−7.3% to +6.2%; ref. 13). In the present study, the tumor measurements were obtained using the same scanning technique, segmentation software, and semi-automated measurement algorithm as in this “rescan” study. Using unidimensional measurement, we found that most measurement changes (54%) were within the 95% limits of agreement, whereas with volumetric measurement, a much smaller portion of measurement changes (25%) were within the 95% limits of agreement. Therefore, measurement variability may have greater impact on unidimensional measurement, potentially limiting its effectiveness in biomarker research. Our data suggest that the use of RECIST-baseline thresholds for dichotomizing tumors according to presence of a sensitizing mutation is inaccurate in this setting. After 3 weeks of therapy, RECIST-based thresholds had a sensitivity of only 14%, although this is likely lower than would be found when applying these thresholds after a conventional 6 or 8 week treatment period. A more accurate approximation of the sensitivity of RECIST response in clinical trials would be derived from the 70% RECIST response rate for EGFR-mutant lung cancers treated with tyrosine kinase inhibitor. This is approximately equivalent to a sensitivity of 70%; although the remaining EGFR-mutant tumors are designated as nonresponders, we believe that most of these still gain benefit from this targeted therapy despite not meeting criteria for a partial response. Just as we found that a response threshold of <30% decrease optimally dichotomizes tumors after 3 weeks of therapy, we suspect that a minor response would be better than a full RECIST response at dichotomizing tumors with and without sensitizing mutations. Others have found that minor response is predictive of better outcome to targeted therapy in colon cancer, renal cell cancer, and gastrointestinal stromal tumor (29–31); whether this is the case for tyrosine kinase inhibitor therapy in NSCLC is unclear and requires further study.

As EGFR mutation status was our tissue biomarker of interest in this study, we used several precautions to ensure minimal error in molecular diagnostic testing. The use of surgical specimens ensured adequate tissue for analysis, and all specimens were reviewed pathologically to ensure an adequate population of cancer cells for tumor DNA extraction. Our PCR analysis, which is specific for exon 19 deletions/insertions and L858R mutations (making up >90% of all EGFR-sensitizing mutations), is extremely sensitive and able to detect a mutation when present in as little as 5% of DNA (16, 32). To identify rare mutations (<10% of total), direct DNA sequencing was done; which is less sensitive but still identifies 80% to 90% of mutations. Together, these steps should therefore identify >98% of sensitizing mutations. Further mass spectrometry analysis was done on four cases with significant tumor regression, identifying one additional rare EGFR-sensitizing mutation. For the remaining 20 patients with no mutation identified, we believe that the likelihood of there being an additional undetected EGFR mutation is <1% to 2%.

Conclusions

This analysis of tumor imaging changes after gefitinib therapy found that volumetric measurement can more accurately distinguish between tumors with and without sensitizing mutations than conventional unidimensional measurement. We recommend further study of tumor volume measurement as a tool for dichotomizing sensitive and resistant tumors to improve tissue biomarker development.

Disclosure of Potential Conflicts of Interest

W. Pao is a consultant of AstraZeneca and MolecularMD and has patent rights on testing for EGFR T790M mutation. M. Kris is a consultant of AstraZeneca, Pfizer, and Boehringer Ingelheim.

Acknowledgments

We thank Carolina Montalvo, study coordinator, for her excellent effort in the data collection for this study over the past 5 years.

Grant Support

Grants R21CA113653, R01CA125143, and P01CA129243 from the National Cancer Institute, the Experimental Therapeutics Center of Memorial Sloan-Kettering Cancer Center, and a grant from the Carmel Hill Fund. The funding organizations of this study had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 01/15/2010; revised 04/25/2010; accepted 05/31/2010; published OnlineFirst 06/09/2010.

References


A Pilot Study of Volume Measurement as a Method of Tumor Response Evaluation to Aid Biomarker Development

Binsheng Zhao, Geoffrey R. Oxnard, Chaya S. Moskowitz, et al.

*Clin Cancer Res* 2010;16:4647-4653. Published OnlineFirst June 9, 2010.