The Imaging Viewpoint: How Imaging Affects Determination of Progression-Free Survival

Daniel Carl Sullivan1, Lawrence H. Schwartz2, and Binsheng Zhao2

Abstract

Tumor measurements on computed tomographic or MRI scans and/or the appearance of new lesions on any of a variety of imaging studies including positron emission tomographic scans are key determinants for assessing progression-free survival as an endpoint in many clinical trials of therapies for solid tumors. Test–retest tumor measurement reproducibility may vary considerably across serial scans on the same patient unless rigorous attention is paid to standardization of image acquisition parameters and unless measurements are made by trained, experienced observers using validated objective methods. Target lesion selection also must be done with care to choose lesions that are or will be reproducibly measurable. Likewise, new lesions will be missed or misinterpreted on follow-up imaging studies unless those imaging studies are obtained using techniques suitable for detecting early, small lesions. Reader variability is clearly a major component of the problem. The increasing availability of semiautomatic image processing algorithms will help ameliorate that issue. In addition, an array of internationally accepted guidelines, standards, and accreditation programs now exist to help address these problems. Clin Cancer Res; 19(10); 2621–8. ©2013 AACR.

Introduction

From an imaging perspective, tumor progression is generally defined as either an increase in tumor burden (as measured on imaging studies) or the detection of new lesions seen on serial imaging studies, and tumor response refers to tumors getting smaller or disappearing. During the era of cytotoxic chemotherapy, which began in the late 1940s and 1950s, “response rate” (RR)—that is, the proportion of tumors that shrunk—was the standard endpoint for clinical trials of cancer therapies. Measurements of solid tumors were generally determined from imaging studies, and therefore guidelines, criteria, and classifications based on measurements from imaging studies were developed in the 1980s and 1990s (1, 2). These guidelines and classifications reflected the emphasis on overall response rate that was the norm at the time.

In contrast, targeted chemotherapies that have emerged in the past 15 years interfere with signaling pathways and inhibit cell growth but do not necessarily lead to cell death or significant tumor shrinkage. Accordingly, oncologists became interested in the length of time that a cancer did not grow or metastasize further, i.e., the length of time before “progression” occurred. In addition, more effective salvage therapies and crossover trial designs have become increasingly used. Thus, over the past decade, progression-free survival (PFS) became a common endpoint for cancer therapy trials (3). The concepts of PFS and methodologic issues associated with PFS in clinical trials are considered in more detail in other CCR Focus articles in this journal issue (4, 5).

During this same period, imaging technology underwent dramatic changes. Cross-sectional and 3-dimensional imaging from computed tomography (CT) and MRI replaced much standard radiography (such as chest radiographs), and sensitive functional imaging methods [e.g., 18F-fluorodeoxyglucose (FDG)-positron emission tomographic (PET) scans] and hybrid devices (e.g., PET/CT scanners) replaced low-resolution nuclear medicine cameras. Anatomic measurements can now be done with much more precision, and early metastatic lesions can be detected with much greater confidence and at earlier stages than was true even a decade ago.

The development of molecularly targeted cancer therapies has stimulated interest in functional imaging modalities to assess tumor change in addition to or instead of measurements based on morphology. However, although dynamic contrast-enhanced (DCE) MRI and PET with promising radiotracers have shown great potential for early identification of drug efficacy because of their ability to depict tumor change at functional and molecular levels, their use for reliably reflecting clinically meaningful biologic change has not yet been satisfactorily validated. CT, on the contrary, is globally accessible; affordable; has mature, reliable, and reproducible technology; and can clearly

Authors' Affiliations: 1Department of Radiology, Duke University Medical Center, Durham, North Carolina; and 2Department of Radiology, Columbia University, New York, NY

Corresponding Author: Daniel Carl Sullivan, Department of Radiology, Duke University, Box 3302, DUMC Durham, NC 27710. Phone: 919-681-8081; Fax: 919-684-7122; E-mail: daniel.sullivan@duke.edu
doi: 10.1158/1078-0432.CCR-12-23936
©2013 American Association for Cancer Research.
depict both normal and abnormal anatomy. For these
reasons, CT is predominantly used in clinical trials and
clinical practice to monitor tumor response and progression
during the course of therapy. Conventional MRI plays
an increasing role in the detection and quantification of
some lesions, especially at the anatomic sites where MRI is
superior to CT (such as the brain), or in patients with
contraindications to CT contrast material or radiation.
Thus, in this article, we focus on CT, with additional com-
ments related to MRI, but the need to improve imaging
measurement reproducibility in clinical trials and clinical
practice applies to all imaging modalities.

Proper assessments of PFS matter to the field of drug
development and to individual patients with cancer. If
“progression” is determined, it usually results in a change
in therapy for a patient in a clinical trial. If there are a large
number of wrong assessments about progression in a clin-
ical trial, the overall conclusion of the trial may be wrong.
Sridhara and colleagues discuss some of these issues in
another article in this CCR Focus edition (6). It is also
increasingly the case that although Response Evaluation
Criteria in Solid Tumors (RECIST) were developed to be a
clinical trial tool, not a clinical practice guideline, oncolo-
gists will use the presence of “progression” as a justification
or trigger to switch to a different therapy in routine clini-
care, now that an increasing number of alternative
therapies are available. Thus, proper assessment of progres-
sion is of paramount importance.

Care of patients with cancer includes elements of both art
and science, but it is becoming more of a science and less of
an art. Therefore, objective, reproducible data are needed on
which to base treatment decisions. And it has always been
true that objective, reproducible endpoints are necessary in
clinical trials. Therefore, the measurements obtained from
imaging scans must also be as reproducible as possible. It
should not matter what hospital a patient goes to, or on
what day, or on which manufacturer’s scanner he or she is
scanned, the result should be the same. To achieve this,
there has to be a rigorous standardization and attention to
detail with respect to all the hardware, software, and human
parameters that can introduce variability.

Imaging Physics and Quantification Issues

Villaruz and Socinski describe the historical backdrop
against which RECIST was developed, as well as some of its
limitations and practical considerations about tumor mea-
surement, especially from the oncologist’s perspective (7). Ac-
According to current RECIST guidelines, the development
of progressive disease is defined as “at least a 20% increase in
the sum of the longest diameter of target lesions, taking as
reference the smallest sum longest diameter recorded since
the treatment started, the appearance of one or more new
lesions, or unequivocal progression of existing nontarget
lesions” (8). As can be inferred from the definition, accurate
and reproducible measurements of lesion size, as well as
size changes and sensitive detection of new lesions over
serial imaging studies, are critical to determining progres-
sion. It should be remembered that not all large primary or
metastatic solid tumors can be measured on CT or MRI and
that, in some cases, size changes or lack of change do not
correlate with progression or regression. Figure 1 shows a
couple of examples. The lung cancer shown in Fig. 1A and B
abuts the pleura, and it is not clear where all of the tumor
margins are. Clinically, this patient was considered stable
over the period of 8 weeks, but it is difficult to tell that
from any measurements obtained from the CT scans.
Figure 1C and D show the development of necrosis in an
adrenal metastasis. Clinically, this patient was consid-
ered a responder, but the measurements do not reflect a
decrease in size.

Target lesion selection may be difficult and can account
for variability in response assessments. In general, it is best
to measure large lesions that are or will be reproducibly
measurable. The RECIST guidelines say that “Target lesions
should be selected on the basis of their size (lesions with the
longest diameter), be representative of all involved organs,
but in addition should be those that lend themselves to
reproducible repeated measurements. It may be the case
that, on occasion, the largest lesion does not lend itself to
reproducible measurement in which circumstance the next
largest lesion which can be measured reproducibly should
be selected” (8). The reasons that the largest lesions might
not always be reproducibly measurable usually have to do
with their location, for example, abutting or invading (or
likely to soon abut or invade) the mediastinum, pleura, or
abdominal organs. There is a bit of an “art” to the process
of target lesion selection and measurement. Care should
be taken to measure lesions in multiple organs, and
lesions should be definitively metastatic. If there is uncer-
tainty about whether a lesion is metastatic or not, it
should be considered non–target disease and excluded
from measurement.

Many factors affect the measurement of target lesions
and the detection of new lesions, ranging from the choice
of imaging modalities, imaging acquisition techniques,
and image reconstruction parameters to variability related
to readers’ differing expertise and measurement methods.
A multicenter renal cancer clinical trial conducted in the
early 1990s reported a 40% major disagreement (de-
defined as, e.g., modifications from response to progression)
between responses assessed by the trial investigators
and an independent evaluation committee (9). Key fac-
tors contributing to such a high degree of inconsistency in
response assessment were quality and reproducibility of
radiologic examinations, selection of target lesions, and
human subjectivity in measuring lesions. Another exam-
ple is a study on non–small cell lung cancer (NSCLC),
which showed that the probability of misclassifying
a tumor could be as high as 43% [by World Health
Organization (WHO) criteria] and 30% (by RECIST) for
progression when tumors were measured by multiple
radiologists. However, the misclassifying rates could be
as low as 3.0% (WHO) and 2.5% (RECIST) for
response if the tumors were measured by the same radiologist (10).
Several other studies have shown similar data on tumor
measurement variability.
The choice of imaging modality and timing of image acquisition can also have an influence on the detection of new lesion(s) (11–14). For instance, density contrast between a hepatic lesion and its surrounding liver parenchyma on contrast-enhanced CT scan images changes during the time following injection. In contrast, as the liver moves from the arterial to the venous system, the phases of contrast are referred to as arterial, equilibrium, and portal. Images acquired in one phase of the vascular distribution (e.g., later phases) can better capture density differences between normal and abnormal tissue and thus better visualize (and quantify as well) the lesion than images acquired at another (e.g., early) phase. Figure 2 is an example showing how imaging modality and acquisition technique can affect the detection of liver metastasis in a patient with rectal carcinoma (11). The liver metastasis (arrow in Fig. 2D) was not depicted on the arterial-phase and equilibrium-phase images of the contrast-enhanced CT scan (Fig. 2A and C). There was a suspicious small, low-attenuation lesion seen on the portal venous phase image of the CT scan (Fig. 2B). However, this small metastasis was clearly captured by the contrast-enhanced portal-phase MRI scans (Fig. 2D and E).

Imaging reconstruction parameters, for example, CT slice thickness and reconstruction algorithm, also affect the detection of new lesions, especially if the lesions are small (15–18). For example, the thicker the slice interval, the more partial-volume artifact there will be. Thus, lesions smaller than the slice thickness may not be visible on images reconstructed with thick slice intervals. Figure 3 presents an example in which 2 small pulmonary nodules, one measuring 2 mm and the other measuring 5 mm (circles in Fig. 3A), were clearly depicted on a thin-section CT image of 1.25 mm. However, on a thick-section image of 5 mm, the smaller (2 mm) nodule in the right upper lobe was missing and the larger one (5 mm) in the left lower lobe was hardly visible (15). The above examples show that improper use of imaging modality, imaging technique, and/or imaging reconstruction parameters in oncology clinical trials can lead to missing or delayed detection of new (metastatic) lesions and thus misinterpretation of the time point at which a disease progresses.

Technical factors in the image acquisition process that are known to influence lesion size measurement and thus anatomic response assessment include scanner differences, intravenous contrast administration, type, volume, timing, injection rate, and CT scan beam settings (19–27). During scan acquisition, such patient-related factors as the phase of respiration during which the image is acquired...
and whether or not the patient can suspend respiration also play a role. If the patient cannot hold his or her breath for the duration of the scan (<30 seconds), the normal and abnormal structures may be blurred. This will generally cause lesions to be measured larger than they truly are and can cause small lesions to be missed completely.

Figure 4 shows an example from a recent study investigating the effects of CT slice thickness on the measurements of solid tumors in phase I/II oncology clinical trials (27). Figure 4 shows a liver metastasis on CT images reconstructed using 3 different slice intervals, i.e., 1.25, 2.5, and 5 mm. Although thin slices can depict more imaging details than thick slices, they have higher noise level, making lesion boundaries harder to identify and thereby decreasing measurement reproducibility. This is especially a problem in measurement of abdominal metastases or visceral lesions, where there is less density contrast between the lesion and the surrounding normal structures. Thus, to reduce measurement variability, it is important that CT scan images be reconstructed with the same slice thickness on serial scans for a given patient.

Currently, hybrid PET/CT scanners have almost completely replaced stand-alone PET scanners. Therefore, patients receiving an FDG-PET scan always receive a CT scan at the same time and the question comes up whether that CT scan could be used for diagnostic interpretations and RECIST measurements. Unfortunately however, the CT component of a PET/CT is usually conducted solely for the purpose of attenuation correction for the PET portion of the scan and is not a diagnostic-quality CT scan. Measurements on this type of CT scan would be subject to considerable inaccuracies. If a diagnostic CT is conducted with the PET scan, which is an option that can be done, then standard measurements could be obtained, assuming the CT scan is conducted with the appropriate slice thickness.

Image Interpretation Issues

Reader-related factors are another significant source of variability in the interpretation and measurement of target lesions (10, 28–32). These factors include (but are not limited to) the radiologist’s expertise and skill in the quantitative and qualitative evaluation of therapy response, measurement bias due to the reader’s systematic over- or underinterpretation of tumor shrinkage, bias due to reader’s knowledge of treatment assignment, random discrepancies in the measurements due to intra- and interreader variability, and human errors that can be caused by tracking different target lesions over time and overlooking development of a new lesion(s). Moreover, choice of measurement technique (e.g., unidimensional, bidimensional, volumetric), measurement method used to determine tumor measurements (e.g., electronic calipers, automated techniques), and measurement environment (e.g., display settings on diagnostic workstations, often referred to as “windowing”) can also affect the accuracy and reproducibility of a radiologist’s measurements (33). Lesions that are difficult to measure (e.g., because of indistinct or obscured lesion margins or lesions...
with heterogeneous densities) can further worsen the accuracy and reproducibility of the lesion measurements.

Recent studies have begun to more systematically evaluate the relative contribution of sources of variability to better understand where corrective efforts should be directed to have the biggest impact. One prominent study is the same-day repeat CT study designed to investigate the minimum change that could be detected by modern

Figure 3. Effect of CT slice thickness on lesion detectability. A, the 1.25-mm thin-slice technique shows 2 confidently seen pulmonary nodules of 2 and 5 mm in diameter (small size) in the anterior segment of the right upper lobe and in segment 6 of the left lower lobe (circles). Two readers diagnosed the nodules confidently in the 1.25-mm slice. B, in 5-mm slice thickness, the nodule in the anterior segment of the right upper lobe is not detectable and the one in segment 6 of the left lower lobe is hardly visible. Both readers diagnosed the patient as negative when reading only the 5-mm slices. Image reproduced with permission from Fischbach et al. (15).

Figure 4. Effects of CT slice thickness on the lesion measurement. From left to right, a liver lesion shown on 1.25-, 2.5-, and 5-mm slice thickness images. The chart shows measurements made by a radiologist blinded to the slice information on the 3 slice thickness images. "Uni" represents the greatest diameter, and "Perp" is the greatest diameter perpendicular to Uni of the lesion that was measured. "Bi" is the product of the 2 diameters, Uni and Perp. In this case, the maximum percentage difference in the greatest diameter between the 2 slice thicknesses is more than 10%. Because truth is not known, it is not possible to determine which of the 3 sets of measurements is more correct. However, to reduce variability, it is apparent that the same slice thickness should be used for each sequential CT scan that a given patient receives. Image reproduced with permission from Zhao et al. (32).
CT scanners using advanced measurement methods (23, 26). The study showed the levels of measurement errors or variability in unidimensional, bidimensional, and volumetric techniques. Table 1 shows the (unidimensional) measurement variability between 2 repeat scans conducted within a few minutes by lesion size; the smaller the lesion, the larger the measurement variability (26). The findings from these studies on sources of variability can not only help to determine the appropriate threshold values to distinguish real (biologic) tumor changes from measurement errors (inherent variability of the technique) but also to identify and suggest the most appropriate imaging acquisition and reconstruction parameters as well as the measurement technique (e.g., unidimensional, bidimensional, or volumetric) to optimally monitor tumor response or progression.

Volumetric techniques show promise for more accurate and earlier detection of tumor changes than the tumor diameter technique does (34, 35). Figure 5 is an example taken from a phase II study correlating early radiographic changes with EGF receptor (EGFR) mutation status in patients with NSCLC. For this EGFR-mutant lung cancer tumor, a significant change was detected using volume measurement (−52.4%) but not using unidimensional measurement (−4.4%).

To reduce variability in the measurement of target lesions and improve detectability of new lesions, continuous efforts are being made to train radiologists, standardize reading workflow, harmonize imaging acquisition techniques and reconstruction parameters, improve image visualization ability, optimize tumor measurement methods, and incorporate complementary imaging modalities. Response assessment methods have been modified and continue to evolve as well (8, 36–40). The Radiologic Society of North America (RSNA) supports the Quantitative Imaging Biomarkers Alliance (QIBA), which has published documents.

Table 1. Measurement variability between 2 repeat CT scans

<table>
<thead>
<tr>
<th>Tumor size category, cm</th>
<th>SD, mm</th>
<th>Size, cm</th>
<th>Range as a result of variability, cm</th>
<th>% change as a result of variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–3</td>
<td>2.0</td>
<td>2</td>
<td>1.6–2.4</td>
<td>±20</td>
</tr>
<tr>
<td>3–5</td>
<td>2.3</td>
<td>4</td>
<td>3.5–4.5</td>
<td>±12</td>
</tr>
<tr>
<td>5–8</td>
<td>3.3</td>
<td>7</td>
<td>6.3–7.7</td>
<td>±9</td>
</tr>
</tbody>
</table>

NOTE: For a lesion that in fact measures 4 cm, for example, the variability of CT imaging can lead to measurements ranging from 3.5 to 4.5 cm. Differences in measurement variability are dependent on lesion size, as calculated from repeat CT scans performed within 15 minutes of each other. Measurement variability is greater for smaller lesions.

*As calculated from the 95% limits of agreement.

Table reproduced with permission from Oxnard et al. (26).

Figure 5. Volumetry as an early metric of tumor response. Volumetric and unidimensional measurements from CT were compared for detection of tumor change on an EGFR mutant tumor at baseline (A) and at 20-day follow-up (B). Computer-delineated tumor contours and diameter lines are superimposed on one image from each study date. Three-dimensional view of each segmented tumor is displayed at the top left corner of each panel. The volumetric measurements at 20-day follow-up were statistically significantly different, whereas the linear dimensions were not. Image reproduced from Zhao et al. (34).
Imaging and Progression-Free Survival

(http://rsna.org/QIBA.aspx) for standardizing image acquisition for CT volumetry, FDG-PET standard uptake value, and dynamic contrast-enhanced MRI (41).

Strategies to minimize these issues and lead to more consistent tumor response assessments by imaging include the following:

1. Implement a scanner calibration and quality assessment program at each clinical site. Two such accreditation programs that can be used are the NCI-supported Centers of Quantitative Imaging Excellence (CQIE) program and the SNMMI-supported Clinical Trials Network Site Qualification and Scanner Validation programs.
2. Ensure that the same single reader evaluates the full set of examinations for any given patient.
3. Provide images to readers in the sequence they were obtained clinically (reader consistency is improved when the reader is presented with images in a time-ordered manner).
4. Ensure that the same representative target lesions are measured at each time point.

The RECIST Working Group continues to meet annually to evaluate new evidence and consider revisions or improvements to RECIST (8). For analyses regarding linear (unidimensional) tumor measurements, the RECIST Working Group relies on databases in which the linear measurements have been obtained from clinical trial case report forms and are subject to all of the potential measurement errors and variables described above. We believe the organizations supporting the RECIST Working Group would be well-served by including in their databases the original CT or MRI scans, so that the quality of the original scan data can be assessed, as well as multiple measurements be obtained from different readers and/or by different means. In this way, the variability associated with each measurement will be known and the strength of conclusions about number of target lesions or measurement thresholds can be better assessed. It is also generally recognized by most oncologists that simple anatomic measurements are not an ideal way, the variability associated with each measurement will be known and the strength of conclusions about number of target lesions or measurement thresholds can be better assessed, as well as multiple measurements be obtained from different readers and/or by different means. In this way, the variability associated with each measurement will be known and the strength of conclusions about number of target lesions or measurement thresholds can be better assessed. It is also generally recognized by most oncologists that simple anatomic measurements are not an ideal indicator of tumor burden and that additional information that may be obtained from functional imaging would likely be helpful. The RECIST Working Group is evaluating the potential of FDG-PET scans to provide additional validated information on tumor response. We believe this is an important future direction. In addition, interest has been expressed in trying to take advantage of the full range of tumor response as a continuous variable rather than categorizing patients into one of a few categories, such as the 4 response categories of RECIST. For example, there is interest in displaying tumor size changes as waterfall plots, but statistical methods for drawing conclusions about response from such plots have not yet been generally accepted.

Conclusions
End point evaluation is influenced by scan variability, by the clinical time point at which the scan was obtained, and by the individual reviewing the image and the ordered sequence in which he or she reviews it. Measurement variability adds noise to clinical trial results, leading to more difficulty and reduced power for detecting true treatment effects. However, important efforts have been made and continue to be made to improve consistency of radiologic endpoint assessment by standardizing image acquisition across sites, devices, and time and by minimizing subjectivity in interpretation by preferentially using trained, experienced observers and more objective measurement methods.

Disclosure of Potential Conflicts of Interest
L.H. Schwartz is a consultant/advisory board member of Novartis, GSK, and Cerulean. The other authors disclosed no potential conflicts of interest.

Authors’ Contributions
Conception and design: D.C. Sullivan, L.H. Schwartz, B. Zhao
Development of methodology: D.C. Sullivan, L.H. Schwartz, B. Zhao
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): L.H. Schwartz, B. Zhao
Writing, review, and/or revision of the manuscript: D.C. Sullivan, L.H. Schwartz, B. Zhao
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): D.C. Sullivan, L.H. Schwartz
Study supervision: D.C. Sullivan, L.H. Schwartz

Received December 3, 2012; revised February 22, 2013; accepted March 25, 2013; published online May 13, 2013.

References


