Translational Oncology
www.transonc.com

Volume 2 Number 4 December 2009 pp. 223–230 223

PET/CT Assessment of Response to Therapy: Tumor Change Measurement, Truth Data, and Error[1]

Paul E. Kinahan*, Robert K. Doot*, Michelle Wanner-Roybal*, Luc M. Bidaut†, Samuel G. Armato III‡, Charles R. Meyer§ and Geoffrey McLennan¶

*Department of Radiology, University of Washington, Seattle, WA, USA; †Department of Imaging Physics, Division of Diagnostic Imaging, UT-MD Anderson Cancer Center, Houston, TX, USA; ‡Department of Radiology, University of Chicago, Chicago, IL, USA; §Department of Radiology, University of Michigan, Ann Arbor, MI, USA; ¶Department of Internal Medicine, School of Medicine, University of Iowa, Iowa City, IA, USA

Abstract

We describe methods and issues that are relevant to the measurement of change in tumor uptake of ^{18}F-fluorodeoxyglucose (FDG) or other radiotracers, as measured from positron emission tomography/computed tomography (PET/CT) images, and how this would relate to the establishment of PET/CT tumor imaging as a biomarker of patient response to therapy. The primary focus is on the uptake of FDG by lung tumors, but the approach can be applied to diseases other than lung cancer and to tracers other than FDG. The first issue addressed is the sources of bias and variance in the measurement of tumor uptake of FDG, and where there are still gaps in our knowledge. These are discussed in the context of measurement variation and how these would relate to the early detection of response to therapy. Some of the research efforts currently underway to identify the magnitude of some of these sources of error are described. In addition, we describe resources for these investigations that are being made available through the Reference Image Database for the Evaluation of Response project. Measures derived from PET image data that might be predictive of patient response as well as the additional issues that each of these metrics may encounter are described briefly. The relationship between individual patient response to therapy and utility for multicenter trials is discussed. We conclude with a discussion of moving from assessing measurement variation to the steps necessary to establish the efficacy of PET/CT imaging as a biomarker for response.

Translational Oncology (2009) 2, 223–230

Introduction: Lung Cancer, Assessment of Treatment Response with Positron Emission Tomography/Computed Tomography

Anatomical imaging with x-ray computed tomography (CT) scanners using the Response Evaluation Criteria in Solid Tumors [1] is the standard technique for evaluating the response of lung cancer to most therapies. The benefits and limitations of this approach are discussed elsewhere in this issue in the article by McNitt-Gray et al. [2]. Imaging of lung cancer with combined positron emission tomography (PET) and CT scanners has recently become a standard component of oncology diagnosis and staging [3]. In particular, PET/CT imaging of ^{18}F-fluorodeoxyglucose (FDG) uptake allows more accurate detection of both nodal and distant forms of metastatic disease [4], and tumor stage is still the most important prognostic factor for predicting the survival of patients with non–small cell lung cancer [5], the most common form of lung cancer. Furthermore, with the development of cytostatic therapies, metabolic status may be a better indicator of response than

Address all correspondence to: Paul E. Kinahan, PhD, Professor of Radiology, Bioengineering and Electrical Engineering, Director, PET/CT Physics, University of Washington, 222 Old Fisheries Center (FIS), Box 357987, Seattle, WA 98195-7987. E-mail: kinahan@u.washington.edu

[1]This work was supported by US National Cancer Institute contract number 24XS036-004 (RIDER), US National Institutes of Health under grant numbers CA74135 and CA115870, and a Society of Nuclear Medicine Student Fellowship awarded to Robert Doot.

Received 15 August 2009; Revised 15 August 2009; Accepted 17 August 2009

DOI 10.1593/tlo.09223

anatomical size changes [6]. For these reasons, and the urgent need for more effective therapies, PET/CT oncology imaging is being increasingly used for quantitative assessment of individual response to therapy and for clinical trials of novel lung cancer therapies.

Although PET imaging has the *potential* to produce quantitatively accurate images of tracer uptake, there is often an unknown global bias. This is likely a consequence of the primary role of PET imaging, which is clinical diagnosis and staging of cancer. For this purpose, it is the relative image fidelity that is of paramount importance. As stated by Coleman in a 2002 editorial [7].

> The answer to the question "Is quantitation necessary for clinical oncological PET studies interpreted by physicians with experience in interpreting PET images?" is "no." Can quantitation be useful in interpreting clinical PET images for the inexperienced observer? The answer is "yes." Image quantitation will become increasingly important in determining the effect of therapy in many malignancies.

Whereas the staging studies used routinely in clinical practice may not depend on accurate image quantification, quantitative measures are essential for the assessment of therapeutic response [8,9]. While inconsistent and non-optimized image quantification has a limited impact on the interpretation of staging FDG PET/CT scans in clinical practice, improper image quantification seriously degrades the utility of FDG PET as a dynamic measure in cancer therapy trials. The disconnection between the use of PET for clinical imaging and its unrealized potential for clinical trials is a point of considerable frustration for both imagers and oncologists [6].

Now, however, there are compelling reasons to understand and improve the quantitative accuracy of PET imaging. As noted above, there is a role for quantitative accuracy of measurements from PET images to determine response to therapy. In addition to CT imaging, pharmaceutical companies are now using quantitative PET/CT imaging to evaluate potential therapies. At the clinical level, the relative tracer uptake by a lesion, called the standardized uptake value (SUV), is now routinely reported. If the SUV of a lesion is reported, it is reasonable to expect that this is an accurate value, or at least that its precision and variance are understood.

In this article, we review the modality-specific factors that distinguish PET from CT and magnetic resonance imaging (MRI), what is known about the bias and variability of PET scanner measurements, both singly and in multicenter combinations, as well as the overall impact on multicenter trials. We summarize contributions made to the Reference Image Database for the Evaluation of Response (RIDER) project as well as some efforts by other groups. We conclude with a brief analysis of what links are missing in the chain of data acquisition and analysis of quantitative PET data used for clinical trials and the assessment of response to therapy.

Measurement Issues Specific to PET/CT

Imaging Characteristics of FDG PET/CT

The fundamental role of a PET/CT scanner is to form an image of the spatially varying concentration of positron emitters, for example, ^{18}F. With a half-life of 110 minutes, ^{18}F is favorable in the time scale of production and the ratio of patient radiation dose to image signal-to-background noise ratio, too. A positron emitter is attached to a biologic substrate of interest, for example, the glucose analog deoxyglucose, to become the radiolabeled tracer FDG. The use of FDG for oncology imaging accounts for approximately 90% of all PET/CT imaging procedures. The effectiveness of FDG stems from it being essentially "trapped" in the glycolytic pathway after the early step of phosphorylation by the enzyme hexokinase. Phosphorylated FDG does not cross cellular membrane barriers owing to the negative charge of the added phosphate group and is not further metabolized by glycolytic enzymes owing to structural differences between FDG and glucose. Thus, the rate of accumulation of FDG conveys information about the rate of glucose metabolism. Because cancer generally has a higher glucose metabolism than surrounding normal tissue [10], FDG is useful in cancer detection and staging. Furthermore, changes in measured FDG accumulation have been shown to be useful as a biomarker for mechanistic effects in response to therapy [6]. As a result, FDG imaging with PET is increasingly being used to guide therapy [11] and as an indicator of response in clinical trials [12,13], although more data on the impact on patient outcomes are still needed [6,8,9].

With corrections for physical effects, notably attenuation and the detection of scattered and random coincidences, accurate estimates of the spatially varying concentration of positron emitters (e.g., FDG) can be obtained. Typical measured units are kilobecquerel per milliliter. The measured concentration, on average, depends on the amount of activity injected and volume of distribution (e.g., patient size). To account for variations in the injected dose and patient size, generally preferred units are SUVs defined as SUV = $R/(D'/\hat{V})$ where R (kBq/ml) is the activity concentration at each point), D' (kBq) is the decay-corrected injected dose, and $\hat{V}$ is a surrogate for the true volume of distribution of tracer inside the body. Typically, patient weight (g) is used as a surrogate for the volume of distribution, in which case the SUV units are grams per milliliter. Because adipose tissue, with the exception of brown fat, does not normally take up significant amounts of FDG, the estimated lean body mass [14] or body surface area [15] is sometimes used instead of weight.

With proper calibration, and if data corrections (attenuation and scatter, etc.) are working properly, the reconstructed PET image represents a quantitatively accurate map of FDG concentration, which in turn is related to relative metabolic activity, that is, a *functional* image. In contrast, a CT image provides an accurate *anatomical* image. In a combined PET/CT scanner, the CT image provides precise localization of regions of FDG uptake, significantly aiding interpretation. The combination has proven so successful that no PET-only scanners used in oncology imaging are now manufactured [16].

The physics of the detection process and consequent scanner designs dictate that PET has good positional accuracy, like CT, but low spatial resolution, compared with CT and MRI. Results are PET images that are relatively blurry. In addition, PET scanner data often suffer from high degrees of statistical noise, particularly in thicker, that is, obese, patients and/or with short acquisition durations. To reduce image noise, additional smoothing is applied, further degrading image resolution and increasing noise correlations in the PET image. Both of these effects lead to size-dependent errors in measuring SUVs, that is, as the size of a focal accumulation, or "hot spot," decreases, the measured SUV decreases. This size-dependent loss in accuracy is often called the partial volume effect or error [17]. With a typical operational resolution (i.e., not the best possible resolution), objects smaller than 2 to 3 cm are susceptible to partial volume errors, as discussed below. Another important property of PET imaging is the high sensitivity, which allows routine measuring of micromolar and nanomolar amounts of radiotracer. In terms of contrast, FDG PET images are mixed in the sense that most normal tissues

have similar levels of uptake, making it difficult to distinguish different organs. In some cases, however, biomolecular activities, such as in most malignant cancers, have a significantly higher level of FDG uptake. Both of these contrast levels are visible in the PET image of Figure 1.

In summary, and to contrast with the other tomographic medical imaging methods (ultrasound, CT, MR, and single photon emission computed tomography), FDG PET images have a high level of sensitivity and very good contrast for most cancers. With proper calibration and corrections for attenuation, scatter, and other effects, PET images accurately represent internal tracer concentrations, expressed as SUVs, which are subject to partial volume errors. In addition, PET images have very good positional accuracy (i.e., no geometrical distortion as can happen in ultrasound or MRI) but low positional precision (i.e., the images have relatively low resolution). The images are typically noisy, with a high level of noise correlations introduced by image reconstruction and smoothing [18]. The combination of PET and CT in a single scanner is particularly felicitous because both modalities have excellent positional accuracy and provide strongly complementary information useful in oncology imaging. In addition, the CT image can be used for attenuation correction of the PET raw data [19], which shortens overall patient scan times and provides other benefits to the PET imaging process, as described elsewhere [20].

Bias and Variance in PET/CT Imaging

In the accompanying article by Meyer et al. [21], precise and accurate definitions of variance and bias are given. With respect to PET imaging, it is possible to use phantoms to accurately measure both features, as described below.

For clinical imaging, determining bias becomes a much more challenging problem due to the lack of ground truth. Less appreciated perhaps is that it is also very challenging to use phantoms to mimic or estimate the bias that occurs in clinical scanners owing to the difficulty of constructing test phantoms that accurately represent both the spatially varying distributions of both radiotracer concentration and attenuation coefficients. In other words, phantom measurements can provide a general guide, but without further confirmation, they cannot be regarded as an accurate representation of bias in clinical imaging.

Variance in clinical imaging can be estimated through the use of repeat scans where it is assumed that there are no changes in the object being imaged. At a minimum, repeat scans on patients have shown that there is a sample SD of approximately 10% to 13% in well-controlled single-scanner tests and when scanning the same patient (not undergoing therapy) within a few days [22–26]. These are akin to the "coffee-break" concept where the patient is imaged, takes a short break, and is reimaged with everything else remaining the same. With PET patient imaging, however, scans occur at a minimum of a day apart to allow for complete radioactive decay of the FDG. The variance in this case is the result of several factors, including variability in the patient, scanner, and procedures. The same studies also showed that the distribution of differences is Gaussian, although this is for the case of no true biologic change. In the case where there has been true biologic change, there may be a different distribution and magnitude of differences, as discussed by Meyer et al. in this issue [21]. Such studies do not address the variance present in longitudinal studies, where there is true change, changes in analysis methods, and multicenter studies.

Sources of Bias and Variance in PET/CT Imaging

A comprehensive survey of the many factors that affect bias and/or variance was recently presented by Boellaard [27]. Here, we summarize the major components: Scanner related issues, patient related factors, protocol variations, analysis methods, and interactions between sources of variance.

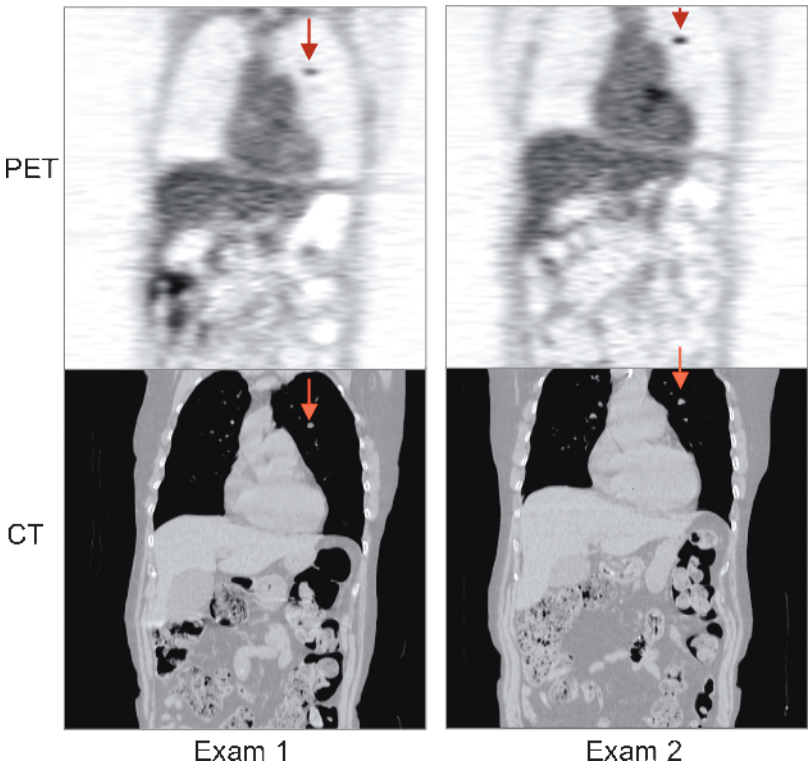

Figure 1. Sample images from the RIDER PET/CT collection. Serial coronal sections of PET/CT images of a patient with lung cancer (arrows).

Scanner components of bias and variance. The major scanner-specific factors that can increase bias and variance are listed in Table 1. Note that the amount of bias and variance introduced by physical effects is a complex interaction between the magnitude of the effect and the correction steps. For example, contamination of raw data by random coincidences adds both bias and noise. Two different methods of estimating the mean levels of random coincidences in the raw data may have different tradeoffs in accuracy and precision. Finally, the manner of using the estimates to remove the effect, for example, simple subtraction *versus* incorporation in the system model, will also impact the bias and variance of the estimate of the uncontaminated raw data.

Several of the effects in Table 1 are characterized by the National Electrical Manufacturers Association NU-2 standard but only for specific reference objects and not for different FDG distributions in different sized patients. In addition, one notable exception is that there is no specification for tracking or testing global scaling accuracy.

Global scaling is determined by several factors, the most sensitive of which is cross-calibration to a known level of activity of ^{18}F-FDG in a syringe, as measured by a dose calibrator. The FDG is then typically injected into a water-filled cylindrical phantom (20-cm diameter × 20-cm height), which is then imaged and reconstructed. The correspondence between the known activity of the syringe and the commensurate measured activity concentration in the uniform cylinder establishes the global scaling factor for the scanner. There are, however, several weak links in this chain, including manual recording of the values (ideally before and after injection) as well as the time of each activity measurement and manual entry of these values into the scanner data system. It is worth noting that these issues are also relevant for each patient scan, and so errors can and do occur on individual patient scans as well. A related approach is to use a long-lived calibration source, typically a 20-cm cylinder with ^{68}Ge (9-month radioactive half-life) embedded in epoxy. However, the accuracy of the activity level within the cylinder is typically ±10%, and the necessary cross-calibration with the dose calibrator is not established. Because global calibration factors are performed periodically, typically monthly or quarterly as per manufacturers' recommendations, every patient scan during the corresponding calibration period is affected by the new global scaling factor. Figure 2 shows the global scaling for a PET/CT scanner during a 18-month period indicating a miscalibration at one point that induced an approximate 25% shift in overall ^{18}F SUV values.

Table 1. Major Scanner-Specific Factors That Affect Bias and Variance in PET Images.

Global calibration method
Attenuation and correction method
Scattered coincidences and correction method
Random coincidences and correction method
Detector efficiency variations and correction method
Geometrical efficiency normalizations
Intrinsic detector resolution
Sensitivity
Live time as a function of count rate
Image reconstruction method (especially smoothing)
Data processing algorithms, e.g., decay correction
Clock synchronization

Another potentially significant error in global scaling is the dose calibrator itself. Recently, this has come under increased scrutiny with the development of an National Institute of Standards and Technology (NIST)–traceable standard [28]. Initial results indicate an approximate 5% variability within manufacturers and a 10% shift in absolute calibration between the two main manufacturers of dose calibrators [29]. As is the case for the scanner global calibration factor, an error in this value is a direct multiplier in terms of bias.

The intrinsic detector resolution, that is, sampling, leads to the well-known partial volume effect, which causes blurring at edges of regions of different tracer uptake [17]. This effect is increased by smoothing applied during the reconstruction process to suppress noise. One consequence of this effect is that "hot" objects smaller than approximately three times the operating resolution have a decreased level of apparent activity. Conversely, small "cold" regions have higher than true activity levels. Because clinical resolution levels are on the order of 1 cm, this means that objects smaller than ~3 cm typically suffer from this effect. This is illustrated in Figure 3.

The relationship of quantitative accuracy with scanner quality assurance and control (QA/QC) procedures is often not clear. Manufacturer QA/QC procedures are generally designed to track a variety of parameters to catch changes in scanner performance parameters that will (hopefully) detect if there is a change in some scanner-operating parameter, for example, reduced sensitivity and/or increased variations

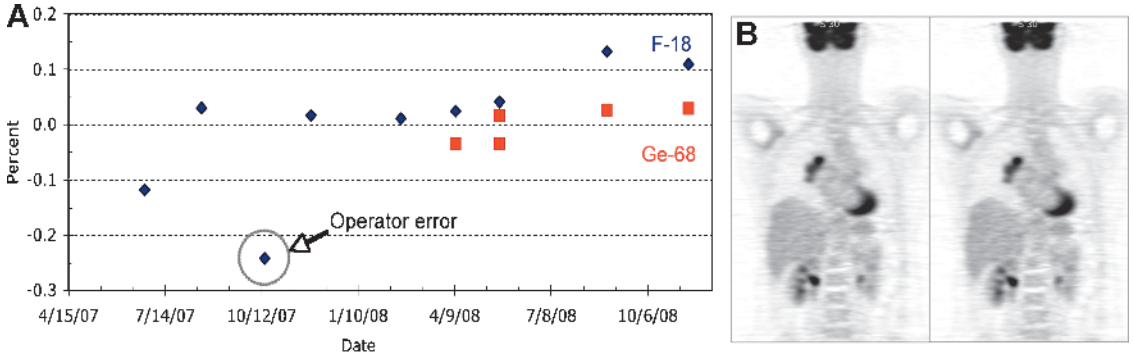

Figure 2. (A) Global calibration factor for a clinical PET/CT scanner during an 18-month period. Indicated is an erroneous calibration factor that was used for approximately one month before detection. Values are indicated for both the standard ^{18}F-FDG calibration method and using long-lived ^{68}Ge sources showing significant variability. (B) FDG PET images using the median (left image) and the outlier (right image) global calibration factors. The images are identical except for an overall scale difference in all SUVs of approximately 25%. This type of periodically vulnerable calibration error has clear consequences for longitudinal studies.

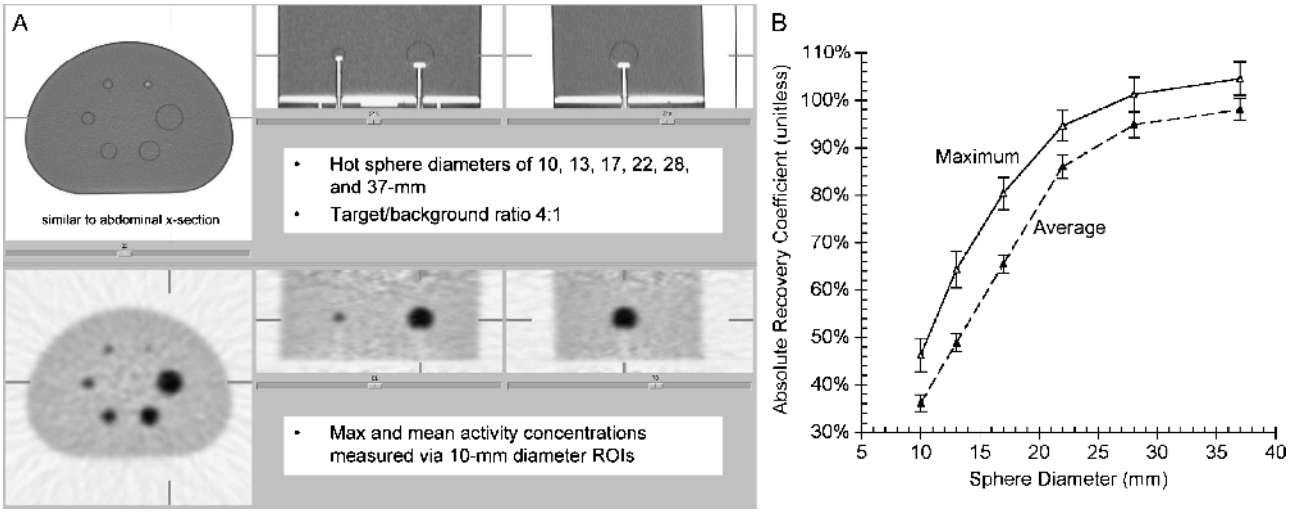

Figure 3. Partial volume and variance effects for a single scanner. (A) CT and PET images of the SNM validation phantom. (B) Absolute recovery coefficient as a function of sphere diameter and method of reporting SUV from an ROI placed over the spheres.

in detector patterns from a uniform source that could indicate failing components such as a detector module. These changes may or may not signify a reduction of image quality. One factor that is generally not tracked, however, is quantitative scaling or accuracy. This arguably important issue has so far received little attention.

Patient and protocol components of bias and variance. There are several effects related to the patient and the imaging protocol that can potentially increase bias and variance in the measurement of FDG uptake. Four of the more recognized factors are 1) the lesion size (i.e., susceptibility to partial volume errors), 2) tumor heterogeneity (e.g., if a tumor has a necrotic center and FDG uptake is localized to the rim), 3) elevated blood glucose levels (which reduces FDG uptake), and 4) the uptake time between FDG injection and scan acquisition, as tumor FDG uptake generally increases for several hours. These factors have been well summarized in several discussions (e.g., [30,31]). It is also possible to include image reconstruction protocols in this category, although we have included it in the prior category because the choice of parameter settings is coupled to the manufacturer-dependent algorithm.

Image analysis methods and impact on bias and variance. Several methods are used to extract SUV values from the reconstruction images. Typically, a region of interest (ROI) is drawn around a lesion or area of interest. The method of drawing can vary widely. Typically, two-dimensional ellipses or circles are drawn in one or more transaxial images, but three-dimensional shapes can also be used. The size of the ROI can be determined manually, for example, based on the PET image and/or the CT image, or set by a thresholding or segmentation procedure. The range of available methods varies among vendors. Once the region is set, then the values within the ROI can be considered as a histogram of SUVs, and one or more derived values can be used to characterize the distribution. Commonly used methods are the average and maximum SUV values as well as the "peak" value, which is the average of a fixed, size ROI (e.g., 15 mm diameter) positioned to obtain the maximum average value while also including the pixel with the maximum value. The most commonly used method for clinical imaging is SUV_{max}, owing in part to its ease of measurement, reproducibility, and reduced sensitivity to partial volume errors. For the purposes of measuring response to therapy, however, there is not yet a consensus on the best method [13,32]. In addition, there have been no efforts to date to test the reproducibility of SUVs measured on different vendors' display stations.

RIDER Supported Investigations into Sources of Error for FDG PET/CT SUVs

Society of Nuclear Medicine Validation Phantom

The Society of Nuclear Medicine (SNM) commissioned a "validation phantom" for comparing quantitation between different PET imaging centers and scanner models [33]. The phantom was filled with ^{68}Ge (271-day half-life) in an epoxy matrix, so that precisely the same object was imaged on different scanners, with the decay of ^{68}Ge being the only difference between scans. In addition, the true activity levels were known to within 10%. As a separate project, the phantom was repeatedly rescanned on a GE and Siemens PET/CT scanner [34] and on each of the three PET/CT scanner manufacturers (GE, Philips, and Siemens) [35]. Several versions of 20-repeat coffee-break experiments were performed. Images of the phantom and illustrative results are shown in Figure 3. Many of the data sets have been uploaded to the National Cancer Imaging Archive (NCIA)–RIDER database in a manner similar to the patient images described below as resource for comparing measurements made from patient images.

The plot in Figure 3*B* illustrates several findings. The first is the reduction in recovery coefficient (measured SUV/true SUV) with decreasing sphere diameter, which is caused by the partial volume error and image smoothing. The second is the difference between SUV_{max} and SUV_{mean}, showing that SUV_{max} is somewhat less sensitive to partial volume error, but that there is also a significant change in bias between the two analysis methods. Third, the coefficient of variation (determined from 20 repeat scans) is relatively small, albeit larger for SUV_{max}, owing to the larger statistical fluctuations in that metric given that it is based on only one pixel.

AAPM/SNM TG145 Calibration Phantom

The SNM validation phantom was not feasible for widespread use owing to cost, weight, and high levels of radioactivity. In an attempt to simplify the process, the Task Group 145 "Quantitative PET/CT Imaging," sponsored by the American Association of Physicists in Medicine (AAPM) and the SNM (AAPM/SNM TG145), devised a modification of the qualification phantom used by the American College of Radiology (ACR). In this case, four adjacent cylinders with diameters ranging from 8 to 25 mm were filled with ^{68}Ge/^{68}Ga in an epoxy matrix, significantly reducing cost, weight, and total radioactivity levels in the package sent to participants. To date, this phantom has been imaged at 10 different PET centers and analysis of the intersite variability of SUV measurements is ongoing [36].

Longitudinal Scanner Calibration Variations

To evaluate longitudinal scanner calibration variations, we evaluated several years worth of scanner calibration data [37]. The results (e.g., Figure 4) indicate 6% to 10% variability if no quality assurance procedures are applied to the calibration process itself. If quality assurance procedures that catch operator errors are applied, then the variability drops to 3% to 4%.

Monitoring the calibration process, however, does not include errors from dose-calibrator measurements, which are used in every patient scan. Recent results indicate that there is approximately a 7% error from dose calibrator measurements for ^{18}F [29]. As part of the RIDER project, we have recently commenced on a multicenter longitudinal measurement of combined scanner and dose-calibrator effects using the new NIST standard for ^{68}Ge as a reference for ^{18}F assays [28].

RIDER PET/CT Images

Under institutional review board approval, 30 sets of clinical PET and 35 sets of PET/CT images of lung cancer patients with one or more follow-up scans were selected. The software application Field Center (and later, the Clinical Trial Processor) developed by the Radiological Society of North America (RSNA) was used to deidentify and transmit the images to the NCIA image database, now called the National Biomedical Imaging Archive. Sample patient images are shown in Figure 1.

During verification that image sets could be uploaded and downloaded, two issues unique to PET/CT were resolved. First, for PET/CT studies, the ability to retrieve concurrent PET and CT image sets was added to the NCIA. Second, the deidentification process had to be adjusted to not remove critical timing and other information that is used for decay correction of the FDG activity levels and SUVs. The necessary information for SUV calculation is listed in Table 2.

Table 2. Public DICOM Header Fields Necessary for Proper SUV Calculation.

DICOM Tag (Group, Element)	Field Name	Used For
(0008,0031)	Series time	Decay correction
(0008,0032)	Acquisition time	Decay correction
(0010,0040)	Patient sex	SUV lean body mass
(0010,1020)	Patient size	SUV lean body mass
(0010,1030)	Patient weight	SUV
(0018,1072)	Radiopharmaceutical start time	Decay correction
(0018,1074)	Radionuclide total dose	SUV
(0018,1075)	Radionuclide half-life	Decay correction
(0054,1102)	Decay correction	Decay correction
(0054,1300)	Frame reference time	Decay correction
(0054,1321)	Decay factor	Decay correction
(7053,1000)	SUV factor	SUV (Philips only)

Note that these are relevant for images acquired by GE or Siemens PET/CT scanners. Other necessary information (e.g., tracer activity) is contained in private fields that are determined by each manufacturer. Philips PET/CT scanners use the private field (7053,1000): SUV Factor only.

As a final step, a subset of the uploaded images were downloaded and analyzed for relative changes using a commercial PET/CT display and analysis package (PET Volume Computed Assisted Reading; GE Healthcare). There were 10 patients with 20 lesions with two or more serial scans. Lesions were tracked from scan to scan using display and analysis software, and several change metrics were recorded. These included lesion volume (from the PET image using a 50% threshold), average SUV, maximum SUV, maximum and average SUV relative to liver SUV, and total lesion glycolysis, also called metabolic volume, which is defined as (average SUV) × volume [38]. The changes for each metric for all 20 lesions are given in Figure 5.

Effects of Bias and Variance in Determining Response for Clinical Trials

Proposed values reported in the literature for the minimum percentage change in PET that reflects true biologic changes between serial FDG-PET measurements range from the 15% to 25% recommended by the European Organization for Research and Treatment of Cancer PET Study Group [39] to 20% [25] to 25% [23] to 50% [40] to 55% [41] and others. A revised definition of these guidelines including

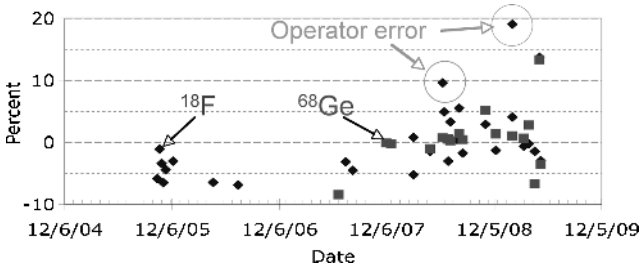

Figure 4. Scanner calibration factors (relative to mean value) during a 4-year period. Data were collected with manufacturer-recommended procedures using ^{18}F in a water-filled cylinder (diamonds) and the same ^{68}Ge in epoxy cylindrical source.

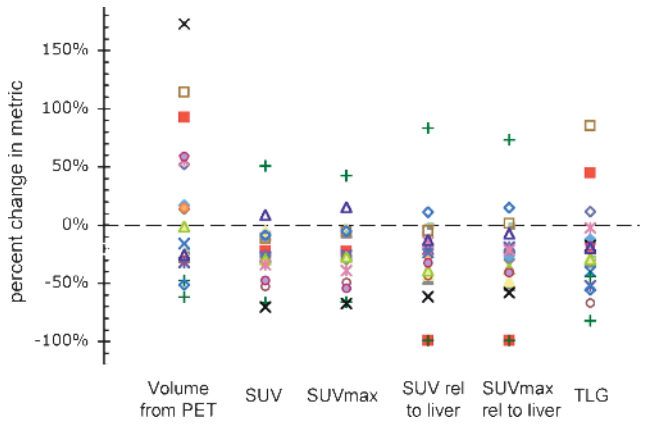

Figure 5. Sample calculation of change metrics from serial PET/CT image sets illustrating a potential use of the RIDER collection. Each symbol represents the change values for an individual lesion for the difference metrics.

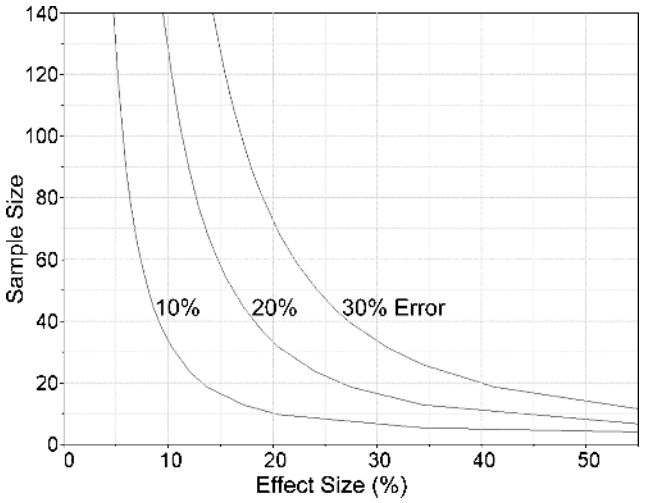

Figure 6. Illustration of sample size *versus* minimum expected effect size for different total noise in measurements for a test power of 80% and significance level (α) = 0.05 (adapted from Doot et al. [47]). Total noise includes, for example, biologic, local calibration, and multicenter measurement effects.

a 30% threshold has recently been proposed by Wahl et al. [13]. There are several constraints on the previously published minimum thresholds for changes in PET metrics, but the issue relevant for the discussion here is that these guidelines were based on the best information at the time of their development, and that they do not necessarily apply to longitudinal studies *or* multicenter trials, where the variance is greater. This is shown in Figure 6, which illustrates the impact of overall PET measurement error on required sample size for a range of expected effect sizes (e.g., ΔSUV). These results were calculated for two-sample *t*-tests and designed to yield 80% power at a significance level (α) of 0.05 [35]. As the noise level doubles from 10% to 20%, the needed number of samples increases more three times for trials that anticipate small effect sizes. These results indicate that if the true levels of variance are underestimated (e.g., owing to increased variance from multicenter comparisons), a study runs the risk of not identifying a significant result because of being underpowered.

Another important consideration is the sensitivity of the selected PET metric to the anticipated true effect change that is provided to the designer of the clinical trial based, for example, from a nonimaging laboratory experiment. Changes in SUV measurements have been reported to be less sensitive than changes in dynamic PET metrics in patients with low baseline SUVs, that is, below 5-to-1 tumor-to-background ratios, in breast cancer patients [42,43]. This reduced SUV sensitivity for some patient populations is due in part to the SUV metric's inability to separate metabolized FDG from unmetabolized background FDG, unlike more sophisticated PET metrics derived for kinetic analyses of dynamic PET scans [9,25,44–46]. The reduced SUV sensitivity impacts trial design by reducing the observed (or measured) effect size and thereby increases the required sample size needed to adequately power the clinical trial [35,47].

Determining the true variance of individual patient measures of response to therapy has not been completed. There are several effects discussed here, such as dose-calibrator calibration, intrascanner and interscanner variability, and longitudinal variability that will also affect multicenter studies. In addition, the effects of intraobserver and interobserver variability and display/analysis workstation concordance across multiple sites have not been determined. Several of these studies are underway as part of the RIDER project or sponsored by the SNM, AAPM, American College of Radiology Imaging Network (ACRIN), and RSNA. In particular, the RIDER project is now focusing on collecting data on multisite longitudinal variability of scanner and dose-calibrator calibration factors.

Initiatives

In addition to the RIDER project, there are several initiatives specific to the use of quantitative PET/CT imaging for response to therapy and for clinical trials. These include (but are not limited to) the publication of ACRIN PET core laboratory procedures, the recent NIST standard for ^{68}Ge, the RSNA Quantitative Imaging Biomarkers Alliance, the joint AAPM/SNM Task Group 145: Quantitative PET/CT Imaging (TG145), the SNM Clinical Trials Network, the NCI-funded Imaging Response Assessment Teams, and European initiatives. These groups are working toward the common goal of determining what is known, what remains to be determined, and what physical standards are feasible for evaluating response to therapy and for clinical trials.

Conclusions

There is growing evidence that FDG PET/CT measures of response to therapy, even early in the therapeutic regime, are linked to patient outcomes for several forms of cancer. As well, PET/CT using FDG and other tracers has been established as a useful tool for mechanistic effects in drug discovery studies. Given the evolving overall picture of the bias and variance in single-center and multicenter imaging studies, however, current studies run the risk of being underpowered. There are several critical areas where bias and variance and the sensitivity of selected PET metric to true change should be estimated and included in initial trial design, including multisite longitudinal variability of scanner and dose-calibrator calibration factors, as well as data analysis methods and effects of variations between display/analysis workstations.

Acknowledgments

The authors thank the support of Larry Clarke and Barbara Croft from the NCI Cancer Imaging Program, Alexander McEwan from the Society of Nuclear Medicine, members of the joint AAPM/SNM Task Group 145, Paul Christian from the Huntsman Cancer Institute at the University of Utah, Joel Karp from the University of Pennsylvania, Fred Fahey from Harvard Medical School, and Cate Lockhart from the University of Washington.

References

[1] Therasse P, Eisenhauer E, and Verweij J (2006). RECIST revisited: a review of validation studies on tumour assessment. *Eur J Cancer* **42**, 1031–1039.

[2] McNitt-Gray MF, Bidaut LM, Armatto SG III, Meyer CR, Gavrielides MA, McLennan G, Petrick N, Zhao B, Reeves AP, Beichel R, et al. (2009). Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error. *Transl Oncol* **2** (4), 216–222.

[3] Vansteenkiste J and Dooms C (2007). Positron emission tomography in non–small cell lung cancer. *Curr Opin Oncol* **19**, 78–83.

[4] Hicks R, Lau E, Alam N, and Chen R (2007). Imaging in the diagnosis and treatment of non–small cell lung cancer. *Respirology* **12**, 165–172.

[5] Brundage M, Davies D, and Mackillop W (2002). Prognostic factors in non–small cell lung cancer: a decade of progress. *Chest* **122**, 1037–1057.

[6] Weber WA (2009). Assessing tumor response to therapy. *J Nucl Med* **50** (Suppl 1), 1S–10S.

[7] Coleman RE (2002). Is quantitation necessary for oncological PET studies? For. *Eur J Nucl Med Mol Imaging* **29**, 133–135.

[8] Spence AM, Muzi M, Graham MM, O'Sullivan F, Link JM, Lewellen TK, Lewellen B, Freeman SD, Mankoff DA, Eary JF, et al. (2002). 2-[(18)F]Fluoro-2-deoxyglucose and glucose uptake in malignant gliomas before and after radiotherapy: correlation with outcome. *Clin Cancer Res* **8**, 971–979.

[9] Lammertsma AA, Hoekstra CJ, Giaccone G, and Hoekstra OS (2006). How should we analyse FDG PET studies for monitoring tumour response? *Eur J Nucl Med Mol Imaging* **33** (Suppl 1), 16–21.

[10] Warburg O (1930). *The Metabolism of Tumors.* London, UK: Constable Press.

[11] Mankoff DA, O'Sullivan F, Barlow WE, and Krohn KA (2007). Molecular imaging research in the outcomes era: measuring outcomes for individualized cancer therapy. *Acad Radiol* **14**, 398–405.

[12] Weber WA (2006). Positron emission tomography as an imaging biomarker. *J Clin Oncol* **24**, 3282–3292.

[13] Wahl RL, Jacene H, Kasamon Y, and Lodge MA (2009). From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med* **50**, 122S–150S.

[14] Zasadny KR and Wahl RL (1993). Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-deoxy-D-glucose: variations with body weight and a method for correction. *Radiology* **189**, 847–850.

[15] Kim CK, Gupta N, Chandramouli B, and Alavi A (1994). Standardized uptake values of FDG: body surface area correction is preferable to body weight correction. *J Nucl Med* **35**, 164–167.

[16] Wahl RL (2004). Why nearly all PET of abdominal and pelvic cancers will be performed as PET/CT. *J Nucl Med* **45**, 82S–95S.

[17] Soret M, Bacharach SL, and Buvat I (2007). Partial-volume effect in PET tumor imaging. *J Nucl Med* **48**, 932–945.

[18] Alessio AM and Kinahan P (2006). PET image reconstruction. In *Nuclear Medicine* (2nd ed., vol. 1), RE Henkin, D Bova, GL Dillehay, JR Halama, SM Karesh, RH Wagner, and AM Zimmer (Eds.), Philadelphia, PA: Mosby, Inc.

[19] Kinahan PE, Townsend DW, Beyer T, and Sashin D (1998). Attenuation correction for a combined 3D PET/CT scanner. *Med Phys* **25**, 2046–2053.

[20] Kinahan PE, Hasegawa BH, and Beyer T (2003). x-ray–Based attenuation correction for positron emission tomography/computed tomography scanners. *Semin Nucl Med* **33**, 166–179.

[21] Meyer CR, Armatto SG III, Fenimore CP, McLennan G, Bidaut LM, Barboriak DP, Gavrielides MA, Jackson EF, McNitt-Gray MF, Kinahan PE, et al. (2009). Quantitative imaging to assess tumor response to therapy: common themes of measurement, truth data, and error sources. *Transl Oncol* **2** (4), 198–210.

[22] Krak N, Boellaard R, Hoekstra O, Twisk J, Hoekstra C, and Lammertsma A (2005). Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur J Nucl Med Mol Imaging* **32**, 294–301.

[23] Minn H, Zasadny K, Quint L, and Wahl R (1995). Lung cancer: reproducibility of quantitative measurements for evaluating 2-[F-18]-fluoro-2-deoxy-D-glucose uptake at PET. *Radiology* **196**, 167–173.

[24] Nahmias C and Wahl L (2008). Reproducibility of standardized uptake value measurements determined by ^{18}F-FDG PET in malignant tumors. *J Nucl Med* **49**, 1804–1808.

[25] Weber WA, Ziegler SI, Thodtmann R, Hanauske AR, and Schwaiger M (1999). Reproducibility of metabolic measurements in malignant tumors using FDG PET. *J Nucl Med* **40**, 1771–1777.

[26] Nakamoto Y, Zasadny K, Minn H, and Wahl R (2002). Reproducibility of common semi-quantitative parameters for evaluating lung cancer glucose metabolism with positron emission tomography using 2-deoxy-2-[^{18}F]fluoro-D-glucose. *Mol Imaging Biol* **4**, 171–178.

[27] Boellaard R (2009). Standards for PET image acquisition and quantitative data analysis. *J Nucl Med* **50**, 11S–20S.

[28] Zimmerman BE, Cessna JT, and Fitzgerald R (2008). Standardization of ^{68}Ge/^{68}Ga using three liquid scintillation counting based methods. *Res Nat Inst Stand Technol* **113**, 265–280.

[29] Zimmerman B, Kinahan P, Galbraith W, Allberg K, and Mawlawi O (2009). Multicenter comparison of dose calibrator accuracy for PET imaging using a standardized source. *J Nucl Med* **50**, 123.

[30] Shankar LK, Hoffman JM, Bacharach S, Graham MM, Karp J, Lammertsma AA, Larson S, Mankoff DA, Siegel BA, Van den Abbeele A, et al. (2006). Consensus recommendations for the use of ^{18}F-FDG PET as an indicator of therapeutic response in patients in National Cancer Institute Trials. *J Nucl Med* **47**, 1059–1066.

[31] Weber W (2005). Use of PET for monitoring cancer therapy and for predicting outcome. *J Nucl Med* **46**, 983–995.

[32] Benz M, Czernin J, Allen-Auerbach M, Tap W, Dry S, Elashoff D, Chow K, Evilevitch V, Eckardt J, Phelps M, et al. (2009). FDG-PET/CT imaging predicts histopathologic treatment responses after the initial cycle of neoadjuvant chemotherapy in high-grade soft-tissue sarcomas. *Clin Cancer Res* **15**, 2856–2863.

[33] Kinahan P, Doot R, Christian P, Karp J, Scheuermann J, Zimmerman R, Saffer J, and McEwan A (2008). Multi-center comparison of a PET/CT calibration phantom for imaging trials. *J Nucl Med* **49**, 63P.

[34] Doot RK, Christian PE, Mankoff DA, and Kinahan PE (2007). Reproducibility of quantifying tracer uptake with PET/CT for evaluation of response to therapy. *IEEE Nucl Sci Symp Conf Rec vol. M12-8 Honolulu, HI*, 2833–2837.

[35] Doot RK (2008). *Factors Affecting Quantitative PET as a Measure of Cancer Response to Therapy [dissertation].* Seattle, WA: Department of Bioengineering, University of Washington.

[36] Fahey F, Kinahan P, Doot R, Snay E, Thurston H, Kocak M, and Poussaint T (2009). Variation in PET quantitation within a multi-center consortium. *J Nucl Med* **50**, 140P.

[37] Lockhart C, MacDonald L, Alessio A, McDougald W, Doot R, Lewellen T, and Kinahan P (2009). Minimizing instrument calibration error to reduce the effect of variability on PET/CT SUV measurements. *J Nucl Med* **50**, 61P.

[38] Larson S, Erdi Y, Akhurst T, Mazumdar M, Macapinlac H, Finn R, Casilla C, Fazzari M, Srivastava N, Yeung H, et al. (1999). Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. The visual response score and the change in total lesion glycolysis. *Clin Positron Imaging* **2**, 159–171.

[39] Young H, Baum R, Cremerius U, Herholz K, Hoekstra O, Lammertsma A, Pruim J, and Price P (1999). Measurement of clinical and subclinical tumour response using [^{18}F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Organization for Research and Treatment of Cancer (EORTC) PET Study Group. *Eur J Cancer* **35**, 1773–1782.

[40] Dose Schwarz J, Bader M, Jenicke L, Hemminger G, Janicke F, and Avril N (2005). Early prediction of response to chemotherapy in metastatic breast cancer using sequential ^{18}F-FDG PET. *J Nucl Med* **46**, 1144–1150.

[41] Schelling M, Avril N, Nahrig J, Kuhn W, Romer W, Sattler D, Werner M, Dose J, Janicke F, Graeff H, et al. (2000). Positron emission tomography using [^{18}F] fluorodeoxyglucose for monitoring primary chemotherapy in breast cancer. *J Clin Oncol* **18**, 1689–1695.

[42] McDermott GM, Welch A, Staff RT, Gilbert FJ, Schweiger L, Semple SI, Smith TA, Hutcheon AW, Miller ID, Smith IC, et al. (2007). Monitoring primary breast cancer throughout chemotherapy using FDG-PET. *Breast Cancer Res Treat* **102**, 75–84.

[43] Doot RK, Dunnwald LK, Schubert EK, Muzi M, Peterson LM, Kinahan PE, Kurland BF, and Mankoff DA (2007). Dynamic and static approaches to quantifying ^{18}F-FDG uptake for measuring cancer response to therapy, including the effect of granulocyte CSF. *J Nucl Med* **48**, 920–925.

[44] Keyes JWJ (1995). SUV: standard uptake or silly useless value? *J Nucl Med* **36**, 1836–1839.

[45] Huang SC (2000). Anatomy of SUV. Standardized uptake value. *Nucl Med Biol* **27**, 643–646.

[46] Freedman NM, Sundaram SK, Kurdziel K, Carrasquillo JA, Whatley M, Carson JM, Sellers D, Libutti SK, Yang JC, and Bacharach SL (2003). Comparison of SUV and Patlak slope for monitoring of cancer therapy using serial PET scans. *Eur J Nucl Med Mol Imaging* **30**, 46–53.

[47] Doot R, Kurland B, Kinahan P, and Mankoff D (2009). Considerations for using PET as a response measure in multi-center clinical trials. *J Nucl Med* **50**, 140P.