Considerations for the Use of Imaging Tools for Phase II Treatment Trials in Oncology

Lalitha K. Shankar,1 Annick Van den Abbeele,2 Jeff Yap,2 Robert Benjamin,3 Scott Scheutze,4 and T.J. FitzGerald5

Abstract

In the context of assessing tumor response, imaging tools have the potential to play a vital role in phase II and III treatment trials. If the imaging test is able to predict potential phase III success in a reliable fashion, it would be a useful tool in phase II trial design as it may provide for a more rapid and timely response assessment. The benefits and challenges of using anatomic imaging measures as well as the promising molecular imaging measures, primarily fluorodeoxyglucose-positron emission tomography, are discussed here. The general issues related to successful implementation of advanced imaging in the context of phase II treatment trials are discussed.

Tumor response as measured by anatomic imaging, which was used to measure therapeutic efficacy in the era of cytotoxic drugs, has been a valuable tool in validating clinical trials. Advanced imaging technologies, including both metabolic and molecular imaging platforms, are becoming important partners in evaluating response as a clinically meaningful study endpoint. Response Evaluation Criteria in Solid Tumors (RECIST) are commonly used for assessing response by computed tomography (CT) or magnetic resonance imaging, to define clinical trial endpoints, and are useful in studies evaluating cytotoxic drugs in which tumor shrinkage is expected (1). RECIST was introduced in 2000 in an effort to increase the accuracy of measuring tumor response to improve comparison of response data between studies. Compared with the older WHO criteria, RECIST streamlined tumor assessment by requiring measurement of only one dimension, defined a minimum measurable tumor size (2 cm), recommended measurement of up to 10 lesions, 5 per organ, and redefined progressive disease as a 20% increase in the sum of the longest diameter. The approach to tumor measurement using RECIST was recently summarized (2). The partial response definitions for RECIST and WHO, at least 30% or 50% decreases in tumor size, respectively, were comparable, indicating a 65% reduction in volume. There was an important difference in the definition of disease progression. The RECIST-defined 20% increase in one dimension translated to a 73% increase in tumor volume, relative to the WHO-defined 25% increase in the sum of the products of bidimensional measurements, which translated to a 40% increase in tumor volume. This effectively increased the time to progression (TTP) rate and may have avoided premature removal of some patients from study. The RECIST were adapted globally as a simplified tumor metric.

Recently, RECIST version 1.1 has been presented, with a number of evidence-based changes recommended based on a review of databases comprising tumor measurements in >6,500 patients.6 These include a reduction to 5 in the number of lesions that must be measured, 2 per organ; a requirement that a 5 mm absolute increase in tumor size be documented, a definition for lymph node measurement including the point that the short axis is most accurate (distinguished from the longest axis for the single dimension of a non-nodal mass); and a comment on inclusion of fluorodeoxyglucose (FDG)-positron emission tomography (PET). RECIST version 1.1 will be described this year in the European Journal of Cancer. These changes are expected to continue to improve the reproducibility of response assessments across clinical trials. Intratrical variability in accurate detection of response and disease progression is considered one major reason for failure of phase II trials in phase III (3, 4). Thus, accurate assessment of anatomic imaging continues to be a holy grail in cancer clinical trials.

Standard anatomic criteria for therapeutic response such as RECIST that are based on changes in tumor size, however, may or may not be applicable for molecular targeted agents, which are being currently being evaluated alone or in combination with cytotoxic drugs in treatment trials. As a result, early changes (indicating progression or response) may remain undetected. Furthermore, tumors may not shrink, although they respond to treatment and traditional objective response

---

Authors’ Affiliations: 1Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland; 2Dana-Farber Cancer Center, Boston, Massachusetts; 3The University of Texas M. D. Anderson Cancer Center, Houston, Texas; 4University of Michigan Health System, Ann Arbor, Michigan; and 5Quality Assurance Review Center, Providence, Rhode Island

Received 10/28/08; revised 12/16/08; accepted 1/15/09; published OnlineFirst 3/10/09.

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.

Requests for reprints: Lalitha K. Shankar, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 6130 Executive Boulevard, Room 6056, Bethesda, MD 20892-7412. Phone: 301-496-9531; Fax: 301-480-3507; E-mail: shankarl@mail.nih.gov.

© 2009 American Association for Cancer Research. doi:10.1158/1078-0432.CCR-08-2030

---

Criteria do not take into account changes in the density of the tumor mass, measured in Hounsfield units (5, 6). As a result of these limitations, standard anatomic response may not correlate with clinical outcome (5) and alternative strategies may be needed to validate response for study outcome analysis.

In a Children’s Oncology Group ongoing clinical trial evaluating targeted therapy with tamoxifen for children with persistent and/or recurrent desmoid tumors, Drs. Kao and Hoffer found that some patients on study classified as progressive disease defined by both the RECIST and the WHO criteria on magnetic resonance imaging (MRI) were clinically stable or even improved when assessed by clinical outcome data. In a retrospective review of the MRI, each of the clinically stable children had a change in magnetic resonance features (such as increase in proportion of low signal intensities on T2W image from maturing fibrosis tissues) independent of tumor size, thus offering another potential variable for response assessment.

Response should be a surrogate for a more meaningful endpoint, such as survival. Often, this has not been the case. In addition to response rate, TTP, and progression-free survival, other endpoints that may be predictive of overall survival have been used. These are sometimes classified as clinical benefit endpoints. This issue of CCR FOCUS is dedicated to the discussion of such endpoints, including novel methods of

---

**Fig. 1.** TTP in good and poor responders to imatinib in 98 patients with GIST by response criteria (32). A, response by Choi et al. criteria. B, response by RECIST. When the tumor response was evaluated based on Choi et al. response criteria, a significant difference was observed in the long-term prognosis between good and poor responders ($P = 0.0002$) for up to 28 mo, but no significant difference was observed between good and poor responders by RECIST ($P = 0.1$). Reprinted and adapted with permission. © 2008 American Society of Clinical Oncology. All rights reserved. Benjamin RS, et al.: We should desist using RECIST, at least in GIST. J Clin Oncol 2007;25:1760–4.

---

7 Personal communication.
analyzing conventional data, as well as identification of novel endpoints and protocol design (4, 7–9).

One of the most commonly used endpoints is stable disease. This endpoint is felt to be a misnomer because rarely is disease stable. The term has become a catch-all between complete response/partial response and progressive disease and is a default creation due to the historical inability to precisely measure lesion change. It includes both responders and nonresponders. Thus, 29% regression to 19% growth for RECIST and 49% regression to 24% growth for WHO are measurements all classified as stable disease. More importantly, there is no agreement on duration of stable disease that is meaningful in most disease.

Specific diseases where RECIST has proven to be less useful in assessing response include gastrointestinal stromal tumors (GIST) and sarcoma. These diseases often require volumetric assessment of the tumor target volume and are difficult to assess using one- or two-dimensional measuring tools.

**GIST.** In evaluating the role of imatinib in the treatment of patients with GIST, Van den Abbeele et al. showed in 2002 that patients who responded by FDG-PET with a decreased standardized uptake value (SUV) to a value <2.5 had longer time to treatment failure than those who did not, whereas those who responded by SWOG criteria had the same time to treatment failure as those without response (10). Choi et al. showed that although the response rate measured was only 45% there were significant decreases in tumor density (reflecting decreased uptake of intravenous contrast by the tumor) and SUV on FDG-PET (11).

Choi et al. devised CT criteria to discriminate between good and poor responses comparable with that determined by FDG-PET and found that, when response was defined as a one-dimensional decrease in tumor size of 10% or a decrease in tumor density by 15%, all patients with poor response by FDG-PET were eliminated and all but one patient with response identified by FDG-PET was correctly identified (6). They subsequently studied a separate group of patients using the Choi et al. criteria and found that the response rate by Choi et al. criteria was 83% compared with 45% by conventional RECIST, that Choi et al. responders had significantly longer TTP than nonresponders, and that the RECIST definition of response could not separate a group with a longer TTP (Fig. 1). In addition, the Choi et al. responders had equivalent TTP as the RECIST responders, suggesting that the degree of apparent tumor shrinkage is not directly related to the extent of cell kill (Fig. 2). The determination of the reproducibility of the Choi et al. criteria and its utility in a prospective multicenter study will help evaluate its role in assessing response in GIST.

**Sarcoma.** Sarcomas of bone and soft tissue are a histologically and clinically heterogeneous group of malignancies. Certain sarcomas, such as synovial sarcoma, and Ewing’s sarcoma are composed principally of malignant cells, whereas other types, such as myxofibrosarcoma and myxoid liposarcoma, contain varying degrees of malignant cells interspersed with fibrous septa, myxoid stroma, and tissue necrosis. Intra-tumoral hemorrhage, inflammatory infiltrate, and cystic degeneration may also contribute to the bulk of the sarcoma mass lesion and confound interpretation of tumor response to treatment assessed by conventional radiography. RECIST may incorrectly classify sarcomas exhibiting significant histologic response to treatment as progressive disease (12). On the other hand, tumor response per RECIST may not correlate with a substantial histologic response to chemotherapy in sarcoma. In a study reported by Schuetze et al., only 3 of 6 cases classified by RECIST as response exhibited <10% residual viable neoplasm after chemotherapy. In addition, RECIST may be difficult to apply to sarcomas located predominantly in bone. The change in FDG uptake in sarcomas during treatment with chemotherapy has been evaluated as a potential surrogate measure of tumor response to treatment. A reduction in the
tumor maximum SUV (SUV$_{\text{max}}$) of FDG to <2% or >50% from pretreatment value has correlated with favorable histologic response of osteosarcoma to chemotherapy (13). A post-chemotherapy tumor SUV$_{\text{max}}$ of ≤2.5 has correlated with significantly better 4-year progression-free survival in Ewing's sarcoma (14) and a reduction in soft-tissue sarcoma SUV$_{\text{max}}$ of >40% from pretreatment levels independently correlated with a lower risk of relapse and improved survival in patients treated with neoadjuvant chemotherapy (15). Change in FDG uptake in sarcomas during chemotherapy appears to be promising as an early (after one to two cycles of chemotherapy) surrogate for clinical outcome in patients with localized disease.

The thresholds for classifying FDG-PET response in sarcoma have been arbitrarily chosen to an extent and need to be validated in appropriate prospective trials. In addition, the threshold for FDG-PET response in osteosarcoma may differ from Ewing's or soft-tissue sarcoma, and little to no data are available correlating sarcoma FDG uptake changes with survival in metastatic soft-tissue sarcoma. This underscores the need for collecting imaging data in a standardized fashion in therapeutic trials across histologies, so that there can be sufficient data for meaningful analyses.

**Functional Response Assessment Using FDG-PET**

Advances in molecular imaging are enabling the noninvasive evaluation of tumor biology, including measures such as the status of a lesion’s metabolism, proliferation, and oxygenation. Because FDG-PET has been the most widely studied in the context of response, the discussion, for the purposes of this article, is restricted to that tracer. FDG-PET has been evaluated in many studies as an indicator of therapeutic response in many tumors with both cytotoxic and cytostatic therapeutic regimens (16–18). The European Organisation for Research and Treatment of Cancer developed response assessment criteria for FDG-PET in a manner analogous to RECIST in 1999 to stratify responders to four categories: complete response, partial response, progressive disease, and stable disease (19). Prospective clinical trials sponsored by the National Cancer Institute are evaluating these criteria.

Studies have shown that the use of FDG-PET has not only identified early metabolic response to molecularly targeted therapy (20, 21) but has been predictive of progression-free survival (5, 22). The use of repeat FDG-PET imaging at select time points matching the pharmacokinetics of the therapeutic agent enables whole-body in vivo pharmacodynamic assessments. For example, this approach has been used to show the metabolic flare (increased FDG uptake in GIST) seen during the “off” periods of treatment cycles that confirmed that the metabolic response seen during the periods when the patient was on the drug was indeed evidence that the drug had hit the target (10–24). This type of noninvasive pharmacodynamic information provided by FDG-PET can potentially provide information that helps with the dosing schedule design of these drugs. Perhaps more importantly, lack of response by FDG-PET identifies patients who are not responding to the drug early after the initiation of the therapy. There is increasing evidence that a small patient population can provide sufficient information regarding “go/no go” decisions in a clinical trial if functional imaging is being used. The use of combined PET/CT scanners allows direct comparisons to be made between changes in anatomic size and glucose metabolism within the same tumor using intrinsically coregistered PET images. As noted before, in the evaluation of molecular targeted therapy, discordant findings can be seen between PET and CT and the clinical response (10, 18, 10–25). This is often due to lack of size change or increase in tumor size on CT despite apparent clinical benefit or lack of significant change in tumor size but appearance of intratumoral nodules. FDG-PET is helpful in resolving these issues by showing metabolic response despite stable anatomic disease or even increase in size (due to intratumoral hemorrhage) confirming response to therapy (23, 24). Conversely, FDG-PET can also show reemergence of glycolytic activity within an intratumoral nodule despite stable tumor size appearance consistent with secondary resistance to therapy. Emerging studies in a variety of tumors have shown...
that metabolic response assessment using changes in FDG-PET SUV do correlate with clinical outcome (17, 18, 22, 25).

Thus, an important role of FDG-PET has been to solve ambiguous findings on CT. Figures 3 and 4 provide examples of the types of problematic images that can benefit from concurrent FDG-PET studies. Problems in imaging following agents such as imatinib, sunitinib, sorafenib, and dasatinib include (a) lack of tumor shrinkage despite clinical benefit with sustained stable disease being predictive of progression-free survival and overall survival (5, 10), (b) false appearance of progressive disease based on increased tumor size due to intratumoral bleeding or appearance of new lesions in the liver following imatinib or sorafenib treatment that are actually consistent with response to therapy, and (c) appearance of new dense lesions within an existing tumor mass consistent with recurrence and/or resistance to treatment despite the lack of significant increase in tumor size (26).

Given that prior work with FDG-PET has been primarily in retrospective studies or prospective trials in single institutions, the clinical cooperative groups have initiated many clinical protocols in multiple disease sites evaluating the role of FDG-PET for both response assessment and validation of radiation therapy target volumes. Disease sites such as non-small cell lung cancer, head and neck cancer, and esophageal cancer have been the initial areas of evaluation in epithelial disease. An ongoing Children’s Oncology Group protocol evaluates children with intermediate risk Hodgkin’s disease with real-time response evaluation of both anatomic and metabolic images. The protocol attenuates therapy for patients deemed both rapid early responders to chemotherapy and those who have a complete response to chemotherapy deemed by imaging defined study criteria. Secondary and tertiary randomization points embedded in the protocol are defined and triggered by the response to therapy. This effort has been extended to both low-risk and high-risk patients with the same disease in studies about to open. Therefore, valuable information will be obtained through these studies concerning the important future role of advanced technology imaging in clinical trials and help better define the circumstances in which anatomic and/or metabolic/molecular imaging strategies should be applied in both epithelial and lymphoma clinical trials.

**Effective Use of Imaging in Phase II Trials**

The accuracy and reproducibility of imaging response measurements in multicenter trials require standardization of acquisition protocols, quality control, and image analysis. Evaluation and qualification of CT, PET/CT, and magnetic resonance imaging scanners are required before the acquisition of trial patient images to show adequate scanner performance,

---

**Fig. 4.** Scans of a 60-year-old woman with mesothelioma before and after treatment with dasatinib. The baseline scan (top left) shows a large pleural-based mass that shows marked decrease in density 1 mo following initiation of dasatinib (top right). This is consistent with therapeutic response despite the apparent increase in tumor size. At the 3-mo follow-up (bottom left), the size of the mass has decreased, but a new nodule has become apparent within the mass. This is consistent with clonal evolution of resistance as confirmed by the 4.5-mo follow-up scan (bottom right), showing further growth of the nodule within the mass. This was also confirmed by FDG-PET (data not shown) with reemergence of glycolytic activity within the intratumoral nodule.
appropriate quality-control procedures, and accuracy of measurements. To obtain comparable images from multiple centers, a standard acquisition protocol must be followed consistently for all patients at all sites. In the case of CT, the acquisition variables such as peak kilovoltage, milliamper, rotation speed, pitch, and collimation as well contrast administration (type, volume, timing, and bolus quality) should be consistent for patients scanned at an individual site and similar across different sites. The use of alternative quantitative measures of response, such as changes in tissue density measurements using CT Hounsfield units, requires greater emphasis on standardization of acquisition technique and patient preparation. Trials using quantitative FDG-PET studies have established requirements for standardization of FDG dose, timing of FDG uptake period, emission scanning duration, and patient preparation (e.g., fasting and reduced physical activity) as has been published in the National Cancer Institute guidelines for FDG-PET in clinical trials (27). Rigorous study monitoring and quality assurance of imaging data are required to ensure protocol compliance and consistency of results. It is equally important to perform central standardized image analysis to achieve consistent response assessments, especially in randomized phase II trials. Although FDG-PET has been successfully approved as a diagnostic modality for staging and restaging many solid tumors and lymphoma, breast cancer is the only histology in which the use of FDG-PET in assessing response is an approved indication by the Centers for Medicare and Medicaid Services. It is hoped that prospective clinical trials done in a standardized fashion will give us definitive answers regarding the relative merit of using FDG-PET as a response assessment tool. The issues that need to be evaluated include both the feasibility of having broad response assessment guidelines, which could work across histologies and therapeutic interventions, and the timing of the response assessment. Current clinical trials in progress in lung cancer, head and neck cancer, esophageal cancer, and lymphoma will help to further establish the role of metabolic imaging in the clinical trials process and likewise determine the nature of the partnership between anatomic imaging platforms and both metabolic and molecular counterparts in clinical trials moving forward. It is likely that fused image platforms will play an important role in response determination moving forward.

Improvements in the quality assurance of image acquisition will provide the optimal platform for uniform image interpretation. Nonuniform target interpretation can lead to ambiguity in clinical trial interpretation. Pediatric Oncology Group protocol 8,725 evaluated patients with advanced Hodgkin’s disease. All patients received chemotherapy with patients randomized to post-chemotherapy radiation therapy intended to be directed to all sites of preexisting disease. Initial evaluation showed no survival benefit to radiation therapy (28). Retrospective review of both anatomic images and radiation therapy treatment objects at Quality Assurance Review Center (QARC), however, showed a 31% noncompliance rate to study objectives concerning the radiation therapy target volume. Patients treated to radiation therapy targets compliant to study had a 10% survival benefit (29), suggesting that interpretation of the diagnostic images was of significant importance to defining the radiation therapy target volume. A similar finding was reported this year at American Society of Clinical Oncology in head and neck cancer (30). Therefore, quality assurance of both image acquisition and interpretation are very important aspects of imaging in the clinical trials process.

**Imaging Core Laboratories**

One of the strategies outlined in the 2005 Clinical Trials Working Group Restructuring the National Cancer Clinical Trials Enterprise Report is to enhance the standardization of tools and processes for trial design, data capture, and data sharing to decrease effort and minimize duplication. Protocol-based measures ensure clarity, consistency, appropriate credentialing, and quality assurance for studies. Consistent guidelines for treatment aims and data submission requirements are essential to the successful conduct of a clinical trial. Appropriate quality assurance of the delivered treatment enhances the validity of study endpoints. Institution-based measures (site credentialing) ensure sites have the equipment, expertise, and tools to participate in clinical trials. Site credentialing is monitored through benchmarks, which are test cases that require the participating facility to show the expertise required to perform the protocol evaluations and treatments.

The QARC was established in 1980 to serve the National Cancer Institute-sponsored cooperative groups by providing radiotherapy quality assurance and diagnostic imaging data management services. QARC conducts assessments of therapy and diagnostic imaging data to address study endpoints such as confirmation of staging, eligibility, response, progression, and/or relapse and correlation of patterns of failure. QARC performs real-time central review of response in several protocols where continued treatment is tailored based on response to initial therapy. This is a significant step forward for clinical cooperative group trials allowing multiple endpoints to be addressed in a trial in real time.

Today, there are several imaging core laboratories that have been established in both academic and commercial settings, underscoring the ever increasing role of imaging in clinical trials and the importance of core laboratories to ensure the quality of imaging in these studies using measures based on the protocol, institution, and subject. These include the core laboratories at the American College of Radiology Imaging Network,11 Cancer and Leukemia Group B,12 and Pediatric Brain Tumor Consortium13 at the cooperative group level as well as the Dana-Farber/Harvard Cancer Center Tumor Imaging Metrics Core. In addition, the Virtual Imaging Evaluation WorkSpace Consortium, which is a collaboration of American College of Radiology Imaging Network, Cancer and Leukemia Group B, and QARC, can potentially be used by any cooperative group to support the imaging needs of any multicenter study.

---

8 http://www.asco.org/
10 www.qarc.org
11 http://www.acrin.org/
12 http://www.calgb.org/
13 http://www.pbc.org/
14 http://tumormetrics.org/
It is expected that Virtual Imaging Evaluation Workspace will address credentialing institutions for imaging clinical trials and provide uniform guidelines and standards for imaging studies and the imaging data acquisition process. Issues associated with image interpretation and application of response criteria are likewise expected to be addressed by this collaboration. This will include computational issues associated with PET and other advanced imaging initiatives.

Conclusion

The effective use of imaging tools in phase II treatment trials can improve by using standardized image acquisition, analysis, and interpretation criteria such as the European Organisation for Research and Treatment of Cancer response criteria for FDG-PET (19, 25, 27). With more molecular and functional imaging techniques showing promise as better evaluators of disease state, strategies for increased standardization, procedures for trial design, and data sharing will help accelerate the evaluation of these tools and future implementation in both trial and clinical settings.

References

20. Van den Abbeele AD; for the GIST Collaborative PET Study Group (Dana-Farber Cancer Institute, Boston, Massachusetts; OHSU, Portland, Oregon; Helsinki University Central Hospital, Turku University Central Hos-

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.
Considerations for the Use of Imaging Tools for Phase II Treatment Trials in Oncology


Updated version  Access the most recent version of this article at:
http://clincancerres.aacrjournals.org/content/15/6/1891

Cited articles  This article cites 29 articles, 15 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/15/6/1891.full.html#ref-list-1

Citing articles  This article has been cited by 12 HighWire-hosted articles. Access the articles at:
/content/15/6/1891.full.html#related-urls

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.